The University Of Sheffield.

DOCTORAL THESIS

Detecting New, Informative Propositions in Social Media

Author:
Nigel Dewdney

Supervisor:
Prof. Robert GAIZAUSKAS

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

The University of Sheffield
Faculty of Engineering
Department of Computer Science

September 19, 2018

Declaration of Authorship

I, Nigel Dewdney, declare that this thesis titled, "Detecting New, Informative Propositions in Social Media" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

Abstract

The ever growing quantity of online text produced makes it increasingly challenging to find new important or useful information. This is especially so when topics of potential interest are not known a-priori, such as in "breaking news stories". This thesis examines techniques for detecting the emergence of new, interesting information in Social Media. It sets the investigation in the context of a hypothetical knowledge discovery and acquisition system, and addresses two objectives. The first objective addressed is the detection of new topics. The second is filtering of non-informative text from Social Media.

A rolling time-slicing approach is proposed for discovery, in which daily frequencies of nouns, named entities, and multiword expressions are compared to their expected daily frequencies, as estimated from previous days using a Poisson model. Trending features, those showing a significant surge in use, in Social Media are potentially interesting. Features that have not shown a similar recent surge in News are selected as indicative of new information. It is demonstrated that surges in nouns and news entities can be detected that predict corresponding surges in mainstream news. Co-occurring trending features are used to create clusters of potentially topic-related documents. Those formed from co-occurrences of named entities are shown to be the most topically coherent.

Machine learning based filtering models are proposed for finding informative text in Social Media. News/Non-News and Dialogue Act models are explored using the News annotated Redites corpus of Twitter messages. A simple 5-act Dialogue scheme, used to annotate a small sample thereof, is presented. For both News/Non-News and Informative/Non-Informative classification tasks, using non-lexical message features produces more discriminative and robust classification models than using message terms alone. The combination of all investigated features yield the most accurate models.

Acknowledgements

This thesis would not have been possible without the help and advice of many people. I would like to extend my thanks to my supervisors Rob Gaizauskas, Yorick Wilks, Mark Stevenson, and Louise Guthrie for putting and keeping me on the right path. Thanks also go to those with whom I've had the pleasure of working and have helped me along the way: Rachel Cotterill, Adam Joinson, Kate Muir, Simon Jones, Miles Osborne, David Guthrie, Boyan Onyshkevych, Samantha Lintott, and Rob Fellows. I would not have been able to undertake this thesis without the support of my friends and loving family, especially Carolyn; wish you were here.

Contents

List of Figures

List of Tables

List of Abbreviations

AI	Artificial Intelligence
FSD	First Story Detection
GATE	General Architecture for Text Engineering
ICWSM	International Conference on Weblogs and Social Media
IDF	Inverse Document Frequency
IE	Information Extraction
IR	Information Retrieval
LVC	Light Verb Construct
MWE	MultiWord Expression
NE	Named Entity
NPMI	Normalised Pointwise Mutual Information
PMI	Pointwise Mutual Information
SVM	Support Vector Machine
TDT	Topic Detection and Tracking
TREC	Text Retrieval Evaluation Competition
URI	Universal Resource Indicator
URL	Universal Resource Locator
VPC	Verb Particle Construct

In memory of John Dewdney 1927-2011, and Carolyn Dewdney 1960-2013

Chapter 1

The Systematic Acquisition of New Knowledge

1.1 Setting the scene

Consider a world in which billions of people are communicating with one another via the medium of online social media. There are multiple reasons for their communications, from simply chatting with one another, through to informing their readership of events they have observed or are involved in. In such a world one could consider these authors as noisy sensors of what is going on. Reading the messages produced may give the reader insight into new events and yield new facts and beliefs. The gathering of intelligence from what is written is an activity found in various domains. A popular image of intelligence gathering is it is a national security function of governments, but it is also to be found in businesses (such as in market assessment), emergency response co-ordination, and journalism. In all these cases there is a need to distill relevant facts from non-curated text, and there is far more material than can be read and analysed by those people engaged in such activities.

The internet contains a wealth of information. In its early days it exhibited rapid growth, which was curtailed somewhat by the "dot com crash" in 2001 and 2002, although the conclusion that such growth is continual may be contentious as the diversity of publishing avenues has increased from simple web-site hosting to encompass various types of content publishing services[1] . What is not so open to argument is that provided content is changing all the time (Brewington and Cybenko, 2000). Not only are there web sites dedicated to providing information on particular topics with reference to people, organisations, places etc. such as journals, papers, and news stories, but also forums, newsgroups, and social network site pages. In these website services, individuals constantly add to the information (and possibly mis-information) on the web.

A significant downside to this growth and diversity in online written content is that there is simply too much material produced to be read, much of which is unlikely to be of interest or relevant to the reader. Search capabilities based on

[1] See www.caslon.com.au/metricsguide.htm for an overview of trends in the Internet

key words and document relevance ranking, often employing relevance feedback (Ruthven and Lalmas, 2003), page rank (Langville and Meyer, 2005), or a combination thereof (Bharat and Henzinger, 1998), provide satisfactory retrieval of documents on a specified subject. However, what if the subject of interest is not well specified? Without the anchors of search keys, how might the vast quantity of text be cut down such that potentially important new information about the world may more easily be discerned?

Information here may be taken to mean that which permits the reader to conclude some fact about the world from a piece of text. That fact may be *explicit*, i.e. directly stated by the text, as in "Alice met Bob". A reader may also infer information. For example, on reading "Celia was also there with Bob", the reader may infer that Alice also met Celia. We may consider this information to be *implicit*.

Information may be considered to be *new* if it permits the reader to establish facts not previous known by the reader. Being potentially new information for a reader is not in itself sufficient to make it newsworthy, however. To be reported in news media, one would expect the information to also be of interest to the readership. For this one might expect the information to have some importance, some potential consequence(s) for the audience. For large mainstream news media organisations, or even smaller specialist domain interests, one would also expect that the audience would be fairly large in number. Global news events, for example, may impact thousands if not millions of people, and therefore are important and of interest to those people.

Information expressed in textual documents could result in a range of potential interest across the potential readership, and its interestingness to that readership may be characterised by whether or not the readers establish new knowledge. The latter becomes less likely as the age of the information increases. Figure 1.1 illustrates these two dimensions. One may also imagine a third dimension, one of importance, which could be described as the scale of consequence for the reader given the new information. One may assume that trivial facts are not important and therefore not of interest. "News" in this thesis, then, should be taken to mean new information that is likely to be of interest to a wide readership. News *events* are those things that happen in the world that are considered to be important enough by reporters to report them.

In the domain of news events, it is easy to find reports of current events as the news has already been published: there are plenty of well known websites that report the news. But news aggregation is not without its dangers. Such is the success of simply republishing stories, that in some instances journalists have found it hard or too demanding to validate their information, as illustrated by the example of the "Black screen of death" story regarding a seemingly fatally flawed security patch from Microsoft, November 2009. This story was widely reported in online technology news sites, but only after several days, and following checks by Microsoft, did

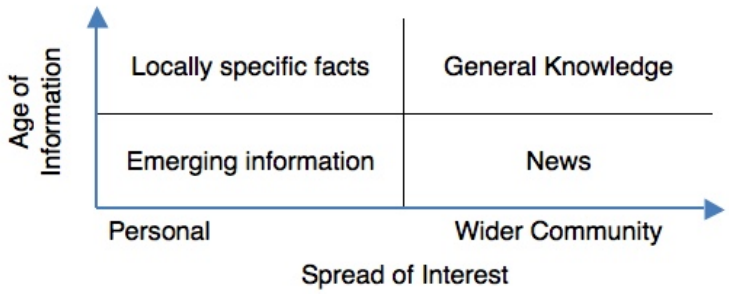


FIGURE 1.1: Newness and interest in Information

it emerge that the story originated from a single blog entry posted by a little known computer security company. Microsoft, which had not been contacted until after the story had been published, found that the patch was not at fault at all.[2] Another potential issue with news aggregation, but not one addressed in this work, is the potential for its manipulation that some commentators have remarked upon.[3].

One might envisage a journalist as an example of the type of user interested in automated detection of new information. One in search of the scoop on the next big story would be interested in the novel: pieces of information that belong together within a domain of interest, that are not well associated already (although parts of the picture may be). There are stories which consist of editorial reproductions of what has appeared in an informal write-up, but these are not likely the source of a scoop. In a similar way, a press release, or a publicised event is unlikely to be the sole source of a unique story for the investigative journalist. In each of these cases the information is already there explicitly for the journalist. All he or she has to do is write it up as a news report. The interesting cases are where the journalist has to piece parts of the story together from multiple sources where the connections may be implicit rather than explicit. For example, different blogs by several people with jobs in drug companies stating that their authors have quit or are looking for alternative employment may be evidence for increasing dissatisfaction with working conditions in the pharmaceutical industry.

The process of establishing that there may be a story to be told is one of collating pieces of related information that have some significance collectively. The facts may be statements that have been made a significant number of times independently but not (yet) reported in the press. Alternatively, the interest may lie in implicit connections between independent pieces of information, as in the establishment of a link between magnesium deficiency and migraine although no such connection had been explicitly made (Swanson, 1988). Another possibility could be viewed as a

[2]Fittingly the history of this episode is reported in a blog:
www.zdnet.com/article/what-the-black-screen-of-death-story-says-about-tech-journalism/
[3]A blog commentary on this potential danger:
www.zdnet.com/article/a-troubling-new-form-of-media-manipulation/

meta-informational basis for a story; that is, aspects of the publishing of connected information rather than the information itself.

A specific example of this phenomenon occurred following the attacks on the World Trade Twin Towers and the Pentagon building in 2001. At first the news stories centred on the attacks themselves. However, in public fora, newsgroups etc. speculation arose about responsibility for the attacks, i.e. conspiracy theories began to emerge, often purporting the U.S. authorities to be the true responsible party. This "angle" on the events of September 11^{th} 2001 then became a story in its own right, i.e. the emergence of conspiracy theories. For example, within two weeks of the attacks the BBC reported on this in an online article "Why we need conspiracy theories"[4], prompted in part, according to the article, by enquiries made by the public to the BBC as to the validity of conspiracy claims.

What are sought after, therefore, are informational patterns present in collections of text corpora that have not been reported in traditional news outlets (or other established reference sources). One could view a process of discovering such patterns as data mining "unstructured" information. Previously undetected patterns would constitute new connections and therefore new information (at least as far as widely accepted knowledge, here represented as that reported in the mass media).

Does such information exist outside of curated news reports? Although, as noted above, much information comes from direct reports of events, press briefings, or from editorial re-issues, there are cases where information of interest to the mass media organisations emerges from other sources before it is picked up and published by them. A study of social media and news coverage covering August and September 2008 looking for emergence of phrases found a few such cases (Leskovec, Backstrom, and Kleinberg, 2009). For example there is a news story about Clarion financing the DVD distribution of the film "Obsession: Radical Islam's War against the West". A significant proportion of the discussion surrounding this story, about 38%, occurred in blogs over a week before the story was reported in the mainstream media. Consequently, the question arises as to whether the emergence of this story could be automatically predicted *before* the mass media picked it up. One particular field that attracts considerable discussion where information and speculation can pre-empt a news story (in part because the the area is known to be of interest) is that of popular sport. In the UK professional football is just one example where discussion surrounding team performances is common place in web fora and blogs. Poor performance will often result in heated debate and continuation of such may lead to managers being sacked, which naturally is a news story. An example of this can be seen in the run-up to Liverpool FC's sacking of their manager in January 2011: speculation was already rife by October 2010 about Roy Hodgson after a poor start to the season. A couple of stories were published with Hodgson publicly expressing

[4] see http://www.bbc.co.uk/1/hi/world/americas/1561199.stm

indignation and the revelation of a £3 million sack clause in his contact in mid October, but it is the amount of speculation later in the year that is interesting. Using Google with the search 'Liverpool "sack Hodgson"' one finds about 495 results for November 2010. Some of these are comments on a media company's website, but the lead story is about a match, not the manager's position. The result is similar for December. In Janauary the results for each day of the first two weeks are shown in Figure 1.2. The first mass media news story picked up was published on the 6^{th} with the 'Daily Telegraph' reporting that many fans were demanding Hodgson go; the 'Guardian' emphasising the lack of official comment from the club. Hodgson was sacked on the 8^{th}. On this day we see the postings peak with mass media contributions, including Reuters', being published. What is of note though is the rise in numbers of postings prior to the event that could lead one to predict it (as was done by some following the match on the 6^{th}). Finding that the concept of "sack" was increasingly used in connection with "Liverpool" and "Hodgson" could potentially lead one to discover this would be a distinct possibility and certainly that an increasing number of fans felt it should have been. It is interesting to note here that the information detected is the amount of belief in an imminent sacking, not the sacking itself, whereas the mass media story is of the actual sacking.

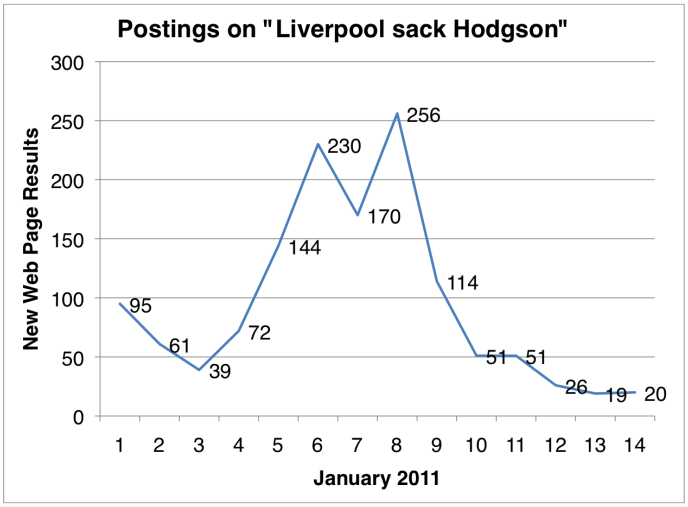


FIGURE 1.2: Google search results by day in January 2011 for "Liverpool sack Hodgson"

Going beyond this, could implicit connections (potentially via chains of relationships) of interest to the press or other investigative bodies, be found automatically? This second question is a harder one for consideration as cases would involve collating information deduced from making (or *discovering*) connections that were not explicitly stated beforehand. It is only after the case (or the consequences thereof) that one can see the evidence was there. The investigator first reporting the new information is the agent of its creation in explicit form. Such knowledge discovery may not occur before consequential events expose the facts though. Information discovery in this context may be viewed as a forensic activity.

Recent high profile examples of stories where indicative information was found to be available before the story broke include the collapses of Enron and Lehman Brothers. Could automated analysis of public communications have predicted that Enron's position was fraudulent, and that the complicity of Andersen consultants would unravel? Could the credit crunch have been predicted? There were commentators who were predicting trouble ahead beforehand, but financial forecasting is notorious for its variety in sentiment and success. Was there anything different in published material that would differentiate the impending crisis from normal negative sentiment?

Techniques for discovering new information through finding patterns of implicit connections between documents would therefore be useful not only for finding potential new stories, but also for detecting abnormal or suspicious patterns in published information. Further, such techniques could therefore be of assistance in forensic examination of textual evidence. Discovery of implicit information was not attempted in this work, the focus being on explicit propositional information. However accumulation of explicit information could be a useful precursor, potentially through chaining common subjects and objects in the form of a graph, say.

Given, then, that new stories and new angles on events emerge, yielding new information, could one construct a retrieval / filtering method that will allow the discovery of new information of interest? A method that with access to the Internet could have allowed one to predict the emergence of the 9/11 conspiracy theories as an interesting story in its own right (beyond the usual noise of those that believe a government, U.S. or other, is behind some atrocity) for example?

One may consider the eventual objective to be the systematic accumulation of knowledge about the world given some domain of interest where recent and current additions are likely to be of most interest. Information would be constantly arriving in the form of textual documents, reports and messages. The user of such a system would be able to quickly identify new relevant information – propositions which were not previously in the knowledge base but related to key entities and concepts.

The relationships between entities and concepts stored in the knowledge base could connect them in the form of a graph. Analysis of this graph could enable the discovery of new information by means of detecting new relational links, and chains thereof. A complete system would extract propositional information from text and populate a knowledge graph, i.e. a graph of connected facts and hypotheses about the world, such that the graph could be used to answer questions. Questions could be about specific entities in the graph, but they could also seek new facts (or assertions) that have been added to the graph since a given point in the graph's history.

It is typically desired that a knowledge base or knowledge graph contains verified facts. However in a developing situation it would also be useful to include assertions such that competing propositions may be compared. Such propositions may

be assertions in arguments for and against some other proposition. In this case the graph may be used as the basis for abstractive summarisation and comparison of people's arguments in stances towards some issue.

Another example application in which facts may not be substantiated but in which accumulation for analysis would be useful is in emergency response co-ordination, where the situation may be rapidly evolving with media outlets, whether social or mainstream, providing multiple "sensory" information. The combination of this information may provide the basis for analysis and decision making about where best to focus relief activities and what those activities need to be.

In situations where information is sought from social media there is a need to reduce the noise in the information that would form the input. There are two main reasons. Firstly there is simply the need to maximise the efficiency of the system by minimising the amount of data presented to the system that is not intended to inform readers. Secondly there is the need to reduce the chance of erroneous information that may arise from processing noisy data. A processing step that identified passages of text that contained propositions, as distinct from those that perform other pragmatic intents or dialogue acts, would therefore be useful as a preprocessor. For example, compare the two following sentences:

> "Acme Industrials has opened a new factory in mainland China."
>
> "Could Acme Industrials expand its operations to tap into the market China offers?"

The first sentence makes an explicit statement, referring to both subject and object. However, the second sentence does not express information, rather it poses a question, and its presence in a discovery system could result in unwarranted attention in the subject concerned. Its removal could improve accuracy in statistical methods for detecting significant trends, and reduce processing in argument and predicate mining for knowledge base and knowledge graph population.

In the overall system envisaged, Social Media is considered to be a source of new information, but one that is very noisy. Including text feature based filtering for explicitly informative messages in the processing flow would increase the likelihood that the filtered text imparts something useful. Established sources and already accumulated knowledge would provide a background of expected references to the world of interest, and facts relating them. Unexpected references might reflect reports of new events or facts. Comparison of references in Social Media to the background could, then, be used as the basis of a 'novelty' filter. The output of this stage would yield the text more likely to contain novel information which could then be promoted to the user and passed to an information extraction stage for ingestion into the knowledge base. The full flow is illustrated in Figure 1.3. This thesis focuses on potential filtering methods in such a system, prior to proposition extraction and ingest into a knowledge base.

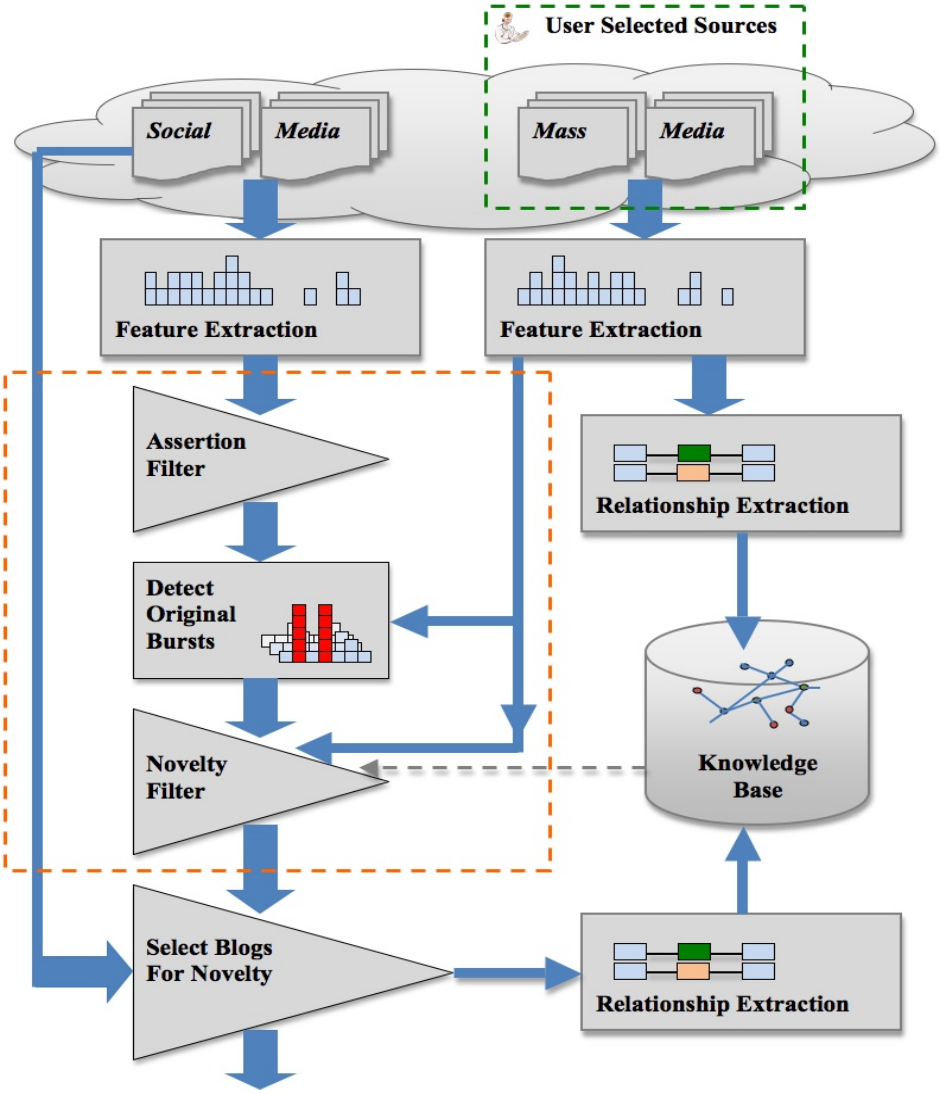


FIGURE 1.3: Detecting new information in a knowledge acquisition system. (Dashed red box outlines thesis focus.)

A question arising from the scenario presented is what constitutes *new* information? What may be new to one user might already be known to another. (Indeed it is presumably known by the author of its written statement!) As Feng and Allan, (2007) notes, in accumulating information on some topic, people remember what they have already ascertained – it is necessary for that task! They are only really interested, then, in messages with information new to them (or providing further evidence for them with which to establish their knowledge). An automated knowledge accumulation system is unlikely to have the same prior knowledge as any one user. However, in an unlimited system, one could envisage a system accumulating knowledge such that statistically what was most new for most users could be estimated, using the mainstream media as a source for what was well known. It could also employ user feedback to increase accuracy and enable 'tuning' to specific user domains.

Another challenge with distilling facts about the world from communications is that human languages allow information to be transmitted in multiple different ways. In discourse one may refer to an established entity indirectly. The context should provide the receiver with sufficient information to resolve who or what the referent is. Co-reference resolution seeks to anchor references to entities, but this is not a focus of the work here. It is assumed that in addressing a wide audience an author will make clear that which is being referred to through necessity, and that resolving arguments to propositions may be carried out as part of subsequent predicate processing.

1.2 Objectives

This thesis seeks to contribute in two ways towards the development of automated knowledge acquisition from social media: (1) by addressing the need to filter out messages that do not contain propositional information, and (2) by proposing a method for detecting potentially new assertions. The veracity of information declared, and the identification of entities to which the information relates, are beyond the scope of this work, being challenging problems in their own right.

The envisaged knowledge discovery and acquisition system shown in Figure 1.3 provided the context for the work carried out. The focal points for the investigation in this thesis are outlined by the central dashed box. Given that, in order to discover what is new to a user, there needs to be some representation of what the user already knows, the expectation is that a user of the system would specify what they know. This could be through some formal representation (e.g. current knowledge graph), a corpus of documents, or a combination thereof. Assuming mainstream news contains information considered to be of wide interest, it was decided to use documents from these sources as a proxy for a user's knowledge. This also has the advantage of being a growing source of information, analogous to a user's expanding knowledge.

As the thesis revolves around finding new interesting information in Social Media, corresponding example source documents were required for comparison with news media. Social Media comprises any media provided by individuals for sharing amongst and communicating with any interested parties. For this thesis, though, it was decided to restrict the source type to blogs as one might expect these to be generally intended to inform a wide readership. It was decided to restrict further to Microblog messages for filtering experiments, as these naturally lend themselves to posts with singular intents.

The first objective of the study was to investigate whether or not it was possible to find documents containing new interesting information through a statistical model of nominal references in the source texts. The second objective was to find out if interesting stories were more likely to be found if references were refined to include just mentions of named entities and concepts. To meet these objectives it was

decided to construct statistical models and use them to prioritise subjects and documents for inspection, investigating how well the models predicted information later found as "interesting" (assuming news to be a reasonable proxy). The hypothesis here is:

$H1$: Some documents containing new information can be found through an unexpected number of references to named entities and concepts.

$H1_{null}$: References to named entities and concepts are no more frequent when related new information emerges than the average rate of mentions.

The next objective was concerned with reducing the amount of processing that would be required for, and reducing the potential noise inherent in, the source material for the discovery system. One might expect the types of words used, and the way in which that are used, would help indicate that a message is intended to impart information, rather than perform some other function (such as posing a question). Features corresponding to these aspects of a message are those other than the surface, lexical, form of the words used. Extracted features of text other than the constituent words themselves, can be characterised, then, as being *non-lexical*. Could an effective filter model be constructed from extracted features to select sentences that are explicitly informative? If so, are such models more effective than those constructed from words alone? To answer these questions, it was decided to employ machine learning methods to build models based on the diverse collection of features extracted form the source text. The main hypothesis tested in developing the approach follows:

$H2$: Sentences asserting information may be distinguished from those intended to perform other functions using non-lexical features.

$H2_{null}$: Non-lexical characteristics of a message carry no information on whether or not that message conveys an explicit statement.

An automated system would require features upon which to operate. Feature extraction from source documents, both from the Social Media that is the focus for this work, and the news that is a proxy for a user's knowledge, was therefore a requirement for the investigation. Text may exhibit many features and there are correspondingly many tools and approaches for processing documents to find and measure them. In this work it was decided to focus on words likely to be references to subjects and objects – things that authors would mention when asserting facts. The rationale for this was that one would need to tell an audience what the information is about and what it is in relation to! Therefore it was necessary to identify and extract mentions of named entities, and nouns in general, from the source texts.

One may observe that the envisaged system also postulates a filter to select only the text that makes assertions. It was expected that textual features regarding the

syntactic roles words played, rather than just the words themselves, would be required to distinguish the intent to inform from other purposes in a sentence. Further shallow syntactic feature extraction was also needed for the study.

As described above, experiments to establish whether or not the proposed methods meet the objectives herein required a collection of features to be extracted and calculated for the input texts. Largely it was possible to find and use established tools. However it was found that no suitable tool to extract multiword nominal references was available at the time of the study. A further objective, therefore, was to construct a feature extractor for multiword expressions.

The objectives of this thesis, then, were to advance approaches to select and discover text in Social Media that conveys new interesting information. To meet these objectives, the goals were to understand whether or not features of text, other than the words used, carry useful information for these purposes.

1.3 Contributions and Publications

During the course of work carried out for this thesis, software was developed, data annotated, and analysis carried out. Some of this work has been published and made available to the wider community. Work was carried out to develop a Multiword Extraction tool in support of feature extraction, described in Chapter 5, and reported in Dewdney, (2017a). The tool itself is available from `https://github.com/NDewdney/extraction`. Some of the results from the analysis of Named Entities and Nouns in Weblogs and Online News described in Chapter 6 were reported in Dewdney, (2012). Some results from the discovery processes analysed were described at CiCling 2015, as reported in Dewdney, (2015). A paper reporting on experiments contributing to the exploration of News filtering was published in Dewdney, (2017b).

1.4 Structure of the thesis

The rest of this thesis is organised into three main parts. The chapters comprising the first part provide the background and set the context for the research. The main body of the work is then presented, detailing the investigations undertaken and the results obtained. Finally a review and conclusion of the thesis is given. Appendices follow.

Chapter 2 provides the relevant computational natural language processing background for the research undertaken for this thesis. Chapter 3 describes the relevant work undertaken in related fields prior to and up to the present day state of the art. Chapter 4 outlines and provides motivation for the approaches taken in the investigation. Chapter 5 describes the methods selected for detecting and extracting the features used in the experiments carried out. It also provides an evaluation of the tool developed to extract Multiword Expressions from Social Media.

Chapter 6 provides an analysis of the use of nominal references in online blogs and mainstream news media. Chapter 7 returns to the theme of news, examining the systematic detection of propositions and information of potential news value in microblogs.

Concluding, Chapter 8 provides a formal summary of the thesis and a discussion of remaining challenges and promising future directions in knowledge discovery.

1.5 Summary

This chapter has introduced the problem of discovering new interesting information from within an ever-increasing amount of online text. Curated reporting and informational articles now form only a fraction of this material; Social media, including weblogs and microblogging, serve more purposes than informing a general readership. Yet important and useful information may be found in amongst this content. If the object to which new information should be relevant is known then searching by keyword(s) is well established; the problem arises when one doesn't know what to look for.

Journalistic research was introduced as an example where information discovery can be useful. The creation of news stories was described as a collation of connected pieces of relevant information, either on or about a related aspect of the central topic. Examples were given that show that sometimes this information exists in online text prior to the creation of the news story. The text may alert the reader to interesting facts, although the significance may only be discovered after a subsequent event, such as the collapse of the Enron corporation. It was argued that automated discovery of potentially significant information could be valuable to journalists, but also in other scenarios involving intelligence gathering such as in disaster response.

A knowledge acquisition system was envisaged in which pieces of information were extracted from text and connected in a knowledge graph. However the challenges that such a system would face include the sheer volume of online text, whether or not the intention behind the text is to inform the readers, and the veracity of imparted information. Filtering streams of online text for that which is more likely to impart information about the world, and discovery of potentially significant and interesting propositions, would therefore be useful initial steps towards such a system. This provides the motivation for the research described in this thesis.

Two main hypotheses explored in the thesis, along with supporting hypotheses, were put forward. These were that documents containing new interesting information can be found through an unexpected number of references to entities and concepts, and that sentences asserting information may be distinguished from those intended to perform other functions using non-lexical features.

The next chapter describes the relevant ideas and techniques in text and natural language processing, providing the context for the research described in the later chapters.

Chapter 2

Communication of knowledge through text

In Chapter 1 the idea of an automatic knowledge discovery and acquisition system was introduced. This provides the context and motivation for this thesis. This chapter explores the relevant ideas and concepts involved in how such a system might function. It considers the function of documents, how they are used to impart information to readers, and temporal effects that may be observed as the world, and its documentation, evolve. It then moves on to approaches to systematically extract and represent that information, before outlining some of the challenges posed in identifying factually true information.

The chapter concludes with a summary.

2.1 Concepts, topics, and interestingness

This thesis posits that there is information that will be new and interesting about things in the world to be found in documents produced in Social Media. People write about things to impart information about them, often in messages, articles, and documents. It seems reasonable therefore that a system intended to find new information about things in the world should be able to capture what it is that people have written about. (It is worth noting that they may write messages for other purposes, such as requests for further information, however these are not of interest here.) What people write about, the subject matter, is often considered to have a focus that is thought of as the *topic* of the written text; the things being written about are typically related to the topic. But what defines a topic?

Information regarding a topic could be connected to (about) an item, or a collection of items, real or abstract; it may refer to relationships or events that would involve them, or qualities ascribable to them. In fact anything that one can refer to could have information associated with it by the act of reference in association with something else. Such things are termed *concepts*. (Note that concepts are not necessarily atomic, nor a specialisation of a more general concept, but are often treated as such. For example one may refer to a “Football Club”. The linguistic expression is made up of many constituent parts, but the concept is the club in its entirety itself.)

As one may refer to something by different means we may consider a concept to be an abstract construct, an *idea* of something. For example we could conceive of the concept of "running" – the action, or "a runner" – the actor, definitions that themselves invoke further concepts, i.e. "action" and "actor". Out of necessity of communicating the idea, the concepts here have labels from a natural language (English), but there is the notion that concepts are somehow independent of the language that is used to refer to them[1]. This model of concepts is similar to that of 'topic map' that postulates topics as existential interconnected features in Plato's ideal world (Pepper, 1999).

Entities can be thought of as physical instantiations of concepts, named or otherwise referred to, be they real or imagined. It might not be possible to physically instantiate all concepts, however, although even more abstract concepts, such as "thinking" for example, could be considered as instantiated in actual events.

If a system is to detect information about things from the documents imparting that information, then it will need to have some notion of how people refer to those things. One may refer to one specific thing by name. One may indicate a generic class. One may also refer to a specific member of the class, often by means of a definite article, as in "the man on the left of the group in the picture". Bertrand Russell considered the specificity of references to things in developing his theory of descriptions (Russell, 1905). He divided denoting phrases into three classes: those that denote non-existent things, those that denote one definite object, and those that denote ambiguously. Denoting phrases can pose an issue because the same object can be referred to by different phrases. *Co-reference resolution* is the task to grouping all denoting phrases that refer to the same thing or concept. Russell defines *definite descriptions* as those denote phrases that uniquely specify an individual. Various uses of definite descriptions have been identified, for example see Hawkins, (2015), and Prince, (1992). Classes include: new entity introduction, direct anaphora where the head noun is identical to that of an antecedent description, and bridging expressions in which there is an antecedent reference but head nouns differ. One might expect definite descriptions to be largely anaphoric, but a study of articles from the Wall Street Journal Poesio and Vieira, (1998) found that they were used to introduce new entities about 50% of the time.

A text, or discourse, that imparts information, then, can be said to be about something, and in so doing makes references to concepts which may be specified by name. This thing is often referred to as the 'topic' of the text and is intuitively appealing as it encapsulates the idea of a unified coherent discourse about that thing. Although we are not attempting to identify 'topic' in attempting to find expressions of new information, it is relevant then. Information pertains to something – a concept or set of concepts.

[1] Although, since conceptualisation can be thought of as a set of labels standing for descriptions of the concepts, just how independent of language they really are is open to debate.

Although, there is no standard definition of a topic that can be precisely computed (in their text on discourse analysis, Brown and Yule, (1983) assert that "formal attempts to identify topics are doomed to failure…', for the purposes of the work here we shall want to consider the set of concepts referred to by the text of a document. We are interested in information imparted by documents irrespective of whether that information is pertinent to what they are about. We are not so much interested in finding multiple documents with information on the 'same topic' as finding information linked to some established topic – i.e. a topic expansion. Similarly we are not interested in sorting and labelling documents into some number of 'topics'. We can therefore treat each document as having a set of concepts, and focus our attention on what information is being imparted.

There is a shortcoming with this understanding, however, which is that it does not necessarily capture the focus of a discourse. 'Topic' is usually understood to be the focal concept that the author wishes the reader to understand the text to be 'about'. However, the focal concept itself may be missing from the text, as may be other relevant concepts that are directly related. Consider the following for example:

> *"The sound of leather upon willow, a gentle warming breeze, and a refreshing pint. Everyone has made it here and it is Sunday tomorrow. What could be better?"*

This is readily understandable but could legitimately have numerous topics assigned. Many of the concepts are implied: "leather upon willow" suggests the game of cricket, "a gentle warming breeze" suggests weather many people find pleasant, and the day is Saturday. More information would certainly help the reader to understand what the focus is, but given some knowledge of the world, the reader is able glean information.

Admittedly this example has been contrived to illustrate the point, and many authors, wanting the reader to understand what he or she is communicating, will be explicit in stating what they are talking about. However, from a perspective of efficiency, it is desirable to communicate only that which needs to be communicated: an author will assume the readership to have sufficient knowledge and capability to understand what he/she has written. The author may also have assumed the reader to have specific prior knowledge – a context in which to consider the information presented that allows inference of the focal concept. Missing information is a problem, then, in traditional topic determination, but as we are not concerned with the author's focus this need not concern us unduly.

Another potential issue is that different readers may have different prior knowledge and different information requirements. This will lead them potentially to have a different focus than the author intended. One could view determination of the focal concept in a document as a summarisation process rather than a classification one, and the idea of focussing a summary towards a user's interest is one that has

been explored (Dang, 2006). However, if interesting or useful information is imparted about something other than the focus, it is useful. Determination of focus is not important for the system envisaged here.

The treatment of a document as a'set of concepts' presented above does give us a potentially appealing interpretation of what constitutes 'relevance' in a document, though. We may have a set of things, and relationships between them, that we are interested in – i.e. a set of related concepts, or here a 'topic'. A document would be relevant to this topic if it imparts information that involves those things. It would be potentially interesting if it imparted information that was not already known (the converse being that a document imparting known information gives us nothing new, bar potential affirmation).

Any piece of new information, i.e. that which extends the reader's knowledge of concepts of interest, has the potential to be useful. Whether that new information is interesting itself, or not, is another matter.

What makes something interesting? It would seem that a necessary condition for a piece of text to be interesting is that it imparts information to the reader that was not previously known. This is not a sufficient condition, however, as a document may express information not previously known but not of value to the reader. A second necessary condition, therefore, is that the new information is relevant to the reader's interests. Conceptually we could see the reading of a document containing interesting information as establishing new links between things the reader cares about. We should also note that one piece of information may not be interesting in itself but be a part of a larger set of informative items that when considered as a whole constitutes interesting information. For example, one person leaving a company may not be interesting, but the observation of many others also leaving the company may be significant.

In the field of data mining, which deals with finding information in structured data, Geng and Hamilton, (2006) note that there is no widespread agreement on a formal definition of interestingness. However they suggest relevant criteria for whether discovered patterns could be considered interesting: conciseness, the relative ease with which information can be understood; coverage, how widely the pattern is repeated; reliability of the information; peculiarity, the distinctness from other known patterns etc.; diversity in constituent elements; novelty, being knowledge not inferable from other sources; surprise, being contrary to existing knowledge or expectations; utility, being contributory to reaching a goal; and applicability, being knowledge that can be acted upon. It may not be necessary for all these factors to be present for discovery of new interesting information in "unstructured" textual data, although novelty, the degree to which the information or knowledge is not already represented for the receiver, would seem to be a key factor.

Interesting information may also come from many sources, however. It could be

well reported and therefore easily found. Discovery of this information would be trivial. Here we are interested in discovering interesting information for the user that is novel, i.e. that is only just beginning to emerge and therefore rare. Time of creation and the amount of reference to the information are two important aspects, then, in measuring novelty of information. For our purposes, information that is already reported in the news-media will not be considered novel, nor that explicitly stated in established reference material such as encyclopaedias.

The process of information discovery here, then, is the finding of documents that when viewed collectively give the user previously unknown information about entities and concepts of interest. We assume the known information is that previously reported and/or present in the mainstream news media.

2.2 Speech acts and social setting

As indicated in 2.1, imparting information isn't the only reason someone may write something. A system for discovering new information from Social Media will need to accommodate the various message types that one would expect to find therein.

In communication by natural language the basic unit of the message being communicated is an utterance that carries some interpretable meaning by the receiver given shared context and knowledge with the originator. It performs some communicative function. In Linguistics such utterances are termed *speech acts*. An utterance may be represented by the written word and so bodies of text may be viewed as a collection of (cohesive) speech acts.

One may assume that there is an intent for an utterance. In the analysis of intents it is useful to categorise the intents that people have for their utterances. Various speech acts, or *dialogue acts* [2], have been proposed. Labels researchers have proposed typically reflect the application domain(s) being studied but usually form a small number of stable categories such as 'Question', 'Command', 'Salutation' etc. Note that in linguistic theory speech acts can be characterised on three different levels:

- the performance of the act, termed the 'locutionary act';
- the pragmatic, intended, function, termed the 'illocutionary act';
- the actual affect of the act, termed the 'perlocutionary act', which may not be the same as that intended;

It is the level of illocutionary act that is most pertinent to the research here, though, and in referring to speech acts or dialogue acts, it is the illocutionary level of analysis that is being invoked. Illocutionary acts include such speech acts as requesting, commanding, questioning, stating and so-forth. Clearly defining a complete set of acts

[2] Strictly dialogue acts are specialised speech acts constrained by the dialogue setting, however one may observe that the terminology has become interchangeable as analysis of dialogue has widened from the spoken form.

is not readily achieved, but Searle, (1969) provided an analysis yielding five classes of illocutionary acts. These are:

- directives - acts that are intended to cause the receiver to perform some action;
- assertives - acts that state propositions that the speaker holds true[3];
- commissives - acts that commit the speaker to future acts;
- expressives - acts that convey the attitude of the speaker;
- declaratives - acts that declare some change in view of the state of the world;

A discourse may involve many speech acts. Interlocutors may seek to achieve shared, different or competitive aims through communication. They may adjust their strategy and choice of words in response to their perception of the speech acts of those communicating with them. Of particular interest to this thesis, though, is where a communicant is explicitly stating some information about the world to the recipient(s). Arguably information is passed to the recipient in any of Searle's classes of illocutionary acts, although explicit statements about the world (beyond the speaker) might be expected to be more likely to be found in assertives and declaratives. Rather than focussing on the class of illocutionary act one needs look at how information at the pragmatic level is structured and passed. The theory of Information structure gives a formal description of how such information is packaged within a sentence.

Languages use a variety of different mechanisms to structure and indicate information, but whatever the method there are three basic notions to information structure that describe aspects of the content of an utterance. These are *focus*, *topic*, and *givenness* (their complements being *background*, *comment* and *newness*). The focus of a sentence is that which is being brought to the attention of the listener or reader, the point of emphasis while the background is that which the focus is relative to. Topic is what a sentence is about, its complement, comment, being what is said about the topic. Givenness describes that which is not expressed in the sentence but is assumed to be known for correct interpretation, i.e. knowledge that is common to the interlocutors. By contrast, words that are not given and correspondingly the information they impart, are by definition new (irrespective of whether the listener/reader is actually already aware).

Searle's categories provide a general description and framework for finer grained distinctions found in dialogue analysis. Various distinct dialogue act categories have been proposed and explored by researchers for specific task and domains, As well as these there have also been attempts at creating standard taxonomies. Examples of adopted schemes include DAMSL (Dialog Act Markup in Several Layers), (Core and Allen, 1997), which was created in the 1990s and used for annotation of the

[3] assuming no mal-intent

Switchboard corpora (Godfrey, Holliman, and McDaniel, 1992), and MRDA (Meeting Recorder Dialog Act), developed in the production of the ICSI meeting corpus (Shriberg et al., 2004). The Tilburg DialogBank project provides a hierarchy of dialogue acts and an annotation standard, ISO 24617-2 (Bunt et al., 2012). Given these and bespoke tagging schemes, some have started to investigate transferability of annotation between schemes, e.g. see (Sureka and Goyal, 2010).

A statement, and in particular an explicit statement, would be an example of an assertive; it makes a proposition about something. These speech acts are therefore particularly relevant for knowledge discovery. Declaratives may also be informative. For example "I declare Alice and Bob to be lawfully married" gives the information that Alice and Bob are *now* married but were not prior to the declaration. An expressive can be characterised as being informative about the originator of the utterance concerned. Typically this may involve an opinion or belief, but it could also be a fact, and imply the fact that the originator holds the belief. Expressives, in imparting a stance towards something, could reveal a speakers belief or opinion. This highlights another potential issue in in knowledge discovery which is a potential requirement to separate opinion from fact. One could simply ignore all Expressives to avoid capturing opinions, assuming information about the reporter's stance is not of interest, but an Expressive isn't the only way to impart an opinion. The reporter could assert his or her opinion as fact, as in "climate change is a hoax".

It is possible to analyse dialogue at a yet deeper level. Dialogue acts capture the function of an utterance in a dialogue, but not the intent of the interlocutor in the wider context of the social interaction. Bracewell, Tomlinson, and Wang, (2012) argue that the emergence of computer mediated dialogue in the form of social media necessitates new models. Arguing that these models need to reflect the social intentions of the participants, they propose 11 *Social Acts* to capture a broad range of interlocutor goals. These can be characterised as pragmatic speech acts. They include Agreement, Challenge, Disagreement, Disrespect, Establish Credibility, Influence, Mediate, Relationship Conflict, Solidarity, Support, and Task Conflict. It is worth noting that different speech acts could be used to achieve the intended social interaction.

2.3 Temporal effects in information flow

The source set of documents from which we hope to find new information is constantly changing as new documents are added (and possibly some deleted). Therefore we may expect new information to be become available, some relevant to established topics, and some pertinent to new emerging topics. Correspondingly, we may expect that the characteristics of the information available in documents we have access to to change with time. This section considers aspects of temporal variance in text.

As we are looking for new emerging information, time of document publication is likely to be a key dimension. Topics have been found to have variable periods of interest and occurrence Wang and McCallum, (2006). Indeed, study of words displaying a periodic occurrence component, and those without, has been shown to be useful in generating rules that allow documents to be dated from their content Clough et al., (2002). Any model of term co-occurrence novelty should take into account a suitable history of past occurrence patterns.

The type of source material in which terms occur also varies with time: evidence suggests that discussion of topics in blogs can pre-empt news-stories as well as coincide or lag behind news stories Lloyd, Kaulgud, and Skiena, (2006). We are interested here in identifying cases which would fall into the first category. Temporal effects may be of significance in identifying potential items that could evolve into stories of interest. These may be evident in other effects than in content: Note that bursts of linking activity have been observed in the evolution of the "Blogosphere" Kumar et al., (2003).

We may expect key terms indicative of a topic to demonstrate a noisy evolution in frequency of occurrence. Key terms, and term sets, could be more noteworthy if they grow in likelihood in the input stream, less significant if merely a 'random' occurrence (resulting in decaying likelihood), but may occur in bursts. Increasing observations of sets of terms may provide a selection scheme for prioritising potential story investigation.

The idea of filtering document streams for topics and then tracking their evolution through time ordered documents began in the late nineties, fostered by DARPA and NIST through the Text Retrieval Evaluation Conference (TREC) competitions and Topic Detection and Tracking (TDT) workshops Voorhees and Harman, (2005). Sensitivity to novelty and the evolution of stories will be critical to discovery of news worthy material. Material that has been reported before or is a development of a story already being followed is not likely to lead to a scoop. A counter example may be where a radical new connection is found, but we could consider this as a completely new story with a connection to the previous one rather than an evolution of it. Recognising that stories evolve, development of a technique to detect a new emerging story may have to take into account story evolution to down-grade or explicitly exclude them. Therefore TDT is quite relevant to the problem in hand.

2.4 Representing knowledge, old and new

Having obtained information, how does one know it is new? One could say that new information is that which extends the knowledge of the recipient; knowledge here being understood as that which is established (or at least believed) as true factual information. If one establishes new knowledge through accumulation of information gleaned from documents, it could be said that the information was there and

new knowledge has been discovered through analysis. This suggests that a representation or model is required through which what is already known and what has been newly acquired may be compared. Knowledge representation is a key area to be considered therefore.

The concept of automated knowledge discovery is not a new one. Knowledge representation, and inference thereon, has been seen as a key enabler for computers to understand the world from textual input. Early approaches in Artificial Intelligence (AI), popular in the 1970s, considered that reality could be represented and modelled, although it was argued whether it could be contained within a formal system or not. The task of inferring and gleaning knowledge from text, or "understanding" it, was considered to be one of unambiguously determining the representation of reality described in the natural language.

As interest in A.I. expanded in the 70s, the field started to fragment into sub-fields, each tackling smaller more focussed tasks such as computer vision, machine learning, and natural language processing. Access to increasing computing power and large amounts of data enabled shallow statistical techniques to outperform those that sought to provide outputs based on knowledge. Work became focussed on developing these techniques driven by applications such as machine translation, speech recognition, and information retrieval. McCorduck, (1979) provides a comprehensive history of the early decades of A.I.

In more recent times, researchers have directed their attention to representing knowledge in resources built for specific tasks such as assistance in word sense disambiguation and topic domain identification. These resources for aiding processing of text at the semantic level are examples of an *ontology*. Originally they were largely hand built, but various researchers have and are looking at automatic acquisition of knowledge for these resources, returning in some sense to AI approaches to understanding natural language. (For examples see Brewster et al., (2007); Cimiano et al., (2005); Fortuna, Mladenič, and Grobelnik, (2006); Magnini and Cavaglià, (2000).)

Although the population of a knowledge base could be the eventual aim of the system envisaged here, it is not part of the focus for this thesis. Ideally we wish to find source text that would be likely to result in the user obtaining new knowledge prior to extracting information from the text. This raises the question of whether a shallower, statistically based, technique would be appropriate.

2.5 Statistical approaches

Whereas the knowledge representational approach seeks to parse knowledge from the text and encode it as a network of inter-related facts, statistical approaches apply mathematical models to surface features. In text the features are typically the words appearing in the text. Feature selection may be employed by which words that are thought to be non-informative are filtered out, and key phrases / named entities

identified to group words together, but their significance assessed through their occurrence rather than their meaning. Typically, due to computational complexity, features are treated as independent, even though it is known that this assumption is not valid for words Nallapati et al., (2004); Church, (2000); Sarkar, Garthwaite, and De Roeck, (2005).

To measure novelty of feature occurrence and co-occurrence we will need a method of scoring which takes into account the background norm and time of document production to differentiate documents containing potential new information from the rest. The scoring method will also need to score with reference to the user's domain of interest, i.e. the method must take relevance into account. Various models have been proposed and used for distinguishing topics and matching documents. The problem here is one of finding significant novel information. Presuming that we can extract features that correlate in some way with information content, what might be the most appropriate scoring methods? In other words, what scoring methods should be applied in order to assess whether the features under investigation can be used to determine whether a document set contains novel information that is likely to be of interest to the user?

Driven by the need to improve Information Retrieval (IR), in which the task is to retrieve documents relevant to a user's information need (Gaizauskas and Humphreys, 2000), research has shifted in focus from heuristic models towards various probabilistic models, due in part to their principled approach. One of the key issues in building probabilistic models is determining what the probability of the feature occurring is. Typically this is estimated by occurrences in training data, appealing to the central limit theorem (Freund, 1971). However, this is problematic when the feature is rare and is particularly acute if not all of the features have not been seen in training data. This is particularly an issue with the construction of probabilistic language models. The need to account for features that have not been seen in the data that is used to estimate probabilities is referred to as the "Out-Of-Vocabulary" problem. In a classification problem a zero count for a feature for a particular class would suggest a zero probability for that feature to occur in an example of that class. Should one occur in the future it could not be classified correctly as the probability of the class given the feature is zero! To deal with this problem, estimates of probabilities are derived from "smoothed" counts whereby some small adjustment is made to avoid any zeroes. Standard methods are well documented - see the introduction given by Jurafsky and Martin, (2000) for example.

A suitable smoothing strategy will be required in our system because new text will present new words and word usage, effects that are likely to occur in document streams. Growth of vocabulary and language use with corpus size has been studied, for example see Bhat and Sproat, (2009), and Wilks and Catizone, (2002) has considered the effect in word senses.

2.6 Machine learning

Our envisaged system is to process natural language utterances which may perform different speech acts, as described in 2.2. Selecting particular speech acts may be advantageous for system performance. However speech acts are not readily indicated in utterances. Could a speech act classifier be created to select utterances likely to express explicit information?

Many approaches for natural language processing make use of machine learning whereby model parameters are learnt from features measured in samples of example data, see Bishop, (2006). There are two main model creation paradigms: *unsupervised learning* assumes no prior labels and seeks to find natural clusters in the data, whereas *supervised learning* seeks to match examples to the appropriate labels having been supplied labelled training data. Within these broad classes there are two further distinct modelling approaches. There are *generative models* which assume features are generated by a random process which has a distribution that corresponds to a particular class. Machine learning here is the optimisation of the parameters governing the distributions. The second class of model is the *discriminative model* which assumes class members are co-located in some distribution space and are separated from other classes. In this case machine learning is the optimisation of the parameters governing the boundaries in feature space between classes.

The statistical approaches to modelling classes can be described as generative models in that they start from some principled assumptions of underlying distributions from which observed features are drawn, or randomly generated. Measuring feature usage in large amounts of data allows the parameters of those models to be estimated. Generative models include such techniques as probabilistic approaches described above and Gaussian mixture models. One of the drawbacks of these models can be the amount of data needed to get reliable estimates of the probabilities and correspondingly the parameters of the model. This is one of the reasons that techniques to develop discriminative models have become popular in machine learning applications. From a pragmatic viewpoint, so long as the feature space is sufficiently large and representative enough to separate classes (to some acceptable margin of error) and you have sufficient examples of classes to generate an estimate of feature distribution, a reliable estimate of the decision boundary can be learnt without knowledge of the what the most appropriate distribution function for each class is.

The approach embodied in discriminative modelling is to use feature values to estimate the boundaries between classes in the feature space. These constitute the discriminative models. Typically they are non-linear in nature and do not assume feature independence. Techniques proposed, developed, and evaluated, have a range of complexity and flexibility in their their treatment of feature spaces. Collectively, along with generative modelling, these constitute the field of machine learning.

Whereas supervised learning requires labelled examples from which to learn models, unsupervised learning attempts to discover classes from unlabelled data. Some assumptions about how many classes there may be, or how they many be distributed, are used as the basis by which examples are clustered. However, there is no guarantee that resultant classes will be useful, stable, or even readily interpretable. Unsupervised learning techniques can be useful for analysing or investigating large corpora of data. Supervised learning is typically used where a known set of classes is required into which new data should be sorted and filtered.

Popular supervised generative model machine learning techniques include Naive Bayes which assumes independent feature association with classes in a closed form (i.e. all possible features are accounted for), and Hidden Markov Modelling which assumes that observed features have some probabilistic dependency on an underlying unseen set of states and transitions between those states. Gaussian Mixture models specify a joint probability distribution over observations and labels, combining some number of Gaussian distributions, the parameters of which are learnt from training data. (Given an infinite number of Gaussians any distribution may be formed, although a correspondingly large amount of training data would be required to estimate the parameters!)

Popular supervised discriminative model machine learning techniques include decision trees, artificial neural networks, and support vector machines. Decision trees work by selection of features by discriminative power between classes and separating examples along two branches according to a condition or threshold on that feature to best separate examples between classes. The process is repeated for further features along each branch recursively until classes are separated at leaf nodes or no further features can be selected. Tree branch 'pruning' may be employed to reduce the risk of over-fitting the data. A popular decision tree technique is C4.5 (Quinlan, 2014). Artificial neural networks (ANN) come in many configurations, but are loosely modelled on a simplified representation of nuerons in the human brain. Typically features values are input at nodes. A further layer of nodes calculates a function of the weighted sums of the input layer values. Further layers may also be applied in the same fashion, ending in an output layer representing output values - typically the desired classes. Weights are learnt by updating according to back-propagation of a proportion of the difference between the output values and the desired output values. Learning is stopped once performance improvements are no longer being obtained. Support Vector Machines (SVM) optimise class separation by maximising the margin between class examples created by hyperplanes, specified by their support vectors, in the feature space. A kernel may be applied to the feature space to introduce non-linear separation of features. SVMs have been shown to be a generalisation of perceptron based ANNs (Andras, 2002; Abe, 2005).

Unsupervised techniques range from relatively simple techniques such as k-means clustering, which groups data by distance in feature space to create k groups

through to those that assume underlying generative models such as the popular Latent Dirichlet Allocation (Blei, Ng, and Jordan, 2003) method. Another class of unsupervised learning is auto-encoding which seeks to create a transform such that feature sets for a class example predict themselves under the transform. Word2vec (Mikolov et al., 2013a) is a recent example of growing popularity.

Machine learning approaches, particularly supervised methods, are, then, a promising area to consider for the development of a speech act classifier tuned for identifying explicitly informative utterances in the filtering stage of our envisaged system.

2.7 Text Mining

This next section considers various applications of text and natural language processing techniques that may be related in some way to the objectives of the system envisaged in this thesis, and therefore worth further investigation.

Text mining applications seek to find, extract, collate, and potentially summarise the information being communicated by the authors. The applications typically leverage one or more of the techniques described above, and have often provided the motivation for their development. Text mining applications that have seen significant development in recent years.

Natural Language Processing may be employed at different levels, but typically a text mining application will employ classification techniques to find and collate text containing the information of interest (or not) and extraction techniques to transform and formally represent that information.

2.7.1 Classification of text

Text classification may be applied at many levels for may purposes. One may wish to class words by the function they perform, or classify documents as belonging to some topic or genre. Techniques such as those described in 2.5 and 2.6 are often applied to textual features hypothesised to be indicative or counter-indicative of the desired classes. See Srivastava and Sahami, (2009) for examples.

Text classification applications often involve sorting at the document or message level. This may be as a process prior to further analysis that is dependent on having the correct language or genre identified. Classification may be for selecting documents on desired topics. Classification has not been attempted for just the nature of the documents; attribution of documents to authors is one such application (Ramnial, Panchoo, and Pudaruth, 2016). Recently characterisation of authors from their documents has become of growing interest. Assessing likely age, gender ("Author Profiling: Predicting Age and Gender from Blogs—Notebook for PAN at CLEF 2013"), language proficiency (Petersen and Ostendorf, 2009), and emotional state (Li

and Xu, 2014) are examples which have important applications in social analysis and security. Document classification techniques can play an important role in building models to assist in such assessment.

It is not just high level categories and general characteristics of documents that classification is useful for. Of particular interest here are the characterisations of phrases, sentences and bodies of connected sentences. People have sought to determine many different aspects of these and the types information they yield. Characterisations at the lexical and syntactical level include such aspects as language, word function (part of speech), formality, and grammatical role. Categories at the semantic level include word and phrase senses, entity types, and discourse topic. See Jurafsky and Martin, (2000) for an introduction. Syntactic roles and semantics at the word and phrase level are critical in understanding sentences and therefore in information extraction (see below).

At the pragmatic level, the intended function of a message is of interest. The purpose of a document may be quite complex, but could be characterised as being built up of sentences which have in themselves a functional purpose. Determination of the smallest functions may be useful not only for the identification of those functions but also for building an understanding of the containing document, for example. These are referred to as 'speech acts' and are discussed in Section 2.2. Some speech acts may serve to give information to the reader. Of interest to many are those that express some opinion about, or express some sentiment towards, some thing or concept. Finding and characterising expressions of opinion and sentiment is described in Section 2.7.3.

2.7.2 Information Extraction

The extraction of information from text is strongly related to classification in that the constituent words and the roles they play in imparting the required information need to be identified. However, the association of roles played with the words used yields the information imparted in the text. This may be relatively simple as in *Named Entity recognition* where the identification of a name leads to the extraction of the information that there is a mention of an entity with the that name in the text. This mention may subsequently be resolved to a canonical reference, a process of identity resolution, and linked to other mentions that may be expressed differently or pronominally, a process of *co-reference resolution*. Reference to, and assertion of, events and relationships pose a greater challenge as there are multiple inter-dependent roles played by the words, and there may be many different, potentially complex, ways to construct their expression. For example "Alice married Bob. Bob is father to Celia." could also be expressed as "Alice and Celia's father, Bob, tied the knot!". See Grishman, (2012) for an introduction to information extraction.

Arguably a mention of some entity or concept is insufficient to express information; some related context is required. The information to be imparted is either implicitly or explicitly related to the mention. (This may also be observed indirectly as in the response "yes" to the question "are Alice and Bob married?") A key objective in information extraction is to identify and formally represent what the author is imparting to the reader.

The type of information sought may be focussed on a closed set of relationships or required attribute types. Expression of these may be mediated through particular verbs or adjectives, providing cues from which to automatically parse sentences. More recently research has turned attention to *open information extraction* which seeks to extract *any* relationship along with the predicate's arguments to form a relationship tuple. Unlike traditional relationship extraction which starts from examples of the desired relation expression, Open Relation Extraction (ORE) has no specific relation constraint, and seeks to leverage patterns of expression independently of the relationship itself. See Banko, Etzioni, and Center, (2008) for a detailed comparison of and the tradeoffs between open and traditional relation extraction.

Whether the relationship, or predicate, is from a closed set or not, the formal expression may not be explicitly represented in the source text. References to the arguments, or even the predicate, may be anaphoric. References may be pronominal or use another naming convention. For example "Alice met Bob last year. He proposed two months later." requires co-reference resolution to equate "he" and "Bob", and an implied reference needs to be inferred in order to extract the predicate-argument tuple "Bob, proposed-to, Alice".

Knowledge Bases (KBs) are important resources for many applications requiring facts about the world. The facts may be an end in themselves (as in question answering) or the basis of AI reasoning systems. As applications have scaled the automatic population of KBs from natural language sources has become increasingly important. NIST have run annual challenges in Knowledge Base Population to support the development of systems (McNamee and Dang, 2009; Ji et al., 2010). Predicate-argument mining is then a key technology in systematic understanding of natural language.

2.7.3 Opinion mining and sentiment analysis

The motivating applications for the work in this thesis could be characterised by their need to extract, or "mine", factual information about the world from text, which only some of the sentences function to provide as such. In "fact finding" one would generally desire objective assertions and seek to disregard subjective ones. In Opinion mining the case is reversed: the goal is to find and extract expressions of sentiment towards some entity or concept, or an aspect thereof. Analysis of opinions of people is important in such fields as marketing and politics where feedback and reception to products and messages is required in order to gain insight into what

has been successful and what may need changing in future projects and campaigns. Opinion mining is a closely related field, therefore, but filtering for opinions is not simply the inverse of filtering for facts as will be shown below.

Opinion mining has become increasingly important as companies seek to gain insight into how their products and services are being received, and customers turn to social media to express their opinions. Social media then potentially provides a rich vein of data for those companies and organisations. In reviewing the field of Opinion Mining, Liu and Zhang, (2012) points out that while one may express opinions in subjective sentences, not all subjective sentences express an opinion, and some opinions may be inferred from objective sentences. The following sentences provide examples of such cases.

> (1) My girlfriend bought a Nokia phone yesterday. (2) The earbuds quickly broke. (3) I want a phone with good sound quality. (4) Blackberry made reliable phones.

While sentences (1) and (2) above are objective, a negative sentiment is implied in (2). Sentence (3) is subjective in that it expresses a desire, but it does not express an opinion. Sentence (4) appears objective but expresses an opinion. Opinion detection is not simply a case of separating subjective sentences or clauses, from objective ones, even though that task, studied by Wiebe and Riloff, (2005) and Esuli and Sebastiani, (2006) for example, can be a useful step.

Sentiment analysis is strongly related to opinion mining in that the statement of opinion often involves some expression of sentiment towards that which the opinion is about. As with the expression of factual information, an opinion may be expressed as an assertion, but whereas a fact is objectively verifiable, an opinion is subjective, i.e. is a belief of the opinion holder. There is an overlap, however, as a statement of fact may imply an opinion: "The car broke down" may imply a negative opinion of a car rental company for example. Implicit sentiment expression, as in sentence (2) above, can be as important as explicit expressions of sentiment in automatic detection of opinion.

Expression of opinion may be direct statement of some sentiment toward something, or be understood through implication. This sentiment may be considered to be positive, negative, or possibly neutral if there is no particular associated emotion. This has led some to develop methods for building lexicons of words and phrases with positive and negative associations, e.g. see Kaji and Kitsuregawa, (2007), and Rao and Ravichandran, (2009). Such resources may help to detect implied option, as in sentence (2) in the example above by a negative connotation with the word 'broke'. Development and use of such lexicons is not without potential issues, however, as sentiment strength and polarity may be context dependent. The word 'carpet', for example, may be discovered to have negative connotations in the domain

of hotel reviews - why mention them if not to complain about their state - but this would be unwarranted in most other domains.

Liu and Zhang, (2012) splits the task of Opinion mining into five sub-tasks: identifying the entities, identifying expressed aspects of those entities, identifying the opinion holder (not always the author), classifying the expressed aspect sentiment, and formal generation of extracted opinion. The last step provides data for the mining application which typically will require many opinions to form an overall picture upon which the user may act. For example one review may find "sound quality is poor" with a product, but the product's manufacturer would only be concerned if many customers were of the same opinion.

Fact finding, or arguably objective assertion mining because determination of the veracity of assertions is not in scope here, may be similarly thought of. The tasks of identifying entities and aspects to which the factual assertions apply are the same, but whereas in Opinion mining words and expression with strong emotional connotations may indicate desired information, for factual information sentiment is less likely to be expressed. However presence of opinion does not preclude presence of fact. In sentence (4) in the example above there is the easily inferable fact that Blackberry made phones from the opinion expressed that they were reliable. (One could argue that the implication that the author holds this opinion is also "a fact", but automatic inference of indirect entailments from sentences would require a level of systematic knowledge and reasoning, another area that poses a wide set of its own research challenges.)

2.8 On truth and the rise of "Fake News"

The separation of subjective and objective assertions is, as described above, one challenge, but there lies another issue with information presented as objective fact: is the information given true? Assuring the veracity of information is something that requires confirmation or rejection. In other words, evidence is required to support or refute the stated fact. This may be from accumulation from independent sources, through logical implication from other facts, or through a combination thereof. Why might a stated fact be false? It may be that the author is simply wrong in believing the assertion. Alternatively the author may be seeking to mis-inform for some reason.

Propaganda and the spread of mis-information are far from new phenomena, but the problem of false information has recently received more attention due to the rise of so called "Fake News" online. This genre of document takes the form of the familiar news story but the stated facts are not actually true (wholly or partially). The purpose of such stories may be to amuse or to spread mis-information. The authors may seek to influence sections of the population, and this has become an

issue for many political parties. For example, the issue was particularly debated as to its possible influence in the 2016 US Presidential election (Kucharski, 2016).

How to assess the truth and veracity of stated information is itself a challenging complex problem and is beyond the scope of this work. However filtering assertions may be a useful step prior to any approach to assess the veracity of the imparted information.

2.9 Summary

This chapter has described some of the central concepts and ideas pertinent to communication of information by means of natural language written and presented online. It began from a characterisation of documents intended to impart information about something, in other words a topic, and argued that discussion of a topic necessarily involves reference to concepts and entities. However lexical reference is typically not canonical, may be ambiguous, may be anaphoric or rely on a reader's prior knowledge. However, given that a document is intended to impart information on a topic, the characterisation of a topic as a set of mentions of things and relationships between them is a useful one in that it may be possible to formally identify new information as sets of mentions not previously seen. There are other challenges and limitations with this model though. A system for detecting new interesting information may not be in possession of the reader's prior knowledge or even interest.

Informing a reader is not the only purpose a document may have. There may be different pragmatic intentions, or a collection thereof, in a text. Individual sentences or utterances may said to perform a speech, or dialogue, act. Searle's Speech Act theory proposes a framework and describes five illocutionary acts: Directives, Assertives, Commissives, Expressives, and Declaratives, from which many Dialogue Act class taxonomies have been developed. One may observe that Assertive and Declarative speech acts are likely to be those invoked when imparting "factual" information, although Expressive speech acts may also give information with respect to information about the originator of the particular utterance. In particular explicit statements are an example of an Assertive.

New documents and new topics are being created all the time, particularly through social media. This results in temporal effects in the language seen in the corpus of available documents. New topics emerge and coverage of topics evolve, and interest in them varies. Bursts of activity around topics has been observed as has been emergence of topics in social media. Topic Detection and Tracking (TDT) attempts to leverage these temporal effects, comparing the features of new documents to old ones, to find and follow new topics despite a noisy lexical background.

The discovery of new knowledge implies prior, old, knowledge, This in turn implies some knowledge representation. Formal representation of knowledge and its

acquisition from text has been a key goal of Artificial Intelligence research since it began. But, with increasing computing power and data, shallow techniques overtook knowledge based systems and work upon the latter became more focussed on building domain resources.

Whereas the knowledge representational approach seeks to formally capture facts from parsing text, statistical approaches apply mathematical models to surface features. The challenge in the latter is to find the relevant features and apply appropriate models. As in Information Retrieval there has been a shift from heuristic models to probabilistic ones as more data has become available. However, with probabilistic models there is still the issue of how one estimates feature probabilities when the population is growing and evolving over time. Due to high dimensionality, features such as words, are often treated as independent, even though this is known not to be true. However, statistical techniques are appropriate for discovery of unexpected, or new, effects given the principled approach.

The advent of Machine Learning, and in particular discriminative modelling, has enabled models incorporating inter-feature dependencies to be incorporated though at the expense of some interpretability. Supervised learning techniques have been successful in creating classifiers and taggers, while unsupervised learning techniques have provided methods for corpora and document collection analysis. Various learning techniques have been developed and become popular in NLP applications, including Support Vector Machines and Latent Dirichlet Allocation. The techniques are appropriate for consideration in developing a filter for particular speech acts from text messages in general therefore.

Given the key ideas in information dissemination through documents and the main techniques explored in finding information in document collections, the chapter then went on to introduce the the most relevant text mining applications. Sorting documents and messages into wanted and unwanted classes is a common requirement. Classification techniques based on statistical models and/or machine learning are often employed, not only for document level classes but also at finer levels for messages, phrases and words. These are key components for information extraction systems. As well as being used for systematic understanding of language the techniques have also been used to assist in characterising those who produced it.

Locating and extracting factual information imparted in text is critical for automatic population of knowledge bases, but is a challenging task because there are many ways of referring to the entities and concepts that are involved in the information given. References may be ambiguous or anaphoric, and may rely upon the assumed reader's prior knowledge. Relationships, or more specifically predicate-argument tuples, are the basic currency of knowledge bases. Information extraction may focus on a closed set of desired, canonical, relationships. Alternatively, Open Information extraction systems seek to extract all relationship expressions (suspending any anchoring of relationship types). However the extraction of information is

not within the scope of this thesis.

Facts are not the only information people impart in what they write. An increasingly important application in text mining is in detecting and characterising opinion about things and aspects thereof. As with factual information extraction, analysis of sentiment and opinion is challenging because it may be nuanced and implied rather than expressed explicitly. The sentiment associated with a word or phrase can also be domain dependent. Discriminating fact from opinion and belief may be an additional challenge, therefore, due to substantial overlap in how each may be expressed.

The chapter finished on a note on the issue of mis-information, and in particular the rise of so-called "fake news". Accurately assessing the veracity of assertions is a large research challenge in its own right, and was not considered in the scope of the work reported here.

Chapter 3

Previous work

This chapter will expand upon the relevant areas introduced in Chapter 2 looking at research carried out into them.

The chapter continues to consider research into word usage, temporal aspects and how they may be leveraged for new topic detection. Different words may perform different functions, though. Research into the use of features determined from words, and particularly named entity mentions, are then considered.

Recall that the applications providing motivation for this thesis are those involving the need for automatic detection and extraction of new, explicit, information from text. The chapter first considers work done in developing techniques for detecting and selecting documents, relevant to a user's needs, from continuous feeds, or *streams*. Language, and features thereof, are known to evolve as what is being talked about evolves. This thesis is concerned with leveraging behaviours in the evolution of linguistic features. The next sections in the chapter therefore review research on word occurrence and how it evolves. Statistical and probabilistic approaches to detecting text of interest are then discussed.

It is recognised that informing the reader is not the only purpose a piece of text may have, and that filtering out that which isn't intended to explicitly inform readers, would therefore be beneficial in the envisaged system. Models for filtering text for different aspects have typically been built using machine learning techniques. The most commonly used techniques are therefore discussed before a review of research more specifically focussed on classifying text by its intended purpose.

The chapter concludes with a summary of the main findings relevant to this work.

3.1 Selecting from document feeds

Information retrieval from streams of documents typically presents the problem of not only having to group related documents together (as in retrieval from static corpora) but also the detection of new topics. The task is one of topic detection and tracking (TDT). Various techniques have been applied to keep different texts from a stream that belong to the same story together, and, as a natural consequence, the problem of detecting a new story.

Allan, Lavrenko, and Jin, (2000) note that topic tracking is analogous to the information filtering task in the TREC competitions (Voorhees and Harman, 2005), the difference being that there is a user provided query for the filtering task, whereas some small number of training articles are used in the tracking task. They show that performance in tracking is what one would expect from information filtering experiments, but that performance in new, or first, story detection (FSD) using tracking techniques is poor. Furthermore the authors carry out a comparison of predicted and measured upper and lower bounds on error rates in FSD. The lower bound for false new topic detection is computed from the assumption that all N topics in a corpus have already been seen, so *N-1* topics could incorrectly track a story. The upper bound is calculated assuming that every story in a topic occurs before those in another topic (i.e. no interleaving). Both consider the probability of missing a first story and that of a false detection, allowing upper and lower bounds to be computed for systems given their actual tracking detection error trade-off (DET) performance, as well as the theoretical performance for different numbers of topics. Considering the tracking of just one topic the authors show that to achieve a desirable FSD performance of less than 10% misses and 1% false alarms using actual system tracking techniques (and therefore standard IR text comparison techniques) would require a 20-fold increase therein. The authors conclude that task specific information about how news topics and events are related and defined will need to be applied to achieve significant progress.

Noting Allan's upper bound on FSD performance, Yang et al. have proposed a two stage approach in which a classifier trained to identify broadly defined topics is first applied before a novelty detection algorithm Yang et al., (2002). They also make use of named entities noting that different events in topics such as "airplane accidents" would share keywords such as 'airplane' and 'crash' but that, intuitively, names would be informative. Using transcripts from broadcast news that had been assigned keywords, they filtered documents into four broad topics and then found separate events in each from random sampling. Streaming the data such that earlier events formed training data they found that the topic label was of little use in determining first story, but that use of weighted named entity and term comparisons conditioned by topic significantly improved performance.

Going beyond the flat collection of topics of TDT, Nallapati et al. considered the structure of news stories, modelling them as a sequence of related events to provide a richer organisation of news stories Nallapati et al., (2004). The model required events to be threaded together in analogous way to email threading Lewis and Knowles, (1997). Here the threads are directed links between events where the links are simply temporal ordering or causal. Feng combined the idea of event threading with news story structure analysis (borrowing the ideas of discourse analysis) to create "incident networks" to organise news stories Feng and Allan, (2007). Traditional clustering techniques when applied to the networks were found to perform well at organising the stories into related groups, whilst simulated annealing was found to

be better at identifying the types of links that related stories to one another. These approaches have sought to address the problem of understanding how news stories are structured and relate to one another. The application is one of organising news stories rather than detecting a new on in data streams.

The TREC evaluations have also run a number of novelty detection tracks Soboroff and Harman, (2005). Rather than detect new topics, the aim of the task was to detect sentences that contained new information on a topic where the topic was known, trained or specified by a query. Source material was also divided into that describing events and that which expressed opinion. Approaches broadly used similarity to detect topic relevance and dissimilarity from previous sentences to measure novelty. A range of statistical, machine learning, and deeper linguistic analysis techniques were employed. When looking at novelty detection alone, most systems were found to consider nearly all sentences as novel![1] The conclusion drawn is that the problem is hard due to the small amount of information available in a sentence, while human assessors make fine distinctions (presumably because they have an inherently deep understanding of what information is being imparted).

Gabrilovich, Dumais, and Horvitz, (2004) have investigated the applicability of several distance metrics in finding novel information within news stories in a system called "Newsjunkie". In their application stories from a news aggregation site are clustered into topics, the first article in a topic taken as having the initial information. Subsequent articles within the topic cluster are ranked by distance from articles already read (higher in the ranking). In evaluation, human judges ranked articles for novelty selected by highest smoothed Kullback-Leiber divergence, the highest proportion of new named entities in the number of named entities seen in within the topic, and the next chronologically ordered as a baseline. Both tested metrics were found to do significantly better at satisfying the judges than the baseline although no significant difference was found between them. The authors also looked at applying the system to breaking news updates using a window approach. To find new information on a topic, bursts of articles with a high divergence from the topic to date (within the window) are looked for, having applied a median filter to eliminate a spurious story re-caps. The authors note new articles can be re-caps, elaborations, offshoots, or irrelevant. The last of these was a particular issue for the users. Initial experiments suggested that the application of the proposed technique within the document could be useful for classifying the document into the four identified types.

Franco and Kawai, (2010) have investigated two approaches to detecting emerging news in blogs. They have looked at what they term as "cascades" of topics through the blogosphere by measuring linking evolution and by clustering the content of posts in the ICWSM dataset (Burton, Java, and Soboroff, 2009), a set of 44 million

[1] Note here that we shouldn't expect novel sentences to equate to novel topics: sentences are used to describe a topic and impart information about that topic. Maintaining interest and stylistic differences alone would still suggest that authors will find many ways to impart the same information.

blog posts and news stories provided by Spinn3r.com. They found that by looking at the shape and size of the cascades over time they could predict the class of the blog (general, political activist, spam...) with a high degree of accuracy. For content they used a clustering method based on merging significant sub-clusters built up over shorter time intervals applied to a vector space term model. In their link analysis cycles are removed and diversity of locations of references was found to be a good indicator of real news blogs. For content they measured diversity by considering the number of distinct sub-domains in the post URL. Just by selecting clusters that had a minimum of half the posts in distinct sub-domains they were able to identify genuine news blogs with a precision of 94% at a recall of 88%.

Petrović, Osborne, and Lavrenko, (2010) have investigated first story detection in Twitter micro-blog feeds. Noting that traditional methods do not lend themselves to scaling well for web scale applications, they developed an algorithm based on Locality Sensitive Hashing Datar et al., (2004) for finding near neighbours in term vector spaces. For documents found to be far from other points, declared new stories, the document is compared with the last 2,000 documents using the traditional cosine distance. A new story is declared if the distance is greater than a threshold. In the streaming context, document entries in the hash 'buckets' are aged off once capacity of each 'bucket' is reached, yielding a constant time and space algorithm. However noting that many tweets do not represent genuine new stories, being personal updates or spam, the authors apply a further step of selecting the fasting growing new cluster, or 'thread', of tweets. On TDT data the system's performance was found to be comparable to the UMass system described above. On a Twitter data corpus Hachey and Osborne, (2010), 1,000 fast growing clusters hand-annotated as "Event", "Neutral" or "Spam", the authors found that ranking the output by the number of users in a thread having first filtered low entropy threads out gave a significant improvement in average precision over a random ordering of the threads. As far as Twitter data is concerned, celebrity deaths are found to be the fastest spreading news events.

3.2 Word occurrence and informativeness

In investigating adaptation of language models (where term weights are adjusted by the occurrence in documents under consideration), Church, (2000) found that some words show a much higher likelihood of re-occurring in a document than its overall expectation (frequency) would suggest (p^2), and that these words tended to be what one would consider as content words. Those words that did not show significant adaptation in the model tended to be those one would consider a function word. "Noriega" is a rare term, but if you see one occurrence in a document you are much more likely to see another. Non-parametric methods were used in the study, using history (first half of a document) as training as well as the concept of "neighbours"

in which words that are in the history are treated as a query and expanded by the top k documents from a document retrieval engine.

Sarkar, Garthwaite, and De Roeck, (2005) investigated term occurrence patterns further, using a double Poisson Bayesian mixture model for term "burstiness", to determine "content" words. The authors demonstrate how the ratio of model parameters (corresponding to the probability of the first occurrence of a term in a document and the probability of a subsequent occurrence) can separate functions words from content words that display similar occurrence frequencies. For example the word "said" shows a fairly even distribution in Associated Press data, whereas "Noriega" does not, thus breaking an independence assumption. Interestingly, in their analysis, the authors note that in a document about George Bush, the complete name may be mentioned only with subsequent references being made by only Bush. Thus "Bush" may be considered by this method as a content word whereas "George" would not. This suggests that some co-reference resolution prior to feature selection may be fruitful.

In considering the prediction of rare events such as new words in a growing corpus, Baroni and Evert, (2007) have taken a different approach to adjust models for term bursts in documents. Rather than create mixture models they propose using document frequency for terms rather than term frequency, although computing over the corpus. They accomplish this by replacing subsequent terms within a document by an "echo" token. Standard methods of estimating frequencies and predicted vocabulary size are applied. Evaluation showed improvements in Zipf law based and generalised inverse Gaussian-Poisson (GIGP) models prediction of vocabulary size in larger corpora. The authors note there was still increasing divergence in the predicted value from the true value as the corpus size N grew however.

Kireyev, (2009) has proposed a Latent Semantic Analysis based metric for estimating the 'informativeness' of terms based on the length of their vector following singular value decomposition of the word/document count matrix and their 'growth'[2] with documents containing them. The metric, LSAspec, is compared to Church and Gale's burstiness in correlation to WordNet senses and hypernyms and found to correspond better. The addition of LSAspec term weighting to standard use of LSA in an information retrieval task showed some improvement in performance.

3.3 Evolution of word use

Language, words and how they are used, is known to evolve over time, but even over shorter timeframes the word use may reflect the evolving world and how people consider it. A study by Dalli and Wilks, (2006) showed that given training documents, the year of a news story's production can be systematically guessed with reasonable accuracy (~88% accuracy, avg. error 1.75yrs reported). Their system was

[2]The vector length is normalised by the number of documents containing the originating word

trained on 67,000 news items selected at random from the GigaWord corpus and evaluated on a further 678,924. Processing involved identifying periodic terms (days of the week, months of the year etc.) that display recurring spikes in use and those that do not. A term-period-frequency matrix is used to predict the most probable period for a test document. It is suggested that non-periodic data may be used to optimise the results.

Variety in style and form in an age of user provided content is much greater than in editorially controlled output, so we may expect significant shifts across sources as well as across time. Such domain difference can prove troublesome for NLP tools trained on one domain when applied in another. Dredze, Oates, and Piatko, (2010) have shown how such domain shifts may be detected using a metric based on distribution change detection combined with classification margins. The basis of the method is the allocation of real values to intervals and looking at the probability that a real value sample falls into one of those intervals. Any large change indicates a shift. The authors use the classification margin as the value. As important features, e.g. significant words, disappear from the stream, the classification margin is expected to decrease and therefore a change in its distribution. The authors evaluated their technique on a number of NLP data sets and showed it to be sensitive to domain shifts with a low false positive rate.

Domain shifts may be seen as the result of sudden significant changes in feature distributions, but term usage observed in multiple domains may also *evolve separately* over time: Lloyd, Kaulgud, and Skiena, (2006) found evidence that suggests that discussion of topics in blogs can pre-empt news-stories as well as coincide or lag news stories. In a system called "Lydia" they classified news and blog text into topical areas, and extracted named entities. Thereafter they attempt to co-reference names and resolve geographical references. They then compared the most popular named entities in news and blogs on a mentions-per-day basis and found a surprisingly low correlation in an examination of the 197 most prevalent entities in both blogs and news stories. It transpired that spikes in numbers of entity mentions could be present in one medium before the other. Examples given include "Hurricane Ophelia", mentioned in blogs much more than in news in days before the storm hit, and "Tim Burton" mentioned in blogs in days after the release of one of his films was covered in the news.

Experiments investigating the combination of content features with temporal and explicit document linkage reported by Zhao et al. show fine grain event detection is possible in social media (e.g. blogs & comments), segmenting topics in time as well as from each other (Zhao, Mitra, and Chen, 2007). Their approach first clusters documents by applying a graph-cut algorithm to the fully-connected graph of documents linked by the cosine similarities of their $tf * idf$ weighted word vector. An adaptive time series algorithm is then applied to the clusters to segment them. Finally groups

of similar authors judged by the intensity of their postings under a topic are identified, again clustering using the graph-cut algorithm. Texts corresponding to an event are then identified as those belonging in the temporal topical segments that have an identified group of contributors.

Kleinberg, (2003) applied a generative model to email data (also paper titles). Using an infinite state automaton (actually limited for time resolution) and observations of the time gaps between term-occurrence (email arrival time), he modelled topic message interest as levels of bursty activity. He revealed bursts of presence in key terms with periods of increased density organised hierarchically. These bursts in email topics seem to coincide with the development of topics and projects of interest to the author.

Temporal effects may be of significance in identifying potential items that could evolve into stories of interest. Bursts of linking activity have been observed by Kumar et al. in the evolution of the "Blogosphere" (Kumar et al., 2003). They first establish blogging communities by extracting dense sub-graphs from a graph of hyperlinked blogs. They then use Kleinberg's technique on each sub-graph. Community association with topics is supported by a comparison of the actual blog graph with a randomly grown graph. The latter becomes well connected but does not exhibit the evolution of distinct sub-graph community structures.

Leskovec, Backstrom, and Kleinberg, (2009) looked at the concept of "memes", short phrases, and how they evolved in news websites and blog publication. They observed a typical lag of 2.5 hours between the median news story posting and blogs quoting the meme thereafter. Although the majority of quotation was found to be in blogs, some "meme" transfer from blog entries to news media was observed. Applying some arbitrary thresholds to filter out trivial quotations and those not picked up widely by the news media, the authors found that 3.5% of quoted phrases originated in blogs.

Whether these short phrases constitute "knowledge" and therefore represent the concept of a meme (Dawkins, 1976) or not may be debatable, however reuse of phrases (if not whole passages) is well observed. Text re-use is a standard practice in news reportage using "copy" – text produced by news agency specifically for that purpose. Clough et al., (2002) have investigated whether a news stories dependence on news agency text can be measured. The authors note that reuse may involve rewrites for style purposes – something not accounted for in Leskovec, Backstrom, and Kleinberg, (2009). To accommodate this they tried three techniques in comparing texts: n-gram overlap, greedy string tiling and sentence alignment. Difficulty was found in distinguishing most prevalent derived material. Similar results were obtained with all three approaches leading the authors to conclude that other lexically based similarity metrics will probably not lead to improved classification results.

These studies suggest that out envisaged system should focus on rapid changes in word use while accommodating slow changes.

3.4 Use of named entities in text processing

It is proposed here that mentions of named entities in particular will be useful in detecting information regarding them. Some attention has been given to whether or not their specific use can assist in text processing tasks.

Thompson and Dozier, (1997) have shown evidence that names can be effective in information retrieval. Their approach was to extract and index names separately from other words appearing in the text. Names in queries were compared to the name index with a proximal search in which a second name needed to appear within two words of the first name. Relevance was calculated by through $tf * idf$ calculated separately for words and names. In a test against 724 Federal Case Law documents, marked for relevance, and 38 queries containing names they found significant improvement in mean average precision.

Searching for names has been shown by Saggion et al., (2005) to be a useful concept in searching archives of news stories. Various natural language processing elements are brought together in their system, "CubReporter", to assist the user in finding and summarising background information from news archives. Structured query representations are used to allow different related angles to be retrieved by wildcarding one or more semantically divided elements of the query.

Note that the ambiguity of a token such as "Bush" is not what is being addressed when considering whether names as distinct features can aid in identification of information. But is ambiguity likely to pose a significant issue? Word sense disambiguation is a function within the problem of identifying names in the first place (and has been studied, e.g. by Qiu and Frei, (1993)). The point here is that "Bush" marked as a name has some semantic attachment to the token. Voorhees, (1993) compared retrieval performance from an index of word senses derived from WordNet to a word stem index and found that although some queries performed better with a word sense index, overall the word stem index gave better results. Sanderson, (1994) found that ambiguity only posed a problem for retrieval systems with short queries. Rosso et al., (2004) have also investigated use of word senses in text classification and information retrieval. They also found they did not help to improve performance in the latter task. They hypothesised that this could have been due to relatively long queries.

In developing a technique for text summarisation, Azzam, Humphreys, and Gaizauskas, (1999) suggested that a document be about something – its topic – and that something would revolve about a central entity, although that entity may not be directly mentioned by full name more than once or twice. Their data comprised individual news stories, therefore having single topics. Even so, news stories may

mention more than one entity. Their technique for selecting sentences for presentation as a summary involved forming entity co-reference chains, i.e. groups of sentences than mentioned the same entity, and then picking the longest co-reference chain as the one indicating the central entity of interest.

3.5 Knowledge representation and inference techniques

Our envisaged system may be expected to accumulate knowledge as it discovers new information with which it may extend its knowledge-base. This thesis is concerned with detection of text bearing the information *before* it is extracted, but what can be learnt from work done with formal representation?

The knowledge representational approach attempts to represent facts and relationships directly, infer further knowledge as a result of applying logical rules over the representation, and expand knowledge through extraction of information from input text. Thus in this approach the system has, and builds, a model of the world.

The AI community largely agreed that knowledge needed to be represented though debated whether logical formalism was a requirement in the model. McCarthy and Hayes, (1969) considered that it should, an approach the progress of which he later reviewed (Mccarthy, 1989). Others did not hold with this and pursued other representations. For example, Minsky, (1974) proposed the idea of frames, being a structure holding stereotypical remembered knowledge, and Schank, (1975) argued that only a canonical unambiguous form is required to represent meaning.[3]. Kowalski, (1986), whilst recognising the criticality of knowledge in an AI system, argued that logic is also important. He proposes a separation of concepts from the formalism that manipulates them, i.e. the rules of inference and deduction.

A problem with natural language is that it is not always unambiguous at the lexical level. A deeper parse may be required in order to determine the intended unambiguous meaning, such as in the approach taken by Wilks, (1975) based on semantic primitives. Formulas were constructed that described how such primitives could be combined into explanatory templates. Patterns falling into more than one template allowed metaphor to be handled.

As agreed by many proponents of AI, Lenat and Feigenbaum, (1991) consider a significant body of knowledge as an essential component in an intelligent system, but that knowledge needs only to be locally consistent. They argue that a critical mass of knowledge must be achieved for learning by discovery to accelerate, but that learning alone will not be an effective means for this. The construction a very large knowledge base for natural language processing is something that has been worked on since 1984 in a project called Cyc (Lenat et al., 1990). Work continues on Cyc today with commercial and open source versions made available[4] for details.

[3]This should not be confused with the fact that words from a natural language may be(and have been) used as the labels in those elements.

[4]See www.cyc.com

The structure used to represent concepts, objects, and relationships between them is given the term *ontology*, derived from Ontology meaning a "theory of existence". Mizoguchi and Ikeda, (1996) identify several categories of ontology broadly divided between task-oriented and task independent. They describe different levels of ontology use, largely focused on knowledge bases, noting that most success has been in domain specific applications where models have been essentially hand built. Reed and Lenat, (2002) conclude that the major barrier to adoption will be requirement for someone to input source schema and protocols. They asserted a need for tools to enable those other than skilled knowledge engineers to perform the mapping.

Various research efforts have approached the systematic extension of ontologies. Matuszek et al., (2005) have investigated using a semi-automated process to extend the facts held within Cyc. Queries are generated from concepts in the knowledge base, and translated into English search strings. These are passed to Google and the resultant documents are parsed for potential new facts. Noting that "considerable human participation" was still required in enlarging ontologies, Reed and Lenat, (2002) developed an approach to automatic knowledge acquisition in ontological semantics. They employ a bootstrapping process starting with seed templates and expand through exemplars found in corpora and validated by a human agent. Fortuna, Mladenič, and Grobelnik, (2006) investigated the use of document clustering from a corpus to suggest topics to a user developing a topic ontology, although user feedback suggested that not all words suggested by the system were relevant. Brewster et al., (2007) have applied an iterative, weakly supervised bootstrapping, approach to automatically acquiring information for learning an ontology, using three sets of resources: the ontology itself, extraction patterns, and the source text, and classifiers to mediate each stage.

Have ontologies been used successfully as a basis for discovering new information in a setting such as the one considered in this thesis? Kass and Cowell-Shah, (2006) developed a knowledge based approach for the problem of spotting news events that could affect a company's business. Here an ontological model of the key concepts, entities, events and inference rules representing the implications of these events is applied to news feeds. As with many onological systems, the model is built by a domain expert, and is therefore focussed on defined domain events.

Mcburney and Parsons, (2001) have developed a dialectical formalism for argumentation as a basis for the discovery of chance (risk and opportunity). Assuming agents with incomplete knowledge, they propose the use of a genetic algorithm to generate potential dialogues, but note that a "realistic application will require considerable effort in coding the domain knowledge".

Although a representation of knowledge is useful, then, it would seem that more would be required in order to systematically detect and select information with which to extend that knowledge.

3.6 Statistical approaches in knowledge discovery

There have been statistical approaches to finding new knowledge; notably that which is implicit in indirect links in document corpora. The problem of making connections between seemingly unrelated documents to find new knowledge is one that has seen particular attention in the biomedical domain due as discovery of new connections can lead to new insights and potential treatments Bekhuis, (2006).

Swanson and Smalheiser, (1999) observed that useful information could go unnoticed, even by the creators, and could only be inferred by "considering two (or more) separate articles neither of which cites the other and which have no authors in common". He showed the existence of this phenomenon with three particular cases discovered in the Medline database (approximately 9 million articles circa 1999) (Swanson, 1986; Swanson, 1988). The basis of the approach, implemented in a system called 'Arrowsmith', was to start with a hypothetical conceptual link, e.g. "Migraine" and "Magnesium", and use each as independent queries. The system then looks for other concepts appearing in the results from both queries.

Lindsay and Gordon, (1999) extended Swanson and Smalheiser's approach in several ways: use of Medline abstracts as well as the titles, use of two and three word phrases as well as single words as concepts, and use of term/corpus metrics to determine potentially connecting concepts. They found their lexical statistics could reasonably find previously known intermediate topics in a list generated from the original query. Relative frequency of concepts co-occurring in the intermediate documents, whilst expected to find novel linking concepts, did not always identify unnoticed indirectly linked concepts because those concepts may already be well represented for other independent topics, although this could be alleviated by combining the results of multiple intermediate literatures.

Weeber et al., (2001) considered both the closed discovery process embodied by Arrowsmith and the open discovery process in considering the requirement to assess results. They argue that many citations can result from even the closed process and that NLP techniques could be used to alleviate this. Their method makes use of the Unified Medical Language System (UMLS) metathesuarus for medical typed concepts to give a reduce search space by synonym resolution and irrelevant term removal. The types associated with concepts are used to filter potential connecting concepts accord to the kind of connection wanted, e.g. a dietary connection between treatment and disease. A part-of-speech tagger and parser are used to identify phrases in the source text which are then mapped to the UMLS concepts.

Pratt and Yetisgen-Yildiz, (2003) similarly proposed the use of a knowledge based system, but in combination with a data mining algorithm to rank potential connections. As with Arrowsmith, starting concepts are provided by the user, but the system, "LitLinker", seeks to find new concepts/connections via intermediate documents. The knowledge base is used to identify key medical terms and concepts,

grouping similar terms by finding highly correlated terms. The system found Swanson migraine example and the authors claimed the system discovered 24 more plausible links.

Srinivasan, (2004) investigated weighting and typing of links found in the Arrowsmith approach. The free text is the subject of the search rather than just document titles, but resultant terms are filtered and typed by occurrence in MeSH, a typed hierarchy of 21,000 medical phrases organised by experts in the domain. Assessment of connections is as before left a manual process.

Petrič, Urbančič, and Cestnik, (2007) have applied Srinivasan's technique to the domain of autism. They built an ontology for this domain from documents returned from PubMed (an online medical paper repository) by the query "autis*". They then selected rarely occurring terms that, considering background knowledge, might be useful for knowledge discovery, i.e. for the intermediate concepts. These were searched for in non-autism related PubMed articles. Resultant terms are typed and filtered using the ontology to give the hypothetical connections.

Das-Neves, Fox, and Yu, (2005) propose an algorithm to address queries that are better answered by a small set of documents when no one document answers it explicitly. Called Stepping Stones and Pathways (SSP), it is aimed at questions such as "What is the relation of topics X and Y?" The approach forms search queries from the centroid of the top ten documents resulting from the two queries run for the topics that the user is looking to relate. Documents returned by both of these two subsequent queries are then connected to the original document clusters by shared terms, and ranked.

Jin, Srihari, and Singh, (2008) have taken Swanson's ideas and extended them to documents in the world wide web. Their approach focuses on Named Entities and connections between them. The potential connections are extracted terms that have co-occurred with either entity, weighted by their (windowed) proximity to the entity mentioned in the text of the top ten documents returned by Google by queries for each of the two entities in question. The result is a connected graph of concepts linking the two entities for the user to analyse.

There are two key observation to make from these studies. The first is that the information they seek is a rarely, if ever, stated connection between two concepts, often named entities. The second is that potential relationships are found through finding a higher number of connections through some intermediate factor than might be expected. In this thesis, though, the object is to discover explicit information rather infer connections.

Hasegawa, Sekine, and Grishman, (2004) have tackled the problem of discovering and extracting relationships explicitly stated in text. They note that prior methods for this required significant amounts of text annotation to develop. Their proposed method looked to extract co-occurring named entity mentions and cluster them by

the words appearing in between. (An early example of Open Information Extraction – see Section 3.12.) Using New York Times articles spanning one year and applying a threshold of 30 co-occurrences they were able to demonstrate a relatively high recall and precision in identifying relationships and labelling them using the intermediary words.

Mooney and Bunescu, (2005) examined and developed information extraction engines for assembling human protein interactions indicated in abstracts in the MEDLINE database. They applied standard data mining algorithms to discovery association rules between slots in a template that have been filled by an information extraction engine. The system, called "DISCO-TEX", has also been applied to such domains as on-line resumes and book descriptions.

Another aspect to be considered when investigating approaches to discovering information in corpora is that of evaluation; how does one assess how good a system is at finding new facts? Yetisgen-Yildiz and Pratt, (2009) note that comparison of the effectiveness of corpus based literature based discovery (LBD) techniques, such as the ones reviewed above, is difficult given the different metrics used in each. They have proposed time-slicing whereby the systems are run over documents dated before a cut-off and the out is assessed by whether their discoveries are to be found in documents afterwards. This is an intuitively attractive idea in that one would expect that once something of interest had been discovered, by whatever means, people would write about it. It also suggests that a rolling time-slice window could be a basis for continuous discovery whereby documents in a current period are compared to those published previously.

3.7 Application of probabilistic models

Probabilistic models are appealing due to their principled approach. The results from applying the model are fully explained and the meaning of their likelihood is understood. Furthermore, given un-weighted features that are counted, such models should be optimal in terms of the information available.[5] Guthrie, Walker, and Guthrie, (1994) have shown this to be the case for document classification based on word frequencies. However, probabilities are not typically known, but have to be estimated.

For Information Retrieval, Robertson and Jones, (1976) introduced a term relevancy weighting that is an estimate of the log-odds that document containing the query term is relevant with counts smoothed by simply adding 0.5. A popular basis for approaches seeking to model relevancy of documents from the probability of relevance given the term has been to combine additional factors as introduced in the

[5]Heuristic models implicitly add further information into their models. Term weighting and feature selection appeal to prior class association that may not be easily obtained as a probability. They may also adapt to non-linear interactions between features whereas probabilistic models tend to treat features as independent (or conditioned upon other feature occurrence), or include non-enumerated features.

BM25 formula introduced by Robertson, Walker, and Beaulieu, (2000). Additional weighting factors, such as the within document frequency, are combined by in a mixture model that introduces mixture and smoothing parameters that also need to be estimated, and, as Svore and Burges, (2009) point out, one of the challenges is the determination of these parameters. (They propose a neural network approach using the same inputs as BM25, but in so doing the probabilistic interpretation is lost.)

Use of language models permits more sophisticated treatments to be employed for smoothing counts using more general information where specific information is not available. The need to adjust frequency estimates where there are unseen terms (or term combinations) is known as the "out-of-vocabulary" issue and is one that has seen much work in the speech recognition community where language models are commonly used. An intuitively appealing technique often used here is that of Katz, (1987) whose method of smoothing counts takes into account what data has been observed by "backing off". The probabilities of unseen n-grams are estimated according to the associated (n-1)-gram. The process is carried out recursively.

A popular smoothing method is that of Kneser and Ney, (1995). Observing that backing off to the probability estimated for a less specific distribution introduces a bias towards words heavily conditioned on preceding words, they proposed backing-off to a different distribution, estimating the parameters by one of two non-equivalent methods. The first method assumes the probability of a seen sequence is known (well represented in the training data), and results in a back-off distribution that does not take into account the frequency of the preceding words, only that it has occurred. The second method is based on the leave-one-out method of estimating the probability of an unseen event resulting in an estimate that only take singleton sequences into consideration. Both variations were found to produce models with ~10% less perplexities than the baseline standard back-off probability estimate, resulting in a 5% reduction in word error rate in speech recognition.

Wu and Zheng, (2000) have proposed enhancing Katz smoothing by interpolating information from lower order n-grams with all n-grams. They evaluated their method in conversion of Chinese Pinyin strings to Chinese text, with test text including over 1500 sentences not appearing in training data. The results show an improvement in character error rate from 6.5% to 4.5% for their method over standard Katz smoothing.

Zhai and Lafferty, (2004) showed that a predictive language model could be viewed in Bayesian terms, with a prior distribution having the smoothing function. The prior distribution gives the expectation of how we'd expect counts to be given whatever knowledge we have of the model before seeing any data. Given no knowledge all we can expect is that counts would be distributed equitably under the model. How that distribution, an "uninformative" prior, should be seen as equitable has been the subject of debate, though, with consideration given to properties, desirable or otherwise, of the prior. (The simplest prior is that of the uniform distribution,

as proposed by Laplace in 1812, in which a single count is assigned to each feature: Each observed count is increased by one.) An "informative prior" could be derived und assumptions of equitability given an assumption of a probability distribution function for terms. Mackay and Peto propose the natural prior to use is the Dirichlet function. The prior distribution when combined with the observed data acts to smooth the counts to give the posterior distribution – our estimated probabilities.

Zhai and Lafferty, (2004) compared three smoothing methods for language models as applied to the Ad-hoc information retrieval problem. They investigated three popular methods: The Jelinek-Mercer mixture model that adjusts the counts by a proportion of counts from a background collection; Bayesian smoothing using a Dirichlet Prior; and Absolute Discounting in which observed frequencies are discounted by a constant rather than proportionally when mixing with a background count. They found that performance at the task was sensitive to the methods and the choice of parameters. In particular they found sensitivity to query length where long queries with the presence of common words, vice short keyword queries, saw marked difference in retrieval precision with smoothing parameter. This led the authors to propose a two stage approach using the Dirichlet prior method in estimation of probabilities followed by Jelinek-Mercer smoothing to adapt to the query and inherent "noise" in its verbosity.

Champion, (2008) has developed an argument for the use of an alternative informative prior to the Dirichlet. He observes that a prior should have a mean that gives rise to the main effect (the expected value) with a distribution shape determined by the behaviour of small interactions (i.e. perturbations of the main effect). Assuming exchangeability the interactions should be similarly distributed. He proposes that one should assume exchangeability in the multiplicative odds domain, and that any single instance should of a feature should be equally discriminative between classes if neutrality is to be observed. This leads to a prior Beta distribution with mean ϕ and variance proportional to $\phi^2(1-\phi)^2$. For word frequencies ϕ is itself drawn from a Beta distribution over the vocabulary.

3.8 Machine Learning in Classification and Prediction

Machine learning has been successfully applied to many areas in natural language processing. Particularly relevant areas are in message classification and in novelty detection. Machine learning relies on training data but, as described in Chapter 2, some methods leverage labelled data in supervised training while others have unsupervised training.

Unsupervised learning is not naturally suited to continuous streams of data, however some techniques have been explored, for example to assist in learning latent feature association in topic modelling, Ramage et al., (2009), and have been approximated for such a setting, Ailon, Jaiswal, and Monteleoni, (2009). Using the Recurrent

Chinese Restaurant Process (an analogy that posits ..) Ahmed et al., (2011) proposed a hierarchical variant of LDA to create time dependent topic clusters in domains such as online news feeds. A two level hierarchy is used by Zimmermann et al., (2012) with the application of *fuzzy c-means* clustering on documents seen to date. Fine grain sub-clusters make use of $tf * idf$ weights. A new cluster from the stream is assigned to an existing or held aside depending on a cosine metric threshold. If sufficient "novel" documents be accumulated the collection (or sub-topic collection) is re-clustered. This process was used to detect burst of activity in news topics.

Supervised learning techniques are typically applied where there are defined classes to be assigned. Owing to its simplicity, Naive Bayes remains a popular choice for text classification tasks, e.g. Gamallo and Garcia, (2014) reports on its use for sentiment analysis, and Behl, Handa, and Arora, (2014) on security risks from software bug reports. As described by Liu et al., (2013), its relative simplicity makes it potentially attractive for very large scale applications. The assumption of feature independence is a limitation, however, but one that some have looked to overcome, for example see Rennie et al., (2003).

Decision Trees have similarly been popular in text classification tasks in the past, and are still used as the basis of some investigations, such as that reported in Polat and Güneş, (2009). However, interest has declined, and most recent work using decision trees now use tree ensembles for improved classification accuracy. For example Koprinska et al., (2007) found decision tree ensemble could outperform other machine learning algorithms in email classification. Ensembles have a compute cost although Shi et al., (2010) has proposed an approach to reduce this.

Support Vector Machines have been widely used as they have proved to be effective in text classification tasks. Naughton, Stokes, and Carthy, (2010) found they outperformed rule based and language modelling approaches in sentence level event detection for example. They have been used in sentiment analysis and opinion mining, (Ye, Zhang, and Law, 2009; Saleh et al., 2011). In a multistage system, Krishnalal, Rengarajan, and Srinivasagan, (2010) used a HMM model to extract features which were then used to train a SVM to classify news topics on the web. Support Vector machines are naturally suited to binary classification tasks, but various methods for creating multi-class classifiers have been proposed and explored. Kumar and Gopal, (2010) provides a comparison of three such approaches and concludes that creating "one vs. all" models is the most effective.

Machine learning has been applied to content using its popularity as a feature in modelling in order to predict various future outcomes. This is relevant here because we are not only seeking material that is new, but also interesting for the user; something being popular or having an impact could be a good correlate.

The use of social media to predict future trends would seem to be a natural area

for investigation given the establishment of evolving trends therein. Relevant studies have recently started to emerge. For example, Gilbert and Karahalios, (2010) have examine whether stock market moves can be predicted from general positive and negative emotion estimated from blog postings. Politics is another area where pundits are active, and political opinion is far from absent from the blogosphere. Munson and Resnick, (2011) found that 25% of political postings were posted on blogs not considered to be political in nature, and Tumasjan et al., (2010) have investigated whether trends in party mentions can predict election outcomes using Twitter data, showing that the number of mentions of political parties correlated well with the actual election results.

Predicting the future commercial success (or otherwise) of products is also of significant interest. Film reviews have been used in sentiment analysis research, for example see Annett and Kondrak, (2008) and Joshi et al., (2010), though there is little work to date on analysing connections between reviews and the success of the actual products. However, Asur and Huberman, (2010) have compared the box office revenues for a film in its opening weekend with the average tweet-rate of Twitter postings mentioning the film over the previous week. They found a high correlation that outperformed the traditional indicator, the Hollywood Stock Exchange index. In similar work, Bothos, Apostolou, and Mentzas, (2010) have combined rating, sentiment and trend analysis of different social media streams focussed on movies, aiming to predict Oscar winners, in an agent based "Prediction Market" paradigm.

3.9 Advances in Machine Learning

The identification of pertinent features for modelling classes is often referred to as *feature engineering*. Not all features are easily isolated. For example, words (or "unigrams" more generally) may be isolated as features by simple tokenisation. Word classes, such as transitive verbs for example, or specific constructs, such as compound nouns, may require deeper processing. Feature engineering is a manual process where the features considered are based on the ideas and insight of those working in the domain of interest. This thesis, for example, is motivated by the idea that concepts as represented by nominal references are significant features in the dissemination of potentially new information through English language. Recently, though, there has been considerable interest in systematic learning of features as well as the models. This field of interest has become known as "Deep Learning".

Deep learning has arisen through progress in the development and application of artificial neural networks, a class of machine learning techniques based loosely on how the brain is thought to learn and associate patterns. The renewed interest in using multi-layered neural networks followed work by Hinton and Salakhutdinov, (2006) showing that pre-training layers in isolation could greatly assist in finding a solution to optimising auto-encoder weights. Recent advances in processing power

and the availability of very large amounts of data have enabled researchers to explore larger and more complex neural models. Such models employ multiple layers, each learning associations from patterns from preceding layers or inputs. It is the depth of these neural models, together with very large amounts of data, that have enabled the co-learning of features and classes. This has allowed modellers to move away from putting effort into feature engineering, although it is now required in network design.

The first significant advances with this class of artificial neural networks were seen in image processing applications. In 2009 the ImageNet corpus of 1.3 million high resolution images with descriptions was released (Deng et al., 2009), providing a large and diverse body training data for object recognition in images. For example see Zhao, Li, and Xing, (2011). As automatic scene and object annotation became more popular, more data was sourced, and solutions were scaled, for example see Weston, Bengio, and Usunier, (2011). With the notable success at the task being shown using deep convolution networks, see Krizhevsky, Sutskever, and Hinton, (2012), deep learning became of interest to many considering machine learning approaches.

Natural language processing often leverages machine learning, so it was natural that researchers would look at deep learning given the availability of large volumes of text. For example, Labeau et al., (2015) applied a convolutional neural network to for learning learn word representations from their constituent characters, with a second stage network to predict the part-of-speech, achieving state-of-the-art performance for German text with the need for feature engineering. Vinyals et al., (2015) have proposed a LSTM-based neural model for domain agnostic parsing, while Andor et al., (2016), using a non-recurrent feed-forward model, have achieved state-of-the-art dependency parsing performance for the Google system "Parsey McParseface". Others have investigated whether a deeper understanding of sentences can be learnt directly, for example Zhou and Xu, (2015) have achieved good results at semantic role labelling using recurrent neural networks.

One NLP challenge for which neural approaches have shown significant advance is that of efficiently representing a word's meaning. The distributional hypothesis holds that one can tell the meaning of a word by the words that co-occur with it, as"words which are similar in meaning occur in similar context" (Rubenstein and Goodenough, 1965). Although not strictly a deep learning method, Word2Vec is a technique developed by Mikolov et al., (2013b) based on a two layer neural network to encode co-occurence. Words are initially assigned random vectors which are used as inputs to the network. The network averages the input vectors and applies a weighted transform to give a single output vector for the target word. The system learns the appropriate networks weights by seeking to minimise the error for target vectors across multiple training contexts.

A alternative non-neural approach to creating word-embedding has also emerged

and become popular. The Global Local Vectors (GLoVe) (Pennington, Socher, and Manning, 2014) algorithm creates the co-occurrence matrix and then performs a Singular Value Decomposition step to reduce the dimensions to the desired representation vector length. However, it has been argued it is attempting to arrive at the same solution by different means Levy and Goldberg, (2014).

Embeddings are becoming increasing popular as a basis for further natural language analysis and processing applications. For example, Soricut and Och, (2015) have proposed an approach to inferring morphology from an embedding model trained on a large sample of the language in question, and Chen and Manning, (2014) have proposed using both word embedding, POS tags and arc label embedding instead of discrete representations in a dependency parser. However, one may observe that use of embeddings is typically aimed at improving NLP functions within, rather than directly solving, user applications such as information discovery and information extraction.

3.10 Speech and Dialog Act Detection

This thesis concerns the explicit statement of information. An explicit statement can be thought of as a particular communicative action, or dialog act – a speech act with a specialised purpose. (For practical purposes the two can largely be treated as synonymous.) Work done to advance methods for the detection of speech and dialog acts and their classification is highly relevant therefore. This section reviews recently published pertinent works.

Recall that speech act theory describes five classes of illocutionary act, and dialogue Acts provide a finer grain classification, typically being domain specific.

Speech act detection has been attempted in various online media settings. For example, Qadir and Riloff, (2011) examined the efficacy of various features in detecting and classifying four of Searle's speech acts in online message board posts, and Arguello and Shaffer, (2015) examined seven speech acts including "Question" and "Answer" in massive open online course discourse forums. Identification of speech acts in dialogue based education systems is important to infer student intentions and thereby provide appropriate feedback. Rus et al., (2012) have take a data driven clustering approach to discover the "natural" speech acts in such systems, showing potential for the method to find utterances consistent with expertly assessed class labels. Tavafi et al., (2013) have studies an SVM-HMM approach to learning dialogue acts in both asynchronous and synchronous settings, including email and meetings. Features used were domain independent but each domain setting had a different set of 11–16 dialogue annotation set.

A major challenge in creating models for recognition of speech acts is often a lack of in-domain annotated data. Jeong, Lin, and Lee, (2009) have proposed using domain adaptation techniques, leveraging labelled Switchboard-DAMSL transcripts to

create models for labelling speech acts in email and forum data. Their method, using features that included extracted phrases and dependency trees, was empirically shown to approach fully supervised learning approach (tested in a five-fold cross validation setting) for a set of 12 speech act tags. Similarly noting a lack of training data for Twitter, Zhang, Gao, and Li, (2012) also investigated semi-supervised machine learning, finding that transductive SVM to be an effective method to leverage unlabelled data in sorting into five speech act categories (Statement, Question, Suggestion, Comment, and Miscellaneous).

Dialog act recognition can be useful for the development of more natural automatic online interaction with humans: Understanding of the intent of someone's message, such as to question or to assert a requirement, is required if a system is to give a sensible response that appears to be human-like. Systems such as Skowron et al., (2013) integrate many approaches to recognise various aspects to an utterance, including dialog act, to assist in forming appropriate responses in human-computer conversation. Observing that the dialog acts in an online interactive system differ from those in two person conversations, Carpenter and Fujioka, (2011) identify more than 40 fine grained dialog acts and describe a rule based system for their identification. Moldovan, Rus, and Graesser, (2011) found words and their POS tags appearing at the start of utterances to be highly predictive of the speech act tag for online chat. Using Naive Bayes and Decision Trees to learn tagging models they achieved F-Measures of around 0.7 for tagging the 15 speech acts identified in the LDC Online Chat corpus Lin, (2007).

Game-play involving natural language can be a useful data source for learning the underlying intents in utterances. Orkin and Roy, (2011) investigated dialog classification in social agent gameplay, and Rus et al., (2012) examined whether or not speech act categories can be automatically discovered in educational games. Wittgenstein's idea of "language games" has been equated with dialogue acts, Schmitz and Quantz, (2013). Although he thought there could be an infinite number of constructs. Despite this, as Traum, (2000) notes in considering issues with Dialogue Act taxonomies, the concept of Dialogue Acts has been found to very useful, but the classes identified usually reflect the domain being studied. Particular acts can be observed in different genres, and there is usually a tradeoff to be made between the power and simplicity of the classification scheme involved. There is no universally agreed set or taxonomy although schemes may be compared. For example, Shriberg et al., (2004) provides a mapping between the SWBD-DAMSL and MRDA dialogue act sets, both of which comprise over fifty fine-grained tags.

3.11 Sentiment Analysis and Opinion Mining

Approaches to sentiment analysis have developed from simple detection of words with positive and negative connotations, such as Kim and Hovy, (2004) and Chesley et al., (2006), to more complex systems that take into account negation and context,

Narayanan, Arora, and Bhatia, (2013) and Agarwal et al., (2015). System development has been encouraged by evaluation tasks run in the now annual SemEval series of international workshops. These workshops bring research groups together to work on shared tasks requiring evaluation of the semantics in textual content. Sentiment analysis as an evaluation task was first introduced as a task to detect affect in news headlines, see Strapparava and Mihalcea, (2007). By 2016, sentiment analysis tasks had extended to work on Twitter data, seek finer grained aspectual sentiment, and introduced assessment of strength of expressed sentiment on a 5 point scale Nakov et al., (2016).

Although a document may present a positive or negative tone it may express both positive and negative sentiment towards different things or aspects. Recent work in sentiment analysis has focussed on associating sentiment with their intended target. For example see Jo and Oh, (2011). While some have focussed on building domain aspect lexicons as resources for targeted sentiment extraction, e.g. Bross and Ehrig, (2013); Park, Lee, and Moon, (2015), others have focussed on jointly modelling aspects and sentiment. For example, Kim et al., (2013) propose using a hierarchical model for extracting aspect related expressions of sentiment in online product reviews. Reviews are available online for many different products and services in many domains. They often contain sentiment towards different aspects and have been used to study aspect sentiment expression. Singh et al., (2013) have investigated using multiple features to focus on aspects in movie reviews before aggregating for an overall net sentiment, and Wallace et al., (2014) have used a probabilistic model to find latent aspect sentiment in online reviews of doctors, and analysed whether these can be used to improve model correlation with state level measures of healthcare.

Another aspect to sentiment is the strength of its expression. Drawing upon Speech Act Theory, Villaroel Ordenes et al., (2016) have found that directive and commissive acts had stronger effects than assertive acts on implicit sentiment expression. They also found that preference in speech act in implicit expression is related to the domain; one might commit to revisit a hotel, for example, but be less likely re-read a book.

Various approaches have been proposed to determine aspect based sentiment. Starting from seed sentiment labelled lexicons Kiritchenko, Zhu, and Mohammad, (2014) generated sentiment labelled lexicons, including negation conditions, via Twitter hashtags. For modelling they combined present sentiment words with various token features in Tweets to create message and aspect level feature vectors. Their system used a linear SVM based model for aspect sentiment labelling and achieved the best results in the SemEval 2013 Twitter Sentiment Analysis task. The approach taken by (Poria et al., 2014), in the Sentic system, is to use a semantic parser to decompose sentences and then use hand-crafted rules to extract concepts and dependencies between them, aligning with then SenticNet sentiment lexicon (Cambria,

Olsher, and Rajagopal, 2014). This approach allows for negation as well as target aspect detection and was found to out perform purely statistical methods. Brychcın, Konkol, and Steinberger, (2014) adopt a machine learning approach in their system, using supervised learning over a wide range of features and extracted concepts as well as latent topics derived through the unsupervised LDA method. Without extension to use external resources their system achieved better than average results in the SemEval 2014 task and subtasks.

As can be inferred from much of the data used its development and analysis, the task of opinion mining is a major application of sentiment analysis, to the point that they are largely synonymous. Although much attention has been given to product and service reviews, opinions can be, and are, expressed about many different things. Besides consumer product reviews, media reviews, opinion mining has been applied to gauging political views, (Asghar et al., 2014), detecting disruptive comments in forums, (Mihaylov and Nakov, 2016), and even in financial market prediction, (Kim, Jeong, and Ghani, 2014).

3.12 Predicate-Argument Mining and Open Information Extraction

Open Information Extraction seeks to extract relational (or factual) information of any type from text without the constraint of a predefined vocabulary. Typically in English an assertion will follow an predicate-argument structure such as "Alice married Bob", in which the predicate is "married" and the arguments are "Alice" and "Bob".

One approach to relationship extraction is to find verbs expressing that relationship and then find the arguments and determine their role. Semantic Role Labelling, as this is known, does not cover all relational expressions however as some my be expressed as a phrase, but may be converted to full Open Relationship Extraction (ORE) (Christensen, Soderland, Etzioni, et al., 2011). Christensen et al. used the SRL approach taken by Johansson and Nugues, (2008). This system relies on a syntactic dependency parser coupled with a classifier which, in addition to the parse tree output, makes use of argument labels and frame core arguments in PropBank/NomBank missing from sequence frames.

Rather than relying on deep analysis, others have sought to achieve open information extraction via shallow parsing and machine learning. The ReVerb system Fader, Soderland, and Etzioni, (2011) creates a POS and noun-phrase chunked sequence and imposes two simple syntactic and lexical constraints. The syntactic constraint requires an extracted relational phrase to match a fixed POS pattern, while the lexical constraint checks the relation against a dictionary of terms known to take a wide range of arguments. The constraints significantly improve the precision by reducing incoherent and uninformative extracted relationships.

The Ollie (Schmitz et al., 2012) system uses a bootstrap system, seeded by relationship tuples from ReVerb, to find potential relationship expressions from which to learn open pattern templates. Observing that some relationships can be mediated by nouns and adjectives, but there usually a verbal form of expression, the system looks to generate training data including non-verbal forms of relationship expression. To reduce errors from the bootstrapping the authors also impose dependency limits via a parser. Extraction patterns are learnt from the dependency tree and the phrase, resulting in templates that may contain lexical as well as syntactic cues, and were found to achieve many more correctly extracted relationships than with ReVerb.

With their system SONEX, Mesquita, (2012) take a clustering approach to open relationship extraction. Pairs of entities co-occurring in sentences are clustered is they share a similar context, Each cluster is then given a label to represent the relationship. The context is formed of weighted features that occur in the sentence with the entity pairs, including lexical unigrams and bigrams and POS patterns. A PMI based automatic evaluation against a very large corpus suggested performance comparable to the state of the art.

Mesquita, Schmidek, and Barbosa, (2013) have examined the trade-off between shallow and deep techniques for open information extraction comparing eight state-of-the-art systems. They note that different systems have been evaluated under different conditions such as relationships at corpus level or at sentence level. (In the former counts are normalised by the number of unique relationships in the corpus, whereas in the latter counts are normalised by the number of sentences expressing a relationship.) Using five different datasets and a sentence level evaluation they found that shallow parsing methods to be faster and potentially as effective than dependency parsers and semantic parsers. However, with their own proposed system EXEMPLAR they found that rule-based approaches can still be competitive with machine learning methods.

3.13 Summary

This chapter has provided a review of recent work in areas relevant to the discovery and filtering of text likely to contain new information. It began with a review of approaches to select documents relevant to a reader's interest.

Section 3.1 reviewed research into selecting and sorting documents by topic from continuous feeds, an area known as Topic Detection and Tracking (TDT). In TDT tasks the aim is to group different texts from a stream that belong to the same story together, and detect when a new story starts. This has been likened to the TREC filtering task but without a user query, and for tracking, performance is similar. However, detection of the first mention of a news story has been found to be hard, and

would require a significant improvement in IR techniques. Some have suggested using multiple stages of classification to help refine the search space.

New information may emerge with a story. Detection of such information has been examined in some TREC evaluation tasks at the sentence level, but these did not include detection of new stories.

Distance metrics have been examined for effectiveness in measuring novelty within news stories. The number of new Named Entities mentions was one metric found to significantly outperform chronological ordering according to human judgement (although new articles were found to include re-caps and elaborations of old stories).

It was then shown how News dissemination is no longer considered to be just in the domain of mainstream news organisations. Research has turned to examine the characteristics of news stories in Social Media. Approaches to detecting emerging news in blogs have included examination of linking behaviour, and content clustering. Research in this area has been supported by the release of Burton, Java, and Soboroff, (2009), a collection of 44 million blog posts and online news stories. With the growing popularity of micro-blogging in Social Media, research has turned to applying techniques for first story detection in services such as Twitter, particularly methods that can cope with very high volume and rate of short messages.

Sections 3.2 and 3.3 examined research into term occurrence, association with topic, and how use changes over time. Words one might consider to be 'content' have been found to show a much higher likelihood of re-occurring in a document than expected from its overall frequency. This phenomenon has been used as the basis to determine content words from bursts in corpora. However, use of a variety of terminology for the same things, as in co-reference to Named Entities, can pose an issue for such techniques.

Modelling distributions of text features has also been shown to highlight differences across domains, but but term usage observed in multiple domains, may also evolve over time within each domain. Evidence has been found that suggests that discussion of topics in blogs can pre-empt news-stories as well as coincide or lag news stories. Spikes in named entity mentions could be present in one medium before the other. Fine-grained event-detection is possible, with bursts in email topics seeming to coincide with the development of topics of interest.

Temporal effects, then, may be of significance in identifying potential items that could evolve into stories of interest. It has been shown that "memes" – short phrases (or even whole passages) – get reused and evolve across media. and how they evolved in news websites and blog publication However it has been noted that within mainstream news media reuse may involve re-writes simply for style purposes.

More generally, some research has found that including named entities in information retrieval based models can be effective. This could be because documents such as news stories may be about some central entity, the focus of the discourse. The usefulness of names in IR, together with the reuse of reference by different people in social media channels indicating popular interest, is a motivating factor for using named entity mentions as feature in the discovery system proposed here.

Discovery of new knowledge implies that one has old, established, knowledge. This could be formally represented. In section 3.5 the chapter therefore went on to discuss work carried out in knowledge representation. It described how in AI that was an agreed need for KR but some argument over whether this needed to be a logical formalism. A highly influential idea has been been that of frames or templates. Language understanding has been cast as the task of filling in these templates from linguistic primitives. It has been argued that logic is still important to be able to carry out inferences, though, but potentially separated from the concepts to be manipulated. These ideas have lead to the development of "ontologies" – structures representing concepts, objects, and relationships between them. General purpose and task-based ontologies have been developed. Business domain specific task-based onotologies have been used successfully but are predominantly hand-built.

Some of the various methods have been proposed and investigated for this task were discussed in the section. Applications of knowledge based approaches have included the problem of spotting news events. However, as with many ontological systems, the model is built by a domain expert, and was therefore focussed on defined domain events.

The scale of building useful ontologies has been, and continues to be, a driver for the development of semi-automated and automated approaches to knowledge acquisition. This includes approaches to learn underpinning ontology resources as for accumulating new knowledge.

There have been statistical approaches to finding new knowledge, particularly for indirect links within the biomedical domain. Swanson's discovery of such links in MEDLINE reports spurred much work. Various teams developed the ideas further or investigated similar approaches. Going beyond word association, ideas have included analysis of phrases, $tf * idf$ scoring, and using NLP techniques to class and select particular types of expression. This has included clustering of the relationship phrases between co-occurring entities, an application of Open Information Extraction, discussed in Section 3.12. Query, or focus, specific ontologies have been constructed as a background against which to find rarely occurring terms which might indicate intermediate concepts indicative of new implicit relationships.

How does one know how effective a method is at discovery when what there is to be discovered isn't known? Splitting a corpus into time periods corresponding to before and after something was discovered has been proposed. This motivates

the idea of using a rolling temporal window within which to measure document features, comparing with the past.

Knowing how likely something is can be correspondingly useful for detecting something new has occurred. A probabilistic model is appealing therefore, and due to a principled approach, the results are fully explained. However there are drawbacks. Feature independence is typically assumed due to the exponential growth in probabilities required in a combinatoric conditional model. Another problem is that probabilities are not typically known, but have to be estimated from the data which may, as in documented language, be incomplete. Rare features may appear subsequently following estimation. The usual mitigation is to account for this when estimating probabilities by a discount in a process known as 'smoothing'. Various approaches proposed and examined for this were reviewed in Section 3.7. Some task performance analysis suggests, though, that choice of method and parameters therein is sensitive.

Machine learning, both supervised and unsupervised, has been successfully applied to many areas in natural language processing. Unsupervised learning is not naturally suited to continuous streams of data, although some techniques have been proposed for learning features and continuous clustering in streaming document scenarios such as in news feeds. Supervised learning techniques are typically applied where there are defined classes to be assigned.

Naive Bayes and Decision Trees have been popular in the past, their relative simplicity having some attractive qualities. More recently Support Vector Machines have been favoured, proving more effective in many text classification tasks. SVMs are most naturally suited to binary classification problems but multi-class models can be built using multiple models where each being trained for one class against all the others has been shown to be an effective approach.

The identification of pertinent features for modelling classes is sometimes referred to as "feature engineering". Recent advances in artificial neural network models has resulted in considerable interest in systematic co-learning of features in models. This field of interest has become known as "Deep Learning". Given significant success arising from availability of very large amounts of data and increases in processing power, interest in applying deep learning to NLP tasks has grown. However, large amounts of training data are required to co-learn features and classes. Applications that have been investigated include POS tagging, dependency parsing, domain agnostic parsing, and semantic role labelling.

Recent work in detection of Speech and Dialogue Acts, of which an explicit statement can be considered an example, was then reviewed. It was shown that detection has been attempted in online settings including message boards, discussion forums, and human computer dialogue systems. Availability of in-domain annotated data has been found to be an issue. Attempts to overcome this have included collecting

data through game-play involving natural language and using semi-supervised machine learning to leverage unlabelled data. Dialogue Act taxonomies proposed have been largely task or domain specific; no universal taxonomy having been agreed.

Filtering utterances in the source text for particular speech acts could be beneficial in the discovery system envisaged here, using models developed through machine learning. The chapter next considered some of the types of information people have sought to derive systematically from utterances.

A particular act of interest to many is the expression of opinion. Recent work in this area has moved on from simply detecting verbs with emotive attachment and assessing whether documents are positive or negative in tone. Approaches have been explored to deal with nuance, negation, and strength in sentiment. Focus has also turned to find targeted sentiment towards particular aspects in expression of opinion. Various techniques proposed for aspectual opinion mining were discussed, including parsing, supervised and unsupervised learning approaches. Opinion mining has been most widely applied to product and service reviews, but also in gauging political views, disruptive forum comments, and even in financial market prediction.

This chapter completed by reviewing recent work in development of techniques in Open Information Extraction, where the objective is to extract predicate-argument expressions without restriction on the set of relationship types to be extracted. A range of shallow and deep parsing techniques have been explored. There is a trade-off to be made as shallow techniques have been found to be faster, and although approaching the accuracy of techniques employing deeper analysis, state-of-the art deep systems can yield the most competitive results.

Chapter 4

Selected Approaches for Investigation

This chapter reflects upon the findings of Chapters 2 and 3 in presenting the methodologies chosen to follow in developing the envisaged system outlined in Chapter 1. It considers the applicability of previous work and techniques in relevant fields for finding information, explicitly stated in social media text, that is new and potentially interesting to the user in the absence of a specific query.

The focal points for the thesis are expanded as the methodology for discovery and filtering is described. Recall that the envisaged system is expected to work on two streams of documents; one from social media and one from mainstream news media (acting as a proxy for known information). The first section details how the processes for the system studied in this thesis uses these two streams to find documents in social media that contain information that is of wide interest and likely to be new for the user. New and old implies a temporal aspect to the stream processing and this is also covered in the first section.

The next section considers the approach to be taken for filtering utterances. A classification scheme is first decided upon, before a consideration of class modelling is given, describing the selected approach.

Both the discovery and filtering approaches require features appearing in the source documents for their operation. The final section describes the features chosen before the chapter is closed with a summary.

4.1 Chosen methodology for finding new interesting information

Recall that the system is expected to work on streams of documents as they are published on social media or mainstream news feeds. The first challenge is to detect significant new information in a stream. But to do this one needs some idea of what is appearing 'now'. A temporal window is appropriate for this, and so it was decided to use a time-slicing method. The second challenge is to filter out what is already known. For this it was decided to use information provided through mainstream media as a proxy.

4.1.1 Selecting a detection approach

Given what one hope are meaningful features and their occurrence, how should one model the data? One may discount the Boolean word occurrence model as having been found too simplistic for standard information retrieval application. Many algorithms for retrieving and ranking documents look for presence of some/all of those words and often weight relevance according to some expectation of occurrence of those words together. The words themselves may also be weighted, as in $tf * idf$ based metrics, by how selective the word might be (or is, in the case of a fixed corpus). But in the case of a story yet to break the words that would be in that story may not yet be well associated and/or may appear in unrelated documents. The standard methods of selecting documents by weighted keyword occurrence are less likely to be effective therefore. Furthermore, possible words that would be of interest to the investigator may not be known.

Latent Dirichlet Allocation, whilst useful for analysing collections, is unlikely to be directly useful for our purposes as one is looking to discover new information; the model to which the document may be allocated has no bearing on whether new information is there. Furthermore, with new documents arriving, and potentially new topics, an increase in the number of topic models may be appropriate which would necessitate maintaining an ever increasing corpus. However, the strength of directly modelling effects known to be present in documents should be considered when turning our attention to discovering new information in arriving documents.

Probabilistic models, in particular generative models, have the advantage of relatively simple computation demands once parameter estimation has been carried out assuming features are independent. They are also routed in statistical principles so outcomes can be explained. The problem lies in effective estimation of the probabilities given sparse data. Attention to smoothing is required to arrive at the best estimates.

However, one knows that words do not occur independently. Generally, referential co-occurrences are not random. Distinguishing specific co-occurrences from those that arise by chance is therefore difficult. However, use of a joint or conditional probabilistic model where independence of features is not assumed, as in an n-gram language model, scales exponentially with the number of dependencies. This would not be practical for our purposes (we have many possible combinations to consider) without some method of eliminating significant amounts of computation. A partial model that allows for a degree of dependency (as in limiting n in an n-gram model) and/or selection of candidate combinations by a greedy algorithm may offer a computationally acceptable compromise.

It was decided to use information from feature behaviour independently, and also conditioned on co-occurrence with another feature. Temporally annotated data was sourced with which to analyse feature behaviour over time.

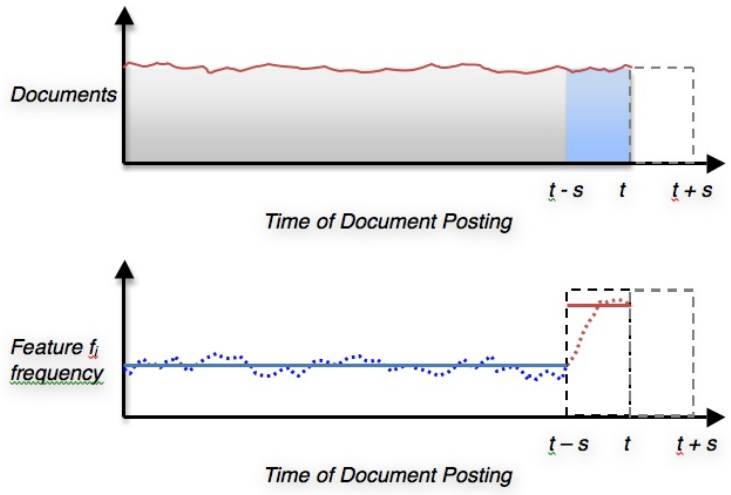


FIGURE 4.1: Time slicing: comparing current document features with those in the recent past

4.1.2 Time slicing - finding what is currently interesting

The idea of time slicing is to take documents that are produced within a period of time, and compare them to documents that were produced prior to that period. Specifically the features measured in those documents are compared. In our system it is expected that the process will be continual. To enable this the point of time-slicing may be stepped forward each time the system is used. The simplest method for this is to step a window forward by the window width as the corresponding time passes. This is illustrated in Figure 4.1.

As documents are produced on the stream, the chosen features are extracted, measured and stored. The comparison of what feature measurements are 'now' compared with what they were before necessitates obtaining measurements from prior documents. This can be achieved by running the system over a period of time without attempting any detection. A window of duration s is used. As each s passes, the frequency of features occurring in documents produced within that s will be calculated and used to estimate the expectation for the corresponding features. At time t the document features measured in the last period, $t-s$, are compared to the expected frequencies. Any feature f_i showing a significant frequency increase is selected as a *trending* feature of potential interest.

After feature selection, the expectation model is updated by the current feature frequencies and the process repeated after stepping forward another period of duration s. In this way the system employs a rolling time-slicing window. In the experiments for this thesis, a period of 1 day was set for s.

The intuition here is that things that are of current interest to people will be written about significantly more often than in general, and correspondingly the features referencing those things will 'spike' in daily frequency.

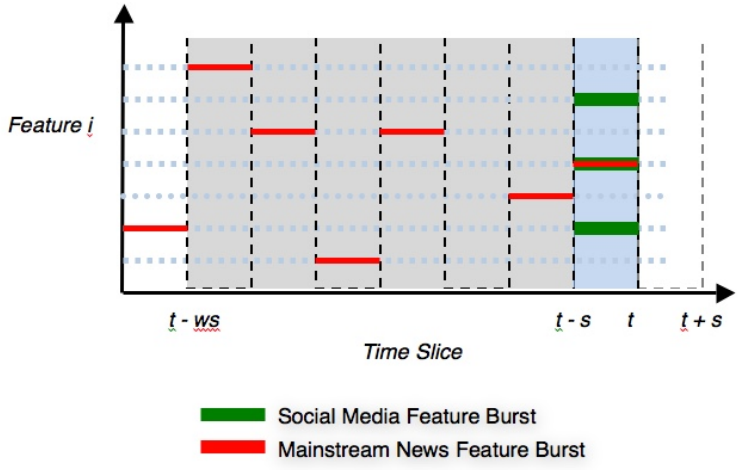


FIGURE 4.2: Time slicing: comparing current document features with those in the recent past

4.1.3 Removing the known

Although time slicing and frequency estimation may permit features potentially corresponding to topics of current interest to be selected through spikes in their use, it doesn't follow that the information being discussed is actually new, i.e. not widely known. As previously discussed, we may not know what an actual user may know, but what is already published in the mainstream news could be taken as a proxy for widely known information. It was therefore decided to use a second stream of documents, this time those being published by mainstream news sources, to represent known information.

The envisaged discovery system posits that comparison of the represented emerging information with established information would permit the detection of new information. The approach taken, therefore, is to process the mainstream news stream in the same way as the social media stream, using an aligned time slicing rolling window.

Spikes in daily feature frequency in the news may be assumed to correspond to news stories involving things referenced by those features. As such we may assume them to correspond to widely known information. Features that trend in both streams may therefore be discounted. However, discussion around a news topic may continue in social media after the news has broken. It therefore makes sense to also discount features that have occurred in recent news. Features trending in the news in a number w of contiguous time slices are removed from the set of features that have trended in social media in the current window. This is illustrated in Figure 4.2

A breaking news story resulting in trending features in the mainstream news media might have started first with chatter in social media. This would be seen as

related features trending first in social media and subsequently in the mainstream news. It was therefore decided to evaluate trending features potential for finding new information by examining how well they predicted their trending in future news media.

4.1.4 Selecting the informative documents

Having selected the features thought to be associated with new interesting information, the final step in the detection approach in the envisaged system is to select the documents thought to contain that information. It was decided to select documents based on co-occurrence of trending features because one might expect information to relate two or more things. Just as one may have an expectation for daily frequency in features, one may also have an expected co-occurrence frequency. Some combinations of features might generally co-occur so it might not be significant that two particular trending features are seen together. Just as the trending features were selected, and ranked, based on how surprising their increase in daily frequency was on the day, the approach decided upon for prioritisation of co-occurrences was to use their estimated conditioned likelihood.

Essentially, each trending feature co-occurrence may be used as a query over the documents published in social media. However the ones that are most surprising probabilistically are the most unexpected, and therefore, it is thought, the documents giving rise to the co-occurrence may be more likely to yield new information.

In IR, queries are known to typically result in irrelevant and non-relevant documents, i.e. covering more than the required topic. Ideally one would wish for queries that resulted in a minimal number of different topics. In IR this would include the desired topic, but since there isn't one in our discovery system – queries having been generated from the data – ideally one would wish for all the selected documents to be related. It was decided, therefore, to use established measures of topic coherence to examine how well the queries picked out related documents, and examine potential methods of ranking in order to prioritise results. Details of the experiments carried out are described in Chapter 6.

4.1.5 Data for discovery investigations

Documents from social media and news media, covering a reasonable number of days for calculating daily feature frequencies, were required for the experiments to be conducted. Fortunately such a dataset already existed, made available or the 3^{rd} International AAAI Conference on Weblogs and Social Media (2009). It was therefore decided to use this, the ICWSM 2009 dataset (Burton, Java, and Soboroff, 2009). Provided by Spinn3r.com, it is a set of 44 million blog posts and news stories made between August 1^{st} and October 1^{st}, 2008. The posts include the text and metadata such as the blog's homepage. Each record is timestamped with the time the article was published.

4.2 Filtering online communications in text

It was suggested in Chapter 1 that the eventual system may require a filtering stage to reduce the noise from text intended to perform actions other that to inform the reader, and to reduce the amount of processing, which could be unnecessarily high. Indeed, the amount of processing for experiments described in Chapter 6 was found to be very high. It was therefore decided to examine a method for filtering text to remove utterances not intended to inform.

One may reasonably expect nouns and indeed named entities to appear in text resulting from pragmatic intents other than one intended to inform the reader. Questions, for example, may refer to entities of interest but not provide information about them, as in " Where are Alice and Bob?". Removal of non-informative speech acts would reduce the noise in data used for novelty estimation.

4.2.1 A set of dialogue acts

Communication may be in one direction as in the publications of articles such as blog posts and news stories, or contribute to multiparty conversations as in online chat. However, in all settings one may presume that there is purpose to each communicative utterance, a dialogue act.

Many dialogue acts are unlikely to be providing information directly. Questions and requests for action for example are not intended to provide information even though one may infer information from them. For example one may infer that an author wants the answer to the question posed or the reader to respond to the request. However, methods for systematic inference were considered to be out of scope, the focus being on detecting information that the author intends to give to a, presumably, uninformed audience.

Classification of text by Dialogue Act could therefore be a useful thing to help select only those utterance that are Assertive. This may be more valuable in shorter communications than large documents which are likely to contain multiple rhetoric devices in discussing a topic. However this may not be enough. An Expressive, which yields information on the author's belief or opinion, may also be factual in respect to the author. Similarly opinion could be asserted, and an assertion could be factually false. Such situations were considered to be unlikely to be separable from factual assertions without prior contextual knowledge. It was therefore decided to generally accept this limitation in the studies conducted. (A small study, described in Chapter 7, investigated the potential for detecting self-referential utterances, but there was insufficient annotated data to draw any conclusions.)

Because many of the dialogue acts distinguished are not of interest when one is wishing to detect explicitly informative utterances, such detailed schemes did not seem warranted. In common with other studies on automatically tagging text in

online settings, a coarser task-specific set of tags seemed more appropriate. Table 4.1 lists, and broadly aligns, some of the tag sets that have been proposed in studies as reviewed in Section 3.10.

VerbMobile 1998	**J**eong 2009	**M**oldovan 2011	**Z**hang 2012	**A**rguello 2015	**P**roposed Acts
Inform Inform-Clarify Inform-Digress Inform-Init Inform-Give Reason	Statement	Statement Clarify	Statement	Issue Issue Resolution	Informative Statement
Suggest	Action Motivator	Emotion	Comment Suggestion		Comment
Request	Rhetorical Question Open Question Closed Question Wh-Question Yes/No Question	 Wh-Question Yes/No Question	Question	Question	Question
Feedback-Accept Feedback-Confirm Feedback-Reject Feedback-Backchannel	Accept Response Reject Response Uncertain Response Acknowledge	Yes Answer No Answer Accept Reject		Answer Positive Ack Negative Ack	
Convention-Politeness Convention-Greet Convention-Thank Convention-Deliberate Convention-Introduce	Polite Mechanism	 Greet Bye			
		Continuation Emphasis Other System	Misc	Other	Misc

TABLE 4.1: Example Dialogue Act categorisations and those selected for study

The *Statement* act is often found within Dialogue Act tag sets, as are forms of *Question*. Some sets provide further distinctions. As argued above, the task in hand does not require such distinctions. It was therefore decided to design a set of categories along similar lines to Zhang's scheme. The chosen categories are also shown in Table 4.1. Whereas Zhang distinguishes *Suggestion*, it was decided not to do so because a suggestion could be put in various forms, such as of a question, a comment or some directive. It was decided to include a category of *Comment/Opinion*, however, to capture general expressions of such. In common with other schemes, a catch-all category of *Other* or *Misc.* was included, plus an *Unknown* category for utterances where no defined act was discernible. Given the domain of Social Media is often used for marketing, a Directive category of Advert was also felt to be warranted. Use of this Dialogue Act tag set is examined in Chapter 7.

The Informative Statement act may be further divided to distinguish those that are explicit, i.e. do not call upon anaphoric or pronominal references to be fully interpretable. It is this particular refinement that is the concern of this thesis.

4.2.2 Building a filtering model

Given presence of features in text corresponding to multiple speech acts, what would be the best approach to filtering out acts other than the explicitly informative text?

The features selected may occur in many different speech acts. No simple rules regarding their presence or absence may be expected to result in an effective filter. Feature occurrence in speech act classes are unlikely to be independent, but interactions between different features may be indicative due to how the author is using them in the language used. Given the desire to identify text corresponding to explicitly informative speech acts from others, the details of how and why feature interactions separates classes is not a concern. A "Black box" filter model is an acceptable solution therefore.

Filtering can be viewed as a classification task. As shown in the previous chapter, many have used Machine Learning techniques to form models for various text classification applications. Unsupervised models, such as LDA and k-means clustering create unlabelled classes which may or may not result in the desired separation. In the case here the classes are known. Supervised learning models are therefore more appropriate, so it was decided to follow this approach to create the explicitly informative message dialogue act selection filter, as illustrated in Figure 4.3

Many supervised techniques have been developed, each with advantages and disadvantages. Performance of techniques may depend on feature spaces, the amount of training data, and the modelling power and assumptions in the technique. Several techniques have been popular in text classification, including Naive Bayes, Maximum Entropy, Decision Trees, Support Vector Machines, and more latterly, Artificial Neural Networks / Deep Learning. Various Machine Learning software packages are available and so it was decided to experiment with the first four of these techniques..

As discussed in Section 3.9, much of the success in deep learning applications has come with the availability of data. Significant amounts of data are required in order to learn the many parameters of the neural models. This is a drawback where data may be readily available but the desired associations are not. Annotation may be too expensive or impractical to address this issue. Unsupervised approaches, which rely of the characteristics of the feature space approaches to arrive at some separation, also require very large amounts of data for stable models. With relatively small amounts of data only small ANNs are warranted. It was decided not to include the application of deep learning techniques, mainly due to the limited amount of labelled data available but also due to the maturity of deep learning techniques during the course of this study.

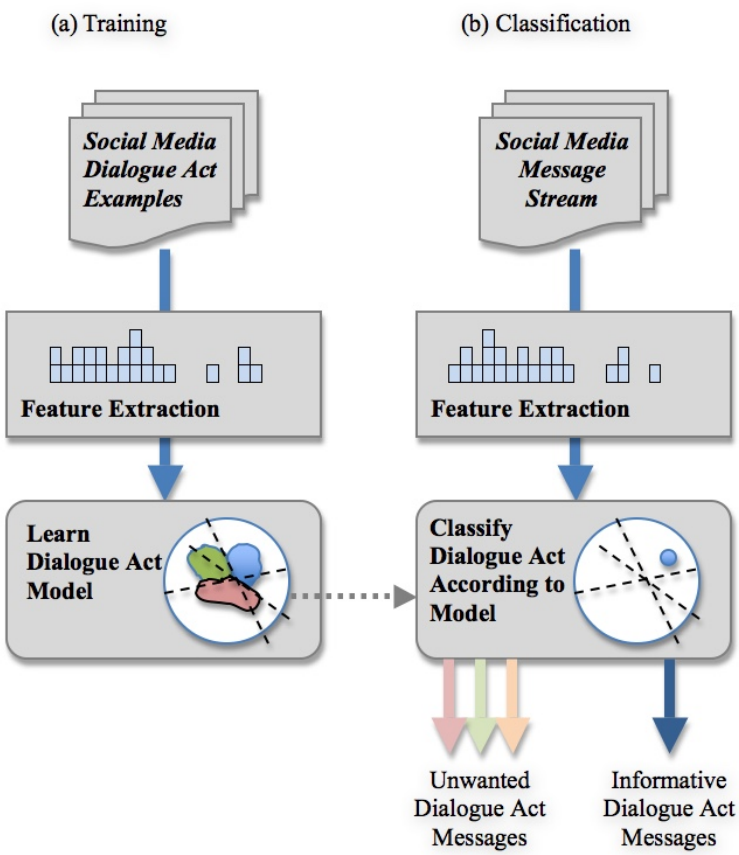


FIGURE 4.3: Creating and using a dialogue act classification model with supervised machine learning

4.2.3 Data for supervised learning

Supervised learning does require, as its name suggests, training data to create models. Data is typically made up of the features extracted from examples of the classes of interest. Creation of the models for filtering explicitly informative text therefore required suitable labelled data to be set aside for training purposes. Availability of appropriately labelled data can be problematic for classes not typically separated (or separable) by people. Finding surrogate labels or carrying out annotation exercises is typically required to form corpora for training and testing models.

It was decided to use a Twitter corpus for Dialogue Act modelling. Twitter messages, Tweets, are short (limited to 140 characters at the time of this study) and so each will typically have a single purpose. The other advantage with Twitter is that various methods for detecting burst of activity and relating Tweets therein into 'events' already exist. Correspondingly it might be possible to start with a Twitter corpus where related messages were already identified. Just such a corpus, where some related event Tweets had been identified as News and Non-News, was made available from the Redites project (Osborne et al., 2014). It was therefore decided to use this as a basis for further annotation and experimentation. Details are given in Chapter 7.

4.3 Selected Features

So far in this chapter, we have considered the approaches taken for discovery and for machine learning based filtering, both of which operate on features appearing in the source documents. This section describes the set of features chosen to be used in the system, and the rationale therein. The methods used for extracting the features, outlined below, are described in Chapter 5.

4.3.1 References to what is being talked about

What terms are informative references to concepts and what combinations should one search for? Terms may be polysemous. However, news often concerns people, places, organisations and the events they are involved in. The field of information extraction is concerned with trying to find references to such entities, events and relationship within text documents. One may consider names and references to entities as having some semantics; one understands the name of a person to refer to a person. If a name tagger has identified a name and the type of the entity it refers to then one has more information regarding the use of the token(s) used to refer to the entity than the term alone. For example the name "Bush" is semantically more informative than the token "bush". If one assumes this to be true then searching for names of people, places, organisations etc., and moreover combinations of entities should yield more precise results in the context of looking for the new news!

Names are not necessarily the only informative features one may consider however. Generic terms, i.e. common nouns and common noun phrases, may also be indicative of what is being communicated. For abstract nouns a verbal form may be used, e.g. "boycotting" rather than "a boycott". However, other words convey little or no information. Pronouns in themselves would be of little use being very common. Conjunctives, adverbs and common verbs similarly are less likely to convey subject matter. Such words are often considered as "function" words and are typically filtered out in feature selection by their high frequency. (Note, however, that not all function words are that common, for example Google has some 80 million documents indexed for "somewhat", but over 295 million indexed for "quite", while the generic common noun "church" has over 270 million entries.) Further "content" terms could be identified by their pattern of usage as in the technique described by Sarkar, Garthwaite, and De Roeck, (2005).

Function words may be thought of as "glue" in forming meaningful phrases. This leads to another way in which language can refer to some entity or concept, and that is by description. Some concepts in a language do not have, or require, a single word to refer to them, rather they are indicated by a compound of words or a phrase. Compound nouns such as "baked beans" and "tomato sauce" are relatively simple examples where the description has entered the language to mean something specific. The meaning may be determinable from the meaning of the constituent words,

as in the two previous examples, but this is not always the case. For example "action man" and "zebra crossing" refer to a toy doll, and a type of pedestrian route across a road, respectively.

Nouns are not the only type of words to form compounds. Verb phrases may have sense or meaning altered by particular particles, such as in the examples of "shake down" and "move on". Longer descriptive constructs and idioms may have entered the language to convey particular concepts. Examples include "take a picture" where the sense of the verb "take" is fixed by the noun "picture", and "taking coals to Newcastle" which is an old metaphor for a seemingly pointless exercise. Compound terms and phrases that have their own distinct meaning are referred to as "multiword expressions". Formation and detection of multiword expressions is an active area of research, attracting focussed efforts such as the series of Multiword Expression workshops (see http://multiword.sourceforge.net/PHITE.php?sitesig=CONF).Multiword expressions are distinct in having a specific meaning as opposed to simply adding the meaning of the constituent words. "A green car" describes a car that is green in colour and not a multiword expression, whereas "a blue moon" is a metaphor for a rare event, and so is one.

Given the desire to find new information about things in the world, Named Entities, Multiword Expressions and Nouns were chosen to be the features to focus on.

4.3.2 Potential features of dialogue acts

Further features besides presence of references are required for a model to select utterances performing an explicitly informative statement because they may occur within various dialogue acts. Also, for the filter model, it is not the references that are of interest but the presence thereof. A wider set of features was therefore chosen for the basis of the dialogue act classification modelling.

Given we might expect two or more references when an author is specifically imparting information relating those things, it was decided to use simple counts of named entity types and nouns, as well as compound nouns and other multiword expressions. To provide some potential differentiation between common or generic references and more specific ones, it was decided to include a set of features based on how unusual features and utterances were, given the document stream, based on the classic term-weighting idea of inverse document frequency.

Other features were also selected as potentially giving clues to the intended purpose of an utterance. These included the number of personal pronouns and domain specific features with a set function (such as identifying a user). Monetary units and date expressions were also counted as potentially useful features.

To evaluate whether or not the chosen features carried useful information for modelling the chosen dialogue acts, it was decided to investigate how well models performed on the classification task compared with models built using just the tokenised words appearing in the training data. Details of the experiments and the results are given in Chapter 7.

4.4 Summary

This chapter has laid out the approaches decided as worth exploring for the system envisaged in this thesis, having considered the relevant areas and previous research therein.

Statistical modelling of document features was selected as the most appropriate approach for information discovery. In particular, temporally sensitive models can capture significant deviation in feature values from those expected which in turn could be indicative of novelty. It was decided to firstly use information from individual feature behaviour independently, and then condition on co-occurrence with another feature.

A rolling time slicing approach was presented where feature frequency across documents appearing in a current window are compared with their expected frequency as calculated by averaging over previous time windows. A window of 1 day was selected for the experiments carried out.

The experiments would require data. The ICWSM 2009 blog dataset was identified as being suitable, being comprised of personal blog posts and news media stories, each document timestamped and tagged by its source type.

Having laid out the methodology adopted for information discovery, the chapter moved on to considering the approach for filtering utterances. An argument that an explicitly informative utterance would be classed as a Statement, or equivalent, in many Dialogue Act schemas was presented, and that fine grain distinctions were not necessary for the tasks being examined. A simple set of five Dialogue Acts including Informative Statement, Comment/Opinion, Advert, Question and Unknown was chosen. (Statement Acts that do not call upon anaphoric or pronominal references are *explicit* as previously defined.)

It was observed that opinion and belief could be expressed in different ways, and that separation of stated opinion from factual information may not be easily achieved without recourse to external knowledge. This is a limitation in the approaches taken here.

Supervised machine learning was identified as the most appropriate method to use to build a Dialogue Act classification model with which to carry out filtering. It was decided to investigate Maximum Entropy, Naive Bayes, Decision Trees and Support Vector Machines as popular techniques. It was decided not to use deep

learning neural network techniques owing to the likely need for a very large amount of training data.

No existing Dialogue Act annotated utterance corpus was identified as suitable for building the intended model(s). It was therefore necessary to build one. It was decided to use a portion of the Redites Twitter corpus as a basis for this because it has already been annotated for News Events, and messages typically have single dialogical intents given their limited size.

Section 4.3.1 covered the features that the information discovery, and the Dialogue Act classification filter, would require to base their models on. It gave a recap on how it had been shown that different words perform different and multiple roles depending on the intention of the author. Information, and in particular new information, about things is the object of interest. It follows, therefore, what the words refer to rather than the words are in their own right, are pertinent to the problem. Nouns, named entities and multiword expressions are all used to refer to worldly concepts and entities.

Whereas nouns and named entities are relatively simple for both people and, to a large extent, automatic techniques, in well formed language, multiword expressions (MWEs) are more problematic.

A wider set of more generic feature types was thought necessary to be the basis of the Dialogue Act classification modelling. It was decided to use simple counts of named entity types and nouns, as well as compound nouns and other multiword expressions. In addition, features based on classic inverse document frequency, were added in the hope this would help de-weight common messages. Personal pronoun occurrences, monetary units and platform specific features such as user identifiers, completed the set of selected features thought to be potentially useful in identifying the selected Dialogue Acts.

Feature extraction, including the potential limitations of detection and extraction of MWEs is explored next in Chapter 5.

Chapter 5

Feature Extraction

The system envisaged in this work is intended to detect new, interesting, information through statistically surprising references to named entities and concepts in social media text. New and interesting implies that this should be distinguished from that which is old or well known. A comparative stream of news media is used here to provide this background information. In order to calculate statistics on the references made in the source documents it is first necessary to detect them. This chapter focuses on that problem. It describes the methods used to extract these features from the data used, and the efficacy of that feature extraction. Some feature classes pose more of a challenge than others for reliable identification. In particular, those that are constituted by multiple words that are also used for non-referential linguistic constructs, multiword expressions, pose an issue due to complex inter-dependencies. Another factor, as will be shown, is that a distinction between what constitutes a multiword expression, and what does not, is not always readily agreed.

As described in previous chapters, research into natural language processing has resulted in many tools to process text and in particular that in English. The work reported here uses well established tools where possible for feature extraction. However, in the case of multiword expressions, no suitable tool was available at the time that this research was undertaken. (Research into multiword expressions was receiving attention by several groups, and tools have subsequently emerged.) A tool was therefore developed, based on shallow parsing – a technique that has been used successfully to extract word sequences in other contexts such as sentiment expression. An evaluation of this tool is described in the discussion on feature extraction efficacy.

The rest of the chapter is organised as follows:

Section 5.1 describes all the features selected for the experiments carried out, the challenge they pose for identification, and their interaction; Section 5.2 describes some of the tools available for feature extraction and in particular those selected for the purposes here, along with their reported efficacy. Section 5.3 describes the tool, called MESME, developed to extract the multiword features, along with an evaluation. Section 5.4 returns to the discussion on features and their identification. In particular it provides an analysis of human judgements on the validity of MWEs

extracted by MESME that were not previously identified by human annotators in a "gold standard" corpus. The results of this analysis are then discussed. Finally, Section 5.5 provides a summary of this chapter's findings.

5.1 Selected Features

The baseline for text processing is typically the words appearing in the text. Identification of lexical units would seem at first to be trivial, but even this needs some processing due to issues such as punctuation, abbreviations, acronyms, etc. Tokenisation aims to resolve these to produce an ordered list of lexical tokens. Text processing frameworks typically provide some default tokenisation as a pre-processing step. Another popular simple technique employed is to split text into tokens based upon white space and punctuation character sequences, often encoded as regular expressions.

As discussed in Chapter 4, the goal of the envisaged system is to find new interesting information from the text produced in Social Media, only a portion of which will be intended to impart information. Ideally, one would only wish to process explicitly informative text in detecting what is being referenced when seeking to find what is potentially new and interesting. Detecting features that correlate with explicitly informative text more closely than the words used alone would be advantageous in the system. It seems a reasonable conjecture that an author will often need to make reference to what the information is about when imparting it, and names, nouns and multiword expressions are all types of words or phrases used to refer to things. It would seem appropriate, therefore, to focus on detecting these features when seeking to find text that imparts information.

Identification of specific types of words and phrases necessitates further analysis beyond tokenisation of the text. The next level of distinction can be thought of as the syntactic role each token performs. The part-of-speech (POS) labels for roles in the grammar yield a set of potentially useful features for this. Various sets of POS tags with varying levels of granularity in distinction have been proposed, but the Penn Treebank tag set is often used as a basis. Nouns are defined as "A word that can be used to refer to a person, animal, place, thing, phenomenon, substance, quality, or idea" [1]. Personal pronouns could also be useful to distinguish reference to the author, reader(s), and other things. Both are readily identified given the output of a POS-tagger.

Other syntactic features may also be readily identified, either within a tagging tool, or by simple rule. These are features that recognise stylistic conventions used by people to indicate types of reference within a domain. For Twitter, and increasingly other Social Media platforms, the prefix of an '@' is used to indicate a person by their user ID on the platform. An embedded '@' may indicate an email address.

[1] Definition as per Wiktionary: https://en.wiktionary.org/wiki/noun

The hashtag, a token prefixed by '#', is another convention made famous by Twitter, and is typically used to indicate a topic or keyword. Another online referent that follows a written protocol is the Universal Resource Location (URL) use to specify a logical address in the word wide web.

More traditional named referents such as People, Locations, and Organisations, as well as expressions of time and quantities of money, require further, deeper analysis. This involves the examination of sequences of POS-tagged tokens and the chunking of constituent words. Fine grain taxonomies of entity types can be employed, however, for the purposes of indication that something is being referred to one may assume that such type distinctions would add little benefit. The commonly used generic named entity types were selected as features to include therefore.

Proper names are not the only references to comprise multiple tokens. Compound nouns such as "baked beans" reference particular things or concepts even though made up of multiple tokens. Multiword expressions have a meaning or, put another way, invoke some concept. They are made up of tokens that may include multiple POS tag types. Such sequences may not necessarily indicate a particular concept or entity however. For example, "long street" would not be considered a concept in its own right. It was decided to include multiword expressions as a feature to investigate because it was thought that their use would be more specific than individual nouns. For example, "the coast road" refers specifically to one thing rather than two. Even though their identification as distinct from words as part of general phrasal construction is a challenge, they may capture some references that would otherwise be lost owing to use of more commonly individually used words.

In information retrieval, weighting terms by term frequency and inverse document frequency – $tf * idf$ – is often employed. The importance of a term to a document is measured by the term frequency. The rationale for the inverse document frequency is that, statistically, how topically selective a term may be can be quantified as the inverse of the number of documents it appears in (Robertson and Jones, 1976) given a sufficiently large corpus. This may be used as the basis for estimating a word's weight in conveying information on a topic, e.g. see Joho and Sanderson, (2007). Therefore, features based on the inverse document frequency of tokens and phrases were included for investigation because they might indicate how informative a piece of text is.

The features selected for the experiments carried out are summarised in Table 5.1. Not all features were used in all experiments. Choice of features used depended on whether they were appropriate for the data and the task (filtering or discovery). Feature counts were used for filtering models whereas the tokens identified as conforming to the feature definitions (i.e. the nouns, the named entity mentions, and the multiword expressions themselves) were used in discovery for finer grain analysis.

Feature	Determination	Description
Unigram	Lexical	Single unannotated token, used as a baseline in comparative experiments
Tokens	Lexical	Simple count of tokens in a text - the length of a document
Average IDF	Lexical Statistic	The inverse document frequency of the text (Macro average)
Mean token IDF	Lexical Statistic	Mean of the IDF of tokens appearing in the text. (Micro average)
Noun	POS	Single tokens identified as a noun
Pronoun	POS	Single tokens identified as a pronoun
UserID	Orthographic	Tokens with a '@' prefix
Hashtag	Orthographic	Tokens with a '#' prefix
URL	Orthographic	Tokens conforming to URL specification
Emoticon	Orthographic	Tokens intended to be decoded to give graphical representation of an emotion
Person	Named Entity	Sequences of tokens tagged as referring to a Person
Organisation	Named Entity	Sequences of tokens tagged as referring to an Organisation
Location	Named Entity	Sequences of tokens tagged as referring to a Location or Address
Miscellaneous	Named Entity	Sequences of tokens tagged as referring to some non-determined named entity
Date	Named Entity	Sequences of tokens tagged as referring to a specific date
Money	Named Entity	Sequences of tokens tagged as referring to a quantity of money
Noun Phrase	Multiword Expression	Sequences of tokens identified as a compound noun expression
Verb Phrase	Multiword Expression	Sequences of tokens identified as a verb-particle or light verb construct

TABLE 5.1: Features chosen for analysis in experiments

5.2 Extraction tools

Given the features chosen for extraction, attention turned to tools to identify and extract those features. As many of these features have themselves been the target of research efforts, there are many text processing systems and tools that are able to do this. This is especially true for English, which has often been the language of focus.

Many of the desired features are either the product of or are derived from POS tags. Available taggers typically include tokenisation in their processing. However, where lexical features were required independently it was necessary to provide tokenisation. It was decided to do this simply within java code created for the experiments using the tokeniser within the SDK parameterised to split tokens on white space characters and punctuation. Otherwise it was decided to derive all features from the output of the tools selected.

Although many POS taggers exist, one of the most popular ones is the Maximum Entropy POS tagger produced by Toutanova et al., (2003a), which is included in the Stanford CoreNLP SDK (Manning et al., 2014). The SDK also includes a CRF based Named Entity extractor and a default model for the basic MUC entity types[2]. The default English POS model was trained on the Penn Treebank and therefore produces the Penn Treebank tags. These tags are limited to what is seen in curated text

[2] Person, Location, Organisation, Date, Time, Monetary, Percent. See Chinchor and Robinson, (1997)

which does not include the kind of conventions seen in Social Media such as Twitter. Extension of the tag set to cover such word use has been shown to be effective, e.g. see Gimpel et al., (2011). Twitter was a data source of interest so it made sense to either train a Twitter-specific POS model or use one already developed if possible.

Two POS taggers specifically created for Twitter were published during the study here, CMU's TweetNLP (Owoputi et al., 2013), and the TwitIE system (Bontcheva et al., 2013). (The former was found after TwitIE had been identified and adopted, and was therefore not considered further.) TwitIE operates with the GATE architecture (Cunningham et al., 2002), and is based on the Stanford POS model as used in GATE. As the author was already familiar with using the GATE framework the TwitIE model was selected for processing Twitter data.

POS tagging is often an important prior process for many NLP tasks, including Named Entity recognition. Most NER system are built within frameworks that already perform tokenisation and POS tagging. Popular frameworks that include NER functionality and models include the aforementioned CoreNLP and GATE. GATE includes an NER configuration known as ANNIE . Like the CoreNLP NER default models, ANNIE models the core entity types as defined in the MUC tasks. Other popular frameworks include OpenNLP (Baldridge, 2005) and the NLTK (Bird, 2006). Given the selection of POS taggers, it made sense to focus on the GATE and OpenNLP frameworks and therefore the OpenNLP and ANNIE NER models.

Dlugolinskỳ, Ciglan, and Laclavík, (2013) found that the ANNIE and CoreNLP NER tools have similar performance in extraction from microblogs, with ANNIE better at Organisation identification, and CoreNLP better at Person identification. Micro-average F1 was measured at 0.60 for ANNIE and 0.63 for CoreNLP for Location, Organisation, and Person entities in the MSM challenge set (Dojchinovski and Kliegr, 2013). In testing, ANNIE processing was found to be less robust to noise in input text than CoreNLP, as well as being slower in execution. It was therefore decided to use the CoreNLP MUC NER model, which also identified the desired entity types, for processing data. As TwitIE also includes Named Entity recognition and it was already being used to produce the microblog-specific POS tags, it made sense to also use it for those features as well. Both NER models covered the desired entity types. The TwitIE NER model extends the ANNIE NER model which, as described above, is also available in the GATE distribution. As both ANNIE and TwitIE models identify the standard MUC entity types, ANNIE was also used to provide a baseline for Named Entity feature extraction from Twitter data in evaluating Multiword expression identification dependency on domain specific POS tagging.

The choice of NER engines did have the limitation that different engines were used with different datasets in the course of this work. However, both frameworks are provided as java based SDKs, and the investigation did not require comparison of entity frequencies across corpus domains.

At the time that experiments were being carried out no multiword expression identification tools were found readily available, although subsequently such tools have been produced. It was therefore deemed necessary to create an extraction tool for multiword expressions, especially compound nouns, as it was thought that such features may be useful in that they make reference in similar fashion to nouns and named entities. It was decided to make use of features already being produced as the basis for an extraction model. This is discussed in the following section.

5.3 MESME: A tool for multiword expression extraction

Multiword expressions are those that are: a) decomposable into multiple simplex words, and b) lexically, syntactically, semantically, pragmatically and/or statistically idiosyncratic (Baldwin and Kim, 2010). They have been defined as "idiosyncratic interpretations that cross word boundaries or spaces" (Sag et al., 2002). Various types of MWE have been identified and described. For example Vincze, Nagy, and Berend, (2011) highlight compound nouns, verb particles, light verb constructs, and idioms as common types, although they find that these classes do not cover all MWEs. Compound nouns include such expressions as "conference paper" and "railway station", verb particles include "follow on" and "run in", and light verb constructs, where the sense of the verb is determined by a nominal argument, are those such as "take a break" and "throw a party". It was decided that these major classes would be suitable as features for the envisaged discovery system. As discussed above, nouns are one of the selected feature types. It may be supposed, therefore, that compound nouns might also provide useful features. Particular emphasis in the development of MESME was therefore put on detecting noun phrases, but verb phrases in the form of verb particles and light verb constructs were also sought in order to capture concepts such as a "break up". Idioms, such as "raining cats and dogs" were not targeted in MESME however.

MESME takes a two-stage approach to extracting the three chosen classes of MWE. The first stage uses rule-based shallow parsing carried out through two state machines. One state machine was designed for compound nouns, the other designed to capture verb particles and light verb constructs. The aim of this stage is to detect potential MWE token sequences without an overwhelming level of over-generation. The second stage applies a filter, developed through machine learning, to to filter out non MWEs suggested by the first stage. For technical details see Appendix A.

Having constructed an extraction tool the question of how well it performs at the task naturally arises. The rest of this section covers the evaluation of MESME.

5.3.1 Evaluation Data

Two corpora were used for the experiments detailed below in 5.3.2. The first, the Wiki50 corpus (Vincze, Nagy, and Berend, 2011), provides a "gold standard" mark-up of MWEs. It comprises 50 pages taken from wikipedia which have been marked-up for compound noun expressions, VPCs and LVCs. This corpus was selected to enable assessment of the efficacy of the extractor with relatively grammatically correct English. The second corpus used, referred to here as 'Tweet-4-MWE', comprised 9,976 Tweets sampled from the 37.5m English Twitter microblog messages in the Redites project corpus (Osborne et al., 2014). This sample corpus was created to examine MWE detection performance in less formal text, as found in microblogs. Details of how the Tweet-4-MWE corpus was created are also given in Appendix A. Table 5.2 gives a summary of the corpora.

MWE Type	**Wiki50**		**Microblog**	
	unique	total	unique	total
Compound Nouns	2408	2926	843	843
VPCs	348	446	31	310
LVCs	343	368	827	8270
Idioms	18	19	86	553

TABLE 5.2: MWE counts in the Wiki50 and the Tweet-4-MWE (Redites sample) Microblog corpora

Although examples of Tweets with MWEs were selected using a limited sample, it is worth noting how often they appear within Twitter messages. Also, the lists of compound nouns, verb particles and light verb constructs are not complete, and do not include all variations. The figures are therefore should be considered a lower bound. The 843 example compound nouns were found in 286,821 Tweets, or 0.76% of the 37.5m Tweets examined. Similarly VPC examples were found in 0.38%, LVC examples in 0.46%, but the example idioms in just 0.02%.

Multiword expression use in Twitter is relatively low compared to mentions of Named Entities. However they do occur, mostly demonstrating a classic Zipf power law distribution in occurrence frequency as shown in Figure 5.1.

There are some compound nouns that might be expected. For example, "high school", "video game" and "phone call" are the most frequent. Speculating, it could be that the frequencies seen for these reflect the medium Twitter is largely used on, and the demographics of those using it. Similarly there seem to be some more commonly used light verb constructs such as "have fun" and "fall in love", possibly reflecting the social nature of microblogging. The top ten MWEs for each class are given in Table 5.3.

5.3.2 MWE Extractor Evaluation

The experiments reported here made use of the two corpora described above. Experiments on the Wiki50 corpus focussed on assessing the impact of POS tagging performance and the efficacy of the extractor with relatively grammatically correct

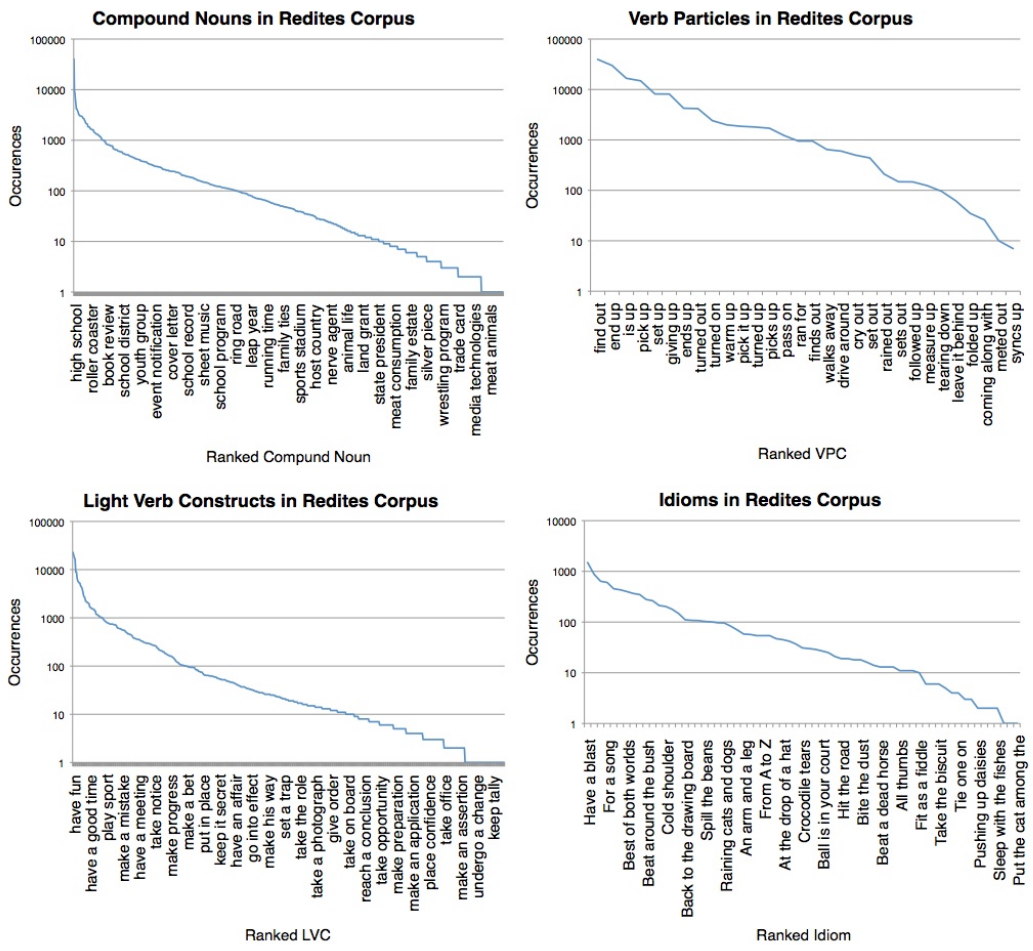


FIGURE 5.1: Frequency of example multiword expressions in Redites Corpus

English. The experiments on the sampled Twitter dataset focussed on the performance of the extractor at finding MWEs in less formal text.

Initial evaluation and dependence on part-of-speech tagging accuracy

Processing of the datasets consisted of applying the TwitIE extractor (Bontcheva et al., 2013) to tokenise the text and obtain the Part-of-Speech tags for each token. For a baseline comparison the datasets were also tagged using the ANNIE extractor (Cunningham et al., 2002). This is included in the GATE NLP framework (Cunningham et al., 2011) which TwitIE uses. Both extractors use the Stanford POS tag set, and TwitIE is 'tuned' to the less formal language found in micro-blogs. MWEs are then obtained by passing the aligned POS tag and token sequences (i.e. the words etc. the POS tags apply to) to the MWE parsers. Both the compound noun and verbal construct parsers accumulate a list of potential MWEs until the sequence has been completed. Defeats, such as single noun extracts, are then filtered out.

The first experiment sought to establish the detection performance on (presumably) gold-standard annotated well formed English. The Wiki50 dataset was used

C.Noun	VPC	LVC	Idiom
high school	find out	have fun	Have a blast
video game	end up	fall in love	Off the hook
phone call	is up	go to bed	Actions speak louder than words
body fat	pick up	take care	For a song
heart attack	set up	take a look	Call it a day
million dollar	giving up	have sex	Through thick and thin
weight loss	ends up	make sense	Best of both worlds
cell phone	turned out	pay attention	Elephant in the room
trust issues	turned on	do homework	Break a leg
credit card	warm up	feel sorry	Beat around the bush

TABLE 5.3: Top 10 occurring class MWEs in Redites corpus

for this. MESME was run with both the TwitIE and ANNIE taggers to give an indication of performance dependence on tagger accuracy. Results for the Wiki50 dataset are shown in Table 5.4. Results are given for the first stage detection and the for the second stage filtered detection in parentheses. The number of full matches and partial matches are given along with the number of spurious (i.e. unexpected or false) extractions. Partial matches are those where a MWE has been identified at the position of an annotated MWE, but the extent (i.e. the number of words in the MWE) is not the same as that annotated for it. Spurious detections are those detections for which there are no corresponding annotations in the corpus. Recall – the proportion of annotated expressions correctly detected – and Precision – the number of correct detections as a proportion of detections – are also given along with the harmonic mean of the two, F1.

Whereas recall for both compound nouns and VPCs is reasonably good at 0.68 and 0.59 respectively, the syntactically more complex and variable LVCs proved harder to identify with a recall of 0.56 but with precision at just 0.08. Despite the text not being formed from microblog messages, MESME accuracy was marginally higher with the TwitIE output, indicating improved tagging performance by TwitIE over ANNIE even for well formed English. However a comparison of tagger performance was not the objective here.

	Com. Noun	VPC	LVC	Idiom
True Markup	3009	446	368	19
Annie				
Full matches	2010	233	202	5
Partial matches	69	7	17	0
Spurious	4509	449	2206	-
Recall	0.67 (0.38)	0.52 (0.25)	0.55 (0.21)	0.26
Precision	0.31 (0.82)	0.34 (0.59)	0.08 (0.61)	-
F1	0.42 (0.52)	0.41 (0.35)	0.14 (0.30)	-
TwitIE				
Full matches	2045	261	206	6
Partial matches	71	8	17	0
Spurious	4311	490	2309	-
Recall	0.68 (0.34)	0.59 (0.34)	0.56 (0.22)	0.26
Precision	0.32 (0.74)	0.35 (0.64)	0.08 (0.65)	-
F1	0.44 (0.47)	0.44 (0.44)	0.14 (0.33)	-

TABLE 5.4: Extractor candidate results on Wiki50 dataset using Annie and TwitIE taggers, 10-fold cross validated predicted performance post SVM filter given in parentheses.

First stage precision is better when using TwitIE over ANNIE for tagging, but it is still low, particularly for LVCs. The syntactic patterns encoded in the state machines produce both true and false MWEs, suggesting that syntactic information in the form of POS tags alone may be insufficient to robustly identify MWEs. This provides the incentive for the second stage of filtering in MESME, based on additional information.

Training and testing of SVM models using 10-fold cross validation on the Wiki50 extracted candidate feature vectors suggests significantly improved precision can be achieved, although with a corresponding drop in recall. F1, the harmonic mean of recall and precision, shows improvement for compound nouns and LVCs but not VPCs. Figures are shown in parentheses in Table 5.4.

Evaluation of MESME on Twitter data

The goal of MESME is to detect MWEs in informal text such as that found in microblogs. The next experiments were aimed at evaluating performance of the approach on such material compared with that observed on well-formed English. They were also intended to give an indication of the contribution of the features used towards MWE detection.

In the first instance just the POS parser was used to do the detection. Results are summarised in Table 5.5. Applying MESME without filtering to the Tweet-4-MWE dataset resulted in 501 (59%) of the expected compound nouns being found (including 77 plural forms of the singular expression given), a 3% absolute decline in performance when working with informal language. Additionally, 273 (32%) were identified as part of longer expressions.

	Com. Noun	**VPC**	**LVC**	**Idiom**	**C.Noun filt**	**VPC filt**	**LVC filt**	**Idiom filt**
True Markup	843	310	8270	553	843	310	8270	553
Annie								
Recall	0.54	0.43	0.27	0.12	0.51	0.32	0.14	0.03
Partial recall	0.39	0.03	0.03	0.12	0.25	0.01	0.01	0.06
Spurious rate	2.47	0.78	0.60	0.22	1.44	0.17	0.15	0.09
Proj. Precision	0.18	0.36	0.30	0.23	0.26	0.65	0.49	0.22
Proj. F1	0.26	0.34	0.28	0.16	0.34	0.43	0.22	0.04
TwitIE								
Recall	0.60	0.50	0.29	0.17	0.56	0.43	0.16	0.05
Partial recall	0.34	0.03	0.03	0.16	0.19	0.02	0.01	0.06
Spurious rate	1.83	0.38	0.43	0.24	1.40	0.14	0.15	0.07
Proj. Precision	0.25	0.57	0.41	0.30	0.29	0.76	0.52	0.26
Proj. F1	**0.40**	0.53	**0.34**	**0.22**	0.38	**0.55**	0.24	0.08

TABLE 5.5: Extractor performance on Twitter messages using Annie and TwitIE taggers, without and with application of 2nd stage SVM filter model trained on the Wiki50 corpus. (Projected metrics are lower bounds.)

The advantage of using a tagger tuned to social media is seen for compound noun detection as this achieves an 5% absolute increase in recall, a greater difference than the 1% seen between use of TwitIE and ANNIE on wikipedia text. Precision was not measured as spurious extracted MWEs were not assessed for accuracy. However, if

one assumes all spurious extracts not to be MWEs then a lower bound of 7% absolute increase in precision over the baseline ANNIE system is observed.

VPC identification in social media similarly shows a drop in performance to that achieved in Wikipedia text: recall down approximately 9% absolute using TwitIE POS tags. (TwitIE maintains a similar advantage over ANNIE). LVC performance drops significantly however, recall falling by more than 25% absolute for both taggers, although TwitIE does marginally better. This may be due to the increased likelihood of POS errors within the longer token sequences of LVCs.

For completeness, MESME was also run over the Tweets where idioms had been found, treating them as a sub-class of LVCs. Unsurprisingly performance was very low. Lower bound precision and F1 were calculated assuming spurious LVCs and partial matches to be incorrect. Given idioms were not targeted in the design of MESME and little data was available to evaluate it for idioms, no meaningful comparison could be made between performance on Wiki50 data and Twitter for idiom extraction.

The experiments were run again taking the results from the second stage filter (built on the Wiki50 data tagged by the appropriate POS tagger). Results are also shown in Table 5.5. There is a small drop in recall in extracting compound nouns, but the corresponding increase in the projected precision is also small. As this is a lower bound (some spurious results may be other MWEs) the true performance may be better. The ML model does improve the precision for VPCs, though, giving an increase in the lower bound F1 projection. LVC detection, which was already a difficult task for the extractor before filtering, saw precision improvement from the ML model but at the cost of over half of the true detections. Spurious extraction rates drop for use of all filter models. Idiom performance is particularly poor using the filtered models, however there was little in the way of idiom training examples.

If one assumes a false detection consistent with the precision observed when running on the Wiki50 corpus, then use of the preceding and subsequent tags, orthographic token classes and lexical stems for verbs and adverbs as features in the second stage filter results in improved precision in the detection of all three MWE classes. However, overall performance, as measured by F1, declines for LVC detection. Information contained in features beyond POS sequences may therefore contribute to the determination of MWEs, but models built on edited English may not be optimal for informal English. Unsurprisingly POS tagging accuracy is still a factor in overall performance.

Each Tweet in the Tweet-4-MWE corpus had been selected for one MWE. However the extractor identified candidates for all three classes of MWE under consideration. For example, 52 VPCs were found within the messages selected for compound nouns. 18 of these appear to be genuine VPCs. Examples are given in Table 5.6.

True VPC		False VPC	
give up	chime in	explains they	look like
sit down	geared up	connected to	imagine if
looking forward	dozes off	startled by	told me
check out	play at	misses me	welcome to
drawn out	wake up	add that	goes directly

TABLE 5.6: Candidate VPCs from Twitter dataset

5.3.3 Discussion

As it operates primarily on the syntactic level, the performance of the extractor is dependent to a large degree on the quality of the part-of-speech tagger. This is shown in the difference in results seen here between the application of a untuned tagger and that of a tagger designed for the target medium (in this case microblogs). Although use of TwitIE does have a performance advantage over ANNIE on Wikipedia, it is more pronounced on the Twitter data. The syntactic parse yields a reasonable performance in extracting compound noun candidates. Furthermore, the application of a filter created through machine learning results in only a small reduction in recall for a gain in precision.

Extraction of VPCs is challenging because there are many common verb particle combinations that are commonly found, such as "spoke to", but are not considered MWEs. (This is reflected in very low inter-annotator agreement found in assessing verb particle combinations as describe in Section 5.4 below).

LVCs proved a more challenging type of MWE to extract reliably. This is true also of idioms and other MWEs (although they are not specifically targeted by the MESME extractor). Not only are these expressions generally longer and therefore more susceptible to POS tagging errors, but there is also the issue that the syntactic form of many is found within regular (i.e. non-MWE) phrasal construction. The features used with machine learning to create a filter did little to help make a distinction as they resulted in LVC recall being significantly reduced. Further non-syntactic information would be required to improve the precision in selecting true LVCs and complex MWEs.

Overall MESME results do not compare favourably with recently reported approaches. Vincze, Nagy, and Berend, (2011) have reported F1 scores of 0.56 for compound nouns and 0.81 for VPCs (Nagy and Vincze, 2014). Riedl and Biemann, (2016) report F1 scores of 0.69, 0.71 and 0.39 for compound nouns, VPCs and LVCs respectively. These figures rise by 0.03 when MWE resources are added into their models. These approaches make use of more feature types including those that are morphologically and semantically related.

Although it was disappointing that MESME did not achieve state-of-the-art performance, its development provided some interesting insights given the approach taken.

5.4 Feature reliability

This section considers how well defined the selected features are. Do people agree on what constitutes a particular class of feature given the role it supposedly performs?

Agreement on what constitutes a lexical token is broadly accepted. Some disagreement might arise around situations such as how contractions such as "don't" should be tokenised, but for the purposes here the "standard" whitespace and punctuation delimited tokenisation is regarded as sufficient to be accepted as correct by people. Statistically determined features, such as inverse document frequency, are not an issue accepting that background estimates come from a sample of a growing corpus. Orthographically determined features follow a morphological protocol, e.g. hashtags are prefixed by '#', so one may assume that people will agree that, within the appropriate medium, tokens following the definition are examples of the feature.

Classes that may be applied to tokens are something that different people may have alternative views on what is correct or appropriate. Part of speech classification can be carried out at various levels of granularity, for example a tag set may or may not distinguish tenses of verbs. POS tags recognised may be dependent upon language and medium. However at a coarse level there is wide agreement on some basic categories found in most natural languages, prompting the proposal of a "universal" tag set, see Petrov, Das, and McDonald, (2011). Inter-annotator agreement on tags for English is very high, although errors may creep into gold standard sets. For example Marcus, Marcinkiewicz, and Santorini, (1993) found disagreement between trained annotators occurred on just 3.5% of the POS tags applied in tagging the Penn Treebank. Proposing methods for finding inconsistencies, Dickinson and Meurers, (2003) found a number of annotations that could not be decided given the tagging guidelines for the Penn Treebank. The Penn Treebank has received much attention and many automatic taggers have used it in development and testing. The tag set itself is commonly used or extended for other domains such as Twitter. Although requiring further tag types, agreement is still found to be high. For example, in developing a POS tag set for Twitter Gimpel et al., (2011) report an inter-annotator agreement of 92%.

Inter-annotator agreement for type of Named Entity is considered to be generally high too. For example, in annotating Named Entities in a Wikipedia sample corpus Balasuriya et al., (2009) obtained a Fleiss' κ score of 0.83 for the Named Entity types Person, Organisation, Location and Miscellaneous. As observed by Doddington et al., (2004), cases where uncertainty often arises where the terms can refer to different but related entity types, Geo-Political entities such as "Europe" often being typical examples. For example, "the White House" could refer to the building, or the organisation that operates from that location. In some cases there may be metonymy where both entities are referenced with the single mention.

The last feature class requiring human judgement to be considered is the multiword expression. Multiword expressions that are metaphorical or "conventionalized" – having a meaning that is distinct from the composition of the constituent words – may not be expected to be controversial. Compositional expressions that have entered the language as having a distinct sense, where potentially the same words could be used simply in composition of a sentence rather than as a reference, could potentially be more open to interpretation. Fazly and Stevenson, (2007) found inter-annotator agreement ranged between 67% and 80% (kappa scores from 0.56 to 0.72) showing a good degree of cohesion, but more uncertainty than found with Named Entity. Given an annotated resource was used in the creation of an automatic MWE tagger, it was decided to explore the effect of opinion on whether or not a phrase has "entered the language" a little further., based upon the idea of human correction of an automatic tagger output. For this exploration, an analysis was conducted of human judgements for "edge-cases" where lexical-syntactic patterns suggested a phrase might possibly be a multiword expression. This analysis is described below.

5.4.1 Human assessment on what word sequences constitute MWEs

MESME's model might be better than the analysis above suggests: An examination of the most frequent spurious MWEs identified, shown in Table 5.7, indicates that there may be a high proportion which have not been identified by the annotators. "Early years" for example could be argued to be a distinct identifiable concept. "Fear and Loathing" is the title of a novel by Hunter S. Thompson, the subject of one of the wiki pages in the corpus and arguably should be considered a named entity. (Although named entities are not the intended target of MESME they may be extracted if not tagged as such). Most spurious LVCs are seen just once.

Com.Noun	No.	VPC	No.	LVC	No.
Fear and Loathing	15	known for	5	used in a group	2
Japanese version	12	was about	4	coached the Sting	2
same time	11	referred to	4	came the idea	2
many years	10	according to	4	obtain the keys	2
following year	10	based on	4	left the station	2
first time	10	reported that	3	includes the full text	2
English version	8	went back	3	credited as the creator	2
same year	7	According to	3		
Numb 3rd	7	is not	3		
European versions	6	do so	3		

TABLE 5.7: Most frequent false positive extracted multi-word expressions

To gauge the level to which the extractor could be identifying MWEs that human annotators might miss, a group of five people were asked to assess whether the additional MWEs the extractor found were genuine MWEs or not. The assumption here is that the annotation of the Wiki50 corpus may be incomplete or have debatable judgements. The spurious extracts therefore may not be false extracts but rather

edge cases. The interest here was a native speaker's "gut instinct" rather than linguistic expert opinion. The assessors were given an informal description of what constituted each of the MWE types under consideration but no formal training.

They were then asked for assessments on all VPCs and LVCs and on those compound nouns appearing more than once (simply to keep the number of assessments manageable). Any expression marked positively by only one assessor was rejected, and any marked by at least four assessors is considered a true MWE for the purpose of creating a MWE pool. The inter-annotator agreement within the pool is given in table 5.8, together with the projected precision and recall (i.e. including the original mark-up).

The most notable aspect of the annotation exercise was the relatively low level of agreement amongst the annotators as to what constituted a MWE. Some judges were more generous in their interpretation of what constituted a MWE than others. No clear agreement was found for 18% of the additionally extracted candidate MWEs. The highest agreement amongst judges was for additional candidate LVCs: All five agreed 78% were not MWEs. (Overall all five agreed that 58% of spurious extracts were not MWEs.) An analysis of inter-annotator agreement using Fleiss's kappa (which takes all decisions into account and factors in any random decision making) scores agreement at 0.13, 0.17 and 0.28 for compound noun, VPC and LVC assessments respectively. A value of less than 0.2 is generally interpreted as slight and below 0.4 as fair. We may conclude that a consistent interpretation of what constitutes a MWE is not easily arrived at as far as edge-cases are concerned.

	Com. Noun	**VPC**	**LVC**	**All**
Candidates	252	430	779	1461
True (@4+)	23	25	11	59
Agreement	0.26	0.17	0.13	0.17
Fleiss's κ	0.13	0.17	0.28	0.24
Projected precision	0.33 (+0.01)	0.41 (+0.03)	0.08 (+0.02)	

TABLE 5.8: Annotation of Wiki50 spurious extracted MWEs and projected precision if included

Low levels of agreement between annotators for the three classes of MWEs targeted by MESME for candidate expression and not already marked-up by the linguists used for the original corpus construction suggest that edge cases are commonplace and subject to opinion. (Although it is not known whether the two sets of judges have a common understanding of what constitutes an MWE.) One possible explanation is that some constructions are within a process of becoming conventionalised through usage. For example the phrases "headed off" and "focused on" were both identified and arguably are becoming, if they have not already become, a prepositional phrasal verb and a particle phrasal verb respectively. Similarly new metaphors may become adopted, or adapted as in "take the graveyard shift". For LVCs the source of some disagreement may lie in whether the verb is indeed "light"

or whether a distinct verb sense is selected by the phrase. “Breaks the news” for example was a contentious candidate.

Although the extractor performs well at identifying likely compound noun constructs, there could be benefit in identifying nested compound noun phrases such as in “equal rights demonstration march”. Furthermore some expressions may be made up of combinations of different MWE types, LVCs taking a compound noun for example. The nesting of MWEs may itself be a source of confusion in assessment. “Brought up a free man" for example is, strictly, ambiguous, but its meaning is clear with the sense of ”brought up” given by “free man”, the expression incorporating a verb particle.

The identification of verbal MWEs here is limited by the forward syntactic parse rules. The tool cannot find verb-final LVCs. The addition of a state sequence to capture these would be beneficial therefore. Finally the tool has a relatively low precision. Future work should focus on improving the filter stage through the use of more features and in-domain training data.

5.5 Summary

This chapter discussed the features chosen for investigation and the options available for their extraction from text. The features were chosen for use in either or both of the filtering and discovery stages investigated for the envisaged system. They included: words, word and MWE inverse document frequencies, nouns, pronouns, four types of named entities, dates, sums of money, emoticons, URLs, and User Identifiers.

The chapter went on to describe the tools selected for extraction of these features, required for the experiments detailed in Chapters 6 and 7, and the rationale for the choice thereof. Tools for the identification of lexical tokens (words), their part of speech, multiword expressions and named entities were all required for features, either directly or for the calculation of derived features.

Much of the core feature identification functionality is contained within the Stanford CoreNLP toolkit, together with a default model for the basic Named Entity types of interest. This SDK was therefore chosen to provide the main extraction tooling. Micro-blog messages, such as found in Twitter, contain additional conventions as well as providing more linguistic variation. An extended model tailored for such messages was selected from the Sheffield GATE NLP framework for processing Twitter data for POS and micro-blog specific features. No suitable tool to extract multiword expressions was identified at the time the experiments were conducted. It was therefore decided to develop a tool for this purpose, making use of the earlier stages of processing.

The chapter went on to describe the main classes of multiword expressions in English and the rationale for a tool to identify them. This tool, called MESME, was

aimed to extract compound nouns, verb particle constructs, and light verb constructs from informal English such as that found in social media. The data used to evaluate the tool's performance at the task and the evaluation experiments were then described. (The design and development of the tool is detailed in Appendix A.)

In sampling Twitter for a test corpus, it was confirmed that MWEs are used in a small but significant number of microblog messages. Use of idiom appears to be much rarer than that of LVCs. VPC and Compound noun usage is most present.

The evaluation experiments, unsurprisingly, showed that POS tagging quality is a significant factor in the identification of MWEs by syntactic patterns. Longer phrases and common constructions pose significant challenges as syntax alone is not sufficient to reliably detect MWEs (even with good quality tagging). MESME showed an ability to identify a majority of compound nouns, suffering only a small degradation (3% in recall) when executed on informal English as exemplified by Twitter messages. However obtaining high rates of identification for VPCs and particularly LVCs was found to be challenging [3]. Although not targeted, some idioms were identified by the tool, but overall, idiom identification performance was found to be very poor.

Subsequent to the tool's development, others have produced and evaluated similar tools. Using a wider range of features and approaches, published results for these tools show superior performance to MESME. However achieving state-of-the art performance was not the principal goal. Only an ability to detect with high precision a significant number of the intended classes of MWE was required, particularly for compound nouns, and this was achieved.

The chapter then discussed the reliability of extracted features with regard to how well agreed examples are given their definition. Low level features, such as what constitutes a lexical term (token) are generally unproblematic. Agreement between people seems to decline as deeper and finer grained interpretation, typically involving multiple terms, is required. Noting this, an analysis of areas of potential disagreement, 'edge cases', in MWEs was described. A wide degree of interpretation of what constitutes a multi-word construction was found with a Fleiss's κ of 0.24 showing only slight to fair agreement amongst annotator assessment of over 1,400 candidate expressions.

The novel contributions presented in this chapter included the MESME MWE extraction tool, a corpus of Twitter messages containing MWEs, and an analysis of the effect POS tagging accuracy has in MWE identification in Twitter messages. Additionally, an analysis of how consistent people are in their notions of what constitutes a multiword expression was also contributed.

[3] It is not known whether or not the same identification performance would be observed with News text as this was not tested owing to the lack of a test corpus. For a complete evaluation, an annotation exercise to identify MWEs in a source of News should have been carried out for further testing of the tool.

Chapter 6

Entity Mentions in Online News and Social Media

6.1 Introduction

The principal motivation for this thesis, presented in Chapter 1, was the discovery of newsworthy information before the information is curated and published as news stories. The envisaged application is an aid to the journalist or intelligence analyst seeking to filter vast amounts of textual information being published and streamed online for useful information when keywords are insufficient. The envisaged system makes use of two streams of documents; 1) social media, from where new interesting information is sought; and 2) news media, which represents well known and established information. This chapter presents an approach to reducing the search space by seeking to filter out established news stories (the well known, established, information) from social media while selecting the features potentially indicative of new interesting information. The types of feature focussed on are nouns and named entities, an analysis of which is provided for documents that have been separated as either from an established news media source or from social media. Selection of features by trends in their daily use is explored, reasoning that a significant rise in popular feature use would correlate with interest in related information. It is examined whether or not that these social media trends indicate information of wider interest by the extent to which they can be used to predict future trends in news media.

In the ideal system, features with predictive qualities would be used to select and prioritise documents for inspection by the user or for information extraction. A potential method for selecting documents is to use co-occurrence of trending features, based on the idea that expressed information would likely relate two (or more) represented things. This idea is examined in an investigation into whether or not it could be used as a basis for selecting documents containing related information.

In particular this chapter addresses the first hypothesis posed:

$H1$: Some documents containing new information can be found through an unexpected number of references to named entities and concepts.

The experiments described in this chapter also investigate the following supporting hypotheses:

$H3$: Documents imparting new information are more likely to contain unusual combinations of named entity mentions than unusual combinations of nouns.

$H3_{null}$: Co-occurring mentions of concepts and named entities are no more likely than co-occurring nouns in documents imparting new information.

$H4$: Mentions of different named entities are less likely to co-occur independently than mentions of different common nouns .

$H4_{null}$: Named entity mentions co-occur no more independently than common nouns co-occur.

The rest of this chapter is organised as follows. Firstly, Section 6.2 gives a recap of how social media differs from mainstream news outlets in disseminating information to the wider community. This provides the context and motivation for the experiments and analysis reported here. Section 6.3 describes related work. The setting for the experiments is described in Section 6.4 and the data used in Section 6.5. The experiments are reported in Section 6.6 with analysis of the results. Section 6.7 concludes the Chapter with a discussion and a summary of the findings.

6.2 Information Dissemination in Mainstream News Media and Social Media

Traditional news media seek to publish stories of interest to consumers. To be of interest and constitute "news" the stories need to present new information in a timely manner. Ideally the information should be checked and presented in clear language, often consistent with the media organisation's preferred style and editorial stance (Cameron, 1996). What information is published is decided by the organisation under its policy and constraints. In other words, information published in the mainstream media is curated.

Social media, on the other hand, is not typically subjected to editorial control[1]. Many more people provide content on social media platforms than the number of journalists providing news media. The result is much more content provision, potentially disseminated faster, than that found in mainstream news. However, social media is used for providing social communications and interaction; only a small proportion may be expected to be new information of wider interest, and the language used may exhibit a wider vocabulary and wider variance from accepted grammar. In other words, social media content is informal.

[1]Some forums may exercise moderation on postings. However where such control is exercised it is often whether a post should or not be published at all rather than to exercise editorial control.

The challenge is to find the nuggets of newsworthy information, or leads, in social media. Does such information arise and could it be found systematically? The next section provides details of data useful for investigation of these questions.

6.3 Related Work

The Internet and social media in particular has provided a rich source of data for those interested in the creation and dissemination of information. News story creation in the past was dominated by professional sources. With wide coverage and many providers, research, as described in Chapter 3, focussed on topic detection and tracking, with the detection of new stories being a key task.

Since the early work analysing news stories, the prevalence of Social Media in the wide dissemination of information has increased significantly. The availability of data regarding not only content but also users' production and consumption thereof, together with data regarding user interaction, has enabled researchers to examine social and temporal aspects of what makes an interesting topic. A popular approach in modelling has been to detect bursts of activity as indications of emerging interest in a topic. Evidence that bursts of activity aligned with interest in a topic was found as soon as wide adoption of online communication became established. For example Kleinberg, (2003) in looking for time gaps between term occurrences in email data found bursts in email topics seemed to coincide with interest to the email authors, and bursts of linking activity have been observed, in the evolution of the "Blogosphere" (Kumar et al., 2003).

Research directions turned to investigate clustering and relating documents contributing to bursts, characterisation of bursting topics being a logical step in news story detection and tracking. Gabrilovich, Dumais, and Horvitz, (2004) investigated the applicability of several distance metrics in finding novel information within clustered news stories in which new bursts of related articles with a high divergence from topic clusters to date are taken to be those containing new information; Franco and Kawai, (2010) investigated two approaches to detecting emerging news in blogs, considering "cascades" of topics through the blogosphere by measuring linking evolution (hyperlink references used in the topic posts) and by clustering the content of postings. Ha-Thuc and Srinivasan, (2008) investigated the use of a log-likelihood estimate of an event within a topic model as an intensity metric.

Glance, Hurst, and Tomokiyo, (2004) have examined bursts in phrases, mentions of people, and hyperlinks in blogs given a background of blogs published in the preceding two weeks. Looking at examples of blogs following product announcements and subsequent sales figures, they go on to hypothesise that product mentions in blogs may have predictive power.

More recently work on emerging topic and trend detection has been focussed on data from the micro-blogging web service Twitter. Micro-blogs, or "Tweets", are

restricted to 140 characters [2], and have been likened to chat rather than publication by Alvanaki et al., (2011), but snippets of (potential) news are also communicated. In their system, named "En Blogue", they detect emerging topics from online media streams by considering pairs of tags (augmented by extracted entities) at least one of which is a popular, i.e. frequent, tag.

Twitter provides its own proprietary trending topics service, but others have sought to provide similar functionality. Petrović, Osborne, and Lavrenko, (2010) have investigated first story detection in Twitter micro-blog feeds; Mathioudakis and Koudas, (2010) describe a system that detects and groups bursting keywords before determining a description of the emergent trend; Cataldi, Di Caro, and Schifanella, (2010) note that Twitter acts as a low level news flash portal where new topics involve emerging terms. They consider a term to be emerging if it frequently occurs in the interval being considered whilst relatively infrequently in a defined prior period, and generate emerging topics from co-occurrence vectors for the considered interval. Benhardus, (2010) has compared different term weighting methods when applied to detecting trends in Twitter data.

Research has also looked at how trends evolve through social media and how content spreads: Cha et al., (2009) studied the structure of the Blogosphere social network and how media content is propagated therein; Lerman and Ghosh, (2010) have studied how news spreads through the Digg and Twitter social networks.

Asur et al., (2011) have examined how trends persist and decay through social media. They found that the majority of trends follow news stories in Twitter, and that social networks and resonance are significant factors in how long a topic trend lasts, it being correlated with the number of authors re-tweeting contributing tweets. Looking at originating sources for re-tweeted items, they found that a large number could be classified as traditional news media providers such as CNN and Reuters. Simmons, Adamic, and Adar, (2011) have examined how quoted text changes as it is communicated through social media networks, finding personal bloggers to be more likely than mainstream news organisations to simply copy exact quotes. This suggests that it should be feasible to filter out mentions of established news from personal blogs to some reasonable degree.

Evidence that social media content could pre-empt publication in the mainstream first began to emerge in 2006. Lloyd, Kaulgud, and Skiena, (2006) found a small percentage of topics discussed in blogs existed before corresponding news-stories were published. Comparing the most popular named entities in news and blogs on a mentions-per-day basis they found that spikes in numbers of entity mentions could be present in one medium before the other. Examples they reported included "Hurricane Ophelia", which was mentioned in blogs much more than in news in the days before the storm hit.

[2] Raised to 280 characters in 2017

Leskovec, Backstrom, and Kleinberg, (2009) in looking at the concept of "memes", short phrases, and how they evolved in news websites and blog publication, found the majority of quotations arose in blogs, typically lagging by 2.5 hours, but some 3.5% of "meme" transfer was from blog entries and quoted in news media.

Many trend analysis approaches analyse simple lexical features, before using other techniques to improve the semantic richness. Techniques such as clustering, and feature co-occurrence analysis, may be employed to determine the topic that is trending, e.g. Mathioudakis and Koudas, (2010), Cataldi, Di Caro, and Schifanella, (2010), Alvanaki et al., (2011). One may observe that trending topics are often about tangible (named) entities. Azzam, Humphreys, and Gaizauskas, (1999) suggested that a document be about something – its topic – and that something would revolve about a central entity, though that entity may not be directly mentioned by full name more than once or twice. Evidence has been found that names can be effective in information retrieval tasks, (Thompson and Dozier, 1997), and searching for names has been shown to be a useful concept in searching archives of news stories (Saggion et al., 2005), which provides the motivation for the focus on named entities and nominal references in the studies reported in this chapter.

Whereas work described above has investigated news story detection and tracking in news streams or through social media, the work presented in the sections below focusses on bursts of interest in topics that first arise in *social media*. It also focusses on use of trending feature co-occurrence as a basis for document selection. The next section describes the application setting for the studies conducted.

6.4 Experimental setting

The setting for the experiments reported in this chapter envisages an application where two streams of textual information are available. One is the output from the mainstream news media outlets. The information from this stream represents stories that have already "broken". The other stream is the mass of social media weblogs where information that could be pertinent to stories of interest may emerge. This is broadly the document input setting as that used by Lloyd, Kaulgud, and Skiena, (2006), but here there is no interest in social media interest in stories that have already appeared in the news media. The downstream process to select the documents is also different in that the system proposed here uses trending feature co-occurrence rather than single trends.

The task of the system is to find new topics in the social media stream that are not already seen in the news stream within some arbitrary timeframe. To do this the system needs to detect bursts of topically related documents in each stream and subtract topics in the mainstream news stream from those in the social media stream, leaving only bursts of documents that are topically new. The system is illustrated in Figure 6.1.

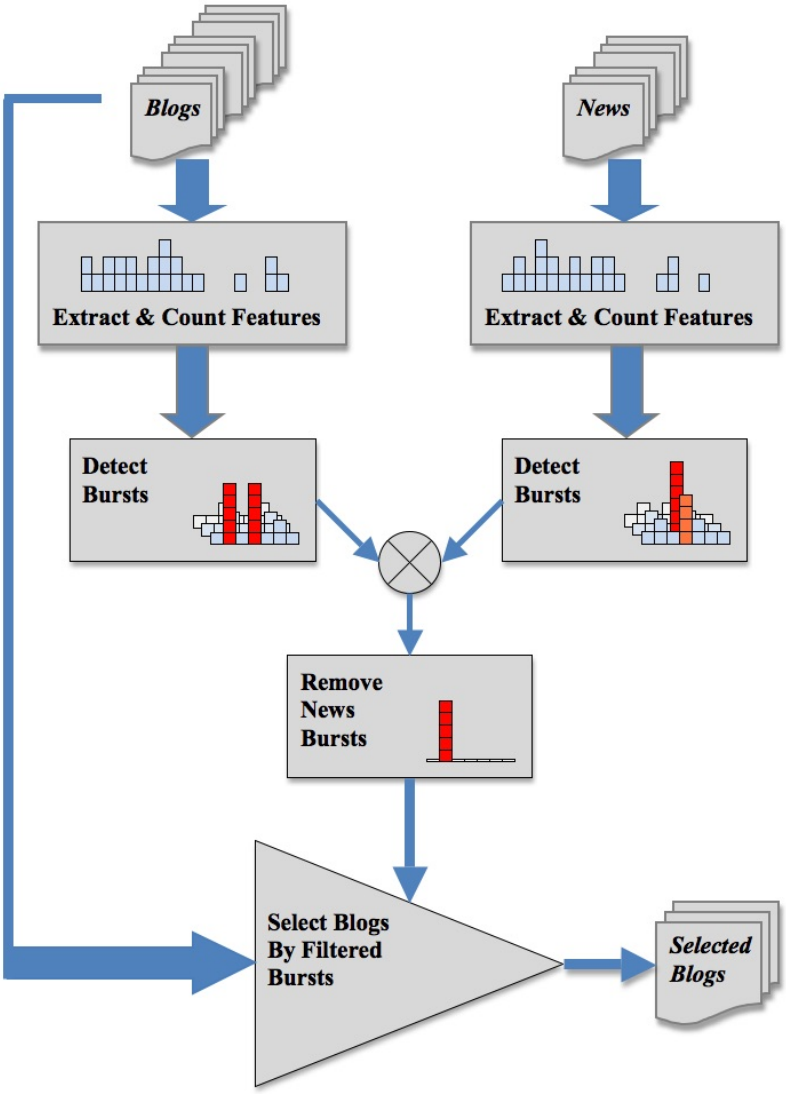


FIGURE 6.1: Illustration of filtering Blogs by occurrence of bursting features other than those bursting in news stories

6.5 Data and modelling

This section describes the data used to simulate the application system described above, the model used to detect topic bursts, the filtering conditions applied, and the features used for the temporal models.

6.5.1 The ICWSM corpus

For the 3rd International AAAI Conference on Weblogs and Social Media (2009), the organisers made a dataset, referred to as the ICWSM 2009 corpus, available to researchers Burton, Java, and Soboroff, (2009). The dataset, provided by Spinn3r.com, is a set of 44 million blog posts and news stories made between August 1st and October 1st, 2008. The records include for each blog post the text and metadata such as the blog's homepage. Each record is timestamped with the time the article was

published. Postings are not filtered for language as many posting do not have a language tag applied and some posts contain words from more than one language. It is possible that a topic related word in one language may be referred to in a post written in another for example, and this may influence the detection of a burst in a topic.

6.5.2 Features and their extraction

The hypotheses explored in this chapter focus on the presence of explicit references to concepts and named entities. Although one may conjure many types and subtypes in a taxonomy of entities, people, places and organisations are often the most commonly occurring classes of entities one might expect to find in news stories. Other types may be grouped as "miscellaneous". (A system working purely at the lexical level might use the words appearing in documents, but maintaining a history of words unlikely to be associated with content, or the topic of a document, would seem only to add unnecessarily to the model size. Motivated by the observation that Church's double Poisson word occurrence model (Church, 2000) found the most significant words were often nouns, simple words other then nouns were therefore discounted.)

Although the features are typically present, documents do not usually mark them out explicitly. Presence needs to be detected and the features then extracted prior to analysis using the model described below. Tagging and extraction tools may be employed for this purpose. Chapter 5 detailed the selected features and the extraction tools selected. Pre-processing for the analysis presented in this chapter made use of the Stanford CoreNLP framework Toutanova et al., (2003b); Finkel, Grenager, and Manning, (2005), using the pre-trained models contained therein. English part-of-speech tagging and named entity recognition was applied to each posting, yielding the features, and the number of their occurrences, associated with the document they occurred in and the time the document appeared on the internet. The MESME tool (Dewdney, 2017a) was used to extract multiword expressions. This was mainly to obtain compound nouns although verb particle and light verb constructs were also identified.

Features were extracted from each blog post and news article in the ICWSM corpus. They were then capitalised to ensure feature frequency calculations were case insensitive.

6.5.3 Temporal modelling of content streams

For simplicity a traditional Poisson model (Haight, 1967) is employed for each feature frequency: This assumes that features occur at random and independently, the intervals between occurrences being Poisson distributed. The reciprocal of the expected interval gives the expected frequency. If a random variable X has a Poisson distribution with expectation $E[X] = \lambda$ then

$$P(X = k|\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}, k \geq 0 \tag{6.1}$$

The mean frequency is simply the inverse of the expected gap between occurrences for the feature k, $1/\lambda$. The variance of the Poisson distribution is also λ so the significance of a trend can be measured as the number of standard deviations the associated gap reduction is from the mean. For feature k with expected frequency $\frac{1}{\lambda_k}$ and observed frequency $\frac{1}{\lambda'_k}$, the strength of a trend in k is given by:

$$T(k) = \frac{\lambda_k - \lambda'_k}{\sqrt{\lambda_k}} \tag{6.2}$$

The daily trend in feature occurrence is measured in standard deviations given by $T(k)$ from the expected frequency which is calculated as average observed frequency over preceding days. This does require a certain amount of "burn-in" time to establish a reasonable estimate of the average frequency $1/\lambda_X$ for each feature $X = 1, 2, ...$ Counts are calculated for each feature in each media category and Laplacian smoothing applied to account for unseen (new) features.

A trend in feature occurrence, then, can be measured in terms of the deviation from its mean frequency within an interval as determined from averaging preceding intervals. Feature frequencies are calculated on a daily basis with average frequencies being calculated on an accumulative basis, i.e. no "window " is applied. In a larger study a rolling interval may be more appropriate to account for long term drifts in language use.

For each experiment a bedding-in time of seven days was arbitrarily chosen. Thereafter, each day's feature counts were calculated and compared to the average, reporting at the start of the next day. So feature counts for 8$^{\text{th}}$ August are compared to the average calculated over the period from 1$^{\text{st}}$ August to 7$^{\text{th}}$ August and reported as at midnight, 00:00hrs, 9$^{\text{th}}$ August. On any one day in the experimental period the top trending features by deviation from the average shown by that feature to date are selected subject to a minimum of one standard deviation above the average. Any feature trending in the news is tagged and is not considered further.

Small changes in occurrences may be expected to be more significant for features that do not occur very often than for high frequency features. The use of Poisson models parameterised by feature counts for each stream affords relative trend measurements, so that deviations in feature frequencies that have different means in each media stream can be compared. Similarly, deviations in the frequencies of different features can be compared.

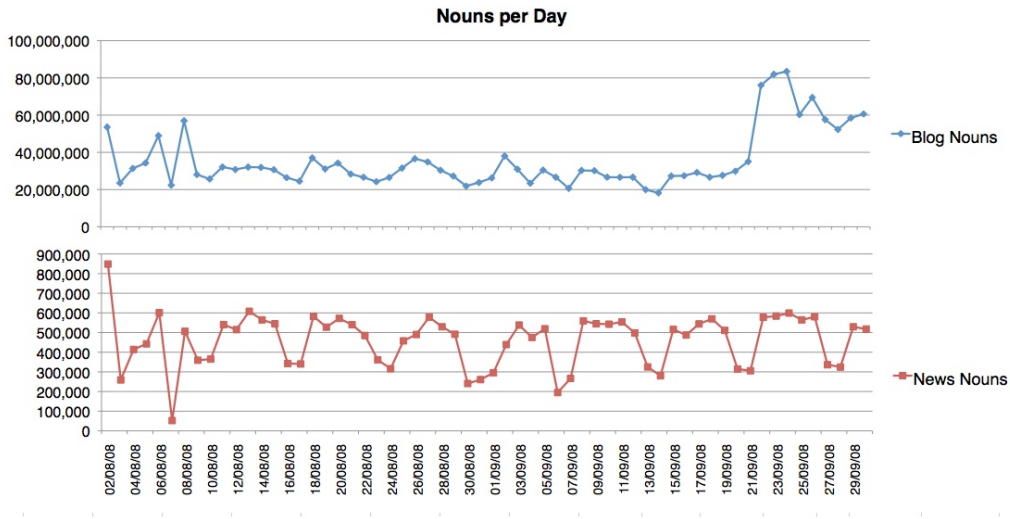


FIGURE 6.2: Total of nouns in blogs and news per day in ICWSM 2009 corpus

6.6 Experiments and Analysis

6.6.1 Initial Analysis

Over the two full months of data in the corpus, August and September 2008, there are 1,593,868 posts from mainstream news sources, while there are 36,740,061 blog postings. Of these, 1,428,482 (89.6%) news stories and 27,074,356 (73.7%) blogs contain at least one named entity (as identified by the Stanford NER tool), and all bar 157 blog postings contain English nouns (although there is no guarantee the post is actually in English).

The amount of material produced each day is not consistent however as can be seen from the graphs shown in figures 6.2, 6.3, and 6.4, although News postings show a periodic nature as one might expect. There is a notable increase in noun output in blogs but not in news towards the end of the period, although this increase is not seen in named entity output. The number of postings made per day shows no significant change suggesting that the rise in noun output is due to a relatively small number of long blog postings that do not mention a correspondingly higher number of named entities.

If News is taken to be representative of what is widely known, then for potential new information, one may focus attention on features that demonstrated a rising trend in occurrence either exclusively or prior to a trend in news articles. In the analysis performed here, a minimum criterion for feature selection was imposed: a minimum of over five occurrences, and a positive deviation of over five standard deviations from their average daily occurrence (calculated over all previous days available), on the day of their maximum positive trend during the two months of the corpus. Trends for features that have trended in news articles within the previous seven days were not considered.

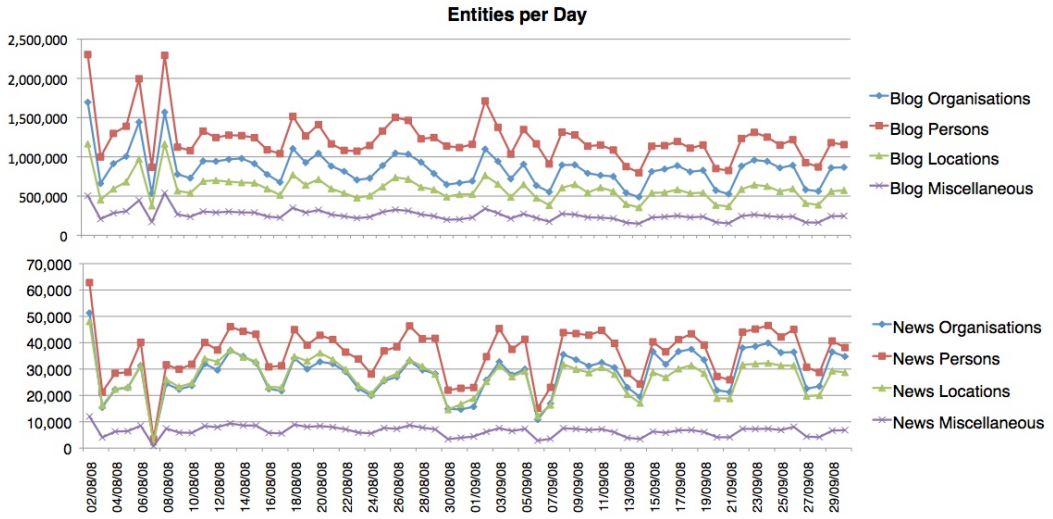


FIGURE 6.3: Total of named entities in blogs and news per day in ICWSM 2009 corpus

Type	No. Trending	No. in News Vocabulary	%
Nouns	7382	6260	84.8%
Misc	11260	5303	47.1%
Location	9809	5331	54.3%
Person	9365	5993	64.0%
Organisation	9823	5889	60.0%
Com. Noun	9780	9751	99.7%
VPC	3648	3574	98.0%
LVC	5736	5709	99.5 %
Totals	66803	47810	71.6%

TABLE 6.1: Number of unique features that have trended on at least one day in social media & amount in news use within ICWSM corpus

No trend analysis was carried out for the first seven days to allow a fair estimate of average daily occurrence to be established, so occurrences on the 8^{th} August were the first to be considered, being reported therefore on the 9^{th}.

This selection process yielded a total of 66,803 features that showed a positive trend originating in social media from the 8^{th} August 2008 to 30^{th} September 2008. An average of 71.6% of these trending features are also to be seen in news articles, though not necessarily trending there. The break-down across feature types is given in table 6.1.

A high proportion of nouns that show trending behaviour originating in blogs, about 85%, were found within the vocabulary of news articles. The lack of editorial control in social media, together with tagger inaccuracies, account for much of the remainder. A much lower proportion of named entities that originally trend within blogs are also seen in news at all. One may conclude that while some people, organisations, and places etc. may be of topical interest in the social media, only about half of them (between 47% and 64%) are also in the sphere of interest of the mainstream media organisations. Blog trending multiword expressions are predominantly found within the vocabulary of news. This is not surprising given their

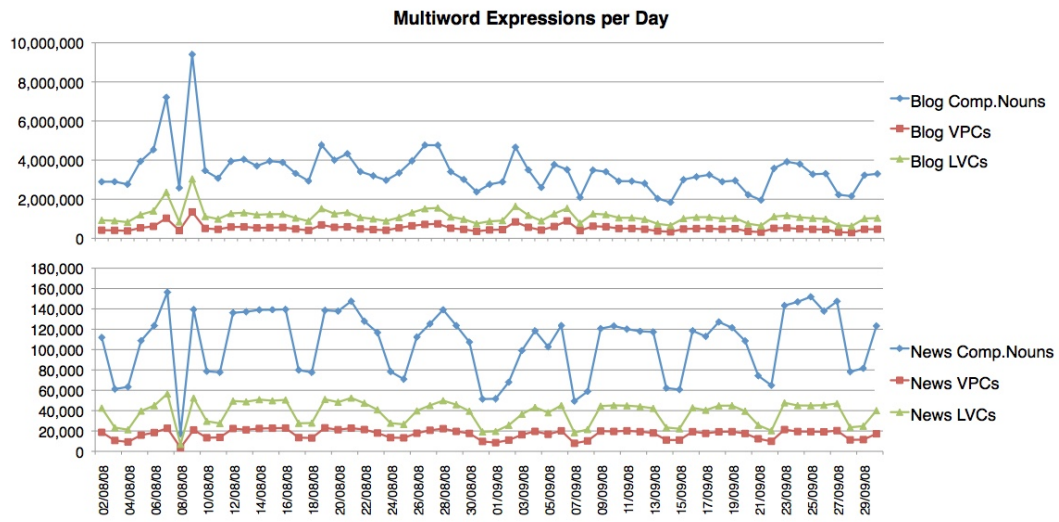


FIGURE 6.4: Total of multiword Expressions in blogs and news per day in ICWSM 2009 corpus

conventional form.

As trend strength is measured relative to the average occurrence of a feature rather than in absolute occurrence numbers, the most popular nouns and entities are not necessarily the same as those that show the strongest trend. Table 6.2 notes the top ten most frequent nouns and the top ten strongest trending nouns with their average frequency and trend strengths at the time of they showed their strongest trend. There are two observations to make: Firstly that a significant number of these "nouns" are not correctly identified by the part-of-speech identifier and named entity tagger, being either broken mark-up or proper nouns; secondly the strongest trends contain unidentified proper nouns.

Top Occurring			Top Trending		
Noun	Avg per Day	Max Trend	Noun	Avg per Day	Max Trend
QUE	67072.4	958.9	WON	184.2	92152.5
%	60444.3	1002.7	– – – – – – – – – – ...–	14.8	83949.8
THINGS	52755.2	410.4	3A	36.1	76574.3
COM	51408.7	5322.4	BEHAR	59.4	75912.3
SOMETHING	48963.0	547.2	PEARCE	87.6	69001.3
DA	46427.7	4269.0	PROP	163.2	68665.7
GIRL	44684.9	1255.9	<BR?/>	810.1	51689.2
MUSIC	44454.6	740.1	PIVEN	705.3	50291.6
DVD	44327.5	4001.1	ANTOFAGASTA	16.2	48510.5
EL	41919.9	666.7	JEUDI	142.1	46578.9

TABLE 6.2: Top ten 'nouns' by average daily occurrence and by trend strength in blogs

Tables 6.3 through 6.9 show the top ten by average occurrence and by maximum trend strength for Named Entity types and multiword expression types. The most frequent named entities are mentioned several thousand times a day (about an order of magnitude less than the most frequent nouns). Trend strengths shown by these named entities range from a few 10's of std. deviations from their average daily occurrence to a few thousand, similar in strength to the top occurring trending nouns. The trend strengths are typically well under those shown by the top ten entities

selected by maximum trend strength, which are in the region of several thousand standard deviations. These too are an order of magnitude less than trend strengths shown by the maximally trending nouns. Overall, mentions of people tend to appear in trends more strongly than those of organisations and places, as well as showing higher average daily occurrences.

Top Occurring			Top Trending		
Organisation	Avg per Day	Max Trend	Organisation	Avg per Day	Max Trend
GOOGLE	10443.0	303.9	ILWU	0.8	5929.3
APPLE	3138.6	577.7	ADM	24.4	5459.2
UA	3083.5	2359.9	OHIO STATE	211.1	5452.2
YAHOO	2279.1	2409.3	STATE FARM	15.4	5147.4
VMWARE	2142.2	1494.6	SOA	243.7	4935.9
HOUSE	2009.3	146.9	IBM	1944.4	4341.8
IDF	1945.2	1663.6	HEALTH MINISTRY	52.9	4122.3
IBM	1944.4	4341.8	BUCS	82.4	4000.3
MCKINSEY	1726.6	1059.5	ACORN	109.3	3967.0
HET	1641.5	1610.4	USAF	115.4	3965.7

TABLE 6.3: Top ten Organisations by average daily occurrence and by trend strength in blogs

Top Occurring			Top Trending		
Person	Avg per Day	Max Trend	Person	Avg per Day	Max Trend
OBAMA	34008.6	64.1	BEHAR	16.4	7119.4
MCCAIN	14677.7	53.3	KWAME KILPATRICK	517.2	6698.0
JOHN MCCAIN	12280.5	68.7	ALICE COOPER	63.2	6345.6
JACK	5415.7	3098.2	FREEMAN	111.9	6155.8
JESUS	3924.7	1994.1	CORSI	166.0	5619.6
JENSEN	2661.4	1675.2	MRS. CLINTON	71.3	5575.4
RYAN	2163.1	2375.9	OLMERT	134.1	5238.0
DAVID	2156.5	1503.5	SANTANA	72.1	5127.6
PETER	1703.3	2613.4	BUFFETT	93.4	5101.8
GOD	1688.5	2767.8	CARL ICAHN	37.0	5028.9

TABLE 6.4: Top ten Persons by average daily occurrence and by trend strength in blogs

Top Occurring			Top Trending		
Location	Avg per Day	Max Trend	Location	Avg per Day	Max Trend
NEW YORK	6267.8	67.2	GOLD COAST	10.6	4005.6
INDIA	6188.3	70.4	BISHKEK	103.6	3912.0
UK	4146.9	293.8	LIMA	358.6	3683.3
FRANCE	3269.7	56.4	WELLS	18.6	3674.3
FLORIDA	3207.1	69.3	WIRED.COM	91.3	3632.4
DELHI	3171.2	1805.5	HAMPTON	30.4	3630.7
IRAN	2755.7	269.6	CALCUTTA	54.8	3613.2
ISRAEL	2750.3	371.8	HULU	167.2	3491.8
LOS ANGELES	2320.7	74.1	YUNNAN	64.9	3463.6
PARIS	2110.3	183.4	TRIPOLI	180.4	3421.1

TABLE 6.5: Top ten Locations by average daily occurrence and by trend strength in blogs

Top Occurring			Top Trending		
Misc	Avg per Day	Max Trend	Misc	Avg per Day	Max Trend
INTERNET	6428.6	85.4	ALPS	42.8	2523.1
WINDOWS	1882.7	1416.4	GENESIS	76.9	2430.3
DEMOCRAT	1669.7	78.0	LITTLE LEAGUE WORLD SERIES	49.4	2263.7
GMT	1479.2	1183.7	SUMMER OLYMPIC	11.0	2050.8
ALS	1267.9	865.3	TEAM	21.3	2029.0
FACEBOOK	1217.0	1385.2	VIETNAM WAR	76.5	1928.7
TWITTER	791.6	1029.4	BOLIVARIAN ALTERNATIVE	3.9	1927.5
CHRISTMAS	786.4	1174.1	BRITONS	71.9	1772.3
JAVA	745.7	923.8	SERIE A	18.3	1765.6
MUSLIMS	703.2	665.0	CHINA OPEN	62.6	1696.2

TABLE 6.6: Top Ten Miscellaneous by average daily occurrence and by trend strength in blogs

To get a sense of any linkage between average daily occurrence and maximum trends strengths in social media originated trends, the two can be plotted against

Top Occurring			Top Trending		
Com. Noun	Avg per Day	Max Trend	Com. Noun	Avg per Day	Max Trend
FIRST TIME	5112.2	23.7	OPENING CEREMONY	24.8	729.2
LONG TIME	3033.3	45.7	PAST WEEKEND	104.0	282.6
FEW DAYS	2911.8	47.5	GAS PRICES	210.7	249.3
MANY PEOPLE	2904.4	47.6	NEXT WEEKEND	227.7	239.4
LAST WEEK	2817.5	29.9	OIL PRICES	247.7	229.1
LAST TIME	2325.5	57.9	BIRTHDAY PARTY	249.9	228.0
NEXT YEAR	1955.9	66.5	YOUNG WOMAN	294.8	209.0
NEXT WEEK	1682.4	82.4	OWN WAY	305.6	205.0
OTHER HAND	1636.5	75.8	CELL PHONES	309.8	203.5
ONLY THING	1411.1	83.9	BLUE EYES	311.9	202.8

TABLE 6.7: Top Ten Compound Nouns by average daily occurrence and by trend strength in blogs

Top Occurring			Top Trending		
VPC	Avg per Day	Max Trend	VPC	Avg per Day	Max Trend
POINTS OUT	501.9	34.9	CHANGED FREQUENTLY	0.0	816.3
CLICK FOR	477.0	35.9	HOLD FIRMLY	0.2	797.4
WORK TOGETHER	475.4	35.9	WOOT OFF	0.0	783.0
GOING DOWN	380.5	40.4	WRITTEN BY	42.4	122.9
COMMENT HERE	347.9	49.6	ROLL OUT	85.0	87.0
SOLD OUT	341.8	42.8	CLIPPED BY	135.6	77.3
FOLLOW UP	328.0	43.7	REFER TO	110.4	76.3
TOOK OVER	296.5	46.1	STOOD OUT	126.7	71.2
RUNNING BACK	290.0	46.6	LEARN FROM	129.3	70.5
POSTED BY	290.0	46.6	ALLOWED ME	130.1	70.2

TABLE 6.8: Top Ten VPCs by average daily occurrence and by trend strength in blogs

Top Occurring			Top Trending		
LVC	Avg per Day	Max Trend	LVC	Avg per Day	Max Trend
SEND TO A FRIEND	2933.7	18.7	'S GOOD	0.0	1540.9
MAKES SENSE	372.9	60.0	MOVIES/OPENSOURCE _ MOVIES	0.0	1537.5
ANNOUNCED TODAY	287.5	68.7	'S FUNNY	0.0	1537.3
HAVE A CHANCE	268.1	71.1	'M SCARED	0.0	1535.9
LOOK GOOD	220.0	78.7	DEL. ICIO	0.0	1535.8
HAVE A LOOK	189.4	84.9	PROVIDES A DETAILED ANALYSIS	0.0	1356.5
FEELS GOOD	156.7	93.4	PAYING OPTIONS	0.0	1262.0
SITTING NEXT	147.8	96.2	INCLUDING MARKET SHARES	0.0	1202.2
CHECK OUT THIS VIDEO	145.2	97.1	PUBLISHED AT DAVEANDSABRINA.COM	0.0	1159.6
ADDED SUPPORT	138.8	85.1	LOVE BLOCK COLORS	0.0	1157.4

TABLE 6.9: Top Ten LVCs by average daily occurrence and by trend strength in blogs

one another. Distributions can be further divided into those features that are unique to language seen in social media, those that are also seen in news articles and those that also trend in news articles after a trend is seen originating from blogs. (Ideally one would wish for these subtype distributions to have a clear separation for easy feature selection based on trend strength alone.) Plots for nouns and each entity type are shown in Figure 6.5 and for multiword expressions in Figure 6.6. For each feature that trends in social media, a point is plotted for its maximum trend strength against its prior average daily occurrence (an indication of how rare, or common, the feature was prior to the trend). A different plot is given for each feature type. The different subtypes (unique, also in news, also trending in news) are shown by colour and mark. From this the distributions of each subtype can be visually compared within a graph, and of types across graphs. Also given in the figures are the mean and standard deviation of the log distributions in occurrences and maximum trend strength.

Trending features that are unique to blogs were found to be fewer and weaker

	Unique		Also In News		Later News trend	
	Log Occurences		Log Occurences		Log Occurences	
	Mean	Std Dev	Mean	Std Dev	Mean	Std Dev
Nouns	0.70	0.87	1.73	0.75	2.84	0.68
Misc	-0.27	0.91	0.39	0.72	1.44	0.75
Locations	-0.13	0.91	0.49	0.72	1.50	0.80
Persons	0.02	0.85	0.74	0.86	1.72	0.66
Organisatio	0.01	0.83	0.66	0.79	1.65	0.69
	Log Max Dev		Log Max Dev		Log Max Dev	
	Mean	Std Dev	Mean	Std Dev	Mean	Std Dev
Nouns	3.37	0.26	3.97	0.17	3.84	0.42
Misc	2.44	0.22	2.59	0.24	2.91	0.21
Locations	2.60	0.21	2.80	0.23	3.08	0.28
Persons	2.72	0.24	2.97	0.21	3.30	0.19
Organisatio	2.65	0.25	2.92	0.21	3.21	0.18

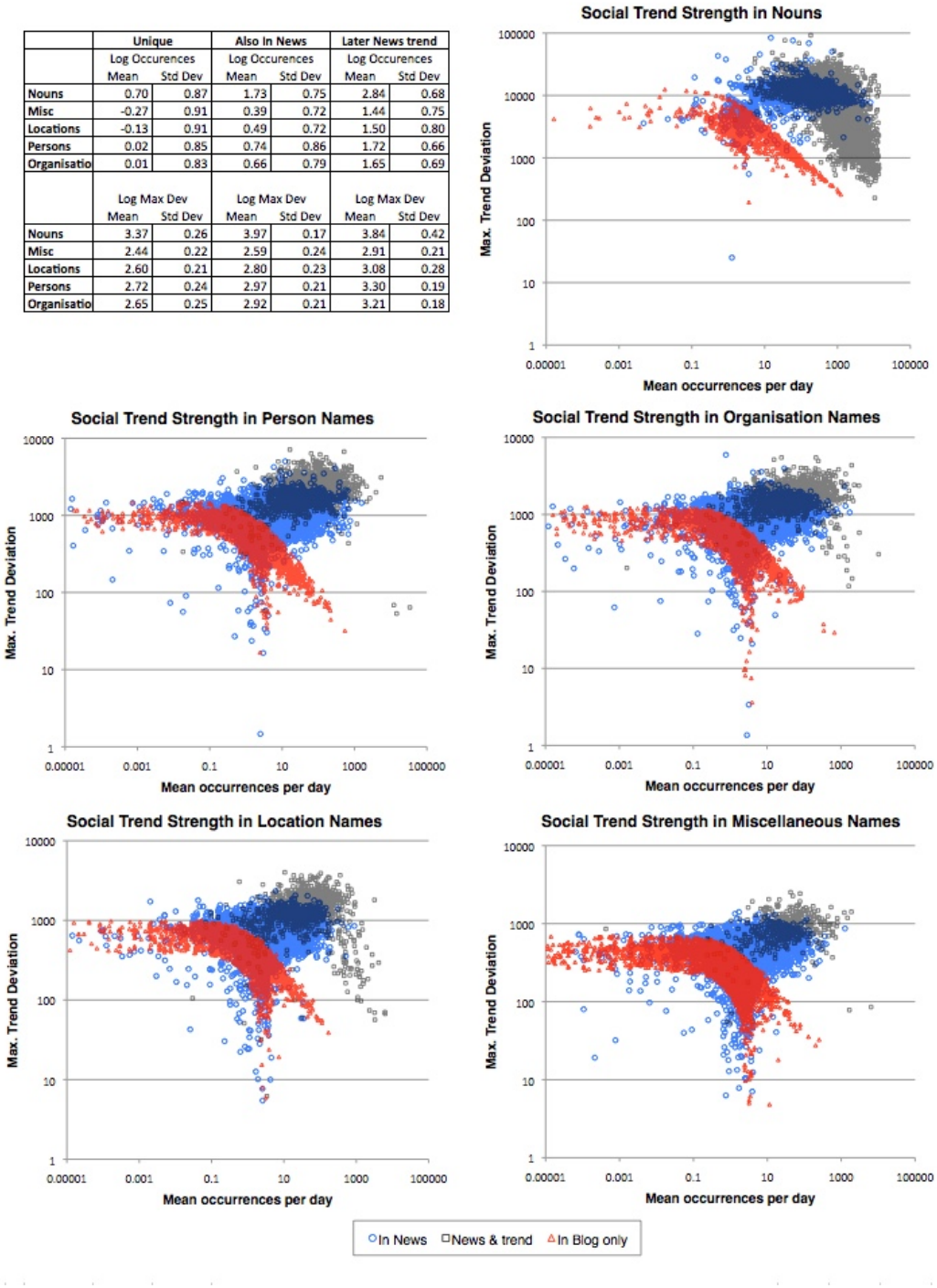
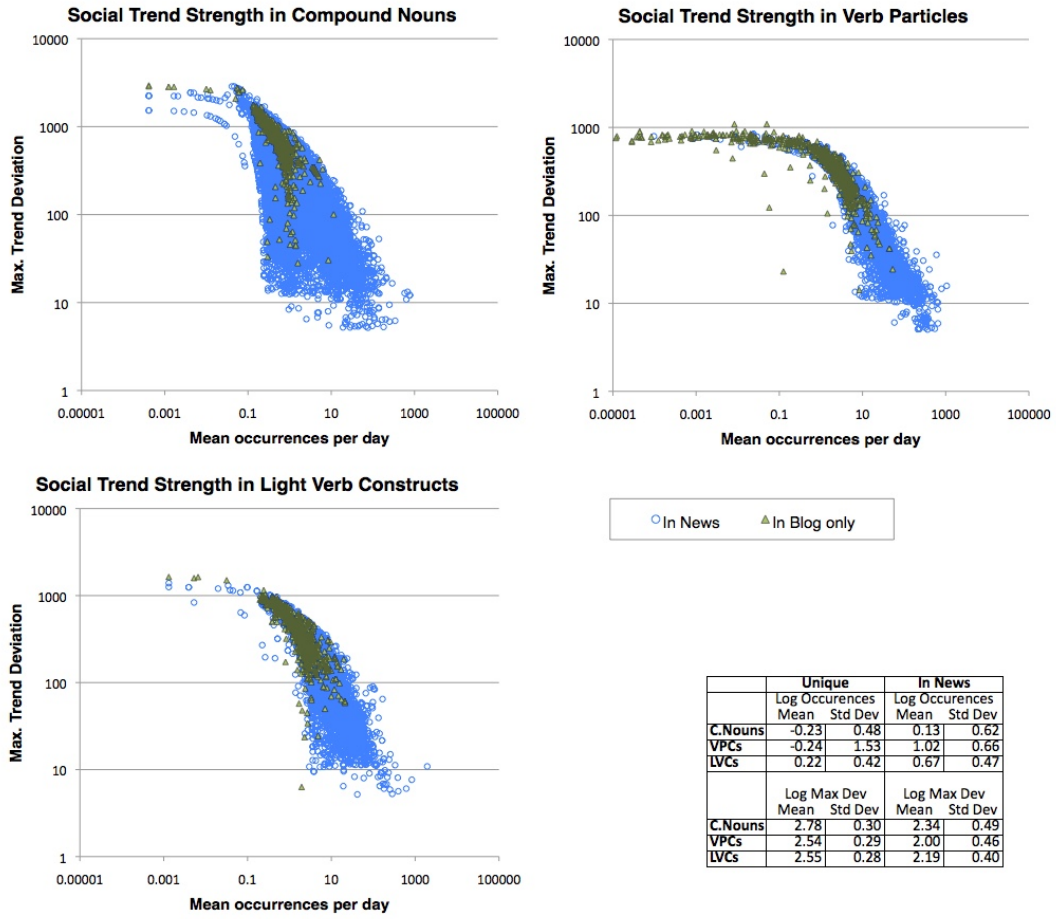


FIGURE 6.5: Distributions of occurrence per day and trend strengths for Noun and Named Entity trends originating in blogs



	Unique		In News	
	Log Occurences Mean	Std Dev	Log Occurences Mean	Std Dev
C.Nouns	-0.23	0.48	0.13	0.62
VPCs	-0.24	1.53	1.02	0.66
LVCs	0.22	0.42	0.67	0.47
	Log Max Dev Mean	Std Dev	Log Max Dev Mean	Std Dev
C.Nouns	2.78	0.30	2.34	0.49
VPCs	2.54	0.29	2.00	0.46
LVCs	2.55	0.28	2.19	0.40

FIGURE 6.6: Distributions of occurrence per day and trend strengths for Multiword Expression trends originating in blogs

than those that also appear in news vocabulary, although the separation is greatest for nouns. However, the presence of unique noun trends may be for the reasons of noise and tagger errors described above. Many trending named entities occurring uniquely in blogs showed average daily occurrence of less than ten per day. This was also found to be the case for multiword expressions.

Features that showed trends in news after the original trend in social media tended to be the most frequently occurring ones. Within these features, named entities also tended to show higher trend strengths and, while some nouns showed high trend strengths, nouns overall had a similar spread in trend strength to those not trending subsequently in the news: a separation in distributions of trend strength for features later trending in news and those not, is not present. Overall, distributions of features showing subsequent trends in news, given the feature type, have significant overlap with the distribution of features of the same type appearing in news articles without subsequent trends therein. The vast majority of those features showing subsequent trends in news articles had an average occurrence of at least one mention per day in blog posts prior to the day of the trend.

Multiword expression blog trends, in contrast to Named Entity and individual nouns, did not show any corresponding later News trend. More common expressions were more likely to be found in the News vocabulary, but otherwise the distributions in trends with vocabulary unique to blogs were within those with vocabulary in mainstream media.

These distributions suggest that entities being written about by bloggers that may be of wider interest at any particular time, tend to show trend strengths of a few hundred standard deviations from their average daily occurrence, although this could be less for very common entities (those with daily occurrence in excess of 1,000). Strengths for nouns in topics of potential wider interest tended to be an order of magnitude higher (average daily occurrence also being about an order of magnitude higher). However, this magnitude difference in trend strength was also true for nouns not subsequently trending in news articles. This suggests that comparisons between feature types would be better made having normalised by the average trend strength within a feature type.

The lack of any multiword expression News trends occurring after trending in blogs suggests that MWEs are unlikely to be useful features in isolation. This is perhaps not surprising for verb phrases as these in themselves do not constitute unambiguous reference. However, the finding for compound nouns seems at first to be at odds with that for singular nouns. The number of compound nouns is naturally less than the number of individual nouns. They also could be considered to be more specific in reference as in, for example, “guide dog” compared with just “dog”. References by singular nouns may be more likely to be unrelated than references by compound nouns. This in turn suggests that, in general, nouns could be very noisy features in new information discovery.

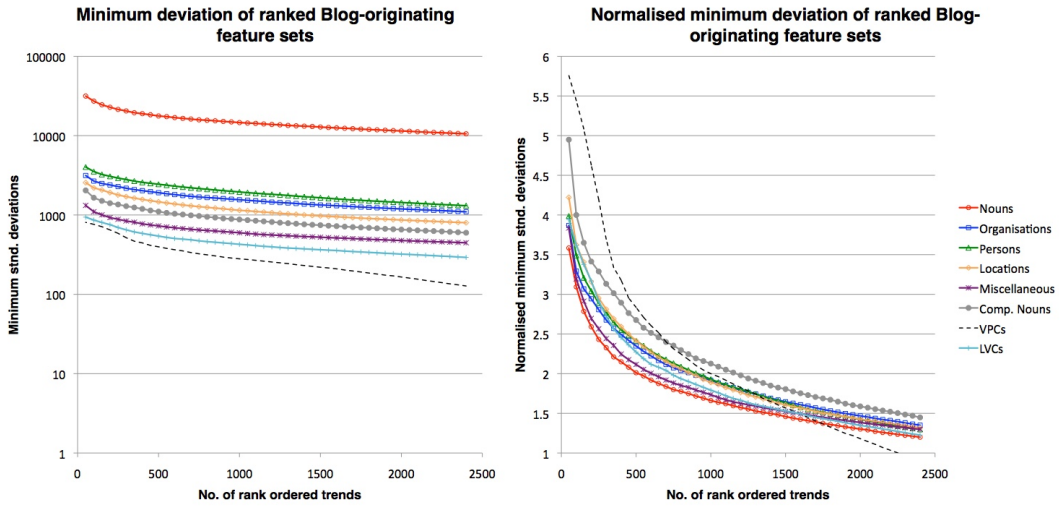


FIGURE 6.7: (a) Minimum trend strength of top n trending features (b) Minimum of top n ranked normalised trend strengths

Different feature types may be more or less prevalent to changes in use. To account for this when comparing trending behaviour across feature types one may calculate a normalised trend strength by dividing the trend strength shown by a feature by the average shown by all features of its type. The normalising factor for each type was calculated here by taking the average of the maximum trend strength shown by each feature, given feature type, over all the days covered in the corpus.

A comparison of maximum trend strengths in feature types given the top n trending features is shown in Figure 6.7: Graph (a) shows the raw trend strengths while Graph (b) shows the normalised trend strengths. Note that normalisation of trend strength by feature type average de-emphasises the dominance of nouns, while the relative difference between entity types shows little change, although Locations have slightly more prominence.

In a monitoring application, it is likely one would wish to select only the most significant trending features. This suggests applying a threshold to observed trend strength, raising the question of how well high blog trending features would predict subsequent trends in News. Graph (a) in Figure 6.8 shows the number of features trending later in News that would be selected from this corpus given a normalised feature trend strength threshold. Each feature type is plotted separately as well as for combined named entities. Note here that although the total number of trending entities outnumbers that for nouns, the spread of entity trend strengths is narrower. Multiword expressions are not plotted as no trends that appear in blogs subsequently trend in the news.

Figure 6.8 Graph (b) shows the trend strength (un-normalised for clarity) for features that also show a later trend in mainstream news articles, against that subsequent news trend strength. This plot shows that there is some correlation between the strength shown in the news trend and that shown in the original social media

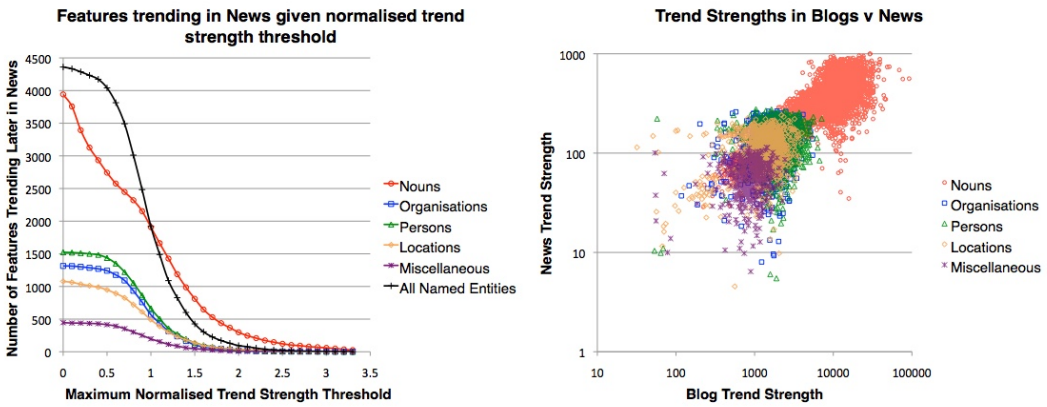


FIGURE 6.8: (a) Number of features selected for given normalised trend strength threshold (b) Relative trend strengths for those seen first in blogs and subsequently in news

trend. If trending in news stories is indicative of wider interest then this suggests that trending features originating in social media are quite likely to be of topical appeal. Furthermore, as shown above, many of these features are likely to be named entities.

6.6.2 Filtering News Trends from Social Media

Having shown that referring features show trends in their usage, and these may originate in social media prior to similar trends in mainstream news, the analysis turned to the potential of these trends for information discovery. Could filtering current news from social media by means of the trending key features could be used as a basis for discovering potential new information of interest? This subsection describes a simple method for doing this and an analysis of the resulting trends found in the ICWSM corpus. The following experiments only used nouns and named entities because no multiword expression trends had been observed to occur in blog posts prior to corresponding trends in news media.

Recall that on any one day in the experimental period trending features are found by positive deviation from their average daily frequency to date. The set of such features can then be ranked according to strength of their deviation from daily average. Further filtering criteria may be applied to reduce noise from very rare features and relatively low deviation. In this analysis trending features are selected subject to a minimum of ten standard deviations above the average, and having more than five occurrences on the day. These thresholds were chosen somewhat arbitrarily having observed the distributions of values for features. The aim was to ensure that features are not selected as trending due to movement caused by natural variation and to have some resilience to poor frequency estimation for those that are very rare.

To filter news from blog trends, any feature trending in the news was tagged and not considered for blogs appearing in the following seven days. If a news-trended

feature did not trend in the news for a period of seven days it was then re-established as potentially blog-trending.

Features examined for trends and compared were, as before, Nouns, Person names, Place names, Organisation names and Miscellaneous entities (others of undetermined type). The numbers of trending features in the period, together with the numbers that subsequently trended in the mainstream news media are shown in Table 6.10. (Features may have trended more than once during the period but are only counted once in these results.) Note that the total number of trends for each feature, excepting nouns, is less than that given in Table 6.1 due to the the more stringent thresholds applied. More nouns are selected because once a feature has not trended in the news for 7 days it is re-qualified as a potential trend.

Type	Trenders	News post trend	%
Nouns	9450	1741	18.4%
Misc	4350	221	5.1%
Location	4650	571	12.3%
Person	5450	740	13.6%
Organisation	5250	589	11.2%
Totals	29150	3862	13.2%

TABLE 6.10: Social media originating trending feature totals & amount subsequently trending in news

Blog trending features may not necessarily trend subsequently in the news, or even appear at all. Since the application in question here is finding features that are likely to trend in the news, some ranking of trending features is required. Trend strength would seem to be a natural choice, but having separated all the trending features into classes of those that subsequently trend in news, those that appear but do not trend in the news, and those that appear uniquely in blog vocabulary, one finds that the rank ordering favours unique vocabulary. Ranking by feature frequency on the day of the trend, however, favours vocabulary that is in the news, including that which trends. The proportions are shown in Figure 6.9 graphs (a) and (b) for nouns, graphs (c) and (d) for named entities. Note that ranking by frequency does not yield a proportion of subsequently trending features that is proportional to the number of ranked features. This is more marked for nouns than named entities.

Graphs (e) and (f) in Figure 6.9 show the proportions of subsequently trending features and unique features respectively broken out by feature type. Miscellaneous entities are most likely to be unique to blogs, whereas Locations are the most likely to trend subsequently in news for the highest ranking features. However, overall, subsequently news-trending nouns are more likely to be selected than any particular entity type while being least likely to be a feature unique to blogs.

It may be that a more sophisticated ranking metric could be employed to improve selection of predictive features. However, feature frequency on the the day it trends is used here for the rest of this study.

For each feature type the top fifty blogs trends, having not trended in the previous seven days of news stories, were further examined. Note that this represents only a

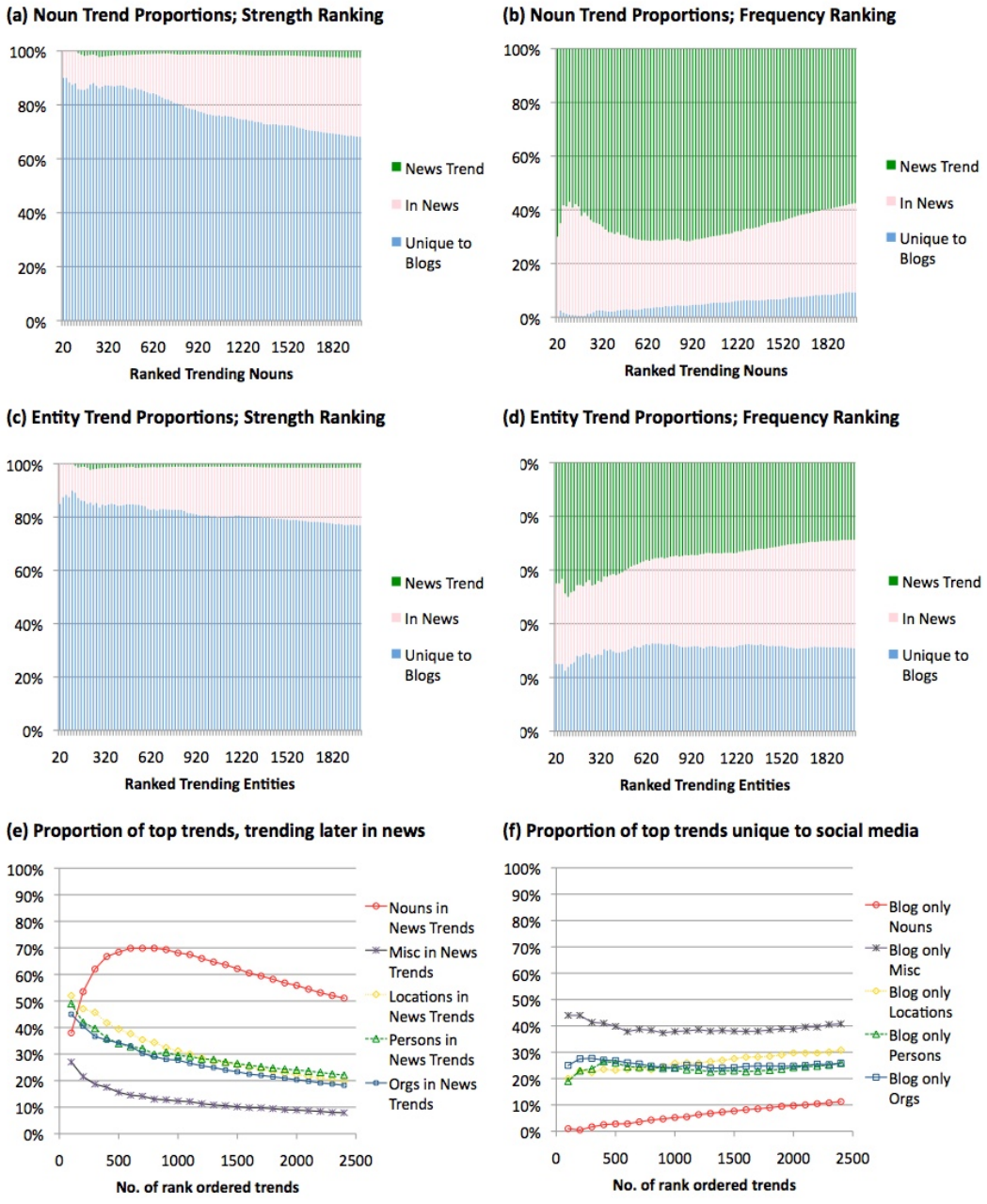


FIGURE 6.9: Proportions of blog trends that are unique to blogs, appear in the news, and subsequently trend in the news, when ranked: (a) Nouns by trend strength, (b) Nouns by frequency, (c) Entities by trend strength, (d) Entities by frequency. (e) Precision in frequency ranked feature types. (f) Unique features in frequency ranked feature types.

minority of all the trends found in the social media, but the most significant for the purposes here. Although only the most significant trends originating in social media are examined, one should not expect all trends to be reflected in subsequent news stories; they may simply not reflect news-worthy material, or have been overlooked by media organisations. This is indeed the case with on average 38% of blog trending features (46% of top fifty by type) seen subsequently trending in news.

Table 6.11 shows the strongest trending features of each type that pre-empt corresponding trends in the news media by occurrence on the day of the maximum trend. Trending features may trend more than once and, during the time they are valid as trends, they may yield different trend strengths. The date of the maximal trend is taken (being the most likely point at which the trend would rank highly). If there is a subsequent news trend for a feature then the difference in date of the news trend and the social media trend gives the time lag in days.

Of the top 50 noun trends originating in blogs, 19 subsequently trended in news stories. However, on inspection one may observe that some of these are not generic nouns at all. (This may also be observed through lower ranks as well.) Part-of-speech tagging errors, e.g. "DIE", and named entity recognition errors, e.g. "OBAMA" suggest that the true number of trending generic nouns is actually less than that measured. Looking at the time difference between when a blog trending noun showed its strongest trend and when the subsequent news trend first occurs, one may observe a spread ranging from 2 to 35 days. Average lag for the top ten nouns is 11 days (9 days for all 19 trending nouns in the top 50). Long lag times are likely to have arisen from unrelated (or at least separate) news events having the featuring the same noun.

For named entities, typing appears sound with only the odd questionable tag type. Person names do not appear to be very specific suggesting that co-reference resolution could be beneficial. Varying counts in different mentions can result in trends being split over periods. This was observed, for example, with "National Convention" and "Republican National Convention" which trended on Aug. 15th and Aug. 16th respectively, with the corresponding news trends occurring on the following two days. As with nouns, there is a wide spread of lag times for named entity mentions (1 to 30 days). Again, features showing long lags times are likely indicative of different underlying events or stories.

Average time lag between maximal trending in blogs and first subsequent news trend for the top ten entities are 9 days for persons, 12 days for organisations, 3 days for locations, and 12 days for miscellaneous. Overall the average lag for all entities is 6 days. If one supposes that news articles containing trending features that originally trended in social media a relatively short time before are more likely than those that trended some time previously to be topically linked, this suggests that named entities are more likely than nouns to be predictive for news topics. Topical linkage is examined below in subsection 6.6.3.

Feature	1st Blog Trend	Avg. Frq. in Blogs	Trend dev.	Max Blog Trend	Max dev.	Posts	News Trend	Avg. Frq. in News	Trend dev.	News Posts	Lag
Nouns											
DESIGN	09/08	396.4	26.3	09/08	26.3	5719	11/08	143.9	30.6	128	2
MOVIE	16/08	540.4	21.5	16/08	21.5	5628	30/08	604.5	10.2	167	14
VIEW	09/08	776.6	17.87	09/08	17.9	5302	18/08	269.9	11.5	147	9
FRIEND	06/09	679.3	17.7	06/09	17.7	21016	08/09	235.2	12.7	158	2
OBAMA	23/08	859.9	16.7	23/08	16.7	1567	27/09	2381.2	12.5	471	35
Persons											
OLLIE	09/08	102.5	13.5	17/09	145.5	53	28/09	9.5	138.6	7	11
RILEY	10/08	185.7	59.2	10/08	59.2	63	13/08	18.5	122.2	2	3
WENDY	10/08	210.7	45.6	10/08	45.6	104	12/08	17.2	114.1	6	2
RODNEY	12/08	300.9	38.2	13/09	44.2	41	19/09	16.5	126.2	1	6
DEXTER	10/08	125.9	23.5	20/09	41.8	54	26/09	14.9	148.8	5	6
Locations											
PUNE	10/08	139.6	79.6	10/08	79.6	60	11/08	112.5	48.0	3	1
CHENNAI	10/08	265.9	54.8	10/08	54.8	75	11/08	153.6	37.6	10	1
MASS.	12/08	199.9	39.8	12/08	39.8	185	14/08	499.2	15.2	17	2
PORTSMOUTH	09/08	213.9	23.5	25/08	33.9	339	08/09	1549.1	17.3	65	14
DELHI	10/08	675.6	29.2	10/08	29.2	157	11/08	198.2	39.4	4	1
Organisations											
STD	10/08	77.2	69.2	13/08	86.6	59	12/09	0.1	157.4	8	30
FANNIE	10/08	187.4	24.7	05/09	72.1	102	06/09	37.3	112.2	8	1
AG	11/08	1.2	71.3	11/08	71.3	543	08/09	10.8	99.7	13	28
MAC	09/08	290.6	44.3	09/08	44.3	450	13/08	109.2	28.0	7	4
NASCAR	09/08	499.9	16.5	01/09	35.7	289	27/09	2496.5	10.7	115	26
Micellaneous											
MISS CONGENIALITY	09/08	10.0	151.7	16/08	188.8	4	29/08	4.5	75.6	6	13
NATIONAL DAY	09/08	716.5	24.0	16/09	119.3	48	29/09	11.2	58.7	11	13
PGA CHAMPIONSHIP	09/08	363.7	21.5	20/08	64.1	22	21/08	170.4	39.1	6	1
TIBETAN	11/08	454.5	11.8	20/09	37.5	23	25/09	158.8	43.1	9	5
TURKS	10/08	357.6	12.2	23/08	29.9	85	31/08	101.4	47.7	5	8

TABLE 6.11: Top 5 trends for each feature type occurring in weblogs prior to mainstream news, showing time and trend strength in standard deviations from average daily occurrences for first trend occurrence, the maximum trend occurrence, and subsequent trend occurrence in news

One may note that the number of mentions of selected terms in blogs per day is greater than that in news stories. However, as the counts have not been normalised for number of posts, this should not be surprising as the number of blog posts is much higher than mainstream media articles (by about 20:1 for this corpus). Blog posts mentioning top trending nouns are particularly higher in number than the number of corresponding news articles. This factor is generally less for named entities and in one case, "MISS CONGENIALITY", there are more news stories at the point of subsequent trend than posts in the original trend. One explanation for this could be the higher number in average use of nouns compared with particular named entity mentions indicating a large background use of the nouns in question. This would suggest that trending named entities are more topically specific than trending nouns.

Figure 6.10 shows the evolution of the top two trends for each of the feature classes. In these graphs, only positive trending behaviour is shown. The trend strength is measured in number of standard deviations from the expected value. Trend activity in news is plotted on the negative y-axis (trend strength x -1). A solid line on the plot indicates when the trend behaviour in the blog is valid, i.e. not filtered from a previous or current news trend. Note that with even within the top selections illustrated one can observe a range in feature trend patterns. It may be possible to identify some patterns for features that would allow elimination or promotion (e.g. periodic trending). This is left for future work.

Just because a feature may trend in social media and then in the news does not

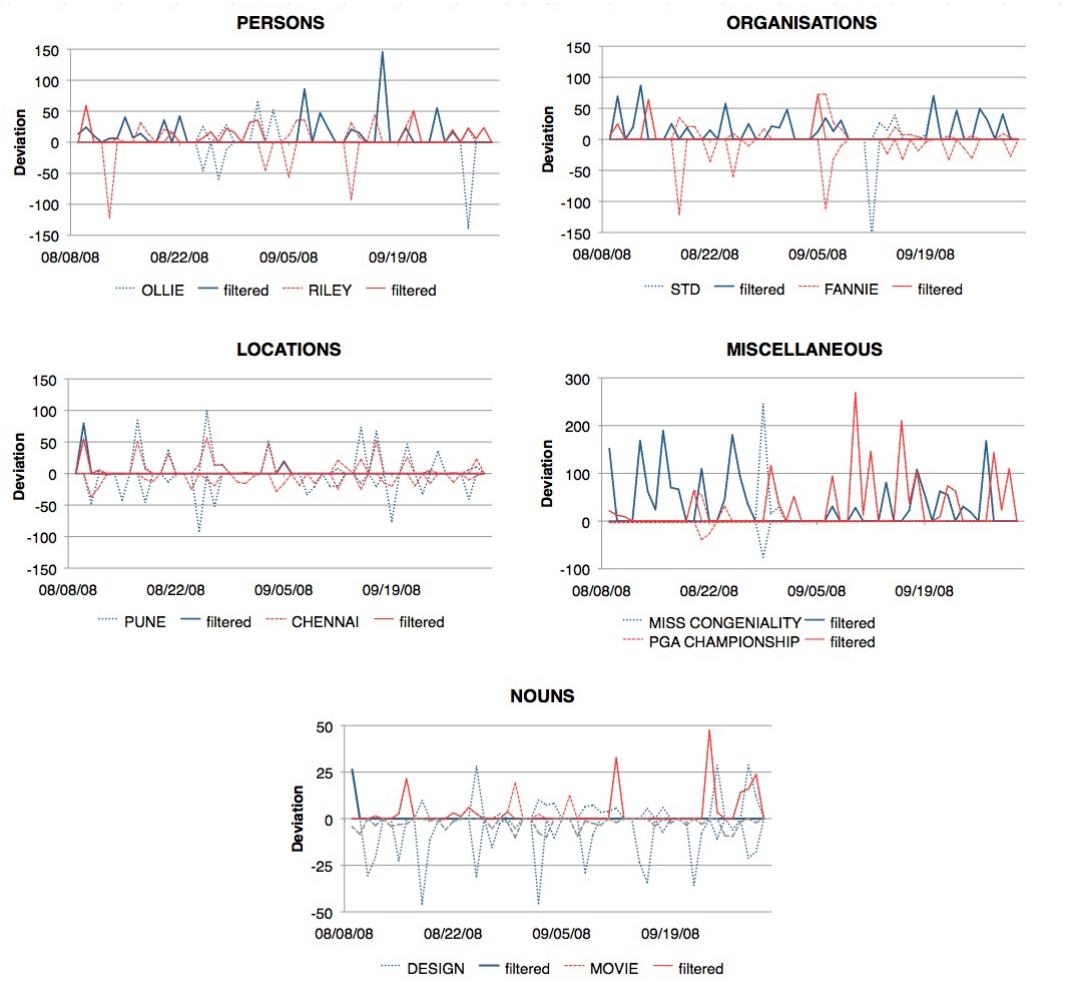


FIGURE 6.10: Trend history, shown as positive deviation from average count, for top two features of each feature type, trending in weblogs prior to news. Weblog on positive y-axis; News on negative y-axis.

mean that they are topically linked. It is well known that some features, be they generic nouns or named entities, are more specific in what they refer to than others. Blog posts, and news articles, that give rise to the trends may have a range of topics themselves, and one may observe this in posts giving rise to the trending features selected here. Unsurprisingly many topics can be found in posts giving rise to trends in generic nouns seen here. For example posts featuring "design" on 9th August include such diverse topics as the Olympics, a design conference, website design, guild badges, and peta-scale computing to name just a few. Subsequent news stories for "design" are similarly diverse, covering jobs, product reviews, etc. Even more specific nouns such as "Movie" arise from multiple topics, with reviews, discussion and even postage of amateur footage amongst blog postings on 16th August, while the mainstream media reports in the trend 14 days later cover unrelated reviews, overall summer box-office success and a Star-Trek convention!

Named entities fare little better. For example, trending location entities often result from a higher than average number of posts that refer to different events from

those places. "Chennai" arises from posts about events in the region as well as travel plans. Similarly "Mass." arises from many posts about activities from multiple locations within Massachusetts. Subsequent news stories also have a wide spread of topics. However, some topics referring to a location are related to later news stories: "Pune" trends from posts of multiple topics including rain storms, Indian celebrities, and personal travel, but also commentary on a gang rape and murder that had occurred there. The subsequent news story is about lack of progress in the case.

Some named entities are more specific, though, and therefore topically more predictive. "Fannie" refers to the U.S. mortgage company Fannie Mae. The economic crash of 2008 had just got underway when ICWSM 2009 corpus was collected, and there was much speculation about whether Fannie Mae (and its counterpart Freddie Mac) would need a U.S. government bailout. The news story trend occurred as various U.S. political figures reacted to market concerns, calling for appropriate aid, the following headline from the Chicago tribune being typical:

> "U.S. plans mortgage bailout"
>
> - http://chicagotribune.com/business/

Another example where an entity name was sufficiently specific to a topic is "PGA Championship" which referred to the 90^{th} PGA golf championship. Blogs centred on discussion surrounding players ahead of the playoffs as in:

> "Jack Nicklaus Isn't Sure if Sergio Garcia Will Ever Win a Major, Colin Montgomerie Stoked"
>
> - http://golf.fanhouse.com/2008/08/20/

while the news articles the next day reported the opening session of the playoffs:

> "Mahan fires 62 to open PGA playoffs"
>
> - http://sportsnet.ca/golf/2008/08/21/

> "Mahan leads by 4 shots at The Barclays"
>
> - http://cbs.sportsline.com/golf/story/10942213/rss

Topic specificity would seem to be important to find meaningful and potentially predictive information. Employing methods to refine trending features into the topics that gave rise to them is likely to be beneficial therefore. The analysis presented next employs a relatively simple method for doing this.

6.6.3 Trend topics

One may argue that to impart information about something one must express some relation between it and some other concept. The collection of relationships between the key concepts, such as entities, expressed in a document could be said, then, to be the document's topic.

The idea of discovering information through co-occurrence of key concepts is one that has seen some notable success, particularly in the medical research domain (e.g. see Swanson and Smalheiser, (1999), Srinivasan, (2004)). Here a similar technique to Alvanaki et al., (2011) is employed. Taking pairs of trending features (a trend bi-gram) to be the key concepts, their co-occurrence in source blog postings is examined. The idea is that frequently co-occurring nouns and named entities would indicate some trending topical information regarding them in connection. But are the documents that have given rise to a trend bi-gram topically connected? The final step is evaluate this for each resultant set of documents given a trend bi-gram, but doing so is difficult because, as discussed in Chapter 2.1, the notion of topic is not well defined. However, techniques have been proposed since document clustering based topic modelling approaches, such as LDA (Blei, Ng, and Jordan, 2003), have been adopted, and the need to evaluate their effectiveness has arisen.

Various approaches to automatically assessing *topic coherence* – how related documents in a cluster are – have been proposed. Wallach et al., (2009) used model perplexity as a basis, while Newman et al., (2010), noting lack of qualitative understanding in this approach, examined various potential measures for topic coherence over several document collections and compared their results with human assessment. They found notable inter-annotator agreement and that a metric using point-wise mutual information (PMI) based scores derived from Wikipedia provided near perfect agreement with humans. Mimno et al., (2011) found that this technique could be improved by using conditioned document frequencies for terms, and co-occurrence thereof, rather than term frequencies derived from reference corpora. These two coherence measures are referred to as the UCI metric and the UMass metric respectively. In evaluating three topic modelling approaches, Stevens et al., (2012) found that the two topic coherence measures often agreed but that the UCI PMI metric induced more separation. Although the techniques were developed for assessing the results of topic models, they may provide the basis for assessing how well particular trend bi-grams select related documents.

Counts of trend co-occurrences are calculated on a by-blog-post basis for the day the individual trends occur. Named entity types are not examined separately here so that the analysis includes finding information across different feature types (e.g. a Person and a Location). If one assumes that a topically related collection of documents would contain more than one feature in common, and that for trending stories this would include more than one trending feature, one can examine topical consistency in posts selected by a single trending feature and those selected by a trend bi-gram [3] (i.e. those posts containing two particular trending features, one of which is the originally selected single trending feature). Correspondingly, this will give an indication of how topically specific any single trend feature is: the greater the number of sub-groups of documents the trend bi-grams split a trend uni-gram set into,

[3] Not to be confused with a traditional word bi-gram which is made up of two sequentially appearing words.

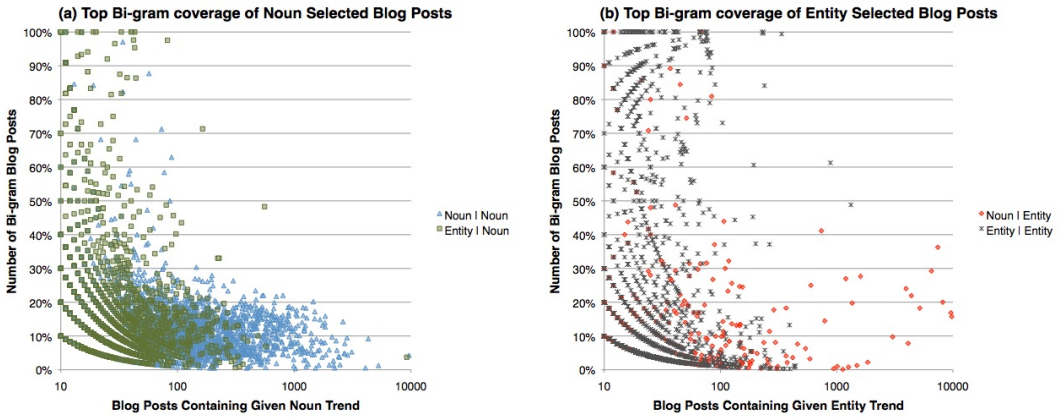


FIGURE 6.11: Distribution in a one week sample of top trend bi-gram coverage of posts selected by constituent unigram: (a) Noun; (b) Entity

the less topically coherent the unigram set is likely to be.

For each trending feature (trend uni-gram) all posts containing it on the day of the trend are retrieved. Trend bi-gram totals are calculated by counting the number of these posts that also contain another trend-unigram. The potential number of trend bi-grams is naturally much higher, being $O(n^2)$ where n is the number of individual trending features. Here, then, only the most likely bi-gram to characterise the collection of posts containing the given uni-gram i.e. that with the highest number of posts (or those with the highest in the case of a tie), are considered. The graphs in figure 6.11 show plots of the number of posts for the bi-gram against the number of posts for the given uni-gram where there are a minimum of ten posts: The graph in (a) shows the distribution for bi-gram types given a noun type uni-gram taken over one week; the graph in (b) similarly shows that for types given an entity type uni-gram. One may observe that there is a spread in all cases indicating that in general selected posts are not topically consistent although some collections are more consistent than others. It seems that bi-gram types containing entities appear further to the top left in each plot suggesting that post collections that are more likely to be topically consistent are more likely to be characterised by entities than nouns.

One may consider whether or not trend bi-grams containing entities are more topically specific further by comparing the number of posts different bi-gram types select, not just those which select the most posts given the unigram. This may be done by rank ordering the proportion of the uni-gram selected posts each bi-gram appears in and comparing the proportion of bi-gram types at varying bi-gram/uni-gram post proportion thresholds; the assumption being that the higher the proportion of bi-grams, the more likely that the sub-set of posts are to be topically linked.

For example, if the trend bi-gram 'Fannie Mae | Obama' appeared in 5 of a set of 100 documents the uni-gram trend 'Obama' selected then its score would be 0.05. If the trend bi-gram 'Freddie Mac | Obama' scored 0.10 given the same documents, it would be ranked higher and therefore considered a better characterisation of the

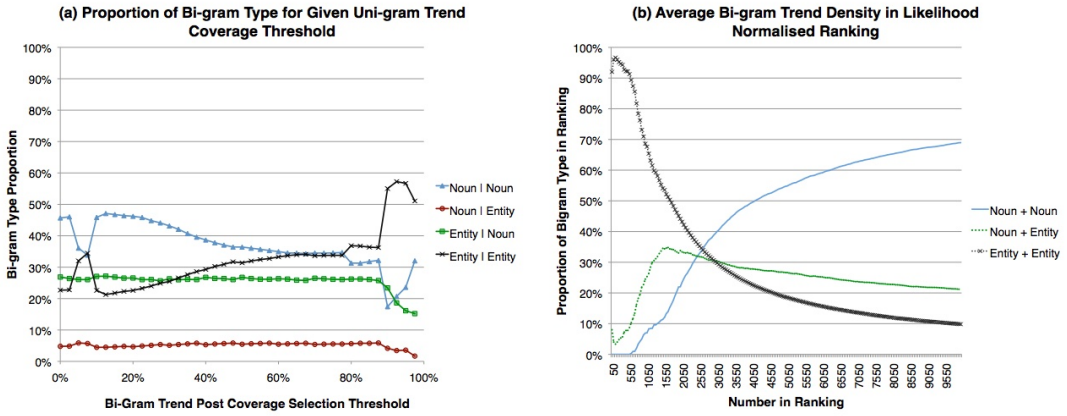


FIGURE 6.12: Bi-gram type densities: (a) when selected as proportion of uni-gram selected blog posts; (b) when ranked by ratio of observed blog post frequency to expected frequency given independent constituent uni-grams

document set. If a threshold on the rank can is applied, only those documents containing the bi-grams with a higher rank would be considered topically linked. So if a threshold of 0.07 is applied in this example, 'Fannie Mae | Obama' would not be considered a characterising bi-gram.

The proportions of trend bi-gram types present in a unigram selected document set following the application of the score threshold were examined to see which types most characterise the more topically linked sets of documents (i.e. those with a higher proportion of at least one common trend bi-gram). To adjust for the fact that there are a different number of trending entities than nouns, the the results are weighted accordingly. Graph (a) in Figure 6.12 shows that when bi-grams are present in more than 90% of the posts containing a given uni-gram trend, there is a better than 55% likelihood that both the trending features will be entities, while there is less than 31% likelihood that both trend features in the bi-gram will be nouns.

Rather than comparing bi-gram selected posts to those selected by one of the constituent uni-grams, one may rank order all the bi-grams of trends by the number of posts they appear in. However, although individually a trend feature is more likely to be an entity than a noun, nouns appear in more posts than entities. Ranking by number of posts would unduly favour bi-grams that contain nouns. A correction for this is to weight the number of posts a bi-gram is observed in by the number one would expect to see by chance co-occurrence of the constituent uni-gram features. This gives a ranking score based on the point-wise mutual information (PMI) metric that is similar to the topic coherence metric proposed by Newman et al., (2010), which sums PMI in term frequency, but is calculated intrinsically using document frequency. (Note that this is the Mimno et al., (2011) UMass metric, but modified slightly to condition on either feature in the co-occurrence pair.)

Denoting by $S_t(a)$ the number of posts containing feature a on day t, in other words the feature's document frequency if considering a post to be a document, the PMI

for a bi-gram feature $\{a, b\}$ is given by:

$$R(a,b) = \frac{S_t(a,b)}{S_t(a)S_t(b)}P_t \tag{6.3}$$

where P_t is the number of posts containing at least one trending feature on day t.

Returning to the example above, if 'Fannie Mae' appeared in 50 documents, and 'Obama' appeared in 100 documents, the bi-gram of the two appearing in 5 documents, then $R('FannieMae',' Obama') = (5/50 * 100)P_t = 0.001P_t$.

Once again, the proportion of each bi-gram type may be examined given a bi-gram trend score minimum threshold, to obtain an indication of which are more selective of more topically coherent document sets. Graph (b) in Figure 6.12 shows the proportion of each bi-gram type in an averaged rank ordering of the daily bi-gram trends. Bi-grams consisting of two entities are found predominantly at the top of the ranking, followed by those with including one entity. Assuming from the analysis above that bi-grams consisting of entities are more topically specific, this suggests that trending topics can be found by selecting posts containing bi-grams of trending entities scored by how unlikely they are to co-occur at random.

But to what extent are features in co-occurrences independent of each of other? This is explored in the next section.

6.6.4 Feature co-occurrence

The point-wise mutual information metric (PMI) compares the frequency of co-occurrence of two random variables with that expected if the probability of each is independent of the other. The metric can be normalised (NPMI) to give a scale of -1 to 1 where 0 indicates independence. From information theory, the amount of information given by the occurrence of random variable B given the occurrence of variable A is zero if $P(B|A) = 1$ or $(P(B|A) = -1$, i.e. occurrences of A and B are completely correlated, or anti-correlated. Maximal NMPI values, given sufficient samples for reasonable estimation, would therefore indicate no information is given by the co-occurrence. However, it does not follow that maximal information is yielded when NPMI equals 0, as information theory relates to information *capacity*. At zero NPMI, feature occurrence is totally independent. While it is possible that two independent features could co-occur due to some relationship, it is not possible to infer this statistically. However, if information is being conveyed relating two things, one might reasonably expect features indicating those things to co-occur more often than expected (as has been found in measuring topic coherence through PMI by Newman et al., (2010)). If new information is given relating two things then, assuming an increase in co-mentions of those two things, there will be an increase in the measured NMPI.

At higher values of NPMI there is less capacity to be informative, but the co-occurrence is more likely to be indicative that information is being conveyed. (It is worth noting that the basis of discovering new information as explored in this chapter is through divergence from expected co-occurrence; there is more capacity for features that more rarely co-occur to convey information than those often seen together.) Across a corpus of text that is presumably comprised of documents intended, in least at part, to convey information, one would expect co-occurring features associated with information being imparted, then, to yield higher NMPI values than those co-occurring through chance.

Whether information is new, or not, is not the key to the analysis here. It is assumed that information has been imparted in the corpus documents. The question is: which co-occurring feature types are more associated with imparting information, i.e. have higher NPMI on average?

Figure 6.13 shows the distributions of NPMI for the most common Nouns and Named Entity mentions in the mainstream news and blog posts in the ICWSM corpus. Sampling the most frequent occurrences of each feature type in each document class resulted in approximately 1 million co-occurrences (excluding co-occurrence with the same feature) for each distribution. As has been shown, the mainstream news contains a smaller vocabulary than found in blogs, and contains a smaller number of focussed topics (reports). Correspondingly the NPMI distributions have higher values in news than the corresponding ones found in blogs. In both cases, though, the distributions found for Named Entity co-occurrences are higher in NPMI value than those found for co-occurrences of Nouns.

As it has been found that PMI, and hence NPMI, can be used to measure topic cohesion across a set of documents, one may expect that topically more cohesive documents to have higher average NPMI for co-occurring features therein. One may also assume that news contains informative text. These distributions therefore provide supporting evidence for hypothesis $H4$ in showing Named Entity co-occurrence is less independent than Noun co-occurrence. The evidence from Blog posts is weaker, however one might expect that only a proportion of these will be informative, and also that a wider range of topics will be covered therein.

6.7 Summary

This chapter has considered whether there is evidence to support the idea that social media content could be used to predict or inform news stories. Analysis was performed using the ICWSM 2009 dataset which comprises categorised postings from the internet. Those categorised as either mainstream news or blogs were used in the study.

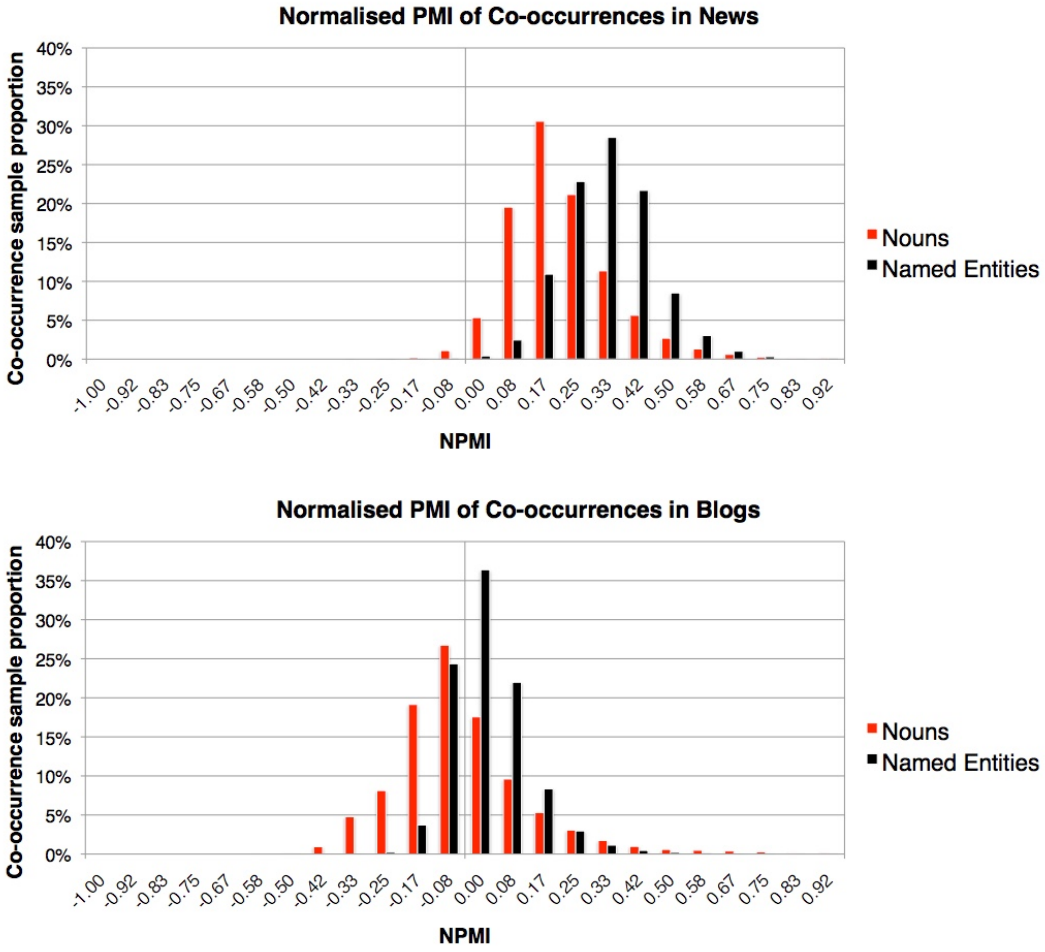


FIGURE 6.13: Distributions in NPMI for Noun and Named Entity Co-occurrences in News and Blogs

The data was prepared by extracting nouns, named entities, and multiword expressions from each blog post and news article in the corpus. These were then capitalised to ensure feature frequency calculations were case insensitive.

An initial study into the type of language in trends originating in social media has shown that although much that is discussed by bloggers is whatever is currently topical in mainstream media, there is a significant amount of material of wider interest that originates in blogs. Furthermore it suggested that a significant proportion of this material may be linked to that which is later topical in news articles. The amount of material produced by bloggers was found to be approximately 20 times greater in number of articles than professional news organisations, and the amount of individual nouns and named entities suggests their postings are also longer. Size of vocabulary is also much greater amongst social bloggers than within the mainstream media (although some of this one would consider to be erroneous or "noisy" text). This established that there is potential, then, for finding material in social media that is of wider interest.

A Poisson based model for tracking feature frequency was used as the basis for trend detection. The features analysed were nouns, four classes of named entities (Persons, Locations, Organisations, and other Miscellaneous) and three types of multiword expressions (compound nous, verb particles and light verb constructs). Trend analysis was carried out for each stream with trending features originating in the news stream being filtered out.

Although maximum trend strengths shown for nouns were found to be considerably greater than those shown in named entities, named entities were marginally more frequent in social media originated trends. Higher trend strengths were observed for those features that were seen, and particularly later trended, in news articles, although this was in strengths measured relative to the distribution seen for the feature type. Noun trends that also trended later in news articles were not separable by trend strength alone.

Compound nouns and multiword expressions in general, were found to give rise to some trends but those originating in blogs were not found in subsequent news trends. Assuming compound nouns generally make more specific references than singular nouns, this seems to be at odds with the finding for the latter, suggesting that unrelated nominal references may be giving rise to false or noisy trends.

Selecting trends purely by highest trend strength was found to be unlikely to be optimal, because many trending entity mentions of potential interest may be missed. A better strategy for selecting trends, likely to be indicative of topics of wider interest, would be to select the strongest trends within classes of nouns and named entities, possibly applying appropriate thresholds. Normalisation of trend strength by average class type trend strength may be another possibility, as this seems to make trend scores for feature types more comparable. Normalised trend scores showed

a narrower distribution around the mean score for entities that subsequently trend in news stories than nouns. This suggested that a threshold could be effectively applied in deciding what should be considered a genuinely trending feature.

An observation of the trending noun features is that many are either untagged entities, or non-English. This suggests that improved entity tagging and filtering by language could be beneficial. Other noun features typically display a high degree of variance with some periodic behaviour suggesting correlation with blogging behaviour rather than specific content. Identification of periodic behaviour could be used to filter these features out. These issues were not observed for compound nouns and may therefore explain why no subsequent trends for these in news were found following blog trends, but were for some singular nouns. From these observation it would seem that trending entity mentions are predictively stronger than simple lexical features.

The results suggest it is more likely that a trending feature in social media, also trending in subsequent news articles, is a named entity than a noun. (Even though the very strongest trend strengths in this analysis were displayed by nouns.) The identification and analysis of named entities as separate features to detect trends in is, therefore, potentially of great benefit when seeking to find emerging topics of interest.

The Named Entity Recogniser was not trained or tuned for social media, but rather well prepared newswire text. One would expect errors to occur both in recognition of named entities and in mis-typing of detected entities, and some errors were observed. However, a sufficiently high accuracy for differences in trends to be detected was observed. The extent to which named entity detection and recognition performance may impact results remains to be determined.

Given that there are a significant number of trends originating from social media, the natural next step was to consider whether one can predict which will go on to be subjects in the news, and what the delay between social media interest and mainstream media interest is.

By filtering out trends originating in news, the study went on to focus on the small percentage ($\tilde{3}$%) of topics that previous studies, Lloyd, Kaulgud, and Skiena, (2006), Leskovec, Backstrom, and Kleinberg, (2009), have found *not* to have originated in mainstream news articles. It has also examined whether named entities could be more useful than common nouns as a feature for predictive trend analysis.

Thresholds were applied to the Poisson trend model and a rolling seven day filtering window applied to potential feature selection. Analysis of the results showed that named entities, and nouns in general, can be found trending in blog postings that have not previously trended in the mainstream media. On average approximately 12% of these features (18.4% of tagged nouns, 11.1% of tagged entities) subsequently trended in news stories (although the total number of trending entities

exceeded that of nouns). Inspection of the top 50 trends has shown that about 30% of the corresponding trends arise from topically connected postings. This suggests that uncorrelated social media trends have some predictive power of mainstream news interest. Note that blog trends that did not subsequently trend in news stories may have warranted interest but were either missed by the mainstream media or not covered in the period of data capture.

Time lag between trending nouns and a subsequent news trend, for those in the top 50 that show this behaviour, is 9 days while for named entities it is 6 days. While being far from conclusive this suggests that if short time differences are indicative of a greater likelihood of topical connection, named entities have better predictive power than common nouns. Overall, only a small proportion of trends originated in social media are reflected in subsequently related news stories.

Assuming 3% of news stories start in social media this suggests that either the majority of significant social media originated trends are not sufficiently interesting to professional news organisations, or are missed. However, pre-emptive trending features do exist, and simple ranking of trending features by the number of occurrences on the day of trending gives reasonable performance in promoting those that subsequently trend in news. One can conclude, then, that there is potential data in blogs with which to investigate methods for prediction of news stories.

Trending mentions of entities may not be the subject of the topic(s) of interest, and subsequent stories may be indirectly related, on developing events, or take a new angle (a "meta-story" if you will). Examples of all these have been seen in the data examined here. However, not all subsequent trending news stories are related in any significant way to the preceding blog topics sharing the mention, and trends may arise from from multiple topics having the feature in common. This seems to be more likely the case with nouns than with named entities. However, it may be prudent to apply some threshold on the number of contributing postings and further characterise topics beyond identification of a trend.

Although there is little evidence that any trending topics originating in social media will be newsworthy, this does not entail that topics are not informative. Preliminary investigation into refining topics from trending features has shown that blog posts that are more likely to be related can be identified through co-occurring trending features, and that such bi-grams are more likely to contain entities than nouns. Selection of such bi-grams can be achieved through a rank ordering of how unexpected the co-occurrence of features within the bi-gram is. Given the finding that bursts, or rising trends, of mentions are correlated with potentially new information, and under the assumption that topic coherence may be measured using PMI and employed to cluster related documents together, evidence was found to support

*H*3: Documents imparting new information are more likely to contain unusual combinations of named entity mentions than unusual combinations of nouns.

Further exploration of selected feature co-occurrences under the assumption that normalised point-wise mutual information (NMPI) can be used as a measure of topic coherence was carried out. Comparison of co-occurrences by feature type revealed that named entity co-mentions contribute more to topic coherence under the NMPI metric than noun co-occurrence. The difference was found to be strongest in news stories. These one would expect to be more topically focussed, and indeed NMPI distributions for both nouns and named entities were found to have higher NPMI values in news than in blogs. These findings provided evidence in support of

*H*4: Mentions of different named entities are less likely to co-occur independently than mentions of different common nouns .

Overall it has been shown that trends in references, particularly named entities, could be used as features in information discovery with co-occurrences thereof being useful in refining topically related information.

The evidence that named entity mentions, and combinations thereof, are more likely to be indicative of coherent topics than common nouns is one of the novel contributions of this work. Another contribution made here is a method to filter social media selected for bursting features by removing those features that have similarly burst in recent mainstream news stories. Using this method evidence that newsworthy material may be found in social media prior to being published by mainstream news organisations.

The techniques explored do not in themselves present a solution for discovery of new interesting information from a text stream such as that found in Social Media. Further work would be required to advance approaches towards this goal. It would be possible to examine whether trend selection or ranking could be improved from detection of characteristic feature behaviour over time. It could also be possible to look further at trend co-occurrence as it seems like this may be a basis for refinement of features into more informative topics. Clustering of posts with trending features may also be a suitable method. These refinement methods could also assist the observer in assessing the likely interest as a news story. Such topic specificity refinement may also allow sufficient characterisation of trending topics originating in social media to make predictions as to which are likely to be picked up by news media.

The rise in popularity of microblogging services such as Twitter has increased both the number of messages and the speed at which they are published. The Poisson burst detection method introduced in this chapter could be adapted to meet this challenge by reducing the the window size, or by sampling more often with overlapping windows. A combination of these two adaptions could also be employed. News agencies, and their journalists, typically have accounts on microblogging services; separation of their accounts from public accounts could be used as a basis for creating the news and social streams for comparison.

However, a great deal of the text processed from social media does not give rise to new information. Many features do not give rise to indicative trends, and co-occurrence statistics are costly to compute. This raises the question of whether it is possible to filter text for that which at least is intended to explicitly provide information, thereby reducing noise in results and unnecessary computation. This idea is explored in the next chapter.

Chapter 7

Winnowing Twitter

7.1 Introduction

There are many reasons that people may have for writing online text. Imparting some information about something is just one. A system to discover new interesting information from Social Media would ideally process only that text which made assertions. Filtering out text that was intended to perform other functions could reduce noise, and would reduce unnecessary processing. This chapter presents an approach to identifying short pieces of text that make explicitly informative assertions which could be used as a basis for such a filter. This approach makes use of machine learning techniques to create models of unwanted and wanted types of text, using features associated with short messages from a Social Media platform. Because the function here is a general filter, the features focussed on are based on the degree of presence of types of selected word tokens and sequences thereof, such as number of nouns, locations, user identifiers etc., rather than specific references (as proposed for information discovery in Chapter 6).

The study presented in this chapter explored the potential of the approach in sorting microblog messages sampled from a corpus of Twitter content into wanted and unwanted classes. Two classification schemes are investigated. Given the use of News as a proxy for known information of wide interest, the first of these is News and Non-News. The second scheme seeks to more directly identify explicitly informative statements within a small set of Dialogue Acts. The acts include Opinion (or comment), Question, and Advertisement, as well as the desired Informative. These acts are described along with a report on an annotation exercise to create suitable training and testing data. A third scheme intended to distinguish messages about the author, those about the social contacts of the author, and those about anyone or anything else – a message's "subjective focus" – was also developed. However, this was not further investigated owing to insufficient annotated examples.

The envisaged application of a filter for explicit statements is that of detecting messages that may yield new information within streams of messages that contain various dialogue acts with multiple intended audiences. As described in Chapter 4, it was decided that a supervised machine learning approach would be appropriate

for constructing classification models that could differentiate desired text from that which did not explicitly provide information.

The chapter reports on the experiments carried out for the study. These experiments compared the performance of the selected features in models compared with that of traditional "bag-of-words", and that of using a combination, in allocating Tweets by the classification schemes. An experiment to compare the utility of the two classification schemes is also reported with a qualitative examination of errors made. Overall, the study aimed to examine evidence for the second hypothesis presented in Chapter 1:

$H2$: Sentences asserting information may be distinguished from those intended to perform other functions using non-lexical features.

$H2_{null}$: Non-lexical characteristics of a message carry no information on whether or not that message conveys an explicit statement.

In supporting this study, the following hypothesis was also investigated:

$H5$: There are more direct references to concepts and named entities in explicitly informative statements than in utterances performing other dialogue acts.

$H5_{null}$: Multiple direct mentions of concepts and named entities are no more common in explicit informative statements than in messages conveying other dialogue acts.

The rest of the chapter is organised as follows: Section 7.2 describes the data annotation schemes used to characterise and differentiate potentially desirable messages from unwanted material; Section 7.3 describes the data used in the classification and filtering experiments; this includes an analysis of the application of the designed annotation schemes on a sample of the sourced data in the creation of a suitable sub-corpus for the experiments. Section 7.4 describes the approach used in the experiments, covering the features used in the experiments, their extraction, and the machine learning methods used to create models. Three sets of experiments were carried out in the investigation: classification of News bearing messages is explored in Section 7.5; Dialogue Act classification in Section 7.6; and comparative experiments using the two approaches are described in Section 7.7. Section 7.8 notes recent related work before Section 7.9 concludes the chapter, providing a summary and a discussion of the results.

7.2 Classification tasks for short messages

The ultimate goal of the envisaged system is to find new interesting information from Social Media text intended to impart that information. The purpose of the filtering stage in the envisaged system is to select only those utterances that are intended to convey information about the world to the recipients. However, identification of

text doing so is a non-trivial task because there are many ways in which information can be conveyed.

In experiments exploring a potential approach to discovering new information, described in Chapter 6, mainstream news was used as a proxy for widely known information. Could new messages be identified as imparting news or not? The task is to separate News from Non-News, and corresponds to that of filtering News from Social media, one which has been increasingly popular with researchers. For example, see Bandari, Asur, and Huberman, (2012), Petrovic et al., (2013), and Madhawa and Atukorale, (2015).

Identifying explicitly informative text, i.e. text from which the expressed information may be understood in isolation, could also be a useful step. The intended purpose for an utterance, often a single sentence or phrase, may be one of many Dialogue Acts, as described in 2.2. For the filtering purposes here, a short message performing a single Dialogue Act is taken to be a reasonable approximation of an utterance.

Social Media microblogging platforms, such as Twitter, facilitate social communication by means of short messages (limited to 140 characters in the case of Twitter at the time of the study). People may communicate (or 'Tweet' in the case of Twitter) for many purposes, and correspondingly their messages perform many different Dialogue Acts. For example, they may wish to ask advice or questions of their followers and contacts, they may wish to express their opinions, or they may simply wish to announce to the world what they are doing or thinking at that moment! Sharing of information amongst groups of contacts is another common purpose for users of Social Media. It is an expression of information that is potentially of interest in a discovery system such as the one envisaged in this thesis.

Twitter is also used by some to disseminate News. Twitter messages therefore provide an opportunity to test both the News / Non-News and the Dialogue Act classification approaches. Efficacy of these approaches for the filtering stage in the overall envisaged system was explored in a set of tasks.

The first task explored the question of whether or not Tweets could be reliably identified as containing News or Non-News content. Experiments in this News / Non-News classification task, are described in Section 7.5. If messages containing News could be reliably identified, would such a classification suffice as a proxy for identifying explicitly informative text in the envisaged system? An answer to this second question requires a comparison between a message's news bearing quality with its intended purpose, or Dialogue Act This requires a Dialogue Act classification.

The second task is to identify a message's Dialogue Act. This may be a better indicator of its potential value than whether it is newsworthy or not, because, as found

in Chapter 6, potentially interesting emerging information might not be widely considered as news. The task was explored using five simple Dialogue Acts for Twitter messages, including one to capture explicitly informative text, as described in 4.2. Experiments exploring classification of Tweets by Dialogue Act are described in Section 7.6.

It was intended to explore a third task, designed to further characterise potentially interesting messages. This task is to to distinguish a message's *subjective focus*, i.e. what the message is about, as one about the author, one about the social contacts of the author, or one about anyone or anything else. However, the task was not fully investigated owing to insufficient annotated examples.

The two classification approaches, outlined above, were compared in experiments described in Section 7.7. These experiments highlighted potential for different interpretations of whether or not information on particular topics constituted News. An annotation scheme for three potentially contentious topics was defined for exploration of this. All of the classification annotation schemes for the tasks described above are described below.

7.2.1 News and Non-News Tweets

Information such as one might expect to see in mainstream news articles has been assumed to be of wide interest for the purposes of this thesis. While it could be argued that new information regarding any event, such as a birth one's family for example, might be considered news, it is not likely that such information would be of wide interest. Defining a boundary that distinguishes what makes something of wide interest is therefore difficult if not impossible. Lack of a formal definition of News by which it could be recognised might seem to preclude classification. However, people are familiar with the topics discussed in mainstream media and have an intuitive sense of what constitutes News. An informal definition of News is therefore used here.

- **News:** A message that conveys information on a topic likely to be found in mainstream news reports.
- **Non-News:** Any messages that do not convey on a topic likely to be found in mainstream news reports.

Examples of News messages may include "The prime minister spoke at the party conference today", "Chancellor: Everyone will have a tax increase next year", and "Manchester United have signed Rooney for 2 more years". However, messages such as "I sang at the party today", " everyone should pay more tax!" and "Rooney is no good at Manchester", are not News. Intuitively one can recognise, in the absence of any contrary contextual information, the difference.

7.2.2 Dialogue Acts in Tweets

Many fine grained Dialogue Acts have been proposed. However, a comprehensive classification scheme is not required for the filter here because many of these acts are not of interest. It was decided to use a simple five act scheme along similar lines to Zhang, Gao, and Li, (2012), as discussed in 4.2.

The Dialogue Acts chosen for the classification scheme used here are:

- **Informative Statement:** A message that, in isolation, unambiguously asserts factual, non opinionated, information about something.
- **Opinion or Comment:** A message that expresses an opinion on, is passing comment on, something.
- **Advertising:** A message that is clearly intended to advertise a product or service. ('Spam' etc.)
- **Question:** A message that is posing a question or request to the reader(s).
- **Other:** Any message not falling into the categories above.

An example of an Informative Statement would be "The prime minister met with the Chinese president for trade talks", whereas "The prime minister should not be talking with the Chinese president" would be classed as an Opinion. One may observe that both these examples explicitly assert information. As discussed in 4.2.1 separation of opinion from fact may be difficult to assess automatically. The choice to distinguish expression of opinion in the dialogue act scheme was made to enable some investigation into any detectable differences expressions of opinion and factual information. Some expression of opinion may be in the form of a comment. An example of this would be "This story is bunkum". A comment may also be a simple description, or caption for media posted alongside, such as "Players celebrating win in 1985 cup final".

Spam, unsolicited messages advertising goods and services, are prevalent in social media. The advertising often funds the running of the social media service. Detecting and filtering out of content such as "Earn up to $1000 per hour in your spare time. Click here to find out how!" is generally desirable to remove noise. Advertising was therefore included in the scheme.

Questions are those messages seeking information or action. Examples such as "Can anyone recommend a good vet in the Springfield area please?", or "What was Cameron thinking?!", one would hope to be often distinct and readily identifiable. Questions were therefore included as a separate class in the Dialogue Act scheme.

Any message intended to perform a function other than those described above is not considered of interest. Such messages are simply classed as Other in this Dialogue Act scheme. However, one may observe other dimensions in social media messaging.

7.2.3 Characterising the subject of a Tweet

Social Media platforms are, by design, *social*, allowing authors to address a wide audience as well as individuals therein. Messages in Social Media, therefore, may be broadcasts or part of a dialogue. Authors could be focussing on themselves, other individuals or the world in general when publishing a message. To facilitate analysis into whether or not different focal subject areas could have a bearing on whether a message was interesting to a wider audience (in the sense described in 1.1), an additional annotation scheme for 'Subjective Focus' was defined:

- **World:** Messages that are about, or are referencing, things other than the author or specified contacts.
- **Contact:** Messages to, or referencing, specified contacts in the Social Media platform.
- **Self:** Messages that are about the author of the message.
- **Unknown:** Any message where the addressed subject is not clearly a specified user of the platform or externally identifiable.

World focussed messages are those one would typically think of when considering reports such as statements of News. "The PM is meeting with the President right now" would be categorised as World focussed because it is about entities other than the message's author or a specific user in social contact with the author.

One can consider a contact focussed message as one specific to, or about, a particular known user. Examples would include "@TwitterUser, you would appreciate this: http://link", and "@TwitterUser will not like that!". Both of these examples would also be categorised as Opinion/Comment. Note that public figures may (and often do) have accounts on Social Media. Their identifiers are widely known. There is a judgement to be made, then, as to whether or not the presence of user ID of a public figure indicates that the individual is a genuine contact of the message's author. In general one would expect is that he or she is not, and the message is therefore not Contact focussed.

Self focussed messages are those such as "I'm going to Paris!". Note this example is also informative, whereas "to me!" would be categorised as Self, but Not Informative.

The chosen Dialogue Acts, and the potential for Subjective Focus, were investigated in experiments described in Section 7.6.

7.2.4 Topics potentially associated with News

Applying the Dialogue Act annotation scheme to a sample of data, as described in 7.3.2, permitted comparison of News classification with Informative Dialogue Act classification. This comparison is described in 7.7. An analysis of the errors made

by classification models suggested that there might be some difference in whether or not certain topics had been considered News or not. To investigate this, an annotation scheme for those topics was devised, and used to re-analyse results of the experiments. The topics chosen for inspection were:

- **Sport:** Messages that are related to sporting events, including results, players, and teams in relation to the sport.
- **Weather:** Messages giving weather reports and forecasts.
- **Travel:** Messages reporting on travel conditions and incidents affecting public and private transport.

These categories should be fairly self-explanatory. Sport related messages include those such as "Man U. go 1-0 up against Man C after 20mins" and "Tokyo to host rugby world cup!". "5cm Snow expected expected overnight" and "Heavy rainfall in N.Virginia" are both Weather related messages. Examples of Travel related messages are "Traffic slow moving on I95 north from 495" and "Flights from Gatwick delayed 3hrs due to high winds".

7.3 Microblog corpora for experiments

A sample of Social Media text messages, suitable for annotation, was required for the experiments. Microblogging services such as Twitter provide a stream of short messages. The purpose the author has in publishing a message on such a platform may be one of many, including providing new information to the world at large (i.e. anyone reading the author's messages). The Twitter API has facilitated the creation of sample corpora. It was therefore decided to use Twitter data to learn models to classify short sections of text using the defined schemes.

The experiments described in this chapter made use of data kindly provided by Miles Osborne from the Redites project. The Redites project Osborne et al., (2014) created a corpus of 1.4 million Tweets by applying a high recall, low precision, "event" detector on 37.5 million Tweets from a Twitter feed. An event, here, is a cluster of related Tweets that has grown rapidly during a short time perdiod. The detector, described by Petrović, Osborne, and Lavrenko, (2010), works by using a locality sensitivity hashing based method (Charikar, 2002) to assign Tweets to a document space divided into 'buckets'. Each new Tweet is compared to Tweets that collide in the hash space and the nearest neighbour found. If the nearest neighbour is further than a threshold, and the 2,000 previous Tweets, as measured by the cosine between the two associated vectors, then the Tweet is declared a new topic. Otherwise, providing the cosine distance between the Tweet and its nearest neighbour is less than a user specified threshold, the Tweet is paired up with its nearest neighbour. This process results in threads of related Tweets. The fastest growing thread at any one time step is output as a burst event.

A sample of Tweets from burst events in this collection was then hand annotated, each event Tweet classified as either a News event or a Non-News event. This resulted in a data set of about 489,000 events. 2,286 Tweets were positively identified as yielding information connected to a news event, leaving approximately 1.03 million Tweets connected to either Non-News events or un-assessed events (outside of the sample period). Although the annotations were made available, no results on any approaches to automatically classifying these Tweets, or the events, as News or Non-News had been published at the time of writing. The selected Tweets covered a time period of 09:58 2/9/13 to 11:16 30/9/13.

Sub-corpora for the tasks described in Section 7.2 were created from this data, as described below.

7.3.1 News/Non-News Tweet sub-corpora

Sub-corpora were selected for the machine learning based experiments described in Section 7.5. A balanced sub-corpus was created by selecting 2,000 News-event Tweets and 2,000 Non-News-events, thereby removing the prior distribution on news and non-news which otherwise be a dominant factor in model optimisation. A further 572 Tweets with later timestamps, 286 of each class, were selected as a balanced evaluation set. Adding all non-news-event Tweets from the same time-period as the 572 resulted in a collection of 173,514 Tweets.

In summary the data selection process resulted in the following sub-corpora:

- **BAL-4000**: 4,000 Tweet balanced training and test sub-corpus
- **LBAL-572**: 572 Tweet balanced evaluation sub-corpus
- **EventStream**: 173,514 unbalanced, representative, evaluation sub-corpus

The corpus was processed to extract the features, particularly Named Entity mentions, using the tools described in Chapter 5. The total number of Named Entity mentions and Noun phrases extracted are summarised in Table 7.1.

	Full Corpus	All-Events	Non-News Events	News-Events
Persons	7344447	287783	286592	1191
Locations	2780899	120951	119191	1760
Organsations	2134051	85796	84931	865
Noun Phrases	70196373	2172404	2165038	7366

TABLE 7.1: Numbers of Named Entities and Noun Phrases extracted from Redites corpus

Ideally one would hope for some feature(s) to be significantly more present in one class of messages than in others. This would make identification of that class easy based on measurement of the feature(s). In challenging classification tasks classes typically have overlapping feature distributions and potentially sparsely occurring features. As has already been described and shown in previous chapters, combinations of features may correlate with desired classes, and machine learning may be

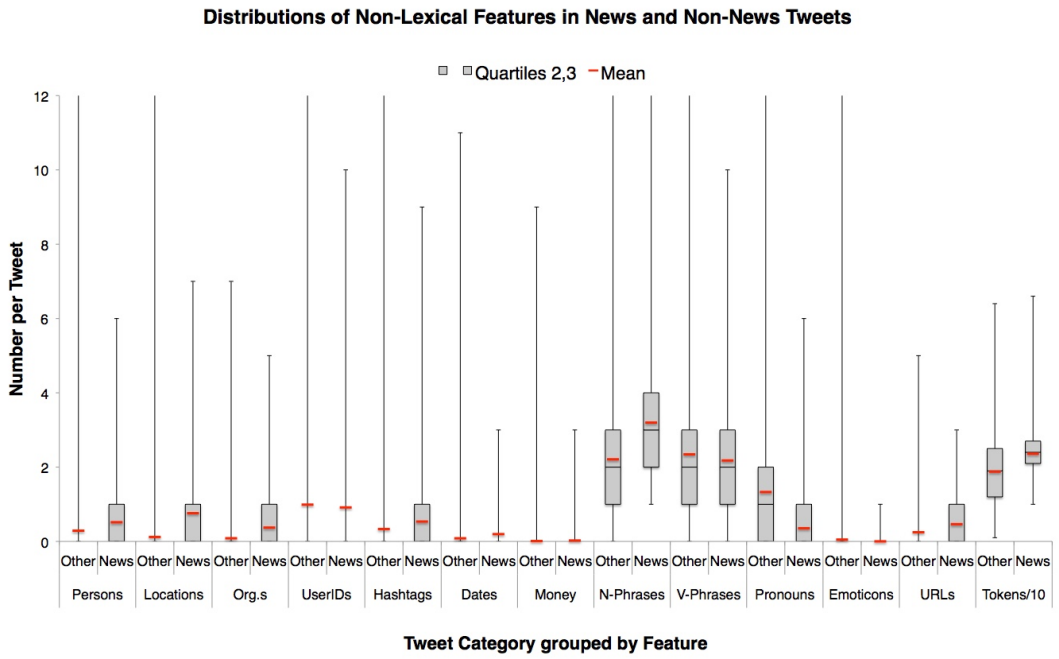


FIGURE 7.1: Non-lexical feature distributions in Redites News/Non-News Event corpus

employed to model these combinations. An indication of how separable the Redites corpus of Tweets could be, using the features described in section 7.4.1, may be gleaned from their distributions over the Tweets annotated as examples of the News and Non-News classes. Comparative box-plots of the non-lexical feature distributions are shown in Figure 7.1.

There are few differences in feature distributions, with some, such as sums of money, rarely present in either class. News items tend to be longer than Non-News items, and mentions of Named Entities, particularly Locations and Organisations, are more frequently found in News Tweets. Personal pronouns, perhaps unsurprisingly, are less frequently found in News Tweets.

7.3.2 Annotating Tweets for Dialogue Acts and Subjective Focus

While it is plausible that Tweets imparting news make explicitly informative statements, it does not follow that an explicitly informative statement is news. Although Tweets containing news may often also be explicitly informative messages, identification of News stories will only capture a limited set of such. The goal in the overall system envisaged in this thesis is to discover new interesting explicit information. News is taken to be that which is already known (and therefore not interesting). Rather than modelling Tweets containing news to represent informative utterances, would it be possible to model explicitly informative utterances more directly? Without the context of a dialogue, how easy is it for people to identify Tweets that make explicit statements, those that reference other Tweets, and those that are examples of other dialogue acts? Are people able to identify the focus of a message without

the context of that message? If it is assumed that statements about the world are potentially newsworthy, then if people can reliably separate Tweets containing such utterances from others it may be reasonable to get a machine to do so. This section explores these notions, giving an analysis of human annotation of Tweets.

To provide data with which to analyse both human and systematic performance at identifying explicit statements about the world, and other dialogue acts with foci, a new corpus was created by sampling previously unclassified Tweets from the Redites corpus. News-event classifiers were built by training models using the same four classifier technologies with the unigram and non-lexical feature sets described in 7.4.1 on all of the 4,572 annotated Tweets described in 7.3. These models were run over the unannotated data. Sampling of the unannotated data was carried out such that a uniform distribution of likeliness of messages to be news-event related Tweets as measured by classifier agreement would be obtained. This ensued a reasonable degree of confidence a coverage of both news-event related and non-news event related Tweets in the corpus.

The Tweets resulting from the sampling were given to assessors via a web-based mark-up tool. The assessors were volunteers from the author's research group. They were asked to tag Tweets with two class labels, one for the pragmatic intent of the message, and one for the focal subject of the message, here referred to as 'Subjective Focus' ('Focus' for short), if determinable as a party to the act of the message, i.e. whether the message was about its author, another Twitter user, or about the world, or anything in it, in general.

- **Message Intent:** Informative Statement, Opinion or Comment, Question, Advertising, Other/Non-informative
- **Focus:** World, Contact, Self, Unknown

Annotation guidelines were kept simple, presenting in the mark-up tool the descriptions of the classes, as given in Section 7.2.2 and Section 7.2.3, with examples of each class. Users were told that the task was assess whether each message fell into one of the message intent categories or gave no information to base a decision on. They were also told to assess who or what the message was focused on "relative to it's author" in making a decision on whether a message fell into one of the focus categories or had no discernible focus. A number of simple examples were made up to illustrate how to apply the categorisations. These were presented in the mark-up tool interface along with each message to be tagged. Illustrative examples included:

The PM has just left to visit China	Informative + World
@ANON Are you going to the pub tonight?	Question + Contact
We are the best bloggers ever!	Opinion/Comment + Self
Our CEO has announced the company intends to buy Apple	Informative + Self

Each Tweet was presented to five assessors. Once it had received five tags for each class label it was added to the new sub-corpus and not presented to another assessor.

A little over 1,700 Tweets were classified by five assessors. Inter-annotator agreement where at least four annotators agreed was 68% and 70% for Dialogue Act and focus respectively. It is possible for assessors to agree by chance. Fleiss's kappa does calculate the level of agreement over that which which would be expected by chance, although there is no natural interpretation for significance. A score $\kappa = 1$ is perfect agreement. Overall, Fleiss's kappa analysis for Dialogue Act annotation gave $\kappa = 0.67$ and for focus gave $\kappa = 0.64$. Details are given in Table 7.2.

	Informative	**Question**	**Comment**	**Ad.**	**Other**	**All**	**IAA**	**Fleiss' κ**
5-way	439	12	141	92	2	686	40%	
4-way	673	24	298	166	7	1368	68%	0.67
	(+234)	(+12)	(+157)	(+74)	(+5)	(+482)		
3-way	783	35	396	226	17	1657	85%	
	(+110)	(+11)	(+98)	(+60)	(+10)	(+289)		
	World	**Contacts**	**Self**	**Unknown**				
5-way	615	25	0	2		642	37%	
4-way	1087	61	41	7		1196	70%	0.64
	(+472)	(+36)	(+41)	(+5)		(+554)		
3-way	1315	139	88	21		923	84%	
	(+228)	(+78)	(+47)	(+16)		(+369)		

TABLE 7.2: Inter-annotator agreement (IAA) for Dialogue Act and focus labels for Tweets. (3-way agreement assumed if alternative annotations differed from each other.)

Table 7.3 gives the number of Tweets for each label combination where at least 3 judges agreed (and 2 did not give a consistent alternative) for *both* labels. This shows that not only is there a very uneven distribution but also, observing the total number of Tweets is less than that obtained for either label independently, the ease by which a Tweet may be judged for a Dialogue Act is not correlated with the ease of judging its focus.

To create a sub-corpus suitable for further machine learning experiments, those Tweets that had had no clear decision on a class amongst annotators (by a margin of two or more) were discarded. This left a collection of 1285 Dialogue Act annotated Tweets, the DAAT-1285 sub-corpus. (1574 Tweets had at least 3 annotators in agreement). There were relatively few examples marked as questions or being other/non-informative, and the majority were considered to have a 'world' or indeterminate focus. There were few clear examples of messages referring to their author or aimed at specific contacts. Creating and evaluating a model through machine learning specifically for the identification of messages with an interlocutor

Focus \Act	**Informative**	**Question**	**Comment**	**Advert**	**Other**
World	605	9	200	89	0
Contacts	14	12	26	11	0
Self	5	0	18	4	0
Unknown	0	0	3	0	6

TABLE 7.3: Combinations of Dialogue Act and focus where at least 3 annotators agreed for both labels

focus would therefore be impractical without the addition of further annotated examples. However as this sub-corpus is marked for informative statements and other speech-acts rather than news and non-news events, it would be possible to compare models built by machine learning on this data with those built on the previously described news-event annotated sub-corpus. A comparison should give some insight into any interdependence between the classes of "news-events" and "explicitly informative statements".

7.4 Classification methodology in experiments

As described in Chapter 4, supervised machine learning was chosen as an appropriate method for developing filtering models. This requires that the classes, into which data should be sorted, be defined, with class-annotated examples from which to learn a model.

The approach reported here uses selected machine learning techniques to build filter classification models. Binary classification models are learnt from example data sampled from Twitter, hand annotated by one of the schemes described in Section 7.2. Data for each example are represented as a vector of feature values corresponding to the message content. These feature values may correspond to the words used in the content, a set of non-lexical aspects of the message (described in Section 7.4.1), or a combination thereof. Feature values are determined using feature extractions tools discussed in Chapter 5. Feature vectors determined from previously unseen messages are compared to the corresponding models to determine the most likely class label. This approach is illustrated in Figure 7.2.

Evaluation of the approach was carried out through a set of experiments, described in Sections 7.5 – 7.7. Various machine learning techniques and feature set combinations were evaluated in the experiments. Data was split into fixed separate training and test sets for each experiment. Evaluation of the models was carried out using annotated examples that were not used in model construction, comparing the predictions of the models with the annotation.

7.4.1 Features

The filtering goal here is to identify particular *types* of text rather than, say, what the text is about. As described in previously in Section 5.1, features were chosen that hopefully would characterise and discriminate the classes defined in Section 7.2 more strongly than just lexical tokens. These features are those that are characteristic of a message's content beyond the lexical level, and break down into four categories:

- **GRP:** Graphically identified features - Emoticons, Hashtags, URLs, and Username Identifiers
- **NE:** Named Entities - Persons, Locations, Organisations, Dates, Sums of Money

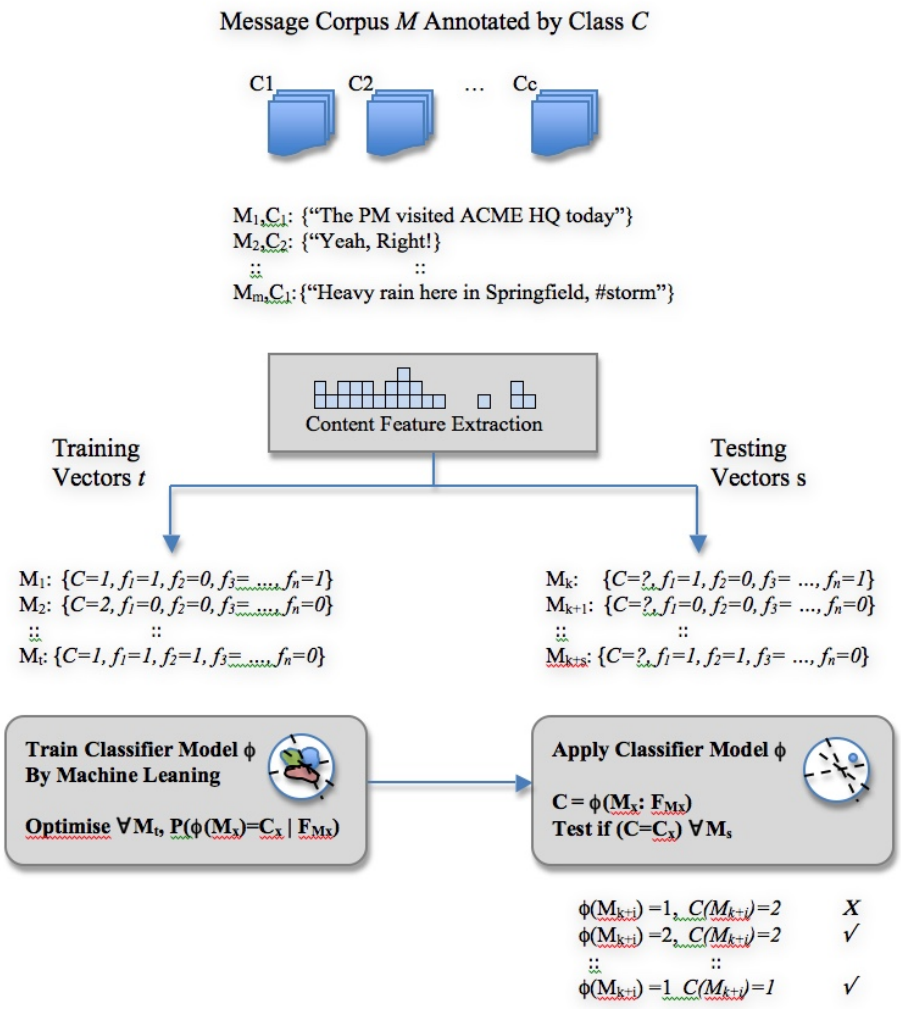


FIGURE 7.2: Creating a classifier model ϕ, optimising over feature vectors in set t, the probability that model predicts the message class given its associated feature vector. Testing the model's prediction of class label C for unseen set s messages operating on their feature vectors. Messages in corpus M are allocated to either training set t or test set s.

- **SYN:** Syntactically determined features - noun-phrases, verb-phrases and pronouns

- **FRQ:** Frequency based features - total word IDF, total phrase IDF, average word IDF and average phrase IDF

The graphically determined features are those token types that are particular to the medium, in this case microblogs. Their use was motivated by the expectation that statements and non-statements would either be more or less likely to contain such features. For example, one might expect a Tweet expressing an author's opinion to be more likely to contain an emoticon than a factual statement about something in the world. An explicitly informative message will state information about something, and this will often be referred to by name. This motivates the inclusion

of Named Entity count features. However, sometimes things are conceptual or referenced pronominally, and the syntactically determined features are designed to capture some of these references. The expectation was that personal pronouns would be more prevalent in conversational messages that in news items. Lastly, inclusion of the frequency based features were motivated by the idea that they might indicate how informative a message is through estimation of word and phrase weight in conveying information.

Traditionally machine learning is applied to the tokenised content of the texts, possibly with the addition of the associated meta-data. As the interest here is in the content of a message alone, a baseline feature set of lexical tokens (unigram) was also employed.

7.4.2 Feature extraction

The graphically indicated features were readily detected by simple pattern matching. A gazetteer look-up is sufficient to identify an emoticon. Hashtags and user identifiers are identified by leading '#' and '@' symbols, and URLs are identified by leading hypertext protocol designator.

Numbers of GRP and NE features in messages were calculated by running the TwitIE extractor Bontcheva et al., (2013) and counting the number of instances of each type of feature. The extractor was also used to detect the presence of pronouns. The other syntactically determined features, counts of noun and verb phrases, were determined using the simple part-of-speech sequence parser, developed for multiword expression extraction, described in Chapter 5. This was used to capture all nominal references, whether compound or not. Similarly verb phrases included all detected verb expressions including both verb-particle expressions and light verb constructs.

The FRQ features include the total and average inverse document frequency (IDF) for words occurring, and extracted phrases, in the message. These were motivated by the idea that the rarer the content of a message, the greater the likelihood that the message is topically specific. IDF for each word (or detected phrase) was calculated across all of the Twitter messages available (see 7.3), approximately 37.5 million Tweets, treating each Tweet as a document.

For unigram model features, tokenisation was carried out by segmentation on white-space and punctuation other than '#' and '@' which have specific meaning in microblog messages.

7.4.3 Machine learning tools employed

The experiments compared four popular machine learning techniques. Naive Bayes and Maximum Entropy classifiers build models based on the the frequency of the features in the training set. Decision trees (Quinlan's C4.5), seek to create segment

classes by separating by each feature in turn. Support Vector Machines optimise a decision boundary margin between class features in a vector space. The LibSVM (Chang and Lin, 2011) implementation is used to create the SVM models while the other classifiers are provided by the Mallet toolkit (McCallum, 2002).

7.5 Detecting news-event Tweets

The task explored in the experiments described in this section is to separate News from Non-News given Tweets selected and clustered into related groups referred to as as 'events'. It is assumed that an initial coarse selection filter, such as proposed by Cui et al., (2012), or Osborne et al., (2014) is used to select event Tweets. The initial experiment compared the classifiers with each of the feature sets using the class-balanced BAL-4000 sub-corpus of 2,000 News-event Tweets and 2,000 Non-News event Tweets. Feature vectors for each feature set under investigation were created for each Tweet. These features were also combined into composite feature vectors.

Tweet vectors were grouped by event and split into ten folds such that any one event was represented in just one fold, minimising potential skew in testing from the small number of duplicates in the form of ReTweets. This was intended to give some control for any event-specific vocabulary when training and testing models using cross-validation methodology. Standard 10-fold cross validation was used to obtain average performance figures for each of the features sets, and combinations thereof, using each of the example classifier technologies. For a given combination of features, each of the ten folds of the feature vectors corresponding to the balanced sub-corpus were held-out in turn and the corresponding nine folds used to train classification models using Maximum Entropy, Naive Bayes, C4.5, and Linear Support Vector Machine methods. The resulting models were then tested on the held out folds giving ten sets of results for each classifier technology. These ten-fold results were then combined to give mean performance figures. Default hyper-parameters were used for classifiers as optimisation of the learning was not the object of the experiment.

Given the balanced sub-corpus, the evaluation metric selected for this experiment was simple accuracy, i.e. percentage of correct answers, and averaged results from across the folds. Also measured was the variance in results to give an indication of how reliable the observed accuracy figures were. Results are shown in Table 7.4. An error margin of 1 standard deviation is shown in parenthesis.

One may observe a strong baseline performance using only unigrams, albeit with a high degree of variance. Only C4.5 unigram model did not achieve an average accuracy over 76%. One reason for the high variance across folds could be a lack of vocabulary coverage to form reliable models for explicit news statements. Another reason could be correlation between particular topics and news dissemination. (One

	Max. Entropy	Naive Bayes	C4.5	Linear SVM
Unigrams	77.6% (26.0%)	76.9% (34.0%)	70.1% (31.0%)	76.9% (23.8%)
GRP	68.3% (1.7%)	68.4% (0.9%)	64.8% (6.3%)	72.6% (6.0%)
NE	70.3% (4.2%)	61.8% (4.2%)	61.8% (13.3%)	77.1% (4.0%)
SYN	50.0% (0.2%)	50.0% (0.0%)	60.2% (6.0%)	72.8% (2.4%)
FRQ	54.2% (1.2%)	61.1% (2.1%)	63.2% (14.6%)	78.6% (2.1%)
GRP+NE+SYN+FRQ	82.7% (3.5%)	73.0% (3.0%)	82.4% (3.8%)	**85.0% (2.3%)**
All-Features	86.9% (3.8%)	**92.1% (4.7%)**	85.2% (7.6%)	90.3% (8.9%)

TABLE 7.4: Accuracy of News/Non-News models in BAL-4000 sub-corpus 10-fold classification using feature set combinations

might expect an overlap given news tends to be about particular topics.) Folds were controlled for events, although independence of events could not be assured.

Individual non-lexical feature sets mostly produced lower than baseline performance but with significantly less variance across the folds. The one exception was an SVM based model using the FRQ feature set, which yielded an average accuracy of 78.6%, an increase of 1.0% over the best unigram models, achieved with Maximum Entropy. The features derived syntactically performed the worst although they did provide some classification power when used with C4.5 or linear SVM. This could be a result of the relative sparsity in the features given to the difficulty of phrase detection in informal texts. The combination of all the non-lexical features gave rise to performance above baseline for three of the classifier technologies, but Naive Bayes (which assumes feature independence) failed to achieve any improvement.

Lexical and non-lexical features capture different contributory information. This can be seen in the performance achieved by using the combination of all the features, including the Naive Bayes classifier. An average 18% relative improvement in accuracy over the baseline is achieved using all features, although with a higher variance shown in results than those for non-lexical feature models. Naive Bayes and SVM models yielded the highest overall accuracies of 92.1 and 90.3% respectively.

The results of this experiment showed some promise that non-lexical, vocabulary independent, features could significantly help in identifying news-bearing Tweets. We next sought to further examine the contribution of our features in the classification task and model performance given new later occurring data.

7.5.1 Feature set contribution

The next set of experiments looked to give an indication of the portability of models and the contribution of each of the non-lexical feature types in classification models using the "hold-out" methodology. All the feature set values were combined into a vector for each Tweet. Each classifier is trained using the vectors for the training data and evaluated on the vectors calculated for the evaluation sub-corpus. Then one set of features was removed to give a set of vectors without that particular set of features, and the train and evaluate procedure carried out again. This process was repeated for each set of features. The contribution of each set of features can then be assessed by any decline in performance resulting from their absence from the model.

The models were built with each type of classifier using each set of feature vectors corresponding to the BAL-4000 sub-corpus. Each model was evaluated on the LBAL-572 Tweet balanced sub-corpus using feature set vectors corresponding to the model being evaluated. A unigram model was also built from the BAL-4000 training corpus and evaluated to give a baseline performance, and to give an indication of how temporally portable a unigram model might be. The results are shown in table 7.5 along with the average performance of each of the classifier types.

	Max. Entropy	**Naive Bayes**	**C4.5**	**Linear SVM**
Unigrams	52.8%	59.3%	59.8%	51.6%
NE+SYN+FRQ	72.7%	70.5%	68.4%	80.4%
GRP+SYN+FRQ	65.2%	68.0%	66.3%	82.4%
GRP+NE+FRQ	72.9%	69.8%	67.1%	83.0%
GRP+NE+SYN	56.3%	67.8%	68.9%	77.6%
GRP+NE+SYN+FRQ	71.3%	69.4%	67.3%	82.7%
Average perf.	65.2%	67.5%	66.3%	76.3%

TABLE 7.5: Feature contribution to classifier model accuracy on held out LBAL-572 sub-corpus data

The first notable observation is that the baseline performance is significantly lower than that observed in the initial closed set experiment. Average classifier accuracy with unigrams is approximately 20% less. The use of non-lexical features on this test set in comparison shows an 8% drop in average classifier accuracy. The difference in the two relative reductions in performance lends support to the idea that apparently good results from lexical models are in some measure due to learnt vocabulary covering current news topics.

How uncommon words in a Tweet are - the informativeness as measured by their frequency in messages - though, is correlated with news-event statements. This can been seen from the 5% drop in accuracy when omitted. The presence of Named Entity features are also a factor given an average 3% drop in accuracy when omitted. Orthographically and syntactically determined features do not seem to provide any significant additional information in this particular experiment.

The results of this experiment suggest that classification models for identifying statements of news in Tweets learnt using non-lexical features are more temporally stable than those learnt using unigrams. The aim of the next experiment was to evaluate example models on a representative set of data from a Twitter event detection stream.

7.5.2 Refining news-event Tweet detection

In the experiments above it was sought to control for the prior expectation of informative event Tweets by training and testing using the same number of positive examples as negative ones. In a stream of arriving Tweets one would expect to receive many more uninformative Tweets than informative ones. Classifier confidence could be used as a means to adjust precision in results allowing the user to control the volume of Tweets received.

The balanced evaluation sub-corpus Tweets were observed to occur over a specific period of time. All the Tweets marked as non-news event Tweets during this time were added to the balanced evaluation sub-corpus to produce the unbalanced evaluation EventStream sub-corpus, a resulting collection of 173,514 Tweets. The results of a classifier on the evaluation Tweets were rank-ordered by classifier confidence that the corresponding Tweet should be tagged as giving news-event content. At each level of classifier confidence the number of Tweets correctly classified and the number incorrectly classified were calculated. The ratio of these values to the total number of Tweets yielding classifier scores at or above the confidence thresh gives the expected true and false positive rates for that threshold. Taking the resulting values for each threshold allows a true/false positive receiver operator characteristic (ROC) to be constructed. The ROC curves for the classifiers using unigram and the non-lexical features under investigation are shown in Figure 7.3.

Absolute performance figures calculated at the Tweet level are naturally low. The 286 news-event Tweets in the held out data covered 187 news-events. By contrast non-news Tweets made up 84,171 non news-events. The ratio of correctly selected Tweets to incorrectly selected gives the precision in selection up to the given threshold. The area under the ROC curve gives an overall performance metric for the associated model. The true positive rate averaged over false positive rates (AUC) for the best performing models with unigram features and the non-lexical features, both using SVM, were 86.1% and 90.9% respectively. A maximum entropy model performs the best with a combination of all the features, despite having poor false positive rates for either feature set alone, with an AUC of 94.5%.

7.5.3 Classification error analysis

While unigram performance within a corpus appears to be very good, comparing the results on the later occurring evaluation sub-corpus to those found in the balanced training and test sub-corpus, one may observe that it drops when the classifiers are run on later data. This does not seem to be as much as a problem for classifiers using non-lexical features. As lexical features are strongly correlated to topic (the basis for most information retrieval methods), this suggests a higher precision than should be expected was achieved in the closed-set corpus through classification of informative topics. Tweets with previously unseen topics are less likely to be classed as informative because their vocabulary will not be that of the training topics. (This may be alleviated by larger training sets and regular retraining but would entail the additional associated cost.)

Furthermore, it may be that stated information is not necessarily considered newsworthy; Tweets that have been classified as news-events may, therefore, be a sub-set of those that are explicitly informative. It may be that models based on non-lexical

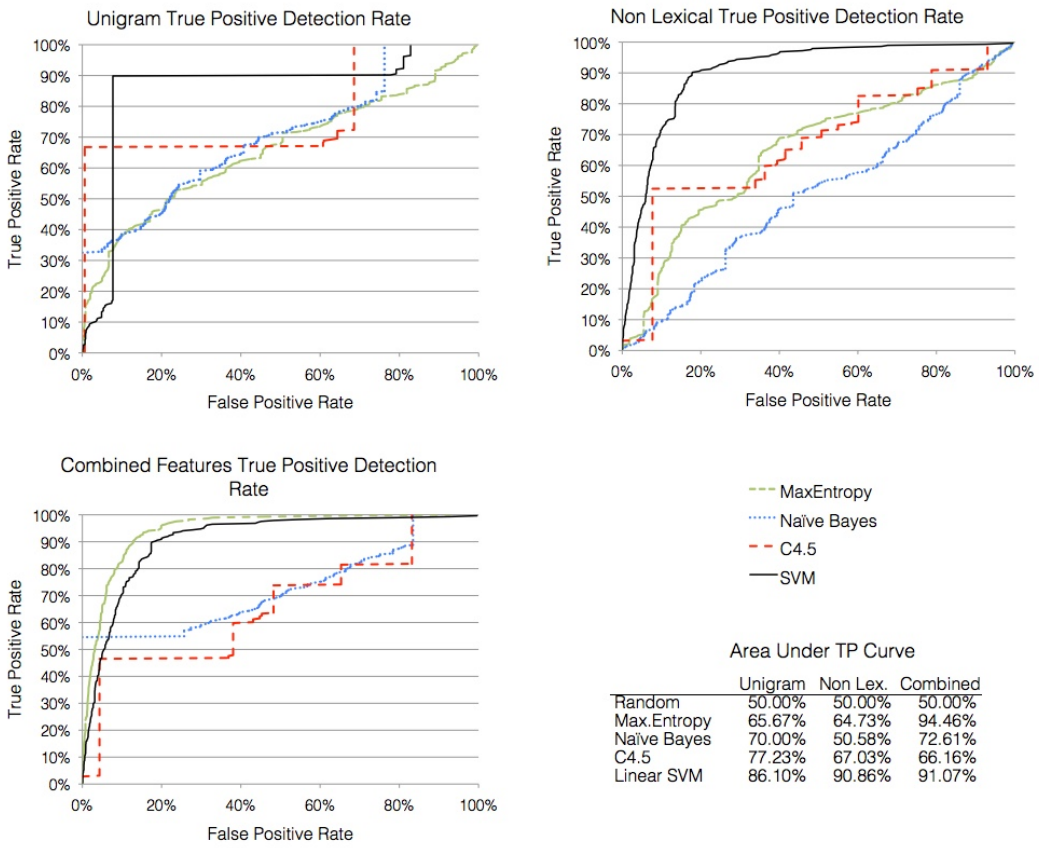


FIGURE 7.3: ROC curves for classifiers on unseen twitter data using unigram and non-lexical features

features, in not capturing topic, more closely correspond to this more general characteristic. The false positive classification results made most confidently by the SVM classifier trained on the non-lexical features are shown below:

- 76 Regime Troops killed by #Syria free army, death toll 4,565, #Bahrain #UAE #Turkey #Iran #Russia #USA #UN cont http://t co/Pt6MbzJ2PS
- RT @———: Obama, Hollande to face off with Putin over Syria http://t co/YDbLDEX7cx #Syria #USA #UK #Iraq #Lebanon #Oman #Qatar #KSA #Y
- Allies against #Syria: US, Australia, Canada, France, Italy, Japan, South Korea, Saudi Arabia, Spain, Turkey and the United Kingdom
- RT @———: PR pros are needed for peace building project #Japan, #Thailand, #India, #Germany, #France, #UK, #USA, #Canada, #Chili
- US urges #Syria to unveil chemical weapons stockpile http://t co/2ri0165iNo #Belgium #egypt #Iraq #news #ABC #Woman #BBC #world #AP #sydney

At least three of these, the first two and last of the Tweets in the list above, seem to be informative (and arguably newsworthy). The third and fourth Tweets above have a high number of locations or hashtags that have fooled the classifier. This observation motivates a closer analysis of intent behind Tweet content and any correlation

with news-event relevancy.

Inspection may also give possible explanations for false negative classifications, i.e. those Tweets that should have been selected as news-bearing but were not. The least confidently predicted news-bearing examples were:

- Papers: United to bid for Coentrao http://t co/trcWis4Mlb
- #Nestl tells us it won't comment on market rumors about the sale of Powerbar gt http://t co/qEx5bHf8QK
- Kumi Naidoo on Greenpeace Activists Detained in Russia http://t co/q0LFrjDRlP
- RT @Polygon: Valve announces Steam Controller http://t co/OwOtt3p2x2
- US braces for possible shutdown http://t co/Uxp4hSQThL

It is not clear on first inspection why these examples were harder for the models to spot. URLs are present but this alone seems unlikely to be the cause. Closer inspection of the associated feature vectors show that the Named Entity tagger did not recognise the entity mentions. Total IDF is also relatively low compared with those messages confidently predicted as news-bearing statements. This suggests that improved performance could be achieved with better Named Entity detection and token specificity estimates.

7.6 Dialogue Act classification

The main filtering goal is to identify text, here Tweets, that explicitly providing information. Could classification by the Dialogue Act scheme described in Section 7.2.2, using models based on message features achieve this? Of particular interest is whether or not information related to the pragmatic intent of a message is carried at a deeper level than its surface lexical form.

While a message's topic may be strongly indicated by information at the lexical level, how informative it is semantically, though, may be better captured by non-lexical features, i.e. features *about* the words rather than the words themselves. Different dialogue acts may use the same words but in different constructs. For example, "He has found the cat" and " Has he found the cat?" contain the same words but one is a statement whereas the other is a question. The first experiment described in this section sought to compare the use of lexical features, in the form of unigram tokens, with the non-lexical features described in Section 7.4.1 in classifying the dialogue acts annotated in the DAAT-1285 collection of Tweets described in Section 7.3.2.

The distributions of the selected non-lexical features in each of the annotated dialogue acts, shown in Figure 7.4, suggest that the task is non-trivial without further information.

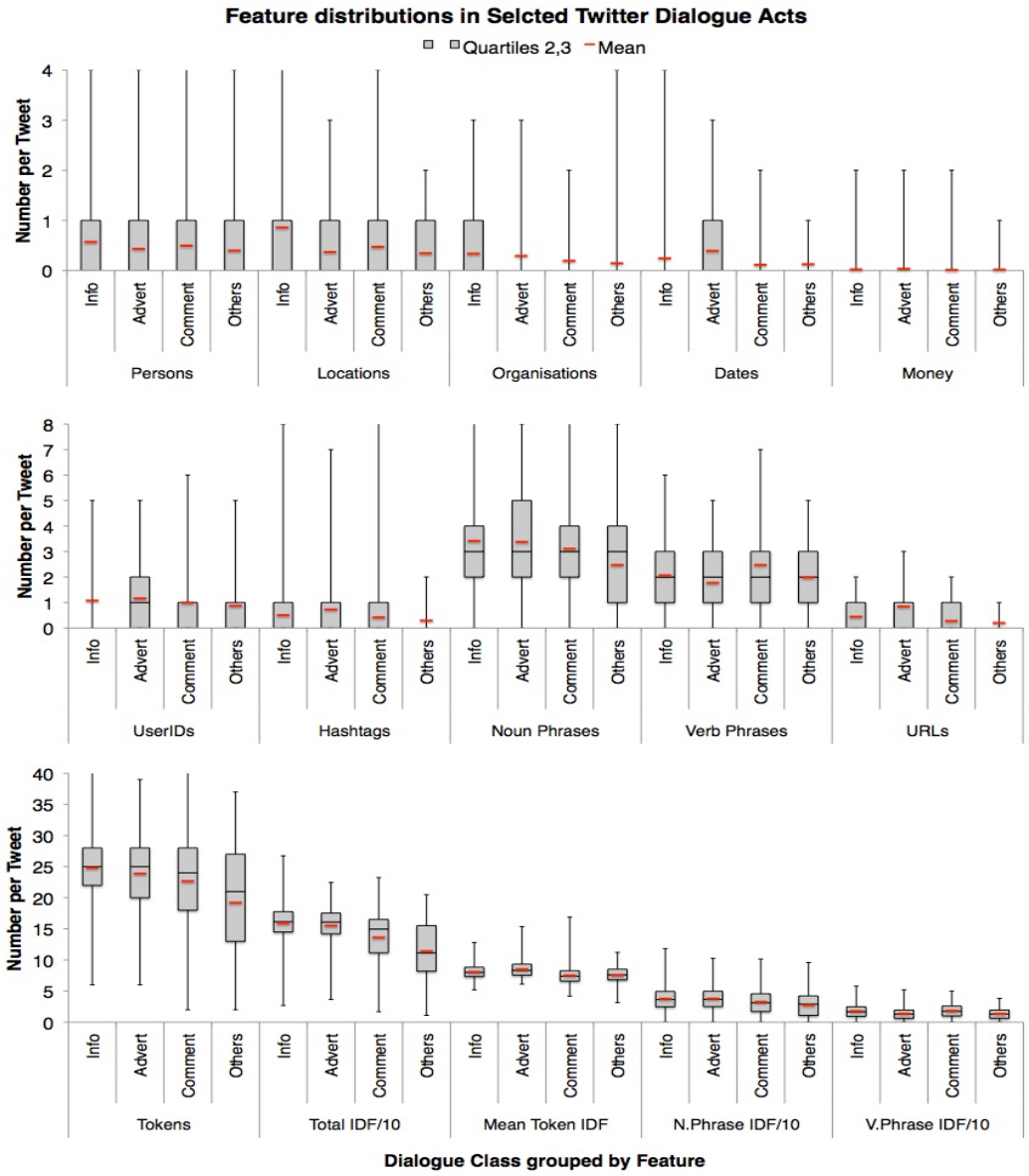


FIGURE 7.4: Non-lexical feature distributions across Dialogue Acts in DAAT-1285 Tweet sub-corpus

Differences in feature distributions across the selected dialogue acts are small. This suggests that before considering any feature interaction, $H5_{null}$ is true. Named Entity mentions, particularly locations, are slightly more frequent in explicitly informative Tweets. Perhaps not surprisingly, URLs and dates are most frequent in advertisements. However, interaction between features may help make class distinctions. For example, as summarised in Table 7.6, the total number of entity mentions, and the variation of their type in any Tweet, do show some differences between the classes.

A multiple binary classifier approach was taken for dialogue act detection, training for one class against all the others. Models were not built for identifying questions or those Tweets with no identifiable form of information because there were

Class	Named Entity Mentions		Named Entity Types	
	Mean	Std.Dev.	Mean	Std.Dev.
Informative	1.77	1.31	1.26	0.78
Advert	1.10	1.08	0.86	0.74
Opinion/Comment	1.17	1.23	0.87	0.82
Other	0.89	1.31	0.65	0.80

TABLE 7.6: Number of Named Entity mentions and types per Tweet in selected Dialogue Acts

insufficient examples of these classes.

Classifier	Lexical Unigram			Non-Lexical			Combined		
	Recall	Prec.	F1	Recall	Prec.	F1	Recall	Prec.	F1
Informative									
Maxim Entropy	0.93	0.64	0.75	0.72	0.65	0.67	0.86	0.62	0.67
Naive Bayes	0.80	0.72	**0.75**	0.73	0.63	0.67	0.80	0.68	0.73
C4.5	0.69	0.64	0.56	0.98	0.58	**0.72**	0.87	0.69	**0.76**
Linear SVM	0.63	0.71	0.63	0.74	0.68	0.71	0.74	0.69	0.70
Average	0.76	0.68	0.67	0.79	0.64	0.69	0.82	0.67	0.72
Opinion/Comment									
Maxim Entropy	0.36	0.70	0.43	0.13	0.64	0.18	0.26	0.54	0.33
Naive Bayes	0.44	0.66	**0.49**	0.46	0.42	**0.44**	0.25	0.66	0.34
C4.5	0.19	0.48	0.26	0.33	0.54	0.38	0.20	0.66	0.28
Linear SVM	0.03	0.65	0.06	0.32	0.65	0.39	0.36	0.79	**0.43**
Average	0.26	0.62	0.31	0.31	0.56	0.35	0.27	0.66	0.34
Advert									
Maxim Entropy	0.85	0.36	**0.50**	0.05	0.13	0.07	0.24	0.68	0.35
Naive Bayes	0.31	0.73	0.41	0.44	0.24	**0.31**	0.35	0.43	**0.36**
C4.5	0.24	0.65	0.34	0.01	0.15	0.02	0.02	0.30	0.03
Linear SVM	0.00	0.00	0.00	0.09	0.56	0.15	0.13	0.51	0.20
Average	0.35	0.44	0.31	0.15	0.27	0.14	0.18	0.48	0.24

TABLE 7.7: Classifier performance in detecting 3 pragmatic intent classes averaged over 10-fold classification using unigram and non-lexical features

Binary classifier performance varied across the Informative, Advert, and Comment classes, in 10-fold cross-validation experiments as summarised in Table 7.7. Since the class representation is not balanced in the sub-corpus, Precision and Recall are used as evaluation metrics:

$$Recall = \frac{tp}{np} \tag{7.1}$$

$$Precision = \frac{tp}{tp + fp} \tag{7.2}$$

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{7.3}$$

where tp is the number of true positive class identifications, fp is the number of false positive class identifications, and np is the number of class examples in the corpus.

Unigrams performed reasonably well for identifying the Informative class when using Naive Bayes or Maximum entropy models, achieving an F1 score of 0.75 with either. However C4.5 and SVM models did not perform so well with unigram features, resulting in an average F1 score of 0.67 across the four classifier technologies.

The difference in classifier performance is also to be observed for the Comment and Advert identification tasks. This is to be expected though given the limited vocabulary and small number of examples resulting from a relatively small corpus. The entropy in feature space would rise significantly when model using unigrams are scaled up to meet the scale and variety in vocabulary seen in practice.

Explicitly informative statements are the most readily determined class examined here, with little performance difference evident between classifiers or feature sets. Performance of non-lexical models is marginally more consistent across classifiers than unigram model, and achieve an average F1 score of 0.69 versus the lexical model average F1 score of 0.67. The combination of features yields an average classifier F1 score of 0.72. These results support hypothesis $H2$, showing evidence that non-lexical characteristics of a message do carry some information on whether or not that message conveys an explicit statement, falsifying $H2_{null}$. The results for the other two classes are less promising though. F1 scores for all models are no more than 0.5 and much more inconsistent across classifiers. Models for adverts appear to be especially hard to learn.

These results suggest that while words and other lexical features are good general purpose features, the non-lexical features more specifically capture information correlating expression of factual information. Neither the vocabulary learnt, nor the non-lexical features examined are sufficient to model expressions of comment or advertisements. The amount of training data for these classes is very small; improved performance might be achieved with more coverage. The small corpus size may also help explain the relatively high performance of unigrams across each class, yielding a small vocabulary and a correspondingly low entropy. One would expect the entropy in unigram feature space to rise significantly when scaling up. Also, as vocabulary and corpus sizes increase, one might expect words to become less specific to particular Dialogue Acts. The entropy of the non-lexical feature space used in this work would not rise with increased corpus size. Given the independence of the non-lexical features from the size of the vocabulary one should expect the performance of the corresponding models to remain at a similar level when scaling up.

The combination of three binary classifiers to give a four way classification model using a winner-takes all approach – based on classifier confidence – was also tried. This set-up always resulted in a classification of either Informative or Other (i.e. Adverts, and Opinion/Comment were mis-classified). This may be a result of having relatively small amounts of example labelled data. However, this is not a significant issue here because the key class of interest is Informative.

In principle the identification of Tweets advertising, passing comments or giving opinions (and indeed asking questions) could be further pursued. However, it would require further annotated data and the interest here is the identification of Tweets explicitly providing information; these particular classes do not do so. Therefore the investigation focussed on the Informative vs. all other classes classification

task.

7.7 News vs. Explicitly Informative Statements

The experiment described in Section 7.6 showed that non-lexical features can help discriminate explicitly informative Tweets from those performing other dialogue acts. The starting assumption was that Tweets providing News are explicitly informative. However, inspection of errors in News classification given in 7.5.3 suggested that not all Tweets containing explicit statements are newsworthy, which has the consequence that if one is wishes to build a model to discriminate utterances that make explicitly informative statements from those that do not, then using News as a proxy for explicitly informative will result in reduced discriminatory power. To what extent does using News and Non-News to represent the desired and undesired utterances, instead of data annotated specifically for the desired task, affect discriminatory models? Can news-bearing Tweets be used to represent informative Tweets? The newly annotated Informative Tweet sub-corpus allowed for further investigation of these questions.

An annotation comparison experiment to test models based on one annotation scheme on data tagged with the other annotation scheme was designed. In this experiment models were built on the original 'news-event' annotated sub-corpus, and also on the Informative Tweet sub-corpus. Each sub-corpus model was tested on the other sub-corpus. The news models were created from the combination of the BAL-4000 and LBAL-572 Tweet sub-corpora using the unigram feature vectors and the non-lexical feature vectors, and the four selected classifier technologies, as used in 7.3.2 above.

The Informative Tweet model was created from 710 Tweets marked as Informative by at least four judges and 570 marked as either Question, Opinion/Comment, Advertising or Other, in the DAAT-1285 sub-corpus. (5 Tweets were dropped for balance in 10-fold cross-validation.) Again, all four classification methods previously used were tested with each feature set. Results are shown in Table 7.8.

Classifier	**Train: News. Test: Informative**				**Train: Informative. Test: News**			
	Recall	Precision	F1	Accuracy	Recall	Precision	F1	Accuracy
Unigram								
Max.Entropy	0.82	0.64	0.72	0.66	0.66	0.83	0.73	0.76
Naive Bayes	0.92	0.65	**0.76**	**0.69**	0.77	0.75	**0.76**	**0.76**
C4.5	0.49	0.67	0.57	0.60	0.20	0.82	0.32	0.58
Linear SVM	0.72	0.65	0.69	0.64	0.18	0.76	0.29	0.56
Average	0.74	0.65	0.51	0.65	0.45	0.79	0.54	0.67
Non-Lexical								
Max.Entropy	0.93	0.62	**0.74**	**0.65**	0.57	0.86	0.69	0.74
Naive Bayes	0.84	0.59	0.69	0.60	0.57	0.84	0.68	0.73
C4.5	0.92	0.60	0.72	0.63	0.91	0.68	0.78	0.74
Linear SVM	0.95	0.61	0.74	0.64	0.63	0.88	**0.73**	**0.77**
Average	0.91	0.61	0.72	0.63	0.67	0.82	0.72	0.75

TABLE 7.8: News event model prediction of Informativeness and Informativeness model prediction of News

Variable results are observed using unigram models for News-event Tweets, testing for explicitly informative statements. Unsurprisingly, this is below the performance expected for selecting news bearing Tweets, as seen in Table 7.4. However Naive Bayes and Maximum Entropy unigram models built on explicitly informative Tweets gave more accurate classification of news bearing Tweets (0.76 accurate versus 0.66 and 0.69 respectively) demonstrating preference for correct selection as reflected in greater precision (0.83 and 0.75 versus 0.64 and 0.65 respectively). C4.5 and SVM models were even more cautious resulting in much lower overall F1 and accuracy scores. Average classifier accuracy, taken across the classifier types, were 65% for prediction of Informative Tweet by unigram News model, and 67% for prediction of News Tweet by unigram Informative model. Factors in this difference may be vocabulary being selective for topics not in the corresponding set, some vocabulary correlated with the act of informing being captured, or a combination thereof.

The non-lexical features yield a more consistent performance across the classifier techniques when tested on informative Tweets having trained models from the News-event Tweets, although none of the models performs as as well as the Naive Bayes unigram model. Average classifier accuracy for non-lexical features trained on News-event Tweets and tested on informative Tweets was 64%. Training with non-lexical features for informative Tweets and testing on News-event Tweets gave an average classifier accuracy of 75%.

Overall, the Informative Tweet models are better at predicting news bearing Tweets than News Tweet models are at predicting explicitly informative Tweets. The best model, whether predicting Informative from News-event training, or predicting News from Informative Tweet training, was a Naive Bayes unigram model with an F1 scores of 0.76 for each. (C4.5 and SVM with non-lexical features were close in predicting News, C4.5 slightly less accurate but with F1 score of 0.78, and SVM slightly more accurate but with an F1 score of 0.73) These results support the idea that news-bearing Tweets are informative but that informative Tweets do not necessarily bear news and, secondly, that our non-lexical feature set is better at capturing whether messages are explicitly informative or not than unigrams.

It is possible to increase detection rates at the cost of precision in the results by applying a threshold to the score or probability a classifier gives that an example belongs to the target class. This allows users to accept or reject according to the cost and type of errors they are willing to accept. A Receiver Operator Characteristic curve shows the cost in false detection errors for a given correct detection rate as the acceptance threshold is reduced. ROC performance curves for the best four classifier models, using the two feature sets, for the two cross-corpus experiments are shown in Figure 7.5, with AUC figures for all classifiers shown in Table 7.9.

It is noteworthy that for a given true positive rate (Recall) below approximately 90% the false positive detection rate is lower in the detection of news bearing Tweets

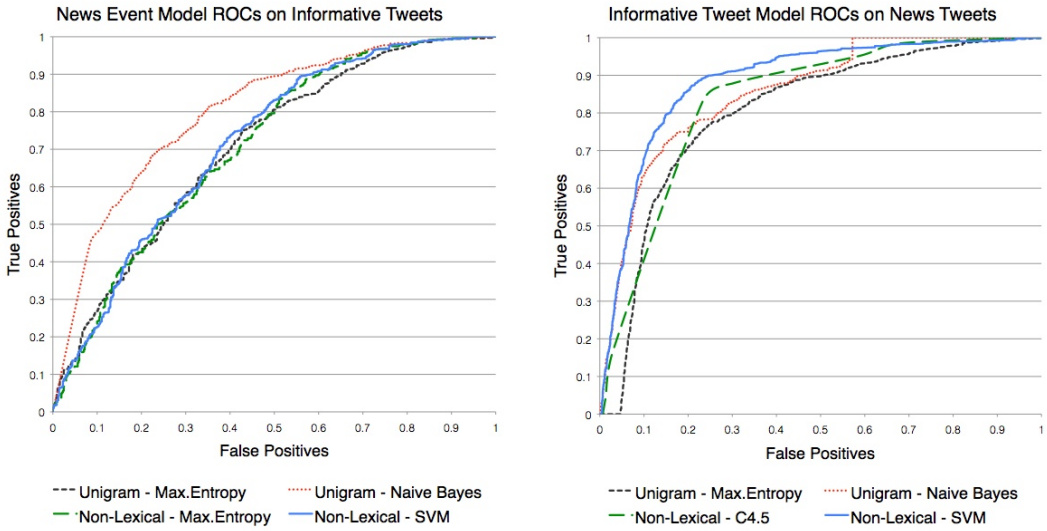


FIGURE 7.5: ROC results for best classifiers across News-Event and Informative Tweet sub-corpora

Classifier	Train: News. Test: Informative		Train: Informative. Test: News	
	Unigram	Non-Lexical	Unigram	Non-Lexical
MaxEnt	0.71	0.71	0.80	0.80
N.Bayes	**0.80**	0.60	0.86	0.84
C4.5	0.59	0.65	0.60	0.83
SVM	0.66	0.72	0.41	**0.89**

TABLE 7.9: AUC figures for ROC results for classifiers across News-Event and Informative Tweet sub-corpora

using the informative statement model than for informative statement detection using the news-event model. This supports the idea that it is easier to distinguish Tweets that make statements, with some of those statements being news, than to distinguish statements of news.

The top News Tweets classified as not informative using non-lexical feature and unigram models were:

By unigram

- RT @TransferNewsCen: Reports in Spain say Hererra has agreed a 5 year deal with Manchester United ahead of his move-Nik
- RT @SportsCenter: SOUTH CAROLINA ANSWER BACK! Connor Shaw finds Nick Jones for a 18-yard TD pass as South Carolina ties Georgia, 17-17. #SC
- RT @GrantWahl: Jozy Altidore cleared to resume full training ahead of Costa Rica-USA WCQ. Has scored in his last 5 internationals, a US rec

By non-lexical features

- RT @realmadriden: .@GarethBale11: "It's a dream come true for me to be here. I would like to thank the club for making it happen." #Welcome

- Wind 0,0 km/h —. Barometer 1014,4 hPa, Rising slowly. Temperature 8,1 C. Rain today 5,1 mm. Humidity 99%
- Microsoft-Nokia Deal Proves Apple Was Right All Along http://t.co/iiZNlJ1n3a

The first of these are Sport related. Sport related terms may occur in more Dialogue Act classes than Explicitly Informative. Misclassifications using a unigram model are presumably due to the words being associated more with uninformative messages. The News Tweets classed as Non-Informative using non-lexical features include a quote containing personal pronouns - a likely indicator of an Opinion or Comment. Arguably this is a reported opinion rather than a News statement. The third non-lexical model example was misclassified due to Name Entity extraction having failed to identify Microsoft, Nokia and Apple as Organisations, illustrating the non-lexical model dependence on accurate feature extraction.

Non-Informative Tweets most strongly classified as News by the News model were:

By unigram

- RT @messileftfoot: Passarella (ex-ARG) Without winning a World Cup, Messi has shown to be on track to surpass Maradona, at least on a club
- RT @footballitalia: Cesare Prandelli was pleased with the win over Bulgaria, as #Italy are one step away from the World Cup. http://t.co/
- World Cup focus: Mexico: Chicharito's team have a battle on their hands to reach Brazil 2014. http://t.co/4mabEGahNN #MUFC

By non-lexical features

- RT @KumbhMelaTravel: Experience of a lifetime. Be a part of Kumbh. Once in 12 years. Largest spiritual gathering on earth. India; http://t.
- RT @Aghorii2: Delhi Fake Media will twists and create stories to blame #UPriots on BJP, the local newspaper media is having correct ground
- RT @UKISS_Venezuela: @HeawenKisses Please support the leader of KISSme Venezuela to go to the Kpop World Festival, vote here->; http://t.co/

The top Non-Informative Tweets selected by the unigram News model are Sport related and were annotated as Opinion. The top false positive Tweets selected by the non-lexical feature model include Adverts and Opinion. Organisations and Locations feature in all of these, features that are often found in News bearing Tweets.

These results lead to questions about the actual news content of the Informative Tweet sub-corpus. To help gain further insight into how news-event related messages and explicitly informative statements differ, the sub-corpus was further annotated for whether items were news or not. In addition to this, annotation was also

added as to whether messages were related to Sport, Travel, or Weather. This was carried out independently of whether the Tweets were deemed to be News or not. This allowed for finer grain distinction or interpretation of what topical domains constitute News. Whether or not all Sport related Tweets had been, or should be, considered News for example was of particular interest given their prevalence in classification errors.

This annotation exercise was carried out by the author on the Tweets in the Informative Tweet sub-corpus and all Tweets that had 3-way agreement for Dialogue Act annotations, a total of 1570 Tweets. The guidelines adopted for annotation were that messages giving facts in regard to evolving events likely to be of interest to an unspecified audience should be tagged as news, but not those expressing opinion with regard to the news, nor those where the facts are intended as advertisements. Commentary on sport events were also deemed not to be news although final results were.

Of the 1570 Tweets, 470 were tagged as News given the above guidelines, 458 of which had been assessed as explicitly informative, leaving just 12 (2.6%) 'uninformative' News Tweets. Inspection of these Tweets suggests that these contain reports of expressed opinion or comment that are themselves the news. All were annotated as "opinion or comment". Where news agencies are expressing comment or speculation on evolving events, are they reporting news? This is perhaps a grey area for making assessments. The Tweets in question are given in Table 7.10 along with the predicted classes given by the two SVM models created from the original News annotated sub-corpus.

364 Tweets that had been annotated as Explicitly Informative were not considered to be News, i.e. 44% of the Informative Tweets.

News Event Tweet	Annotation	Unigram SVM Model	Non-lexical SVM Model
@——— Somerville officials order review after alleged sexual assaults that took place ?yards, if not feet away? from ... @bieberrkfc	Opinion	Other	News
RT @FootyRelated: "@SkySportsNews: Arsene Wenger will do 'all he can' to sign Mesut Ozil and Angel di Maria from Real Madrid: " Also known	Opinion	Other	News
RT @CFCtransferlive: BREAKING: Sky Sports understands that Chelsea are looking to bring in a big-name striker in the next few hours. #cfc #	Opinion	Other	News
RT @————_: Had a chat with Mourinho yesterday. He claims Mata's body is "Astonishingly tired" and it could take up to a month until	Opinion	News	Other
RT @timesofindia: #Gujarat encounter cops followed #NaMo and Amit Shah s policy, says jailed DIG Vanzara http://t.co/ABd4zsOUO1 http://t.co	Opinion	News	News
RT @CNBC_Pakistan: Intelligence agencies solution to #Karachi problem: #Karachi problem: decentralize Police on community level	Opinion	Other	News
RT @loudobbsnews: Republican Party establishment has again aligned itself against rank and file, and with Pres. Obama on attacking Syria. T	Opinion	News	News
RT @DailyMtnEagle: Cullman Superintendent offers no apology in her statement about AHSAA sanctions. Still thinks her team is classy. http:/	Opinion	News	Other
RT @NewStatesman: The Trussell Trust hits out at Cameron: the coalition has broken its agreement with foodbanks, from @RowennaDavis http://	Opinion	News	News
RT @————: Government's attempt to stifle #BarrettBrown's free speech rights may hint at their misconduct, mistakes: http://t.co/	Opinion	News	News
US Secretary of State John Kerry: "There is no military solution... but to enforce the standard with respect to the use of chemical weapons"	Opinion	News	News
RT @DamonBruce: Terrible news on what was a great day there. @RicSal80: @cnnbrk: Fan falls to his death at Candlestick Park. http://t.co	Opinion	News	News

TABLE 7.10: News event Tweets not annotated as explicitly informative

7.7.1 Prediction of News Tweets in new Informative Tweet sub-corpus

How well did the News models, described in Section 7.5, predict News Tweets in the News annotated Informative Tweet sub-corpus (NAIT-1570)? The predictions made by the news models created on News-event sub-corpus Tweets were compared against the news/non-news annotations in the NAIT-1570 sub-corpus. The full sub-corpus, i.e. those Tweets where at least three annotators agreed on the Dialogue Act class, was tested. The results are shown in Table 7.11.

	Recall	Precision	F1	Accuracy
Random	50.0%	29.9%	37.5%	50.0%
Unigram				
Max.Entropy	81.0%	35.4%	49.3%	50.3%
Naive Bayes	95.1%	37.5%	53.7%	51.2%
C4.5	52.1%	42.1%	46.6%	64.3%
Linear SVM	72.9%	36.4%	48.6%	54.0%
Non-Lexical				
Max.Entropy	96.2%	35.5%	51.8%	46.8%
Naive Bayes	83.1%	32.0%	46.3%	42.4%
C4.5	95.1%	33.8%	49.9%	43.1%
Linear SVM	97.0%	34.9%	51.4%	45.2%
Combined				
Max.Entropy	95.1%	35.7%	51.9%	47.4%
Naive Bayes	98.1%	35.1%	51.7%	45.5%
C4.5	88.9%	35.0%	50.2%	47.5%
Linear SVM	92.7%	35.4%	51.2%	47.3%

TABLE 7.11: News event Tweet prediction results for News Tweets in the News Annotated Informative Tweet sub-corpus

There are several outcomes may be observed in these results: firstly that performance at filtering out using unigram features scores only about 0.50 in F1, about 0.06 less than the results for the earlier data shown in Table 7.5. These are figures for the 1570 Tweets that received at least 3-way agreement in annotation. Using just those with a top annotation choice receiving at least two more votes than any second choice (1280 Tweets) results in an average 5 point increase in performance across classifiers. Similarly, there is a slightly higher performance increase when testing on the Tweets that achieved 4-way agreement. Full figures are given in Appendix B. Unlike the previous experiments in detecting news, though, neither the non-lexical features nor the combination of the feature sets performed any better (although C4.5 and SVM models using unigrams were weaker at identifying News Tweets than Naive Bayes and Maximum Entropy, as measured by Recall). This may be due to the sampling method used to get representation of the dialogue acts of interest vice the previous sampling focus of news and non-news. The accuracy of classifiers making use of non-lexical features appears to be worse than random. There is a slight bias towards classifying Tweets as News which may have resulted from non-lexical feature differences being over represented in the (relatively small amount of) training data. Although classifier accuracies are not high, tending to over-classify Tweets as News, they do all filter better than random choice, reflected in higher F1 scores. No classifier performs well at filtering out Non-News Tweets, despite having been trained on balanced News and Non-News classes.

A possible factor in the low performance could arise from any inconsistency in annotation of what constitutes News. The training data and the test data were annotated independently. As described above, sub-corpus used for testing was also annotated for relatedness to the categories of Sport, Weather and Travel. Any selection correlation with these categories may reveal some difference in whether annotators felt these topics were generally news or not.

Of the 470 News Tweets in the sub-corpus, 168 were annotated as being Sport related (36%), 14 as Travel related (3%) and 7 as Weather related (1%). In the 1100 Non-News set there were 219 Sport related (20%), and no Travel or Weather related Tweets. Table 7.12 summarises. How did the models classify these Tweets? To focus on those that posed consistent problems for the models, those that at least three models mis-classified given a feature set were selected and their proportions of the sub-corpus calculated. These are shown in Table 7.13. As well as the number of consistently misclassified Tweets, the proportion of the Tweets with the corresponding annotations is also given. For example, the number of Tweets annotated as both Sport and News, that were consistently mis-classified as Non-News by unigram models, was 7. This represents 3.2% of the 219 Sport News Tweets. A high proportion may indicate that Tweets falling into those classes may pose an issue for the News models, possibly owing to some difference in annotation guidelines followed for News training data.

Class	Sport	Travel	Weather	Other
News	168	14	7	281
Non-News	219	0	0	881

TABLE 7.12: Number of Sport, Weather and Travel related Tweets annotated as News and Non-News in Informative Tweet sub-corpus

Weather and Travel were not often (i.e. consistently across models) misclassified as Non-News by the models, although the number of examples is relatively very low. Sport related Tweets, on the other hand, were predominantly classified as News. Few of the News Sports related Tweets were consistently mis-classified, but a significant proportion of the Non-News Sports related Tweets were consistently misclassified, more so using non-lexical features. A fifth of the Non-News examples in the sub-corpus are Sports-related. Typically a quarter of the Tweets misclassified as News were sports related. The sports related Non-News Tweets, therefore, were a little harder to identify as such. Overall, this suggests that Sports related Tweets in the training data were predominantly annotated as being News.

Arguably Tweets expressing opinion on sports news are related to the events being reported in the news, but equally it is arguable that the expressed opinions do not constitute News in themselves. The latter position was assumed for the annotation of the test set.

To see what effect tagging Sports News Tweets as News in the test set had, the classification results were re-calculated with all sport related Tweets re-annotated as

Error Type	Total Errors	Sport	Sport=News	Weather	Travel	Other
Unigram						
False Pos.	433	119 (70.8%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	314 (35.6%)
False Neg.	11	7 (3.2%)	14 (3.6%)	-	-	4 (1.6%)
Non-Lexical						
False Pos.	674	161 (95.8%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	513 (56.6%)
False Neg.	8	6 (2.7%)	11 (2.8%)	-	-	2 (0.8%)
Combined						
False Pos.	653	160 (93.5%)	0 (0.0%)	0 (0.0%)	(0.0%)	496 (54.7%)
False Neg.	2	2 (0.9%)	8 (2.1%)	-	-	0 (0.0%)

TABLE 7.13: Number of News/Non-News Tweets consistently misclassified across classifier types. Proportion in topic category in error given in parenthesis.

News (totalling 387 Tweets), and all Travel and Weather re-annotated as Non-News. No false positive Sports News are then possible in this case, but false negatives are. (Also shown in Table 7.13) Similarly false negative Travel and Weather Tweets are not possible with this rule. All the re-annotated Weather and Travel Tweets were misclassified by at least one model, but only Travel Tweets showed consistent misclassification. This suggests that travel related news Tweets in the training data were indeed considered News by the annotators, but that generally weather information was not. A reason for this could be that travel information may often relate to an incident or event affecting travel, whereas weather, barring extreme out-of-the-ordinary conditions, is not considered eventful. Overall, the distinction between News and Non-News Sport related Tweets in the evaluation accounts for about 10 – 13 points in the F1 metric. The Naive Bayes unigram model achieved an F1 score of 0.65 (versus 0.54) where all Sports are considered News, for example.

Table 7.14 provides an examination of whether or not consistent News misclassifications were Informative Tweets. Using a non-lexical model 23.8% of the false positives were annotated as Informative. The proportion for unigram models was 51.0% and 43.4% for the combined feature models. The total number of false positives was highest when using non-lexical feature models, suggesting they produce lower precision models for News identification. More than half of the Informative Non-News Tweets were consistently misclassified if non-lexical features were used, under half if unigram features alone were used. All bar one of the News Tweets consistently given false negative classification had been annotated as Informative. (459 of the 470 Tweets annotated as News were also annotated as Informative). These results support the idea that Tweets bearing News are explicitly informative, but that explicitly informative Tweets are not necessarily bearing News.

Error Type	Unigram		Non-Lexical		Combined	
	Total	Informative	Total	Informative	Total	Informative
False Pos.	433	221 (48.7%)	674	161 (61.2%)	653	284 (62.6%)
False Neg.	11	11 (3.0%)	8	7 (1.9%)	2	2 (0.5%)

TABLE 7.14: Number of consistently misclassified News/Non-News Tweets annotated as Informative. Proportion of the class in error given in parenthesis.

The analysis above was carried out on the full sub-corpus, i.e. those Tweets with

at least 3-way agreement between annotators for the selected dialogue acts. The same analysis was also carried out for those Tweets with the majority annotation vote at least two more than the second most popular annotation, and also for those with at least 4-way inter-annotator agreement. The findings for these divisions of the sub-corpus yielded similar results (given in Appendix B). One observation from the results obtained when constraining evaluation to Tweets with a higher inter-annotator agreement, though, was that the proportion of Non-News Sport related Tweets often falsely classified as News declined if non-lexical features were used. This suggests that those Sports related Tweets that were harder for annotators to decide a Dialogue class for (Informative or Opinion for these Tweets) were also hard for the models to distinguish at non-News.

The final take-away from this analysis is that caution should be exercised with regard to annotation standards for what constitutes an item of News. Not only are there the issues of domain dependency and temporal effects, but also receiver interpretation. All of these may adversely affect performance and evaluation when moving from one setting, or corpus, to another.

7.7.2 Comparison of non-lexical feature distributions in Informative Tweet sub-corpus

This section provides statistics of the non-lexical features in the classes annotated in the sub-corpus created for Dialogue Act analysis. Box plots for each of the non-lexical features extracted from the Tweets are shown for those that were considered news-bearing and those not news bearing in Figure 7.6. Figures are also summarised in Table 7.15. One may observe that there are few differences in feature distributions between the two categories, and that the distributions are broadly similar to that found for the complete Redites corpus (shown in Figure 7.1). Location and Organisation names are found a little more often in News Tweets than non News Tweets. Pronouns are less likely to be found in News Tweets.

There is a similar lack of difference between the feature distributions over the Explicitly Informative and other Dialogue Act categories, although again, Location and Organisation names are a little more likely to be found in Explicitly Informative Tweets than in others, and pronouns are less frequently found. Box plots for these distributions are shown in Figure 7.7. Again Figures are summarised in Table 7.15 (including IDF features not shown in the box-plots owing to the relative scale).

Explicitly informative Tweets are slightly longer on average than other classes, and have slightly more selective terms as measured by IDF. This difference is also seen between News and Non-News Tweets.

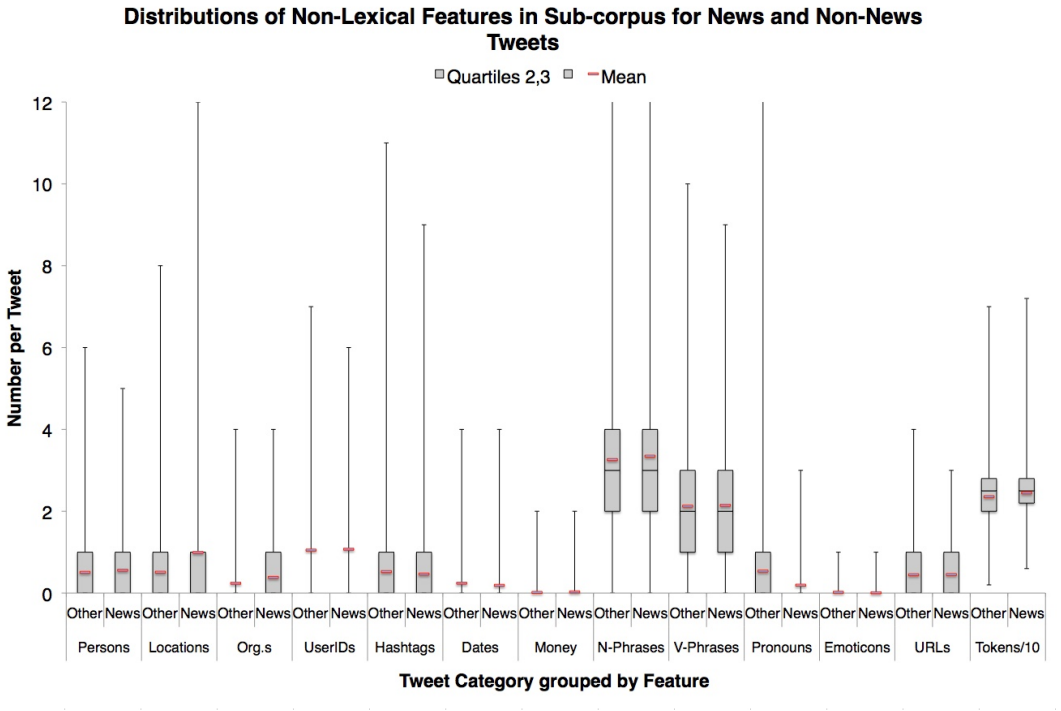


FIGURE 7.6: Non-lexical feature distributions in Informative Tweet sub-corpus divided by News v. Non-News

7.8 Contributions

The work presented in this chapter represents several novel contributions. In exploring three potential classification schemes, by which social media could be filtered prior to analysis for for information discovery, a simple Dialogue Act classification scheme and a Subjective Focus classification scheme were introduced, complementing a standard News / Non-News scheme. A data set annotated for Dialogue Act and News, along with topic relevance to Travel, Weather and Sport, has also been contributed.

One of the main contributions of this work has been the set of results of classification experiments in which a control for topic was applied. Often, such a control has not been considered for text classification tasks other than topic. This may be an important factor because content words are known to be useful in determining topic and bias in data towards topic could mask underlying effectiveness for the task in hand.

Evidence has been shown that the presence of named entity mentions is a useful feature in determining whether or not a microblog message contains explicit information. Similar evidence has also been found recently by Edouard et al., (2017) in examining whether or not external knowledge bases could be leveraged in identifying messages containing news material. They replaced Named Entities by their types as specified in DBpedia (Bizer et al., 2009) or the YAGO (Suchanek, Kasneci, and Weikum, 2008) taxonomies. However, they also discard features such as those

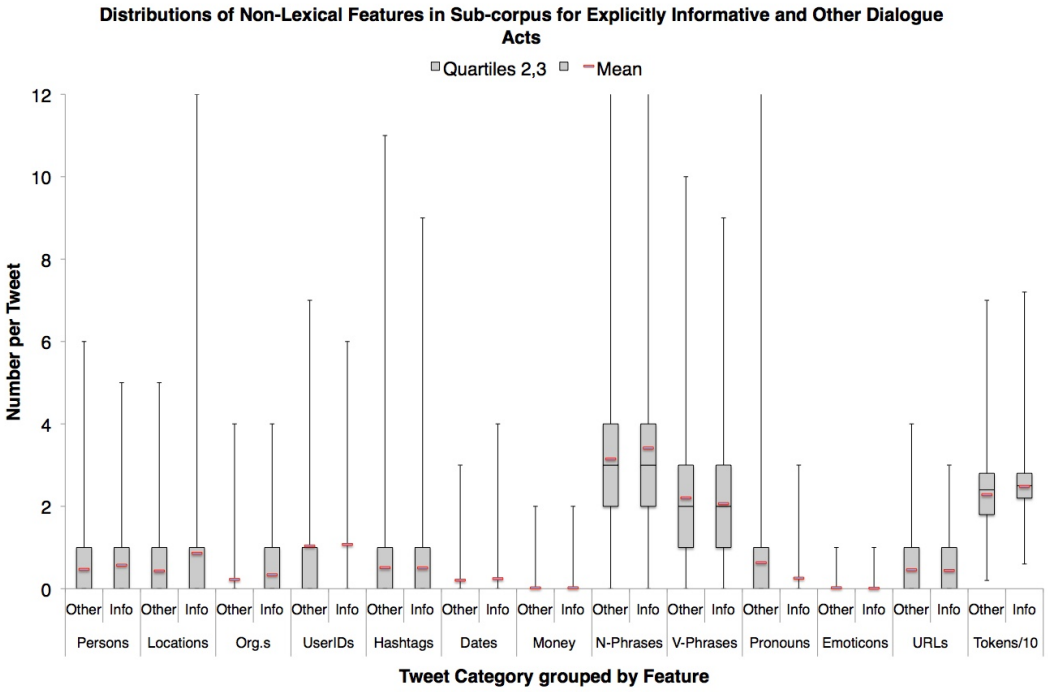


FIGURE 7.7: Non-lexical feature distributions in Informative Tweet sub-corpus divided by Informative v. Other/Non-Informative

in the GRP feature set described in Section 7.4.1, also shown to be useful in the experiments described in this chapter.

7.9 Summary

In a knowledge discovery system that seeks new information from Social Media, it would be useful to process only that text which makes assertions. This chapter has described research examining an approach for filtering streams of short messages, as exemplified by Twitter, for those that make explicitly informative statements.

Two classification tasks were investigated. The first was to separate messages into News and Non-News related posts. The second task was to separate messages into one of a set of five Dialogue Acts: Informative, Opinion or Comment, Question, Advert, and Non-Informative. The classification schemes were described in Section 7.2 along with a scheme to separate messages by what is termed here as a message's subjective 'focus': that is the relationship of a message's subject with the author. The idea of subjective focus is separation of messages that are about the author from those that refer to personal contacts of the author, and from those that referred to neither.

Data for the studies, described in Section 7.3 was obtained from the Redites project (Osborne et al., 2014). This provided a sample of 1.4 million Tweets from 37.5million English language Tweets produced between 09:58 2/9/13 and 11:16 30/9/13. Within this corpus, bursts of temporally related Tweets had been identified and

	Non News		News		Other Dialogue		Inform	
Feature	Mean	Std.Dev.	Mean	Std.Dev.	Mean	Std.Dev.	Mean	Std.Dev.
Person	0.51	0.79	0.56	0.72	0.47	0.76	0.57	0.77
Location	0.51	0.52	0.99	0.65	0.43	0.51	0.86	0.61
Organization	0.24	0.81	0.39	1.07	0.22	0.73	0.34	1.03
UserID	1.05	0.89	1.07	0.67	1.04	0.93	1.08	0.73
Hashtags	0.53	1.04	0.47	0.84	0.51	1.02	0.50	0.96
Dates	0.24	0.54	0.19	0.49	0.21	0.51	0.24	0.55
Money	0.02	0.16	0.03	0.18	0.02	0.17	0.02	0.16
N-Phrases	3.26	1.71	3.35	1.55	3.15	1.73	3.41	1.59
V-Phrases	2.13	1.35	2.14	1.16	2.20	1.39	2.06	1.20
Pronouns	0.54	0.96	0.19	0.46	0.63	1.06	0.26	0.56
Emoticons	0.02	0.14	0.01	0.09	0.02	0.15	0.01	0.11
URLs	0.45	0.56	0.46	0.56	0.46	0.58	0.44	0.54
Tokens	23.57	6.96	24.49	5.52	22.81	7.37	24.79	5.59
AVG-IDF	146.97	40.00	158.00	28.03	140.92	42.15	158.80	29.49
AVG-AVG-IDF	7.89	1.39	8.15	1.05	7.85	1.48	8.09	1.11
AVG-NP-IDF	35.42	19.87	37.01	17.97	34.02	19.96	37.60	18.58
AVG-VP-IDF	16.50	21.88	18.99	21.01	16.64	20.59	17.80	22.63

TABLE 7.15: Summary of non-lexical feature distributions in Informative Tweet sub-corpus

grouped into 'events' using a streaming clustering algorithm (Petrović, Osborne, and Lavrenko, 2010) to detect related Tweets. Tweets in a sample of the events had been further hand-annotated by the Redites team as News related events or Non-News related events. Examples were drawn from this data to create sub-corpora appropriately balanced for training and testing classifier models. Although the data contained messages that contained News or Non-News, it was not annotated for Dialogue Act. A new sub-corpus was created by applying the Informative Dialogue Act scheme defined in Section 7.2, annotating a sample of 1,700 Tweets. The annotation process, described and analysed in Section 7.3.2 resulted in a sub-corpus of 1,285 Tweets, annotated for Dialogue Act, where the majority agreed class received at least 2 more votes than the second most popular choice amongst 5 judges. Fleiss's κ was measured at 0.67 which has a recommended interpretation of "substantial agreement". The Tweets were also annotated for focus (Fleiss's κ = 0.64), however this showed that the significant majority of the messages were deemed to have a World focus. Too few Tweets were identified as Self or Contact focussed to investigate their systematic identification.

The approach, outlined in Section 7.4, uses supervised machine learning to create discriminative models based upon features of example messages from desirable and undesirable classes. Naive Bayes, Maximum Entropy, Decision Tress, and Linear Support Vector Machines were selected as exemplar machine learning techniques for use in exploring the approach.

Traditional text classification models often use "bag-of-words" features. The experiments reported here used this approach as a baseline for comparison, but sought to examine the efficacy of features that are characteristic of a message's content beyond the lexical level. To this end four categories of non-lexical features were selected and described. The GRP set of features included counts of features that could be determined from the characters used to denote them. They included Emoticons,

Hashtags, URLs and User Identifiers. The NE set of features included counts of Person names, Location names, Organisation names, Dates, and sums of Money. The SYN set of features included counts of noun phrases, verb phrases, and pronouns determined from part-of-speech tag sequences. Finally the FRQ set of features included total and average Inverse Document Frequency scores for words and phrases appearing in a message. The IDF scores used were determined from the 35.7 million Tweets of the Redites corpus, treating each Tweet as a document.

Having described the class schemes, data and the feature set used, the chapter then described the experiments carried out in investigating the proposed filtering approach.

Section 7.5 described experiments to explore the efficacy of the proposed approach to identify News event Tweets from Non-News event Tweets, and compare the use of the non lexical features described in Section 7.4.1 with more traditional lexical unigrams. The initial experiment compared the classifiers with each of the feature sets using the class-balanced News/Non-News BAL-4000 sub-corpus described in Section 7.3. Tweet vectors were grouped by event and split for 10-fold cross-validation, such that any one event was represented in just one fold, giving some control for any event specific vocabulary.

Naive Bayes, Maximum Entropy, C4.5, and SVM classifiers were trained with unigram words, the non-lexical feature sets, and a combination thereof. Results of the 10-fold cross-validation experiment, shown in Table 7.4, indicated that Lexical and non-lexical features capture different contributory information. Unigram feature models achieved average fold accuracies of between 70.1% (C4.5) and 77.6% (Maximum Entropy), albeit with high variance. Non-lexical feature set based models did not perform as well generally, average fold accuracies ranging between 50.0% (Maximum Entropy with SYN features) and 78.6% (SVM with FRQ features). However, use of all of the non-lexical features resulted in classifiers that out-performed corresponding unigram models, except for Naive Bayes. The SVM model achieved the highest accuracy, with all non-lexical features, at 85.0%. The combination of lexical and non-lexical features yielded the most accurate models overall, ranging from 85.2% (C4.5) to 92.1% (Naive Bayes), an average 18% relative improvement over unigram models.

The next experiment, described in Section 7.5.1, used a 'leave-one-out' methodology to examine feature set contribution to overall classification performance. Model portability was also examined by training on the 4,000 Tweets used in the cross validation experiment, and testing on 572 Tweets from later occurring News and Non-News events. Average classifier accuracy using unigram models was shown to drop by 20% from the closed setting described above, while an an average 8% accuracy drop was observed for non-lexical feature based models. The difference in the two relative reductions in performance suggest that News Tweet identification models using non-lexical features are more temporally stable than using unigrams.

Features making use of message token frequency, and the presence of various types of Named Entity were found to be the most useful. Although simple orthographically determined features, such as hashtags, had some discriminatory power they were not always effective. Compound phrase features which were detected using the MESME simple part-of-speech parser described in Chapter 5, were not found to be useful, although this may have been as the result of feature sparsity and low detection rates from the parser.

Section 7.5.2 provided an analysis of classifiers, and features sets, in a more realistic, unbalanced, filtering setting. 173,514 Tweets were tested, having added all the Tweets marked as Non-News event to the News event Tweets within the annotation time.

Classifier confidence was used to rank order the predictions made by the models, trained on the balanced sub-corpus, for these Tweets. ROC curves, shown in Figure 7.3, were produced for each model by applying a threshold to classifier confidence. AUC analysis showed the best performing unigram and non-lexical feature models were achieved using SVM, at 86.1% and 90.9%, and the best model overall was a Maximum Entropy model using all the features, at 94.5%. This provided further evidence that non-lexical features carry useful information for identifying Tweets imparting News.

An analysis of mis-classifications, in Section 7.5.3, showed that some Tweets falsely classified as News were explicitly asserting information, and that information might actually be considered newsworthy. Less than perfect feature extraction, particularly of Named Entities, was one possible explanation for false negatives for News Tweet identification when using non-lexical feature based models.

Section 7.6 described experiments carried out to examine how well trained models could distinguish the Dialogue Acts annotated for the Informative Tweet subcorpus. Classification performance in identifying three of the Dialogue Acts, using lexical and non lexical features, was tested in a 10-fold cross validation experiment. Too few Tweets were annotated as Question or Non-Informative for training, leaving Informative, Advert and Opinion/Comment Tweets to be used. A multiple binary classifier approach, using 1,280 3-way+ agreed Tweets, was employed.

Explicitly Informative Tweets were the most readily identified class, with little difference in performance evident between classifier types and feature set used. Non-lexical models performed marginally better than lexical models, achieving an average F1 score of 0.69 across classifiers, versus the lexical model average F1 score of 0.67. The combination of features gave an average classifier F1 score of 0.72. These results provided support for the second of the principle hypotheses in this thesis:

*H*2: Sentences asserting information may be distinguished from those intended to perform other functions using non-lexical features.

Performance for Advert and Opinion/Comment was shown to be less promising with inconsistent F1 scores of no more than 0.5 across classifiers. This could be a result of the relatively small amounts of training data used. The combination of three binary classifiers to give a four way classification model using a winner-takes all approach – based on classifier confidence – was also tried. This set-up always resulted in a classification of either Informative or Other (i.e. Adverts, and Opinion/Comment were mis-classified).

Section 7.7 described experiments to compare the News models with the Informative Tweet models. The Informative Tweet sub-corpus was used for testing the News/Non-News discriminative models developed in the work described in Section 7.5. Results of this experiment showed that use of non-lexical features gave improved performance over unigrams in discriminating explicitly informative statements from Tweets performing other dialogue acts. Both feature sets gave worse results than those achieved in experiments using just the original News-event corpus, indicating that explicitly informative Tweets exist in both news and non-news classes. New models were created by training with feature vectors derived from Tweets in the Informative Tweet sub-corpus and tested on discrimination of Tweets, in the original corpus, marked as related to News events from those marked as related to Non-News events. This showed Informative Tweet models were better able to identify News Tweets from Non-News Tweets than News Tweet models were able to identify Informative Tweets from others. Average classifier accuracy, taken across the classifier types, was 65% for prediction of Informative Tweet by unigram News model, and 64% for prediction by non-lexical Informative model. Average classifier in predicting News event Tweets using unigram Informative Tweet models was 67%, but was 75% when using non-lexical feature models.

However, results observed in this experiment were variable across classifier and feature set combinations. The best model, whether predicting Informative from News-event training, or predicting News from Informative Tweet training, was a Naive Bayes unigram model with an F1 scores of 0.76 for each. (C4.5 and SVM with non-lexical features were close in predicting News, C4.5 slightly less accurate but with F1 score of 0.78, and SVM slightly more accurate but with an F1 score of 0.73) However, the use of non-lexical feature based models, in general, marginally out-performed unigram based models. Performance at detecting News-event related Tweets when trained on Informative Tweets, with non-lexical features, was only marginally less than that achieved with models trained on News-event Tweets. This result provides some further evidence in support of $H2$.

The Informative Tweet sub-corpus was further annotated for whether the content was disseminating news or not, and also for topics related to Sport, Travel and Weather. Having established that the classes of News and Explicitly Informative are not synonymous, this further annotation allowed for the original News classification models to be further evaluated at News classification. Results of the evaluation

showed weaker performance than the previous evaluation. Average F1 score was about 0.05 less at 0.50 when using unigram models. However, better performance was observed for those Tweets that were easier for annotators to agree were explicitly informative or not; these were easier to automatically classify as News or Non-News. Non-lexical feature set News models did not perform any better. This drop in performance may have resulted from the sampling for Dialogue Act rather than News. For example, opinion is often expressed on News topics, especially Sport. Analysis of Sport related Tweets in the sub-corpus found that the distinction between News and Non-News Sport related Tweets in the evaluation accounted for a drop of about 10–13 points in the F1 metric.

All of the Tweets in the Informative Tweet sub-corpus that were consistently given false positive classification as News bearing were annotated as Informative, and Between 40% and 49% of the Tweets consistently given false negative classifications were annotated as Informative, indicating that Tweets bearing News are explicitly informative, but that Explicitly Informative Tweets include Non-News items.

The analysis of annotation with respect to Sport in News and Non-News Tweets showed that caution should be exercised with regard to annotation standards for what constitutes an item of News. Not only are there the issues of domain dependency and temporal affects, but also receiver interpretation. All of these may adversely affect performance and evaluation when moving from one setting, or corpus, to another.

Although this chapter has shown that information as to whether a message is making an explicitly informative statement is carried at a deeper level than the message's lexical surface (falsifying $H2_{null}$), the experiments reported found no evidence to support similar hypotheses for other pragmatic classes. In particular, the number of mentions of named entities and other concepts were not found to be significantly different in Tweets classified as Informative or some other Dialogue Act, supporting:

$H5_{null}$: Multiple direct mentions of concepts and named entities are no more common in explicitly informative statements than in messages conveying other Dialogue Acts.

The Dialogue Act categories annotated were found to be unevenly represented in the random sample of Twitter event burst messages. Further work would require the annotation of more data to enable these classes to be modelled more effectively. More annotated data from a longer collection window would also allow temporal stability of models to be assessed.

The novel contributions the work presented in this Chapter include a simple Dialogue Act classification scheme and a corpus of Microblog messages annotated using said scheme. The corpus is also annotated for wether or not messages contain News along with identification of Sport, Weather, and Travel related topics. Results were presented from experiments that were controlled for topic and provided evidence

that the presence of named entity mentions, amongst other features, is a useful feature in determining whether or not a message makes an explicit statement.

Chapter 8

Conclusion

The work in this thesis has addressed how one may detect new interesting information intentionally imparted by people by means of the written word via online Social Media.

Chapter 1 set out the context and the motivations for the thesis, setting out two objectives: These were to find out (1) whether or not documents containing new interesting information could be found through an unexpected number of specific references, and (2) whether or not sentences intended to assert information could be distinguished from those intended to perform other functions using non-lexical features. Chapter 2 described the relevant concepts and Chapter 3 provided a review of relevant research carried out in these areas. Chapter 4 set out the approaches taken to address the two principle objectives of the thesis. It introduced the idea of using a rolling time-slicing window for comparing "current" feature frequencies with past frequencies as a basis for detecting significant changes in feature use. It also introduced a simple Dialogue Act classification scheme for selecting text intended to inform the reader from that intended for other purposes. Selected supervised machine learning techniques for building classification models were described. Chapter 5 provided details of the text features used in the methods explored, and the tools used for their extraction. Consideration was given to the performance of extraction tools, which included an evaluation of a multiword expression extraction tool developed for the work herein. Chapter 6 explored the first objective of the thesis in detail. It described an analysis of nouns, named entities and multiword expressions in a sample of blogs and news articles, and experiments to determine if frequency variation thereof in Social Media could be used to predict later news stories. Chapter 7 explored the second objective, the filtering of text to select messages intended to inform the reader. It described experiments carried out in classifying microblog messages for news content and explicitly informative statements.

This chapter first provides a recap of the thesis and its motivation in Section 8.1. Section 8.2 summarises the background and the methods adopted for the investigation. Section 8.3 provides a summary of the experiments carried out and the conclusions drawn. The outcomes and contributions of this thesis are summarised in Section 8.4. The chapter concludes in Section 8.5 with a discussion on the next steps

to be taken toward building the information discovery system envisaged at the outset of this thesis.

8.1 Thesis motivations and objectives

This thesis began by introducing the problem of discovering new interesting information from within an ever-increasing amount of online text. It described how curated reporting and informational articles now only form only a fraction of this material as Social Media, including weblogs and microblogging, has become increasingly popular. Some of this content may contain useful information. It was described how that, whereas there are well established information retrieval methods for scenarios where the topic is known, discovering new information with respect to *any* topic remains a challenge. Could new interesting content be found in Social Media, as it is published, when no associated keywords are known?

Journalistic research was introduced as an example where information discovery can be useful. The chapter described how news stories could be considered as a collation of connected pieces of relevant information, either on or about a related aspect of the central topic. Examples were given that show that sometimes this information exists in online text prior to the creation of the news story. It was argued that automated discovery of potentially significant information could be valuable to journalists, but also in other scenarios involving intelligence gathering such as in disaster response.

The detection of informative text, and discovery of new facts therein, was framed as an automatic knowledge acquisition system. It was postulated that such a system would have at its heart a graph representing things and their relationships – a knowledge base. The knowledge base would be grown with new facts as they were discovered from online media as it was produced. However, there are challenges in obtaining asserted information from the sheer volume of online text, and in discovering what therein is new and interesting. These challenges motivated the work reported in this thesis.

Two main hypotheses were put forward. These were that documents containing new interesting information can be found through an unexpected number of references to entities and concepts, and that sentences asserting information may be distinguished from those intended to perform other functions, such as asking a question and making a suggestion, using non-lexical features.

Formally the main hypotheses were stated thus:

$H1$: Some documents containing new information can be found through an unexpected number of references to named entities and concepts.

$H1_{null}$: References to named entities and concepts are no more frequent when related new information emerges than the average rate of mentions.

$H2$: Sentences asserting information may be distinguished from those intended to perform other functions using non-lexical features.

$H2_{null}$: Non-lexical characteristics of a message carry no information on whether or not that message conveys an explicit statement.

A system to discover and acquire new information from textual communications intended for humans must necessarily process natural language. Chapter 2 described some of the central concepts and ideas pertinent to communication of information by means of natural language written and presented online. It started with the idea of a topic – what a document's focus is about. It was argued that in writing about a topic, references to concepts and entities pertinent to the topic are often necessary, although such references could be polysemous or anaphoric, especially when the author can assume the presumed reader has prior knowledge.

Informing the reader is not the sole purpose a document or message may have. Speech Act theory and the idea of Dialogue Acts characterise the pragmatic intent behind utterances. It was described how, under such a characterisation, one might expect an author to make an assertion when seeking to impart information to a readership. Such an act would be most pertinent to the task of information discovery.

One may assume that in seeking information, one is looking to extend one's knowledge. It was observed that knowledge representation has been a key goal of Artificial Intelligence since its inception, and has been approached typically with logical formalism, or, as computing power increased, with relational graphs. Whatever the representation, though, the requirement is detection and discovery, and the chapter went on to consider statistical models of novelty. It was explained that as more data became available it became possible to better estimate background probabilities of textual features and therefore significant deviations from those. Feature selection and inter-feature dependency though present a challenge, owing to the computation required for the number of model parameters resulting from increased feature space dimensionality. A trade-off between computability and model sophistication often has to be made. For example, it was explained how words have often been treated as independent features even though it is known that they are not.

Machine learning techniques have been used to address the feature modelling issue. It was described how supervised learning techniques have been successful in creating classifiers and taggers, and unsupervised learning techniques have provided methods for corpora and document collection analysis. Various learning techniques have been developed and become popular in NLP applications, including Support Vector Machines and Latent Dirichlet Allocation.

Having considered the pertinent ideas, relevant text analysis and mining applications were considered. Sorting documents and messages into wanted and unwanted classes was found to be a common requirement in text processing. Classification

techniques baed on statistical models and/or machine learning have often been employed, not only for document level classes, but also at finer levels for messages, phrases and words. These are key components in information extraction systems.

In turning to information extraction, the exploration of concepts in text analysis returned to the theme of acquiring knowledge from its expression in natural language. This processing is critical for automatic population of knowledge bases, but is a challenging task. There are many ways of referring to the entities and concepts that are involved in the information given. References may be ambiguous or anaphoric, and may rely upon the assumed reader's prior knowledge. Resolving these references may be necessary in order to find predicate-argument tuples in which the entities and concepts are connected with the extracted relationship. Some systems focus on a closed set of desired, canonical, relationships; others engage in Open Information extraction, seeking to extract all relationship expressions (suspending any anchoring of relationship types).

In concluding the exploration of relevant concepts in text analysis, the review returned to the purposes authors may have for the text they write. Assertion of fact is not the only Dialogue Act. A similar and increasingly important purpose is the expression of opinion. It is similarly challenging and its expression may be difficult to distinguish from factual assertion. Opinions could be expressed as if they were facts. A further complication is that expressed fact may or may not be true. The issue of "fake news" was noted although it was not addressed in this work.

8.2 Background to, and decisions made in adopted approaches

Chapter 2 provided a review of published research relevant to the tasks described in this thesis.

It began with research into selecting and sorting documents by topic from continuous feeds, an area known as Topic Detection and Tracking (TDT). The aims of this area of study are to group together different texts from a stream, typically of news reports, that belong to the same story, and also to detect when a new story starts. The latter, however, has been found to be a significant challenge. Although some TREC (Voorhees and Harman, 2005) tasks have tackled sentence level detection of new information within a story, these tasks did not include detection of new stories.

The idea of measuring novelty has been explored, with various distance metrics proposed. One metric, which was found to significantly outperform chronological ordering, was the number of new Named Entities mentioned.

Review of more recent work showed how News dissemination is no longer confined to mainstream news organisations as Social Media has become popular. Approaches to detecting emerging news in blogs have included examination of linking behaviour, and content clustering. The ICWSM corpus (Burton, Java, and Soboroff, 2009), a collection of 44 million blog posts and online news stories, was used in much

of the research undertaken for this thesis. Social Media platforms have developed though, and research in first story detection has turned to examining micro-blogs, as services such as Twitter have become more popular.

Many techniques for first story detection rely on detecting changes in vocabulary use. The review therefore turned to research that has examined term occurrence and evolution therein over time. Words one might consider to be 'content' have been found to show a much higher likelihood of re-occurring in a document than expected from its overall frequency, although use of different words for the same thing may mask this. Rarer terms in document collections are potentially more selective. Weighting models such as $tf * idf$ have therefore been popular and successful in Information Retrieval tasks.

Bursts in email topics, characterised by associated words, have been found to coincide with the development of people's interest the topic, and evidence has been found suggesting topic discussion in blogs can be pre-emptive of news-stories as well as coincidental. Particular "memes" – short phrases associated with a topic – have been found to be reused across media, and increases in the mentions of particular named entities have been observed in linked blog posts and stories. More generally, some research has found that including named entities in information retrieval based models can be effective. This could be because documents such as news stories may be about some central entity, the focus of the discourse. These observations provided some of the motivation for approach taken in this thesis.

News has not been the sole focus for Knowledge Discovery research. Knowledge is important in many other domains such as biomedical research, and many have sought to formally represent such domain knowledge. It was argued that discovery of new knowledge implies that one has old, established, knowledge. Therefore work carried out in knowledge representation was reviewed.

A highly influential idea in AI has been been that of frames or templates, and language understanding has been cast as the task of filling in these templates from linguistic primitives. These ideas have lead to the development of "ontologies" - structures representing concepts, objects, and relationships between them. General purpose and task-based ontologies have been developed, and have been used successfully, but domain ontologies have been built manually in the main. The scale required for ontologies to be useful has been a driver for the development of semi-automated and automated approaches to knowledge acquisition, but it was found that this is needed for the underpinning ontology resources as much as for accumulating new domain knowledge.

Whether detection of new knowledge is carried out prior to, or post representation, though, novelty models are required. Given the assumption of no domain specific ontology, and therefore no resource with which to carry out inferencing, attention was turned towards investigations into knowledge discovery using the application of statistical models. The biomedical domain has been a rich vein for research. Following Swanson's discovery of indirect links in MEDLINE reports, work has examined the potential of concept co-occurrence frequencies as a basis for discovery. Other ideas have included examination of phrases and types of expressions. Moving partially into formal representation, some have applied clustering techniques to extracted relationships, and some have used a domain ontology to create a background knowledge base.

Accurate frequency estimation is an important factor in statistical models. Since estimation in the open text domain necessarily involves sampling, accounting for rare and as yet unseen features is required. Methods for smoothing with various degrees of sophistication were reviewed, however it was found that performance could be sensitive to model assumptions and the amount of data available for estimation, with overfitting a potential outcome for those smoothing models with a large number of parameters.

The research review then moved on to consider what machine learning could contribute towards the development of the envisaged system. Machine learning has been successfully applied to many areas in natural language processing, so the main approaches that have been taken were examined. These fall into two approaches – unsupervised and supervised model training. (Some have also used a combination of the two to try to boost performance in supervised training scenarios). Many unsupervised methods were found not to be naturally suited to filtering, although some have proposed approaches for creating document clusters on the fly. Supervised methods have been widely used where defined classed are desired. It was found that Naive Bayes and Support Vector Machines have been two popular and successful approaches in text classification tasks.

Machine learning techniques have been traditionally applied to defined feature sets. During the course of this work, research on Artificial Neural Networks has developed to the point where features and classification models can be co-learnt, given sufficient data, in an approach known as "Deep Learning". The success in this area, particularly with co-occurrence embedding through auto-encoding through algorithms such as Word2Vec, has resulted in notable popularity in the development of natural language processing tools using large neural models. However they typically require significant computation time and amounts of data to train, as well as careful network design.

It was suggested that some classes of Dialogue Act would be most likely to express new information. Work applying text classification through machine learning

to Speech and Dialogue Acts was therefore reviewed. Detection of various sets of Dialogue Act classes has been attempted in online settings including message boards, discussion forums, and human computer dialogue systems. However, no universal taxonomy has been agreed and those used or proposed have been mainly domain or task specific. One particular act that has become important for researchers is the expression of Opinion.

Work in Opinion and Sentiment analysis was found to have moved on from simple detection of expressive verbs through dealing with nuance, negation, and strength, and then to finding the objects and aspects thereof that the expressed opinion was targeted at. Techniques including parsing, supervised and unsupervised learning approaches have been used in such areas as product and service reviews as well as in discussion forums, and political debate.

Whether fact or opinion, a final step in converting from bodies of text into a formal representation, is the extraction of the proposition. A predicate-argument form has often been used. While early relationship extraction focused on particular relationships for predicate anchors, later Open Information Extraction has looked to relax this restriction. Again tradeoffs have been found between shallow and deeper processing techniques.

Following the review of previous work, Chapter 4 outlined the approaches chosen for exploration as potential methods for the discovery and filtering stages in the envisaged knowledge discovery system. Statistical modelling of document features was selected as the most appropriate approach for information discovery. A rolling time slicing approach was presented where feature frequencies across documents appearing in a current window are compared with their expected frequencies as calculated by averaging over previous time windows. A window duration of 1 day was chosen.

The ICWSM 2009 dataset (Burton, Java, and Soboroff, 2009) was identified as a suitable corpus for experiments, containing timestamped blog posts and news articles covering 2 months of online publishing.

The question of the approach to take for filtering was then addressed. It was argued that an explicitly informative utterance would be classed as a Statement, or equivalent, in many Dialogue Act schemas. A simple set of five Dialogue Acts was defined, comprising Informative Statement, Comment, Advert, Question and Unknown/Non-Informative, the first of these being "explicitly informative" text. A potential limiting factor in the approach was observed in that opinions and beliefs can be expressed in different ways; separation of fact from belief without external knowledge may not be possible.

Supervised machine learning was identified as the most appropriate method to use to build a Dialogue Act classification model with which to carry out filtering.

Four popular methods – Maximum Entropy, Naive Bayes, Decision Trees and Support Vector Machines – were selected for use in experiments.

It was noted that it was necessary to create a new corpus of data annotated with the selected Dialogue Acts. A sample of the Redites Twitter corpus (Osborne et al., 2014) was chosen for this purpose, as it contained short, utterance-like messages with assessments for News content.

The chapter went on to discuss the features selected for use in the approaches explored. It was noted that information, and in particular new information, about things, is the object of interest for the discovery system. Therefore, nouns, named entities and multiword expressions, were selected as key features of interest. For the filtering stage it was argued that a wider set of more generic feature types would be required to capture Dialogue Acts. As well as simple counts of these feature types, counts of other token types such as personal pronouns and monetary units, were selected, along with features based on inverse document frequency.

Chapter 5 examined the features chosen to be used in the envisaged system, and the tools selected for their extraction. These features, described previously in Chapter 4, have a varying range of complexity for identification or computation. They included: words, word (and phrase chunk) inverse document frequencies, nouns, pronouns, four types of named entities, dates, sums of money, emoticons, URLs, and User Identifiers.

The classification and discovery algorithms used in the investigation operate upon the selected features extracted from text. Their performance therefore relies on the efficacy of feature extraction. It was decided to use the Stanford CoreNLP toolkit, together with a default model for the basic Named Entity types required, as it contained much of the core desired feature identification. An extended model tailored for microblog messages was also selected from Sheffield's GATE NLP framework for processing Twitter data for POS and micro-blog specific features.

No suitable tool for extracting multiword expressions (MWEs), required for nominal compound features, was readily available at the time of the study. It was therefore decided to create one. Details of the design and development of the tool, called MESME, are given in Appendix A.

Although simple features are readily identified, the systematic identification of non-surface features may rely on models with parameters learnt through training examples. The efficacy of such models relies upon the quality of the data used in training, and therefore on how well humans identify and agree upon the features. Given the focus of this work was on presence of non-lexical features as cues to authors imparting information, the chapter gave some consideration to how well these features were defined and detected by the selected extraction tools. For most features it was clear that they were sufficiently recognised and extracted by the chosen tools. This was less clear in the case of MWEs and the MESME tool. An analysis of

MWEs identified by MESME, summarised in Section 8.3 below, was therefore performed to complement this.

8.3 Experimental work

This section summarises the experiments carried out in support of this thesis, and the findings thereof. It provides a recap of the work done on developing techniques for discovering new information, and then on filtering text for explicitly informative statements. Before doing so, though, it reviews the evaluation carried out with regard to the MWE features and their extraction.

MWE extraction was carried out using the tool MESME, developed specifically for this purpose. MESME predominantly makes use of syntactic information in the form of POS sequence. Unsurprisingly, testing of MESME showed that POS tagging quality is a significant factor in the identification of MWEs. It was shown that longer phrases and common constructions pose significant challenges. It was also found that syntax alone is not sufficient to reliably detect MWEs (even with good quality tagging). However, a reasonable proportion of compound nouns, the principal target MWE feature type, could be extracted from Social Media. Test data was successfully collected by sampling a corpus of Twitter message for known MWEs. MESME showed useful performance in compound noun detection in Social Media, suffering only a small degradation (3% in recall) when compared with that on a Wikipedia text sample. However, obtaining high rates of identification for verb particle constructs (VPCs) and particularly light verb constructs (LVCs) was found to be challenging.

Although alternative tools, subsequently developed in the field, have been demonstrated to have better performance, MESME showed the capability to reliably extract the majority of nominal compounds sampled in Twitter messages with an acceptable level of false positives.

The chapter's examination of features concluded with a study on the boundary between what people consider to be MWE and non-MWE constructs. This study used the spurious MESME extractions from Wikipedia text. A wide degree of interpretation of what constitutes a MWE was found, a 0.24 Fleiss's κ showing only slight to fair agreement amongst annotator assessment of over 1,400 'spurious' candidate expressions. 'Accurate' identification of MWEs, at least for the application of knowledge discovery, might not be a significant issue, therefore, because whether or not a phrasal reference should be considered a MWE would be irrelevant providing the phrase was indeed a reference to something.

Having established that features and tools for their extraction were sufficiently reliable for use, investigations turned to examining their presence and distribution in Social Media and News dissemination.

Chapter 6 examined the proposed method for the envisaged system's discovery stage, in which features predictive of interesting new information, and posts containing it, are sought. The idea here was that trends in feature occurrences might be useable as predictors of newsworthy stories. Analysis was performed using the ICWSM 2009 dataset which includes two months of internet postings categorised as either mainstream news or blogs. Within this corpus, the volume of articles, size of vocabulary, and range of topics covered in Social Media was found, unsurprisingly, to be much greater than that found in News. The corpus was therefore ideal for testing the proposed feature frequency-based discovery method.

A Poisson-based model for tracking feature frequency was adopted as the basis for trend detection. The selected features, extracted from the corpus documents using the tools described in Chapter 5, were: nouns, four classes of named entities (Persons, Locations, Organisations, and other Miscellaneous) and three types of multiword expressions (compound nous, verb particles and light verb constructs). Frequencies were calculated and updated on a daily basis using the rolling window time-slicing method described in Chapter 4. Those features demonstrating significant positive deviation in daily occurrence from that expected were selected as Trending. Trending features were found that appeared to be in common between the two types of postings, suggesting there could be links.

Focus was turned to the trends originating in Social Media and an examination of their relative strength. Although maximum trend strengths shown for nouns were found to be considerably greater than those shown for named entities, the latter were marginally more frequent in social media originated trends. Higher trend strengths were observed in Social Media for those features that were also seen in, and particularly later trended, in news articles. (Trend strengths were measured relative to the distribution of trend strengths observed for the feature type.) Although some noun and named entity mention trends originating in Social Media were found to subsequently trend in News, this was not found to be the case for multiword expressions. Therefore multiword expressions were not considered any further for information discovery. (The observation also cast doubt on whether subsequent noun trends in News were related to the trends found in blogs because one might expect compound nouns to be more specific than singular nouns in general.)

An examination of the distributions in trend strength and trending feature frequency for those trending features unique to Social Media, those found to also occur in News, and those subsequently trending in News, was carried out in order to determine potential trend selection strategies. Absolute trend strength was shown to be poor for predicting features likely to trend later in News. Normalisation by average trend strength across each feature type, or by application of feature type thresholds, was found to be a strategy more likely to select features trending in both classes. It was shown that these Social Media originating trend features were more likely to be named entity mentions than common nouns.

Having established feature types that seemingly displayed trends in both Social Media and News, and a trend ranking/selection strategy, attention turned to whether or not News articles could be predicted from Social Media; could one find the small proportion of News stories that don't originate in the mainstream media before they appear there?

An experiment was described (Section 6.6.2) in which a 7-day rolling window to filter out features trending in the News was used. In this experiment, arbitrary minimum feature frequency and minimum trend strength thresholds (5 occurrences and 10 std. deviations) were applied and Social Media originating trends were successfully selected. It was found that on average approximately 12% of trending features (18.4% of tagged nouns, 11.1% of tagged entities) subsequently trended in news stories. Assuming 3% of news stories start in social media, as found by Lloyd, Kaulgud, and Skiena, (2006), this suggests that either the majority of significant social media originated trends are not sufficiently interesting to professional news organisations, or are missed. However, inspection of the top 50 selected Social Media trends showed that about 30% predicted subsequent topically connected news articles. This suggests that trends originating in Social Media do have some predictive power.

A trending feature does not necessarily constitute the topic of the posts giving rise to it. Given that topic models may be constructed from words that appear together in topically related documents, it was decided to examine the co-occurrence of trending features in source documents as a basis of refining topics. Using an implicit measure of topic coherence, based on normalised point-wise mutual information (NPMI), it was shown that bigrams of trends were more likely to contain named entity mentions than nouns, providing evidence in support of:

*H*3: Documents imparting new information are more likely to contain unusual combinations of named entity mentions than unusual combinations of nouns.

Examining NPMI use further, through application to selected feature types irrespective of trend behaviours, it was found that co-mentions of different named entities yielded higher NPMI values than co-occurrences of different nouns. The difference was found to be strongest in news stories (which one might expect to be more likely to be topically focussed than Social Media posts). These findings provided evidence in support of:

*H*4: Mentions of different named entities are less likely to co-occur independently than mentions of different common nouns .

Trends in references, particularly through named entity mentions, could be used as features in information discovery with co-occurrences thereof being useful in refining topically related information. However, only a small percentage of that originating in Social Media was found to constitute a subsequent News story (such stories themselves a very small percentage of the News). Although the techniques explored showed potential for refining possible source documents they did not constitute a

complete approach for information discovery. Furthermore, much of the text processed did not give rise to any new information and co-occurrence counts are costly to compute. Identifying and removing text unlikely to carry information could help reduce unnecessary processing. The question of whether or not uninformative text could be removed in a filtering stage was addressed in the following chapter.

Chapter 7 presented the proposed approach for filtering in the envisaged information discovery system. The method uses supervised machine learning to create discriminative models with which to separate text that is explicitly informative from that which is not. Two main classification schemes for this separation were described. The first distinguished News and Non-News from one another. It was selected because News messages were assumed to be predominantly explicitly informative. The second scheme was a set of five Dialogue Acts including Informative, Opinion or Comment, Question, Advert, and Non-Informative. The Informative class was defined to capture explicitly informative statements.

The studies presented used Twitter microblog posts, called Tweets, as a source of short messages. The reasons for this were that each message could be assumed to perform a single Dialogue Act, and that there existed corpora where whether or not messages contained News had already been assessed. Section 5.3.1 described one such corpus, known as Redites (Osborne et al., 2014). A sample of this corpus was further annotated for five Dialogue Acts, using the simple scheme described in Section 7.2. This exercise, described in Section 7.3.2, showed the scheme to be sufficiently stable with substantial agreement between annotation judges, yielding a Fleiss's κ score of 0.67.

The machine learning techniques selected for the experiments described were Naive Bayes, Maximum Entropy, Decision Trees, and Linear Support Vector Machines. These operate on vectors of feature values derived from the data to be put into classes. The features selected for the classification experiments included the traditional "bag-of-words" features often used in text classification applications, plus a selection of non-lexical features. Non-lexical features included: Emoticons, Hashtags, URLs and User Identifiers, (designated GRP); noun phrase count, verb phrase count, and pronoun count (designated SYN); person count, location count, organisation count, (designated NE); and total and average Inverse Document Frequency scores (designated FRQ).

Three main sets of experiments were carried out. The first of these aimed to examine the efficacy of the approach in discriminating News Tweets from Non-News Tweets. The second set of experiments examined classification of Tweets by the chosen Dialogue Act scheme, and the third set of experiments sought to compare the two.

In separating News Tweets from Non-News Tweets, as reported in Section 7.5, it was found that unigram word feature models achieved accuracies between 70.1%

(C4.5) and 77.6% (Maximum Entropy) using 4,000 Tweets in a 10-fold validation experiment. However the best non-lexical feature model, using SVM, achieved an accuracy of 85.0%. The best combined feature models, showing similar accuracy across folds, achieved 90.3% (SVM) and 92.1% (Naive Bayes). Model tests on later posted 572 Tweets showed a reduction in accuracy however. The reduction was an average 20% for unigram models, and 8% for non-lexical models, suggesting that they are more stable than unigram models for News Tweet identification. An examination of the different non-lexical feature types, carried out using a "hold-out" method, showed that message features based on inverse document frequency, and Named Entity type counts were the most useful, while the SYN set of features had little discriminatory power for the data used.

Placed into a representative filtering setting, where Non-News Tweets were significantly more common than News Tweets (at a ratio of 605:1 in the sample), analysis of classifier confidence determined ROC curves showed AUC of 86.1% and 90.9% for unigram and non-lexical feature models (both using SVM), and 94.5% for a combined feature model (using Maximum Entropy). This provided further evidence that non-lexical features carry useful information for identifying Tweets imparting News.

The chapter then turned to systematic identification of Dialogue Acts in Section 7.6, using the Tweets annotated for this purpose. Too few Tweets were annotated as Question or Non-Informative for training, leaving Informative, Advert and Opinion/Comment Tweets. 1280 Tweets, where 5 judges had agreed the class by a least two votes over the second choice, were used in a multi-classifier approach, again using 10-fold cross validation. Explicitly Informative Tweets were the most readily identified class, and non-lexical models performed marginally better than lexical models, with F1 scores of 0.69 and 0.67 respectively. Combined features gave an average classifier F1 score of 0.72. Performance for the other two classes was poor, possible due to the relatively small amounts of example data.

Further experiments, comparing Dialogue Act classification with News classification, considered only the Informative Dialogue Act class - the principal Act of interest. Section 7.7 described these experiments, in which the Informative Tweet sub-corpus was used for testing the News/Non-News discriminative models from the first set of experiments described in the chapter. The goal here was to get a better understanding of the limitations of using News classification for finding Informative messages, and of using Informative message classification for finding statements of News (a proxy for 'interesting' here).

New models were created by training with feature vectors derived from Tweets in the Informative Tweet sub-corpus. These models were tested on discrimination of News Tweets from Non-News Tweets in the original corpus. The analysis showed

that Informative Tweet models were better able to identify News Tweets from Non-News Tweets (i.e. selecting News Tweets as Informative) than News Tweet models were able to identify Informative Tweets from others (i.e. selecting Informative Tweets as News). Average classifier accuracy, taken across the classifier types, were 65% for prediction of Informative Tweet by unigram News model, and 67% for prediction of News Tweet by unigram Informative model. Average classifier accuracy in predicting Informative Tweets using non-lexical feature models trained on news-event Tweets was 64%, and 75% predicting News event Tweets from models trained on Informative Tweets. However results were variable across classifier and feature set combinations. The best model, whether predicting Informative from News-event training, or predicting News from Informative Tweet training, was a Naive Bayes unigram model with an F1 score of 0.76 for each. (C4.5 and SVM with non-lexical features were close in predicting News, C4.5 slightly less accurate but with F1 score of 0.78, and SVM slightly more accurate but with an F1 score of 0.73.)

The Informative Tweet sub-corpus was further annotated for News, as well as topics related to Sport, Travel and Weather. (The latter was carried out for possible insights into any differences in annotator interpretation of whether or not these topics constituted News.) The performance of the News models created from the initial sample of News and Non-News Tweets in the Redites corpus were tested again on the Informative Tweet sub-corpus, but this time for News prediction. Results of the evaluation showed weaker performance than the previous evaluation. Average F1 score was about 0.05 less at 0.50 when using unigram models. However, sampling may have had an effect. The section then went on to consider whether or not there might have been some discrepancy in interpretations of what constituted News between the original News/Non-News sub-corpus annotations and the later Informative Tweet sub-corpus annotations; in particular with respect to the topics of Sport, Travel Information, and Weather Reports. Tweets annotated as one of the selected topics were further analysed with respect to their News category. It was found that Sport related Tweets in the sub-corpus annotated as Non-News accounted for a drop of about 10–13 points in the F1 metric.

It was concluded that not only are there the issues of domain dependency and temporal affects, but also receiver interpretation to consider in deciding whether or not a message imparts news.

All of the Tweets in the Informative Tweet sub-corpus that were consistently given false positive classification as News bearing were annotated as Informative, and between 40% and 49% of the Tweets consistently given false negative classifications were annotated as Informative, indicating that Tweets bearing News are explicitly informative, but that Explicitly Informative Tweets include Non-News items.

Section 7.8 concluded the investigations with a review of recent related work which had found similar results with respect to the usefulness of Named Entity mentions as features in detection of Tweets related to News events.

8.4 Thesis outcomes

Two principal objectives were set out in this thesis: firstly, to investigate whether or not new interesting information could be found in Social Media through an unexpected number of references to entities and concepts; and secondly, to investigate whether or not explicitly informative text could be distinguished text intended to perform other functions using non-lexical features. Two corresponding hypotheses were proposed:

$H1$: Some documents containing new information can be found through an unexpected number of references to named entities and concepts.

$H1_{null}$: References to named entities and concepts are no more frequent when related new information emerges than the average rate of mentions.

$H2$: Sentences asserting information may be distinguished from those intended to perform other functions using non-lexical features.

$H2_{null}$: Non-lexical characteristics of a message carry no information on whether or not that message conveys an explicit statement.

This thesis has demonstrated that surges in daily frequency – trends – of nominal references, particularly mentions of Named Entities, occur in streams of online text. It has proposed a two-stream time-slicing based approach to select trending features originating in Social Media as a basis for information discovery, and provided evidence that such features could predict a corresponding trend in future news stories. It has therefore found contrary evidence to $H1_{null}$, providing supporting evidence for the first hypothesis.

The thesis has provided evidence for the utility of Named Entities in detecting new informative text. It was shown that blogs detected as containing new information, were more likely to be detected through co-occurrence of trending named entity mentions than through co-occurrence of trending nouns. There was also evidence shown suggesting co-occurrence of different named entity mentions to be less independent than co-occurrence of different nouns, and therefore more topically informative. Further evidence for the utility of Named Entity mentions in detecting informative text was found in a sample of Dialogue Act annotated Tweets. In this sample, NEs were found to be slightly more frequent in explicitly informative Tweets, than in non-informative ones, particularly with Location entities.

It has been shown that models to distinguish explicitly informative text messages are possible through the application of machine learning techniques on example Twitter messages. Through comparison of a News v. Non-News model and an Informative v. Non-Informative model, evidence was found that News Tweets were also Informative, but Informative Tweets may also provide Non-News information.

All the classification experiments were controlled for topic. It was shown that non-lexical feature based models were generally more accurate and more robust than lexical models for both classification schemes, particularly for identifying Informative Tweets, while a combination of all the features examined gave the best performance. It has therefore been shown that non-lexical characteristics of a message can carry information as to whether or not that message is explicitly informative, contrary to $H2_{null}$, in support of the second principal hypothesis.

In addition to the studies reported herein, this work has also contributed a corpus of Twitter data annotated for Dialogue Act, and for News content. It has also proposed an annotation scheme for finer analysis, characterising the subject of a message in relation to the message's author.

In developing software for the experiments carried out, a multiword extraction tool called MESME was developed and made publicly available. A corpus of Tweets containing examples of compound nouns, verb particles, light verb constructs and idioms has also been provided.

8.5 Potential Future Directions

The knowledge discovery system envisaged in this thesis is one that would be comprised of many elements. The work presented here has contributed to the development of two separate stages within the system. Chapter 6 presented an investigation into a proposed method for discovering text containing assertions of new interesting information, and Chapter 7 presented an investigation into how Social Media text might be filtered such that only text making explicitly informative statements is presented to the discovery process. The natural next step would be to examine the efficacy of using the two stages together.

It was shown that finding significant increases in daily mentions of things, particularly by name, in Social Media could predict a similar corresponding later increase in mentions in news media (assumed here as a proxy for known interesting information). However, there are many trends in Social Media that do not, and processing is expensive. Would the application of the proposed explicitly informative message classifier as a text filter prior to the calculation of feature occurrence for trend analysis significantly reduce the amount of processing, and what effect would it have on the trends detected? Presuming the number of features seen trending would decrease, what would be the effect on the density of predictive features found? It was not possible to answer these questions in this study because the time available did not permit the necessary processing to be carried out. It would, however, be a logical next step to combine the two approaches reported here in any further development of the envisaged Social Media knowledge discovery system.

Two classification schemes were developed. One for five Dialogue Acts, and one – "Subjective Focus" – for characterising the relationship between the subject of a

message and its author. However, insufficient examples of some of the classes were found in the data annotated for meaningful classification experiments to be carried out. Further annotation would rectify this and permit an investigation into whether these schemes could provide useful distinctions in text characterisation or not.

Disappointingly, little utility was found for multiword expressions as features in the methods adopted for information discovery and informative statement filtering. However, the tool developed for their extraction did not demonstrate a particularly high rate of detection. The use of multiword expressions in text imparting information could be further studied through use of better performing extraction tools. It would also be possible to further develop the MESME tool.

Appendix A

MESME: A Multiword Extraction Tool

This appendix describes the design of the multiword extraction tool developed in order to assist in extraction of the features required for experiments done to investigate the theses herein. The tool, called MESME, is publicly available at https://github.com/NDewdney/extraction and may be further developed.

A description of the main classes of multiword expressions as found in English, and a review of work related to methods to identify and extract them, are given in Sections A.1 and A.2. It was decided to take a two stage approach in MESME: the first stage taking a rule-based shallow parsing approach encoded in state machines to detect potential MWE token sequences; the second stage applying a filter model developed through machine learning to filter out non MWEs. These two stages are described in detail in Section A.3 Annotated data was required for training and evaluation. Section A.3.1 describes the corpora used.

A.1 The need for identification of Multi-word expressions

Languages such as English allow a vast range of ideas and concepts to be expressed. However, the totality of concepts far exceeds the number of individual words. English does not have a name for everything in the world we may wish to reference or talk about. Whether such concepts are of physical objects (concrete) or abstract ideas, multiple words are required to make reference to them. For example, a "town hall" is a compound noun for a physical object, i.e. a particular type of building used for civic purposes. A "busman's holiday" is an abstract example, being a metaphorical description of a vacation that is strongly related to the profession of the individual(s) partaking in it. As with mentions of Named Entities, the identification of multi-word expressions within sentences is non-trivial. However, the capability to identify these expressions, particularly compound noun phrases, would be helpful in systematic analysis of references to different concepts made in a sentence. This section considers multi-word expressions (MWE) and describes an approach to their extraction from informal English sentences.

Used to convey concepts that have some form of singular semantics or pragmatics as a whole, MWEs often have a meaning that is different from that which might be thought given the meanings of the constituent words. They may also be idiosyncratic to the language. They can, therefore, pose problems for language learners (De Cock et al., 2000) and can lead to mis-translation if not properly recognised. Similarly they pose a challenge for natural language processing systems. As a result, research into MWEs and their identification has seen an increasing amount of research in recent years. There has also been an increasing interest in moving towards processing at the conceptual level in language modelling as the limits of so called "bag-of-words" models are being reached.

MWEs cover a wide range of syntactic constructions, from relatively simple constructs such as compound nouns (e.g. "tomato paste" and "conference paper") to more complex idiomatic and metaphoric phrases (e.g. "raining cats and dogs" and "blowing in the wind"). The approach described here focuses on three types of multi-word expressions: compound noun constructions, verb particle constructs (VPCs), such as a "break up", and light verb constructions (LVCs) where the noun determines the meaning of the verb, as in "taking a photograph" [1]. One of the challenges in multiword expression detection is that not every instance of the constituent words will be instance of the expression. For example, "taking a photograph to the editor" does not contain the example LVC, whereas "taking a photograph of the editor" does.

Some expressions have metaphorical or idiomatic meanings that are quite distant from the semantics of the composite words. For example, "barking up the wrong tree" has the meaning of looking in the wrong place. Typically such phrases and expressions are culturally dependent. Idiomatic MWEs are not sought as an explicit class here but may be included when they also conform to the VPC, LVC, and compound noun grammatical patterns.

With the rise of social media as a source for information, natural language processing tools increasingly need to be more robust to grammatically inconsistent language. Applications such as sentiment analysis have a need to identify aspects of targets such as products coupled with expressions of sentiment. For example "The *power winder* on this camera will *take your breath away*!" contains a sentiment expressed as a metaphor towards a compound noun object aspect.

Although methods and tools for MWE identification and extraction are now beginning to emerge, they were not available at the time of the studies reported herein. It was therefore necessary to create a suitable MWE extraction tool, with particular emphasis on concepts expressed in noun phrases and verbal forms, to enable their inclusion as features in the identification of explicit statements. This tool has been designed for contiguous[2] MWE extraction from Social Media in English (MESME)

[1] Note that some VPCs may also be classified as LVCs.

[2] Non contiguous expressions are typically those that can take some arbitrary argument within, such as in "bringing [X] to a conclusion".

and other similarly informal discourse. The approaches taken towards achieving the principal objectives in this theses focus on references to things in the world because one would expect new interesting information to be about something mentioned. Of particular importance for the experiments into detection of explicit statements, therefore, was the ability to extract references to concepts expressed as compound nouns, although those expressed in verbal constructs, e.g. "shooting a film" are also of interest.

A.2 Related Work

The identification of MWEs is a challenging but important task not only for NLP systems but also for assisting language learners. There have therefore been efforts in recent years to collect and analyse MWE lexicons, e.g. Brooke et al., (2015), Farahmand, Smith, and Nivre, (2015), Vincze, Nagy, and Berend, (2011). Idiomatic MWEs are similarly challenging for machine translation. Correspondingly there have been and continue to be efforts to identify idiomatic paraphrases and comparable translations, e.g. Pershina, He, and Grishman, (2015).

Various methods have been proposed for identifying potential MWEs. Early work focussed on statistical techniques to find significant co-occurrences, based on pairwise association measures, e.g. Church and Hanks, (1990), Smadja, (1993). Schone and Jurafsky 2001 evaluated many of these along with a Latent Semantic Analysis based method for recomputing an association score for semantically related expressions. They find significant room for improvement. Observing limitations with scaling corpus based association measure approaches, Watrin and François 2011 have proposed augmenting association measurement techniques with an n-gram frequency database. However, Villavicencio et al. 2007 found that ranking potential MWEs by different collocation metrics had a marked effect on identification performance.

Purely statistical measures may miss many MWEs through low frequency in corpora and common phrasing may not itself constitute a MWE. Researchers have suggested the use of linguistic cues in the form of syntactic, morphological and lexical information to parse out likely MWEs.

Ramisch et al. 2008 showed that the addition of syntactic information in the form of POS sequences significantly improved the identification of English verb-particle constructions and German adjective-noun constructions. Starting from a corpus of known constructs they generate potential variations conforming to a set a syntactic variations. They have shown that an entropy based metric incorporating these variations gives improved precision-recall over entropy of the candidate alone. Sangati and van Cranenburgh 2015 have proposed using recurring tree fragments – kernels derived from a treebank – to identify potential MWEs. Using the hierarchy in parse trees enables some MWEs to be identified where words have been inserted (as in

"taking [it] for granted"). Rajagopal et al. 2013 have proposed a syntactic graph parsing approach to extract compound concepts.

With the respective limitations of statistical methods and linguistic models, researchers have employed machine learning to train MWE classifiers.

Tsvetkov and Wintner 2014 used a Bayesian network to combine multiple sources of linguistic information to identify candidate MWEs consisting of two words. They generate a number of linguistic features for each MWE such as word inflection variation, capitalisation, contextual variation, and syntax. They have shown a handcrafted, linguistically motivated, Bayesian network to outperform a PMI baseline, using an automatically derived Bayesian network and a discriminative SVM model on data generated from English-French and English-Hebrew bitexts. Dubremetz and Nivre 2014 have applied supervised machine learning techniques to nominal MWE extraction in French, and Gayen and Sarkar 2014 have used a Random Forest supervised classifier for identifying Bengali compound nouns.

Recently hybrid approaches have been proposed whereby linguistic information is used in combination with machine learning techniques to discriminate MWEs from token sequences that have similar grammatical structure.

Nagy and Vincze 2014 examined syntactic parsing for VPC identification within a two-stage tagger. Their method uses a dependency parser to select candidate verb particle and verb prepositional sequences followed by a classifier trained on the syntactic information along with orthographic, lexical, and semantic features.

Rondon et al. 2015 have proposed a continuous learning approach for extraction of MWEs. In their system, NEMWEL, they extract candidate token sequences based upon POS tags and lemmas matching regular expressions which have been determined from a tagged corpus. Following this they apply a classifier to promote "true MWEs". The novelty is that periodically these expressions are used to retrain the classifier.

A.3 Extractor Design

MESME takes a modular approach to identifying potential multi-word expressions, following the hybrid paradigm. It first executes a forward parse of the text at the syntactic level. This is a similar approach to that taken by Poria et al., (2014) and Rajagopal et al., (2013) to extract concepts in sentiment analysis. It applies a series of dependency rules via state machines to the part-of-speech (POS) tag sequence determined for the input text. This POS tag sequence may be supplied by any English tagger conforming to the Penn Treebank tag-set[3]. However, the performance of early stages in a language processing pipeline can have a marked impact on the overall

[3]In principle the technique could be applied to other languages although it is expected that MWEs of other languages would be covered by different syntactic patterns and therefore require different states.

performance Derczynski et al., (2013). As the source material for MESME is intended to be informal language, use of appropriately "tuned" tokenisation and POS tagging tools is likely to be important. We used the TwitIE system Bontcheva et al., (2013) as this was designed to work on English found in micro-blog social media platforms such as Twitter.

There are two state machines within MESME, one to identify candidate noun constructs, and one to identify verbal constructs.

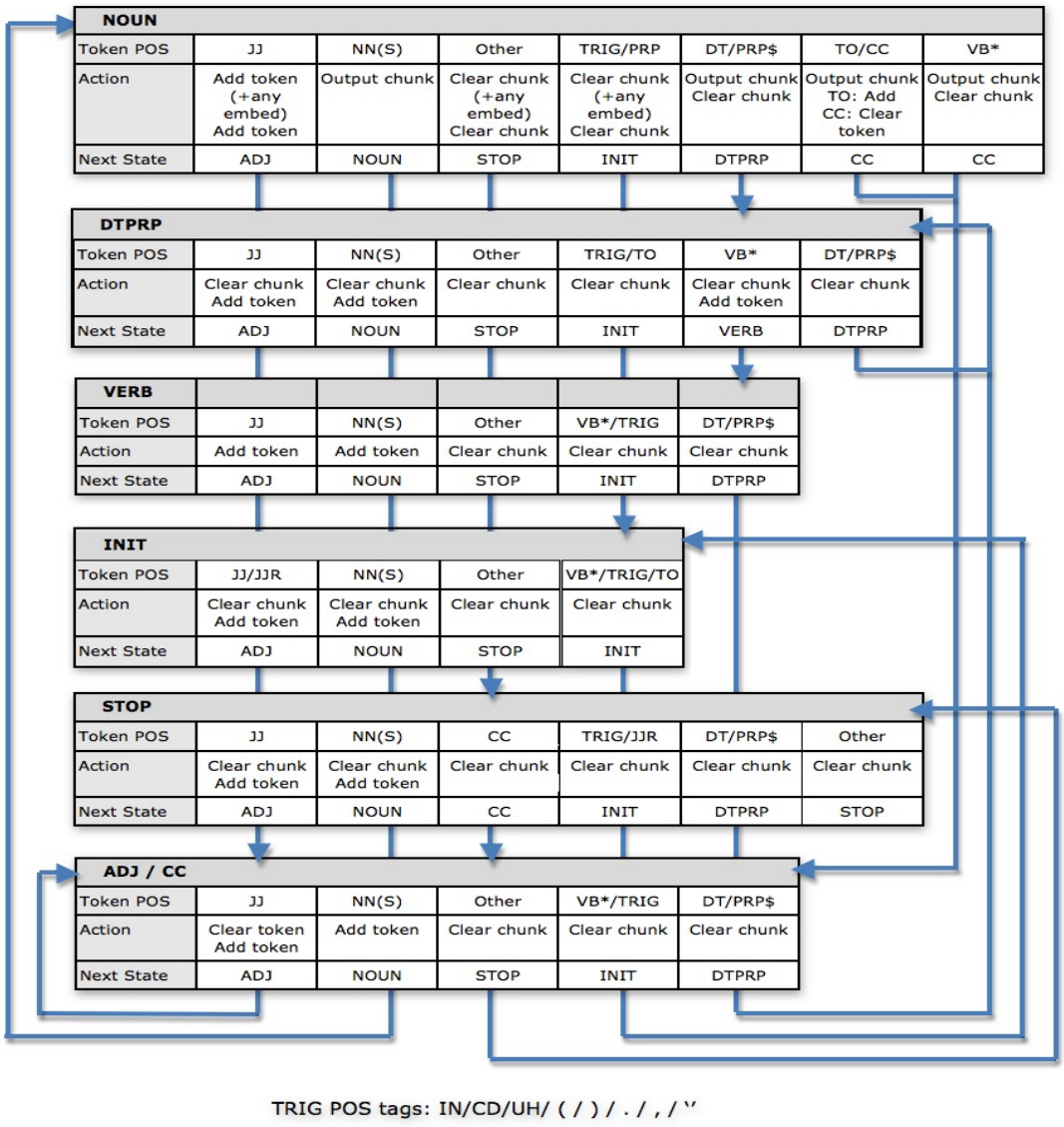


FIGURE A.1: Compound Noun identification parse states. A POS tag is read on transition between states. Within a state an action is taken and the next state is indicated.

The state machine flow for chunking compound nouns is illustrated in Figure A.1. That for VPC and LVC chunking is more complex and nests some conditions owing to the higher number of POS tag sequence combinations one may observe. It is illustrated in Figure A.2. The state machines start in the STOP state, and carry out actions according to the first POS tag observed. The state machines then proceed

to the next state as indicated by tag type. On entering a state, the next POS tag for the input text is observed and the process continues in the same fashion until all the text has been parsed. For example the text "in a buffet car," would have the POS sequence " IN DT NN NN , " which would yield the state sequence "INIT INIT NOUN NOUN STOP" and accumulate a chunk of " buffet car".

STATE	POS TAG	ACTIONS	NEXT STATE
STOP	VB*	Clear Chunk, Add token	VERB
	**	Clear Chunk	STOP
VERB	NN(S)/JJ*/CD	Add token	NOUN
	PRP	Add token	VB-PRP
	PRP$	Add token	VB-PRP$
	DT	Add token	VB-DT
	RB	Add token	RP
	RP	Add token	RB
	TO	Add token	TO
	IN	Add token	IN
	VB*	Clear Chunk, Add token	VERB
	**	Clear Chunk	STOP
VB-PRP	RB/RP	Add token, Output VPC, Clear Chunk	STOP
	TO	Output VPC, Clear Chunk	STOP
	VB*	Clear Chunk, Add token	VERB
	**	Clear Chunk	STOP
VB-PRP$	NN(S)/JJ	Add token	NOUN
	VB*	Clear Chunk, Add token	VERB
	**	Clear Chunk	STOP
VB-DT	NN(S)/JJ*	Add token	NOUN
	VB*	Clear Chunk, Add token	VERB
	**	Clear Chunk	STOP
DT	NN(S)	Add token	NOUN
	CC/,/.	Output VPC, Clear Chunk	STOP
	VB*	Clear Chunk, Add token	VERB
	**	Clear Chunk	STOP
PRP	NN(S)	Add token	NOUN
	CC/,/.	Output VPC, Clear Chunk	STOP
	VB*	Clear Chunk, Add token	VERB
	**	Clear Chunk	STOP

STATE	POS TAG	ACTIONS	NEXT STATE
TO	TO/CC/,/."	Output VPC, Clear Chunk	STOP
	DT	Add token	DT
	PRP	Output VPC, Add token	PRP
	VB*	Clear Chunk, Add token	VERB
	**	Clear Chunk	STOP
IN	NN*/TO/CC/IN JJR/RB/RP/,/."	Output VPC, Clear Chunk	STOP
	DT	Output VPC Add token	DT
	PRP/PRP$	Skip	PRP
	VB*	Clear Chunk, Add token	VERB
	**	Clear Chunk	STOP
RB	NN*/TO/CC/ IN/UH/,/."	Output VPC, Clear Chunk	STOP
	DT	Add token	DT
	PRP/PRP$	Add token	PRP
	VB*	Clear Chunk, Add token	VERB
	**	Clear Chunk	STOP
RP	NN*/TO/CC/ IN/UH/RP/,/."	Output VPC, Clear Chunk	STOP
	DT	Add token	DT
	PRP/PRP$	Output VPC, Add token	PRP
	VB*	Clear Chunk, Add token	VERB
	**	Clear Chunk	STOP
NOUN	NN(S)	Add token	NOUN
	TO/IN/CC/,/(/)	Sub-state test: If True then Output LVC, Clear chunk	STOP
	VB*	Clear Chunk, Add token	VERB
	**	Clear Chunk	STOP

NOUN-SUBSTATE: Test Tag before Chunk	RESULT
NN*/VB*/TO/RB/RP/PRP/PRP$/CC/,/.	True
**	False

FIGURE A.2: VPC and LVC identification parse states. A POS tag is read on transition between states. Within a state an action is taken and the next state is indicated.

The second step is to identify which of the candidate MWEs found by their syntactic patterns are actual MWEs and which are just part of a larger phrase or sentence, without a particular meaning in isolation. There are a very small set of hand-crafted lexical filters placed on the output of the first stage. These are derived from the parser limitations. LVCs are characterised by selection of the sense of the main constituent verb but POS tagging does not distinguish verbs which have multiple senses from those which do not. For example the LVC parser cannot distinguish the verb "to be" from other verbs in the POS tag sequence and so may incorrectly identify

patterns that are objective in a subject-object clause as LVCs. Patterns starting with various forms of 'being' are therefore filtered out. This also seems to apply for VPCs so the same filter is applied. In addition however those candidate VPCs ending with "this" or "that" are removed as these determiners are not commonly found to be part of a VPC. (Even though there may be, arguably, a small number of exceptions such as the exclamation "take that!")

To improve performance this step, additional filtering may be achieved via a model learnt through the application of machine learning to annotated data. MESME employs a linear support vector machine using the LibSVM Chang and Lin, (2011) implementation. Features used in the model include POS tags of the candidate MWE, the two preceding tags, the two successive tags, and the orthographic class of each token, as well as the stemmed token itself if it is a verb, adverb or particle (also prepositional "TO").

The MESME filters were created by training SVM models using the feature vectors associated with the MWEs extracted from the wiki50 dataset using the first stage state machines. The numbers of true and false examples in the training data were balanced between those that were annotated as MWEs in the wiki50 corpus and those that had no corresponding annotation (i.e. spurious false positives). Models were created for the appropriate POS tagger

MESME is encoded as a java program. The source code is freely available from https://github.com/NDewdney/extraction

A.3.1 MWE Data

Two corpora were used for development and evaluation of MESME. The first, the wiki50 corpus (Vincze, Nagy, and Berend, 2011), provides a "gold standard" mark-up of MWEs. It comprises 50 pages taken from wikipedia which have been marked-up for compound noun expressions, VPCs and LVCs. This corpus was selected to enable assessment of the efficacy of the extractor with relatively grammatically correct English. The second corpus used comprised a sample of Twitter microblog messages collected for the Redites project (Osborne et al., 2014). This sample corpus was created to examine MWE detection performance in less formal text, as found in microblogs.

The microblog data set was created by searching the Redites corpus for examples of known MWEs. This was done by firstly taking the 1,042 compound noun MWE dataset created by Farahmand, Smith, and Nivre, (2015) and selecting Twitter messages that contained any of these expressions from the 37.5 million English Tweets in the Redites corpus. Each expression found is represented by just one Tweet (selected by first occurrence). This yielded 843 Tweets, 199 of the MWEs not having been used in any of the source messages. The expressions in this corpus are compound noun

constructions but have been further classed as to whether or not they are compositional in their construction and if they have been "conventionalized". "Board game" for example is considered to be compositional but to have entered into the English language as a conventional concept, whereas"action figure" is not compositional and thereby conventionalized. "Action movie", in contrast, is considered compositional but still conventionalized.

A second selection of Tweets for LVCs was created by the same technique, this time using a collection of verb constructs created by Vincze, Nagy T, and Zsibrita, (2013). 827 different LVCs were found in the Twitter data. 10 examples of each were selected at random rather than one owing to the expectation that these would be harder to detect and variation in context might be a factor. No independent list of VPCs in a readily parsable data format was available at the time of this experiment, so a final selection of Tweets including likely VPCs was collected in the same way using the small number of VPCs identified in the wiki50 corpus. Given the small number, 10 Tweets were selected where possible for each VPC. This yielded 310 Tweets.

Although idioms are not a specific target for the MESME extractor, for completeness Tweets that contained one of the idioms listed under the wikipedia page for "English-language idioms" were selected from the Redites corpus. This selection only yielded a total of 8139 Tweets making use of 86 idioms however, 52 of which occurred in 10 or more Tweets. After removing duplicate content this left 553 Tweets.

The combination of the results from the four MWE type sampling yielded a corpus, herein referred to as 'Tweet-4-MWE', of 9,976 Tweets. Table A.1 provides a summary.

MWE Type	**Unique MWEs**	**Example Tweets**
Compound Nouns	843	843
VPCs	31	310
LVCs	827	8270
Idioms	86	553
Total	1787	9976

TABLE A.1: MWE counts in the Tweet-4-MWE microblog corpus

Appendix B

Twitter Annotation Analysis

Chapter 7 reported on experiments in filtering Twitter for News and explicitly informative messages. Section 7.7 examined differences between News and Explicit Information classification, for which a new sub-corpus was developed, with annotation for a small number of dialogue acts including Inform (explicitly). Each Tweet received five independent assessments. The Tweets were also subsequently annotated by the author for whether they bore News or not.

The guidelines used for News annotation were that Tweets should be annotated as News if they gave information about events or entities in the world other than in relation to the author of the Tweet (unless they author was an organisation an it was not advertising or otherwise promoting itself.) Sporting results were to be annotated as News, but not commentary. Reports of travel conditions were also to be considered News. All other Tweets, including expressions of opinions about news events, were to be annotated as Non-News.

News identification models were evaluated for News identification on 1570 Tweets from newly annotated corpus that at least 3 assessors agreed upon for dialogue act. The analysis was also carried out with restriction to 1280 Tweets with a Dialogue Act label vote margin of 2 or more, and to 1160 Tweets that at least 4 annotators agreed upon. The results of these runs are shown in Tables B.1 and B.2 respectively.

Considering the possibility that a different interpretation of what constituted a News story may have been applied in annotating the new sub-corpus, a finer grain annotation for Sport, Weather, and Travel related messages was also applied. Annotation for relevance for these topics were simply that Tweets should be annotated for Sport if the content referenced a sporting event or sporting entity, Weather if the Tweet reported weather conditions or forecast weather conditions, and Travel if the Tweet reported conditions in any travel infrastructure.

The consistent errors the News models made, i.e. those Tweets posing the most difficultly for models, were analysed using the full 1570 Tweet sub-corpus. The analysis was also conducted using the those Tweets that achieved higher levels of inter-annotator agreement. The results are shown in Tables B.3 and B.4.

	Recall	Precision	F1	Accuracy
Unigram				
Max.Entropy	81.0%	39.7%	53.2%	52.1%
Naive Bayes	95.1%	42.1%	58.4%	54.3%
C4.5	55.0%	46.1%	50.2%	63.2%
Linear SVM	72.6%	41.4%	52.7%	56.2%
Non-Lexical				
Max.Entropy	95.8%	39.8%	56.2%	49.8%
Naive Bayes	83.1%	35.9%	50.1%	44.4%
C4.5	92.8%	38.4%	54.3%	47.4%
Linear SVM	96.8%	39.3%	55.9%	48.5%
Combined				
Max.Entropy	87.5%	39.0%	54.0%	49.8%
Naive Bayes	97.9%	39.4%	56.2%	48.5%
C4.5	89.3%	39.0%	54.3%	49.4%
Linear SVM	92.3%	40.0%	55.9%	50.9%

TABLE B.1: News event Tweet prediction results for News Tweets clear majority agreed for explicit informativeness

	Recall	Precision	F1	Accuracy
Unigram				
Max.Entropy	81.1%	42.7%	56.0%	53.4%
Naive Bayes	95.0%	45.0%	61.0%	55.8%
C4.5	90.3%	42.6%	57.9%	52.2%
Linear SVM	70.7%	44.2%	54.4%	56.7%
Non-Lexical				
Max.Entropy	95.7%	42.6%	59.0%	51.5%
Naive Bayes	83.0%	38.8%	52.9%	46.1%
C4.5	92.7%	41.5%	57.4%	49.7%
Linear SVM	96.7%	42.3%	58.8%	50.6%
Combined				
Max.Entropy	90.5%	43.1%	58.4%	52.9%
Naive Bayes	97.9%	42.3%	59.1%	50.6%
C4.5	86.1%	42.5%	56.9%	52.5%
Linear SVM	92.2%	42.7%	58.4%	52.1%

TABLE B.2: News event Tweet prediction results for News Tweets 4-way and 5 way agreed for explicit informativeness

There is little to be observed from results of the focussing on those Tweets with less disagreement in their Dialogue Act annotation. Tweets most often classified as News by the models are predominantly Informative. Very few News Tweets are classified as Non News that that were accepted as Informative by annotators. (Unsurprising given finding that almost all News Tweets were annotated as Informative).

Weather related Tweets and Travel related Tweets that were well agreed as Informative did not pose an issue for the News models.

Error Type	Total	Sport	Sport=News	Weather	Travel	Other	Informative
Unigram							
False Positive	223	72 (40.0%)	0 (0.0%)	0 (0.0%)	-	151 (22.5%)	111 (39.1%)
False Negative	13	8 (5.1%)	120 (37.6%)	0 (0.0%)	-	5 (1.9%)	13 (4.0%)
Non-Lexical							
False Positive	513	136 (75.6%)	0 (0.0%)	0 (0.0%)	-	377 (56.1%)	216 (76.1%)
False Negative	7	6 3.(8%)	11 (3.4%)	0 (0.0%)	-	1 (0.4%)	6 (1.8%)
Combined							
False Positive	471	116 (64.4%)	0 (0.0%)	0 (0.0%)	-	355 (52.8%)	215 (75.7%)
False Negative	4	3 (1.9%)	5 (1.6%)	0 (0.0%)	-	1 (0.4%)	4 (1.2%)

TABLE B.3: Number of News/Non-News Tweets, majority margin 2 for Dialogue Act, consistently misclassified across classifier types. Proportion of topic category in error given in parenthesis.

Error Type	Total	Sport	Sport=News	Weather	Travel	Other	Informative
Unigram							
False Positive	342	98 (59.8%)	0 (0.0%)	0 (0.0%)	-	244 (42.2%)	179 (71.0%)
False Negative	7	6 (3.9%)	24 (15.5%)	0 (0.0%)	-	1 (0.6%)	7 (1.7%)
Non-Lexical							
False Positive	446	120 (73.2%) 0 (0.0%)		0 (0.0%)	-	326 (56.4%)	198 (78.6%)
False Negative	6	4 (2.6%)	10 (6.5%)	0 (0.0%)	-	2 (1.3%)	6 (1.4%)
Combined							
False Positive	397	110 (67.1%)	0 (0.0%)	0 (0.0%)	-	287 (49.7%)	186 (73.8%)
False Negative	4	3 (1.9%)	5 (3.2%)	0 (0.0%)	-	1 (0.6%)	4 (1.0%)

TABLE B.4: Number of News/Non-News Tweets, majority margin 3 for Dialogue Act, consistently misclassified across classifier types. Proportion of topic category in error given in parenthesis.

Bibliography

Abe, Shigeo (2005). *Support Vector Machines for Pattern Classification (Advances in Pattern Recognition)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc. ISBN: 1852339292.

Agarwal, Basant, Namita Mittal, Pooja Bansal, and Sonal Garg (2015). "Sentiment analysis using common-sense and context information". In: *Computational intelligence and neuroscience* 2015, p. 30.

Ahmed, Amr, Qirong Ho, Choon Hui Teo, Jacob Eisenstein, Alex Smola, and Eric Xing (2011). "Online inference for the infinite topic-cluster model: Storylines from streaming text". In: *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 101–109.

Ailon, Nir, Ragesh Jaiswal, and Claire Monteleoni (2009). "Streaming k-means approximation". In: *Advances in Neural Information Processing Systems*, pp. 10–18.

Alexandersson, Jan, Bianka Buschbeck-Wolf, Tsutomu Fujinami, Michael Kipp, Stephan Koch, Elisabeth Maier, Norbert Reithinger, Birte Schmitz, and Melanie Siegel (1998). *Dialogue acts in Verbmobil 2*. DFKI Saarbrücken.

Allan, James, Victor Lavrenko, and Hubert Jin (2000). "First story detection in TDT is hard". In: *CIKM '00: Proceedings of the ninth international conference on Information and knowledge management*. McLean, Virginia, United States: ACM, pp. 374–381. ISBN: 1-58113-320-0. DOI: http://doi.acm.org/10.1145/354756.354843.

Alvanaki, Foteini, Michel Sebastian, Krithi Ramamritham, and Gerhard Weikum (2011). "EnBlogue: emergent topic detection in web 2.0 streams". In: *Proceedings of the 2011 international conference on Management of data*. SIGMOD '11. Athens, Greece: ACM, pp. 1271–1274. ISBN: 978-1-4503-0661-4. DOI: http://doi.acm.org/10.1145/1989323.1989473. URL: http://doi.acm.org/10.1145/1989323.1989473.

Andor, Daniel, Chris Alberti, David Weiss, Aliaksei Severyn, Alessandro Presta, Kuzman Ganchev, Slav Petrov, and Michael Collins (2016). "Globally normalized transition-based neural networks". In: *arXiv preprint arXiv:1603.06042*.

Andras, Peter (2002). "The equivalence of support vector machine and regularization neural networks". In: *Neural Processing Letters* 15.2, pp. 97–104.

Annett, Michelle and Grzegorz Kondrak (2008). "A comparison of sentiment analysis techniques: polarizing movie blogs". In: *Proceedings of the Canadian Society for computational studies of intelligence, 21st conference on Advances in artificial intelligence*. Canadian AI'08. Windsor, Canada: Springer-Verlag, pp. 25–35. ISBN: 3-540-68821-8, 978-3-540-68821-1. URL: http://dl.acm.org/citation.cfm?id=1788714.1788717.

Arguello, Jaime and Kyle Shaffer (2015). "Predicting Speech Acts in MOOC Forum Posts." In: *ICWSM*, pp. 2–11.

Asghar, Muhammad Zubair, Ahmad B RahmanUllah, Aurangzeb Khan, Shakeel Ahmad, and Irfan Ullah Nawaz (2014). "Political miner: opinion extraction from user generated political reviews". In: *Sci. Int (Lahore)* 26.1, pp. 385–389.

Asur, Sitaram and Bernardo A. Huberman (2010). "Predicting the Future with Social Media". In: *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*. WI-IAT '10. Washington, DC, USA: IEEE Computer Society, pp. 492–499. ISBN: 978-0-7695-4191-4. DOI: http://dx.doi.org/10.1109/WI-IAT.2010.63. URL: http://dx.doi.org/10.1109/WI-IAT.2010.63.

Asur, Sitaram, Bernardo A. Huberman, Gábor Szabó, and Chunyan Wang (2011). "Trends in Social Media : Persistence and Decay". In: *CoRR* abs/1102.1402.

Azzam, Saliha, Kevin Humphreys, and Robert Gaizauskas (1999). "Using coreference chains for text summarization". In: *CorefApp '99: Proceedings of the Workshop on Coreference and its Applications*. College Park, Maryland: Association for Computational Linguistics, pp. 77–84.

Balasuriya, Dominic, Nicky Ringland, Joel Nothman, Tara Murphy, and James R Curran (2009). "Named entity recognition in wikipedia". In: *Proceedings of the 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources*. Association for Computational Linguistics, pp. 10–18.

Baldridge, Jason (2005). "The opennlp project". In: *URL: http://opennlp. apache. org/index. html,(accessed 2 February 2012)*.

Baldwin, Timothy and Su Nam Kim (2010). "Multiword Expressions." In: *Handbook of natural language processing* 2, pp. 267–292.

Bandari, Roja, Sitaram Asur, and Bernardo A Huberman (2012). "The pulse of news in social media: Forecasting popularity". In: *arXiv preprint arXiv:1202.0332*.

Banko, Michele, Oren Etzioni, and Turing Center (2008). "The Tradeoffs Between Open and Traditional Relation Extraction." In: *ACL*. Vol. 8, pp. 28–36.

Baroni, Marco and Stefan Evert (2007). "Words and Echoes: Assessing and Mitigating the Non-Randomness Problem in Word Frequency Distribution Modeling". In: *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Prague, Czech Republic: Association for Computational Linguistics, pp. 904–911. URL: http://www.aclweb.org/anthology/P07-1114.

Behl, Diksha, Sahil Handa, and Anuja Arora (2014). "A bug mining tool to identify and analyze security bugs using naive bayes and tf-idf". In: *Optimization, Reliabilty, and Information Technology (ICROIT), 2014 International Conference on*. IEEE, pp. 294–299.

Bekhuis, T (2006). *Conceptual biology, hypothesis discovery, and text mining: Swanson's legacy Biomed Digit Libr*.

Benhardus, J. (2010). *Streaming trend detection in twitter*. Tech. rep.

Bharat, Krishna and Monika R. Henzinger (1998). "Improved algorithms for topic distillation in a hyperlinked environment". In: *SIGIR '98: Proceedings of the 21st*

annual international ACM SIGIR conference on Research and development in information retrieval. Melbourne, Australia: ACM, pp. 104–111. ISBN: 1-58113-015-5. DOI: http://doi.acm.org/10.1145/290941.290972.

Bhat, Suma and Richard Sproat (2009). "Knowing the unseen: estimating vocabulary size over unseen samples". In: *ACL-IJCNLP '09: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1*. Suntec, Singapore: Association for Computational Linguistics, pp. 109–117. ISBN: 978-1-932432-45-9.

Bird, Steven (2006). "NLTK: the natural language toolkit". In: *Proceedings of the COLING/ACL on Interactive presentation sessions*. Association for Computational Linguistics, pp. 69–72.

Bishop, Christopher M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc. ISBN: 0387310738.

Bizer, Christian, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann (2009). "DBpedia-A crystallization point for the Web of Data". In: *Web Semantics: science, services and agents on the world wide web* 7.3, pp. 154–165.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan (2003). "Latent dirichlet allocation". In: *J. Mach. Learn. Res.* 3, pp. 993–1022. ISSN: 1532-4435. DOI: http://dx.doi.org/10.1162/jmlr.2003.3.4-5.993.

Bontcheva, Kalina, Leon Derczynski, Adam Funk, Mark A. Greenwood, Diana Maynard, and Niraj Aswani (2013). "TwitIE: An Open-Source Information Extraction Pipeline for Microblog Text". In: *Proceedings of the International Conference on Recent Advances in Natural Language Processing*. Association for Computational Linguistics.

Bothos, Efthimios, Dimitris Apostolou, and Gregoris Mentzas (2010). "Using Social Media to Predict Future Events with Agent-Based Markets". In: *IEEE Intelligent Systems* 25 (6), pp. 50–58. ISSN: 1541-1672. DOI: http://dx.doi.org/10.1109/MIS.2010.152. URL: http://dx.doi.org/10.1109/MIS.2010.152.

Bracewell, David B, Marc T Tomlinson, and Hui Wang (2012). "Identification of Social Acts in Dialogue." In: *COLING*, pp. 375–390.

Brewington, Brian E. and George Cybenko (2000). "How dynamic is the Web?" In: *Proceedings of the 9th international World Wide Web conference on Computer networks : the international journal of computer and telecommunications netowrking*. Amsterdam, The Netherlands: North-Holland Publishing Co., pp. 257–276. DOI: http://dx.doi.org/10.1016/S1389-1286(00)00045-1.

Brewster, C, J Iria, Z Zhang, F Ciravegna, L Guthrie, and Y Wilks (2007). "Dynamic iterative ontology learning". In: *In Recent Advances in Natural Language Processing (RANLP 07*.

Brooke, Julian, Adam Hammond, David Jacob, Vivian Tsang, Graeme Hirst, and Fraser Shein (2015). "Building a Lexicon of Formulaic Language for Language Learners". In: *Proceedings of NAACL-HLT*, pp. 96–104.

Bross, Juergen and Heiko Ehrig (2013). "Automatic construction of domain and aspect specific sentiment lexicons for customer review mining". In: *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*. ACM, pp. 1077–1086.

Brown, Gillian and George Yule (1983). *Discourse Analysis*. Cambridge Textbooks in Linguistics. Cambridge, UK: Cambridge University Press.

Brychcın, Tomáš, Michal Konkol, and Josef Steinberger (2014). "UWB: machine learning approach to aspect-based sentiment analysis". In: Citeseer.

Bunt, Harry, Jan Alexandersson, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Volha Petukhova, Andrei Popescu-Belis, and David R Traum (2012). "ISO 24617-2: A semantically-based standard for dialogue annotation." In: *LREC*. Citeseer, pp. 430–437.

Burton, Kevin, Akshay Java, and Ian Soboroff (2009). "The ICWSM 2009 Spinn3r Dataset". In: *Third Annual Conference on Weblogs and Social Media (ICWSM 2009)*. http://icwsm.org/2009/data/. San Jose, CA: AAAI.

Cambria, Erik, Daniel Olsher, and Dheeraj Rajagopal (2014). "SenticNet 3: a common and common-sense knowledge base for cognition-driven sentiment analysis". In: *Twenty-eighth AAAI conference on artificial intelligence*.

Cameron, Deborah (1996). "Style policy and style politics: a neglected aspect of the language of the news". In: *Media, Culture & Society* 18.2, pp. 315–333. DOI: 10.1177/016344396018002008. URL: https://doi.org/10.1177/016344396018002008.

Carpenter, Tamitha and Emi Fujioka (2011). "The role and identification of dialog acts in online chat". In: *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*.

Cataldi, Mario, Luigi Di Caro, and Claudio Schifanella (2010). "Emerging topic detection on Twitter based on temporal and social terms evaluation". In: *Proceedings of the Tenth International Workshop on Multimedia Data Mining*. MDMKDD '10. Washington, D.C.: ACM, 4:1–4:10. ISBN: 978-1-4503-0220-3. DOI: http://doi.acm.org/10.1145/1814245.1814249. URL: http://doi.acm.org/10.1145/1814245.1814249.

Cha, Meeyoung, Juan Antonio, Navarro Pérez, and Hamed Haddadi (2009). "Flash Floods and Ripples: The Spread of Media Content through the Blogosphere". In: *ICWSM 2009: Proceedings of the 3rd AAAI International Conference on Weblogs and Social Media*. AAAI.

Champion, Colin (2008). "Models for interactions in contingency tables". GCHQ internal tech report B14Inf571.

Chang, Chih-Chung and Chih-Jen Lin (2011). "LIBSVM: a library for support vector machines". In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 2.3, p. 27.

Charikar, Moses S (2002). "Similarity estimation techniques from rounding algorithms". In: *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*. ACM, pp. 380–388.

Chen, Danqi and Christopher D Manning (2014). "A Fast and Accurate Dependency Parser using Neural Networks." In: *EMNLP*, pp. 740–750.

Chesley, Paula, Bruce Vincent, Li Xu, and Rohini K Srihari (2006). "Using verbs and adjectives to automatically classify blog sentiment". In: *Training* 580.263, p. 233.

Chinchor, Nancy and Patricia Robinson (1997). "MUC-7 named entity task definition". In: *Proceedings of the 7th Conference on Message Understanding*. Vol. 29.

Christensen, Janara, Stephen Soderland, Oren Etzioni, et al. (2011). "An analysis of open information extraction based on semantic role labeling". In: *Proceedings of the sixth international conference on Knowledge capture*. ACM, pp. 113–120.

Church, Kenneth W. (2000). "Empirical estimates of adaptation: the chance of two noriegas is closer to p/2 than p2". In: *Proceedings of the 18th conference on Computational linguistics*. Saarbrücken, Germany: Association for Computational Linguistics, pp. 180–186. ISBN: 1-55860-717-X. DOI: http://dx.doi.org/10.3115/990820.990847.

Church, K.W. and P. Hanks (1990). "Word association norms, mutual inform ation and lexicography". In: *Computational Linguistics* 16.1, pp. 22–29.

Cimiano, P, A Pivk, L Schmidt-Thieme, and S Staab (2005). "Learning taxonomic relations from heterogenous sources of evidence". In: *Ontology Learning from Text: Methods, Evaluation and Applications*. IOS Press, pp. 59–73.

Clough, Paul, Robert Gaizauskas, Scott S. L. Piao, and Yorick Wilks (2002). "METER: MEasuring TExt Reuse". In: *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Philadelphia, Pennsylvania: Association for Computational Linguistics, pp. 152–159. DOI: http://dx.doi.org/10.3115/1073083.1073110.

Core, Mark G and James Allen (1997). "Coding dialogs with the DAMSL annotation scheme". In: *AAAI fall symposium on communicative action in humans and machines*. Vol. 56. Boston, MA.

Cui, Anqi, Min Zhang, Yiqun Liu, Shaoping Ma, and Kuo Zhang (2012). "Discover Breaking Events with Popular Hashtags in Twitter". In: *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*. CIKM '12. Maui, Hawaii, USA: ACM, pp. 1794–1798. ISBN: 978-1-4503-1156-4. DOI: 10.1145/2396761.2398519. URL: http://doi.acm.org/10.1145/2396761.2398519.

Cunningham, Hamish, Diana Maynard, Kalina Bontcheva, and Valentin Tablan (2002). "GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications". In: *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*.

Cunningham, Hamish, Diana Maynard, Kalina Bontcheva, Valentin Tablan, Niraj Aswani, Ian Roberts, Genevieve Gorrell, Adam Funk, Angus Roberts, Danica Damljanovic, Thomas Heitz, Mark A. Greenwood, Horacio Saggion, Johann Petrak, Yaoyong Li, and Wim Peters (2011). *Text Processing with GATE (Version 6)*. ISBN: 978-0956599315. URL: http://tinyurl.com/gatebook.

Dalli, Angelo and Yorick Wilks (2006). "Automatic dating of documents and temporal text classification". In: *ARTE '06: Proceedings of the Workshop on Annotating and Reasoning about Time and Events*. Sydney, Australia: Association for Computational Linguistics, pp. 17–22. ISBN: 1-932432-81-7.

Dang, H T (2006). "Overview of DUC 2006". In: *In Proceedings of the Document Understanding Workshop*, pp. 1–10.

Das-Neves, Fernando, Edward A. Fox, and Xiaoyan Yu (2005). "Connecting topics in document collections with stepping stones and pathways". In: *CIKM '05: Proceedings of the 14th ACM international conference on Information and knowledge management*. Bremen, Germany: ACM, pp. 91–98. ISBN: 1-59593-140-6. DOI: http://doi.acm.org/10.1145/1099554.1099573.

Datar, Mayur, Nicole Immorlica, Piotr Indyk, and Vahab S. Mirrokni (2004). "Locality-sensitive hashing scheme based on p-stable distributions". In: *Proceedings of the twentieth annual symposium on Computational geometry*. SCG '04. Brooklyn, New York, USA: ACM, pp. 253–262. ISBN: 1-58113-885-7. DOI: http://doi.acm.org/10.1145/997817.997857. URL: http://doi.acm.org/10.1145/997817.997857.

Dawkins, R (1976). *The Selfish Gene*.

De Cock, Sylvie et al. (2000). "Repetitive phrasal chunkiness and advanced EFL speech and writing". In: *Language and Computers* 33, pp. 51–68.

Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei (2009). "Imagenet: A large-scale hierarchical image database". In: *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, pp. 248–255.

Derczynski, Leon, Diana Maynard, Niraj Aswani, and Kalina Bontcheva (2013). "Microblog-genre noise and impact on semantic annotation accuracy". In: *Proceedings of the 24th ACM Conference on Hypertext and Social Media*. ACM, pp. 21–30.

Dewdney, Nigel (2012). "Named Entity Trends Originating from Social Media". In: *Proceedings of the Workshop on Information Extraction and Entity Analytics on Social Media Data*. Mumbai, India: The COLING 2012 Organizing Committee, pp. 1–16.

– (2015). "Finding Potential News from Trends Originating in the Blogosphere". In: *Research in Computing Science* 90, pp. 251–263.

– (2017a). "MESME: Multi-word Expression Extraction for Social Media in English". to appear.

Dewdney Nigel Cotterill, Rachel (2017b). "Just the Facts: Winnowing Microblogs for Newsworthy Statements using Non-Lexical Features". to appear.

Dickinson, Markus and W Detmar Meurers (2003). "Detecting errors in part-of-speech annotation". In: *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1*. Association for Computational Linguistics, pp. 107–114.

Dlugolinskỳ, Štefan, Marek Ciglan, and Michal Laclavík (2013). "Evaluation of named entity recognition tools on microposts". In: *Intelligent Engineering Systems (INES), 2013 IEEE 17th International Conference on*. IEEE, pp. 197–202.

Doddington, George R, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie Strassel, and Ralph M Weischedel (2004). "The Automatic Content Extraction (ACE) Program-Tasks, Data, and Evaluation." In: *LREC*. Vol. 2, p. 1.

Dojchinovski, Milan and Tomás Kliegr (2013). "Datasets, GATE Evaluation Framework for Benchmarking Wikipedia-Based NER Systems." In: *NLP-DBPEDIA@ ISWC*.

Dredze, Mark, Tim Oates, and Christine Piatko (2010). "We're not in Kansas Anymore: Detecting Domain Changes in Streams". In: *EMNLP '10: Proceedings of the Conference on Empirical Methods in Natural Language Processing*. MIT, Massachusettes USA.

Dubremetz, Marie and Joakim Nivre (2014). "Extraction of nominal multiword expressions in french". In: *EACL 2014*, p. 72.

Edouard, Amosse, Elena Cabrio, Sara Tonelli, and Nhan Le Thanh (2017). "Semantic Linking for Event-Based Classification of Tweets". CICLING 2017 short paper, to appear.

Esuli, Andrea and Fabrizio Sebastiani (2006). "Determining Term Subjectivity and Term Orientation for Opinion Mining." In: *EACL*. Vol. 6, p. 2006.

Fader, Anthony, Stephen Soderland, and Oren Etzioni (2011). "Identifying relations for open information extraction". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 1535–1545.

Farahmand, Meghdad, Aaron Smith, and Joakim Nivre (2015). "A multiword expression data set: Annotating non-compositionality and conventionalization for english noun compounds". In: *Proceedings of NAACL-HLT*, pp. 29–33.

Fazly, Afsaneh and Suzanne Stevenson (2007). "Distinguishing subtypes of multiword expressions using linguistically-motivated statistical measures". In: *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions*. Association for Computational Linguistics, pp. 9–16.

Feng, Ao and James Allan (2007). "Finding and linking incidents in news". In: *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*. Lisbon, Portugal: ACM, pp. 821–830. ISBN: 978-1-59593-803-9. DOI: http://doi.acm.org/10.1145/1321440.1321554.

Finkel, Jenny Rose, Trond Grenager, and Christopher Manning (2005). "Incorporating non-local information into information extraction systems by Gibbs sampling". In: *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. Ann Arbor, Michigan: Association for Computational Linguistics, pp. 363–370. DOI: http://dx.doi.org/10.3115/1219840.1219885. URL: http://dx.doi.org/10.3115/1219840.1219885.

Fortuna, Blaž, Dunja Mladenič, and Marko Grobelnik (2006). "Semi-automatic Construction of Topic Ontologies". In: *Semantics, Web and Mining*. Ed. by Markus Ackermann, Bettina Berendt, Marko Grobelnik, Andreas Hotho, Dunja Mladenic, Giovanni Semeraro, Myra Spiliopoulou, Gerd Stumme, Vojtech Svátek, and Maarten

van Someren. Vol. 4289. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, pp. 121–131. URL: http://dx.doi.org/10.1007/11908678_8.

Franco, Leandro and Hideki Kawai (2010). "News Detection in the Blogosphere: Two Approaches Based on Structure and Content Analysis".

Freund, John E. (1971). *Mathematical Statistics*. 2nd. Englewood Cliffs, New Jersey: Prentice Hall.

Gabrilovich, Evgeniy, Susan Dumais, and Eric Horvitz (2004). "Newsjunkie: providing personalized newsfeeds via analysis of information novelty". In: *WWW '04: Proceedings of the 13th international conference on World Wide Web*. New York, NY, USA: ACM, pp. 482–490. ISBN: 1-58113-844-X. DOI: http://doi.acm.org/10.1145/988672.988738.

Gaizauskas, Robert and Kevin Humphreys (2000). "A Combined IR/NLP Approach to Question Answering Against Large Text Collections". In: *In Proceedings of the 6th Content-Based Multimedia Information Access Conference (RIAO-2000*, pp. 1288–1304.

Gamallo, Pablo and Marcos Garcia (2014). "Citius: A naive-bayes strategy for sentiment analysis on english tweets". In: *Proceedings of SemEval*, pp. 171–175.

Gayen, Vivekananda and Kamal Sarkar (2014). "A Machine Learning Approach for the Identification of Bengali Noun-Noun Compound Multiword Expressions". In: *arXiv preprint arXiv:1401.6567*.

Geng, Liqiang and Howard J. Hamilton (2006). "Interestingness measures for data mining: A survey". In: *ACM Comput. Surv.* 38.3, p. 9. ISSN: 0360-0300. DOI: http://doi.acm.org/10.1145/1132960.1132963.

Gilbert, Eric and Karrie Karahalios (2010). "Widespread Worry and the Stock Market". In: *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*. URL: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/view/1513.

Gimpel, Kevin, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A Smith (2011). "Part-of-speech tagging for twitter: Annotation, features, and experiments". In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*. Association for Computational Linguistics, pp. 42–47.

Glance, Natalie S., Matthew Hurst, and Takashi Tomokiyo (2004). "BlogPulse: Automated trend discovery for weblogs". In: *WWW 2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*. ACM.

Godfrey, John J, Edward C Holliman, and Jane McDaniel (1992). "SWITCHBOARD: Telephone speech corpus for research and development". In: *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*. Vol. 1. IEEE, pp. 517–520.

Grishman, Ralph (2012). "Information Extraction". In: *The Oxford Handbook of Computational Linguistics*. Oxford University Press. ISBN: 9780191743573. DOI: 10.1093/oxfordhb/9780199276349.013.0030.

Guthrie, Louise, Elbert Walker, and Joe Guthrie (1994). "Document classification by machine: theory and practice". In: *Proceedings of the 15th conference on Computational linguistics*. Kyoto, Japan: Association for Computational Linguistics, pp. 1059–1063. DOI: http://dx.doi.org/10.3115/991250.991322.

Ha-Thuc, Viet and Padmini Srinivasan (2008). "Topic models and a revisit of text-related applications". In: *PIKM'08: Proceedings of the 2nd PhD workshop on Information and knowledge management*. Napa Valley, California, USA: ACM, pp. 25–32. ISBN: 978-1-60558-257-3. DOI: http://doi.acm.org/10.1145/1458550.1458556.

Hachey, Ben and Miles Osborne, eds. (2010). *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media*. Los Angeles, California, USA: Association for Computational Linguistics. URL: http://www.aclweb.org/anthology/W/W10/W10-05.

Haight, F.A. (1967). *Handbook of the Poisson distribution*. Publications in operations research. Wiley. URL: https://books.google.co.uk/books?id=l8Y-AAAAIAAJ.

Hasegawa, Takaaki, Satoshi Sekine, and Ralph Grishman (2004). "Discovering relations among named entities from large corpora". In: *ACL '04: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*. Barcelona, Spain: Association for Computational Linguistics, p. 415. DOI: http://dx.doi.org/10.3115/1218955.1219008.

Hawkins, John A (2015). *Definiteness and indefiniteness: A study in reference and grammaticality prediction*. Vol. 11. Routledge.

Hinton, Geoffrey E and Ruslan R Salakhutdinov (2006). "Reducing the dimensionality of data with neural networks". In: *Science* 313.5786, pp. 504–507.

Jeong, Minwoo, Chin-Yew Lin, and Gary Geunbae Lee (2009). "Semi-supervised speech act recognition in emails and forums". In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3*. Association for Computational Linguistics, pp. 1250–1259.

Ji, Heng, Ralph Grishman, Hoa Trang Dang, Kira Griffitt, and Joe Ellis (2010). "Overview of the TAC 2010 knowledge base population track". In: *Third Text Analysis Conference (TAC 2010)*. Vol. 3. 2, pp. 3–3.

Jin, Wei, Rohini Srihari, and Abhishek Singh (2008). "Generating hypotheses from the web". In: *WWW '08: Proceeding of the 17th international conference on World Wide Web*. Beijing, China: ACM, pp. 1211–1212. ISBN: 978-1-60558-085-2. DOI: http://doi.acm.org/10.1145/1367497.1367731.

Jo, Yohan and Alice H Oh (2011). "Aspect and sentiment unification model for online review analysis". In: *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM, pp. 815–824.

Johansson, Richard and Pierre Nugues (2008). "Dependency-based syntactic-semantic analysis with PropBank and NomBank". In: *Proceedings of the Twelfth Conference on Computational Natural Language Learning*. Association for Computational Linguistics, pp. 183–187.

Joho, Hideo and Mark Sanderson (2007). "Document frequency and term specificity". In: *Large Scale Semantic Access to Content (Text, Image, Video, and Sound)*. LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE, pp. 350–359.

Joshi, Mahesh, Dipanjan Das, Kevin Gimpel, and Noah A. Smith (2010). "Movie reviews and revenues: an experiment in text regression". In: *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Los Angeles, California: Association for Computational Linguistics, pp. 293–296. ISBN: 1-932432-65-5. URL: http://dl.acm.org/citation.cfm?id=1857999.1858037.

Jurafsky, Daniel and James H. Martin (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Upper Saddle River, NJ, USA: Prentice Hall PTR. ISBN: 0130950696.

Kaji, Nobuhiro and Masaru Kitsuregawa (2007). "Building Lexicon for Sentiment Analysis from Massive Collection of HTML Documents." In: *EMNLP-CoNLL*, pp. 1075–1083.

Kass, Alex and Christopher Cowell-Shah (2006). "Using Lightweight NLP and Semantic Modeling to Realize the Internet's Potential as a Corporate Radar". In: *AAAI Symposium*.

Katz, Slava M. (1987). "Estimation of probabilities from sparse data for the language model component of a speech recognizer". In: *IEEE Transactions on Acoustics, Speech and Signal Processing*, pp. 400–401.

Kim, Soo-Min and Eduard Hovy (2004). "Determining the sentiment of opinions". In: *Proceedings of the 20th international conference on Computational Linguistics*. Association for Computational Linguistics, p. 1367.

Kim, Suin, Jianwen Zhang, Zheng Chen, Alice Oh, and Shixia Liu (2013). "A hierarchical aspect-sentiment model for online reviews". In: *Twenty-Seventh AAAI Conference on Artificial Intelligence*.

Kim, Yoosin, Seung Ryul Jeong, and Imran Ghani (2014). "Text opinion mining to analyze news for stock market prediction". In: *Int. J. Advance. Soft Comput. Appl* 6.1.

Kireyev, Kirill (2009). "Semantic-based estimation of term informativeness". In: *NAACL '09: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Boulder, Colorado: Association for Computational Linguistics, pp. 530–538. ISBN: 978-1-932432-41-1.

Kiritchenko, Svetlana, Xiaodan Zhu, and Saif M Mohammad (2014). "Sentiment analysis of short informal texts". In: *Journal of Artificial Intelligence Research* 50, pp. 723–762.

Kleinberg, Jon (2003). "Bursty and Hierarchical Structure in Streams". In: *Data Min. Knowl. Discov.* 7.4, pp. 373–397. ISSN: 1384-5810. DOI: http://dx.doi.org/10.1023/A:1024940629314.

Kneser, Reinhard and Hermann Ney (1995). "Improved Backing-off for M-Gram Language Modeling". In: *Proceedings of the IEEE International Conference on Acoustics and Speech Processing*. Vol. 1, pp. 181–184.

Koprinska, Irena, Josiah Poon, James Clark, and Jason Chan (2007). "Learning to classify e-mail". In: *Information Sciences* 177.10, pp. 2167–2187.

Kowalski, Robert (1986). "The limitation of logic". In: *CSC '86: Proceedings of the 1986 ACM fourteenth annual conference on Computer science*. Cincinnati, Ohio, United States: ACM, pp. 7–13. ISBN: 0-89791-177-6. DOI: http://doi.acm.org/10.1145/324634.325168.

Krishnalal, G, S Babu Rengarajan, and KG Srinivasagan (2010). "A new text mining approach based on HMM-SVM for web news classification". In: *International Journal of Computer Applications* 1.19, pp. 98–104.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton (2012). "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems*, pp. 1097–1105.

Kucharski, Adam (2016). "Post-truth: Study epidemiology of fake news". In: *Nature* 540.7634, pp. 525–525.

Kumar, M Arun and Madan Gopal (2010). "A comparison study on multiple binary-class SVM methods for unilabel text categorization". In: *Pattern Recognition Letters* 31.11, pp. 1437–1444.

Kumar, Ravi, Jasmine Novak, Prabhakar Raghavan, and Andrew Tomkins (2003). "On the bursty evolution of Blogspace". In: *Proceedings of the 12th international conference on World Wide Web*. New York, NY, USA: ACM, pp. 568–576.

Labeau, Matthieu, Kevin Löser, Alexandre Allauzen, and Rue John von Neumann (2015). "Non-lexical neural architecture for fine-grained POS Tagging." In: *EMNLP*, pp. 232–237.

Langville, Amy N. and Carl D. Meyer (2005). "A Survey of Eigenvector Methods for Web Information Retrieval". In: *SIAM Rev.* 47.1, pp. 135–161. ISSN: 0036-1445. DOI: http://dx.doi.org/10.1137/S0036144503424786.

Lenat, Douglas B. and Edward A. Feigenbaum (1991). "On the thresholds of knowledge". In: *Artif. Intell.* 47.1-3, pp. 185–250. ISSN: 0004-3702. DOI: http://dx.doi.org/10.1016/0004-3702(91)90055-O.

Lenat, Douglas B., R. V. Guha, Karen Pittman, Dexter Pratt, and Mary Shepherd (1990). "Cyc: toward programs with common sense". In: *Commun. ACM* 33.8, pp. 30–49. ISSN: 0001-0782. DOI: http://doi.acm.org/10.1145/79173.79176.

Lerman, Kristina and Rumi Ghosh (2010). *Information Contagion: An Empirical Study of the Spread of News on Digg and Twitter Social Networks*. URL: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/view/1509.

Leskovec, Jure, Lars Backstrom, and Jon Kleinberg (2009). "Meme-tracking and the dynamics of the news cycle". In: *KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. Paris, France: ACM,

pp. 497–506. ISBN: 978-1-60558-495-9. DOI: http://doi.acm.org/10.1145/1557019.1557077.

Levy, Omer and Yoav Goldberg (2014). "Neural word embedding as implicit matrix factorization". In: *Advances in neural information processing systems*, pp. 2177–2185.

Lewis, David D. and Kimberly A. Knowles (1997). "Threading electronic mail: A preliminary study". In: *Information Processing and Management* 33.2. Methods and Tools for the Automatic Construction of Hypertext, pp. 209 –217. ISSN: 0306-4573. DOI: DOI:10.1016/S0306-4573(96)00063-5. URL: http://www.sciencedirect.com/science/article/B6VC8-3SWVHFP-7/2/46fec0be28cb81c5d4bfc7f85fc4a908.

Li, Weiyuan and Hua Xu (2014). "Text-based emotion classification using emotion cause extraction". In: *Expert Systems with Applications* 41.4, Part 2, pp. 1742 –1749. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2013.08.073. URL: https://www.sciencedirect.com/science/article/pii/S0957417413006945.

Lin, Jane (2007). "Automatic author profiling of online chat logs". PhD thesis. Monterey, California. Naval Postgraduate School.

Lindsay, Robert K. and Michael D. Gordon (1999). "Literature-based discovery by lexical statistics". In: *J. Am. Soc. Inf. Sci.* 50.7, pp. 574–587. ISSN: 0002-8231. DOI: http://dx.doi.org/10.1002/(SICI)1097-4571(1999)50:7<574::AID-ASI3>3.0.CO;2-Q.

Liu, Bing and Lei Zhang (2012). "A survey of opinion mining and sentiment analysis". In: *Mining text data*. Springer, pp. 415–463.

Liu, Bingwei, Erik Blasch, Yu Chen, Dan Shen, and Genshe Chen (2013). "Scalable sentiment classification for big data analysis using naive bayes classifier". In: *Big Data, 2013 IEEE International Conference on*. IEEE, pp. 99–104.

Lloyd, Levon, Prachi Kaulgud, and Steven Skiena (2006). "Newspapers vs. blogs: Who gets the scoop". In: *AAAI spring symposium on Computational Approaches to Analyzing Weblogs*, pp. 117–124.

Madhawa, PKK and Ajantha S Atukorale (2015). "A robust algorithm for determining the newsworthiness of microblogs". In: *Advances in ICT for Emerging Regions (ICTer), 2015 Fifteenth International Conference on*. IEEE, pp. 135–139.

Magnini, Bernardo and Gabriela Cavaglià (2000). "Integrating subject field codes into wordnet". In: *2nd International conference on Language Resources and Evaluation*, pp. 1413–1418.

Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Prismatic Inc, Steven J. Bethard, and David Mcclosky (2014). "The Stanford CoreNLP natural language processing toolkit". In: *In ACL, System Demonstrations*.

Marcus, Mitchell P., Mary Ann Marcinkiewicz, and Beatrice Santorini (1993). "Building a Large Annotated Corpus of English: The Penn Treebank". In: *Comput. Linguist.* 19.2, pp. 313–330. ISSN: 0891-2017. URL: http://dl.acm.org/citation.cfm?id=972470.972475.

Mathioudakis, Michael and Nick Koudas (2010). "TwitterMonitor: trend detection over the twitter stream". In: *Proceedings of the 2010 international conference on Management of data*. SIGMOD '10. Indianapolis, Indiana, USA: ACM, pp. 1155–1158. ISBN: 978-1-4503-0032-2. DOI: http://doi.acm.org/10.1145/1807167.1807306. URL: http://doi.acm.org/10.1145/1807167.1807306.

Matuszek, Cynthia, Michael Witbrock, Robert C. Kahlert, John Cabral, Dave Schneider, Purvesh Shah, and Doug Lenat (2005). "Searching for common sense: populating Cyc™ from the web". In: *AAAI'05: Proceedings of the 20th national conference on Artificial intelligence*. Pittsburgh, Pennsylvania: AAAI Press, pp. 1430–1435. ISBN: 1-57735-236-x.

Mcburney, Peter and Simon Parsons (2001). "Chance Discovery Using Dialectical Argumentation". In: *New Frontiers in Artificial Intelligence*. Springer, pp. 414–424.

McCallum, Andrew Kachites (2002). "Mallet: A machine learning for language toolkit". In:

Mccarthy, John (1989). "Artificial intelligence, logic and formalizing common sense". In: *Philosophical Logic and Artificial Intelligence*. Kluwer Academic, pp. 161–190.

McCarthy, John and Patrick J. Hayes (1969). "Some Philosophical Problems from the Standpoint of Artificial Intelligence". In: *Machine Intelligence*. Edinburgh University Press, pp. 463–502.

McCorduck, Pamela (1979). *Machines Who Think*. New York, NY, USA: W. H. Freeman & Co. ISBN: 0716710722.

McNamee, Paul and Hoa Trang Dang (2009). "Overview of the TAC 2009 knowledge base population track". In: *Text Analysis Conference (TAC)*. Vol. 17, pp. 111–113.

Mesquita, Filipe (2012). "Clustering techniques for open relation extraction". In: *Proceedings of the on SIGMOD/PODS 2012 PhD Symposium*. ACM, pp. 27–32.

Mesquita, Filipe, Jordan Schmidek, and Denilson Barbosa (2013). "Effectiveness and efficiency of open relation extraction". In: *New York Times* 500, p. 150.

Mihaylov, Todor and Preslav Nakov (2016). "Hunting for troll comments in news community forums". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL*. Vol. 16, pp. 399–405.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean (2013a). "Distributed representations of words and phrases and their compositionality". In: *Advances in neural information processing systems*, pp. 3111–3119.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean (2013b). "Efficient estimation of word representations in vector space". In: *arXiv preprint arXiv:1301.3781*.

Mimno, David, Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum (2011). "Optimizing Semantic Coherence in Topic Models". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. EMNLP '11. Edinburgh, United Kingdom: Association for Computational Linguistics, pp. 262–272. ISBN: 978-1-937284-11-4. URL: http://dl.acm.org/citation.cfm?id=2145432.2145462.

Minsky, Marvin (1974). *A Framework for Representing Knowledge*. Tech. rep. Cambridge, MA, USA.

Mizoguchi, Riichiro and Mitsuru Ikeda (1996). *Technical Report AI-TR-96-1, I.S.I.R., Osaka Univ Towards Ontology Engineering*.

Moldovan, Cristian, Vasile Rus, and Arthur C Graesser (2011). "Automated Speech Act Classification For Online Chat." In: *MAICS* 710, pp. 23–29.

Mooney, Raymond J. and Razvan Bunescu (2005). "Mining knowledge from text using information extraction". In: *SIGKDD Explor. Newsl.* 7.1, pp. 3–10. ISSN: 1931-0145. DOI: http://doi.acm.org/10.1145/1089815.1089817.

Munson, Sean and Paul Resnick (2011). "The Prevalence of Political Discourse in Non-Political Blogs". In: *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*. URL: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2871.

Nagy, István and Veronika Vincze (2014). "VPCTagger: Detecting Verb-Particle Constructions With Syntax-Based Methods". In: *EACL 2014*, p. 17.

Nakov, Preslav, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, and Veselin Stoyanov (2016). "SemEval-2016 task 4: Sentiment analysis in Twitter". In: *Proceedings of the 10th international workshop on semantic evaluation (SemEval 2016), San Diego, US (forthcoming)*.

Nallapati, Ramesh, Ao Feng, Fuchun Peng, and James Allan (2004). "Event threading within news topics". In: *CIKM '04: Proceedings of the thirteenth ACM international conference on Information and knowledge management*. Washington, D.C., USA: ACM, pp. 446–453. ISBN: 1-58113-874-1. DOI: http://doi.acm.org/10.1145/1031171.1031258.

Narayanan, Vivek, Ishan Arora, and Arjun Bhatia (2013). "Fast and accurate sentiment classification using an enhanced Naive Bayes model". In: *International Conference on Intelligent Data Engineering and Automated Learning*. Springer, pp. 194–201.

Naughton, Martina, Nicola Stokes, and Joe Carthy (2010). "Sentence-level event classification in unstructured texts". In: *Information retrieval* 13.2, pp. 132–156.

Newman, David, Jey Han Lau, Karl Grieser, and Timothy Baldwin (2010). "Automatic evaluation of topic coherence". In: *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 100–108.

Orkin, Jeff and Deb Roy (2011). "Semi-automated dialogue act classification for situated social agents in games". In: *Agents for games and simulations II*. Springer, pp. 148–162.

Osborne, Miles, Sean Moran, Richard McCreadie, Alexander Von Lunen, Martin D Sykora, Elizabeth Cano, Neil Ireson, Craig Macdonald, Iadh Ounis, Yulan He, et al. (2014). "Real-time detection, tracking, and monitoring of automatically discovered events in social media". In:

Owoputi, Olutobi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A Smith (2013). "Improved part-of-speech tagging for online conversational text with word clusters". In: Association for Computational Linguistics.

Park, Sungrae, Wonsung Lee, and Il-Chul Moon (2015). "Efficient extraction of domain specific sentiment lexicon with active learning". In: *Pattern Recognition Letters* 56, pp. 38–44.

Pennington, Jeffrey, Richard Socher, and Christopher D Manning (2014). "Glove: Global Vectors for Word Representation." In: *EMNLP*. Vol. 14, pp. 1532–1543.

Pepper, Steve (1999). "Navigating haystacks and discovering needles: introducing the new topic map standard". In: *Markup Lang.* 1.4, pp. 47–74. ISSN: 1099-6621. DOI: http://dx.doi.org/10.1162/109966299760283201.

Pershina, Maria, Yifan He, and Ralph Grishman (2015). "Idiom Paraphrases: Seventh Heaven vs Cloud Nine". In: *Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics (LSDSem)*, p. 76.

Petersen, Sarah E. and Mari Ostendorf (2009). "A machine learning approach to reading level assessment". In: *Computer Speech & Language* 23.1, pp. 89 –106. ISSN: 0885-2308. DOI: https://doi.org/10.1016/j.csl.2008.04.003. URL: http://www.sciencedirect.com/science/article/pii/S0885230808000272.

Petrič, Ingrid, Tanja Urbančič, and Bojan Cestnik (2007). "Discovering Hidden Knowledge from Biomedical Literature". In: *Informatica* 31, pp. 15–20.

Petrov, Slav, Dipanjan Das, and Ryan McDonald (2011). "A universal part-of-speech tagset". In: *arXiv preprint arXiv:1104.2086*.

Petrovic, Sasa, Miles Osborne, Richard McCreadie, Craig Macdonald, Iadh Ounis, and Luke Shrimpton (2013). "Can Twitter Replace Newswire for Breaking News?" In: *Seventh International AAAI Conference on Weblogs and Social Media*.

Petrović, Saša, Miles Osborne, and Victor Lavrenko (2010). "Streaming First Story Detection with application to Twitter". In: *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Los Angeles, CA, USA: Association for Computational Linguistics, pp. 181–189. URL: http://www.aclweb.org/anthology/N/N10/N10-1021.

Poesio, Massimo and Renata Vieira (1998). "A corpus-based investigation of definite description use". In: *Computational linguistics* 24.2, pp. 183–216.

Polat, Kemal and Salih Güneş (2009). "A novel hybrid intelligent method based on C4. 5 decision tree classifier and one-against-all approach for multi-class classification problems". In: *Expert Systems with Applications* 36.2, pp. 1587–1592.

Poria, Soujanya, Erik Cambria, Gregoire Winterstein, and Guang-Bin Huang (2014). "Sentic patterns: Dependency-based rules for concept-level sentiment analysis". In: *Knowledge-Based Systems* 69, pp. 45–63.

Pratt, Wanda and Meliha Yetisgen-Yildiz (2003). "LitLinker: capturing connections across the biomedical literature". In: *K-CAP '03: Proceedings of the 2nd international conference on Knowledge capture*. Sanibel Island, FL, USA: ACM, pp. 105–112. ISBN: 1-58113-583-1. DOI: http://doi.acm.org/10.1145/945645.945662.

Prince, Ellen F (1992). "The ZPG letter: Subjects, definiteness, and information-status". In: *Discourse description: diverse analyses of a fund raising text*, pp. 295–325.

Qadir, Ashequl and Ellen Riloff (2011). "Classifying sentences as speech acts in message board posts". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 748–758.

Qiu, Yonggang and Hans-Peter Frei (1993). "Concept based query expansion". In: *SIGIR '93: Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*. Pittsburgh, Pennsylvania, United States: ACM, pp. 160–169. ISBN: 0-89791-605-0. DOI: http://doi.acm.org/10.1145/160688.160713.

Quinlan, J Ross (2014). *C4. 5: programs for machine learning*. Elsevier.

Rajagopal, Dheeraj, Erik Cambria, Daniel Olsher, and Kenneth Kwok (2013). "A graph-based approach to commonsense concept extraction and semantic similarity detection". In: *Proceedings of the 22nd international conference on World Wide Web companion*. International World Wide Web Conferences Steering Committee, pp. 565–570.

Ramage, Daniel, David Hall, Ramesh Nallapati, and Christopher D Manning (2009). "Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora". In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*. Association for Computational Linguistics, pp. 248–256.

Ramisch, Carlos, Paulo Schreiner, Marco Idiart, and Aline Villavicencio (2008). "An evaluation of methods for the extraction of multiword expressions". In: *Proceedings of the LREC Workshop-Towards a Shared Task for Multiword Expressions (MWE 2008)*, pp. 50–53.

Ramnial, Hoshiladevi, Shireen Panchoo, and Sameerchand Pudaruth (2016). "Authorship Attribution Using Stylometry and Machine Learning Techniques". In: *Intelligent Systems Technologies and Applications: Volume 1*. Ed. by Stefano Berretti, Sabu M. Thampi, and Praveen Ranjan Srivastava. Cham: Springer International Publishing, pp. 113–125. ISBN: 978-3-319-23036-8. DOI: 10.1007/978-3-319-23036-8_10. URL: https://doi.org/10.1007/978-3-319-23036-8_10.

Rao, Delip and Deepak Ravichandran (2009). "Semi-supervised polarity lexicon induction". In: *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 675–682.

Reed, Stephen L. and Douglas B. Lenat (2002). *Mapping Ontologies into Cyc*.

Rennie, Jason D, Lawrence Shih, Jaime Teevan, David R Karger, et al. (2003). "Tackling the poor assumptions of naive bayes text classifiers". In: *ICML*. Vol. 3. Washington DC), pp. 616–623.

Riedl, Martin and Chris Biemann (2016). "Impact of MWE Resources on Multiword Recognition". In: *ACL 2016*, p. 107.

Robertson, S E and K Sparck Jones (1976). "Relevance weighting of search terms". In: *Journal of the American Society for Information Science* 27.

Robertson, S. E., S. Walker, and M. Beaulieu (2000). "Experimentation as a way of life: Okapi at TREC". In: *Inf. Process. Manage.* 36.1, pp. 95–108. ISSN: 0306-4573. DOI: http://dx.doi.org/10.1016/S0306-4573(99)00046-1.

Rondon, Alexandre C, Helena de Medeiros Caseli, and Carlos Ramisch (2015). "Never-Ending Multiword Expressions Learning". In: *Proceedings of NAACL-HLT*, pp. 45–53.

Rosso, P., E. Ferretti, D. Jiminez, and V. Vidal (2004). "Text categorization and information retrieval using WordNet senses". In: *Proceedings of the 2nd Global Wordnet Conference (GWC'04)*. Brno, Czech Republic, pp. 299–304.

Rubenstein, H. and J. Goodenough (1965). "Contextual correlates of synonymy". In: *Communications of the ACM* 8.10, pp. 627–633.

Rus, Vasile, Arthur Graesser, Cristian Moldovan, and Nobal Niraula (2012). "Automatic discovery of speech act categories in educational games". In: *Educational Data Mining 2012*.

Russell, Bertrand. (1905). "On Denoting". In: *Mind* 14, pp. 479–493. URL: https://en.wikisource.org/wiki/On_Denoting.

Ruthven, Ian and Mounia Lalmas (2003). "A survey on the use of relevance feedback for information access systems". In: *The Knowledge Engineering Review* 18.02, pp. 95–145. DOI: 10.1017/S0269888903000638. eprint: http://journals.cambridge.org/article_S0269888903000638. URL: http://journals.cambridge.org/action/displayAbstract?fromPage=online&aid=187423&fulltextType=RA&fileId=S0269888903000638.

Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger (2002). "Multiword Expressions: A Pain in the Neck for NLP". In: *Computational Linguistics and Intelligent Text Processing: Third International Conference, CICLing 2002 Mexico City, Mexico, February 17–23, 2002 Proceedings*. Ed. by Alexander Gelbukh. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 1–15. ISBN: 978-3-540-45715-2. DOI: 10.1007/3-540-45715-1_1. URL: https://doi.org/10.1007/3-540-45715-1_1.

Saggion, Horacio, Emma Barker, Robert Gaizauskas, and Jonathan Foster (2005). "Integrating NLP tools to support information access to news archives". In: *Proceedings of the 5th Int'l conference. on Recent Advances in Natural Language Processing*.

Saleh, M Rushdi, Maria Teresa Martín-Valdivia, Arturo Montejo-Ráez, and LA Ureña-López (2011). "Experiments with SVM to classify opinions in different domains". In: *Expert Systems with Applications* 38.12, pp. 14799–14804.

Sanderson, Mark (1994). "Word sense disambiguation and information retrieval". In: *SIGIR '94: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*. Dublin, Ireland: Springer-Verlag New York, Inc., pp. 142–151. ISBN: 0-387-19889-X.

Sangati, Federico and Andreas van Cranenburgh (2015). "Multiword Expression Identification with Recurring Tree Fragments and Association Measures". In: *Proceedings of NAACL-HLT*, pp. 10–18.

Santosh, K, Romil Bansal, Mihir Shekhar, and Vasudeva Varma. "Author Profiling: Predicting Age and Gender from Blogs—Notebook for PAN at CLEF 2013". In: *In Forner et*, p. 10.

Sarkar, Avik, Paul H. Garthwaite, and Anne De Roeck (2005). "A Bayesian mixture model for term re-occurrence and burstiness". In: *CONLL '05: Proceedings of the Ninth Conference on Computational Natural Language Learning*. Ann Arbor, Michigan: Association for Computational Linguistics, pp. 48–55.

Schank, Roger C. (1975). "The primitive acts of conceptual dependency". In: *TINLAP '75: Proceedings of the 1975 workshop on Theoretical issues in natural language processing*. Cambridge, Massachusetts: Association for Computational Linguistics, pp. 34–37. DOI: http://dx.doi.org/10.3115/980190.980205.

Schmitz, Birte and Joachim J Quantz (2013). "Dialogue acts in automatic dialogue interpreting". In:

Schmitz, Michael, Robert Bart, Stephen Soderland, Oren Etzioni, et al. (2012). "Open language learning for information extraction". In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, pp. 523–534.

Schone, Patrick and Daniel Jurafsky (2001). "Is knowledge-free induction of multiword unit dictionary headwords a solved problem". In: *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, pp. 100–108.

Searle, John R (1969). *Speech acts: An essay in the philosophy of language*. Vol. 626. Cambridge university press.

Shi, Lei, Mei Weng, Xinming Ma, and Lei Xi (2010). "Rough set based decision tree ensemble algorithm for text classification". In:

Shriberg, Elizabeth, Raj Dhillon, Sonali Bhagat, Jeremy Ang, and Hannah Carvey (2004). *The ICSI meeting recorder dialog act (MRDA) corpus*. Tech. rep. DTIC Document.

Simmons, Matthew, Lada Adamic, and Eytan Adar (2011). "Memes Online: Extracted, Subtracted, Injected, and Recollected". In: *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*. URL: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2836.

Singh, Vivek Kumar, Rajesh Piryani, A Uddin, and P Waila (2013). "Sentiment analysis of movie reviews: A new feature-based heuristic for aspect-level sentiment classification". In: *Automation, Computing, Communication, Control and Compressed Sensing (iMac4s), 2013 International Multi-Conference on*. IEEE, pp. 712–717.

Skowron, Marcin, Mathias Theunis, Stefan Rank, and Arvid Kappas (2013). "Affect and Social Processes in Online Communication–Experiments with an Affective Dialog System". In: *IEEE Transactions on Affective Computing* 4.3, pp. 267–279.

Smadja, F (1993). "Retrieving collations from text: Xtract". In: *Computational Linguistics* 19.1, pp. 143–177.

Soboroff, Ian and Donna Harman (2005). "Novelty detection: the TREC experience". In: *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Vancouver, British Columbia, Canada:

Association for Computational Linguistics, pp. 105–112. DOI: http://dx.doi.org/10.3115/1220575.1220589.

Soricut, Radu and Franz Josef Och (2015). "Unsupervised Morphology Induction Using Word Embeddings." In: *HLT-NAACL*, pp. 1627–1637.

Srinivasan, P (2004). "Text Mining: Generating hypotheses from MEDLINE". In: *American Society for Information Science and Technology* 55.5, pp. 396–413.

Srivastava, Ashok and Mehran Sahami (2009). *Text Mining: Classification, Clustering, and Applications*. 1st. Chapman & Hall/CRC. ISBN: 1420059408, 9781420059403.

Stevens, Keith, Philip Kegelmeyer, David Andrzejewski, and David Buttler (2012). "Exploring Topic Coherence over Many Models and Many Topics". In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. EMNLP-CoNLL '12. Jeju Island, Korea: Association for Computational Linguistics, pp. 952–961. URL: http://dl.acm.org/citation.cfm?id=2390948.2391052.

Strapparava, Carlo and Rada Mihalcea (2007). "Semeval-2007 task 14: Affective text". In: *Proceedings of the 4th International Workshop on Semantic Evaluations*. Association for Computational Linguistics, pp. 70–74.

Suchanek, Fabian M, Gjergji Kasneci, and Gerhard Weikum (2008). "Yago: A large ontology from wikipedia and wordnet". In: *Web Semantics: Science, Services and Agents on the World Wide Web* 6.3, pp. 203–217.

Sureka, Ashish and Atul Goyal (2010). "Insights on Transferability of Dialog-Act Cue-Phrases Across Communication Domains, Modality and Semantically Similar Dialog-Acts". In: *8th International Conference on Natural Language Processing (ICON)*, pp. 28–37.

Svore, Krysta M. and Christopher J.C. Burges (2009). "A machine learning approach for improved BM25 retrieval". In: *CIKM '09: Proceeding of the 18th ACM conference on Information and knowledge management*. Hong Kong, China: ACM, pp. 1811–1814. ISBN: 978-1-60558-512-3. DOI: http://doi.acm.org/10.1145/1645953.1646237.

Swanson, D R (1988). "Migraine and Magnesium: eleven neglected connections". In: *Perspectives in Biology and Medicine, 31:526â€“557*.

Swanson, D R and N R Smalheiser (1999). "Implicit text linkages between Medline records: Using Arrowsmith as an aid to scientific discovery". In: *Library Trends* 48, pp. 48–59.

Swanson, Don R (1986). "Fish Oil, Raynaud's syndrome, and undiscovered public knowledge". In: *Perspectives in Biology and Medicine*, pp. 7–18.

Tavafi, Maryam, Yashar Mehdad, Shafiq Joty, Giuseppe Carenini, and Raymond Ng (2013). "Dialogue act recognition in synchronous and asynchronous conversations". In: *Proceedings of the 14th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL 2013), SIGDIAL*. Vol. 13.

Thompson, P and C Dozier (1997). "Name Searching and Information Retrieval". In: *In Proceedings of Second Conference on Empirical Methods in Natural Language Processing*, pp. 134–140.

Toutanova, Kristina, Dan Klein, Christopher D Manning, and Yoram Singer (2003a). "Feature-rich part-of-speech tagging with a cyclic dependency network". In: *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*. Association for Computational Linguistics, pp. 173–180.

Toutanova, Kristina, Dan Klein, Christopher D. Manning, and Yoram Singer (2003b). "Feature-rich part-of-speech tagging with a cyclic dependency network". In: *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology 1*. Edmonton, Canada: Association for Computational Linguistics, pp. 173–180. DOI: http://dx.doi.org/10.3115/1073445.1073478. URL: http://dx.doi.org/10.3115/1073445.1073478.

Traum, David R (2000). "20 questions on dialogue act taxonomies". In: *Journal of semantics* 17.1, pp. 7–30.

Tsvetkov, Yulia and Shuly Wintner (2014). "Identification of multiword expressions by combining multiple linguistic information sources". In: *Computational Linguistics* 40.2, pp. 449–468.

Tumasjan, Andranik, Timm Sprenger, Philipp Sandner, and Isabell Welpe (2010). "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment". In: *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*. URL: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/view/1441.

Villaroel Ordenes, F, S Ludwig, D Grewal, K de Ruyter, and M Wetzels (2016). "Analyzing Online Reviews Through the Lens of Speech Act Theory: Implications for Consumer Sentiment Analysis". In: *Journal of Consumer Research*.

Villavicencio, Aline, Valia Kordoni, Yi Zhang, Marco Idiart, and Carlos Ramisch (2007). "Validation and Evaluation of Automatically Acquired Multiword Expressions for Grammar Engineering." In: *EMNLP-CoNLL*, pp. 1034–1043.

Vincze, Veronika, István Nagy, and Gábor Berend (2011). "Multiword Expressions and Named Entities in the Wiki50 Corpus." In: *RANLP*, pp. 289–295.

Vincze, Veronika, István Nagy T, and János Zsibrita (2013). "Learning to detect English and Hungarian light verb constructions". In: *ACM Transactions on Speech and Language Processing (TSLP)* 10.2, p. 6.

Vinyals, Oriol, Łukasz Kaiser, Terry Koo, Slav Petrov, Ilya Sutskever, and Geoffrey Hinton (2015). "Grammar as a foreign language". In: *Advances in Neural Information Processing Systems*, pp. 2773–2781.

Voorhees, Ellen M. (1993). "Using WordNet to disambiguate word senses for text retrieval". In: *SIGIR '93: Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*. Pittsburgh, Pennsylvania, United States: ACM, pp. 171–180. ISBN: 0-89791-605-0. DOI: http://doi.acm.org/10.1145/160688.160715.

Voorhees, Ellen M. and Donna K. Harman (2005). *TREC: Experiment and Evaluation in Information Retrieval*. Digital Libraries and Electronic Publishing. MIT Press. ISBN: 0262220733.

Wallace, Byron C, Michael J Paul, Urmimala Sarkar, Thomas A Trikalinos, and Mark Dredze (2014). "A large-scale quantitative analysis of latent factors and sentiment in online doctor reviews". In: *Journal of the American Medical Informatics Association* 21.6, pp. 1098–1103.

Wallach, Hanna M, Iain Murray, Ruslan Salakhutdinov, and David Mimno (2009). "Evaluation methods for topic models". In: *Proceedings of the 26th annual international conference on machine learning*. ACM, pp. 1105–1112.

Wang, Xuerui and Andrew McCallum (2006). "Topics over time: a non-Markov continuous-time model of topical trends". In: *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. Philadelphia, PA, USA: ACM, pp. 424–433. ISBN: 1-59593-339-5. DOI: http://doi.acm.org/10.1145/1150402.1150450.

Watrin, Patrick and Thomas François (2011). "An n-gram frequency database reference to handle MWE extraction in NLP applications". In: *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*. Association for Computational Linguistics, pp. 83–91.

Weeber, Marc, Henry Klein, Lolkje T. W. de Jong-van den Berg, and Rein Vos (2001). "Using concepts in literature-based discovery: simulating Swanson's Raynaud-fish oil and migraine-magnesium discoveries". In: *J. Am. Soc. Inf. Sci. Technol.* 52.7, pp. 548–557. ISSN: 1532-2882. DOI: http://dx.doi.org/10.1002/asi.1104.abs.

Weston, Jason, Samy Bengio, and Nicolas Usunier (2011). "Wsabie: Scaling up to large vocabulary image annotation". In: *IJCAI*. Vol. 11, pp. 2764–2770.

Wiebe, Janyce and Ellen Riloff (2005). "Creating subjective and objective sentence classifiers from unannotated texts". In: *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, pp. 486–497.

Wilks, Yorick (1975). "An intelligent analyzer and understander of English". In: *Commun. ACM* 18.5, pp. 264–274. ISSN: 0001-0782. DOI: http://doi.acm.org/10.1145/360762.360770.

Wilks, Yorick and Roberta Catizone (2002). "Lexical Tuning". In: *CICLing '02: Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*. London, UK: Springer-Verlag, pp. 106–125. ISBN: 3-540-43219-1.

Wu, J and F Zheng (2000). *On enhancing katz-smoothing based back-off language model*.

Yang, Yiming, Jian Zhang, Jaime Carbonell, and Chun Jin (2002). "Topic-conditioned novelty detection". In: *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. Edmonton, Alberta, Canada: ACM, pp. 688–693. ISBN: 1-58113-567-X. DOI: http://doi.acm.org/10.1145/775047.775150.

Ye, Qiang, Ziqiong Zhang, and Rob Law (2009). "Sentiment classification of online reviews to travel destinations by supervised machine learning approaches". In: *Expert Systems with Applications* 36.3, pp. 6527–6535.

Yetisgen-Yildiz, Meliha and Wanda Pratt (2009). "A new evaluation methodology for literature-based discovery systems". In: *Journal of Biomedical Informatics* 42.4, pp. 633 –643. ISSN: 1532-0464. DOI: https://doi.org/10.1016/j.jbi.2008.12.001. URL: http://www.sciencedirect.com/science/article/pii/S1532046408001482.

Zhai, Chengxiang and John Lafferty (2004). "A study of smoothing methods for language models applied to information retrieval". In: *ACM Trans. Inf. Syst.* 22.2, pp. 179–214. ISSN: 1046-8188. DOI: http://doi.acm.org/10.1145/984321.984322.

Zhang, Renxian, Dehong Gao, and Wenjie Li (2012). "Towards scalable speech act recognition in twitter: tackling insufficient training data". In: *Proceedings of the Workshop on Semantic Analysis in Social Media*. Association for Computational Linguistics, pp. 18–27.

Zhao, Bin, Fei Li, and Eric P Xing (2011). "Large-scale category structure aware image categorization". In: *Advances in Neural Information Processing Systems*, pp. 1251–1259.

Zhao, Qiankun, Prasenjit Mitra, and Bi Chen (2007). "Temporal and information flow based event detection from social text streams". In: *AAAI'07: Proceedings of the 22nd national conference on Artificial intelligence*. Vancouver, British Columbia, Canada: AAAI Press, pp. 1501–1506. ISBN: 978-1-57735-323-2.

Zhou, Jie and Wei Xu (2015). "End-to-end learning of semantic role labeling using recurrent neural networks." In: *ACL (1)*, pp. 1127–1137.

Zimmermann, Max, Irene Ntoutsi, Zaigham Faraz Siddiqui, Myra Spiliopoulou, and Hans-Peter Kriegel (2012). "Discovering global and local bursts in a stream of news". In: *Proceedings of the 27th Annual ACM Symposium on Applied Computing*. ACM, pp. 807–812.