
The Detection of Contradictory Claims in Biomedical Abstracts



Author:

Abdulaziz D. Alamri

Supervisor:

Dr. Mark Stevenson

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

in the

Natural Language Processing Group
Department of Computer Science

December 20, 2016

Declaration of Authorship

I, Abdulaziz D. Alamri, declare that this thesis titled, “The Detection of Contradictory Claims in Biomedical Abstracts” 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:

THE UNIVERSITY OF SHEFFIELD

Abstract

Faculty of Engineering

Department of Computer Science

Doctor of Philosophy

The Detection of Contradictory Claims in Biomedical Abstracts

by Abdulaziz D. Alamri

Research claims in the biomedical domain are not always consistent, and may even be contradictory. This thesis explores contradictions between research claims in order to determine whether or not it is possible to develop a solution to automate the detection of such phenomena. Such a solution will help decision-makers, including researchers, to alleviate the effects of contradictory claims on their decisions.

This study develops two methodologies to construct corpora of contradictions. The first methodology utilises systematic reviews to construct a manually-annotated corpus of contradictions. The second methodology uses a different approach to construct a corpus of contradictions which does not rely on human annotation. This methodology is proposed to overcome the limitations of the manual annotation approach.

Moreover, this thesis proposes a pipeline to detect contradictions in abstracts. The pipeline takes a question and a list of research abstracts which may contain answers to it. The output of the pipeline is a list of sentences extracted from abstracts which answer the question, where each sentence is annotated with an assertion value with respect to the question. Claims which feature opposing assertion values are considered as potentially contradictory claims.

The research demonstrates that automating the detection of contradictory claims in research abstracts is a feasible problem.

Acknowledgements

I would like to thank my advisor, Dr. Mark Stevenson, for guiding and supporting me over the past few years. Without his guidance and feedback, completing this PhD would not have been possible.

I would also like to thank my thesis committee members for all of their guidance throughout this process; your discussions, ideas and feedback have been absolutely invaluable.

I would like to thank my amazing family for the love, support, and constant encouragement I have received over the years. In particular, I would like to thank my parents, my brothers, my sister, my wife, and my children Rand and Khalid.

I would finally like extend my thanks to my employer, the Ministry of the Interior in Saudi Arabia, for the funding received towards my PhD studies.

Contents

Declaration of Authorship	iii
Abstract	vii
Acknowledgements	ix
1 Introduction	1
1.1 Background	3
1.2 Definition of Contradiction (See <i>Section (3.2)</i>)	5
1.3 Research Aim and Objectives	7
1.4 Contributions	8
2 Related Literature	9
2.1 Introduction	9
2.2 Contradiction	9
2.2.1 Contradiction Typology	11
2.2.2 Contradiction Recognition Approaches	15
2.2.3 Contradiction in Biomedical Domain	20
2.3 Paraphrasing and Textual Entailment Recognition	22
2.3.1 Methods Based on Lexical-Syntactic Similarity	24
2.3.2 Methods Based on Semantic Similarity	26
2.3.3 Methods Based on Rules	27
2.3.4 Methods Based on Logic	29
2.3.5 Methods Based on Machine Learning	30
2.3.6 Paraphrasing and Textual Entailment Corpora	31
2.4 Argumentation Mining	33
2.4.1 Argumentation Mining in Non-Scientific Text	33
2.4.2 Argumentation Mining in Scientific Text	34
2.4.3 Argumentation Mining Corpora	37
2.5 Negation and Speculation	38
2.6 Information Extraction	39
2.6.1 Named Entity Recognition	40

2.6.2	Relation Extraction	42
2.6.3	Event Extraction	43
2.7	Question Answering	44
2.7.1	Medical QA	44
2.7.2	Biological QA	46
2.8	Evaluation Methods Overview	47
2.8.1	Intrinsic vs Extrinsic Evaluation	47
2.8.2	Accuracy, Precision, Recall and F1-score Measures	48
2.9	Conclusions	49
3	Two Corpora of Contradictory Research Claims	51
3.1	Introduction	51
3.2	Definitions	52
3.2.1	Contradiction Definition	52
3.2.2	Claim Definition and Types	54
3.3	Manually Annotated Contradiction Corpus	56
3.3.1	Corpus Data Collection	56
3.3.2	Question Formulation	59
3.3.3	Corpus Annotation	60
3.3.4	Results and Discussion	62
3.4	Automatically Annotated Contradiction Corpus	69
3.4.1	SemMedDB	69
3.4.2	Hypothesis to Collect Contradictory Sentences	72
3.4.3	SemMedDB Browser	74
3.4.4	Corpus Construction	76
3.4.5	Results and Discussion	77
3.5	Conclusions	82
4	Identification of Research Claims in Biomedical Abstracts	83
4.1	Introduction	83
4.2	Previous Work Related to Claim Zoning Component	84
4.2.1	Methods	85
4.2.2	Conditional Random Fields (CRFs)	85
4.2.3	Features	86
4.2.4	Data	88
4.2.5	Results and Discussion	88
4.3	Previous Work Related to Answer Selection Component	91

4.3.1	Methods	92
4.3.2	Support Vector Machines (SVMs)	93
4.3.3	Features	93
4.3.4	Data	95
4.3.5	Results and Discussion	96
4.4	Pipeline System	98
4.5	Conclusions	101
5	Identification of Contradictory Claims in Biomedical Abstracts	103
5.1	Introduction	103
5.2	Related Work	103
5.3	Methods	105
5.4	Lexicons	106
5.5	Fact Extraction	107
5.5.1	Relation Extraction	108
5.5.2	Biomedical Concept Identification	110
5.5.3	Relation Relatedness	111
5.6	Fact Assertion Value Detection	112
5.7	Results and Discussion	114
5.7.1	Fact Extraction Stage	114
5.7.2	Fact Assertion Value Detection	115
5.7.3	Error Analysis	121
5.8	Evaluation	123
5.9	Research Claims Highlighter	124
5.10	Conclusions	128
6	Conclusions	129
6.1	Summary of Contributions	130
6.2	Future Work and Open Questions	130
A	Corpus Annotation Guidelines:	133
A.1	Definitions	133
A.1.1	Formulation of PICO Questions	134
	Notes	134
A.1.2	Identification of Claims	134
	Notes	135
A.1.3	Annotation of Claims Assertion Values	135
A.1.4	Annotation of the Claim Type	135

Notes	135
B The Questions Formulated for ManConCorpus	137
C Lexicons	139
C.1 Negation	139
C.2 Directionality	140
C.3 Sentiment	144
Bibliography	145

List of Figures

2.1	The semantic representation of statements (34) and (35)	21
2.2	Dependency trees of statements (49) and (50) by Malakasiotis and Androutsopoulos (2007)	25
2.3	Grammatical dependency relations of statements (49) and (50)	26
2.4	Paraphrasing and textual entailment using semantic similarity by Malakasiotis (2009)	26
2.5	Tree skeletons used for textual entailment recognition	28
3.1	An example of a forest plot diagram. The dataset was retrieved from Viechtbauer (2010)	58
3.2	Examples of formatted claims	63
3.3	A diagram of how <i>SemRep</i> extracts a semantic predicate	71
3.4	The main interface of the <i>SemMedDB</i> Browser system	75
3.5	Sentences that contain a particular tuple	76
3.6	Examples of formatted sentences	79
4.1	Abstract (19414832) is an example of how label values can differ from the <i>nlmCategory</i> values	88
4.2	The maximum margin separating the hyperplane within two dataset classes using a linear SVM	93
5.1	Research Claims Highlighter Architecture	125
5.2	A sample of a formatted XML file	126
5.3	Research Claims Highlighter System interface	127
5.4	The system configured to display the unstructured abstracts as structured	127
5.5	An abstract featuring a claim that agrees with the query	128
5.6	An abstract featuring a claim that disagrees with the query	128

List of Tables

1.1	Examples of contradictory claims from the biomedical domain	6
2.1	An example of contradictory answers to a question (Harabagiu, Hickl, and Lacatusu, 2006a)	10
2.2	An example of contradictory statements due to antonym used by (Marneffe, Rafferty, and Manning, 2008)	11
2.3	The first relation tuple is an example of a negated event (<i>tested</i>). the second relation tuple is an example of a negated entity (<i>judge</i>) used by (Harabagiu, Hickl, and Lacatusu, 2006a).	12
2.4	An example of contradictory statements due to negation used by (Harabagiu, Hickl, and Lacatusu, 2006a)	12
2.5	An example of contradictory statements due to number mismatch used by (Marneffe, Rafferty, and Manning, 2008)	13
2.6	Examples of contradictory statements due to factivity used by (Marneffe, Rafferty, and Manning, 2008)	13
2.7	Modality rules for textual entailments and contradiction used by (MacCartney et al., 2006)	14
2.8	Examples of contradictory statements due to lexical (16 & 17) and syntactical structure (18 & 19) used by (Marneffe, Rafferty, and Manning, 2008)	14
2.9	Examples of apparent contradictory statements due to background used by (Ritter et al., 2008)	15
2.10	Examples of the contradictory sentences used by Harabagiu, Hickl, and Lacatusu (2006a) dataset	15
2.11	The system performance achieved by Harabagiu, Hickl, and Lacatusu (2006a)	16
2.12	The system performance achieved by Marneffe, Rafferty, and Manning (2008)	17
2.13	Assumptions considered in Ritter et al. (2008) system	18
2.14	Contradiction over functional relationship by (Ritter et al., 2008)	18

2.15	Pham, Nguyen, and Shimazu (2013) system performance	19
2.16	Sentences (34 & 35) are examples of contradictory statements and (36 & 37) are examples of contrastive statements used by Sarafraz (2011)	21
2.17	Examples of paraphrases and entailments generated by Androutsopoulos and Malakasiotis (2010)	23
2.18	Examples of paraphrases and entailments templates by Androutsopoulos and Malakasiotis (2010)	23
2.19	Textual entailment using lexical similarity by Malakasiotis and Androutsopoulos (2007)	24
2.20	Textual entailment using syntactical similarity by Malakasiotis and Androutsopoulos (2007)	25
2.21	Paraphrasing and textual entailment using rules by Malakasiotis (2009)	27
2.22	Generating new rules using synonyms for textual entailment recognition	28
2.23	The LEDIR rules of Bhagat, Pantel, and Hovy (2007)	29
2.24	A rule for textual entailment using logic-based approach (Kamp and Reyle, 1993)	29
2.25	Examples of paraphrases sentences from MSRP corpus	31
2.26	Examples of instances from RTE-6/RTE-7 corpus	32
2.27	Examples of instances in Bowman et al. (2015a) corpus	32
2.28	A claim that belongs the explicit type based on Blake (2010) scheme	36
2.29	Examples of comparative claims from (Park and Blake, 2012)	36
2.30	A question formulated in PICO format	45
2.31	Confusion table for a binary classification problem	48
3.1	Examples of potentially contradictory sentences	53
3.2	Claims Typology	55
3.3	Example of studies associated with a systemic review	59
3.4	A sample of the questions formulated for the final corpus	64
3.5	Claims classes and type distribution among the groups	65
3.6	Claims extracted from the abstracts of the studies listed in Table (3.3)	66
3.7	Potential answers to a formulated question from the same abstract	66
3.8	The contingency table of claims identification	67
3.9	Multiple inferences derived from two claims	67
3.10	The contingency table of annotating whether the claims agreed or disagreed with the questions	68
3.11	The contingency table of annotating the claims types	68
3.12	<i>SemMedDB</i> database tables	70

3.13	A sentence extracted from a PubMed abstract	71
3.14	<i>SemMedDB</i> predicates	73
3.15	Examples of Incompatible relation tuples extracted from sentences in Table (3.16)	74
3.16	The sentences extracted from <i>SemMedDB</i> based on the incompatibility of their relation tuples described in Table (3.15)	74
3.18	Non-contradictory sentences that were included in the corpus as contradictory	77
3.17	<i>AutConCorpus</i> topics, relation tuples and sentence distributions	78
3.19	A title included in <i>AutConCorpus</i>	78
3.20	Examples of sentences in <i>ManConCorpus</i> from which <i>SemRep</i> failed to extract any relation tuples from.	81
3.21	Examples of relation tuples extracted from a set of claims in <i>ManConCorpus</i> . These tuples do not exist within <i>AutConCorpus</i>	81
4.1	The performance of <i>Claim Zoning</i> system. The baseline is 40.3%, which is the accuracy percentage of annotating all sentences with class <i>Results</i>	89
4.2	The <i>Answer Selection</i> component performance using different sets of features. The average F-score of the component, under the best setting, to annotate both <i>potential-answer</i> and <i>non-potential answer</i> achieved 97%. (The baseline score is 91%)	96
4.3	the pipeline system performance using <i>ManConCorpus-unst</i> . The baseline is 87%, which is the accuracy percentage of annotating all sentences with class <i>non-potential answer</i>	98
4.4	An example of a potential answer that was missed by the pipeline system due to its rhetorical label	99
4.5	The distribution of the <i>Conclusions</i> sentences among the <i>ManConCorpus-unst</i> abstracts after the <i>Claim Zoning</i> annotation. The first and the third columns show the number of abstracts that contain a specific number of <i>Conclusions</i> sentences and the second and the fourth columns show the number of <i>Conclusions</i> sentences in these abstracts	99
4.6	two examples from two abstracts where the pipeline annotated multiple sentences from the same abstract as potential answers	100
4.7	An example of an error generating by the pipeline system due to Z-scores	101
5.1	A sample of negation terms used in the negation lexicon	106
5.2	A sample of terms used in the directionality lexicon	107

5.3	A sample of terms used in the sentiment lexicon	107
5.4	An example of a claim sentence where ReVerb and WOE could not find relations to extract, while Ollie managed to extract at least one tuple. . .	109
5.5	Contradictory claims (2) and (3) with respect to the question (1).	110
5.6	The relation tuples extracted from the question and claims Table (5.5). . .	110
5.7	Annotation of claims tuples using <i>MetaMap</i>	111
5.8	Annotating negation terms in a claim tuple	113
5.9	Annotating directionality terms in a claim tuple	113
5.10	A claim sentence extracted from abstract (9412879) to be annotated by negation, directionality and sentiment	114
5.11	A fully annotated claim tuple	114
5.12	A complex claim, where relation extraction systems failed to correctly extract relations	115
5.13	The relation tuples extracted from the claim in Table (5.12)	115
5.14	The contradiction detection performance using <i>ManConCorpus</i> (left) and <i>AutConCorpus</i> (right). The baseline of <i>ManConCorpus</i> is 68% and that of <i>AutConCorpus</i> is 69.6%. The baseline score represents the accuracy percentage when annotating all sentences with class <i>yes</i>	116
5.15	The learning curves of using <i>ManConCorpus</i> and <i>AutConCorpus</i> using feature sets (3), (5) and (7)	120
5.16	The result of using training the classifier on the dataset of <i>AutConCorpus</i> to predict the assertion values of the claims in <i>ManConCorpus</i> (the baseline is 68%)	121
5.17	A relation tuple that was extracted from a claim and made the classifier to choose the wrong assertion value.	122
5.18	Another example of the classifier errors due to choosing the relation tuple based on its relatedness score with the question.	122
5.19	The performance of the <i>Contradiction Detection</i> system using <i>ManConCorpus-unst</i>	123
5.20	Estimated performance of the contradiction detection components in combination.	124
B.1	A list of the 24 questions formulated for the final corpus	138
C.1	The negation lexicon	139
C.2	Directionality lexicon	143
C.3	The sentiment lexicon	144

Chapter 1

Introduction

The biomedical research literature is vast, and rapidly increasing. It encompasses a substantial number of claims, including those supporting the effectiveness of treatments or reporting potential causes of diseases. These claims, however, are not always consistent and may even be contradictory, thus making it difficult for researchers and practitioners to understand current thinking about a research question without reading all research literature associated with it.

The contradiction between research claims in the biomedical field was noted early by Horwitz (1987). The existence of contradictory claims is not uncommon, and it was found that seven research claims out of forty-five from a highly-cited original studies were contradicted by subsequent studies addressing the same problems (Ioannidis, 2005). Decision-makers such as clinicians, researchers and even patients are confused by this issue, which makes it difficult to rely solely on such studies for making decisions when the results are contradictory. Ioannidis and Trikalinos (2005) reported that the contradiction between research claims is caused by the tendency of investigators to reproduce the outcomes of original research sometimes with contradictory findings. Editors and publishers are attracted by these results and tend to publish them more swiftly than those with other findings; an observation termed the *Proteus phenomenon*.

That has led the community to develop certain protocols which are generally considered tedious and time-consuming. For example, clinicians utilise point-of-care resources (e.g. *DynaMed Plus* and *UpToDate*) to summarise thousands of medical topics based on the evidence available in the literature. Such systems are regularly updated by qualified editors who follow a protocol to maintain the quality of the contents. Such systems require editors to spend a great deal of time on search engines to find, evaluate and summarise the studies of these topics. Similarly, researchers conduct systematic reviews (Higgins and Green, 2008) to summarise the results of multiple studies to reach a final conclusion about a specific question. Unfortunately, this process has proven to be a challenging and time-consuming task, since medical literature is growing exponentially and includes a large number of searchable databases.

Semi-automated systems are often used to reduce the workload when following such protocols. For example, Tsafnat et al. (2013) described a system which works at the stage of screening abstracts. The system uses natural language processing (NLP) methods capable of detecting sentences or phrases which are particularly important for appraisal. O'Mara-Eves et al. (2015) presented two approaches reliant on NLP, which could contribute to reducing the workload of such tasks. The first is to prioritise the abstracts of studies returned from a search engine based on their relevance to the review question, where the top studies in a ranking are those more likely to be relevant to the review question. The second is to apply machine learning methods to automate the task of including/excluding studies by learning from the decisions made by reviewers when including/excluding studies in a review (Thomas, McNaught, and Ananiadou, 2011). However, none of these approaches discuss the use of NLP to detect contradictory research claims, which is one of the key reasons for establishing reviewing protocols.

Tools that support the automatic identification of contradictory claims may be of benefit to those who rely on the biomedical literature, and could be used to highlight research claims that are contradictory to other research claims, and assist in the creation of systematic reviews and point-of-care resources. Such tools would also be useful for automatic text mining applications, which generally accept claims made within the research literature as *prima facie* correct.

Nevertheless, little research into this issue has been conducted in the NLP domain to date. Previous work (Sarafraz, 2011) focused on descriptions of molecular events by combining the events in the BioNLP09 corpus (Kim et al., 2009) with a subset of the events of the GENIA corpus (Ohta, Tateisi, and Kim, 2002). That work was restricted to a single indicator of contradiction/contrast, the use of negation. Outside of the biomedical domain, few researchers have studied the problem of contradiction identification, independently of the more general problem of textual inference (Bowman et al., 2015b; Harabagiu, Hickl, and Lacatusu, 2006a; Marneffe, Rafferty, and Manning, 2008).

This research studies contradiction between research claims in the biomedical domain in order to find out how well it is possible to automatically recognize such a phenomenon using the current NLP methods and technologies.

This chapter provides an introduction to the contradiction problem in biomedical research claims. This includes description of three examples of controversial topics which researchers published contradictory claims about. These topics have remained controversial until recently which impacted both the biomedical community and ordinary people.

Furthermore, the chapter uses one of these examples to explain how contradiction

between research claims can be captured. This is useful here as a preliminary definition of the contradiction concept considered in this research. The last sections in this chapter describe the aim and objective of this research including the specific goals to achieve the research goal. The remainder of this thesis is organized into five chapters.

Chapter 2 reviews the main literature related to this research. The first three sections (2.2), (2.3) and (2.4) describe related literature on topics which were found to be highly relevant to the main problem: contradiction, paraphrasing and textual entailment and argumentation. The rest of the sections (2.5 - 2.8) highlight topics which are considered in the practical section of this study, including the construction of the corpora and the development of the proposed system: negation and speculation, information extraction, question answering, and evaluation of NLP tasks.

Chapter 3 presents two corpora, constructed for the purpose of understanding the phenomenon of contradiction in biomedical research and the automation of the identification of such phenomenon. This chapter consists of two main sections, the first of which describes the procedure followed in order to construct *ManConCorpus*. This construction process follows the standard NLP approach of constructing a corpus for machine learning systems. The second section discusses the construction of *AutConCorpus*, which is larger and can be produced without the need for human annotation. Both corpora are used to validate the proposed solution, in order to identify potential contradictions between research claims.

Chapter 4 describes a pipeline system, using machine learning to detect authors' claims within research abstracts. It consists of two subsystems, the first of which uses techniques imported from the argumentation literature to detect the claim zone within each abstract. From each claim zone, the second subsystem identifies the sentences most relevant to the research question which could therefore represent research claims.

Chapter 5 presents a machine learning system to discover potential contradictions between claims. The corpora described in *Chapter 3* are used to train and evaluate the system, and a comparison and discussion of system performance when using the two corpora is provided.

Chapter 6 discusses the summary of the contributions and future work.

1.1 Background

In the 1970s, the public were advised that margarine was healthier than butter, eggs raised cholesterol levels in blood, and that teeth should be brushed thoroughly following the consumption of carbonated soft drinks. However, recent evidence shows that margarine is high in hydrogenated fats, the consumption of eggs has little impact on

cholesterol levels, and that brushing teeth immediately after drinking carbonated soft drinks can destroy tooth enamel and damage gums. Although this information has predominantly appeared in the media, the source of such conflicts can be found in the research literature, with different researchers providing contradictory answers to particular research questions. It is not only controversial topics which researchers disagree about; the domain of biomedical research features a myriad of other cases. Ritter et al. (2008) describe three cases.

The first concerns the value of mammography in the detection of breast cancer. Strax et al. (1967) conducted an experiment in the USA to evaluate the influence of mammography on reducing the mortality rate in those who underwent examination versus those who did not. The results showed that, after several years, fewer deaths were observed in those who underwent screening. In the 1980s, mammography became a widely-accepted mechanism for the early detection of cancer. In 1993, *The Lancet* published results from five research centres in Sweden which supported the positive effects of mammography (Nyström et al., 1993). The same outcomes were subsequently reproduced in Scotland, Canada, and the US. In 1999, however, a second Swedish research team (Mayor, 1999) found no evidence of a decreased risk of death from breast cancer in those who had undergone a mammography. This led a research group in Denmark (Gøtzsche and Olsen, 2000) to investigate the positive results earlier reported in *Lancet*. They found that six of eight trials were of poor quality, and that the two acceptable trials showed no correlation between mammography and mortality rate. These results put the value of mammography into doubt. Ultimately, it was discovered that the conflict between these findings was due to the use of different outcome measures, and that mammography is in fact beneficial (Ritter et al., 2008). The degree of benefit is arguable from a financial perspective, however, as the survivors in the trials tended to be older, and consequently the number of years restored was relatively low.

The second case concerns a contraceptive method called the Dalkon Shield which was popular during the 1970s. The Dalkon Shield was an intrauterine device (IUD), placed in the uterus to prevent pregnancy. The device was relatively inexpensive, safe and easy to use and remove. Problems emerged, however, when Christian (1974) reported the deaths of ten pregnant women who had used the shield. At the same time, other researchers argued that the device was responsible for an increased incidence of pelvic inflammatory disease (PID). In 1975, the manufacturer of the Dalkon Shield was advised to withdraw it from the market. In 1976, the National Institutes of Health (NIH) conducted a study which uncovered a relationship between IUD devices and PID. The results (Burkman, 1981) showed that those who used the Dalkon Shield were at higher

risk of contracting PID. However, Kronmal, Whitney, and Mumford (1991) re-analysed the study and concluded that the NIH study contained several flaws, including the interpretation of the results themselves. Furthermore, Mumford and Kessel (1992) investigated the researchers who believed the Dalkon Shield caused PID. They concluded that the accusations were erroneous, and that no correlation existed between the device and PID incidence. Although IUDs provide the highest satisfaction rate among contraceptive devices, only 1% of women in the USA use them (Ritter et al., 2008).

The third case involves aspirin. Although aspirin has not undergone the rigorous clinical testing required of modern medicines, it has been universally accepted as a painkiller. However, in addition to its analgesic properties, it has gained popularity for its efficiency in preventing blood clots. When heart disease causes the narrowing of the arteries, even a small blood clot can cause a heart attack. Medical research has shown that, in individuals who have previously had a heart attack, a daily dose of aspirin can prevent the occurrence of a second heart attack. The question thus arose as to whether it could prevent the first heart attack. A research study conducted by the Boston Collaborative Drug Surveillance program (1974) showed that aspirin had a significant benefit in preventing heart attack. Conversely, Hennekens, Karlson, and Rosner (1978) reported only a small difference in the risk of heart attack between those who did not use aspirin and those who frequently did. Furthermore, Paganini-Hill et al. (1989) supported previous findings, and concluded that aspirin failed to show a preventive role in heart attack. Another study found that the initial claims about the benefit of aspirin in heart attack were supported (Ritter et al., 2008). The contradictions within aspirin research lasted for twenty years, until researchers finally concluded that though aspirin certainly reduces the risk of non-fatal heart attacks, its effects on other problems, such as strokes, remain unclear (Brotons et al., 2015).

1.2 Definition of Contradiction (See Section (3.2))

In an attempt to link these cases to the contradiction definition considered in this research, which slightly differs from the logical strict definition of contradiction as known in linguistics, we use the research abstracts used in the third case to show how their research claims become contradictory in the light of the contradiction definition considered in this research. This definition states that two research claim sentences, T_1 and T_2 , are said to be contradictory when, for a given proposition F , information inferred about F from T_1 is unlikely to be true at the same time as information about F inferred from T_2 . A more thorough discussion about why this definition is chosen can be found later in Section (3.2).

Table (1.1) shows a research question (1), and three claims (2)¹, (3) and (4) that answer the question and were extracted from the abstracts of research Boston Collaborative Drug Surveillance program (1974), Hennekens, Karlson, and Rosner (1978), and Paganini-Hill et al. (1989). Claim (2) agrees with the research question, while claims (3) and (4) disagree.

	Claim	Agree?
1	In the elderly, does aspirin prevent the risk of coronary diseases?	–
2	The data are consistent with the hypothesis that aspirin protects against non-fatal myocardial infarction disease.	yes
3	These data provide no evidence for a preventive role of regular aspirin intake in coronary deaths	no
4	The daily use of aspirin increased the risk of kidney cancer and ischaemic heart disease	no

TABLE 1.1: Examples of contradictory claims from the biomedical domain

In linguistics, claim (2) might not be considered contradictory to claim (3) or (4) since they linguistically describe different relationships; but in biomedicine, claim (2) is considered contradictory to both claims since the information inferred from claim (2) is not compatible with the information inferred from claim (3) or (4), regarding the role of aspirin in coronary disease. Note that *myocardial infarction* mentioned in claim (2) is one type of coronary disease and *ischaemic heart disease* mentioned in claim (4) is a synonym of coronary disease. Furthermore, the information about *kidney cancer* in claim (4) did not affect the decision of considering claim (2) contradictory to claim (4), because that decision was based on common information between the claims. This definition ensures that the contradiction problem discussed in this research is tractable and is common with other work on contradiction detection in non-biomedical domains (Marneffe, Rafferty, and Manning, 2008).

¹The original claim sentence in the paper used the term *this* to refer to *non-fatal myocardial infarction* mentioned in the previous sentence. It was modified here for ease explanation

1.3 Research Aim and Objectives

The motivation of this research is to alleviate the consequences of contradictory claims on those who rely upon medical research to make informed decisions. The aim of this research is to study *how well contradictions between research claims in the biomedical domain can be recognized the using current NLP methodologies and tools*.

To achieve this goal we propose a solution which automatically detects research claims from abstracts likely to contradict each other with respect to a given research question. Such a solution is supposed to reduce the workload required from researchers or editors in the abstract screening stage when reviewing or summarising literature related to a particular question. The solution is developed under the assumption that it will be deployed as a component of a search engine.

The solution is assumed to improve search engines' capabilities by detecting contradictory claims in sentences from the abstracts that were returned by the search engine as relevant to the query. The solution will be useful for both editors, who update point-of-care resources, and researchers, who conduct systematic reviews, to minimise the personal effort required to summarise the studies of interest.

The research adopts the approach of developing systems to assist human users in accomplishing tasks rather than automating the task itself. Such an approach is intended to create a better partnership between human effort and the machine (Elliott, 2013), while simultaneously being a more reasonable goal to achieve.

To achieve the goal of this research, specific aims need to be addressed:

1. Exploration of current NLP literature on the subject of contradiction.
2. Exploration of the linguistic characteristics and features that can be used to locate research claim in abstracts.
3. Exploration of linguistic characteristics which can be used to determine potential contradictions between claims.
4. Construction of a corpus or corpora of contradictory research claims which can be applied to the testing of an automatic system developed to detect contradictory claims.
5. Development of a system to identify the sentences containing claims within research abstracts.
6. Development of a system to identify potential contradictions between research claims on a given question.

This research make one main assumption in order to achieve these goals; that the input of the system described in (5) above, which is responsible for identifying research claims, is a list of research abstracts which answer the same research questions. Because the proposed solution in this research is assumed to be used in a search engine setting, the search engine is expected to provide the system with abstracts. The system functions under the assumption that all of these abstracts contain answers to the same question.

1.4 Contributions

Work described in this thesis has resulted in the following publications:

- Alamri, Abdulaziz and Mark Stevenson (2016). A corpus of potentially contradictory research claims from cardiovascular research abstracts. In: *Journal of Biomedical Semantics* 7.1, pages 19. (**Chapter 3**)
- Alamri, Abdulaziz and Mark Stevenson. Introducing a New Methodology to Construct a Potentially Contradictory Corpus From Biomedical Domain, Proc. ACM DTMBIO Workshop (Submitted) (**Chapters 3 and 5**)
- Alamri, A. and Stevenson, M. (2015). Automatic detection of answers to research questions from MEDLINE abstracts. In *Proceedings of BioNLP 15, pages 141 - 146, Beijing, China*. Association for Computational Linguistics. (**Chapter 4**)
- Alamri, A. and Stevenson, M. (2015). Automatic identification of potentially contradictory claims to support systematic reviews. *2015 IEEE International Conference on Bioinformatics and Biomedicine*. (**Chapter 5**)

Chapter 2

Related Literature

2.1 Introduction

This chapter reviews the literature of various topics that are either directly relevant to the contradiction problem or to the approaches and methods used to explore it in this thesis. The next three sections explore the literature on contradiction (most of which does not focus on the biomedical domain), paraphrasing and textual entailment, and argumentation. Techniques found to be useful for exploring the contradiction problem are: negation and speculation, information extraction (IE), question answering (QA) and the chapter concludes with an overview of evaluation methods for NLP tasks.

Most of the literature on contradiction initially emerged from textual entailment tasks, which do not focus on biomedical text. The paraphrasing and textual entailment literature is therefore important for this research. It also discusses methods for inferring answers expressed differently in texts, which is very relevant to the topic of the thesis (Haghighi, Ng, and Manning, 2005).

Argumentation is considered a central aspect of human communication in order to verify the truth of a given hypothesis. With the rapid growth of digital communication and advances in NLP mining techniques, the subject of argumentation becomes important to automate the recognition of the argumentative structure of a document (e.g. premises and claims/conclusions) to be used in NLP applications. This topic is discussed here as this research mainly addresses contradictions that occur between research claims; it is therefore important to understand the argumentative structure of biomedical abstracts in order to be able to extract claims sentences.

2.2 Contradiction

The discovery of contradictory statements is important to support NLP applications. Question Answering systems for instance (generally) provide a ranked list of the N top candidate answers to a particular question and disregard the others, even if they are

contradictory. However, this type of system should be capable of returning answers that display semantic values that differ to the question.

For example, Table (2.1) shows a question (1), which has two answers: answer (2), which denies that Pakistan has tested Shaheen-2 missiles, and answer (3), which shows that Pakistan has performed such testing in 2004. Although they both are lexically similar, they provide opposite answers to the same question. The ability to detect such phenomena in information access applications is important to enable further investigation to validate such information (Harabagiu, Hickl, and Lacatusu, 2006a). Condoravdi et al. (2003) recognised the importance of detecting contradiction in text, and considered this as a minimum necessary criterion to understand language.

	Text
1	When did Pakistan test its Shaheen-2 ballistic missiles?
2	The source noted that the Shaheen-2 with a range of 2,400 km, has never been tested by Pakistan
3	Pakistan has performed several tests of its Shaheen-2 missiles in 2004

TABLE 2.1: An example of contradictory answers to a question (Harabagiu, Hickl, and Lacatusu, 2006a)

Harabagiu, Hickl, and Lacatusu (2006a) stated that contradiction between texts occurs when information is incompatible, or when one text asserts a proposition and the other negates it. Moreover, they showed that the recognition of contradiction can be achieved by two methods: the first method is to measure the textual entailment (see Section (2.3)) between texts after removing negation propositions; if entailment holds true, then the pair is considered contradictory. The second method involves derivation of linguistic features such as negations, contrasts and antonyms.

Marneffe, Rafferty, and Manning (2008) provided a looser definition whereby two sentences are considered contradictory when they involve the same event but are extremely unlikely to be true at the same time. An example of this is *Sally sold a boat to John* and *John sold a boat to Sally*. The corpus annotation guidelines in that research, showed that a pair of texts T and H is annotated as contradictory if the assertion in the hypothesis H appears to directly refute portions of the text T .

Bowman et al. (2015a) discussed the indeterminacies of event coreference and entity coreference in the definition of contradiction and its impact on those who provide annotation for inference resources. They showed that sentences such as *A boat sank in the Pacific Ocean* and *A boat sank in the Atlantic Ocean* can be considered contradictory, if the annotator assumes that both describe the same single event. However, it is remains reasonable that the sentences be annotated as neutral (not contradictory) if that

assumption is not considered. If that assumption is made, however, counter-intuitive contradictions can be found. For example, sentences such as *Ruth Bader Ginsburg was appointed to the US Supreme Court*, and *I had a sandwich for lunch today*, would be annotated as contradictory since they do not meet the assumption of statements referring to the same event.

2.2.1 Contradiction Typology

Marneffe, Rafferty, and Manning (2008) described various linguistic constructions that lead to contradiction: antonym, negation, numeric, factive, structure, lexical and world knowledge.

Antonyms are pairs of words with opposite meanings, for example *good* and *bad*. They are applicable to both gradable adjectives such as *hot* and *cold*, or non-gradable such as *life* and *death* (Mohammad, Dorr, and Hirst, 2008). WordNet, an important resource for antonyms, contains more than 7,000 antonym relationships. It groups lexical items into sets of synonyms called synsets; thus, if a pair of antonyms belongs to different synsets (*A* and *B*, for instance); then every word in synset *A* can be considered an antonym to every word in synset *B* (Harabagiu and Moldovan, 1998). Antonyms have been used as a feature to recognise contrastive information between text which may consequently cause contradiction, as in the sentences (4) and (5) in Table (2.2).

	Text
4	Capital punishment is a <i>catalyst</i> for more crime
5	Capital punishment is a <i>deterrent</i> to crime

TABLE 2.2: An example of contradictory statements due to antonym used by (Marneffe, Rafferty, and Manning, 2008)

Harabagiu, Hickl, and Lacatusu (2006a) used the antonyms available in WordNet, in addition to other lexical items that take the form of IS-A relationships with the antonyms and their definitions, to recognise contrastive information in aligned predicate-arguments. Marneffe, Rafferty, and Manning (2008) followed similar approach, but also included oppositional verbs from VerbOcean (Chklovski and Pantel, 2004) to identify contradictory text. That research showed that contradiction due to antonyms could be readily detected using automated methods. Andrade et al. (2013) and Kawahara, Inui, and Kurohashi (2010) used antonyms to detect contrastive information for the purpose of discovering contradiction in Japanese text. Furthermore, Pham, Nguyen, and Shimazu (2013) used antonyms to identify contradictory information in aligned semantic frames rather than predicated arguments in order to detect contradiction in text.

Negation is another indicator for contradiction. Harabagiu, Hickl, and Lacatusu (2006a) showed two types of negation markers: directly licensed such as *don't*, *no* and *never*, and indirectly licensed such as *deny*, *fail* and *refuse*. These markers were used to detect different types of negations including: negated events and negated entities.

Negated events are predicates that fall within the syntactic scope of the negation term existing in a predicate field. For example, Table (2.3) shows two relation tuples, the first one was extracted from sentence (2) above; because the predicate *tested* is within the scope of the negation term *never*, which covers the entire predicate field, the predicate is annotated with false truth value. The second relation tuple was extracted from sentence (7) in table (2.4), which shows that the scope of negation terms in the noun phrases covers its noun phrase rather than the entire argument field.

L-Argument	Predicate	R-Argument
the Shaheen-2	had never been tested by	Pakistan
juries and not judges	must impose	a death sentence

TABLE 2.3: The first relation tuple is an example of a negated event (*tested*). the second relation tuple is an example of a negated entity (*judge*) used by (Harabagiu, Hickl, and Lacatusu, 2006a).

	Text
6	The Supreme Court decided that only <i>judges</i> can impose the death sentences
7	A closely divided Supreme Court said that juries and <i>not judges</i> must impose a death sentence

TABLE 2.4: An example of contradictory statements due to negation used by (Harabagiu, Hickl, and Lacatusu, 2006a)

Marneffe, Rafferty, and Manning (2008) used negation markers to detect polarity differences in dependency graphs. The scope of negation in a dependency graph covers nodes that have negation dependency; however, further checkpoints are included to ensure that these nodes are not antonyms. Note that the negation scope was important for the detection of negation in statements such as *no bullet penetrated* and *the bullet did not penetrate*. Andrade et al. (2013), Kawahara, Inui, and Kurohashi (2010), and Pham, Nguyen, and Shimazu (2013) used negation to detect contradiction in Japanese text.

Numeric or temporal mismatch is another indicator for contradiction. Table (2.5) shows sentences (8) and (9), which are contradictory due to a mismatch in death numbers. Marneffe, Rafferty, and Manning (2008) normalised numeric information within text into ranges in order to detect contradiction; for example, *over 100* and *170* is not considered a mismatch.

	Text
8	The tragedy of the terrorist attack in Paris that killed 130 civilians has left Frances leadership facing a dilemma
9	An investigation into the incident in Paris found 89 confirmed dead thus far

TABLE 2.5: An example of contradictory statements due to number mismatch used by (Marneffe, Rafferty, and Manning, 2008)

Temporal expression is another term for textual phrases describing a potentially complex time, date or duration. Andrade et al. (2013) calculated the cost of the minimum alignment between normalised temporal expressions as a feature to generate a machine learning system for identifying contradiction with Japanese text.

Factivity is an implicit assumption about the truth of a certain fact which may cause contradiction. Marneffe, Rafferty, and Manning (2008) used three types of factivity verbs to indicate contradiction: factive verbs, implicative verbs and non-factive verb. An example of factive verbs is *realised* in “John realised that he was in debt”, as the sentence presupposes that John was in fact in debt; an example of an implicative verb is *forget* in “Bill forgot to take his wallet”, which presupposes that Bill did not take his wallet; and an example of a non-factive verbs is *believe* in “Bill believed that he took his wallet”, which does not presuppose any assumption about the information provided. The verbs embedded in such sentences can cause contradiction. For example, Table (2.6) shows that sentence (10) contradicts sentence (11), however, sentence (11) does not contradict sentence (12) (Marneffe, Rafferty, and Manning, 2008).

	Text
10	Bill forgot to take his wallet
11	Bill took his wallet
12	Bill did not forget to take his wallet

TABLE 2.6: Examples of contradictory statements due to factivity used by (Marneffe, Rafferty, and Manning, 2008)

Modality is a type of expression used to express possibility and necessity. For example, lexical items such as *may*, *might*, *can* and *could* are called possibility modals, while *must*, *should* and *have to* are necessity modals. MacCartney et al. (2006) used six modality markers (possible, not possible, actual, not actual, necessary, not necessary) to create entailment rules and map these rules onto judgements: *yes*, *weak yes*, *no*, *weak no* and *dont know*. The rules were used to test pairs of modalities extracted from two expressions. For example, in Table (2.7), rule (13) shows *not possible* entails *not actual* has the value *yes*, which is entailment, whilst rule (14) is *weak no*. Marneffe, Rafferty, and

Manning (2008) used the same approach to create rules for contradiction, for example, rule (15) in the same table, shows *possible* entails *not possible* is true.

	Text
13	$(\textit{not possible} \models \textit{not actual})? \implies \textit{yes}$
14	$(\textit{possible} \models \textit{necessary})? \implies \textit{weak no}$
15	$(\textit{possible} \models \textit{not possible})? \implies \textit{yes}$

TABLE 2.7: Modality rules for textual entailments and contradiction used by (MacCartney et al., 2006)

Lexical discrepancies may give rise to contradiction, as demonstrated in sentences (16) and (17) in Table (2.8). Similarly, syntactic structures may also lead to contradiction; for example, the subject of sentence (18) in the same table overlaps with the object of sentence (19).

	Text
16	Bush called for U.S troop to be withdrawn from the Balkans
17	He cites such missions as example of how America must “stay the course”
18	<i>Jacques Santer</i> succeeded <i>Jaques Delor</i> as president of the European Commission 1995
19	<i>Delors</i> succeeded <i>Santer</i> in the presidency of the European Commission

TABLE 2.8: Examples of contradictory statements due to lexical (16 & 17) and syntactical structure (18 & 19) used by (Marneffe, Rafferty, and Manning, 2008)

Background knowledge is a key factor for the detection of contradiction by human intuition. Marneffe, Rafferty, and Manning (2008) considered background knowledge as essential to remove pairs of texts that described non-coreferent events in order to detect contradiction. For instance, *Pluto’s moon* in statement (20) differs from *The moon Titan* in statement (21). Ritter et al. (2008) showed that background knowledge was important for distinguishing between likely contradiction from genuine contradiction. For example, statements (22) and (23) appear contradictory, as *Salzburg* differs lexically from *Austria*. However, if we know that *Salzburg* is a city in Austria, we can determine that statement (22) does not contradict (23).

	Text
20	Pluto’s moon, which is only about 25 miles in diameter, was photographed 13 years ago
21	The moon Titan has a diameter of 5100 km
22	Mozart was born in Salzburg
23	Mozart was born in Austria

TABLE 2.9: Examples of apparent contradictory statements due to background used by (Ritter et al., 2008)

2.2.2 Contradiction Recognition Approaches

Harabagiu, Hickl, and Lacatusu (2006a) demonstrated a contradiction detection system relying on three forms of linguistic information, negation, antonyms, and contrast of discourse relations. They considered two methods in order to detect contradiction. The first method is the removal of negation markers, textual inputs are passed to an alignment module which takes advantage of lexical alignment and textual paraphrases (Hickl et al., 2006). The alignment module was designed for a textual entailment system, which was based on the assumption that if T entails H , they can paraphrase each other. In the second method, the predicate-argument structure of texts were extracted. Next, an alignment module was used to identify negations, antonyms, contrasts, dependencies and semantics features. These features were used to train a classifier to detect contradiction in the same way textual entailment is detected.

	Text
24	John Bok, who has been on a hunger strike since Monday, says he wants to increase pressure on Stanslav Gross to resign
25	A hunger strike was attempted
26	A hunger strike was <i>not</i> attempted
27	A hunger strike was <i>called off</i>

TABLE 2.10: Examples of the contradictory sentences used by Harabagiu, Hickl, and Lacatusu (2006a) dataset

That research used the RTE2 dataset (see Section (2.3)), which consists of 1,600 pairs of sentences, 800 pairs of which were annotated as true entailments and the other 800 were not. For example, Statements (24) and (25) in Table (2.10) were annotated as entailment. In order to convert that dataset into one that could be used for contradiction, two annotators were asked to negate one sentence from pairs annotated as entailment in order to produce contradiction due to negation, as shown in statement (26). Similarly, the same task was repeated but the negated sentences were paraphrased in order

to produce contradiction due to contrast, as in statement (27). The result of that process was 400 contradictory pairs. Table (2.11) shows the system performance using different features.

Features	Accuracy	Precision
Negation only	75.63 %	68.07 %
Paraphrase only	62.55 %	67.35 %
Negation and Paraphrase	64.0 %	75.74 %

TABLE 2.11: The system performance achieved by Harabagiu, Hickl, and Lacatusu (2006a)

Subsequent work by Marneffe, Rafferty, and Manning (2008) demonstrated that a contradiction detection system requires more refined distinctions than the textual entailment systems could provide since some types of contradiction may require even deeper inference to identify the capability of such a system. In that work, they converted a pair of texts H and T to their representations in dependency graphs using the Stanford parser (Marneffe and Manning, 2008). Collocations and named entities were collapsed to a single node to increase the semantic representation of the graphs. Next, each node in the hypothesis graph was mapped to a unique node in the text according to its similarity score with that node, using the MIRA algorithm (Crammer and Singer, 2003), or otherwise assigned as null. In the next stage, pairs of text were ensured to represent the same event (coreference event), with the purpose of removing non-coreferent pairs. Event coreference was measured by computing the topicality of the hypothesis and text, which were considered related if their scores were above a tuned threshold. This was important as textual contradiction usually requires that the pair of texts describes the same event, which is different from textual entailment tasks which only consider hypotheses that are not supported by text as non-entailment.

The system described in that work was trained on RTE1 and RTE2 corpora in addition to 131 contradictory pairs extracted from the general domain to reflect ‘real life’ contradiction. Several features were extracted from these corpora, including polarity, number/date/time, antonyms, structure, factivity, modality and relational features (see Section 2.2.1). The system performance varied according to the dataset used for evaluation, but a clear decline in performance was typically observed during testing of a new dataset, demonstrating the complexity of the contradiction task and the difficulty of constructing a broad coverage system that could operate on different domains.

Contradiction Type	RTE3_dev	RTE3_test
Antonym	25.0 %	42.9%
Negation	71.4 %	60.0 %
Numeric	71.4 %	28.6%
Factive/Modal	25.0%	10.0%
Structure	64.2 %	21.1%
Lexical	13.3 %	0.0%
Background	18.2%	8.3%

TABLE 2.12: The system performance achieved by Marneffe, Rafferty, and Manning (2008)

Table (2.12) illustrates recall according to contradiction type. The scores of the first three types of contradictions were better than the performances of the other types. Moreover, the lowest performance by the system was related to contradiction, as a result of lexical and background knowledge. The authors reported that the detection of contradiction from number mismatch was relatively straightforward, although it was difficult to achieve a high precision score due to challenges in distinguishing between the meanings of numbers, e.g. *10% increase* versus *4 million increase*. Contradiction was demonstrated to be key in text comprehension, whereby certain aspects of contradiction can be automatically resolved while others require additional investigation and research.

Ritter et al. (2008) investigated contradiction over functional relationships such as $BornIn(Person) = Place$. In that relationship, the *BornIn* relation maps peoples names to their unique birthplace. However, other relationships such as *visited* do not reflect such uniqueness. That work used the web to automatically generate seemingly contradictory pairs rather than using manually chosen sentences. Various reasons may cause apparent contradictions, including synonyms like *Mrs. Bush* and *Laura Bush*, hypernyms like *renal failure* and *kidney disease*, and ambiguity such as *Mr. Smith* in *Mr. Smith was born in 1990*.

In contrast to the work of Marneffe, Rafferty, and Manning (2008), who considered that T and H were contradictory when T entailed the negation of H as in (28) in Table (2.13), Ritter et al. (2008) suggested that T and H alone are mutually consistent as in (29), and that contradiction could only be detected with the benefit of background knowledge (B), as in (30) in the same table.

	Rule
28	$T \models \neg H$
29	$T \not\models \neg H \wedge H \not\models \neg T$
30	$((B \wedge T) \models \neg H) \vee ((B \wedge H) \models \neg T)$

TABLE 2.13: Assumptions considered in Ritter et al. (2008) system

Relation tuples were extracted from text using *TextRunner* system (Banko et al., 2007), which generates tuples in the form of $R(x,y)$ as shown in Table (2.14), where R is the predicate and X and Y are the arguments. The system developed by Ritter et al. (2008) considered tuple (31) as an example of a functional relationship since the distribution of Y (PLACE) to X (Mozart) is roughly unambiguous. Tuple (32) is an example of a non-functional relation due the ambiguity of the values of Y (PLACE). Tuple (33) is an example of a functional relation, but the system developed by Ritter et al. (2008) considered it as non-functional due to the distribution of Y (PLACE) to X (Mozart).

	Relation
31	was_born_in(Mozart, PLACE): Salzburg(66), Germany(3), Vienna(1)
32	lived_in(Mozart, PLACE): Vienna(20), Prague(13), Salzburg(5)
33	was_born_in(John Adams, PLACE): Braintree(12), Quincy(10), Worcester(8)

TABLE 2.14: Contradiction over functional relationship by (Ritter et al., 2008)

The system described by Ritter et al. (2008) consists of three components: *Extractor* to extract relation tuples, *Functional learner* to learn functional relations and *Contradiction detector* to detect which pairs of relationships represent genuine contradictions. The authors found that about half of the errors (49%) generated by the system occurred in tuples where the distribution of the argument (Y) to argument (X) was ambiguous. For example, an error occurs with tuple (33), which shows that the distribution of the places (i.e. Braintree, Quincy and Worcester) are roughly ambiguous to *John Adams*. Furthermore, a considerable amount of errors (34%) were due to missing meronyms and synonyms (14%). These results suggested that background knowledge is an important element in the contradiction detection task. Moreover, lexical resources beyond WordNet and the Tipster Gazetteer (Gee, 1998) seem important for such task.

Pham, Nguyen, and Shimazu (2013) integrated a shallow semantic representation

with binary relationships to identify contradiction. That system consists of two modules. The first module considers contradiction occurs when an event described in a semantic role labelling (SRL) frame in H is incompatible with an event described in an SRL frame in T . The authors used *SENNA* (Collobert et al., 2011) to extract SRL frames from H and T , where each SRL frame consists of a verb (predicate) and a list of SRL elements. The SRL frames of H and T were subsequently aligned by calculating the similarity of the two SRL elements using the local lexical level matching method described by Dagangan and Massimo (2007), along with coreference resolution using a unified term to describe the equivalent elements.

Following the alignment stage, incompatibility between the two frames was measured using 1) relatedness measures, including WordNet and WordNet::Similarity (Pedersen and Patwardhan, 2004), to measure the relatedness between verbs, and 2) local lexical level matching to measure the relatedness between SRL elements. Contradiction was considered to arise when the relatedness between the two SRL frames was below a certain threshold.

In the second module, contradictions occur when incompatible pairs of relation tuples from H and T are found after they were extracted using ReVerb (Fader, Soderland, and Etzioni, 2011). Thus, the task of the second module was to search for incompatible tuples from pairs of texts. Multiple criteria were considered to measure incompatibility, including the mismatch of relations due to antonyms, the mismatch of the second argument if the first argument and the relation matched, the roles of the arguments were exchanged, or the arguments have the same types (number, date etc.) but different values.

That work used three datasets, RTE-3, RTE-4 and RTE-5. These sets were annotated to identify pairs that represented *entailment*, *unknown* and *contradiction*. The non-contradiction pairs were re-annotated as *non-contradiction*. Table (2.15) shows the result of system performance using each module separately and in combination. The SRL-based module was shown to outperform the relation-based module, an outcome which may have arisen from the use of shallow semantic representations which provided greater information than the relation tuple approach.

	RTE-3			RTE-4			RTE-5		
	P	R	F1	P	R	F1	P	R	F1
SRL-based	13.4	15.27	14.28	22.41	17.33	19.55	22.72	16.67	19.23
Relation-based	22.58	9.72	13.59	26.3	10.0	14.49	19.48	16.67	17.96
Combination	14.0	19.44	16.27	23.0	23.67	22.82	21.14	28.89	24.4

TABLE 2.15: Pham, Nguyen, and Shimazu (2013) system performance

2.2.3 Contradiction in Biomedical Domain

In the biomedical domain, Sanchez-Graillet and Poesio (2007) conducted a preliminary study on the discovery of contradictions about protein-protein interaction. These occur when an author argues that protein *X* interacts with protein *Y*, while another argues that protein *X* does not interact with protein *Y*. Contradiction is detected by mapping each text to a semantic representation, each of which consists of multiple attributes. Contradiction between statements is confirmed when an incompatibility between attributes values is found. These attributes are: *protein names*, *cue word* and *manner*. The *cue word* is a word expressing the interaction between the proteins, and includes verbs and their normalisations. A cue word has three attributes; a *semantic relation* with multiple possible values including: activate (e.g. *transactivate*) or inactivate (e.g. *decrease*), *polarity* to describe whether the interaction is positive or negative and *direction* to show direction of the interaction; e.g. + (e.g. *generate*), - (e.g. *release*) or neutral (e.g. *substitute*). The *manner* attribute is an adjective or adverb which affects the direction attribute, such as terms which imply speculation, e.g. the term potential in “*there is a potential interaction*”; and consists of one attribute - *manner polarity* - to indicate whether the manner has a positive, negative or neutral effect on a cue word.

The study evaluated thirty-one pairs of sentences from articles extracted from the Journal of Biological Chemistry. The pairs of sentences were annotated by the system, biologists and non-biologists. The kappa score (inter-annotator agreement) among the biologists was 37%, and similar score was obtained between the automatic system and biologists, however, the kappa score among the non-biologists was 22%, and between the automatic system and non-biologists was 19%. That research showed that the system achieved similar performance to biologists and outperformed non-biologists. One of the reported findings was that biologists concluded that protein-protein contradictions are rare.

Sarafraz (2011) developed a system to identify conflicting biomedical events in biomedical text. The event, in this case, is a chemical interaction between certain types of organic molecules, where conflicts may occur due to contrast or contradiction. That system defined contradiction as two events which share (1) interaction type, (2) cause, (3) theme, (4) anatomical location, and are both (5) assertive, but have different (6) polarity. For example, Table (2.16) shows statements (34) and (35) are contradictory as they share all these attributes and the second statement negates the first one by using the negation mark *no*. The semantic representations of statements (34) and (35) are shown in Figure (2.1). Contrastive events were defined in a manner similar to contradiction but the cause, theme and anatomical location are not required to be the same in both

events, such as statements (36) and (37).

	Relation
34	Positive regulation of CXCR4 expression interleukin-7 in CD4+ mature thymocytes correlates with their capacity to favor human immunodeficiency X4 virus replication
35	In contrast, in intermediate CD4(+) CD8(-) CD3(-) thymocytes, the other subpopulation known to allow virus replication, TEC or IL-7 has little or no effect on CXCR4 expression and signaling
36	In addition, cloning efficiencies were acceptable (over 30%) when IL 2 produced spontaneously from the leukaemic cell Jurkat (M-N) was used
37	However, IL-2 is not normally synthesized by solid tumor cells

TABLE 2.16: Sentences (34 & 35) are examples of contradictory statements and (36 & 37) are examples of contrastive statements used by Sarafranz (2011)

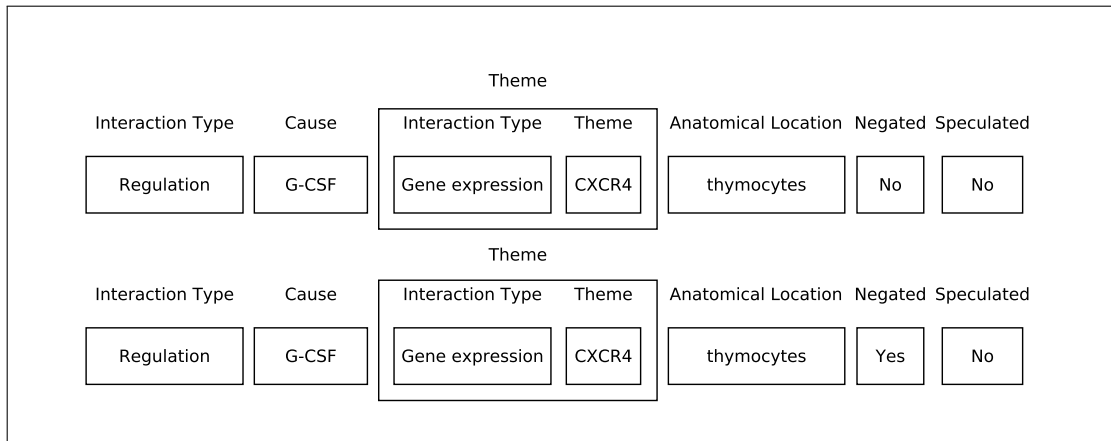


FIGURE 2.1: The semantic representation of statements (34) and (35)

Strict and relaxed modes of conflict between events were considered in that system; for example, if two events shared all attributes except that the anatomical location was unknown and polarity was different, they become contradictory in a relaxed mode but not in a strict mode. The system was constructed using rule-based and machine learning techniques to extract events and their contexts from the biomedical text; events that negated other events under specific conditions were subsequently identified as contradictory or contrasting. The corpus used in that work was constructed by combining the events in the BioNLP09 corpus with a subset of the events of the GENIA corpus, where the newly-constructed corpus contained the attributes described in the definition of contradiction. Two main contributions described in that work are: methods to extract biomedical events and methods to extract negated events.

2.3 Paraphrasing and Textual Entailment Recognition

The variability of language proves challenging for NLP tasks which require language understanding, particularly when the same meaning can be inferred from different texts. For instance, in QA systems, it is important to recognise when the meaning of a text can be inferred from the meaning of another text (textual entailment), and furthermore, when two fragments of text contain almost the same meaning (paraphrasing).

Textual entailment is defined as a directional relationship between the entailing *Text* T and the entailed *Hypothesis* H . Dagan, Glickman, and Magnini (2005) stated that “We say that T entails H if, typically, a human reading T would infer that H is most likely true”. Paraphrase recognition can be defined as the process of identifying that a pair of texts conveys almost the same meaning. Table (2.17) shows that statement (38) can be inferred from statements (39), (40) and (41) (textual entailment) and that statement (39) is a paraphrase of statements (40) and (41). Textual entailment is (uni-)directional relationship. For example, while statements (40) and (41) cannot be inferred from (38), statement (38) can be inferred from statements (40) and (41). In contrast, paraphrasing is a bidirectional textual entailment in that T entails H and, simultaneously, H entails T as in statements (40) and (41).

	Text
38	Shakespeare is a writer
39	William Shakespeare wrote Romeo and Juliet
40	Romeo and Juliet was written by William Shakespeare
41	Shakespeare is the writer of Romeo and Juliet

TABLE 2.17: Examples of paraphrases and entailments generated by Androutsopoulos and Malakasiotis (2010)

Textual entailment topic has been investigated for some time within the NLP field following its establishment through a series of workshops known as PASCAL Recognizing Textual Entailment (RTE) Challenges (RTE1-RTE7) (Dagan, Glickman, and Magnini, 2005). Multiple systems which require a pair of texts or templates as input have been used for RTE tasks, with the output presented as a judgement or probability to state whether one text entails another. The workshops RTE4-RTE7 implemented a three-way entailment decision, rather than a binary decision as in RTE1-RTE3. A three-way decision requires systems to determine whether a text T entails H (entailment), T contradicts H (contradiction), or T does not contradict or entail H (unknown). The NLP literature includes much research on paraphrasing and textual entailment tasks. The following sections provide more details on these two tasks, including definitions and approaches to recognition.

Both tasks may operate using templates of expressions, which usually contain slots filled with arbitrary nouns or noun phrases, or with a specified syntactic or semantic category if required (Androutsopoulos and Malakasiotis, 2010). Templates, as in Table (2.18), are important since information extraction systems often use such patterns to identify information of a particular type to extract the entities involved (Grishman, 2003; Moens, 2006). Templates (42-44) are suitable for paraphrasing, and can also be used as T and (45) as a hypothesis H for textual entailment. However, many current recognition approaches use other methods that rely on lexical, syntactical, semantic representations and logic (particularly in entailment) to recognise paraphrasing and textual entailment.

	Text
42	X wrote Y
43	Y was written by X
44	X is the writer of Y
45	X is a writer

TABLE 2.18: Examples of paraphrases and entailments templates by Androutsopoulos and Malakasiotis (2010)

2.3.1 Methods Based on Lexical-Syntactic Similarity

Several research studies have used similarity measures to recognise paraphrases and textual entailment. The three main approaches to this type are lexical, syntactic, and semantic similarity.

For lexical similarity, Wan et al. (2006) used a metric imported from machine translation tasks (Papineni et al., 2002) to determine the similarity between a pair of texts for paraphrasing. That approach used the precision measure, but on n-grams in order to compute the similarity as shown in (Equation 2.1).

$$precision = \frac{ngrams - overlap(s_1, s_2)}{ngrams - counts(s_1)} \quad (2.1)$$

Malakasiotis and Androutsopoulos (2007) attempted several similarity measures to recognise whether the entailment held true between T and H . These measures included Levenshtein distance (Levenshtein, 1966), Euclidean distance and cosine similarity (Manning, Raghavan, and Schütze, 2008). That approach must consider other factors that can potentially influence the decision of the system such as negation, which may cause T or H to not retain the truth value.

Another approach is to measure the similarity between a sliding window in T that is of the same size of H , where T in the textual entailment task is typically longer than H . The greatest similarity between H and a particular window of T is a good indication of entailment, as exemplified in statements (46-48) in Table (2.19). However, the use of fixed window size may not be suitable for entailment detection in cases such as (46) and (47). An attempt to resolve that problem Burchardt et al. (2009) aligned words in T with words in H and then used the shortest span of T that contains the words in H for similarity measurement.

	Text
46	<u>John Kennedy was assassinated</u> in Dallas on November 22, <u>1963</u> , while on a trip to Texas to smooth over frictions between ..
47	<u>John Kennedy</u> was on a trip to Texas to smooth over frictions between some politicians before he was <u>assassinated</u> on November 22, <u>1963</u> ...
48	John Kennedy was assassinated in 1963

TABLE 2.19: Textual entailment using lexical similarity by Malakasiotis and Androutsopoulos (2007)

Lexical similarity is useful for paraphrasing and textual entailment tasks, however, it has some limitations. For example, it does not take into account the syntactic characteristics of words within contexts. Statements (49) and (50) in Table (2.20), for instance,

have a high similarity score when using a sliding window of size H words, which consequently may mislead the judgement of paraphrasing or textual entailment recognisers. One approach which can be used to overcome such limitations is to assess the similarity between texts at a syntactic level, at which multiple representations can be found.

	Text
49	The national Institute of Psychology in Israel was established in 1979
50	Israel was established in 1979

TABLE 2.20: Textual entailment using syntactical similarity by Malakasiotis and Androutsopoulos (2007)

Wan et al. (2006) and Malakasiotis (2009) used a dependency relation tuples representation to count the common dependency tuples between texts, in order to measure their similarity for paraphrasing where a similarity score above a certain threshold indicates potential paraphrasing. An alternative approach to that was to calculate the tree edit distance between their dependency trees (Zhang and Shasha, 1989). The tree edit distance is the minimum cost to compute the sequence of operations required to add, replace or remove a node or edge that transforms one tree into another, where each transformation operation is assigned an appropriate cost to achieve a reasonable result (Mehdad, 2009). For example, the cost to replace a word with one of its synonyms should be less than the cost of replacing it with an unrelated word (Haghighi, Ng, and Manning, 2005).

Figures (2.2) and (2.3), show the dependency tree and the grammatical dependency relations, respectively of (49) and (50), which show that the tree of statement (50) (H) has no similarity with any subtree or relation of (49) (T).

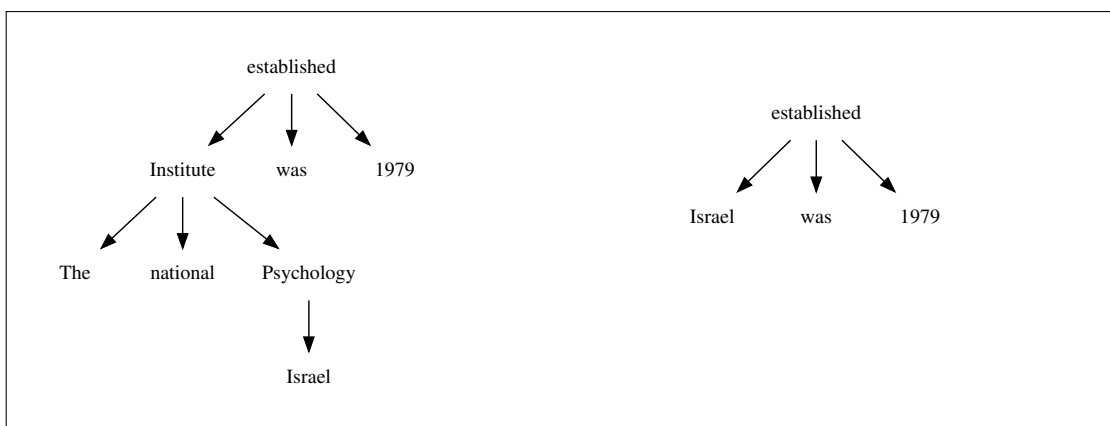


FIGURE 2.2: Dependency trees of statements (49) and (50) by Malakasiotis and Androutsopoulos (2007)

det(Institute, The)	nsubjpass(established, Israel)
amod(Institute, national)	auxpass(established, was)
nsubjpass(established, Institute)	root(ROOT, established)
case(Psychology, of)	case(1979, in)
nmod:of(Institute, Psychology)	nmod:in(established, 1979)
case(Israel, in)	
nmod:in(Psychology, Israel)	
auxpass(established, was)	
root(ROOT, established)	
case(1979, in)	
nmod:in(established, 1979)	

FIGURE 2.3: Grammatical dependency relations of statements (49) and (50)

2.3.2 Methods Based on Semantic Similarity

Although the comparison of text at syntactic level exploits information that cannot be detected at a lexical level, other information cannot be detected at that level, such as the relationships between synonyms and hypernyms (e.g. *arrest* & *capture*, or *computer* & *artifact*). It can be beneficial to exploit the semantic relationships between words within texts, in order to recognise paraphrasing and textual entailment. For example, Haghighi, Ng, and Manning (2005) used *WordNet::Similarity* (Pedersen and Patwardhan, 2004) to measure the semantic relationship between edges on tree graphs that do not correlate to each other, for example, *establish* and *found* in Figure (2.4).



FIGURE 2.4: Paraphrasing and textual entailment using semantic similarity by Malakasiotis (2009)

Semantic Roles Labelling (SRL) is an alternative approach to dependency trees, which can be used to identify words not corresponding to each other in order to measure their semantic similarity. SRL uses a predicate-argument structure to assign role labels to each argument associated with a predicate. For example, the SRL engine of Punyakanok, Roth, and Yih (2008) analysed the sentence *Mary sold the book to John, sell* as the predicate, *Mary* as the seller, the *book* as the item sold, and *John* as the buyer.

In this case it is straightforward to measure the semantic similarity between texts that belong to the same semantic role.

FrameNet (Baker, Fillmore, and Lowe, 1998) is an example of an application that functions as an SRL engine, and is formulated from the notion that the meaning of most words is based on semantic frames; for example, the concept of cooking usually involves a person who cooks (*Cook*), the food to be cooked (*Food*), a container to hold the food (*Container*), and a heat source (*Heating_Instrument*). In *FrameNet*, this is represented as *Apply_heat* frame, and the other elements involved in the cooking process are called frame elements (FEs).

Burchardt et al. (2007) used *FrameNet* to analyse texts according to their roles, and then used *WordNet* to measure the similarity between words that belonged to the same role. Several measures using *WordNet* can be considered in this case, such as those described by Leacock, Miller, and Chodorow (1998), Lin (1998a) and Resnik (1999). Such measures generally evaluate similarity based on the length of the path that connects two words (senses) in *WordNet*, and their frequency in a predefined collection of text; less commonly-used words tend to have a higher score since they typically represent important information.

2.3.3 Methods Based on Rules

Template-based rules can be used to recognise paraphrasing and textual entailment. DIRT (Discovery of Inference Rules from Text) (Lin and Pantel, 2001) is an example of a set of inference rules used in textual inference tasks, including paraphrasing and entailment. These rules were constructed based on an extended version of Harris's Distributional Hypothesis (DH), which states that words in the same context tend to have similar meanings. However, that work applied the rules on dependency trees, and the extended hypothesis becomes paths that connect the same set of words tend to have similar meanings. The output of that work was a set of inference rules that were extracted from a large corpus according to the extended hypothesis, for example, sentences (51) and (52) in Table (2.21).

	Text
51	X wrote $Y \approx X$ is the author of Y
52	X caused $Y \approx Y$ is blamed on X

TABLE 2.21: Paraphrasing and textual entailment using rules by Malakasiotis (2009)

Wang and Neumann (2007) used DIRT rules to recognise textual entailment without the use of an external knowledge resource (i.e. *WordNet*). Their method was based on applying the rules to tree skeletons rather than to a dependency tree. A tree skeleton is a subtree of the lowest common root nodes of two dependency trees of T and H , where the inner paths of the subtrees are ignored. In such a case, a textual entailment is detected when the left or right path of the tree skeleton (or the root nodes) contains a DIRT inference rule. Figure (2.5), for instance, illustrates the dependency trees of a pair of texts representing a textual entailment case.

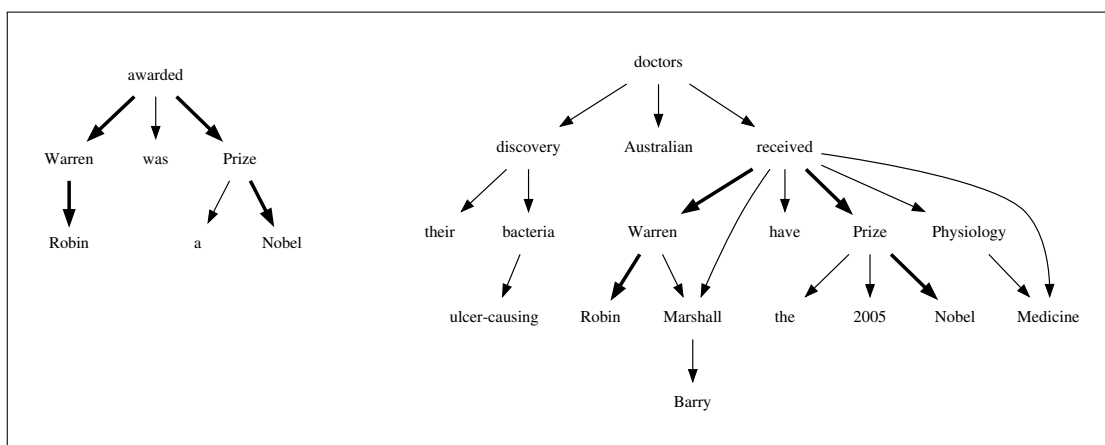


FIGURE 2.5: Tree skeletons used for textual entailment recognition

Dinu and Wang (2009) used the same set of rules in addition to other rules, which were generated by replacing words in the existing DIRT rules with their synonyms from *WordNet*. For example, *face* in rule (53) in Table (2.22) is a synonym of *confront*, therefore, a new rule (54) was generated from (53). That study found that only a portion of the DIRT rules were suitable for use as inference rules, and that 50% of these rules were lexical rules while the other 50% can be recognised using *WordNet*.

	Text
53	X face the threat of Y
54	X confront the threat of Y

TABLE 2.22: Generating new rules using synonyms for textual entailment recognition

Although they are not directional, DIRT rules are beneficial in textual entailment. In an attempt to identify the directionality of DIRT rules, Bhagat, Pantel, and Hovy (2007) developed an unsupervised algorithm called Learning Directionality of Inference Rules (LEDIR) to classify DIRT rules into four classes: paraphrases, $P1$ entails $P2$, $P2$ entails $P1$ and no plausible inference. Note that $P1$ and $P2$ were slot fillers that belong to

concepts such as person or location. The algorithm was based on a Directionality Hypothesis, which states that when two binary semantic relations exist in similar contexts and the first occurs in significantly more contexts than the second, then the second relation is more likely to imply the first relation and not vice versa, for example, rules (55) and (56) in Table (2.23). When someone likes something it does not necessary mean that it can be eaten, but when someone eats something, he/she most probably likes it.

	Text
55	$X \text{ eats } Y \iff X \text{ likes } Y$ (DIRT)
56	$X \text{ eats } Y \implies X \text{ likes } Y$ (LEDIR)

TABLE 2.23: The LEDIR rules of Bhagat, Pantel, and Hovy (2007)

2.3.4 Methods Based on Logic

Logic has been used to recognise paraphrase and textual entailment. The input in such an approach is a pair of texts (T, H) , that are mapped to first-order logic semantic representations such as Discourse Representation Structure (DRS) (Kamp and Reyle, 1993). Next, a theorem prover like *Vampire* (Riazanov and Voronkov, 2002) is applied to find proof that text T implies hypothesis H , using a set of rules extracted from external background knowledge resources B (such as *WordNet* or *FrameNet*). The prover finds a proof that $(\phi_T \wedge B) \models \phi_H$ where ϕ_t and ϕ_H are the logic semantic representations (DRS) of T and H ; and the background knowledge B is represented by a set of axioms where each represents a rule such as passive-active transformation or an is-a relationship (hypernym). As an example of such a rule, *assassinate* is a hypernym of *kill* in *WordNet*; therefore, axiom (57) in Table (2.24) is added to B to enable the prover to recognise the relationship between *assassinate* and *kill*.

	Text
57	$\forall_x \forall_y \text{assassinate}(x, y) \implies \text{kill}(x, y)$

TABLE 2.24: A rule for textual entailment using logic-based approach (Kamp and Reyle, 1993)

The logic-based approach usually uses background knowledge to recognize textual inferences. Bos and Markert (2005) used that approach to recognize textual entailment. However, they had to use other features, such as automated reasoning (Claessen and Sorensson, 2003) and shallow semantic features, to reduce the effect of using background knowledge on the system performance. Rinaldi et al. (2003) used a similar

approach for paraphrasing tasks to improve QA systems for finding answers to user questions.

2.3.5 Methods Based on Machine Learning

Machine learning algorithms have also been used for textual inference tasks. Each input expression in a pair of texts is represented as a vector that contains different similarity measures on different levels of text (lexical, syntactical and semantic), in addition to other features such as negation and modality. A machine learning algorithm is trained on manually classified vectors (entailment versus non-entailment, or paraphrasing versus non-paraphrasing) to build a classifier to classify unseen pairs of text.

Iftene and Balahur-Dobrescu (2007) used multiple features, such as edit tree distance, acronyms and negation, to calculate the transformation of H to T to recognise textual entailment. These features were extracted using several resources, including a syntactic analysis tool such as *Minipar* Lin, 1998b, in order to convert pairs of text to dependency trees; lexical resources such as *WordNet* were also used to recognise named entities along with semantic resources (such as *DIRT*) to identify paraphrases.

Burchardt and Frank (2006) showed a textual entailment system based on the lexical, syntactical and semantic overlap between text T and hypothesis H . The system relied on two main components: probabilistic LFG grammar (Cahill et al., 2004), and frame semantics using *FrameNet*. The overlap scores between T and H at various levels were used as features of a machine learning algorithm to measure their textual entailment.

Wang and Neumann (2007) exploited the syntactic representation of text without the use of any external knowledge resource to recognize textual entailment. That research used tree skeletons extracted from the dependency trees of T and H (see *Figure (2.5)*), whereby the bold paths represented the corresponding tree skeletons. Four main features were extracted from each tree skeleton: left spine difference (LSD), right spine difference (RSD), verb consistence (VC), and verb relation consistence (VRC). The left and the right spines represent the contents of the left and right paths obtained by the two tree skeletons; the VC feature shows whether the two root nodes of the trees were similar or different, and the VRC shows whether the relationships between the root nodes were contradictory. The system used an SVM algorithm to determine whether a pair of texts represents an entailment, non-entailment or contradiction. That research demonstrated that the skeleton features were useful in constructing a textual entailment system without the need to use external knowledge resources.

Bos and Markert (2005) used logic inference, in addition to other lexical and semantic features, to measure textual entailment. The logic-based feature was extracted using the *Vampire* theorem prover mentioned previously. If the prover finds that T implies H , entailment is present; if the prover finds these to be inconsistent, no-entailment is inferred. The prover output, in addition to other lexical and semantic features, was used as a feature in a machine learning algorithm (decision trees) to recognise textual entailment. The authors found that a reliance on logic-based features enhanced the system precision score but lowered the recall score, due to the lack of a suitable background knowledge resource for performing such a task.

2.3.6 Paraphrasing and Textual Entailment Corpora

Microsoft Research Paraphrase (MSRP) is a significant paraphrasing corpus. It consists of a collection of 5,801 pairs of sentences where each pair is annotated as either constituting a paraphrase or not. The corpus not only considers paraphrase pairs that are strictly semantical (such as those which have become paraphrases due to simple lexical synonym or local syntactic changes), but also considers complex paraphrases such as those illustrated in sentences (58) and (59) in Table (2.25). Such considerations have enriched the corpus with a high level of complex variation, the complexity of which was left to annotators to decide the degree of which can be considered as paraphrasing.

	Text
58	Charles O. Prince, 53, was named as Mr. Weills successor
59	Mr. Weills longtime confidant, Charles O. Prince, 53, was named as his successor

TABLE 2.25: Examples of paraphrases sentences from MSRP corpus

For textual entailment tasks, the RTE-1 to RTE-7 datasets are widely used benchmarks. The RTE datasets were constructed to reflect several application scenarios, including QA, relation extraction, information retrieval and multi-document summarisation, to apply automated entailment judgement. In a QA system, for example, the entailment judgement was based on ensuring that the candidate answer was entailed by the corresponding text passage; similarly, for relation extraction, the aim was to ensure that an extracted relation was indeed entailed from the corresponding text.

With the exception of RTE-4, the RTE-1 to RTE-5 datasets include distinct development and testing sets, with each set containing from 600 to 800 entailment pairs. RTE-4 represents a single set, consisting of 1,000 pairs. All datasets were balanced (i.e. 50% YES, 50% NO), and the negative entailments in RTE-4 and RTE-5 were further divided

into two categories: Unknown and Contradiction, a process which introduced the contradiction detection task.

The distribution balance, which was artificially created by the annotators in the RTE-1 to RTE-5 datasets, rendered them unnatural for actual NLP applications. To address that issue, RTE-6 and RTE-7 implemented another approach to generate a new dataset that considered the actual distribution of entailment in real life documents. That dataset was constructed from previous multi-document summarisation systems used in Text Analysis Conference (TAC) (Dang and Owczarzak, 2008). Such systems use multiple clusters of documents on different topics to automatically produce short summaries, whereas RTE-6 and RTE-7 datasets use the summaries to generate the hypothesis dataset. Sentence (60), for example, was broken into smaller units and rephrased as standalone sentences to be used as hypotheses such as (61-63); the RTE task therefore becomes the recognition of sentences in document clusters that are entailed by hypotheses.

	Text
60	Merck, the maker of Vioxx, which was approved by the FDA in 1999, voluntarily took the drug off the market in September. [Summary sentence]
61	Merck is the maker of Vioxx
62	Vioxx was approved by the FDA in 1999
63	Merck withdrew Vioxx from the market

TABLE 2.26: Examples of instances from RTE-6/RTE-7 corpus

Recently, Bowman et al. (2015a) released the *Stanford Natural Language Inference corpus*, developed for textual inference tasks such as entailment and contradiction. The corpus consists of 570,152 pairs, making it suitable for use in neural network-based models. The pairs were annotated by five individuals using three labels: entailment, neutral and contradiction, and the gold standard label was selected by at least three annotators (see Table (2.27)). If a pair with no such agreement exists, as was the case for approximately 2% of the entire corpus, it is labelled with “-”.

T	Annotation	H
A man inspects the uniform of a figure in some East Asian country.	contradiction CCCCC	The man is sleeping.
A soccer game with multiple males playing.	entailment EEEEEE	Some men are playing a sport.
An older and younger man smiling	neutral NNNECN	Two men are smiling and laughing at the cats playing on the floor.

TABLE 2.27: Examples of instances in Bowman et al. (2015a) corpus

2.4 Argumentation Mining

Argumentation is one of the central aspects of human communication that is used to convey tendency, attitude or opinion and to attempt to make the partner or reader accept or adopt the same attributes (Peldszus and Stede, 2013). Argumentation refers to the process of constructing a set of coherent and relevant arguments with the aim of arriving at a conclusion which conveys a certain opinion to the reader. With the arrival of computational argumentation, the term argumentation has been used to represent arguments and their interactions within texts, and to distinguish legitimate from invalid arguments. An argument, in the context of structured argumentation framework (Besnard et al., 2014), is defined as a set of propositions or statements that consist of a premise, a conclusion and an inference from the premise to the conclusion (Walton, 2009). The argumentation literature includes multiple proposals for structured argumentation, such as the Toulmin model (Toulmin, 2003) and Freeman model (Freeman, 2011). The abundance of electronic text and advances in computational linguistics have increased interest in the extraction of arguments from text and converting them into structured format for further analysis and processing, thereby forming the impetus for a new area of research called *Argumentation mining*.

Argumentation mining has been used to describe the process of automating the discovery of an argument in a document by identifying its units and the relationships between them (Mochales and Moens, 2011). It consists of multiple subtasks: text segmentation, which is mainly to split text into smaller fragments called argumentative discourse units; segment classification to identify the role of each discourse unit; and relation identification to identify the relationships between individual discourse units.

Researchers have used argumentation mining to investigate methods for automatic classification of sentences in different domain. For example, in the scientific domain to understand the role of sentences (Liakata et al., 2012; Teufel, 2010), in the law domain to differentiate between argument types (e.g. counter and rebuttal) (Moens et al., 2007), in online debates by adopting textual entailment methods (Cabrio and Villata, 2013), and in ideological debates to understand which argument is for or against (Somasundaran and Wiebe, 2010).

2.4.1 Argumentation Mining in Non-Scientific Text

Examples of argumentative text include news editorials and law reports, which tend to influence readers' opinions on certain topics. Bal and Saint-Dizier (2009) used news editorials to analyse argumentation structure and strength in order to determine their

inherent persuasiveness. The ultimate result of that work was to produce semantic representations of arguments in the interest of identifying their attitudes (positive vs negative) towards a specific topic. The semantic representation consists of a *root*, which is the conclusion of an argument, and one or more *support* statements to support the conclusion and its *relations*. The *root* consists of one attribute, polarity, which has three possible values: positive, negative or neutral. The *support statement* has multiple attributes to describe its characteristics such as *orientation_support*, which has two possible values (For and Against), and *persuasive_effect*, which has three possible values (Low, Average and High). The *support relation*, which represents the rhetorical relation between supports, has multiple possible types; for example, *contrast* to show a partial contradiction between two supports, and *paraphrase* to show an alternative approach to the support or conclusion. That study showed that such annotation was challenging at two levels: determination of the polarity of the conclusions and determination of support strength attributes.

Law reports are also a domain of interest in argumentation mining. The importance of such reports derives from the role that precedent plays in English law. Legal experts typically summarise these reports to facilitate rapid observation and examination for the benefit of students or other lawyers. Manual summarisation is a time consuming task, so the automation of this process is of importance in that context. Grover, Hachey, and Korycinski (2003) developed an automatic summarisation system to classify sentences in law reports according to their argumentative roles. They categorised arguments into three roles: *Background*, *Case* and *Own*. Moreover, each role could describe different subcategories; the background role, for example, has the following subtypes: *Precedent* to describe a previous case, and *Law* to describe whether the sentence contains public statutes.

2.4.2 Argumentation Mining in Scientific Text

Teufel and Moens (2002) applied rhetorical structure theory (RST) (Mann and Thompson, 1987), which states that adjacent segments of text hold a semantic relationship to each other, to identify the role of sentences within a scientific paper. The aim of that work was to improve information retrieval and support automatic text summarisation applications. They designed seven categories to annotate the sentences: *Aim*, *Background*, *Basis*, *Contrast*, *Other*, *Own* and *Textual* (Teufel, Siddharthan, and Batchelor (2009) added another 15 categories). That system implemented a supervised learning algorithm to categorise the sentences, and the system performance varied according to

the zone. For example, system performance upon identifying sentences that belonged to the *Contrast* zone was 26% and 86% for the *Own* zone.

Mizuta and Collier (2004) also adopted the same direction and produced an annotation scheme to classify biological articles into zones, although their work identified text which contained research findings and outcomes related to an author's own work.

Liakata (2010) used a scheme called Core Scientific Concepts (CoreScs) to annotate sentences in scientific articles into 11 categories: *Hypothesis*, *Motivation*, *Goal*, *Object*, *Background*, *Method*, *Experiment*, *Model*, *Observation*, *Result* and *Conclusion*. The main purpose of that scheme was to enable the identification of specific parts of text such as *Results* and *Conclusions*, in order to distinguish between positive and negative results and thus evaluate the confidence in any conclusion drawn.

Jimeno Yepes, Mork, and Aronson (2013) used a scheme adopted from the values used in *NlmCategory* field of MEDLINE XML files to convert unstructured abstracts into a structured format. They developed multiple classifiers to label sentences or paragraphs according to their argumentative roles. The results showed that the task of labelling paragraphs with their role achieved higher performance scores compared with the labelling of sentences.

Green (2015b) designed ten argumentation schemes from genetic research articles, semantically distinct in term of their premises and conclusions. For example, the scheme *Effect to Cause* occurs when event X is unknown, event Y is observed, a potential relationship between X and Y has been previously established, and the conclusion describes that X occurred and caused Y . Another scheme is *Failed to Observe effect of Hypothesised Cause*, which occurs when event X is hypothesised, event Y is not observed, a potential relationship between X and Y has been established, and the conclusion describes that X did not occur and did not cause Y . The author demonstrated that such schemes could be identified by adhering to prescribed annotation guidelines.

Blake (2010) developed an annotation scheme (claim framework) specifically for the portions of text that represent authors' claims. The framework used full-text articles to differentiate between three types of claims found in biomedical articles: explicit, implicit and under-specified claims (including observations, correlations and comparisons). That study identified claims by capturing four facets: two concepts, a nature of change, and the basis of a claim. The two concepts can function as agent-object, where the agent concept has a change influence on the object concept, the nature of change is captured by the change term, and the basis of change illustrates what the author measures to prove his/her claim.

Under that framework, claims were categorised by the number of facets they contain; for example, explicit claims must contain agents, objects and natures of change, while implicit claims only require agents and objects. For instance, sentence (64) in Table (2.28) is an explicit claim because *trauma* is the agent, *hematopoietic progenitor cells* is the object, and *increases* is the nature of change. Blake’s study used a supervised learning system to automate the process of detecting explicit claims. Multiple features were used in the system, including lexico-syntactic and semantic features (the directionality of the change term, e.g. increase represents \uparrow and inhibit represents \downarrow). The results of the study showed that explicit claims could be captured by combining semantic and syntactic features.

	Text
64	Trauma reportedly also increases the frequency of hematopoietic progenitor cells

TABLE 2.28: A claim that belongs the explicit type based on Blake (2010) scheme

Park and Blake (2012) extended that work by automating the detection of comparative claims (comparisons), one of the categories defined by the claim framework. A sentence describing at least one similarity or difference in relationship between two entities, such as that shown in sentence (65) in Table (2.29), was considered a comparative claim by the authors. Moreover, they followed Jindal and Liu’s (2006) work by categorising comparative claims into three types: gradable and non-gradable similarity, and non-gradable difference. A gradable claim usually expresses the order of entities with respect to a certain aspect, as shown in sentence (66). Non-gradable similarity claims only state the similarity between entities sentence (67) while non-gradable difference claims only state the differences between entities as in sentence (68).

	Text
65	The plasma concentration of nm23-H1 was higher in patients with AML than in normal controls (P = 0.0001)
66	The number of deaths was higher for rats treated with the Emulphor vehicle than with corn oil and increased with dose for both vehicle
67	Mean maternal body weight was similar between controls and treated group just prior to the beginning of dosing
68	Body weight gain and food consumption were not significantly different between groups

TABLE 2.29: Examples of comparative claims from (Park and Blake, 2012)

Several classifiers employing different algorithms were implemented in that work: Naive Bayes (NB), Support Vector Machine (SVM) and Bayesian Network (BN), using multiple features to exploit the characteristics of comparative claims. Examples of these are lexical features that show comparison (e.g. *more* and *less*), directionality (*increase* and *decrease*), and terms indicating similarity and difference along with other syntactic features. The study demonstrated that the accuracy and F1 scores of the BN algorithm were statistically higher than those obtained from NB and SVM; furthermore, they demonstrated that comparative claims were found in about 12% of the corpus sentences.

2.4.3 Argumentation Mining Corpora

Constructing annotated documents (corpora) for argumentation mining is a relatively complicated and possibly controversial task, because it requires a certain level of homogeneity and consistency in identifying the argument components, their boundaries and the relationships between them (Lippi and Torroni, 2015). The literature contains a few argumentation corpora that have been used in different fields.

For example, AraucariaDB (Reed and Moens, 2008) is a corpus that consists of a set of argumentative examples extracted from different sources such as newspaper editorials, judicial summaries and discussion boards, originating from different English-speaking regions to allow for a wide range of argumentative styles. The original examples were extracted and annotated using Araucaria, a graphical tool for argument structure analysis. The overall corpus consists of almost 4,000 atomic propositions and 1,500 premises.

The European Court of Human Rights (ECHR) corpus (Mochales and Moens, 2011) consists of 257 arguments distributed over 47 legal documents. The annotation scheme categorises argument elements into premises or conclusions, where the premise is categorised into either supporting or against.

NoDE (Cabrio and Villata, 2014) is a set which consists of 792 pairs of arguments extracted from a variety of sources including Debatepedia¹ and ProCon², and annotated to reflect the positive and negative relationships between them. Positive relationships represent the support relation in bipolar argumentation (Cayrol and Lagasquie-Schiex, 2005) and negative relationships represent the attack relation in Dung argumentation framework (Dung, 1995).

¹<http://idebate.org/debatabase>

²<http://www.procon.org>

Stab and Gurevych (2014) constructed a corpus from *essayforum*³, which consists of 90 persuasive essays written in English. The annotation scheme includes arguments and relations. Arguments consist of premises, major claim (the central argument of an essay) and claims (specific arguments discussed within a specific scope in the essay). Relations consists of support relation and attack relation, where a support relation occurs if the conclusion of the argument is equivalent to the argument's premisses and an attack relation occurs when the conclusion of the argument is contradictory to the premisses.

For biomedical text, Green (2014, 2015a,b) presented research on the construction of biomedical/biological corpora, which may be useful to link between symptoms and diseases or genes and diseases. However, none of these corpora are publicly accessible.

2.5 Negation and Speculation

Capturing the semantics of text is important in biomedical text mining. For instance, the identification of negation and hedging is essential for applications that describe biomedical events and rely on factual rather than speculative knowledge.

Negation occurs when a linguistic marker is used to reverse the meaning of part of a statement, for example, "*Aspirin is **not** associated with more toxicity than other ...*". Several systems have been developed to detect negation within biomedical text, e.g. *NegExpander* (Aronow, Fangfang, and Croft, 1999), *NegEx* (Chapman et al., 2011; Chapman et al., 2001) and *NegFinder* (Mutalik, Deshpande, and Nadkarni, 2001).

NegExpander is a tool to distinguish between negative and positive evidence. The tool detects negation terms and conjunctions in order to identify negated noun phrases, including those existing in conjunctions. As an example of this, "*no suspicions masses, suspicious classification ... etc.*" becomes "*no_suspicious_masses, no_suspicious_classification...etc*". Such expansion is useful for text indexing, given that negation terms such as *no* are usually considered stop words.

NegEx, a simple regular expression algorithm, uses UMLS resources and certain patterns to detect negated biomedical concepts in sentences. For instance, in the phrase "*This is not an infection*", the negated concepts is *infection*. That algorithm uses 136 phrases and terms such as *unlikely* and *was ruled out* to indicate probable or definitive negated concepts.

NegFinder is another tool to detect negated biomedical concepts. The system operates in a pipeline fashion, where the input of a component is an output of the previous

³<http://www.essayforum.com>

component. The system input is a discharge summary document and the output is colour-coded text that highlights the negated phrases or concepts. One significant finding from that research was that the negative terms *no*, *not*, *without* and *denied/denies* were found within 92.5% of the negated patterns of documents.

Speculation is the use of tentativeness and possibility language to reduce the strength of certain information (Hyland, 1995). An example of this is the use of the word *possibly* in “*Rhabdomyolysis was observed possibly because of a drug interaction between once-daily ticagrelor ..*”. Two main methods have been applied to detect speculation: use of speculation cues (substring matching), such as *suggest*, *likely* and *possibly* to identify speculative sentences, and use of statistical learning. Light, Qiu, and Srinivasan (2004), who introduced the speculation problem in NLP, have used both approaches to recognise speculation and reported that the machine learning method outperformed substring matching in terms of precision, and that substring matching outperformed the machine learning method in terms of recall. Furthermore, Agarwal and Yu (2010) developed another statistical model to detect hedge cues and their scope using a condition random field (CRF) algorithm (Lafferty, McCallum, and Pereira, 2001), and used the publicly-available corpus BioScope (Szarvas et al., 2008) to evaluate the model. Their system achieved 77.6% F1-score in detecting uncertain sentences and 77.44% F1-score in detecting hedging cues, however, they only achieved 19.27% F1-score in identifying the hedge scopes.

2.6 Information Extraction

Information extraction (IE) is the process of identifying instances of information from unstructured or semi-structured text. It has a wide range of applications in the biomedical domain. For instance, researchers frequently need to explore a large number of research texts to identify biomedical entities and their relationships, for further analysis and comprehension (Hobbs, 2002). A simple search-based approach may be insufficient for such a task, since biomedical entities frequently have synonyms and ambiguous terms. This section discusses two fundamental tasks in IE for the resolution of such problems, entity recognition (NER), which is responsible for recognising biomedical concepts, and relation extraction, which attempts to identify the relationships between concepts.

2.6.1 Named Entity Recognition

The goal of NER within the biomedical domain is to extract named entities such as genes, proteins, and cell and drug names. Within the biomedical literature, however, such an undertaking is challenging for various reasons (Shatkay and Craven, 2012). First, many biomedical terms, such as gene and protein names, appear in multiple forms. For instance, the short form of *breast cancer type 1 susceptibility protein* is *BRAC1*. Second, some biomedical concepts are homonyms of ordinary English words. For example, the fruit fly *Drosophila melanogaster* has gene names such as *And* and *lot*, and other gene names such *Sunday driver*. A third reason is that some concepts names are composed from other concept names, for instance, *MAP kinase 1* protein and *MAP kinase kinase 1*. Finally, the language used in the biomedical domain does not follow strict naming conventions, e.g. *Nacetylcysteine* vs *N-acetyl-cysteine* (Neustein, 2014s). Three key approaches have been followed in the development of NER systems: dictionary-based, rule-based, and machine learning-based systems.

Dictionary-Based Named Entity Recognition

A simple approach to develop an NER is to look for entities in the text that match a predefined list of entities (from a dictionary). The dictionary used in this case is a list of names, along with their types. However, dictionary-approach NER systems usually suffer from two essential limitations (Shatkay and Craven, 2012).

First, NER systems assume that the given dictionary is complete, however, this is difficult to achieve in the biomedical domain. Although repositories exist that contain an extensive list of concepts, such as gene names, the literature still contains uncatagorised concepts, especially those that remain under investigation. Moreover, there are uncatagued variants of concept names that are already catalogued. Bunescu et al. (2005) attempted to resolve this limitation through the use of a generalised dictionary which uses conventional patterns of certain concepts that have been indexed in the biomedical databases to recognise unseen concepts. For instance, the pattern *interleukin-⟨num⟩ ⟨gre⟩* of the protein name *interleukin-1 beta* can be used to recognise other protein names. In this example, the placeholder *⟨num⟩* refers to numbers and *⟨gre⟩* to a Greek letter, such as *beta*.

Second, dictionary-based NER is unable to address the homonym problem previously mentioned. For instance, a dictionary-based system cannot distinguish the gene name *And* in *Drosophila* from the ordinary English word that connects words or phrases.

Although it can be claimed that such an example can be resolved by detecting the occurrence of capital letters at the beginning of tokens that indicate names, additional contextual features may be required to resolve such a scenario.

Rule-Based Named Entity Recognition

Another approach to constructing NER systems for biomedical concepts is to apply certain rules to recognise concepts. These rules exploit certain morphological and lexical features to recognise named entities. For instance, Fukuda et al. (1998) applied two levels of rules to build a system for recognising gene and protein names. In the first level, orthographic, morphological and lexical rules are used to identify terms that are potentially part of protein names; in the second level, other lexical and part-of-speech rules are used to recognise the sequence of terms that represent the protein name. For example, to recognise that *focal adhesion kinase (FAK)* is a protein name, the system identified that *kinase* and *FAK* are common terms in protein names (level 1 rules); then, the system extended to the left until it reached the determiner *the*, since the speech tags between the words are adjectives (i.e., *focal* and *adhesion*) (level 2 rules). The rule-based approach to NER systems has the advantage that experts can easily understand the rules which consequently can be easily adjusted and modified.

Machine Learning-Based Named Entities Recognition

An alternative approach is the use of machine learning. For example, Bunescu et al. (2005) used several algorithms to develop a learning NER system and found that the use of a maximum entropy algorithm (Berger, Pietra, and Pietra, 1996) outperformed the other approaches in a specific study.

Machine learning NER systems treat the recognition of named entities as a classification problem. In reality, however, certain dependencies may exist between sequence of tokens that represent a concept. Therefore, it might be useful to employ machine learning algorithms that exploit the sequential nature of language to recognise biomedical concept names. Two popular sequence models that have been used in NER are hidden Markov models (HMMs) (Collier, Nobata, and Tsujii, 2000) and Conditional Random Fields (CRF).

An HMM is a joint probability distribution over paired observations and a sequence of labels. The Viterbi algorithm (Ryan and Nudd, 1993) can be used to train the parameters of the model in order to maximise the joint likelihood of a training set. It identifies the most likely label sequence in the state space of the possible label distribution to

align a sequence of tokens to states, and when these states correspond to named entities, they are extracted as concepts. HMMs have been successfully applied to NER tasks (Hinton, Brown, and London, 2001); however, they are limited in being unable to process multiple features that may require several dependencies. A CRF model, however, is able to learn from a large set of features without the need to explicitly encode the dependencies between them (Shatkey and Craven, 2012) (see Section (4.2.2)).

2.6.2 Relation Extraction

Relationship extraction aims to automatically discover the relationships between named entities in text. In the biomedical domain, relation extraction has been applied to a range of problems such as to discover relationships between proteins (Ramani et al., 2005), proteins and sub-cellular locations (Craven and Kumlien, 1999), genes drugs and cells (Rindflesch et al., 2000), genes and diseases (Rinaldi et al., 2003) or diseases and treatments (Bundsusch et al., 2008). Common methods employed to extract relationships are rule-based and machine learning-based approaches.

Rule-Based Methods

This approach uses a set of regular expression rules over words or part-of-speech tags to discover relationships between biomedical entities. Proux et al. (2000), for example, demonstrated a system to detect protein-protein interactions using linguistic patterns such as *gene product acts as a modifier for gene*, where the predicate of this pattern is *act*. This pattern covers sentences such as “*Eg1 protein acts as a repressor of BicD*”, where *act* is the predicate and *Eg1* and *BicD* are the arguments of that relationship. Ono et al. (2001) used another set of rules for the same task, formulated from syntactic features such as $\{PROTEIN1.*\ not\ (interact|associate|bind|complex).*\ PROETIN2\}$.

Relationship extraction using a rule-based method requires human effort to generate rules for a specific domain (such as protein-protein); thus, it is not easily adapted to other domains, such as the relationship between genes and diseases. Furthermore, it is difficult to formulate rules that cover all possible relationships existing between particular entities.

Machine Learning-Based Methods

An alternative approach is to employ machine learning methods, which treat the relationship extraction task as a classification problem. A protein-protein interaction within a sentence can be recognised by classifying the sentence as to whether it represents an interaction or not. In such cases, a classifier uses a set of features extracted from training

examples to construct a model for predicting the interactions within a sentence. Examples of these features include the two entities, their POS tagging, the word sequence between them, the POS tagging of the sequence, and the dependency path between the entities.

Miwa et al. (2009), for example used Bag-of-Words, Kim, Yoon, and Yang (2008) used Shortest Path (SP) and Airola et al. (2008) used graph features to recognize protein-protein relations. Other systems used kernel representations as features to predict sentences containing relationships, an approach which uses the similarity between an example and other examples to learn certain statistical features. For example, Mooney and Bunescu (2006) used a subsequence kernel to extract protein-protein relationships, where the subsequence kernel is a function that calculates the similarity between common patterns in sentences (such as subsequence of words between the entities in two sentences). Other examples of kernel features include the string kernel (Lodhi et al., 2002) and graph kernel (Airola et al., 2008).

2.6.3 Event Extraction

Although binary relationships are capable of describing many key biological interactions, other relationships describing nested events cannot be captured in this manner. For instance, the phrase “*..the phosphorylation of TRAF2 inhibits binding to the CD40 cytoplasmic domain*” describes two relationships: the phosphorylation of *TRAF2* (relation 1) and the interaction between *TRAF2* and *CD40* (relation 2) (Zhou, Zhong, and He, 2014). To capture the events in this example, it is important to follow a more expressive representation than in binary relations. For example, an event may consist of a type, trigger and additional arguments that describe the entities used in the event such as theme and cause, where theme describes the entity that was the subject of the events action, and cause is the entity that causes the event. Based on this representation, the previous example includes two event types, phosphorylation and binding. The first event is triggered by phosphorylation, and in second by binding. The first event contains only one argument, which is *TRAF4* (the theme), while the second event has the entity *CD40* as the theme and *TRAF4* as the cause.

Several systems have been developed to extract events, using different approaches such as NER and relationship extraction methods, in addition to machine learning. These systems are typically developed within a pipeline in order to assemble the event constituents, such as event type, trigger and arguments. For example, Björne et al. (2010) described a pipeline consisting of a NER system to recognise biomedical entities,

a dependency parser, event detection which uses the output of the NER and the dependency parser to predict the event triggers and polarity, and uncertainty detection for the produced events. Such systems, events and arguments are extracted independently; other researchers such McClosky, Surdeanu, and Manning (2011) and Riedel and McCallum (2011) apply various methods to extract both types simultaneously rather than independently.

2.7 Question Answering

Question Answering (QA) is a form of information retrieval system which aims to provide direct and precise answers to queries instead of providing users with a large set of documents that could potentially be relevant to the query. A QA system generally consists of five main processing stages (Hirschman and Gaizauskas, 2001): question analysis, document collection pre-processing, candidate answer document selection, candidate answer document analysis and answer extraction and generation.

In the question analysis stage, the question is assessed and classified based on its linguistic features to determine the corresponding answer type, and keywords and relationships that will be used to identify potential answers. In the second stage, documents are pre-processed into different forms to enable QA to select the best candidate answers at a later stage; this process includes shallow parsing and POS tagging. The document answer selection stage requires the choice of a mechanism for retrieving answers, such as choosing between a Boolean based search engine and a vector space model engine to retrieve relevant documents; further decisions may be also required to identify the parameter chosen for the selected retrieval method. The candidate answer document analysis stage uses the output of the document pre-processing stage to select the most relevant passage of text with the potential to have a correct answer. The answer extraction stage uses the output of the question analysis stage and the candidate answer analysis stage to select and rank the set of documents produced by the previous stage, likely to contain answers to the question.

Several research studies have examined different types of biomedical QA, including medical QA and biological QA (Athenikos and Han, 2010).

2.7.1 Medical QA

Researchers in the medical field typically use the medical literature to enable decision makers (such as clinicians) to draw conclusions. Such information can be retrieved using a specific method of formulating questions, such as the PICO format. The PICO

format is commonly used to formulate such a question, and it consists of four components. The first, the population or problem (P), in terms of specific demographic information such as age range or sex. The second is the intervention or treatment of interest, including procedures, diagnostic tests, and risk factors (I). The third is the comparator or control, which could be an intervention used for comparison (C). The fourth is the outcome, which describes the effect of the intervention (O). Table (2.30) shows the components of the PICO question *“In patients with recurrent furunculosis, do prophylactic antibiotics, compared to no treatment, reduce the recurrence rate?”*.

P (Population or Problem)	patients with recurrent furunculosis
I (Intervention)	prophylactic antibiotics
C (Comparator)	no treatment
O (Outcome)	reduction in recurrence rate of furunculosis

TABLE 2.30: A question formulated in PICO format

Ely et al. (2000) adopted a different taxonomy to classify medical questions. This taxonomy categorises questions into two main sectors, clinical (this includes evidence-based, non-evidence and specific questions) and non-clinical. The clinical evidence-based questions can relate to intervention, such as *“What is the drug of choice for condition x?”*, or non-intervention, *“How common is depression after infectious mononucleosis?”*. An example of a clinical non-evidence question is *“What test is indicated in situation x?”*, while a clinical-specific might include *“What is the cause of symptom x?”*. An example of a non-clinical question is *“How should I manage condition x (not specifying diagnostic or therapeutic)?”*. That research suggested that such a taxonomy is important to address the real questions that occur in practice, in order to retrieve the relevant documents. Furthermore, they found that non-evidence, specific, and non-clinical questions were generally not answerable.

Huang, Lin, and Demner-Fushman (2006) studied the adequacy and suitability of mapping clinical questions into the PICO framework, and found that such a framework is useful for questions that are primarily centred on therapy; however, it is less suitable for other types of questions.

For QA systems, Niu et al. (2003) developed a QA system that considers a question, identifies the four roles in the PICO framework, and identifies potential answer text according to the semantic roles of the question. They found that semantic role identification of both questions and answers was an effective approach for locating answers in QA systems.

Yu et al. (2007) developed *MedQA* (medical definitional question answering), which takes a definitional question and returns a short, coherent answer. The system integrates information retrieval, extraction and summarization techniques to generate paragraph-level text to response to the query. The system takes a question and classifies it into one of the categories described by Ely et al. (2000) in order to identify the definitional one. The noun phrases in the question are then used to retrieve relevant documents using Lucene[®]. Next, lexicon-syntactic patterns are extracted from the retrieved documents to recognise the definitional answers. Finally, summarisation techniques are applied to definitional answers to present them in shorter forms while preserving the information content. The *MedQA* system did not exploit the semantic information which may play an important role in identifying relevant answers to a question, especially at the answer extraction and summarisation stages.

2.7.2 Biological QA

In contrast to the medical domain that benefits from a structured framework to formulate questions (i.e., PICO), or from taxonomical questions described previously that can be exploited in medical QA systems, the biological domain lacks such a structured approach. Thus, apart from the TREC Genomic track described by Hersh et al. (2006), only a few biological QA systems have been described.

Takahashi, Koike, and Takagi (2004), for example, developed a specific QA system that answers biological questions based on Medline and UMLS semantic types. The system used different biomedical resources such as UMLS, GENIA (Ohta, Tateisi, and Kim, 2002), a family name dictionary (Koike, Niwa, and Takagi, 2005), and a thesaurus to resolve other semantic issues. The system takes a question and returns the most relevant answer. It starts by analysing the question using UMLS or another thesaurus to identify the type of answer appropriate to the question. Stemmed question terms are subsequently expanded using the previously-mentioned resources, along with other heuristics, to render it applicable to the full text functionality supported by MySQL. The output of this phase is a list of documents pertaining to the question. The system then identifies all terms in the documents that belong to the same semantic type or superclass as the answer type. When those terms are found in relation to any of the question terms, such as subject-object, those terms become candidate answers to the question. Along with their IDs, sentences, and supporting evidence, these candidates are entered into a voting system to determine the most likely response to the question. The final output of the system is the evidential sentence and associated abstracts.

Lin et al. (2008b) developed a factoid biological QA system to address biomolecular events such as gene-protein interactions. Factoid questions take multiple forms, such as *what is ..* and *when is ..*; the answers to such questions tend to be short pieces of information like time, location or biomedically-named entities. The system consists of four components, question processing, passage retrieval, candidate extraction and feature generation, and answer ranking. The question processing component includes named entity recognition (NER), semantic role labelling (SRL), question classification and query modification. NER identifies named entities within the question and SRL extracts the predicate and its corresponding arguments, where these are transformed into features to be subsequently used by the answer ranking component. The question classification step uses hand-crafted patterns to recognise the required named entity type (e.g. protein, cell or DNA). The query modification step expands queries to generate additional synonyms and other tenses related to the main verb in the question, in order to improve the search in Google.

The passage retrieval component comprises an interface that sends queries to Google (only results indexed from PubMed are used) and returns relevant web pages. In the candidate extraction and feature generation components, NER and SRL are used to extract NEs, predicates, and their corresponding arguments (similar to the question processing component), and transform these into features. In addition to those features derived from the question, the features are used to match the question with the passages. Moreover, the GENIA tagger (Usami et al., 2011) is used to extract biomedical events in a nominal form. In the answer ranking stage, each name entity is treated as an answer candidate, and the ranking score of each candidate is obtained using a linear model to calculate its score based on its features.

2.8 Evaluation Methods Overview

This section presents some key methods for evaluation commonly used in NLP.

2.8.1 Intrinsic vs Extrinsic Evaluation

In intrinsic evaluation, a component is directly evaluated against a set of predefined criteria related to the desired functionality of the component. For example, a relationship extraction component in a semantic search engine is evaluated using a gold standard dataset which contains sentences along with their manually-extracted relationship tuples. A portion of this test dataset is used to evaluate the level of agreement between

the component output and the *ground truth* or the correct tuples. The optimum performance for the component is achieved when it is able to produce exactly the same tuples as in the test data set. In this type of evaluation, questions such as how accurately can the component extract the correct relation tuples from sentences parsed by the search engine can be answered.

In extrinsic evaluation, however, a component is indirectly evaluated by assessing its contribution on an external task to the component itself (Mollá and Hutchinson, 2003). For example, the relationship extraction component is evaluated by measuring how it contributes to improving search engine performance. An example of a question that can be answered using extrinsic evaluation is how much value does the relationship extraction component add to improving the accuracy of search engine results?.

2.8.2 Accuracy, Precision, Recall and F1-score Measures

Table (2.31) shows the confusion matrix that allows visualisation of the performance of a binary classification problem. The columns in the table represent the actual classes, and the rows represent the classifier predictions. Each cell contains the number of instances predicted by the classifier that falls into that category. The terms True_Positive and True_negative, False_Positive and False_Negative compare the results of the classifier against trusted external judgement. The terms Positive and Negative refer to the classifier's prediction, and the terms True and False refer to whether the classifier prediction corresponds to the external judgement.

	Positive	Negative
Positive	<i>True_Positive</i>	<i>False_Positive</i>
Negative	<i>False_Negative</i>	<i>True_Negative</i>

TABLE 2.31: Confusion table for a binary classification problem

The most intuitive method for evaluating a binary classification problem is to calculate the accuracy Acc (see Equation (2.2)), the proportion of correct results that the classifier was able to achieve. However, the accuracy score alone may be somewhat misleading as it is possible to achieve high accuracy by predicting all instances as Positive. Precision, recall, and F-score measures (see Equations (2.3)) are thus used to obtain further insight into system performance. The precision (P) is the fraction of retrieved instances that are relevant, recall (R) is the fraction of relevant instances that are retrieved and the F-score (F) is a measure that balances the information gained from both precision and recall.

$$Acc = \frac{True_Positive + True_Negative}{True_Positive + True_Negative + False_Positive + False_Negative} \quad (2.2)$$

$$P = \frac{True_Positive}{True_Positive + False_Positive} \quad R = \frac{True_positive}{True_Positive + False_Negative} \quad (2.3)$$

$$F = 2 \times \frac{precision \times recall}{precision + recall}$$

2.9 Conclusions

This chapter highlighted various topics that are directly relevant to either the contradiction problem itself or to the tools and approaches used to develop the proposed system. The topics considered relevant to the contradiction problem and described here were: contradiction, paraphrasing and textual entailment and argumentation. These topics are useful in developing the conceptual understanding of the phenomenon of contradiction in the field of biomedicine. The topics relevant to the tools and approaches used to develop the proposed system and discussed in this chapter were: negation and speculation, information extraction, question answering and evaluation methods. Such topics were found to be beneficial in terms of understanding the various available methods by which to extract the linguistic features of claims, and formulating and evaluating the proposed system.

Chapter 3

Two Corpora of Contradictory Research Claims

3.1 Introduction

Progress on exploring the contradiction problem in biomedical research is hampered by a lack of appropriate resources for the development and testing of potential strategies. The development of these resources may be relatively complex, given the volume of research that has been published and the difficulty in identifying contradictory claims within it.

This chapter presents two methodologies for developing contradiction corpora: manual and automatic. The manual approach follows standard NLP methodology to develop a corpus using human annotators, employing published systematic reviews. Such methodology, however, requires considerable effort and time to collect, synthesise, and annotate the dataset. The second methodology presents an alternative approach to automate the process of constructing a contradiction corpus, using a biomedical knowledge resource repository called *SemMedDB* (Kilicoglu et al., 2012). Unlike the manual methodology, which explicitly uses claim sentences to construct the corpus, the automatic method makes use of any type of sentence, including claims derived from research abstracts. The corpus is generated by employing incompatibility characteristics between the relation tuples extracted using *SemRep* and stored in *SemMedDB* (see Section (3.4.1)).

Compared with existing work (see Section 2.2), which focused on negation, both corpora are intended to include a wider range of linguistic phenomena and topics from the biomedical literature to identify contradiction.

This chapter describes four contributions: development of a methodology to generate a corpus of contradictory claims by making use of systematic reviews; generation of a contradiction corpus (*ManConCorpus*) using that methodology; development of

another methodology to automate the corpus generation process without the need of human annotation by making use of the *SemMedDB* database; and implementing that methodology to construct another contradiction corpus (*AutConCorpus*).

The remainder of this chapter is organised into three main sections. The next section provides definitions for contradiction and claims. Section (3.3) shows the process of generating the contradiction corpus *ManConCorpus* using manual annotation. Section (3.4) describes the process of generating the new corpus *AutConCorpus* without the need for human annotation.

3.2 Definitions

3.2.1 Contradiction Definition

Contradiction has been defined as the existence of two or more incompatible propositions that describe the same proposition (Oxford-Dictionary, 2003). In other words, two fragments of text, T_1 and T_2 , are contradictory when they each assert different information about the same proposition that cannot simultaneously be true.

The problem of contradiction has previously been explored within work on textual entailment (Bowman et al., 2015b; Giampiccolo et al., 2008; Harabagiu, Hickl, and Lacatusu, 2006b) where a common approach is to consider T_1 and T_2 to be contradictory when one of them entails the negation of the other. Marneffe, Rafferty, and Manning (2008) used a looser definition which was intended to be less restrictive: two fragments of text are contradictory when they are extremely unlikely to be true at the same time.

Sarafraz (2011) defined contradiction as two texts that describe events sharing certain attributes (e.g., theme, cause and anatomical location) but with different polarity. That work was restricted to statements about a very specific type of information (chemical interactions) and one way of expressing contradiction (negation).

This research focuses on the same domain but targets the piece of text that describes the research claims rather than only molecular interaction; moreover, this research uses a less restrictive definition of contradiction which covers further expressions in addition to negation.

The definition of contradiction used here is as follows: Two research claim sentences, T_1 and T_2 , are said to contradict when, for a given proposition F , information inferred about F from T_1 is unlikely to be true at the same time as information about F inferred from T_2 .

This definition is based on inferences from statements being *unlikely to be true at the same time* rather than being *logically inconsistent*. This method evades the overly restrictive definition of contradiction employed in previous research (Marneffe, Rafferty, and Manning, 2008). Research findings in scientific documents are often expressed cautiously, e.g. using hedges (Hyland, 1995), to reduce the chance of statements being logically inconsistent with one another. Nevertheless, researchers are often interested in obtaining as much information as possible about a research question of interest, and are therefore likely to be interested in statements which are unlikely to be simultaneously true.

This research defines contradiction as three-way relation (between proposition F and two sentences T_1 and T_2) rather than two-way relation (between two propositions) as in logic (Chierchia and McConnell-Ginet, 2000) for one main reason. The language used in biomedical documents tends to involve complex sentence structures with multiple propositions described within the same sentence, so it is therefore important to consider contradiction relative to a particular research proposition. Table (3.1) shows two sentences, (1) and (2), that may be considered contradictory in relation to some proposition but not to others. Sentence (1) states that fish consumption does not prevent heart failure, without providing information about the types of fish or population groups the assertion applies to. Sentence (2) asserts that fish consumption *does* prevent heart failure in a particular population group of “*older adults*” and with a specific type of fish, “*tuna, broiled or baked*”. The sentences would not be considered contradictory, relative to the proposition *consumption of fried fish prevents heart failure*, since both suggest that it does not. However, they would be considered contradictory if the proposition being considered was *eating tuna prevents heart attack in older adults*, since sentence (2) suggests that it does while sentence (1) suggests that it does not.

	PMID	Text
1	19789394	Our findings do not support a major role for fish intake in the prevention of heart failure
2	15963403	Among older adults, consumption of tuna or other broiled or baked fish, but not fried fish, is associated with lower incidence of CHF

TABLE 3.1: Examples of potentially contradictory sentences

It is possible that contextual information may affect whether pairs of statements are considered contradictions (e.g., there would be no contradictions between sentences (1) and (2) if sentence (1) only applied to teenagers and fried fish). However, we consider the contradiction in isolation and do not take account of its context. This strategy

ensures that the problem does not become intractable, and is in line with approaches adopted by other research into contradiction detection (Marneffe, Rafferty, and Manning, 2008).

3.2.2 Claim Definition and Types

The identification of claims, and the contradictions between them, is made difficult by the range of different types of claims that can occur in the biomedical literature. The term *claim* has been previously used within argumentation literature (see Section 2.4) as a synonym for the term *conclusion*. Toulmin (2003) used *claim* in terms of a position being argued for or the conclusion of an argument.

This research focuses on contradiction between claim sentences rather than on any sentence for one main reason. Claim sentences in a research abstract tend to describe the take-home message of the research, which is one of the most important outcomes of the research. Thus, the contradiction between these types of sentences represent conflicting central findings of entire studies which will be of more interest to researchers than potential contradiction over small and possibly unimportant details. The ability to automatically recognize such information and then to identify contradictory claims would be very useful for people using research documents to make decisions.

This research defines *claim* as: the summary of the main points presented in a research argument; these points can either introduce new knowledge to readers or update their knowledge on a topic. A claim contains the most important piece of information that authors want to communicate to the reader, as it contains the research outcomes. In the biomedical literature, claims tend to summarize the authors' findings and are usually presented at the end of the research article.

Blake (2010) identified five types of claims: explicit, implicit, correlations, observations and comparisons. These types were formulated based on the availability of certain information (facets): two concepts, a change and the basis of the claim. Although this classification system provides a framework through which a biomedical claim can be automatically analysed, it is not clear how a judgemental claim, such as effectiveness of a drug or a technique (as in sentence (5) in Table (3.2)) can be analyzed.

	PMID	Type	Text
3	25645463	Factual	Ghrelin and its synthetic analog hexarelin are specific ligands of growth hormone secretagogue (GHS) receptor
4	25651296	Recommendation	Therefore, it is recommended to treat such fractures at institutions with medical staff experienced in their management
5	22942294	Evaluative	Combined clopidogrel and aspirin overcome single drug resistances, are <i>safe</i> for bleeding
6	21050973	Evaluative	Aspirin plus clopidogrel is <i>more effective</i> in venous graft patency than aspirin alone in the short term after CABG, but further, long-term study is needed
7	16308009	Causal	Autologous stem cell transplantation led to significant improvement in cardiac function in patients undergoing off-pump coronary artery bypass grafting for ischemic cardiomyopathy
8	21146675	Causal	Routine use of postoperative aspirin after coronary artery bypass grafting (CABG) reduces graft failure and cardiovascular events
9	10198739	Causal	In the Spanish Mediterranean area, the presence of antigens B-15 and DQ3 would be associated with advanced DCM

TABLE 3.2: Claims Typology

This research presents another framework which has been adopted from a general linguistic analysis rather than biomedical point of view (Mayberry, 2009). This framework was found suitable for this research for two reasons: first, to show the most common types of claims that a researcher may find in a biomedical abstract, and second, to assist the corpus annotators during the corpus annotation task.

Factual claims assert the truth of a statement, such as that shown in sentence (3) of Table (3.2). This type of claim is usually used to support the authors argument rather than forming the conclusion of an argument.

Recommendation claims suggest a particular course of action rather than providing new information, as shown in sentence (4). This type of claim is employed by authors in order to verify their understanding of a given issue. Words such as *should*, *would*, *must*, *ought* and *needs to be* are good indicators of such claims.

Evaluative claims occur when an author expresses a judgement about the value of a biomedical concept (e.g., drug, procedure, equipment, gene, or protein). This type of

claim is often used as an interpretation of evidence presented in the research, and is usually expressed by either making a statement about the properties of a concept, e.g. sentence (5), or by comparing the concept with another, e.g. sentence (6).

Causal claims suggest a relationship between two concepts and assert that one concept influences the other. Hashimoto et al. (2012) described three types of influences: excitatory, inhibitory and neutral. Excitatory influence indicates a direct activation or enhancement, e.g. sentence (7), which shows that *Autologous stem cell transplantation* had an excitatory influence on *cardiac function* under certain conditions. An inhibitory influence is the opposite of excitatory and indicates direct deactivation or suppression. For example, sentence (8) is a casual claims which asserts that *Routine use of postoperative aspirin* has an inhibitory effect on *graft failure and cardiovascular events*. The final type of causal claim, neutral, is neither excitatory nor inhibitory. For example, sentence (9) asserts a relationship between *presence of antigens B-15 and DQ3* and *advanced DCM (Dilated Cardiomyopathy)*, but doesn't explicitly state whether it is excitatory or inhibitory.

3.3 Manually Annotated Contradiction Corpus

The traditional approach used to construct a corpus for textual inference tasks, including contradiction, is to compose artificially created pairs of text T_1 and T_2 (Bentivogli et al., 2010). This type of strategy, however, disregards the proposition that these components typically comprise a larger system, such as a Question Answering system, which therefore separates the component prediction from the context and environment of the larger system.

The construction of *ManConCorpus* corpus is considered within the content of search engines, which take a query/question and return a list of relevant documents such as research abstracts. In such an environment, a user searching for information about a certain problem, e.g. the effect of aspirin on heart attack, enters a closed question in a specific format (see Section 3.3.1) as the query, and expects the search engine to return a list of research abstracts where each abstract contains at least one claim that answers the question. Thus, the corpus consists of multiple groups of abstracts, with each group consisting of a question and a list of claim sentences that answer the question.

3.3.1 Corpus Data Collection

ManConCorpus was constructed based on the hypothesis that a systematic review, which aggregates multiple research findings to reach more reliable and accurate answer to a

particular research question, can be used as a potential resource to collect contradictory research claims about a particular problem.

Systematic Review

A *systematic review* is an approach to identify, collect, evaluate and assess the current evidence related to a particular research question (including those providing contradictory claims). The systematic review generally involves five stages: (1) formulating a research question, (2) identifying all studies relevant to the question, (3) assessing the quality of the studies, (4) summarizing the findings, and (5) interpreting the outcomes. Stages (1) and (4) are directly relevant to the methodology used for constructing *Man-ConCorpus*.

A systematic review necessitates a well-formulated question in order to determine whether a study is directly applicable and significant to the research problem. The PICO format is commonly used to formulate such a question (see *Section (2.7.1)*). This research employs the PICO questions as a reference to determine whether the claims are contradictory or not. If a claim sentence in a research abstract agrees with a certain PICO question and if another claim sentence in another abstract disagrees with the question, then these claims are considered potentially contradictory.

Many systematic reviews are based on *Meta-analysis*, a quantitative, formal and epidemiological method that uses a statistical approach to aggregate the outcomes of the studies used in the systematic review in order to obtain a conclusion that provides a better understanding of study findings. The results of the meta-analysis are graphically presented using a forest plot diagram, which shows the findings of multiple quantitative studies addressing the same problem (Lewis and Clarke, 2001). These diagrams play an important role to predict the systematic reviews that likely contain contradictory studies and consequently contradictory research claims.

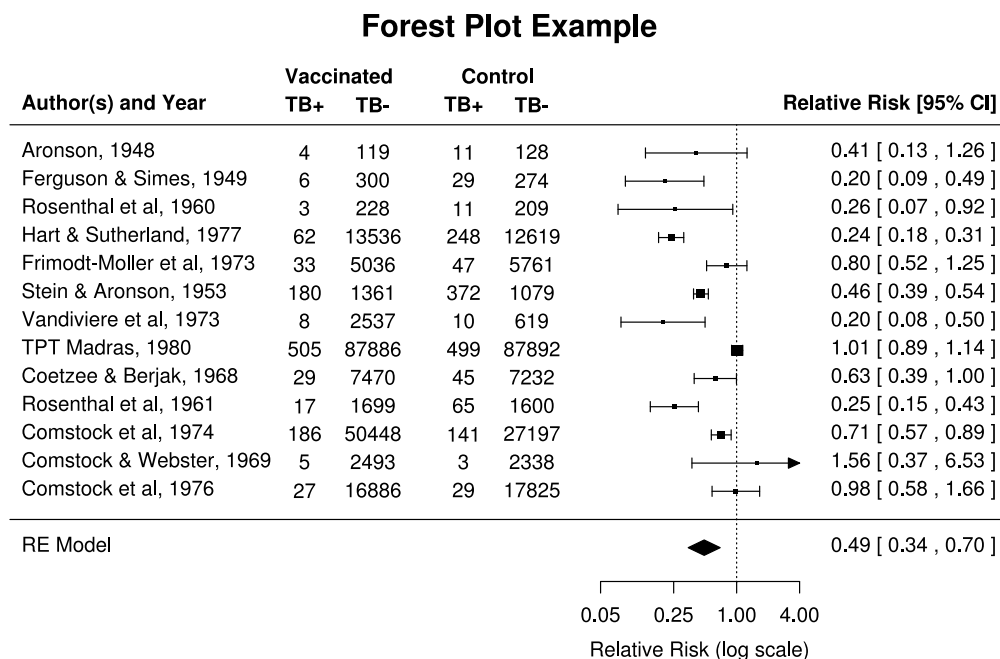


FIGURE 3.1: An example of a forest plot diagram. The dataset was retrieved from Viechtbauer (2010)

Figure (3.1) shows an example of a forest plot diagram illustrating the findings of studies that evaluated the impact of a vaccine, known as BCG, to prevent the development of tuberculosis (TB). The left side of the vertical column shows studies that favoured the vaccine (treatment) and the right side shows those that favoured the placebo. One study with findings that statistically favoured the placebo is represented (Comstock & Webster, 1969), two studies (TPT Madras, 1980) and (Comstock et al., 1976) show no significant difference, and the rest show favourable outcomes for the vaccine. Consequently, the diagram shows that there is at least one study (Comstock & Webster, 1969) that may contain a claim contradictory to the others.

Systematic Reviews Collection

A corpus was compiled using systematic reviews on the topic of cardiovascular disease, a major contributor to global mortality and whose causes are frequently the subject of research papers (Fuster et al., 2010). Given the volume of research published on the topic, it was expected that some contradictory claims would be found.

Four types of cardiovascular disease were chosen: *Cardiomyopathy*, *Coronary artery*, *Hypertensive* and *Heart failure*. The systematic reviews of these topics were retrieved using the *Pubmed*[®] search engine. For example, the query “*Cardiomyopathy*”[title] AND “*meta-analysis*”[title] was used to search for systematic reviews discussing cardiomyopathy disease, and the same procedure was applied for the other disease types. The

modifier *[title]* was used to ensure that the search keywords occurred within the title of the article.

Only systematic reviews that contain a meta-analysis were considered for inclusion in the corpus. Meta-analysis typically uses visual diagrams (i.e., forest diagrams) to statistically compare the findings of the studies included in the review. Although the differences between findings may not be significant, forest diagrams are good indicators of whether the studies used in the review may contain contradictory claims. However, forest plot diagrams only serve as an initial filtering mechanism to decide whether to include the studies from that review in the next stage of corpus construction.

Following completion of this stage, multiple systemic reviews had been identified. For example, Table (3.3) shows the titles of a systemic review (10) and its associated studies. These studies were included as candidate datasets, since the forest diagram of the review showed disagreement between at least one of the study findings.

	PMID	Text
10	23489806	Fish consumption and incidence of heart failure a meta-analysis of prospective cohort studies
11	18954578	Incident heart failure is associated with lower whole-grain intake and greater high-fat dairy and egg intake in the Atherosclerosis Risk in Communities (ARIC) study
12	19789394	Intake of very long chain n-3 fatty acids from fish and the incidence of heart failure: the Rotterdam Study
13	20332801	Fatty fish, marine omega-3 fatty acids and incidence of heart failure
14	15963403	Fish intake and risk of incident heart failure
15	21610249	Fish intake and the risk of incident heart failure: the Women's Health Initiative

TABLE 3.3: Example of studies associated with a systemic review

3.3.2 Question Formulation

Research claims may express multiple statements about the same or different propositions, even within the same sentence. This may confuse annotators when annotating claims to construct a contradiction corpus. Therefore, it was considered essential to create a question for each group of studies collected previously in section (3.3.1) to identify the common proposition, in order to determine the potentially contradictory claims.

An annotator with an advanced degree in medicine was asked to use the titles of the study abstracts included in a systemic review, in addition to the review abstract itself, to formulate a suitable question that could be answered by each study in the review.

The questions were formulated to be closed (i.e., requiring either a *yes* or *no* response), and written in the simple present tense. Moreover, the questions were compiled with the PICO standard to include information about the patient problem or population (P), intervention (I), comparison (C) and outcomes (O). *“In the elderly, is n-3 fatty acid from fish intake associated with reduction in the risk of developing heart failure”* was formulated for the studies described in Table (3.3).

The purpose of this approach was to enable the annotators at later stages to measure the assertion values of claims with respect to the question, which is the proposition in this situation. Thus, when two claims provide different assertion values or conclusions to a question, they are considered potentially contradictory. The instructions that were given to the annotator to formulate the questions were as follows:

1. Read the title of the review abstract and its content to understand the objective of the review.
2. Read the title of each study abstract associated with the review and examine its content, particularly the conclusion sections, to ensure that the study is directly relevant to the question addressed in the review. Exclude any studies that are found not to be directly associated with the main objective of the review, or where the association is unclear.
3. Formulate a PICO question for each review. The question should be a closed question; in other words, it can be answered with either a *yes* or *no*.

Moreover, the annotator was advised that there may be cases where there is incompatibility between the populations considered in the studies or there are studies that use alternative terms to refer to the same or similar concepts (e.g. cardiovascular disease and myocardial infarction). In these cases the question may be formulated using either (a) a generic term covering all the concepts, or (b) list all terms via the use of or, e.g. *“In patients with X condition, is y associated with cardiovascular disease or myocardial infarction”*.

3.3.3 Corpus Annotation

The final stage of corpus construction was to identify and annotate the claims in the study abstracts. Two annotators (other than the one recruited in the question formulation stage) were recruited, each with native level English fluency, an advanced degree in a field related to medicine, and employment in a medical role (one in an academic department, and the other in a hospital). Both were familiar with medical research literature and evidence-based medical research. The annotators were asked to carry out

three tasks in each group of abstracts: choose a claim sentence from each research abstract in the group, annotate the claim with an assertion value (*yes/no*) with respect to the question formulated for the group, and annotate the claim type (*causal/evaluative*) according to the claim types described earlier.

The first task was to evaluate each study abstract with respect to the question formulated for its group, and subsequently identify the best sentence that provided the best answer to the group question. The motivation behind identifying only the best sentence rather than identifying all possible sentences that may answer the given question was to encourage the annotators to focus on the sentence that may describe the contribution of the abstract with respect to the given question and at the same time may contradict other claims in the other abstracts in the same group. The instructions that were given to the annotators were that for each abstract associated with a review:

1. Carefully read the question associated with the review.
2. Examine each study abstract and identify the best sentence that serves as an answer to the review question.

Moreover, the annotators were advised that the claim sentence can usually be found in the conclusions section of the study abstract. This can be identified by the use of the explicit label (*Conclusion/Conclusions*) or implicitly by the use of signal words such as *In conclusion...*, *We found that...* and *Our work suggests...*. In cases where no sentence providing an answer to the question is found in the conclusion section, a sentence from the results section can be chosen; provided the sentence answers the question and can be considered as a claim. If no suitable sentence can be identified the study abstract should be excluded from the set of abstracts associated with that particular review.

In cases where more than one sentence that could potentially serve as the answer to the review question is identified, the annotator should choose the sentence that provides the clearest answer to the question considering all of the information contained in the study abstract.

The second task was to annotate the claim with either *yes* (to indicate the claim agrees with the question) or *no* (to indicate that it was not). Following the initial annotation phase, the pair of annotators met to resolve disagreements and decide on the final annotation. The instructions that were given to the annotators were to provide an assertion value for each claim with respect to the question. Two possible values can be assigned: *yes* and *no*, where *yes* should be used when the claim asserts a positive answer to the question and *no* if it does not. If the claim neither asserts nor denies the question then the assertion value should be *no*.

In the final task, the annotators were required to determine claim type, and once again met to resolve disagreements and determine the final claim type. The instructions that were given to the annotators were to annotate each claim as either *causal* or *evaluative* (see the definitions above), where *causal* claims should be annotated as CAUS and *evaluative* claims as EVAL.

Moreover, they were advised that claims in biomedical abstracts tend to be complex and a claim can be interpreted as causal and evaluative at the same time. For example, “Among our population of largely low or asymptomatic HCM patients, the presence of scar indicated by CMR is a good independent predictor of all-cause and cardiac mortality.” This claim states that the scar indicated by CMR is a predictor of all causes and cardiac mortality, which shows an indirect causal relationship between the scar and cardiac mortality. However, at the same time the claim evaluates this relation using the term “good”. In such cases, the annotator should consult the abstract content to determine whether the purpose of the study is to identify an association between the scar and mortality or to evaluate to what degree the scar can be used as a predictor for cardiac mortality. Appendix (A) shows the guidelines given to the annotators.

3.3.4 Results and Discussion

A total of 40 suitable systematic reviews were identified, and a question formulated for each review and its associated studies. After retrieving the abstracts for studies cited in the reviews (397 in total), annotators were asked to identify a claim from each study abstract that answered the question, determine whether or not it agreed with the question and its claim type.

From the initial retrieved abstracts, 19 were removed since the annotators were unable to identify a claim that provided a clear answer to the question. Once the annotation process was complete, it was found that there were 16 systematic reviews for which no contradictions were identified in the corresponding abstracts (i.e., the annotators had annotated all the claims as either agreeing or disagreeing with the question). These reviews, and the abstracts associated with them, were also excluded.

The final corpus consisted of 259 abstracts associated with 24 systematic reviews. The corpus itself is formatted as XML for ease of processing. Figure (3.2) shows examples of two formatted contradictory claims. The corpus is available in http://staffwww.dcs.shef.ac.uk/people/M.Stevenson/resources/bio_contradictions/.

```
<CORPUS>
<REVIEW REVIEW_PMID="23602289" REVIEW_TITLE="The effects of statins on blood
  pressure in normotensive or hypertensive subjects—A meta-analysis of
  randomized controlled trials">
<CLAIM ASSERTION="NO" PMID="2010437" QUESTION="In patients undergoing coronary
  bypass surgery , does Aspirin usage , compared to no aspirin , cause bleeding"
  TYPE="CAUS">Patients taking 85–325 mgm of aspirin with a normal bleeding time
  undergoing elective CABG did not have increased RBC loss or increased
  transfusion requirements.</CLAIM>
<CLAIM ASSERTION="YS" PMID="16153930" QUESTION="In patients undergoing coronary
  bypass surgery , does Aspirin usage , compared to no aspirin , cause bleeding"
  TYPE="CAUS">Aspirin pretreatment revealed no beneficial effects and resulted
  in increased postoperative bleeding and requirement for blood product
  transfusions after coronary artery bypass grafting in patients with stable
  angina.
</CLAIM>
</REVIEW>
</CORPUS>
```

FIGURE 3.2: Examples of formatted claims

Table (3.4) shows a sample of the questions used to annotate the claims (see Appendix (B) for the full list of questions).

Review-PMID	Question
22498326	In patients with HCM, does using imaging technique, compared to conventional techniques, serve as a predictor for adverse prognosis?
23623290	In patients with chronic heart disease, does Bone marrow Stem cell transplantation or injection, compared to none, improve cardiac function?
21556773	In patients with dilated cardiomyopathy, are HLA genes associated with development of Dilated Cardiomyopathy?
24040766	In Han Chinese population, is SNP T-778C of apolipoprotein M associated with risk of developing Diabetes or stroke?
24212980	In patients undergoing coronary bypass surgery, does Aspirin usage, compared to no aspirin, cause bleeding?
24035160	In patients undergoing coronary artery bypass, does the combination of aspirin and clopidogrel, compared to aspirin alone, prevent graft occlusion or improve patency?
24172075	In patients undergoing coronary by pass surgery, is Off-pump, compared to conventional on pump coronary artery bypass grafting, more beneficial?
24135644	In patients with coronary artery disease, is mutation or polymorphisms in endothelial nitric oxide synthase gene associated with CAD or MI or ACS development?
24036021	In patients with atherosclerotic plaque or myocardial infaction, does -463G or -463A polymorphism in MPO gene influence MI or CAD development?
24039708	In patients with coronary artery disease (CAD), is C242T polymorphism of P22(PHOX) gene associated in development of CAD?
24090581	In patients with coronary artery diseases, does combining CABD and CEA, compared with CABG or CEA alone, reduce morbidity?

TABLE 3.4: A sample of the questions formulated for the final corpus

Table (3.5) shows the number of study abstracts associated with each review and their distribution according to the assertion values and the claims types. The table shows that the corpus contains 180 claims that agree with their questions, and 79 claims that disagree. Moreover, the table shows that 165 sentences were causal claims and 94 were evaluative.

Topic	Review-PMID	#Abstracts	Assertion		Type	
			yes	no	caus	eva
Cardiomyopathy	22498326	4	3	1	2	2
	23623290	9	7	2	3	6
	21556773	15	12	3	12	3
Coronary artery	24035160	5	3	2	2	3
	24135644	20	13	7	13	7
	24036021	4	2	2	4	0
	24212980	20	12	8	14	6
	24039708	18	11	7	17	1
	24172075	7	2	5	0	7
	24090581	8	4	4	2	6
	24040766	5	4	1	5	0
Heart failure	23489806	4	3	1	4	0
	23181122	5	4	1	5	0
	23962886	15	13	2	14	1
	24163234	29	22	7	13	16
	24165432	6	4	2	2	4
	23219304	10	6	4	5	5
	21521728	11	7	4	6	5
Hypertensive	23602289	17	14	3	10	7
	22795718	14	13	1	9	5
	23435582	6	4	2	5	1
	22854636	5	3	2	2	3
	23223091	7	6	1	6	1
	22086840	15	7	8	10	5
TOTAL		259	180	79	165	94

TABLE 3.5: Claims classes and type distribution among the groups

Table (3.6) shows examples of the claim sentences extracted from the abstracts of the studies described in Table (3.3).

	PMID	Claim	Value	Type
11a	18954578	In this large, population-based sample of African-American and white adults, whole-grain intake was associated with lower HF risk, whereas intake of eggs and high-fat dairy were associated with greater HF risk after adjustment for several confounders	yes	caus
12a	19789394	Our findings do not support a major role for fish intake in the prevention of heart failure	no	caus
13a	20332801	Moderate consumption of fatty fish (1-2 servings per week) and marine omega-3 fatty acids were associated with a lower rate of first HF hospitalization or death in this population	yes	caus
14a	15963403	Among older adults, consumption of tuna or other boiled or baked fish, but not fried fish, is associated with lower incidence of CHF	yes	caus
15a	21610249	Increased baked/broiled fish intake may lower HF risk, whereas increased fried fish intake may increase HF risk in postmenopausal women	yes	caus

TABLE 3.6: Claims extracted from the abstracts of the studies listed in Table (3.3)

The main reason for disagreement in the first annotation task involved cases where there was more than one claim in the same study abstract that provided a potential answer to the question formulated from the systematic review. For example, Table (3.7) lists two sentences, (16) and (17), which were extracted from the same abstract and which are both potential answers to the question, *“In women with pre-eclampsia, is polymorphism in angiotensin gene associated with pre-eclampsia?”*. In such cases, the annotators were asked to favour sentences found in the conclusion sections of the abstracts.

	PMID	Sentence	Value	Type
16	15082899	The frequency of T allele of angiotensinogen T174M gene was slightly increased, but not significantly, in preeclampsia (0.11) than in controls (0.07)	yes	caus
17	15082899	In conclusion, a molecular variant of ACE, but not angiotensinogen, gene is associated with preeclampsia in Korean women	yes	caus

TABLE 3.7: Potential answers to a formulated question from the same abstract

The contingency table (3.8) shows the distribution of annotations for the first task (see Section (3.3.3)). The observed agreement (A_0), was remarkably high at 0.98; the chance agreement score (A_e) was 0.83; and the kappa score (k) was 0.88.

		Annotator#1		
		Claim	Non-Claim	
Annotator#2	Claim	239	20	259
	Non-Claim	20	2530	2550
		259	2560	2809

TABLE 3.8: The contingency table of claims identification

Disagreement between annotators in the second annotation task arose from claims that described opposite outcomes related to the same proposition. Although the formulation of a question for each group eliminated most examples of this problem, in some situations the question was not specific enough and multiple inferences leading to opposite assertion values could be derived from the same claim. For example, the question “*In the elderly, is n-3 fatty acid from fish intake associated with reduction in risk of developing heart failure?*” queried the association between *n-3 fatty acid from fish* and the risk of *developing heart failure*. Note that the question did not specify the type of fish (boiled/baked/fried). Table (3.9) shows multiple inferences derived from claims (14a) and (15a) in Table (3.6). Inferences (14a-1) and (14a-2) (Table (3.9)) were extracted from claim (14a) (Table (3.6)), and inferences (15a-1) and (15a-2) extracted from claim (15a). Each inference implies a different assertion value with respect to the question.

	PMID	Conclusion	Value
14a-1	15963403	consumption of tuna or other broiled or baked fish is associated with lower incidence of CHF	yes
14a-2	15963403	fried fish is not associated with lower incidence of CHF	no
15a-1	21610249	Increased baked/broiled fish intake may lower HF risk	yes
15a-2	21610249	Increased fried fish intake may increase HF risk in postmenopausal women	no

TABLE 3.9: Multiple inferences derived from two claims

In such situations, the annotators were asked to select the assertion value of the claim based on the inference that was the best fit for the question. In this case, the assertion value of inference (14a-1) was used as the assertion value to claim (14a), since it was deemed more general than the second inference which only included information about fried fish. The assertion value of claim (15a) was extracted from inference (15a-1), since (15a-2) only contained information about one gender (female).

The contingency table (3.10) shows the distribution of annotations for the second task, which determined whether the claim agreed with the question or not. The observed agreement (A_0), was remarkably high at 0.97; the chance agreement score (A_e) was 0.48 ; and thus, the kappa score (k) was 0.94. This high score indicates that the problem is well defined.

		Annotator#1		
		Yes	No	
Annotator#2	Yes	176	5	181
	No	4	74	78
		180	79	259

TABLE 3.10: The contingency table of annotating whether the claims agreed or disagreed with the questions

The main reason for this low score was due to the disagreement arose from claims that could potentially be interpreted as causal or evaluative at the same time. For example, the claim “*These results suggest that HLA-DR4 antigen may be a genetic marker for susceptibility to dilated cardiomyopathy*”, can be considered a causal claim, since it describes an association relationship between the two concepts *HLA-DR4 antigen* and *dilated cardiomyopathy*. At the same time, however, this claim can be considered evaluative, since it can be understood as the authors evaluation of whether that gene can be used as a marker for *susceptibility to dilated cardiomyopathy* or not.

In these scenarios, the annotators were reminded that *evaluative* claims typically describe the author’s judgement of, for example, certain medical processes, events or entities; where this was not the case, claims should be annotated as *causal*.

The contingency table (3.11) shows the annotation distributions for the annotations of the third task, which determined the claim type (causal/evaluative), among the annotators. The observed agreement (A_0) was 0.86; the chance agreement (A_e) score was 0.57; and the kappa score (k) was 0.67.

		Annotator#1		
		Causal	Evaluative	
Annotator#2	Causal	158	24	182
	Evaluative	13	64	77
		180	79	259

TABLE 3.11: The contingency table of annotating the claims types

The high inter-annotator agreement figures observed for the corpus annotation, particularly for identification of the claim and whether it agreed with the question or not, indicate that the annotations contained in the corpus are reliable and form a sound basis for further analysis. Although agreement for the claim type identification is lower, the

information provided by this annotation may still be relevant for further exploration. Moreover, analysis of systematic reviews and their associated forest plot diagrams was found to be a useful way of identifying potentially contradictory claims.

Identification and analysis of contradictory claims is a complex issue, and a number of associated challenges were identified during the construction of our corpus. First, claims tend to appear at the end of a study abstract and authors may often use shorter word forms and acronyms when referring to concepts, such as the example shown in the claim “*Our observations indicate a significant relationship between p22phox C242T and PARP-1 Val762Ala polymorphisms, CAD and its severity, but not with occurrence of MI in T2DM individuals with significant coronary stenoses*”. This posed an additional challenge to claim identification, particularly since acronyms are often ambiguous in medical text (Okazaki, Ananiadou, and Tsujii, 2010; Stevenson et al., 2009). Second, authors use a range of terms for similar concepts, making it difficult to identify connections between statements. For example, *statin*, *atorvastatin* and *rosuvastatin* were all used to refer to drugs that lower cholesterol levels in studies included in our corpus.

The construction methodology of *ManConCorpus* has two main limitations. First, the time and effort required to construct the corpus was considerable although this is normal in the NLP domain. Second, the corpus size may be insufficient to develop a reliable machine learning system to predict contradictory claims. Therefore, it would be of value to identify an alternative approach to construct a larger corpus than the *ManConCorpus* in a shorter time.

3.4 Automatically Annotated Contradiction Corpus

This section presents an alternative construction approach to generate another contradiction corpus called *AutConCorpus*. The corpus is automatically annotated using the incompatibility that arises between the relation tuples extracted by *SemRep* (Rindfleisch and Fiszman, 2003) and stored in *SemMedDB* (Kilicoglu et al., 2012) (see Section (3.4.1)). *AutConCorpus* is approximately twice the size of *ManConCorpus* but was generated more quickly.

3.4.1 SemMedDB

SemMedDB is a repository that contains all relation tuples extracted from the entire set of PubMed abstracts (processed up to June 2015) using *SemRep*. The repository consists of 82,239,652 tuples extracted from 25,027,441 abstracts and stored in a MySQL relation database. Table (3.12) shows the description of the database tables in the repository.

Table Name	Content
Citations	Metadata relevant for each PubMed citation
Concepts	Relevant information about UMLS Metathesaurus concepts.
Concept_Semtype	One-to-many relationships between concepts and their semantic types from UMLS semantic network.
Predication	Unique predicate.
Predication_Argument	Links between each predicate and its subject and object contained in <i>Concept</i> table.
Predication_Aggregate	Convenience table that aggregates information from all of the tables above for more efficient access.
Sentence	Sentences from each PubMed citation.
Sentence_Predication	Links between a sentence and a predicate extracted from it.

TABLE 3.12: *SemMedDB* database tables

SemRep (Rindfleisch and Fiszman, 2003) is a rule-based system that extracts relation tuples from biomedical text. A relation tuple takes the form of *subject-predicate-object*. The system extracts the tuples using multiple resources including the *SPECIALIST* lexicon (McCray, Srinivasan, and Browne, 1994), *Semantic Network* (McCray, Burgun, and Bodenreider, 2001), *MedPost* (Smith, Rindfleisch, and Wilbur, 2004) and *Metathesaurus* (Bodenreider, 2004). The *SPECIALIST* lexicon and *MedPost* are used in the first step of the extraction of tuples from a sentence, in order to generate a shallow syntactic analysis of the sentence and identify syntactic elements such as simple nouns phrases and verbs. The noun phrases are then mapped to their equivalent concepts in *Metathesaurus* using *MetaMap* (Aronson, 2001), and considered as potential arguments for a relation tuple. The other syntactic elements including the verbs, normalization and prepositions are processed further using a set of predefined rules to match them with the predicates available in *Semantic Network*. Among the rules are negation rules to recognize negated biomedical concepts.

SemRep relies on *NegEx* to identify negated biomedical concepts. *NegEx* considers three types of lexical cues to identify negated concepts: terms such as *denies* that indicate a negation, negation terms such as *no increase* that contain a negation term but do not negate the concepts; and terms such as *but* that terminate the scope of a negation (Chapman et al., 2013). *NegEx* applies different actions for each cue. For example, negation terms and pseudo-negation modify the information to the right of the term, and termination terms stops the negation scope otherwise the scope continues to the end of the sentence. *NegEx* has been evaluated on clinical text. The performance varied between an F-score of 75%-95% and an recall of 73%-85%. Errors reported were mainly

due to lexicon coverage and scope detection.

Table (3.13) shows an example of a sentence processed by *SemRep*. the system will identify multiple elements including *without*, *decrease* and noun phrases such as “*fish oils*” and “*cholesterol*”. The noun phrases are processed by *MetaMap* to identify their semantic types, which would be *Biologically Active Substance (bacs)* and *Lipid (lipd)* for the first phrase and *Biologically Active Substance* and *Steroid* for the second phrase. Furthermore, the term *decrease* is processed by the *Semantic Network* relations to find its match, which is in this case the predicate *INHIBITS*. Because *Semantic Network* contains the pattern *Biologically Active Substance-INHIBITS-Biologically Active Substance* in its semantic network, which is equivalent to the noun phrases and the verb extracted from the sentence, moreover since the term *without* in the sentence implies negation, *SemRep* recognizes the tuple *Fish Oils-NEG_INHIBITS-Cholesterol* as a correct relation tuple extracted from the sentence. Figure (3.3) shows a visual diagram to explain how *SemRep* constructs a relation tuple.

	PMID	Text
18	11077510	Intake of fish, fish oils and alpha-linolenic acid has positive effects on several clinical end points, often without marked decrease in serum cholesterol

TABLE 3.13: A sentence extracted from a PubMed abstract

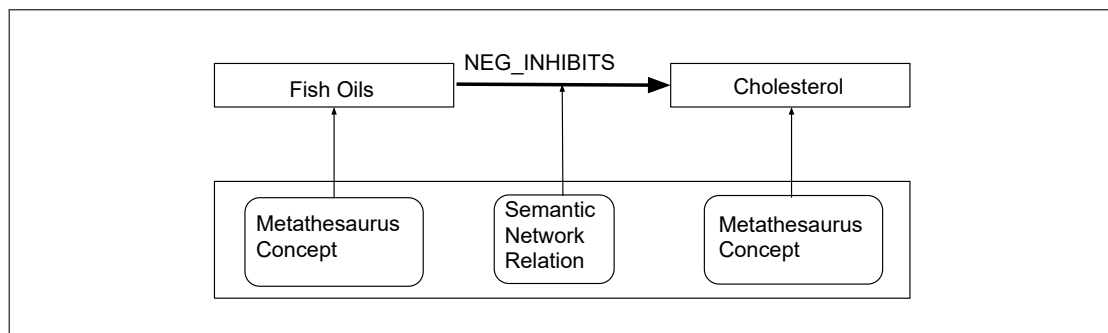


FIGURE 3.3: A diagram of how *SemRep* extracts a semantic predicate

SemRep has not been formally evaluated, due to the lack of a gold standard corpus; however, many task-based evaluations have been reported. For example, Rindfleisch et al. (2003) evaluated 830 instances of *ISA* predicates and reported 83% precision. Rindfleisch and Fiszman (2003) also evaluated 1,124 sentences containing tuples on genetic aetiology of disease and reported 76% precision. Ahlers et al. (2007) constructed a reference standard for relation tuples on pharmacogenomics and annotated 623 of them, reporting 55% recall and 73% precision. *SemRep* generates some errors but provides output that has proved useful for downstream processing.

3.4.2 Hypothesis to Collect Contradictory Sentences

The identification of incompatible predicates is relatively straightforward. *SemMedDB* includes a table called *Predication* which contains 58 predicates, the majority of which are listed in Table (3.14). The other predicates used by *SemMedDB* are: *than as*, *ISA*, *same as*, *compared with*. These predicates are categorised into six groups: group (A) shows excitatory predicates, group (AN) is the negation of group (A), group (B) shows inhibitory predicates, group (BN) is the negation of group (B), group (C) shows predicates that are neither excitation or inhibitory and group (CN) is the negation of group (C) (except the last row).

Predicates P_1 and P_2 are considered incompatible in the following cases:

1. if $P_1 \in group(A)$ and $P_2 \in group(AN)$ or if $P_2 \in group(A)$ $P_1 \in group(AN)$.
(e.g. *PRODUCES* and *NEG_AUGMENTS*)
2. if $P_1 \in group(B)$ and $P_2 \in group(BN)$ or if $P_2 \in group(B)$ $P_1 \in group(BN)$.
(e.g. *DISRUPTS* and *NEG_PREVENTS*)
3. if $iP_1 \in group(A)$ and $P_2 \in group(B)$ or if $P_2 \in group(A)$ $P_1 \in group(B)$.
(e.g. *AUGMENT* and *PREVENT*).
4. if $P_1 \in group(C)$ and $P_2 \in group(CN)$ or if $P_2 \in group(C)$ $P_1 \in group(CN)$,
where:
 $P_1 = Neg_P_2$ or $P_2 = Neg_P_1$.
(e.g. *PROCESS_OF* and *NEG_PROCESS_OF*)

Group(A)	Group(AN)	Group(B)	Group(BN)
AUGMENTS	NEG_AUGMENTS	DISRUPTS	NEG_DISRUPTS
CAUSES	NEG_CAUSES	INHIBITS	NEG_INHIBITS
COMPLICATES	NEG_COMPLICATES	PREVENTS	NEG_PREVENTS
PREDISPOSES	NEG_PREDISPOSES		
PRODUCES	NEG_PRODUCES		
STIMULATES	NEG_STIMULATES		
Group(C)		Group(CN)	
ADMINISTERED_TO		NEG_ADMINISTERED_TO	
AFFECTS		NEG_AFFECTS	
ASSOCIATED_WITH		NEG_ASSOCIATED_WITH	
COEXISTS_WITH		NEG_COEXISTS_WITH	
CONVERTS_TO		NEG_CONVERTS_TO	
DIAGNOSES		NEG_DIAGNOSES	
INTERACTS_WITH		NEG_INTERACTS_WITH	
LOCATION_OF		NEG_LOCATION_OF	
MANIFESTATION_OF		NEG_MANIFESTATION_OF	
METHOD_OF		NEG_METHOD_OF	
OCCURS_IN		NEG_OCCURS_IN	
PART_OF		NEG_PART_OF	
PRECEDES		NEG_PRECEDES	
PROCESS_OF		NEG_PROCESS_OF	
TREATS		NEG_TREATS	
USES		NEG_USES	
higher_than		NEG_higher_than	
lower_than		NEG_lower_than	
higher_than		lower_than	

TABLE 3.14: *SemMedDB* predicates

Table (3.15) shows two pairs of examples of incompatible tuples and Table (3.16) shows the sentences from which they were extracted. The first pair of relation tuples are incompatible because of the excitation factor: one predicate (AUGMENT) signifies an excitatory relationship while the other (DISRUPTS) indicates an inhibitory one. Consequently, the sentences related with them ((19) and (20)) are considered potentially contradictory. Similarly, the second pair relation tuples are incompatible because of negation, with the predicates negating each other and therefore sentences (21) and (22) are also considered potentially contradictory.

Subject	Predicate	Object
aspirin	AUGMENT	blood pressure
aspirin	DISRUPT	blood pressure
spironolactone	INHIBITS	blood pressure
spironolactone	NEG_INHIBITS	blood pressure

TABLE 3.15: Examples of Incompatible relation tuples extracted from sentences in Table (3.16)

	Sentence	PMID
19	.. the blood pressure of SHR below 160 mmHg was increased by aspirin	8401941
20	A highly significant blood pressure reduction was, however, observed in the patients who received aspirin before bedtime...	12732586
21	.. spironolactone appears to lower blood pressure compared to placebo ..	20687095
22	.. antagonist spironolactone was administered in a subtherapeutical dose, not lowering the blood pressure, and hydralazine ..	23104102

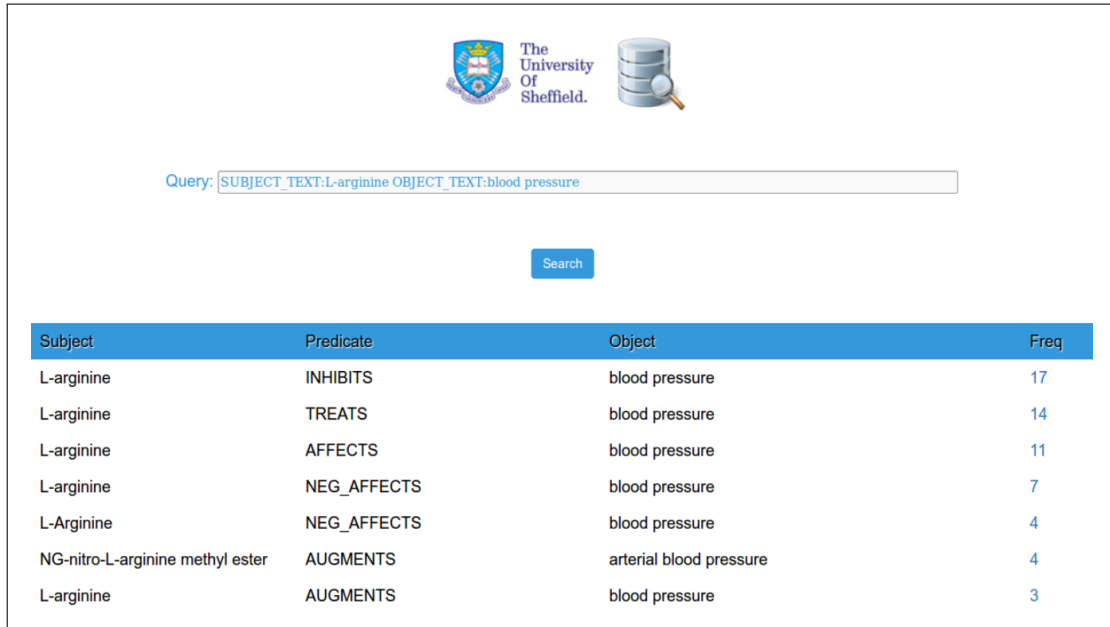
TABLE 3.16: The sentences extracted from *SemMedDB* based on the incompatibility of their relation tuples described in Table (3.15)

3.4.3 SemMedDB Browser

The *SemMedDB* repository is stored in a MySQL database environment. This environment is considered unsuitable for optimised browsing and exploring of the database to compile contradictory datasets, especially when the database contains millions of records distributed in multiple tables. Searching for such information requires complex syntax queries that may include queries to join multiple database tables in order to link the information in different tables into one consolidated view.

These issues led us to the development of *SemRebDB Browser*, a web-based application that enables researchers to browse and retrieve potential contradictory sentences from *SemMedDB* in an efficient and convenient way. The browser consists of two main components: a search engine to store the part of the *SemMedDB* content relevant to the corpus construction task, and interfaces to allow users to interact with the search engine. The search engine employs *Apache Lucene*[®] technology, an open source information retrieval library commonly used for its text indexing and search capabilities. The engine indexes the following information: the subject argument, the predicate, the object argument, and the sentence used to extract the tuple. This information is retrieved by joining the tables *Predication_Aggregate* and *Sentence* in *SemMedDB*. Other meta-data information such as the *predicate id* and the *abstract id* are also stored in the engine. The search engine acts as an information holder, thus providing much faster performance

compared with the database. The interfaces use *Java Servlet*[®] technology, a well known programming language for web applications. Two interfaces are used to access the search engine.



Subject	Predicate	Object	Freq
L-arginine	INHIBITS	blood pressure	17
L-arginine	TREATS	blood pressure	14
L-arginine	AFFECTS	blood pressure	11
L-arginine	NEG_AFFECTS	blood pressure	7
L-Arginine	NEG_AFFECTS	blood pressure	4
NG-nitro-L-arginine methyl ester	AUGMENTS	arterial blood pressure	4
L-arginine	AUGMENTS	blood pressure	3

FIGURE 3.4: The main interface of the *SemMedDB* Browser system

Figure (3.4) shows the main interface, which consists of a query field that accepts three types of information in any combination: subject, predicate and object. This information is entered using a specific pattern. For example, to search for a subject, the keyword `SUBJECT_TEXT:` followed by the subject text is entered; similarly, to search for an object, the keyword `OBJECT_TEXT:` followed by the object text is entered, and the keyword `PREDICATE:` to search for predicate text. Furthermore, the query accepts patterns like `NEG_*` to search for negative predicates such as `NEG_CAUSES` or `NEG_AFFECTS`. The results of the query appear in the same interface, consisting of a list of aggregated relation tuples along with their frequencies in the index, in the form of *subject-predicate-object-frequency*, where the frequency number is a hyperlink to the sentences that contain that tuple.

Figure (3.5) shows the second interface displayed after clicking the frequency hyperlink. The interface displays the sentences of a particular relation tuple along, with other meta-data information such as the predicate id and the abstract id. This information is partitioned using pipe delimiters to permit the user to copy and paste the content into a CSV file for downstream processing.

Search

Sentences

7608279_22984865|L-arginine|INHIBITS|blood pressure||The iv administration of L-arginine, a precursor of endothelium-derived relaxing factor/nitric oxide, is known to decrease blood pressure in humans by its direct vasodilatory effects.

7608279_22984874|L-arginine|INHIBITS|blood pressure||The mechanism by which L-arginine infusion decreases blood pressure can be at least in part explained by inhibition of the renin-Ang system.

7635532_23109951|L-arginine|INHIBITS|blood pressure||Compared with rats adapted to a high salt diet, those adapted to a low salt diet were more sensitive to the reductions in blood pressure and renal vascular resistance (threshold dose of L-arginine for renal vascular resistance: low salt, 2.9 +/- 0.9 mmol . kg-1 versus high salt, 20.0 +/- 6.2; P < .025), but the maximal changes in renal vascular resistance were similar (low salt, -43 +/- 5% versus high salt, -34 +/- 5%; P = NS).

7795361_22916336|L-arginine|INHIBITS|blood pressure||The infusion of L-arginine resulted in a decrease in blood pressure.

FIGURE 3.5: Sentences that contain a particular tuple

3.4.4 Corpus Construction

AutoConCorpus was developed based on the same research questions used in *ManConCorpus* corpus. Queries were formulated for each of the research questions contained in the corpus and used to interrogate *SemMedDB* using the interface described in the previous Section. For example, the query [SUBJECT.TEXT:L Arginine OBJECT.TEXT:blood pressure] is formulated from the question “In women with pre-eclampsia, does treatment with L Arginine as compared to placebo reduce blood pressure”. Figure (3.4) above shows the results of the above query. The first tuple in the results includes the predicate *INHIBITS*, which represents an inhibitory effect from the subject *L-arginine* on the object *blood pressure*; this tuple is generated from a total of seventeen sentences. Moreover, the seventh tuple shows the predicate *AUGMENTS*, which represents an excitatory effect of *L-arginine* on *blood pressure* and this tuple is generated from three sentences. Although the tuples share the same arguments their predicates are incompatible. Consequently, the sentences associated with them are considered potentially contradictory and are included in the corpus.

Assertion values of the sentences are assigned based on a reference predicate that is selected randomly from one of the predicates used in the these tuples, for example, in this case the reference predicate was *INHIBIT* and thus sentences that hold tuples that contain predicates that fall within its category (group (B)) are assigned the assertion value *yes*, and sentences containing incompatible predicates such as those from group (A) or group (BN) are assigned the assertion value *no*.

The construction of *AutConCorpus* consisted of two stages: retrieval and annotation. The retrieval stage was manually carried out only to find sentences that match topics used in *ManConCorpus*. However, the annotation stage’s automation was extremely easy given that the information related to relation tuples was stored in table

Predicate_Argument, which consists of multiple columns including *Subject_text* column, *Object_Text* column and *Predicate_Text* column, and given that accessing those columns to discover similar relation tuples was straightforward. The annotation stage was automatically carried out using the incompatibility cases described in Section (3.4.2).

3.4.5 Results and Discussion

Composed of 526 sentences, *AutConCorpus* features approximately double the number contained in *ManConCorpus*. The corpus data were collected from 13 topics. Table (3.17) shows the distribution of the corpus sentences among the topics and their associated assertion values. For example, *Topic-1* in the table consists of three groups of sentences that describe the effect of *L-arginine* on *blood pressure*. The first group in that topic consists of 17 sentences, the second group, three sentences, and the third group, four sentences. Since the first group shows a deactivation effect which was incompatible with the second and third groups, their sentences were assigned the value 1, while the sentences belonging to the second and the third groups were assigned the value 0.

Some errors were found in *AutConCorpus*. First, some relation tuples were inaccurately extracted by *SemRep*. For example, Table (3.18) shows that sentences (23) and (24) were included in the corpus as contradictory even though they were not. This problem arose because the predicates extracted using *SemRep* were incompatible (*NEG_TREATS* and *TREAT*). However, a closer look at sentence (23) reveals that the predicate *NEG_TREATS* was inaccurately extracted, and that the correct predicate should be *TREAT*. Such errors may cause a reduction in the quality of the corpus.

	PMID	Sentence
23	18650598	Surprisingly, the <i>ACE inhibitors</i> proved to be effective <i>not</i> only in patients with high renin <i>hypertension</i> , but also in many patients with normal levels of plasma renin activity.
24	3154305	<i>ACE inhibitors</i> have emerged as important pharmacologic agents for treatment of <i>hypertension</i> and heart failure

TABLE 3.18: Non-contradictory sentences that were included in the corpus as contradictory

Furthermore, it was observed that the research abstract titles featured among the information processed *SemRep*. Titles usually do not report information, and therefore titles may not explicitly describe a relationship between biomedical concepts. For example, Table (3.19) shows a title (25) that was included in the corpus as *SemRep* extracted from it the tuple *ACE inhibitors-TREATS-hypertension*.

	Subject	Predicate	Object	Value	#Sen.
Topic-1	L-arginine	INHIBITS	blood pressure	1	17
	L-arginine	AUGMENTS	blood pressure	0	3
	L-Arginine	CAUSES	blood pressure	0	4
Topic-2	ACE inhibitors	TREATS	hypertension	1	58
	ACE inhibitors	NEG.TREATS	hypertension	0	5
Topic-3	statin	AUGMENTS	blood pressure	1	1
	statin	DISRUPTS	blood pressure	0	1
Topic-4	aspirin	TREATS	bleeding	1	36
	aspirin	CAUSES	bleeding	0	31
	aspirin	INHIBITS	bleeding	1	8
	aspirin	PREDISPOSES	bleeding	0	17
	aspirin	AUGMENTS	bleeding	0	22
	aspirin	NEG_AUGMENTS	bleeding	1	8
Topic-5	insulin	AUGMENTS	glucose transport	1	13
	insulin	NEG.AFFECTS	glucose transport	0	6
Topic-6	insulin	AFFECTS	lipogenesis	1	15
	insulin	NEG.AFFECTS	lipogenesis	0	4
	insulin	AUGMENTS	lipogenesis	1	13
	insulin	DISRUPTS	lipogenesis	0	5
Topic-7	fibrosis	PROCESS_OF	mice	1	59
	fibrosis	NEG.PROCESS_OF	mice	0	25
Topic-8	bone marrow	LOCATION_OF	cells	1	84
	bone marrow	NEG.LOCATION_OF	cells	0	12
Topic-9	insulin	TREATS	rats	1	9
	insulin	NEG.TREATS	rats	0	8
Topic-10	fish oil	STIMULATES	cholesterol	1	4
	fish oil	INHIBITS	cholesterol	0	2
Topic-11	statin	AFFECTS	cancer	1	2
	statin	NEG.AFFECTS	cancer	0	2
Topic-12	spironolactone	INHIBITS	blood pressure	1	17
	Spironolactone	NEG.AFFECTS	blood pressure	0	5
	Spironolactone	NEG.INHIBITS	blood pressure	0	2
Topic-13	atorvastatin	STIMULATES	cholesterol	1	5
	atorvastatin	INHIBITS	cholesterol	0	21

TABLE 3.17: *AutConCorpus* topics, relation tuples and sentence distributions

	PMID	Sentence
25	11243672	ACE inhibitors for hypertension

TABLE 3.19: A title included in *AutConCorpus*

AutConCorpus was also formatted in XML for ease of processing. Figure (3.6) shows examples of two formatted contradictory sentences. The sentences are grouped in topics, where each topic has three main attributes as a reference to identify the contradictory sentences. According to the hypothesis described in section (3.4.2), the left argument (*L_ARGUMENT*) and the right argument (*R_ARGUMENT*) in all sentences that belong to a topic should be the same. However, the tuples might be different.

Therefore, a tuple from the sentences in the topic is randomly chosen as the *REFERENCE_PREDICATE* to identify the target values of the sentences within that topic. For example, the reference tuple in the examples in figure (3.6) is *AUGMENTS*, thus, the target value of the first sentence is 1 since it uses the same predicate, however, the target value of the second sentence is 0 since its predicate (*NEG_AFFECTS*) is incompatible with the reference predicate. The corpus is available from <https://drive.google.com/open?id=0B5sg2-DPQTMwbnBFdkExNlpBczg>.

```

<CORPUS>
<TOPIC MODIFIED_PREDICATE="INHIBITS">
<SENTENCE LARGUMENT="aspirin" PMID="16490462" PREDICATE="CAUSES" RARGUMENT="
bleeding" SEMREPDB_SID="82452704" TARGET="0">Because aspirin can cause major
bleeding, the appropriate dose is the lowest dose that is effective in
preventing both MI and stroke because these two diseases frequently co-exist.<
/SENTENCE>
<SENTENCE LARGUMENT="aspirin" PMID="19628366" PREDICATE="INHIBITS" RARGUMENT="
bleeding" SEMREPDB_SID="86938437" TARGET="1">The hypothesis of the present
study is that aspirin will decrease the rate of operative site bleeding
without increasing thromboembolic events when aspirin is used for VTE
prophylaxis after major orthopaedic surgery.</SENTENCE>
</TOPIC>
</CORPUS>

```

FIGURE 3.6: Examples of formatted sentences

ManConCorpus was mainly developed for use in the development of a contradiction detection system (see *Chapter (5)*). However, because the corpus size of *manConCorpus* was not that large, and it would have been expensive to generate another dataset following the same methodology, this research sought for a second methodology (the automatic methodology) capable of generating another dataset that could **assist** the development of the contradiction detection system using *ManConCorpus*, rather than replacing *ManConCorpus* itself. In other words, the purpose of *AutConCorpus* was to ensure that the linguistic features suggested by *ManConCorpus* were indeed reliable features to predict contradictory claims. One approach to determining reliability was to examine the capability of these features to recognize the annotations of *AutConCorpus*, which originally were obtained from *SemRep*.

Some may claim that the capability of the linguistic features suggested by *AutConCorpus* is not indicative of the quality of the features themselves given that the annotations were not manually evaluated. Notably, however, the automatic methodology did

not involve any kind of processing to *AutConCorpus*. It only involved grouping sentences stored in *SemMedDB* that share certain identical information (*Subject* and *Object* arguments), and then annotating the sentences based on the semantic meaning of the predicates associated with their relation tuples stored in the repository.

Thus, the quality of *AutConCorpus* and the quality of *SemRep* are assumed to be interdependent. If *SemRep* offers accurate relation tuples from *PubMed* sentences, then the information stored in *SemMedDB* is reliable as well. The available literature on *SemRep* reported that the quality of this tool varied between F-scores of 0.73 and 0.83 (see Section (3.4.1)); moreover, *SemRep* is a known biomedical relation extraction system and has been used for literature-based discovery and hypothesis generation applications¹. This research thus assumed that the information stored in *SemMedDB* is reliable as well and hence used that information to form a contradictory dataset to support the experiments conducted using *ManConCorpus*.

To ensure that the automatic methodology is reliable enough to generate a contradiction corpus, a validation stage should be incorporated after generating the corpus to examine its quality. The validation stage requires recruiting annotators to conduct two tasks that are far easier than the effort required using the manual methodology. The validation stage aims to validate that (1) the sentences grouped in a topic are actually describing the same preposition and that (2) these sentences are annotated with the correct assertion values. The annotators are required only to examine the quality of existing information; they do not have to produce any information.

Another approach to consider for evaluating *AutConCorpus* would be to measure its overlap with *ManConCorpus*. However, that approach may not be possible for two main reasons. First, *SemMedDB* stores only information about sentences and their associated relation tuples extracted by *SemRep*; therefore, sentences that do not contain such relation tuples are not included in *SemMedDB*. *SemRep* extracts only relation tuples of specific patterns of sentences (see Section (3.4.1)), and those patterns are not necessarily available in many of the sentences used in *ManConCorpus*. Therefore, it is difficult to compare *ManConCorpus* with *AutConCorpus*². For example, Table (3.20) shows a list of claims that belong to a systematic view and were included in *ManConCorpus*. These claims were supposed to provide information to answer the question “*In patients with HCM, does using imaging technique serve as a predictor for adverse prognosis*”. When *SemRep* was run over these sentences to extract relation tuples, the tool failed to extract any relation tuples from them.

¹<https://skr3.nlm.nih.gov/>

²*SemRep* was unable to extract relation tuples from more than 60% of *ManConCorpus*.

	Claim sentence
19808288	These data suggest an important role for myocardial fibrosis in the clinical course of HCM patients but are not sufficient at this time to consider DE as an independent risk factor for adverse prognosis.
19850699	If replicated, LGE may be considered an important risk factor for sudden death in patients with HCM.
20667520	Among our population of largely low or asymptomatic HCM patients, the presence of scar indicated by CMR is a good independent predictor of all-cause and cardiac mortality.
20688032	In patients with HCM, myocardial fibrosis as measured by late gadolinium enhancement cardiovascular magnetic resonance is an independent predictor of adverse outcome.

TABLE 3.20: Examples of sentences in *ManConCorpus* from which *SemRep* failed to extract any relation tuples from.

Second, even if the *SemRep* was able to extract some relation tuples from *ManConCorpus*, it is unlikely that these tuples will overlap with the relation of *AutConCorpus*. For example, Table (3.21) shows a list of relation tuples extracted from another set of claims that were supposed to describe the same information in *ManConCorpus*, however, none of them share the same *Subject* and *Object* arguments, therefore, it becomes difficult to apply the same strategy used in the automatic methodology to the sentences used in *ManConCorpus*.

Claim PMID	Subject	Predicate	Object
19559191	Aspirin	TREATS	patient
21050973	clopidogrel	higher_than	Aspirin
22942294	Aspirin	TREATS	Hemorrhage
22942294	clopidogrel	TREATS	Hemorrhage

TABLE 3.21: Examples of relation tuples extracted from a set of claims in *ManConCorpus*. These tuples do not exist within *AutConCorpus*.

Chapter (5) found *AutConCorpus* useful in the development of a contradiction system using *ManConCorpus*. This finding supports the hypothesis that the *automatic methodology* could be a potential alternative for the manual methodology. This research does not suggest use the corpus *AutConCorpus* itself as an alternative for *ManConCorpus*, but it does suggest use of the automatic methodology to construct a contradiction corpus. Constructing this corpus may include a validation stage to ensure the quality of the generated instances.

3.5 Conclusions

The methods employed in the construction of two corpora to develop machine learning systems for the detection of contradictory claims in PubMed research abstracts are detailed in this chapter. *ManConCorpus* was generated using a standard NLP approach, which included annotators. The corpus contains contradictory claim sentences extracted from abstracts from systematic reviews. The analysis showed that the agreement between annotators is reliable, suggesting that the information in the corpus will be useful in the development of machine learning to detect contradictory claims.

AutConCorpus was automatically generated from the *SemMedDB* repository without annotators. The corpus was retrieved with the support of the *SemMedDB Browser* application, developed to facilitate the process of constructing the corpus. The corpus was constructed to examine the suitability of an automatically generated corpus for the development of a contradiction detection system. The methods used for the development of both corpora may be of benefit in the establishment of other larger corpora for the same objective.

Chapter 4

Identification of Research Claims in Biomedical Abstracts

4.1 Introduction

The initial stage of our approach to identify contradictory claims is to allocate sentences containing the claims within abstracts. This chapter presents a pipeline system that automatically identifies the sentences containing research claims in abstracts that are relevant to a given query. The pipeline consists of two components: one to identify the abstract sections likely to contain the sentences describing the claims of the research (claim zone), and another to determine which sentences within that zone are likely to be relevant to the research claim and capable of answering the user's question. The claim zone in an abstract is the segment of text that is mostly likely to contain the sentences describing the research claims. Such sentences are largely found within the *Conclusions* sections and occasionally within the *Results* sections (see next *Section 4.2*). Therefore, this research assumes that the claim zone is the sentences within the *Results* and *Conclusions* sections.

The first component takes an abstract as an input, assuming it is relevant to the user's query, and returns the sentences within the *Results* and *Conclusions* sections as an output. The abstract can be either structured or unstructured. In structured abstracts, the sentences that belong to the *Results* and *Conclusions* are automatically extracted (see *Section 4.2*) as the claim zone. For unstructured abstracts, a machine learning classifier labels each sentence with its rhetorical role (Teufel and Moens, 2002). Sentences belonging to the *Results* and *Conclusions* roles are extracted as the sentences in the claim zone.

The second component is another machine learning classifier that considers sentences in the claim zone as input, with the purpose of obtaining the most informative sentence(s) likely to represent a claim and contain information that addresses the given

query.

This chapter describes three contributions: the development of a classifier to identify the rhetorical roles of sentences and the comparison of it with a current state of art system; the development of a classifier to detect research claims that are potential answers to a given query and the introduction of a novel feature termed *Z-score*, beneficial for the comparison of similarity scores.

4.2 Previous Work Related to Claim Zoning Component

Rhetorical status, a characteristic that is often assigned to a text segment based on its role in the overall textual context, has been a topic of interest for some time. In summarization for example, Teufel and Moens (2002) developed a system to summarize scientific articles using the rhetorical status of the articles' sentences (see Section 2.4.2). One assumption that was made in that research was that the intellectual attribution was clearly described in the text, and that readers should have no difficulty in identifying and understanding that information (for example, claims made by the research authors).

Agarwal and Yu (2009) used rhetorical status to classify sentences in biomedical articles into four roles: Introduction, Methods, Results and Discussion (IMRAD). Their main motivation was that other text mining applications could benefit from categorizing sentences based on their rhetorical role. For example, question answering systems can target a particular category of sentences to find answers. The best performance was achieved using a Multinomial Naive Bayes approach.

Ruch et al. (2007) used rhetorical status to extract key information from biomedical abstracts. They regarded information that appeared in the *Conclusions* section as the *key* information to determine the abstract topic. That research used a Naive Bayes classifier to label sentences according to one of the following four categories: Purpose, Methods, Results and Conclusion.

Lin et al. (2009) considered sentences that appeared in the *Results* and *Conclusions* sections as the most important information, since they describe the main contribution of the research. They used a Conditional Random Fields (CRFs) algorithm and multiple features including position, named entity, tense and word frequency to sequentially annotate sentences to one of the three labels: Objective, Methods and Result-Conclusion.

Chung (2009) regarded sentences referring to Intervention, Participants and Outcome Measures in Randomized Control Trials (RCT) abstracts as the key sentences. They extended previous research on sentence labelling, based on the rhetorical roles, to label sentences in RCT abstracts. Their work demonstrated that CRFs were superior to

SVMs, using a range of elements such as word features, normalization of complex numerical and mathematical notation, POS, position and windowed features (for previous and next sentences).

Structured abstracts are those with distinct headings. Structured abstracts in PubMed use a variety of labels as headings. The National Library of Medicine (NLM) reported that 2,779 headings have been used as section heading labels (Ripple et al., 2012). These abstracts represent only 30% of the entire repository stored in PubMed, meaning that the other 70% in the repository are unstructured. Researchers therefore are required to devote more resources to the determination of key information within these abstracts.

Hirohata et al. (2008) used the labels assigned to the header sections of structured abstracts to label sentences in unstructured abstracts. However, this approach is not optimal, since such labels are not unified across all of PubMed and different abstracts may use different sets of labels.

Jimeno-Yepes et al. (2013) solved this issue using the values assigned to the *nlm-Category* attributes in each section of the abstracts in XML. The NLM annotates each section of a structured abstract with one of five values: Objective, Background, Methods, Results and Conclusions. These values are consistently used across all structured abstracts in PubMed.

4.2.1 Methods

The task of extracting the claim zone from structured abstracts becomes feasible, since it requires only the extraction of the sections with the values *Results* and *Conclusions* in the NLM category attribute. A machine learning classifier was thus developed to label sentences in unstructured abstracts according to the same categories as those used by Jimeno-Yepes et al. (2013). As the *Background* and *Objective* sections tend to overlap each other (Lin et al., 2009), and often appear sequentially, merging them into a new category called *Introduction* is advisable. Thus, the *Claim Zoning* component labels the sentences in unstructured abstracts with one of four possible categories: *Introduction*, *Methods*, *Results* and *Conclusions*.

4.2.2 Conditional Random Fields (CRFs)

Overall, structured abstracts typically feature some structural and sequential characteristics. For example, the *Introduction* section usually appears at the beginning of an abstract and the *Conclusions* section at the end; furthermore, it is unlikely that the *Results* section will appear after the *Conclusions*. In order to model these characteristics during

construction of the *Claim Zoning* component, a CRF (Sutton and McCallum, 2006) algorithm is used, an approach which has successfully been used for similar tasks (Hirohata et al., 2008; Jimeno Yepes, Mork, and Aronson, 2013; Lin et al., 2009)

CRFs represent a discriminative model that describes the conditional distribution of the observed features over a set of observed labels. Given a set of sentences in an abstract $x = (x_1..x_n) \in x^n$, CRF (Equation 4.1) computes the probability $p(y|x)$ of a possible label sequence $y = y_1..y_n \in Y^n$; where the labels in this case are the four possible categories described earlier.

$$p(y|x) = \frac{1}{Z(x)} \exp\left(\sum_{j=1}^n \sum_{i=1}^m \lambda_i f_i(y_{j-1}, y_j, x_j)\right) \quad (4.1)$$

The function $f_i(y_{j-1}, y_j, x_j)$ returns the value 1 if its corresponding feature is activated when moving from label y_{j-1} to y_j after observing x , otherwise it returns the value zero; λ_i is the weight of the feature f_i . If $\lambda_i > 0$ and $f_i = 1$, then the probability of moving from label y_{j-1} to y_j is increased; otherwise if $\lambda_i < 0$, then the probability of moving from label y_{j-1} to y_j is decreased.

4.2.3 Features

The *Claim Zoning* system employs various features that exploit the structural, sequential and lexical/syntactical features of sentences in abstracts, in order to recognize their roles.

N-grams: N-grams are lexical features shown to be of benefit for capturing the general context of text (Turney, 2002; Yu and Hatzivassiloglou, 2003). For every sentence, the uni-grams and bi-grams are extracted from the abstract's title t , the current sentence s_n , the previous sentence s_{n-1} , and the next sentence s_{n+1} . Although, it looks redundant to consider the n-grams of s_{n-1} and s_{n+1} as features in a sequential learning algorithm, previous research by Chung (2009) and Hirohata et al. (2008) showed that such features improved the overall performance.

Cosine Similarity $sim(s,t)$: We hypothesise that the lexical similarity between a sentence and the title of an abstract is a good indicator for the relevance of the sentence with respect to the research topic. The value of such similarity has already been shown in the context of summarization tasks (Teufel and Moens, 2002). Lexical similarity is captured using the cosine metric, which computes the angle between the sentence vector s and the title vector t in the vector space. These vectors are generated by computing the terms' *tf.idf*. Equation (4.2) shows the cosine similarity function. The scores are binned into 11 values, ranging from 0 to 10.

$$sim(s, t) = \frac{\sum_{i=1}^n (s_i \times t_i)}{\sqrt{\sum_{i=1}^n (s_i)^2} \times \sqrt{\sum_{i=1}^n (t_i)^2}} \quad (4.2)$$

Obtaining accurate similarity scores between sentences in biomedical abstracts is not a straightforward task. Authors in this domain frequently use the long form of concepts in the title and start of an abstract, but subsequently resort to the short form (e.g., abbreviation). For example, abstract (26860956), entitled *Comparison of diagnostic evaluations for cough among initiators of angiotensin converting enzyme inhibitors and angiotensin receptor blockers*, used the concept *angiotensin converting enzyme inhibitors* in the title and at the beginning of the abstract, with the abbreviation *ACEI* used thereafter. This approach can present a challenge, as the similarity score between the long form phrase and its abbreviation will be zero (despite being related to the same concept).

In an attempt to minimize this issue, we automatically unify the appearance of the concepts used in abstracts prior to computing the $sim(s, t)$, using an algorithm developed by Schwartz and Hearst (2003). That algorithm was mainly developed to extract the correct long-form of a concept that was mentioned in a text using its abbreviation. The algorithm uses patterns (i.e. *long-form(short-form)* and *short-form (long-form)*) to identify the correct long-form of an abbreviation. The *Claim Zoning* component uses this algorithm to replace all biomedical concepts mentioned in abbreviations with their long-forms.

Relative Location (*loc*): The location of a sentence within a document may provide some information about its rhetorical role. Rather than using the original location of the sentence as a feature, we use the relative location compared to other sentences by employing Equation (4.3). This function adjusts all sentences locations in an abstract to adopt the same scale, from 1 to 10.

$$loc = \frac{sentence_location \times 10}{abstract_size} \quad (4.3)$$

Main Verb Tense (*tense*): The tense of verbs used in sentences often correlates with its rhetorical status (Teufel and Moens, 2002). For example, authors often use the present perfect tense in the *Introduction* section but the past simple tense in the *Conclusions* section. This feature is extracted using the *Stanford parser* (Marneffe and Manning, 2008), which extracts the tense of the main verb (ROOT-0, verb) from every sentence. For example, the main verb of the sentence “*Therefore, we concluded that the original eins of dulse ACE inhibitory peptides were phycobiliproteins.*” is *concluded* and the tense is *VBD*.

4.2.4 Data

The *Claim Zoning* component is built using a collection of 10,000 structured abstracts retrieved from PubMed, using the query *cardiovascular disease*, to match the corpus constructed in Chapter (3). Figure (4.1) shows a part of a structured abstract which illustrates how heading section labels can differ from the labels assigned by NLM. The dataset uses the *NlmCategory* values assigned to each section to label the corpus. The corpus consists of 110,138 sentences, 21,319 of which belong to the *Introduction* section, 26,642 sentences to the *Methods* section, 42,758 to the *Results* section and 19,419 to the *Conclusions* section.

```

<ABSTRACT>
  <ABSTRACT.TEXT Label="BACKGROUND" NlmCategory="BACKGROUND">Little ..
</ABSTRACT.TEXT>
  <ABSTRACT.TEXT Label="OBJECTIVE" NlmCategory="OBJECTIVE">To..
</ABSTRACT.TEXT>
  <ABSTRACT.TEXT Label="DESIGN" NlmCategory="METHODS">Randomized ..
</ABSTRACT.TEXT>
  <ABSTRACT.TEXT Label="SETTING" NlmCategory="METHODS">10 European..
</ABSTRACT.TEXT>
  <ABSTRACT.TEXT Label="PARTICIPANTS" NlmCategory="METHODS">140..
</ABSTRACT.TEXT>
  <ABSTRACT.TEXT Label="INTERVENTION" NlmCategory="METHODS">Stent ..
</ABSTRACT.TEXT>
  <ABSTRACT.TEXT Label="RESULTS" NlmCategory="RESULTS">Forty-six ..
</ABSTRACT.TEXT>
  <ABSTRACT.TEXT Label="CONCLUSION" NlmCategory="CONCLUSIONS">Stent ..
</ABSTRACT.TEXT>
</ABSTRACT>

```

FIGURE 4.1: Abstract (19414832) is an example of how label values can differ from the *nlmCategory* values

4.2.5 Results and Discussion

The *Claim Zoning* component was implemented using a CRF algorithm (CRFSuite package (Okazaki, 2007)) and combinations of different sets of features to categorise sentences in unstructured abstracts into four groups according to their rhetorical role: *Introduction*, *Methods*, *Results* or *Conclusions*. The system was trained using a set of 7,000 abstracts, and tested on a different set consisting of 3,000 abstracts.

	Features	Classes	Precision	Recall	F1-score
1	uni+bi-grams of sentence s_n	Introduction	0.93	0.94	0.93
		Methods	0.87	0.88	0.87
		Results	0.88	0.88	0.88
		Conclusions	0.87	0.86	0.87
		Micro-Average	0.88	0.88	0.88
2	uni+bi-grams of sentence s_{n-1} , s_n and s_{n+1}	Introduction	0.94	0.95	0.94
		Methods	0.87	0.87	0.87
		Results	0.89	0.88	0.88
		Conclusions	0.89	0.89	0.89
		Micro-Average	0.89	0.89	0.89
3	uni+bi-grams of sentence s_{n-1} , s_n , s_{n+1} and the title t ,	Introduction	0.95	0.94	0.95
		Methods	0.86	0.88	0.87
		Results	0.89	0.88	0.89
		Conclusions	0.89	0.89	0.89
		Micro-Average	0.90	0.90	0.90
4	uni+bi-grams of sentence s_{n-1} , s_n , s_{n+1} and the title t and relative location	Introduction	0.96	0.95	0.95
		Methods	0.87	0.89	0.88
		Results	0.90	0.89	0.89
		Conclusions	0.91	0.90	0.91
		Micro-Average	0.90	0.90	0.90
5	uni+bi-grams of sentence s_{n-1} , s_n , s_{n+1} and the title t and relative location and the tense of the main verb	Introduction	0.95	0.95	0.95
		Methods	0.87	0.89	0.88
		Results	0.90	0.89	0.90
		Conclusions	0.92	0.90	0.91
		Micro-Average	0.91	0.91	0.91
6	uni+bi-grams of the top 200,000 that have high $\tilde{\chi}^2$ (chi-squared) in s , s_{n-1} and s_{n+1} and relative location (Hirohata et al., 2008)	Introduction	0.95	0.95	0.95
		Methods	0.87	0.88	0.87
		Results	0.89	0.90	0.89
		Conclusions	0.91	0.89	0.90
		Micro-Average	0.90	0.90	0.90

TABLE 4.1: The performance of *Claim Zoning* system. The baseline is 40.3%, which is the accuracy percentage of annotating all sentences with class *Results*

Table (4.1) shows the system results using different sets of features in micro-averaging. The baseline of the system is 40.3%, which is the average result of annotating all sentences with class *Results*. The leftmost column describes the set of features used in each configuration, while the following four columns show the system precision, recall

and F-score of each class. The first five configurations (1-5) show the set of features attempted to improve the performance of the system, while the configuration of (6) used the set of features proposed by Hirohata et al. (2008) and was implemented using our corpus. This system was implemented to compare our system with a state-of-the-art system.

The first configuration was developed using only simple n-gram features (unigrams and bi-grams). Such features were reported as effective in combination with text classification systems, and the absence of one may result in diminished system performance (Pang, Lee, and Vaithyanathan, 2002). The system produced an average F-score of 0.88.

The second configuration incorporated the n-grams of the previous and the next sentences s_{n-1} and s_{n+1} , in addition to the current sentence s_n . The performance slightly improved by 1% in terms of F1-score, as a result of improvements in detecting *Introduction* and *Conclusions* sentences. The improvement in detecting these sections may suggest a codependent relationship between the features of the previous/next sentences and those of current sentences, and the use of such features might be able to capture the sequential relationship existing between them.

The third configuration added the title (t) n-grams in addition to the previous features. This approach marginally enhanced the performance of detecting *Introduction* and *Results* sentences. Overall performance improved by 1% compared with configuration (2). Implementation (4) incorporated the relative location of sentences in addition to the existing features. The precision and recall scores for the *Methods* and *Conclusions* sections, but not the *Introduction* and the *Results*, were shown to have improved slightly. The overall performance of configuration (4) did not improve. The fifth configuration incorporated the tense of the main verb in addition to the other features, and overall performance improved by 1% compared with configuration (4). The average precision, recall and F1-score achieved using this configuration is 91%.

For evaluation purposes, the best configuration performance (5) was compared with configuration (6), a state-of-the-art system implemented by Hirohata et al. (2008). This system was considered the closest to *Claim Zoning*, since it uses the same algorithm and employs similar features. The results show that configuration (5) slightly outperformed configuration (6), particularly in detecting the *Results* and *Conclusions* sections which is essential to *Claim Zoning*.

The results of these configurations show that the difference between using simple features such as uni+bi-grams or more complicated features such as set (5) was not significant (88% vs. 91%). This finding suggests that significant improvement in such a system may require the incorporation of features other than those described here or in

previous work.

The best configuration (5) is considered to be relatively efficient for identifying claim zones in abstracts. This is due to the fact that many of the annotation errors occurred at section boundaries. For example, the errors occurring at the boundary of the *Introduction* and *Methods* sections are not problematic here since they both are not part of the claim zone. Similarly, the misclassification occurring in sentences at the boundary of the *Results* and *Conclusions* sections are also not problematic since they still exist within the claim zone and in any case both sentences will be passed to the *Answer Selection* Component.

4.3 Previous Work Related to Answer Selection Component

The *Answer Selection* component is an established component of QA (Hirschman and Gaizauskas, 2001). It mainly performs two primary tasks (Athenikos and Han, 2010): matching of the expected answer type, and ranking of qualified answers.

This work employs an *Answer Selection* component which may resemble that used in QA, but features some differences. First, the ultimate goal of this component is to identify the best sentence(s) from the claim zone that may answer a given question. This is distinct from QA, which tends to extract precise information such as names, locations, numbers or dates as answers. Second, its input is a list of sentences and its output is one or more sentences that are potential answers. In QA, however, the output is the top N qualified answers ranked according to a predefined matching score (Athenikos and Han, 2010). The *Answer Selection* in this work is less similar to those used in biomedical QA and more similar to the task *Answer Selection in Community Question Answering*, as presented in SemEval 2015 (Nakov et al., 2015).

The SemEval 2015 evaluation is an ongoing series of evaluations that develops multiple tasks, including one on answer selection in community question answering. The task is composed of two subtasks, one of which relates to the classification of answers (with respect to a given question) that appear in open community forums into: *good*, *potentially relevant* or *bad*. The automation of this type of process is beneficial for users striving to review all posts that answer a given question. The *Answer Selection* component is similar to this for two reasons. First, one of the goals of this research is to minimize the cognitive effort required by users when searching for abstracts relevant to a particular research question. Second, this component considers the task as a classification problem, but with the two-way classes *potentially answer* versus *non-potentially answer* rather than three-way classes.

The best three systems in this evaluation employed various algorithms and features, but mainly relied on SVM classifiers and features such as n-grams, text similarity, sentiment analysis, word vector representation, topic modelling and translation features.

Tran et al. (2015) has described the best system, developed as a classifier based on SVMs with multiple features including translation based features, topic model based features and word vector representation based features. The translation features were used, as they were found to be of benefit in matching similar questions in the QA archives (Zhou et al., 2011). The topic modelling features were used to compute the cosine similarity between the topic vectors of the question and its answers. The word vector representation features component was used to model the relevancy between questions and answers. These features were extracted using the pre-trained Word2Vec model (Mikolov et al., 2013). The system achieved an F-score of 0.57 and accuracy of 72.52%.

Hou et al. (2015), who described the second best available system, proposed an alternative solution based on ensemble learning and hierarchical classifiers (including SVMs) with more simple features than those in the best system. These features include n-grams, POS, named entities, word length and sentence length. The system achieved an F-score of 0.56 and an accuracy of 68.67%.

Nicosia et al. (2015), who reported the third best system, used SVMs and also applied a set of features including n-grams, lexical similarity, syntactic similarity, semantic similarity and sentiment analysis features. That system achieved an F-score of 0.53 and an accuracy of 70.5%.

4.3.1 Methods

Developed as a classification system, the *Answer Selection* component classifies sentences from the claim zone as *potential answer* versus *non-potential answer*. Three assumptions were considered in the design of this component. The first assumption is that sentences in an abstract that share many words with the title tend to express important information about the research topic. The second is that sentences that have a similarity with the title within a certain threshold, and exist within the *Results* and *Conclusions* sections, tend to be key sentences concerning the research topic (Lin et al., 2008a, 2009; Otani and Tomiura, 2014; Ruch et al., 2007). The third assumption is that sentences with the previous two characteristics that have a high similarity with a given question tend to contain information related to the question.

4.3.2 Support Vector Machines (SVMs)

Support Vector machines (SVMs) (Vapnik, 1982) have been shown to be highly effective for text classification problems. In two way classification problems, SVMs attempt to find a linear separation between the hyperplanes defined by the two classes, in which this separation (margin) is as large as possible. In other words, in a given set of training examples that belong to two classes, a SVM algorithm builds a model based on assigning these examples to n dimensional input vectors in a high dimensional feature space with a maximum separation. Figure (4.2) shows the hyperplane that separates the two different classes with the maximum margin. The examples closest to the hyperplane are marked with square shape called *Support Vectors*. The *Answer Selection* component uses a linear SVM algorithm available in the *Scikit* library package (Pedregosa et al., 2011) to predict sentences classes for a given query.

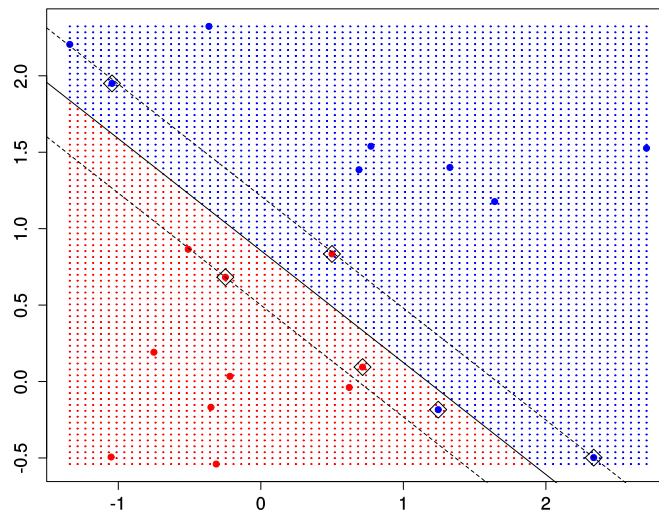


FIGURE 4.2: The maximum margin separating the hyperplane within two dataset classes using a linear SVM

4.3.3 Features

The *Answer Selection* component uses different sets of features extracted from each sentence in the claim zone. In addition to the n -grams and the similarity $sim(s,t)$ features used in the *Claim Zoning* component (see Section 4.2.3), a further four features are employed:

Rhetorical Role (*role-label*): This feature derives from either the *nlmCategory*, if the abstract is structured, or from the *Claim Zoning* classifier if the abstract is unstructured. This feature can hold one of the two values: Results or Conclusions.

Cosine Similarity $sim(s,q)$: This feature captures the relationship between the research question and sentences in an abstract. Sentences with a high degree of lexical similarity are more likely to provide potential answers.

Z-score $Z(s,t)$ and $Z(s,q)$: The *Z-score* is a standard score that shows the number of standard deviations (σ) the value X is above or below the mean (μ) in normally distributed datasets (Wonnacott and Wonnacott, 1990). This measure is useful to compare the similarity scores of sentences across the entire dataset, since the score is relative to the average (mean) score, which therefore makes it comparable with the scores from other abstracts. Equation (4.4) shows the formula to compute the *Z-score* of each sentence in corpus.

$$Z = \frac{X - \mu}{\sigma} \quad (4.4)$$

For example, suppose that the similarity score of sentence s_a in abstract $A = 70$, $\mu = 60$ and $\sigma = 15$; and the similarity score of sentence s_b in abstract $B = 75$, $\mu = 68$ and $\sigma = 12$. Despite $\text{sim}(s_b, q)$ having a higher score than $\text{sim}(s_a, q)$, it is not necessarily more relevant to q as they derive from different abstracts (distribution). Therefore, in order to determine whether $\text{sim}(s_b, q)$ is higher than $\text{sim}(s_a, q)$, the *Z-score* of each is computed as shown in equation (4.5). The results show that, although $\text{sim}(s_b, q)$ is larger than $\text{sim}(s_a, q)$, the value of $Z(s_a, q)$ is larger than $Z(s_b, q)$ which makes sentence s_a more relevant to the query q than sentence s_b shows.

$$Z(s_a, q) = \frac{70 - 60}{15} = 0.67 \quad Z(s_b, q) = \frac{75 - 68}{12} = 0.58 \quad (4.5)$$

4.3.4 Data

The *Answer Selection* component uses 255 abstracts previously collected for *ManConCorpus* (see Section 3.3). The claim sentences in the corpus are considered as the *potential answers*. The corpus is divided into two subsets *ManConCorpus-str* (184 abstracts), which contains structured abstracts and *ManConCorpus-unst* (67 abstracts), which contains unstructured abstracts.

ManConCorpus-str is used to train and test the *Answer Selection* component. It consists of 2,108 sentences, 184 of which are *potential answers* and the rest are not *potential answers*. The dataset is distributed into 382 sentences that belong to the *Introduction* section, 556 sentences belong to the *Methods* section, 842 sentences belong to the *Results* section and 321 sentences belong to the *Conclusions* section.

ManConCorpus-unst is used to evaluate the pipeline system (the *Claim zoning* and *Answer Selection*). It consists of 660 sentences, 67 of which are *potential answers* and the others are considered *non-potential answers*.

4.3.5 Results and Discussion

The *Answer Selection* component predicts the most informative sentence(s) from the claim zone of an abstract in relation to a given query. The system was developed using a SVM classifier and five-fold cross validation.

	Features set (1): uni+bi-grams+loc			Features set (2): uni+bi-grams+section_label		
Fold-#	Precision	Recall	F1-score	Precision	Recall	F1-score
Fold-1	0.71	0.58	0.64	0.73	0.87	0.80
Fold-2	0.65	0.53	0.58	0.63	0.89	0.74
Fold-3	0.59	0.53	0.56	0.64	0.78	0.70
Fold-4	0.86	0.69	0.77	0.86	0.86	0.86
Fold-5	0.61	0.57	0.59	0.62	0.83	0.71
Micro-Av.	0.68	0.58	0.63	0.70	0.85	0.76
	Features set (3): uni+bi-grams+section_label+sim(s,t)			Features set (4): uni+bi-grams+section_label+Z(s,t)		
Fold-#	Precision	Recall	F1-score	Precision	Recall	F1-score
Fold-1	0.68	0.71	0.69	0.69	0.76	0.72
Fold-2	0.65	0.82	0.72	0.70	0.79	0.74
Fold-3	0.70	0.78	0.74	0.76	0.86	0.81
Fold-4	0.85	0.92	0.88	0.85	0.94	0.89
Fold-5	0.63	0.74	0.68	0.63	0.77	0.69
Micro-Av.	0.70	0.79	0.74	0.73	0.82	0.77
	Features set (5): uni+bi-grams+section_label+sim(s,q)			Features set (6): uni+bi-grams+section_label+Z(s,q)		
Fold-#	Precision	Recall	F1-score	Precision	Recall	F1-score
Fold-1	0.74	0.82	0.78	0.72	0.89	0.80
Fold-2	0.67	0.74	0.70	0.71	0.76	0.73
Fold-3	0.69	0.75	0.72	0.72	0.78	0.75
Fold-4	0.89	0.86	0.87	0.85	0.94	0.89
Fold-5	0.67	0.74	0.70	0.63	0.74	0.68
Micro-Av.	0.73	0.78	0.75	0.73	0.82	0.77
	Features set (7): uni+bi-grams+label+sim(s,t)+sim(s,q)			Features set (8): uni+bi-grams+label+Z(s,t)+Z(s,q)		
Fold-#	Precision	Recall	F1-score	Precision	Recall	F1-score
Fold-1	0.68	0.71	0.69	0.68	0.74	0.71
Fold-2	0.65	0.82	0.72	0.74	0.82	0.78
Fold-3	0.70	0.78	0.74	0.79	0.86	0.83
Fold-4	0.85	0.92	0.88	0.85	0.92	0.88
Fold-5	0.63	0.74	0.68	0.65	0.74	0.69
Micro-Av.	0.70	0.79	0.74	0.74	0.82	0.78

TABLE 4.2: The *Answer Selection* component performance using different sets of features. The average F-score of the component, under the best setting, to annotate both *potential-answer* and *non-potential answer* achieved 97%. (The baseline score is 91%)

Table (4.2) shows the variation of the system performance in micro-averaging using different sets of features (1)-(8). The baseline is 91%, which is the average accuracy of annotating all sentences with class *potential-answer*. The table shows only the performance for the detection of *potential answer* sentences. The detection of *non-potential*

answer sentences is not described here since they were remarkably high (between 97 and 99%), as a result of the imbalance between the corpus classes (9% vs. 91%) which biases the system during learning; also, we are most interested in identifying *potential answers*. The table consists of eight groups of scores, and each group belongs to one set of features.

Features sets (1) and (2) show the system performance using uni+bigrams and relative location (*loc*) features versus uni+bigrams and rhetorical roles (*role-label*) features. The use of *role-label* features outperformed the relative location features. The performance using feature set (2) achieved a 76% F1-score, compared to a 63% F1-score using feature set (1).

Features sets (3) and (4) show the system performance using $sim(s,t)$ features in addition to the features used in set (2) versus the use of $Z(s,t)$ features. Features set (3) lowered the system performance to a 74% F1-score from 76%; substituting the $sim(s,t)$ features with the $Z(s,t)$ increased system performance to 77%, which in turn outperformed the best performance thus far.

Features sets (5) and (6) compare the use of $sim(s,q)$ versus $Z(s,q)$ in addition to feature set (2). The results matched the results of feature sets (3) and (4); the use of $sim(s,q)$ produced a lower score compared to the use of feature set (2) alone; however, the replacement of that feature with the $Z(s,q)$ features enhanced the performance to 77%, a similar result to the system performance when using feature set (4).

Features sets (7) and (8) compare system performance using the combination of $sim(s,t)$ and $sim(s,q)$ versus the combination of $Z(s,t)$ and $Z(s,q)$. The results were consistent with the scores of previous groups. *Z-scores* always produce better results than those achieved by the use of similarity scores. The best performance for the *Answer Selection* component was achieved using *n-grams*, *role-label*, $Z(s,t)$ and $Z(s,q)$ features, which achieved a 78% F1-score.

Two main observations were noted regarding the system results. Firstly, the combined use of *n-grams* and *role-label* features achieved a better average recall score than incorporating any of the cosine similarity scores features ($sim(s,t)$ or $sim(s,q)$ or a combination thereof). The reason for this is that the majority of the *potential answers* in the corpus belong to the *Conclusions* section. Thus, the classifier may associate the decision of predicting whether a sentence is a *potential answer* or not, based on the value of the *role-label* feature which includes 85% of the *potential answers*. However, incorporating the cosine similarity scores in feature sets (3), (5) and (7) resulted in the classifier excluding *potential answers* in the *Conclusions* section with relatively low similarity scores, which caused a drop in the recall score of around 6%.

Secondly, it was observed that the use of *Z-score* features improves the recall score of the system more than the use of cosine similarity scores; however, they were still lower than the score achieved using feature set (2) (82% vs. 85%). Moreover, the use of *Z-score* features improved the system precision more than the use of feature set (2) (70% vs. 73-74%). This may suggest that the *Z-scores* are reliable features and can be substituted for the cosine similarity scores.

4.4 Pipeline System

The *Claim Zoning* and *Answer Selection* components have been evaluated individually. The best feature set used in the first component was set (5), as shown in Table (4.1), while the best feature set used in the second component was set (8), as shown in Table (4.2). The next stage is to combine both components in a pipeline to predict the answers of given queries from unstructured abstracts (*ManConCorpus-unst*). Note that the *Answer Selection* component requires the rhetorical roles of sentences as features; however, these were not available in *ManConCorpus-unst*, since it contains unstructured abstracts. Therefore, the abstracts had to be processed by the *Claim Zoning* component in order to assign the appropriate rhetorical roles, before applying the *Answer Selection* component to identify sentences which answer the queries.

Class	Precision	Recall	F-score
potential answer	0.62	0.60	0.61
non-potential answer	0.96	0.96	0.96
Micro-Average	0.93	0.92	0.92

TABLE 4.3: the pipeline system performance using *ManConCorpus-unst*. The baseline is 87%, which is the accuracy percentage of annotating all sentences with class *non-potential answer*.

Figure (4.3) shows the pipeline performance, an F-score of 0.61 was achieved for detecting *potential answer* sentences and 96% for *non-potential answer* sentences. The high score for the detection of *non-potential answer* sentences was attributed to bias in the distribution of the classes in both the *ManConCorpus-str* and *ManConCorpus-unst* corpora described earlier.

Three main observations were noted during the evaluation process. Firstly, the *Claim Zoning* component in the pipeline labelled 58 out of 67 of the *potential answers* in the *ManConCorpus-unst* corpus as *Conclusions*, while the remaining nine sentences were annotated as *Results*. The *Answer Selection* component in the pipeline had been

trained in *potential answer* instances where 91% of them belong to the *Conclusions* section. Consequently, the *Answer Selection* component would always be biased towards predicting those belonging to the *Conclusions* section as *potential answers* and ignoring the *potential answers* which belong to the *Results* section. Because of the *Claim Zoning* annotation and the bias of the *Answer Selection* decision, the pipeline system would always predict any *Results* sentence as a *non-potential answer*, therefore, this caused a reduction in pipeline performance by 13% recall.

For example, Table (4.4) shows a question (1) and its potential answer (2), which were extracted from abstract (18324526). Sentence (2) in the corpus was the actual *potential answer* to the question and was missed by the *Answer Selection* component as it was mislabelled by the *Claim Zoning* component as *Results*.

	role-label	Sentence
1	X	In patients with coronary artery disease (CAD), is C242T polymorphism of P22(PHOX) gene associated in development of CAD?
2	Results	The C242T variant was associated with CHD risk in women.

TABLE 4.4: An example of a potential answer that was missed by the pipeline system due to its rhetorical label

The second observation was that the *Answer selection* component was able to correctly detect the *potential answers* in many of the abstracts containing only one *Conclusions* sentence. However, in abstracts containing more than one *Conclusions* sentence (see Table(4.5)), the component tends to occasionally annotate more than one sentence as a *potential answer*. The analysis showed that many of these false positive sentences were in fact carrying information similar to the information in the actual potential answers in their abstracts.

# of Abstracts	# of sent. in Conclusion sections
36	1
20	2
7	3
3	4
1	5

TABLE 4.5: The distribution of the *Conclusions* sentences among the *ManConCorpus-uns* abstracts after the *Claim Zoning* annotation. The first and the third columns show the number of abstracts that contain a specific number of *Conclusions* sentences and the second and the fourth columns show the number of *Conclusions* sentences in these abstracts

Table (4.6) shows two examples of false positive sentences. The first example was identified in abstract (3257376), where the *Answer Selection* component annotated sentences (4) and (5) as *potential answers* to question (3), though only sentence (4) was the actual *potential answer* to the question. The second example was found in abstract (8733865), where the *Answer Selection* component annotated sentence (7) and (8) as *potential answers* to question (6), though only sentence (7) was the actual *potential answer*. Although, these errors caused a reduction in the overall pipeline performance, they are still useful in achieving the main goal. This is because many of these errors still carry similar information to the actual *potential answers*.

	PMID	role-label	Sentence
3	X	X	In patients undergoing coronary bypass surgery, does Aspirin usage, compared to no aspirin, cause bleeding?
4	3257376	Conclusions	We conclude that aspirin ingestion increases postoperative blood loss and transfusion requirements and we recommend discontinuation of aspirin therapy before cardiac procedures.
5	3257376	Conclusions	This finding suggests that a subset of patients are particularly sensitive to aspirin and have significantly prolonged bleeding times after aspirin ingestion.
6	X	X	in patients with coronary artery diseases, does combining CABD and CEA , compared with CABG or CEA alone, reduce morbidity?
7	8733865	Conclusions	These data demonstrate that the performance of simultaneous CABG and CEA procedures is associated with increased neurologic morbidity (14.3%) both ipsilateral and contralateral to the side of carotid surgery in contrast to staged CABG and CEA (3.4%).
8	8733865	Conclusions	In addition when staged carotid surgery preceded coronary revascularization in those with severe coronary artery disease the combined cardiac complication and mortality rate was significantly higher than when coronary revascularization preceded CEA.

TABLE 4.6: two examples from two abstracts where the pipeline annotated multiple sentences from the same abstract as potential answers

The final observation was that in a small number of abstracts the *Answer Selection* component annotated actual *potential answer* sentences as *non-potential answers* and annotated *non-potential answers* as *potential answers*. For example, Table (4.7) shows two sentences (10) and (11). The first sentence was annotated as *potential answer* although it was not as per the corpus annotation, and the second sentence was falsely annotated as

a *non-potential answer*. Both sentences were found to hold the same answer to question (9). Such annotations caused a further reduction in the pipeline performance. A manual investigation revealed that the errors were due to the *Z-scores* of sentence (10) being higher than those of sentence (11) which caused the classifier to favour it. After manual inspection, it was noted that similar discrepancies existed in many sentences and could act as alternative potential answers since these sentences contained similar information to the actual potential answer. Therefore, such sentences are still considered useful for the main aim of this research.

	Section-label	Sentence
9	X	In patients with dilated cardiomyopathy, Are HLA genes associated with development of Dilated Cardiomyopathy?
10	Conclusions	The reported association of HLA-DR4 with idiopathic dilated cardiomyopathy in the Caucasian population does not apply to the Omanis.
11	Conclusions	The lack of any HLA antigen association in Omanis would argue against the proposed HLA-linked autoimmune pathology of idiopathic dilated cardiomyopathy.

TABLE 4.7: An example of an error generating by the pipeline system due to *Z-scores*

4.5 Conclusions

The development of a pipeline system to identify research claims relevant to a given query was described in this chapter. The system consists of two components: the first is responsible for identifying the abstract section that may contain the research claim sentences, and the second is responsible for selecting the most relevant sentence(s) from that section containing the *potential answer* to the query. Both systems were evaluated in a pipeline setting, and the performance was found to be reliable given the fact that many of the errors produced by the pipeline were on sentences that contain the same information of the actual potential answers.

Chapter 5

Identification of Contradictory Claims in Biomedical Abstracts

5.1 Introduction

The previous chapter described the first stage of a system to detect contradiction between claims. This chapter discusses the second stage which identifies contradictory claims from a given set of claims sentences.

This chapter proposes an approach to identify contradiction between claims by following a two-stage process: extraction of facts from claims relevant to a given question (termed *fact extraction*), and detection of assertion values of these facts with respect to the given question, assuming that the question is formulated as a closed question that can be answered with either *yes* or *no*. This latter stage is called *fact assertion value detection*. The approach followed to identify contradictory claims is based on the definition of contradiction (see [Section \(3.2.1\)](#)), in which two claims are considered contradictory when the assertion value of the fact extracted from a claim differs from the assertion value of the fact extracted from the other claim, with respect to the same question.

The main contribution of this chapter is the development of a supervised machine learning system to detect contradictory claims for a given question. The approach uses features inspired by claim typologies (see [Section 3.2.2](#)). The system is evaluated using *ManConCorpus* and *AutConCorpus* (see [Chapter \(3\)](#)). Furthermore, the system evaluated using *AutConCorpus* is also used to predict the assertion values of *ManConCorpus* and the results are reported and analysed.

5.2 Related Work

For the purposes of the fact extraction stage, a *fact* is expressed using a composition of a predicate and its arguments. For example, in the statement “*aspirin increased blood pressure*”, the arguments are *aspirin* and *blood pressure* and the predicate is *increased*. The

term *fact* is used in this research as a synonym to the term *event*, which is widely used in textual inference research (Harabagiu, Hickl, and Lacatusu, 2006a; Marneffe, Rafferty, and Manning, 2008). However, that term tends to be used in the biomedical domain to mainly describe interactions between molecules, a meaning which differs in the non-biomedical domain; we avoided that problem by using the term *fact*.

Scientific text tends to employ complex text structure and may describe multiple facts within a single sentence. For example, the claim “*Although a bedtime dose of doxazosin can significantly lower the blood pressure, it can also increase left ventricular diameter, thus increasing the risk of congestive heart failure*”¹ includes three facts, the first of which is the effect of doxazosin on blood pressure, the second is the effect of doxazosin on left ventricular diameter and the third is the effect of doxazosin on the risk of congestive heart failure. It is therefore critical to determine which fact is relevant to the potential contradiction.

Various research studies on similar topics, such as textual entailment, have considered the same stage for identification of the fragment of text in (T, H) that describe the same fact. For example, Harabagiu, Hickl, and Lacatusu (2006a) extracted the predicate-argument structures from sentences to find the possible alignment between the text T and the hypothesis H , in order to determine whether the pairs described the same event.

Ritter et al. (2008) used *TextRunner* (Yates et al., 2007), a relation extraction system, for the same purpose in order to extract information from text in the form of tuples. That work showed that the use of such tool simplified the task of ensuring that both T and H represented the same event, since multiple syntactic problems (such as anaphora) and other semantic challenges (such as counter-factuals) were delegated to the relation extraction system.

Marneffe, Rafferty, and Manning (2008) used an event coreference component to filter for pairs of texts that do not describe the same event. However, that component was more complex than the approach used in previous research. Three methods were adopted in that component in order to select the non-coreferent events. The first method checks that the entities in the pairs represent the same thing, using the *background knowledge*, the second method verifies that the root verbs of the pairs are the same and the third method calculates the topicality of the noun phrases in the pairs to ensure that they both describe the same topic and, consequently, the same event.

Kawahara, Inui, and Kurohashi (2010) used a similar approach to align the predicates and the arguments within Japanese statements in order to identify contradictory

¹from Pubmed abstract pmid#18551024

and contrastive relations. Andrade et al. (2013) followed the same approach using a tool called *Syncha* (Iida and Poesio, 2011), a predicate-argument structure analyser for Japanese text.

5.3 Methods

This research involves the construction of a supervised machine learning system using the two contradictory corpora described in Chapter (3): *ManConCorpus* and *AutConCorpus*, in order to detect potentially contradictory claims. The two corpora are used here to (1) compute and compare their performances and (2) evaluate whether the classifier developed from the automatically generated corpus is reliable for detecting the contradiction of claims in *ManConCorpus*.

This research use three methods to detect the assertion values of facts: the use of negation, measurement of the semantic orientation of the adjectives, and measurement of the semantic orientation of the predicates. The use of negation and the semantic orientation of adjectives have been previously reported (see Section (2.2.1)).

The semantic orientation of the predicates is another indicator for contradiction. Hashimoto et al. (2012) used the term *excitation* to describe that property; and described three semantic values that can be assigned to predicates: excitatory, inhibitory and neutral (see Section 3.2.2). For example, the predicate *improve* is an excitatory, *destroy* is an inhibitory and *evaluate* is a neutral. If two claims therefore describe the same fact, but one employs an excitatory predicate while the other includes an inhibitory one, they are considered likely to be contradictory.

One important study that is relevant to our research was conducted by Niu et al. (2005), who developed a system to predict the polarity of the clinical outcomes at the sentence level. They collected four types of words: words indicating *more*, words indicating *less*, words indicating *good* and words indicating *bad*. If one of the words in the groups appears in the text, its value is attached to that word and to the following words, until the next punctuation mark is reached. Following this process, multiple patterns were identified in the text and were used as features to determine text polarity.

The fact assertion value detection stage uses a similar approach; however, rather than collecting the four groups randomly, this research adopts a more systematic methodology to collect and use the group of words. For example, only the top most frequently used adjectives found in the *Conclusions* sections of PubMed abstracts are used to collect words indicating *good* and *bad*, and the same approach is used to collect verbs indicating *more* or *less*. Furthermore, instead of attaching a word value to adjacent words until the next punctuation mark, this research uses the predicate argument boundaries.

Unlike that approach, however, which measured the semantic orientation of a clinical result text in terms of patient case improvement, this research measures the semantic orientation of a research claim with respect to a given question. In addition to these differences, this research incorporates a fact extraction stage which was not implemented in the work of Niu et al. (2005).

5.4 Lexicons

Three lexicons are constructed to detect terms that may indicate contradictions: *Negation* to detect negation terms, *Directionality* to detect the semantic orientation of predicates and *Sentiment* to detect the semantic orientation of adjectives.

Negation: In biomedical information extraction tasks, detection of the negation status of a research finding is an essential step for demonstrating the absence of a certain medical condition. This research searches only for negation markers rather than identifying the negated findings (see Section (2.5)).

A negation lexicon is constructed using 16 terms, mainly collected from the lexicon of Chapman et al. (2011), in addition to the terms *lack* and *against*. Table (5.1) shows a sample of the negation lexicon; the full list can be found in Appendix (C.1). The lexicon should serve to identify the negation terms in claims sentences.

Term	Value	Term	Value
absence	<i>not</i>	lack	<i>not</i>
cannot	<i>not</i>	negative	<i>not</i>
deny	<i>not</i>	never	<i>not</i>
fail	<i>not</i>	no	<i>not</i>

TABLE 5.1: A sample of negation terms used in the negation lexicon

Directionality: The directionality lexicon was created based on the excitatory, inhibitory and neutral influence concepts described in Section (3.2.2). The term *directionality* was also used by Blake (2010) to express changes occurring to biomedical concepts; for example, the use of the term *increased* shows that the effect on a concept is *upward*, while *decreased* is *downward*.

The directionality lexicon was constructed by analysing the *Conclusions* sections of 1,338,368 Medline abstracts. An annotator with an advanced degree in medicine was asked to choose verbs that represented the highest frequencies with either an excitatory or inhibitory influence. Each verb was annotated with its directionality value, which was either *more* or *less*. Verbs that show a neutral influence were not included in the

lexicon since they have no influence on their arguments. The lexicon consists of 345 terms, 202 terms represent excitatory effect and 143 terms represent inhibitory effect. Table (5.2) shows a sample of the terms used in the lexicon, and the full list can be found in Appendix (C.2).

Term	Value	Term	Value
alleviate	<i>less</i>	abuse	<i>more</i>
ameliorate	<i>less</i>	accelerate	<i>more</i>
attenuate	<i>less</i>	accumulate	<i>more</i>
block	<i>less</i>	activate	<i>more</i>
cut	<i>less</i>	add	<i>more</i>
decrease	<i>less</i>	augment	<i>more</i>

TABLE 5.2: A sample of terms used in the directionality lexicon

Sentiment: The sentiment lexicon was constructed based on previous research using the semantic orientation of adjectives (lexicon-based systems) to detect the semantic orientation of text (e.g., Hatzivassiloglou and McKeown (1997), Taboada et al. (2011), and Turney (2002)).

The same group of abstracts used for the directionality lexicon was employed for the extraction of the semantic orientation of adjectives. The annotator was asked to choose adjectives with the highest frequencies that either describe positive (*good*) such as *favourable* or negative (*bad*) such as *abnormal*. The lexicon consists of 56 adjectives, 27 of which had the value *good* while 29 had the value *bad*. Table (5.3) shows a sample of the sentiment lexicon, and the full list can be found in Appendix (C.3).

Term	Value	Term	Value
abnormal	<i>bad</i>	acceptable	<i>good</i>
abusive	<i>bad</i>	advantageous	<i>good</i>
adverse	<i>bad</i>	appropriate	<i>good</i>
aggressive	<i>bad</i>	beneficial	<i>good</i>
bad	<i>bad</i>	best	<i>good</i>
harmful	<i>bad</i>	better	<i>good</i>

TABLE 5.3: A sample of terms used in the sentiment lexicon

5.5 Fact Extraction

Fact extraction is the stage where the most relevant piece of information in a claim sentence to a question is extracted in order to identify its assertion value at a latter stage.

This information is extracted in the form of a predicate and its associated arguments i.e. a relation tuple. Questions in this research are written in PICO format, which has a specific writing style. Such questions contain only one relation tuple, which is usually between the Intervention (I) and the Outcome (O). On the other hand, a claim may contain multiple relation tuples.

The fact extraction stage identifies which relation tuple in a claim is more associated with the relation tuple extracted from its question. The most associated claim tuple is considered the fact in that claim that may contain an answer to the question. The following sub-sections describe the three main stages of the fact extraction process.

5.5.1 Relation Extraction

This stage takes a list of claims from different research abstracts that supposedly answer a research question, and generates a list of relation tuples, where each claim has one or more tuples. The relation tuples are extracted using available relation extraction systems. Two types of relation extraction systems were evaluated at this stage: *SemRep*, a relation extraction system designed for biomedical documents (see [Section 3.4.1](#)) and three open information extraction systems, *ReVerb* (Wu and Weld, [2010](#)), *WOE* (Fader, Soderland, and Etzioni, [2011](#)) and *Ollie* (Mausam et al., [2012](#)), designed for general domain text. The evaluation was carried out to evaluate the capability of such tools for extraction of relation tuples from *ManConCorpus*.

Evaluation of *SemRep* (Rindfleisch and Fisman, [2003](#)) showed that the system was only capable of extracting a few relation tuples from the corpus. This outcome was clear, given that *SemRep* extracts relation tuples only for specific patterns available in the *Semantic Network*, and that these patterns only consider biomedical concepts that belong to certain semantic types in which the nature of research claims may not necessarily fit these patterns.

For general extraction systems, it was found that two systems (*ReVerb* and *WOE*) extract relation tuples by identifying only the relation phrases that satisfy certain lexical and syntactical constraints and the appropriate NP argument pairs for each relation phrase. A fundamental limitation of these tools is that they only extract relation phrases that are mediated by verbs, and ignore relations that are mediated by other syntactic categories such as nouns and adjectives (Mausam et al., [2012](#)). [Table \(5.4\)](#) shows a claim sentence where the two relation systems could not find relations to extract. The *Ollie* system, on the other hand, managed to extract relation tuples from these examples as shown in the table by expanding the syntactic scope to include other expressions that belong to the noun and adjective categories. The evaluation results showed that *Reverb*

appeared more reliable, in terms of generating accurate relation tuples. However, none of the three versions were capable of extracting relation tuples from all research claims in the corpus.

Text	L-Argument	Predicate	R-Argument
We conclude that in women with preeclampsia, prolonged dietary supplementation with l-arginine significantly decreased blood pressure through increased endothelial synthesis and/or bioavailability of NO.	l-arginine	decreased	blood pressure..

TABLE 5.4: An example of a claim sentence where ReVerb and WOE could not find relations to extract, while Ollie managed to extract at least one tuple.

Therefore, it was decided to use *ReVerb* at the outset to extract relation tuples from claims; if it failed to extract any relation from a claim, *WOE* was applied to that particular sentence; if *WOE* failed, *Ollie* is used. The purpose of following such a sequence is to extract at least one relation tuple from each claim sentence. The results of this strategy showed that *ReVerb* was able to extract at least one relation from about 95% of the corpus; the other systems were used to extract relations from the remaining 5% of the corpus.

Table (5.5) shows two claims (1) and (2) that provide answers to question (1); Table (5.6) shows the relation tuples extracted from these sentences. The relation tuple (1a) was extracted from the question, tuples (2a) and (2b) were extracted from claim (2), and tuple (3a) was extracted from claim (3).

	PMID	Text
1	Question	In women with pre-eclampsia, does treatment with L Arginine, compared to placebo, reduce blood pressure or pre-eclampsia?
2	15638817	We conclude that in women with preeclampsia, prolonged dietary supplementation with l-arginine significantly decreased blood pressure through increased endothelial synthesis and/or bioavailability of NO
3	14678093	Oral L-arginine supplementation did not reduce mean diastolic blood pressure after 2 days of treatment compared with placebo in pre-eclamptic patients with gestational length varying from 28 to 36 ...

TABLE 5.5: Contradictory claims (2) and (3) with respect to the question (1).

	L-Argument	Relation	R-Argument
1a	L Arginine	reduce	blood pressure
2a	We	conclude	that in women with preeclampsia, prolonged dietary supplementation with
2b	l-arginine	decreased	blood pressure; through increased endothelial synthesis
3a	Oral L-arginine supplementation	did not reduce	mean diastolic blood pressure

TABLE 5.6: The relation tuples extracted from the question and claims Table (5.5).

5.5.2 Biomedical Concept Identification

When a claim sentence contains multiple relation tuples, such as tuple (2a) and (2b) listed above, one is usually more relevant to the research question than the other. In such a situation, it is important to find a mechanism to automate the process of identifying which tuple is more relevant to the question. An option to resolve that problem is by using *UMLS-Similarity* (McInnes, Pedersen, and Pakhomov, 2009), a system that implements various measures to compute the semantic similarity or relatedness between biomedical concepts using the UMLS resources. Because the biomedical concepts in claims sentences are not known, *MetaMap* (Aronson, 2001) is used to map biomedical

phrases in the sentences to their equivalent concepts in the UMLS *Metathesaurus* using a Concept Unique Identifier (CUI).

Seven semantic types of biomedical concepts (out of the total 15 types available in UMLS *Metathesaurus*) were used with our corpora: *Anatomy, Chemicals & Drugs, Devices, Disorders, Genes & Molecular Sequences, Living Beings and Physiology*. These groups were included because they describe diseases, genes, proteins and medical equipment, categories to which a significant proportion of biomedical entities are ascribed. The other semantic groups were not considered, in order to avoid annotating words such as *increased* which are significant for the classifier to understand the directionality terms in a sentence. For example, *MetaMap* assigns the CUI (C0004057) to *ASA, aspirin* and *acetylsalicylic acid*, since they represent the same concept despite being lexically different. Table (5.7) shows the relation tuples described in Table (5.6) following annotation of the concepts with their CUIs.

	L-Argument	Relation	R-Argument
1a	C0003765	reduce	C1271104
2a	We	conclude	that in C0043210 with C0032914, prolonged C0242297 with ;
2b	C0003765	decreased	C1271104; through increased C0014257 synthesis;
3a	C0003765 C0242297	did not reduce	mean C1305849

TABLE 5.7: Annotation of claims tuples using *MetaMap*

5.5.3 Relation Relatedness

UMLS-Similarity differentiates between semantic similarity and semantic relatedness between concepts. Semantic similarity is identified by measuring the distance between two concepts in a UMLS source such as MeSH using a predefined relation such as parent/child (is-a). For example, the similarity score between *nose* and *head* is 0.33, since *nose* comes under the body region *head* (and both are in the same path). The use of predefined relations to measure similarity may sometimes hinder the discovery of very similar concepts such as *statins* and *atorvastatin*, where *statins* are a group of medicines that helps to lower the level of low-density lipoprotein cholesterol in the blood and *atorvastatin* is one of these medicines. The score for measuring the similarity of these two concepts was -1 as they were not on the same path in the MeSH hierarchy. Therefore, it is important to identify an alternative method to identify relevant concepts.

Semantic relatedness between concepts is another approach that is evaluated using information outside of the example of the parent/child relationship. *UMLS-Similarity*

uses two measures to compute the relatedness between concepts: Lesk and Vector. This research uses the Vector measure, which computes relatedness using second-order co-occurrence vectors from the UMLS extended definitions of concepts (Pedersen and Patwardhan, 2004). The score that resulted from computing the relatedness between *statin* and *atorvastatin* using the vector measure was 0.9089, a significant relatedness score.

The relatedness score between a claim tuple and a question tuple is computed using Equation (5.1) as follows: first, the CUIs of the claims and the question arguments are gathered into two sets $cuis_c$ and $cuis_q$. Then the relatedness scores between every possible pair of $cuis_c$ and $cuis_q$ are added together to compute the overall relatedness between the claim and the question. The claim tuple that achieves the highest relatedness score with the question tuple is considered the one containing the fact, with respect to the question.

$$sim(cuis_c, cuis_q) = \forall cui \in cuis_c, \forall cui \in cuis_q \sum_{c \in cuis_c} \arg \max rel(cui_c, cui_q) \quad (5.1)$$

For example, the relatedness score between the question tuple (1a), and the claims tuples (2a), (2b) and (3a) were 1.66, 4.13 and 3.4, respectively. Since the tuple (2a) and (2b) were derived from the same claim text, and because (2b) achieved a higher score than (2a), tuple (2b) is selected as the fact of that claim, and progresses to the stage of assertion value detection. However, because claim (3) has only one tuple (3a), the relatedness score is not important and therefore is selected to represent its claim, regardless of its relatedness score.

5.6 Fact Assertion Value Detection

The previous stage, fact extraction, is mainly performed in order to extract only the segments of text from claims relevant to the question. The output of that stage is the question tuple and a list of claims tuples, where each tuple represents one claim. The next stage is to detect the assertion values of the claims tuples (facts) with respect to the question tuple, in order to detect claims that contain incompatible facts, which is considered potentially contradictory.

Unlike contradiction in linguistics, which directly compares the meaning of one text with another, this research indirectly compares the research claims to a specific question. For example, $Claim_a$ and $Claim_b$ contain information to answer the question Q , which has two possible answers, *yes* or *no*. If the assertion value extracted from the fact in $Claim_a$ with respect to *question* is *yes*, and the assertion value of the other fact

extracted from $Claim_b$ that answers the same Q is *no*, then the claims are considered potentially contradictory.

Fact assertion value detection component is considered in this research as a two-way classification problem, where the claims that agree with the question are labelled *yes* and those that do not agree labelled *no*. An SVM classifier that uses the linear SVM algorithm in the Scikit library (Pedregosa et al., 2011) is applied on the two contradiction corpora *ManConCorpus* and *AutConCorpus* using four-fold cross validation on four features: n-grams, negation, directionality and sentiment.

N-grams (uni+bi-grams): These features were previously employed in the *Claim Zoning* component (see Section (4.2)) and the Answer Selection component (see Section (4.3)) and found useful. These features consist of the uni-grams and bigrams in the claims' text with a minimum frequency of 30 across the whole corpus.

Negations: Terms from the questions and claims tuples that have a match in the negation lexicon are replaced with the label *not*, and the label *not_* is subsequently attached to every word after *not* in that particular field. For example, the relation field of tuple (3a) contains the word *not*, which is in the negation lexicon. Therefore, the terms are prefixed with *not_*, as shown in (5.8).

	L-Argument	Relation	R-Argument
3a	Oral L-arginine supplementa..	did <i>not not_reduce</i>	mean diastolic blood pressure

TABLE 5.8: Annotating negation terms in a claim tuple

Directionality: Verbs in the relation fields of questions and claims tuples that have a match in the directionality lexicon are replaced with their corresponding values. For example, Table (5.9) shows the relation field of tuple (2b) contains the verb *decreased*, which has a match in the lexicon. Therefore, that verb is replaced with its value, *less*. If no directionality term is identified in the relation field, but a directionality term is found in the right argument field (*R-Argument*), the value of that term is added to the relation field. For example, if the relation field of tuple (2b) has no directionality term, the value of the term *increased* in the second argument field would have been added to the relation field.

	L-Argument	Relation	R-Argument
2b	l-arginine	<i>less</i>	blood pressure; through increased endothelial synthesis;

TABLE 5.9: Annotation of directionality terms in a claim tuple

Sentiment: Adjectives in the argument fields (*L-Argument* and *R-Argument*) of question and claim tuples are replaced with their corresponding values in the sentiment lexicon. For example, Table (5.10) shows a claim extracted from abstract (9412879), while Table (5.11) shows the relation tuple extracted from that claim before and after annotation. The term *loss* in the second argument field of the tuple is replaced with its corresponding value in the sentiment lexicon. Similar to the directionality annotation, if no adjective term was identified in that field but an adjective that has a match in the lexicon is found in the relation field, the value of that term is added to the second argument field.

	Text
4	In patients undergoing a first CABG and with no known factors affecting their coagulation, ASA therapy did not appear to increase blood loss, reopening for bleeding, or blood products usage requirements during the hospital stay.

TABLE 5.10: A claim sentence extracted from abstract (9412879) to be annotated by negation, directionality and sentiment

	L-Argument	Relation	R-Argument
4a	ASA therapy	did not appear to increase	blood loss.
4a (annotated)	ASA therapy	did <i>not</i> <i>not</i> _appear <i>not</i> _to <i>not</i> _more	blood <i>bad</i>

TABLE 5.11: A fully annotated claim tuple

5.7 Results and Discussion

5.7.1 Fact Extraction Stage

A single relation tuple was generated for approximately 41% of the claims in *ManConCorpus*, and between two and five relations were generated for the remainder. However, in *AutConCorpus*, a single relation tuple was extracted from approximately 56% of claims, two relation tuples were extracted from 32% of the sentences, and the remainder included between three and six relation tuples. The relation extraction tools performed efficiently, and the items inaccurately extracted were due to the complexity of the structure of the text. For example, Table (5.12) shows a claim sentence extracted from abstract (18711405), and Table (5.13) shows the relation tuples extracted from that claim. The first two tuples, (5a) and (5b), were correctly extracted; however, tuples (5c) and (5d) overlooked the negation term *not*. If the *UMLS-similarity* tool selected one of these relations as the most relevant to the question, the classifier would predict the opposite value, since *not* is missing. Such errors could have been mitigated by using

NegExpander (see Section (2.5)), however, it was decided not to use it here to avoid the possibility of producing false positive examples.

	Text
5	Injection of autologous BMCs directly into the scar or into the artery supplying the scar is safe but does not improve contractility of nonviable scarred myocardium, reduce scar size, or improve left ventricular function more than CABG alone.

TABLE 5.12: A complex claim, where relation extraction systems failed to correctly extract relations

	L-Argument	Relation	R-Argument
5a	the scar	is	safe
5b	the scar	does not improve	contractility of nonviable scarred myocardium
5c	the scar	reduce	scar size
5d	scar size	improve left	ventricular function

TABLE 5.13: The relation tuples extracted from the claim in Table (5.12)

Fact extraction is crucial for detecting common facts between claims in order to identify contradictions. However, current biomedical extraction systems such as *SemRep* appear to be extremely specific and can only extract relations under certain conditions, lowering their potential for extracting relations from claims texts. As such, generic relation extraction systems functioned better than *SemRep* in the extraction of common facts between claims.

5.7.2 Fact Assertion Value Detection

The annotation stage showed that many sentences contained negation terms, directionality terms or a combination thereof; however, only a few sentences contained sentiment terms. This could suggest that the influence of sentiment terms on detection of the assertion value of claims may be higher than that of other features since its representation in the vector space as *tf:idf* becomes important.

Table (5.14) shows the system performance using *ManConCorpus* and *AutConCorpus*. Each row in the table shows the system performance when using the two corpora. The column *Features* shows the features used for each evaluation. The *negation* features means that only the negation annotations were used as features to train and test the classifier, *dire+sent* means that both the directionality and the sentiment annotations were used as features to evaluate the classifier, etc.

		ManConCorpus			AutConCorpus			
	Features	Class	P	R	F1	P	R	F1
1	negation	No	0.83	0.55	0.66	0.59	0.21	0.25
		Yes	0.83	0.96	0.89	0.72	0.88	0.79
		Micro-Average	0.83	0.83	0.82	0.68	0.67	0.62
2	dire+sent	No	0.89	0.31	0.43	0.26	0.18	0.21
		Yes	0.76	0.97	0.85	0.72	0.93	0.81
		Micro-Average	0.80	0.76	0.72	0.58	0.70	0.63
3	negation+dire+sent	No	0.82	0.60	0.69	0.63	0.23	0.31
		Yes	0.85	0.95	0.90	0.74	0.94	0.82
		Micro-Average	0.84	0.84	0.83	0.70	0.72	0.66
4	negation+uni+bi-grams	No	0.83	0.55	0.66	0.64	0.52	0.56
		Yes	0.83	0.96	0.89	0.81	0.87	0.83
		Micro-Average	0.83	0.83	0.82	0.75	0.76	0.75
5	negation+uni+bi-grams+sent	No	0.82	0.60	0.69	0.65	0.54	0.58
		Yes	0.85	0.95	0.90	0.81	0.87	0.84
		Micro-Average	0.84	0.84	0.83	0.76	0.77	0.76
6	negation+uni+bi-grams+dire	No	0.80	0.57	0.66	0.66	0.58	0.62
		Yes	0.84	0.95	0.89	0.83	0.87	0.85
		Avg	0.82	0.83	0.82	0.78	0.78	0.77
7	All Features	No	0.80	0.60	0.69	0.67	0.58	0.62
		Yes	0.85	0.94	0.89	0.82	0.87	0.85
		Micro-Average	0.83	0.83	0.83	0.78	0.78	0.78

TABLE 5.14: The contradiction detection performance using *ManConCorpus* (left) and *AutConCorpus* (right). The baseline of *ManConCorpus* is 68% and that of *AutConCorpus* is 69.6%. The baseline score represents the accuracy percentage when annotating all sentences with class *yes*.

The main goal in presenting the results of the contradiction detection system using *ManConCorpus* next to the results achieved from *AutConCorpus* was to determine whether the change pattern in the performance of both systems would be similar to the feature set changes or not. In other words, this table is not presented to compare the contradiction detection system performance using the two corpora, rather, the table examines only the change of the performance patterns in the two systems using a different set of features. The assumption here is that both systems should experience a similar performance change i.e. if the performance of the system using *ManConCorpus* increased after changing from feature set from (3) to (4), the performance of the system using *AutConCorpus* should increase as well.

The table indicates that the performance of using *ManConCorpus* was superior to using *AutoConCorpus*. This observation is unsurprising since the dataset used to evaluate the system was manually collected, while the corpus used for the other system was automatically collected and consequently experienced some errors generated by *SemRep*.

The baseline performance of the system using *ManConCorpus* was 68% which represents annotating the entire dataset with the assertion value *yes*. Similarly, the baseline performance of the system using *AutConCorpus* was 69.6% on annotating the entire dataset with the assertion value *yes*. These baselines were used as benchmarks to compare system performance using different sets of features as shown in the table.

The table shows that the system performance using the *ManConCorpus* corpus under any set of features outperformed its baseline score, however, the differences between the scores using any of the feature sets were only marginal. The best scores in this setting were achieved when using feature set (3) and feature set (5).

The system performance using *ManConCorpus* revealed that the negation features are typically beneficial for enhancing system performance. Moreover, the incorporation of directionality and sentiment features in addition to the negation as in feature set (3) has slightly improved the performance to F-score of 0.83. This score was also achieved when using the negation, n-grams and sentiment features as in feature set (5). This may prove that although the sentiment features are not as common as negation and directionality, they are still important features to measure the assertion values of sentences. The use of all features as in feature set (7) provided the same F-score as in feature set (3) and (5) however the precision and recall scores were lower by 0.1.

The main observation from the scores of using *ManConCorpus* was that adding more features may not necessarily produce better results. For *AutConCorpus*, in contrast, adding more features would improve the system performance. This outcome was interesting because the initial estimation was that *ManConCorpus* would show an improvement in the performance as the features increase, and possibly with the use of *AutConCorpus*. However, Table (5.14) shows exactly the opposite outcome.

Moreover, the results of using *ManConCorpus* suggested that the best performance could be achieved using only negation, directionality and sentiment (feature set (3)). This was interesting since adding n-grams to the feature set (feature set (7)) did not enhance the performance, yielding a result which is not compatible with what has been reported in the literature that n-grams are useful to capture the context of text (see Section (4.3.3)) and should improve the performance. The results achieved using *AutConCorpus* suggested that the use of n-grams features in addition to the other three features would

improve the system performance. The results achieved using *AutConCorpus* were more intuitive than the results achieved using *ManConCorpus*.

Therefore, it appears that a discrepancy exists between the results of the two systems. This discrepancy is not compatible with the initial hypothesis that the change in the performance pattern of both systems should be similar. Without presenting the system performance scores using *AutConCorpus*, it would have been difficult to discover such inconsistency despite the quality of *AutConCorpus*.

One interpretation of the discrepancy between the performance patterns of the two systems could be related to the variation in corpus size used by the classifier, whereby the *ManConCorpus* is roughly half the size of the *AutConCorpus* corpus. The size of *AutConCorpus* might enable its classifier to capture more features that helped to improve the performance scores. In order to explore this problem, we have drawn the learning curves of both systems under the feature sets that showed incompatibility in their results, i.e. feature sets (3), (5) and (7).

Figure (5.15) presents the learning curves of the system using the two corpora (the left figure shows the learning curves of the system using *ManConCorpus*, while the right figures shows the learning curves of the system using *AutConCorpus*) considering the feature sets (3), (5) and (7) as previously presented in Table (5.14).

A learning curve is a plot used to show the improvement of a classifier while increasing the number of data points. This is useful to better understand the classifier's behaviour when increasing the size of the corpus. The learning curve figure shown here uses two curves: a training validation curve (red) and a K-folds cross-validation curve (green).

The training curve displays the classifier's accuracy as the dataset increases, using the same subset for training and testing. The cross-validation curve shows the classifier's accuracy as the dataset increases by using four-fold cross-validation, where 75% of the data in each fold is used for training and 25% for testing; moreover, the cross-validation score using a particular subset is the average accuracy of all K-runs in that subset.

The expected curve behaviour in the training validation is achieved by starting with a high accuracy when using a very small dataset; as the dataset size increases, the accuracy gradually decreases until certain level is reached where the classifier stays approximately constant even after the dataset size is increased. However, in the cross-validation curve, the ideal curve is to start with a low accuracy when the dataset size is small, and then gradually increase it as the dataset size increases. The cross-validation curve achieves its best performance where it converges with the score achieved by the

training validation curve.

Table (5.14) showed that the best performance using *ManConCorpus* was achieved using feature set (3); however, the learning curves of the corpus (see *Figure (3)*-left) showed that the best accuracy that can be achieved using the feature set was around 84%. The subsequent curves using feature sets (5) and (7) suggested an improvement in the overall accuracy as the dataset increases. This indicated that system reliability increased when feature set changed, and that the optimum system (in terms of F1-score and learning curve) was achieved using feature set (7). Moreover, the learning curves in these figures suggested that additional datasets may increase performance further. On the other hand, the reliability of the system using *AutConCorpus* increased when incorporating more features. The learning curves of the system tend to become more reliable as the feature sets change, and the best system performance and reliability is depicted using feature set (7) (see *Figure (7)*-right).

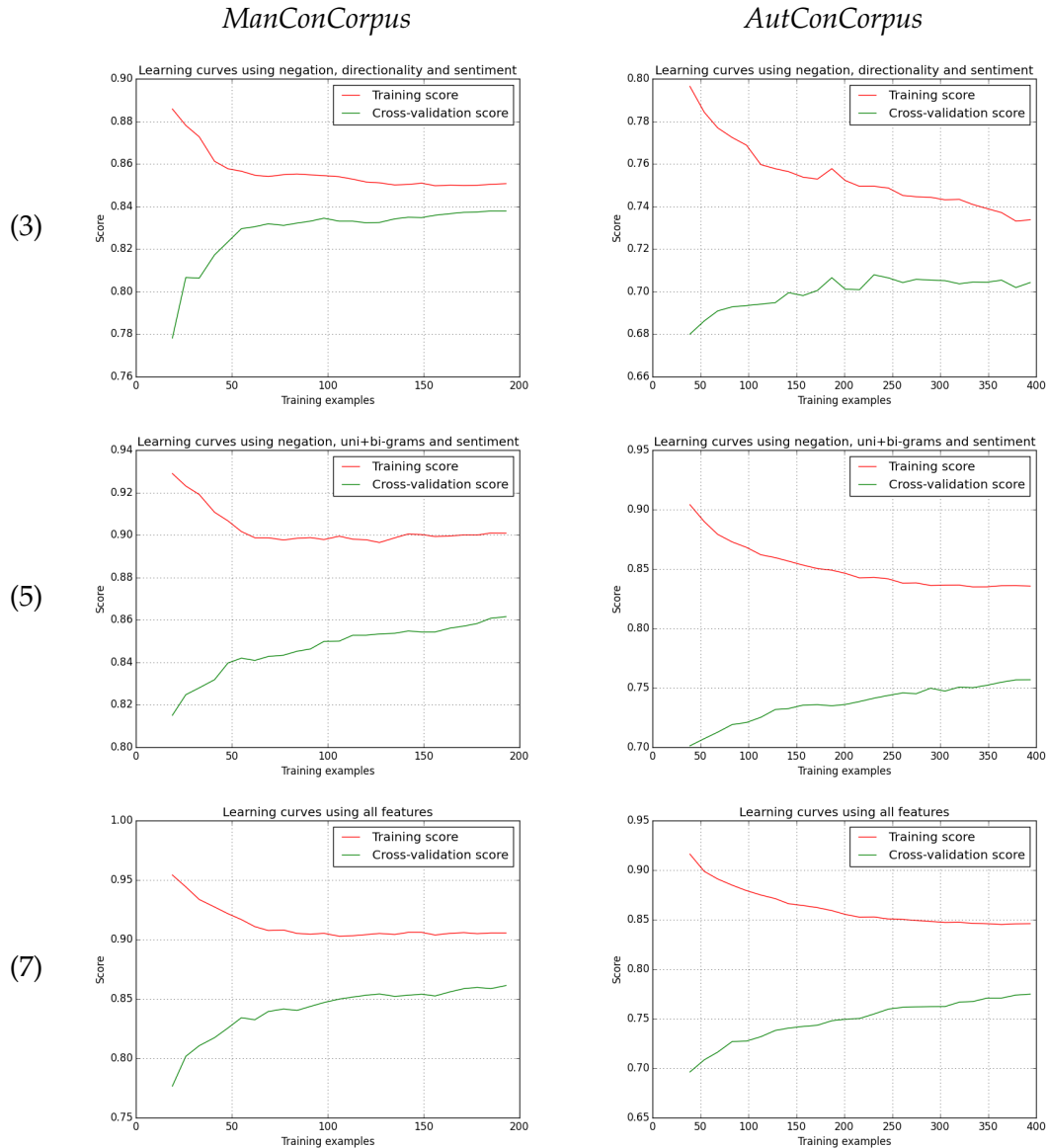


TABLE 5.15: The learning curves of using *ManConCorpus* and *AutConCorpus* using feature sets (3), (5) and (7)

Unlike the results of Table (5.14) which showed some discrepancy between the results of the contradiction detection system using *ManConCorpus* and *AutConCorpus*, the learning curves of the systems showed similar behaviour i.e the reliability of both systems increased as the feature set changed from feature set (3) to (5) and then to (7). At feature set (3) the reliability of both systems was not good; the reliability increased after using feature set (5) and the best reliability was achieved using feature set (7). This consistency suggests some homogeneity between the two corpora, which therefore supports our initial hypothesis that the automatic methodology could be a potential alternative to the manual methodology to construct a contradiction corpus.

To examine that hypothesis further, the classifier trained using *AutConCorpus* was also used to predict the annotations of *ManConCorpus*. Table (5.16) shows the results of that experiment. The system developed from the automatic corpus was able to achieve an average F-score of 72%, 4% higher than the baseline of *ManConCorpus*. The difference between the classifier performance and the baseline was not dramatically high, but, it was surprising that the classifier was able to exceed the baseline at all given that the instances of *AutConCorpus* were not restricted to sentences describing claims as in *ManConCorpus* and included titles that were not explicitly useful for such a task. Moreover, some instances included in the corpus were assigned the wrong assertion values due the errors generated by *SemRep* (see Section (3.4.5)).

Class	Precision	Recall	F1-Score
No	0.54	0.60	0.57
Yes	0.81	0.76	0.78
Micro-Average	0.72	0.71	0.72

TABLE 5.16: The result of using training the classifier on the dataset of *AutConCorpus* to predict the assertion values of the claims in *ManConCorpus* (the baseline is 68%)

The result of table (5.16) in addition to the result obtained from learning curves presented in Figure (5.15) support the hypothesis which suggests that the automatic methodology could be a potential alternative for the manual methodology. Note that this research does not suggest the use of *AutConCorpus* alone to develop a contradiction detection system, however, it suggests the use of the automatic methodology to construct a contradiction corpus.

5.7.3 Error Analysis

The reduction in the results scores were analysed, and two reasons were found to be responsible for this. Firstly, in some claims the relation extraction tools only identified a single relation tuple, which did not contain the information required to identify the correct assertion value. This therefore led the classifier to predict the assertion value based on incorrect information. Table (5.17) shows a claim and the relation tuple extracted as relevant to the question “*In patients undergoing coronary bypass surgery, does Aspirin usage, compared to no aspirin, cause bleeding?*”. This relation caused the classifier to predict the assertion value as *no*, which disagrees with the question. However, the correct information relevant to the question was “*There was more postoperative blood loss...*”, which agrees with the question.

Claim ID	L-Argument	Relation	R-Argument
There was more postoperative blood loss, on average, in patients treated with aspirin, but the difference was not .	the difference	was not	significant

TABLE 5.17: A relation tuple that was extracted from a claim and made the classifier to choose the wrong assertion value.

Secondly, when the relation extraction tool managed to extract multiple relation tuples from a claim, the relatedness score of the irrelevant tuple was sometimes higher than the one which was more relevant, causing the classifier to make its judgement based on the wrong tuple. Table (5.18) shows an example of two relation tuples that were extracted from the claim sentence “Binary logistic regression analysis with adjustments for age, gender, triglycerides, total cholesterol, low-density lipoprotein, high-density lipoprotein, apoAI, apoB, and LP(a) indicated that the TC and CC genotypes in SNP T-778C were not significantly associated with the development of CAD ..” in abstract (PMID#20360902). Based on the *UMLS-similarity* measurement, the first tuple scored 0.966, which means it is more related to the question “In the Han Chinese population, is SNP T-778C of apolipoprotein M associated with risk of developing diabetes or stroke?” than the second tuple. However, the second tuple in fact contains information that is more relevant to the question. This led the classifier to select the first relation tuple as the fact and to make the judgement according to that selection. However, a closer look at the second relation tuple may show that the tuple is more relevant to the question and contains linguistic features such as the negation term *not* that may change the judgement of the classifier.

Score	L-Argument	Relation	R-Argument
0.966	age	triglycerides	,
0.552	SNP T-778C	were not significantly associated with	the development of CAD

TABLE 5.18: Another example of the classifier errors due to choosing the relation tuple based on its relatedness score with the question.

Though the classifier performed reasonably well, it seems that the generic relation extraction tools used in this study were the main cause for the reduced performance. The development of a customised relation extraction tool for biomedical text seems an important step toward producing a reliable system to identify contradictory claims. Moreover, it seems that using the *UMLS-similarity* alone to measure the relatedness of relation tuples is not enough, given the fact that there are cases such as the relation

tuples of Table (5.18) in which such a tool failed to identify the tuple most relevant to the given question.

5.8 Evaluation

An early assumption considered in this research was that the components developed for the contradiction detection task are supposed to be integrated in larger information retrieval systems such search engines (see *Section (1.3)*). However, the performance of such components under a search engine system may be impacted. That is because the performance will be attributable to the quality of the search engine results. If an engine returned irrelevant abstracts as relevant, then the *Answer Selection Pipeline* component may be misled and recognize some sentences within the irrelevant abstracts as *potential answer* and the component performance may be therefore drops and consequently drops the performance of the *Contradiction Detection* component.

An alternative approach to estimate the performance of the components as one integrated system is to use the previous reported scores on each component. The performance of *Answer Selection Pipeline*, as reported in *Chapter (4)*, achieved a precision score of 0.62, a recall score of 0.6 and an F-score of 0.61 using *ManConCorpus-unst* as test data. The performance of the *Contradiction Detection* component, as reported in *Chapter (5)*, achieved a precision score of 0.83, a recall score of 0.83 and an F-score score of 0.83; however, these scores were computed using the cross-validation of the entire *ManConCorpus* rather than only using *ManConCorpus-unst* as in *Answer Selection Pipeline*. Thus, to integrate its performance with *Answer Selection Pipeline*, it is important to compute the performance of *Contradiction Detection* using only *ManConCorpus-unst*.

Table (5.19) shows that evaluation. A slight drop has occurred in the performance, possibly due to reducing the training data size, but the performance is still comparable with the performance reported in Table (5.14).

Class	Precision	Recall	F1-Score
No	0.84	0.57	0.67
Yes	0.81	0.93	0.87
Micro-Average	0.82	0.81	0.81

TABLE 5.19: The performance of the *Contradiction Detection* system using *ManConCorpus-unst*

The overall estimated performance of the overall system is computed by multiplying the performance scores of detecting *potential answers* in *Answer Selection Pipeline* by the average scores of the *Contradiction Detection* in table (5.19). Note that the scores of detecting the *non potential answers* were excluded here since these scores were biased (due to the data imbalance (see Section (4.3.5)). Table (5.20) shows the overall performance, which is lower than the performance of the individual components. This was expected given that the performance of the *Contradiction Detection* component was impacted by the results of *Answer Selection Pipeline*.

	Precision	Recall	F1-Score
Answer Selection Pipeline	0.62	0.60	0.61
Contradiction Detection	0.82	0.81	0.81
Overall Performance	0.51	0.49	0.49

TABLE 5.20: Estimated performance of the contradiction detection components in combination.

5.9 Research Claims Highlighter

PubMed search engine is a well-known IR system offering access to more than 25 million abstracts in the biomedical literature, including Medline[®]. *PubMed* takes a query and returns a list of abstracts that are determined as relevant or partially relevant to the query. As a result, the user has to review each abstract for further analysis and evaluation. Researchers conducting a systematic review (Gough, Oliver, and Thomas, 2012), for example, typically follow this approach when collecting studies of interest; however, significant effort is usually devoted to identification of relevant studies.

Research Claims Highlighter is an intelligent web search engine that highlights the most relevant claims (if any) in each abstract returned by the engine in response to the query. The system overcomes the limitation of *PubMed*, since it highlights the sentences that are likely to be claims but at the same time may contain information to answer the query. Furthermore, the system provides different colour codes to claims that agree with the query versus claims that disagree in order to identify those that are potentially contradictory. Thus, the user can focus on the highlighted piece of information instead of scanning the entire abstract to evaluate its relevancy to the query and to identify whether the claims agree or disagree with the query.

The system thus allows users to increase their efficacy and reduce their cognitive load when searching for abstracts. The motivation behind developing this system is to demonstrate how *Claim Highlighter*, which consists of the *Answer Selection Pipeline* component and *Contradiction Detection* component can be used to resolve the limitations

of current biomedical IR systems. Figure (5.1) shows the Research Claims Highlighter system architecture, which is composed of five main components: *Claim Zoning*, *Search Engine* (Ounis et al., 2005), *Answer Selection*, *Contradiction Detection* and *User Interface*.

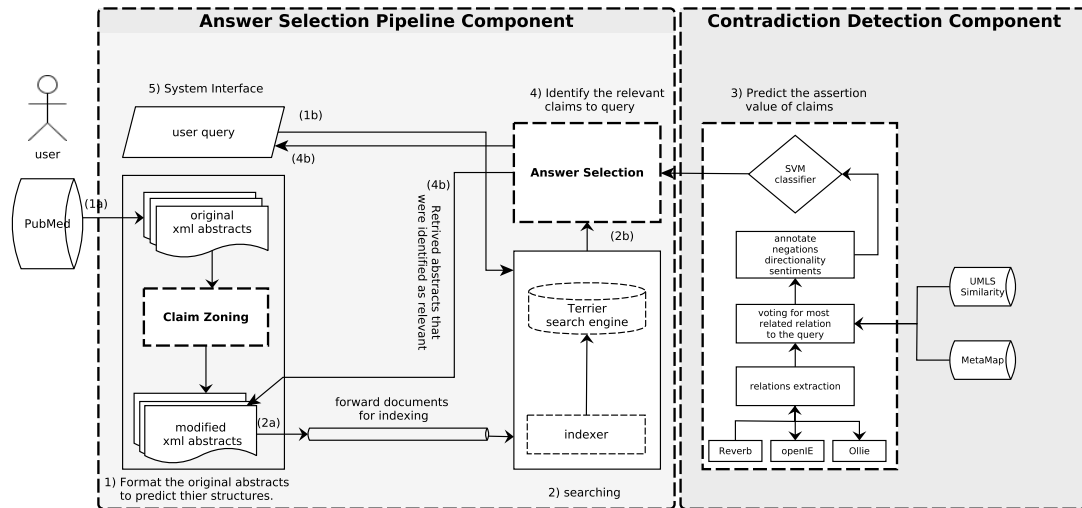


FIGURE 5.1: Research Claims Highlighter Architecture

The system can be viewed at two levels: system and user. At the system level, the system takes a collection of PubMed XML abstracts, extracts the sentence roles and computes the *Z-scores* of their $sim(s,T)$, and $sim(s,Q)^2$, and generates a newly formatted XML file as shown in Figure (5.2). At the user level, the user enters a query to the search engine, and the engine returns a list of relevant abstracts. However, before displaying the abstracts to the user, the *Answer Selection* reads the content of each abstract to predict which sentences are likely to provide a potential answer to the query. Sentences that were predicted as *potential answers* are highlighted in blue.

² $sim(s,Q)$ is computed on the fly.

```

<ABSTRACT>
<DOCUMENT>
  <DOCNUM>44</DOCNUM>
  <PMID>309032</PMID>
  <UNSTRUCTURED>TRUE</UNSTRUCTURED>
  <TITLE>Relation of preoperative use..</TITLE>
  <INTRODUCTION>
    <span id="1" z_simt="290.0">To evaluate the ..</span>
  </INTRODUCTION>
  <RESULTS>
    <span id="1" z_simt="29.0" >Preoperative ..</span>
    <span id="2" z_simt="126.0">Mean ..</span>
    <span id="3" z_simt="-1.0">than in ..</span>
  </RESULTS>
  <CONCLUSIONS>
    <span id="1" z_simt="179.0">The degree ..</span>
    <span id="2" z_simt="51.0">In addition..</span>
  </CONCLUSIONS>
  <INTRODUCTION.TEXT>To evaluate the potential..</INTRODUCTION.TEXT>
  <RESULTS.TEXT>Preoperative prothrombin ..</RESULTS.TEXT>
  <CONCLUSIONS.TEXT>The degree of ..</CONCLUSIONS.TEXT>
</DOCUMENT>
</ABSTRACT>

```

FIGURE 5.2: A sample of a formatted XML file

The system interface consists of two pages. The first one takes the user query and displays its relevant abstracts. The second displays the details of each abstract. Figure (5.3) shows the user query and the first two results in descending order. Each abstract is associated with meta information such as the title (hyperlink), PMID, a flag to show whether the abstract is structured or not, and a snippet of its textual information. When the user selects a particular abstract, a new page containing the abstract details is displayed. The system highlights the sentences predicted as *potential answers* to the query, thus allowing the user to focus on the portion of text relevant to the query.

Figure (5.4) shows an example of an abstract that was originally unstructured. The highlighter system has structured it and highlighted the most informative sentence to answer the query “*In patients with heart disease, does aspirin cause postoperative bleeding?*”

The University Of Sheffield.
Research Claims Highlighter

In patient with heart disease, does aspirin cause postoperative bleeding

Highlighter Off On

Results for In patient with heart disease, does aspirin cause postoperative bleeding, displaying 1- 11 of 1000

Document	
	Effect of aspirin on postoperative bleeding in coronary artery bypass grafting.
	PMID - 18818571 UNSTRUCTURED - FALSE
	Our study suggests that contrary to the commonly held beliefs in our setup, the use of aspirin till the date of surgery does not increase the risk of postoperative bleeding after CABG. In contrast, our data show reductions in the bleeding incidence of those in whom aspirin was not withheld prior to surgery. Therefore we strongly recommend its continued use of aspirin until the date of surgery.
	Prediction of the excessive perioperative bleeding in patients undergoing coronary artery bypass grafting: role of aspirin and platelet glycoprotein IIIa polymorphism.
	PMID - 16153930 UNSTRUCTURED - FALSE

FIGURE 5.3: Research Claims Highlighter System interface

Unstructured?:	TRUE
PMID	22781376
TITLE	[Aspirin in primary prevention of cardiovascular diseases: how to balance risks and benefits].
INTRODUCTION	While the use of aspirin in the secondary prevention of cardiovascular atherothrombotic disease is well established, many aspects of primary prevention are still unclear. Uncertainties mostly depend on a doubtful risk-benefit ratio, because of the low atherothrombotic risk of populations involved on the one hand, and the non-negligible bleeding risk of treatment on the other. Areas of specific doubt are those of diabetes and asymptomatic peripheral arterial disease, where neither single trials nor meta-analyses allow issuing high-grade specific recommendations at the moment. The present review aims at giving an account on this topic, highlighting areas for further studies, but also attempting at providing a rationale for what to do practically now, while awaiting more conclusive evidence.
METHOD	
RESULTS	
CONCLUSIONS	Based on the results of a number of clinical trials and meta-analyses, and especially considering the absolute figures of the benefit (major cardiovascular events avoided) and of the harm (major bleeding events occurred related to aspirin), the authors recommend to limit primary cardiovascular prevention with aspirin (in apparently healthy subjects with no previous cardiovascular events) to subjects with an estimated global cardiovascular risk ≥ 2 major cardiovascular events per 100 patients-year, as assessed by the risk score assessments proposed in the Italian "Progetto Cuore" (www.progettocuore.it). This cut-off should also be adopted for primary prevention in patients with type 2 diabetes and/or asymptomatic peripheral arterial disease.

FIGURE 5.4: The system configured to display the unstructured abstracts as structured

To further develop the highlighting capability, the *Contradiction Detection* was integrated into the system in order to colour code claims in terms of agreement or disagreement with the question. Figure (5.5) and (5.6) illustrate two abstracts obtained by querying the search engine with the phrase, "in patients with heart disease, does aspirin cause preoperative bleeding?". The first abstract features a claim sentence which was automatically highlighted in green, indicating that the claim agrees with the query. The second abstract shows a claim sentence that was automatically highlighted in yellow, suggesting that the claim disagrees with the query.

Unstructured?:	FALSE
PMID	22271021
TITLE	Aspirin use and bleeding risk after endoscopic submucosal dissection in patients with gastric neoplasms.
INTRODUCTION	The risk of bleeding after endoscopic submucosal dissection (ESD) in patients with early gastric neoplasms who do not discontinue aspirin for the procedure has not been established. We aimed to investigate whether post-ESD gastric bleeding is increased in patients who take aspirin.
METHOD	Patients who underwent ESD for early gastric neoplasms at the National Cancer Center Hospital, Korea, between November 2008 and January 2011 were enrolled. The risk of post-ESD bleeding was evaluated using Poisson regression analysis.
RESULTS	We categorized 514 patients into three groups according to aspirin intake at the time of the procedure: patients who never used aspirin (n=439), patients who interrupted aspirin use for 7 days or more (n=56), and patients who continuously used aspirin (n=19). Post-ESD bleeding occurred in 4.1% (21/514) overall, and was more frequent in continuous aspirin users (4/19 [21.1%]) than in those who never used aspirin (15/439 [3.4%]) (P=0.006) and those with interrupted aspirin use (2/56 [3.6%]) (P=0.033). Multivariate analysis showed that use of aspirin by itself was associated with post-ESD bleeding (relative risk [RR] 4.49; 95% confidence interval [95%CI] 1.09-18.38). The resumption of clopidogrel combined with aspirin use (RR 26.71, 95%CI 7.09-100.53), and increased iatrogenic ulcer size (RR 1.52, 95%CI 1.14-2.02), were significantly associated with post-ESD bleeding.
CONCLUSIONS	Continuous aspirin use increases the risk of bleeding after gastric ESD. Aspirin use should be stopped in patients with a low risk for thromboembolic disease to minimize bleeding complications.

FIGURE 5.5: An abstract featuring a claim that agrees with the query

Unstructured?:	FALSE
PMID	12842510
TITLE	Effect of preoperative aspirin use in off-pump coronary artery bypass operations.
INTRODUCTION	The effect of preoperative aspirin use until the day of operation on mortality rate and bleeding risks in patients who had on-pump coronary artery bypass operation has been well documented. However, the effect of aspirin use in patients undergoing off-pump coronary artery bypass operation (OPCAB) with regard to postoperative blood loss and morbidity has not been studied. We aimed to determine the effects of continuing aspirin therapy preoperatively.
METHOD	We performed a retrospective study of 340 patients who had first-time OPCAB between January 1998 and September 2001. A propensity score for receiving aspirin until the day of operation was constructed from core patient characteristics. All aspirin users (n = 170) were matched with unique 170 nonaspirin users by identical propensity score. The primary outcome measures were in-hospital mortality rate and hemorrhage-related outcomes (postoperative blood loss in the intensive care unit, reexploration for bleeding, and blood product requirements). Secondary outcome measures were stroke, myocardial infarction, gastrointestinal bleeding, and sternal wound infections.
RESULTS	There were no differences in patient characteristics between aspirin users and nonaspirin users. The average postoperative blood loss (845 mL versus 775 mL; p = 0.157) and the rate of reexploration for bleeding (3.5% versus 3.5%; p > 0.99) were similar in aspirin users and nonaspirin users. We found no significant difference between blood product requirements for the two groups. Similarly, we found no significant difference in the incidence of the secondary outcomes.
CONCLUSIONS	Preoperative aspirin did not increase bleeding-related complications, mortality rate, or other morbidities in patients who had off-pump coronary artery operation.

FIGURE 5.6: An abstract featuring a claim that disagrees with the query

5.10 Conclusions

Automatic identification of potentially contradictory claims would be extremely useful to individuals working with biomedical literature. This chapter described a supervised machine learning system to automatically identify contradictory claims. The system was assessed using manually and automatically annotated corpora, both of which indicated that the combined use of negation, n-grams, directionality and sentiment features can produce a reliable system to detect contradictory claims. Furthermore, the classifier trained on the automatic corpus was used to predict the assertion values of claims in the manually constructed corpus and the results supports the hypothesis that generation of a corpus from *SemMedDB* is useful for developing a classifier to detect contradiction between claims.

Chapter 6

Conclusions

Contradictions between research claims are not uncommon in the field of biomedical research. This hinders the ability of decision-makers to make informed decisions, which could adversely affect human lives. Though multiple approaches and systems have emerged to minimise the problems raised by contradictory claims, many of them are both time-consuming and difficult to use. Such approaches require researchers or editors to manually screen research abstracts in order to identify their relevancy to a research question and determine which studies agree or disagree with a given question.

This study aimed to explore the contradiction problem in biomedical abstracts using NLP techniques. The study proposed an automatic system to identify contradictory claims in biomedical abstracts. The implementation of such a system will help to minimise the workload of researchers or editors during the process of screening research abstracts. This study resulted in three main outcomes:

1. A novel methodology for constructing a corpus of contradictions using systematic reviews and forest plot diagrams. This methodology was found to be useful if there are sufficient resources to collect, annotate and evaluate the dataset.
2. A novel methodology for automating the construction of a corpus of contradictions corpus using the *SemMedDB* knowledge resource. This methodology was found to be useful for generating a large corpus in a short period of time.
3. A pipeline to detect contradiction between research claims in abstracts. The pipeline is composed of three machine learning classifiers. The first classifier annotates sentences in an abstract with their associated rhetorical labels. The second classifier annotated sentences that belong to the *Results* and *Conclusions* sections to identify those considered potential answers for a research question. The third classifier predicts the assertion values of the potential answers in order to identify contradictory claims. The results of these experiments showed that the automatic detection of contradictory claims in the domain of biomedicine using NLP methods is a tangible problem.

6.1 Summary of Contributions

The main objectives listed at the beginning of this study have been met as follows:

1. This study constructed two corpora, using novel approaches. The first corpus was gathered from systematic reviews, and their forest plot diagrams were used to assess the annotators in terms of whether the studies used in a review were likely or unlikely to contain contradictory claims. The second corpus was automatically generated without human annotation using the repository of *SemMedDB*.
2. This study developed a novel approach to identify research claims in abstracts. Human behaviour was synthesised when screening an abstract, where a portion of text likely to contain the author's claim is identified, followed by the sentences in the abstract which represents the claim relevant to the given question. A pipeline system was developed for this purpose, which consisted of two machine-learning classifiers. The first classifier achieved an F1-score of 91% ; the second classifier achieved an F1-score 78%, and both classifiers were integrated into a pipeline, which achieved a reasonable F1-score of 61% for annotating potential answers.
3. This research utilised a novel approach to develop a system to detect contradictory claims using their assertion values with respect to a query. The exploration of linguistic features to develop the system revealed that four simple features were found to be useful: n-grams, negation, sentiment and directionality. The system consists of two stages: fact extraction and fact assertion value detection. The system achieved an F1-score of 83% using *ManConCorpus*, and an F1-score of 78% using *AutConCorpus*. Furthermore, the latter system was used to annotate the claims of *ManConCorpus* and achieved an F1-score of 72%.

6.2 Future Work and Open Questions

This thesis has explored the problem of identifying contradictions between research claims in biomedical abstracts. However, there are still a lot of opportunities for extending the scope of this thesis remain. The *ManConCorpus* corpus was constructed based on choosing only one claims sentence from an abstract that answers a given question. However, the observations of the previous experiments described in sections (4.2) and (4.3) showed that it is possible for an abstract to contain more than one sentence which answers the same question. Moreover, experiments in Chapter (5) revealed that there was no significant difference between the classifier trained using claims sentences or

general sentences to detect contradictions between claims. Therefore, it will be useful to integrate those sentences in the *ManConCorpus* even if they are not claims in order to determine how the performance of the pipeline system described in section (4) will be affected given the fact that some of the errors found there could be potential answers.

Secondly, the *ManConCorpus* corpus was constructed using a single attribute to show whether the claim agrees (*yes*) or disagrees (*no*) with the question. It would be interesting to discover the feasibility of annotating such a corpus on a detailed level to better understand the underlying reasons behind the contradictions. For example, new annotation attributes could be added to each claim sentence in order to highlight linguistic characteristics such as negation, sentiment, directionality or negative directionality, etc., which cause a sentence to either agree or disagree with the given question.

Thirdly, the methodology followed to construct *AutConCorpus* makes it possible to automate the process of identifying all possible incompatible relation tuples in *SemMedDB* repository, and consequently extract their sentences as potentially contradictory. This will likely generate a much larger corpus than that developed in this study. Furthermore, it is pertinent to evaluate how the size of such a corpus might increase the efficiency of detecting contradictory claims.

Fourthly, one bottleneck in the contradiction detection system is the relation extraction systems used to extract the common facts between claims for a particular question. The current tools are either highly specific (such as *SemRep*), or overly generic (such as *OpenIE*). Therefore, it is important to determine the feasibility of developing a relation extraction system tailored for contradiction detection tasks, which can recognise biomedical concepts and preserve the scopes of negation, directionality and sentiment while extracting relation tuples.

Finally, an intelligent search engine that automatically highlights research claims in biomedical abstracts relevant to a particular query can be developed. The search engine can make use of the *Claim Zoning* and *Answer Selection* components described in Chapter (4) to highlight relevant claims. Such a system will allow users to increase their efficacy and reduce their cognitive load during searching for abstracts since they will focus on highlighted piece of information rather than an entire abstract. To enhance the capability of the search engine further, the component described in Chapter (5), responsible for detecting the assertion value of claims, can be incorporated into the engine in order to colour code claims in term of agreement or disagreement with the query, in which two claims that are coded with different colours are considered contradictory or in disagreement.

Appendix A

Corpus Annotation Guidelines:

A.1 Definitions

Please read and carefully consider the following definitions before proceeding with the annotation task:

- A review abstract is the abstract of a systematic review.
- A study abstract is an abstract of a study used in a systematic review to answer the review question.
- A PICO question is a well-defined question that includes four parts: population, intervention, comparator and outcome.
- A research claim is the most important point that research authors want to present to the reader. It is the overall conclusion or outcome that can be understood from the research findings/results. Thus, a claim is not a result but the interpretation of the results.
- A causal claim is a claim that suggests a relationship between two concepts and asserts that a concept has an influence on the other concept. The relationship can be direct (e.g. cause, increase, decrease and protect) or indirect (e.g. is associated with). An example of a causal claim based on a direct relationship is MN-BMC transplantation improves cardiac function in ischemic heart failure patients during CABG. A example causal claim based on an indirect relationship is These results suggest that there is no HLA association with ischemic heart disease.
- An evaluative claim is a claim that expresses a value judgement about a treatment or process. This can be expressed by stating the value directly or by comparing it against something else. An example of an evaluative claim is Combined clopidogrel and aspirin are safe for bleeding. Another example is The reduction in hospitalizations achieved using standardized telephonic case management in the

early months after a heart failure admission is greater than that usually achieved with pharmaceutical therapy. Annotation Process

The annotation process consists of four stages, to be carried out in turn.

A.1.1 Formulation of PICO Questions

Formulate a PICO question for each review abstract. This question will be used in the later stages of the annotation. Please follow this process to formulate the question:

1. Read the title of the review abstract and its content to understand the objective of the review.
2. Read the title of each study abstract associated with the review and examine its content, particularly the conclusion sections, to ensure that the study is directly relevant to the question addressed in the review. Exclude any studies that are found not to be directly associated with the main objective of the review, or where the association is unclear.
3. Formulate a PICO question for each review. The question should be a closed question; in other words, it can be answered with either a yes or no.

Notes

There may be cases where there is incompatibility between the populations considered in different studies or studies use alternative terms to refer to the same or similar concepts (e.g. cardiovascular disease and myocardial infarction). In these cases the question may be formulated using either (a) a generic term covering all the concepts, or (b) list all terms via the use of or, e.g. in patients with X condition, is y associated with cardiovascular disease or myocardial infarction.

A.1.2 Identification of Claims

The objective of this stage is to identify the best sentence within the each study abstract that answers the question formulated in the previous stage. For each abstract associated with a review:

1. Carefully read the question associated with the review.
2. Examine each study abstract and identify the best sentence that serves as an answer to the review question.

Notes

The claim sentence can usually be found in the conclusions section of the study abstract. This can be identified by the use of the explicit label (Conclusion/Conclusions) or implicitly by the use of signal words such as In conclusion,, We found that... and Our work suggests... In cases where no sentence providing an answer to the question is found in the conclusion section, a sentence from the results section can be chosen; provided the sentence answers the question and can be considered as a claim. If no suitable sentence can be identified the study abstract should be excluded from the set of abstracts associated with that particular review. In cases where more than one sentence that could potentially serve as the answer to the review question is identified, the annotator should choose the sentence that provides the clearest answer to the question considering all of the information contained in the study abstract.

A.1.3 Annotation of Claims Assertion Values

Provide an assertion value for each claim with respect to the question. Two possible values can be assigned: YS and NO. YS should be used when the claim asserts a positive answer to the question and NO if it does not. (If the claim neither asserts nor negatives the question then the assertion value should be NO).

A.1.4 Annotation of the Claim Type

Annotate each claim as either causal or evaluative (see the definitions above). Causal claims should be annotated as CAUS and evaluative claims as EVAL.

Notes

Claims in biomedical abstracts tend to be complex and a claim can be interpreted as causal and evaluative at the same time. For example, Among our population of largely low or asymptomatic HCM patients, the presence of scar indicated by CMR is a good independent predictor of all-cause and cardiac mortality. This claim states that the scar indicated by CMR is a predictor of all causes and cardiac mortality, which shows an indirect causal relationship between the scar and cardiac mortality. However, at the same time the claim evaluates this relation using the term good. In such cases, the annotator should consult the abstract content to determine whether the purpose of the study is to identify an association between the scar and mortality or to evaluate to what degree the scar can be used as a predictor for cardiac mortality.

Appendix B

The Questions Formulated for ManConCorpus

Review-PMID	Question
22498326	In patients with HCM, does using imaging technique, compared to conventional techniques, serve as a predictor for adverse prognosis?
23623290	In patients with chronic heart disease, does Bone marrow Stem cell transplantation or injection, compared to none, improve cardiac function?
21556773	In patients with dilated cardiomyopathy, are HLA genes associated with development of Dilated Cardiomyopathy?
24040766	In Han Chinese population, is SNP T-778C of apolipoprotein M associated with risk of developing Diabetes or stroke?
24212980	In patients undergoing coronary bypass surgery, does Aspirin usage, compared to no aspirin, cause bleeding?
24035160	In patients undergoing coronary artery bypass, does the combination of aspirin and clopidogrel, compared to aspirin alone, prevent graft occlusion or improve patency?
24172075	In patients undergoing coronary by pass surgery, is Off-pump, compared to conventional on pump coronary artery bypass grafting, more beneficial?
24135644	In patients with coronary artery disease, is mutation or polymorphisms in endothelial nitric oxide synthase gene associated with CAD or MI or ACS development?
24036021	In patients with atherosclerotic plaque or myocardial infaction, does -463G or -463A polymorphism in MPO gene influence MI or CAD development?
24039708	In patients with coronary artery disease (CAD), is C242T polymorphism of P22(PHOX) gene associated in development of CAD?
Continued on next page	

24090581	In patients with coronary artery diseases, does combining CABD and CEA, compared with CABG or CEA alone, reduce morbidity?
24165432	In elderly patients with CHF, does physical exercise or cardiac rehabilitation, compared to no exercise, improve cardiac function?
23962886	In patients with heart failure, do statin drugs treatment, compared to non statin drug, treatment improve cardiac function or prevent cardiac morbidity?
23219304	In patients with renal or cardiovascular disease, does treatment with ACE inhibitors, compared with placebo, improve renal function or protect against cardiovascular incidents respectively?
23181122	In the elderies, is n-3 fatty acid from fish intake associated with reduction in risk of developing heart failure?
23489806	In the elderlies, does omega 3 acid from fatty fish intake, compared with no consumption, reduce the risk of developing heart failure?
24163234	In patients with CHF, does care giving or teleguidance-telecare, compared to usual care, reduce morbidity?
21521728	In patients with advanced diabetes, does treatment with antihypertensives, compared with placebo, improve renal function or protect against cardiovascular incidents?
22854636	In patients with hypertension, does revascularisation, compared with medical therapy, improve blood pressure?
22795718	In patients with hypertension, does treatment with ACE inhibitors, compared to placebo, reduce risk of cardiovascular event or improve blood pressure?
23602289	In patients with hypertesion or hypercholesterolemia, does statin drugs, compared to placebo, reduce blood pressure or lipid levels?
23435582	In women with pre-eclampsia, does treatment with L Arginine, compared to placebo, reduce blood pressure or pre-eclampsia?
22086840	In women with pre-eclampsia, is Polymorphism in angiotensin gene associated with pre-eclampsia?
23223091	In women with pre-eclampsia, is mutation in renin-angiotensin gene associated with pre-eclampsia?

TABLE B.1: A list of the 24 questions formulated for the final corpus

Appendix C

Lexicons

C.1 Negation

Term	Value
absence	<i>not</i>
cannot	<i>not</i>
deny	<i>not</i>
negative	<i>not</i>
never	<i>not</i>
no	<i>not</i>
nor	<i>not</i>
not	<i>not</i>
nothing	<i>not</i>
out	<i>not</i>
without	<i>not</i>
similar	<i>not</i>
lack	<i>not</i>
none	<i>not</i>
unlikely	<i>not</i>
fail	<i>not</i>

TABLE C.1: The negation lexicon

C.2 Directionality

Term	Value	Term	Value	Term	Value
alleviate	less	shallow	less	improve	more
ameliorate	less	short	less	include	more
antagonize	less	shorten	less	increase	more
arrest	less	shortening	less	incremental	more
attenuate	less	silence	less	induce	more
benign	less	silent	less	inducing	more
block	less	slight	less	infect	more
blocking	less	slow	less	infective	more
brief	less	slowing	less	inflate	more
compress	less	small	less	influence	more
conservative	less	smaller	less	intense	more
conserve	less	soft	less	intensify	more
cut	less	suppress	less	intensive	more
decrease	less	terminate	less	intervene	more
decreased	less	unaffected	less	intoxicate	more
deficient	less	underestimate	less	invade	more
deficit	less	undermine	less	larger	more
degenerate	less	underscore	less	lead	more
degenerative	less	undetected	less	lengthening	more
degrade	less	weak	less	lengthy	more
delay	less	weaken	less	lethal	more
delete	less	withdraw	less	lift	more
delineate	less	withhold	less	longer	more
deplete	less	withholding	less	macroscopic	more
depress	less	worse	less	major	more
depressive	less	worsen	less	malignant	more
deprive	less	worsening	less	manifest	more
descend	less	abuse	more	many	more
destroy	less	accelerate	more	massive	more
deter	less	accumulate	more	maximal	more
deteriorate	less	activate	more	maximise	more

Continued on next page

Table C.2 – continued from previous page

Term	Value	Term	Value	Term	Value
devastate	less	activating	more	maximize	more
diminish	less	acute	more	mediate	more
disable	less	add	more	metastasis	more
disabled	less	addictive	more	modulate	more
disappear	less	additive	more	more	more
discourage	less	adequate	more	most	more
dissatisfy	less	advance	more	motivate	more
disturb	less	aged	more	motivating	more
down	less	aggravate	more	much	more
dysfunctional	less	aggregate	more	multiple	more
eliminate	less	allergenic	more	necessitate	more
emptying	less	allow	more	numerous	more
eradicate	less	amplify	more	outstanding	more
exclude	less	anxious	more	outweigh	more
fall	less	ascending	more	overactive	more
falls	less	aseptic	more	overall	more
few	less	assist	more	overcome	more
fewer	less	augment	more	overestimate	more
hamper	less	benefit	more	overload	more
hinder	less	big	more	overweight	more
ignore	less	bigger	more	overwhelm	more
impair	less	broad	more	pacing	more
impede	less	broaden	more	pandemic	more
inaccurate	less	bulky	more	persist	more
inactivate	less	cause	more	plus	more
inactive	less	combine	more	popular	more
inadequate	less	conspicuous	more	potent	more
incomplete	less	contribute	more	potentiate	more
inferior	less	develop	more	power	more
infrequent	less	diffuse	more	powerful	more
inhibit	less	dominant	more	predominant	more
insignificant	less	dominate	more	predominate	more
Continued on next page					

Table C.2 – continued from previous page

Term	Value	Term	Value	Term	Value
insufficient	less	double	more	prevalent	more
lack	less	doubling	more	produce	more
least	less	elevate	more	progress	more
less	less	elevated	more	proliferate	more
limit	less	emphasis	more	prolong	more
limited	less	emphasise	more	prominent	more
little	less	emphasize	more	promote	more
lose	less	emphasizing	more	provoke	more
low	less	empower	more	quick	more
lower	less	enable	more	raise	more
lowering	less	encourage	more	raising	more
lowest	less	endemic	more	rapid	more
marginal	less	enforce	more	regenerate	more
mere	less	enhance	more	reinforce	more
microscopic	less	enlarge	more	repeat	more
mild	less	enormous	more	repetitive	more
minimal	less	enrich	more	replicate	more
minimise	less	epidemic	more	reproduce	more
minimize	less	escalate	more	reward	more
minimum	less	exacerbate	more	rich	more
minor	less	exaggerate	more	rise	more
mitigate	less	exceed	more	rising	more
myopic	less	excess	more	robust	more
negligible	less	excessive	more	speed	more
obscure	less	excite	more	spread	more
obstruct	less	exert	more	spreading	more
obviate	less	exhaustive	more	stimulate	more
occasional	less	expand	more	strengthen	more
preserve	less	expedite	more	strengthening	more
prevent	less	explosive	more	stressful	more
preventive	less	extend	more	stretching	more
protect	less	extensive	more	strong	more
Continued on next page					

Table C.2 – continued from previous page

Term	Value	Term	Value	Term	Value
quiet	less	extra	more	substantiate	more
rare	less	extrapolate	more	succeed	more
recessive	less	facilitate	more	sufficient	more
reduce	less	fast	more	superior	more
reducing	less	favor	more	support	more
reject	less	favour	more	supporting	more
relapsing	less	fortify	more	sweeten	more
relieve	less	fruitful	more	swelling	more
reluctant	less	full	more	thicken	more
remove	less	gain	more	traumatize	more
resist	less	greater	more	trigger	more
resistant	less	grow	more	ultrasonic	more
resistive	less	hard	more	uncontrolled	more
restrict	less	hasten	more	uninfected	more
restrictive	less	healthier	more	up	more
retard	less	heavy	more	upper	more
retarded	less	heighten	more	widespread	more
reverse	less	high	more	yield	more
scant	less	higher	more		
scarce	less	huge	more		
sedentary	less	implement	more		

TABLE C.2: Directionality lexicon

C.3 Sentiment

Term	Value	Term	Value
normal	bad	advantageous	good
abusive	bad	appropriate	good
adverse	bad	beneficial	good
aggressive	bad	best	good
bad	bad	better	good
damage	bad	competent	good
debilitate	bad	creative	good
defective	bad	effective	good
deleterious	bad	efficacious	good
detrimental	bad	efficient	good
disadvantage	bad	favorable	good
disrupt	bad	favourable	good
fatal	bad	good	good
harmful	bad	healthy	good
hazardous	bad	helpful	good
ill	bad	optimal	good
infectious	bad	positive	good
infertile	bad	protect	good
injurious	bad	protective	good
invasive	bad	right	good
metastatic	bad	safe	good
poor	bad	successful	good
severe	bad	suitable	good
toxic	bad	valid	good
unfavorable	bad	valuable	good
unfavourable	bad	viable	good
unsuccessful	bad	safety	good
worst	bad		
wrong	bad		

TABLE C.3: The sentiment lexicon

Bibliography

- Agarwal, Shashank and Hong Yu (2009). "Automatically classifying sentences in full-text biomedical articles into Introduction, Methods, Results and Discussion." In: *Bioinformatics* 25.23, pages 3174–3180.
- Agarwal, Shashank and Hong Yu (2010). "Detecting Hedge Cues and Their Scope in Biomedical Text with Conditional Random Fields". In: *Biomedical Informatics* 43.6, pages 953–961.
- Ahlers, Caroline, Marcelo Fiszman, Dina Demner-Fushman, François-Michel Lang, and Thomas Rindflesch (2007). "Extracting semantic predications from Medline citations for pharmacogenomics." In: *The Pacific Symposium Biocomputing*, pages 209–220.
- Airola, Antti, Sampo Pyysalo, Jari Bjorne, Tapio Pahikkala, Filip Ginter, and Tapio Salakoski (2008). "All-paths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning". In: *BMC Bioinformatics* 9.Suppl 11, S2.
- Andrade, Daniel, Masaaki Tsuchida, Takashi Onishi, and Kai Ishikawa (2013). *Detecting Contradiction in Text by Using Lexical Mismatch and Structural Similarity*. Proceedings of the 10th NTCIR conference.
- Androutsopoulos, Ion and Prodromos Malakasiotis (2010). "A Survey of Paraphrasing and Textual Entailment Methods". In: *Artificial Intelligence Research* 38.1, pages 135–187.
- Aronow, David, Feng Fangfang, and W. Bruce Croft (1999). "Ad Hoc Classification of Radiology Reports". In: *Journal of the American Medical Information Association* 6.5, pages 393–411.
- Aronson, Alan (2001). "Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program." In: *The American Medical Informatics Association Annual Symposium Proceedings*, pages 17–21.
- Athenikos, Sofia J. and Hyoil Han (2010). "Biomedical Question Answering: A Survey". In: *Computer Methods and Programs in Biomedicine* 99.1, pages 1–24.
- Baker, Collin, Charles Fillmore, and John Lowe (1998). "The Berkeley FrameNet Project". In: *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*. ACL '98. Stroudsburg, PA, USA: Association for Computational Linguistics, pages 86–90.

- Bal, Bal Krishna and Patrick Saint-Dizier (2009). "Towards and Analysis of Argumentation Structure and the Strength of Arguments in News Editorials". In: *AISB Symposium on Persuasive Technologies, Edinburgh*. <http://www.aisb.org.uk/>: The Society for the Study of Artificial Intelligence and Simulation of Behaviour (SSAISB), pages 55–63.
- Banko, Michele, Michael J. Cafarella, Stephen Soderland, Matt Broadhead, and Oren Etzioni (2007). "Open Information Extraction from the Web". In: *Proceedings of the 20th International Joint Conference on Artificial Intelligence. IJCAI'07*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., pages 2670–2676.
- Bentivogli, Luisa, Peter Clark, Ido Dagan, Hoa T. Dang, and Danilo Giampiccolo (2010). "The sixth PASCAL recognizing textual entailment challenge". In: *The Text Analysis Conference (TAC 2010)*.
- Berger, Adam L., Vincent J. Della Pietra, and Stephen A. Della Pietra (1996). "A Maximum Entropy Approach to Natural Language Processing". In: *Computational Linguistics* 22.1, pages 39–71.
- Besnard, Philippe, Alejandro Garcia, Anthony Hunter, Sanjay Modgil, Henry Prakken, Guillermo Simari, and Francesca Toni (2014). "Introduction to structured argumentation". In: *Argument & Computation* 5.1, pages 1–4.
- Bhagat, Rahul, Patrick Pantel, and Eduard Hovy (2007). "LEDIR: An Unsupervised Algorithm for Learning Directionality of Inference Rules". In: *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*. Prague, Czech Republic: Association for Computational Linguistics, pages 161–170.
- Björne, Jari, Filip Ginter, Sampo Pyysalo, Jun'ichi Tsujii, and Tapio Salakoski (2010). "Scaling Up Biomedical Event Extraction to the Entire PubMed". In: *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing. BioNLP '10*. Stroudsburg, PA, USA: Association for Computational Linguistics, pages 28–36.
- Blake, Catherine (2010). "Beyond genes, proteins, and abstracts: Identifying scientific claims from full-text biomedical articles." In: *Biomedical Informatics* 43.2, pages 173–189.
- Bodenreider, Olivier (2004). "The Unified Medical Language System (UMLS): integrating biomedical terminology." In: *Nucleic Acids Research* 32.Database issue, pages D267–D270.
- Bos, Johan and Katja Markert (2005). "Recognising Textual Entailment with Robust Logical Inference." In: *MLCW*. Edited by Joaquin Quinonero Candela, Ido Dagan,

- Bernardo Magnini, and Florence d'Alch Buc. Volume 3944. Lecture Notes in Computer Science. Springer, pages 404–426.
- Boston Collaborative Drug Surveillance program (1974). "Regular aspirin intake and acute myocardial infarction." In: *BMJ* 1.5905, pages 440–3. ISSN: 0007-1447.
- Bowman, Samuel, Gabor Angeli, Christopher Potts, and Christopher Manning (2015a). "A large annotated corpus for learning natural language inference". In: *CoRR* abs/1508.05326.
- Bowman, Samuel R., Gabor Angeli, Christopher Potts, and Christopher D. Manning (2015b). "A large annotated corpus for learning natural language inference". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Brotons, Carlos, Robert Benamouzig, Krzysztof J. Filipiak, Volker Limmroth, and Claudio Borghi (2015). "A systematic review of aspirin in primary prevention: is it time for a new approach?" eng. In: *American Journal of Cardiovascular Drugs* 15.2, pages 113–133.
- Bundschus, Markus, Mathaeus Dejori, Martin Stetter, Volker Tresp, and Hans-Peter Kriegel (2008). "Extraction of semantic biomedical relations from text using conditional random fields". In: *BMC Bioinformatics* 9.1, page 207.
- Bunescu, Razvan, Ruifang Ge, Rohit J. Kate, Edward M. Marcotte, Raymond J. Mooney, Arun K. Ramani, and Yuk Wah Wong (2005). "Comparative experiments on learning information extractors for proteins and their interactions." In: *Artificial Intelligence in Medicine* 33.2, pages 139–155.
- Burchardt, Aljoscha and Anette Frank (2006). "Approximating Textual Entailment with LFG and FrameNet Frames". In: *Proceedings of the second PASCAL Recognizing Textual Entailment Workshop*, pages 92–97.
- Burchardt, Aljoscha, Nils Reiter, Stefan Thater, and Anette Frank (2007). "A Semantic Approach to Textual Entailment: System Evaluation and Task Analysis". In: *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*. RTE '07. Stroudsburg, PA, USA: Association for Computational Linguistics, pages 10–15.
- Burchardt, Aljoscha, Marco Pennacchiotti, Stefan Thater, and Manfred Pinkal (2009). "Assessing the Impact of Frame Semantics on Textual Entailment". In: *Natural Language Engineering* 15.4, pages 527–550.
- Burkman, R. T. (1981). "Association between intrauterine device and pelvic inflammatory disease." In: *Obstetrics & Gynecology* 57.3, pages 269–276.

- Cabrio, Elena and Serena Villata (2013). "Detecting Bipolar Semantic Relations among Natural Language Arguments with Textual Entailment: a Study". In: *Joint Symposium on Semantic Processing (JSSP-2013)*. Trento, Italy.
- Cabrio, Elena and Serena Villata (2014). "NoDE: A Benchmark of Natural Language Arguments". In: *Computational Models of Argument - Proceedings of COMMA 2014, Atholl Palace Hotel, Scottish Highlands, UK, September 9-12, 2014*, pages 449–450.
- Cahill, Aoife, Michael Burke, Josef Van Genabith, and Andy Way (2004). "Long-distance dependency resolution in automatically acquired wide-coverage PCFG-based LFG approximations". In: *In Proceedings of the 42nd Meeting of the ACL*, pages 320–327.
- Cayrol, Claudette and Marie-Christine Lagasquie-Schiex (2005). "On the Acceptability of Arguments in Bipolar Argumentation Frameworks." In: *ECSQARU*. Edited by Lluís Godo. Volume 3571. Lecture Notes in Computer Science. Springer, pages 378–389.
- Chapman, Brian E., Sean Lee, Hyunseok Peter Kang, and Wendy W. Chapman (2011). "Document-level classification of CT pulmonary angiography reports based on an extension of the ConText algorithm." In: *Biomedical Informatics* 44.5.
- Chapman, Wendy, Dieter Hilert, Sumithra Velupilai, Maria Kvist, Maria Skeppstedt, Brian Chapman, Mike Conway, Melissa Tharp, Danielle Mowery, and Louise Deleger (2013). "Extending the NegEx Lexicon for Multiple Languages". In: *Proceedings of the 14th World Congress on Medical and Health Informatics (MedInfo 2013)*, pages 677–681. URL: <http://www.ncbi.nlm.nih.gov/pubmed/23920642>. published.
- Chapman, Wendy W., Will Bridewell, Paul Hanbury, Gregory F. Cooper, and Bruce G. Buchanan (2001). "A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries". In: *J Biomed Inform* 2001, pages 34–301.
- Chierchia, Gennaro and Sally McConnell-Ginet (2000). *Meaning and Grammar An Introduction to Semantics*. The MIT Press.
- Chklovski, Timothy and Patrick Pantel (2004). "VerbOcean: Mining the Web for Fine-Grained Semantic Verb Relations". In: *Proceedings of EMNLP 2004*. Edited by Dekang Lin and Dekai Wu. Barcelona, Spain: Association for Computational Linguistics, pages 33–40.
- Christian, C. D. (1974). "Maternal deaths associated with an intrauterine device." In: *Obstetrics and Gynecology* 119.4, pages 441–444.
- Chung, Grace Y. (2009). "Sentence retrieval for abstracts of randomized controlled trials." In: *BMC Med Inform Decis Mak* 9, page 10.

- Claessen, K. and N. Sorensson (2003). "New Techniques that Improve MACE-style Finite Model Finding". In: *Proceedings of the CADE-19 Workshop: Model Computation - Principles, Algorithms, Applications*. Edited by P. Baumgartner and C. Fermueller.
- Collier, Nigel, Chikashi Nobata, and Jun ichi Tsujii (2000). "Extracting the Names of Genes and Gene Products with a Hidden Markov Model". In: pages 201–207.
- Collobert, Ronan, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa (2011). "Natural Language Processing (Almost) from Scratch". In: *Machine Learning Research* 12, pages 2493–2537.
- Condoravdi, Cleo, Dick Crouch, Valeria de Paiva, Reinhard Stolle, and Daniel G. Bobrow (2003). "Entailment, Intensionality and Text Understanding". In: *Proceedings of the HLT-NAACL 2003 Workshop on Text Meaning - Volume 9*. Stroudsburg, PA, USA: Association for Computational Linguistics, pages 38–45.
- Cramer, Koby and Yoram Singer (2003). "Ultraconservative Online Algorithms for Multiclass Problems". In: *Machine Learning Research* 3, pages 951–991.
- Craven, Mark and Johan Kumlien (1999). "Constructing Biological Knowledge Bases by Extracting Information from Text Sources". In: *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, pages 77–86.
- Dagan, Ido, Oren Glickman, and Bernardo Magnini (2005). "The PASCAL Recognising Textual Entailment Challenge". In: *Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment*.
- Dagangan, Ido and Fabio Massimo (2007). *A tutorial on textual entailment*.
- Dang, Hoa Trang and Karolina Owczarzak (2008). "Overview of the TAC 2008 update summarization task". In: *In TAC 2008 Workshop - Notebook papers and results*, pages 10–23.
- Dinu, Georgiana and Rui Wang (2009). "Inference rules for recognizing textual entailment". In: *In Proceedings of the IWCS*.
- Dung, Phan Minh (1995). "On the Acceptability of Arguments and Its Fundamental Role in Nonmonotonic Reasoning, Logic Programming and N-person Games". In: *Artificial Intelligence* 77.2, pages 321–357.
- DynaMed Plus. <http://www.dynamed.com/home/>. Accessed: 2016-03-01.
- Elliott, Julian H (2013). *Re: The automation of systematic reviews*. <http://www.bmj.com/content/346/bmj.f139/rr/625503>.
- Ely, J. W., J. A. Osheroff, P. N. Gorman, M. H. Ebell, M. L. Chambliss, E. A. Pifer, and P. Z. Stavri (2000). "A taxonomy of generic clinical questions: classification study." In: *BMJ* 321.7258, pages 429–432.

- 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*. EMNLP '11. Stroudsburg, PA, USA: Association for Computational Linguistics, pages 1535–1545.
- Freeman, James (2011). *Argument Structure: Representation and Theory*. Springer.
- Fukuda, K., A. Tamura, T. Tsunoda, and T. Takagi (1998). "Toward information extraction: identifying protein names from biological papers." In: *Pacific Symposium on Biocomputing*, pages 707–718.
- Fuster, V., B.B. Kelly, C.P.G.E.C.D.M.C.D. Countries, B.G. Health, and I. Medicine (2010). *Promoting Cardiovascular Health in the Developing World: A Critical Challenge to Achieve Global Health*. The National Academies Press. ISBN: 9780309147743.
- Gee, Ruth (1998). "The TIPSTER Text Program Overview". In: *Proceedings of a Workshop on Held at Baltimore, Maryland: October 13-15, 1998*. TIPSTER '98. Stroudsburg, PA, USA: Association for Computational Linguistics, pages 3–5.
- Giampiccolo, Danilo, Bernardo Magnini, Ido Dagan, Bill Dolan, Hoa Trang Dang, and Elena Cabrio (2008). "The fourth PASCAL recognizing textual entailment challenge". In:
- Gøtzsche, P. C. and O. Olsen (2000). "Is screening for breast cancer with mammography justifiable?" In: *Lancet* 355.9198, pages 129–134.
- Gough, David, Sandy Oliver, and James Thomas (2012). *An introduction to systematic reviews*. Sage Publications.
- Green, Nancy (2014). "Towards Creation of a Corpus for Argumentation Mining the Biomedical Genetics Research Literature". In: *Proceedings of the First Workshop on Argumentation Mining*. Baltimore, Maryland: Association for Computational Linguistics, pages 11–18.
- Green, Nancy (2015a). "Annotating Evidence-Based Argumentation in Biomedical Text". In: *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. Washington, USA.
- Green, Nancy (2015b). "Identifying Argumentation Schemes in Genetics Research Articles". In: *Proceedings of the 2nd Workshop on Argumentation Mining*. Denver, Colorado.
- Grishman, Ralph (2003). "Information Extraction. The Oxford Handbook of Computational Linguistics". In: edited by Ruslan Mitkov. Ruslan Mitkov, editor, Oxford University Press. Chapter 30.
- Grover, Claire, Ben Hachey, and Chris Korycinski (2003). "Summarising Legal Texts: Sentential Tense and Argumentative Roles". In: *Proceedings of the HLT-NAACL 03*

- on Text Summarization Workshop - Volume 5*. Stroudsburg, PA, USA: Association for Computational Linguistics, pages 33–40.
- Haghighi, Aria D., Andrew Y. Ng, and Christopher D. Manning (2005). “Robust textual inference via graph matching”. In: *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*. HLT '05. Stroudsburg, PA, USA: Association for Computational Linguistics, pages 387–394.
- Harabagiu, Sanda M., Andrew Hickl, and V. Finley Lacatusu (2006a). “Negation, Contrast and Contradiction in Text Processing.” In: *AAAI*. AAAI Press, pages 755–762.
- Harabagiu, Sanda M., Andrew Hickl, and V. Finley Lacatusu (2006b). “Negation, Contrast and Contradiction in Text Processing”. In: *National Conference on Artificial Intelligence*.
- Harabagiu, S.M. and D.I. Moldovan (1998). “Knowledge processing on an extended wordnet”. In: *WordNet-An Electronic Lexical Database*, pages 379–405.
- Hashimoto, Chikara, Kentaro Torisawa, Stijn De Saeger, Jong-Hoon Oh, and Jun’ichi Kazama (2012). “Excitatory or Inhibitory: A New Semantic Orientation Extracts Contradiction and Causality from the Web.” In: *EMNLP-CoNLL*. ACL, pages 619–630.
- Hatzivassiloglou, Vasileios and Kathleen R. McKeown (1997). “Predicting the Semantic Orientation of Adjectives”. In: *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*. ACL '98. Stroudsburg, PA, USA: Association for Computational Linguistics, pages 174–181.
- Hennekens, CH, LK Karlson, and Bernard Rosner (1978). “A case-control study of regular aspirin use and coronary deaths.” In: *Circulation* 58.1, pages 35–8.
- Hersh, William, Aaron Cohen, Phoebe Roberts, and Hari Rekapalli (2006). “TREC 2006 Genomics Track Overview”. In: *Proceedings of the Fifteenth Text REtrieval Conference, TREC 2006, Gaithersburg, Maryland, USA, November 14-17, 2006*.
- Hickl, Andrew, Jeremy Bensley, John Williams, Kirk Roberts, Bryan Rink, and Ying Shi (2006). “Recognizing textual entailment with lccs groundhog system”. In: *Proceedings of the Second PASCAL Challenges Workshop*.
- Higgins, J. P. T. and S. Green, editors (2008). *Cochrane Handbook for Systematic Reviews of Interventions*. 5th edition. The Cochrane Collaboration.
- Hinton, G. E., A. D. Brown, and Queen Square London (2001). *Training Many Small Hidden Markov Models*.

- Hirohata, Kenji, Naoaki Okazaki, Sophia Ananiadou, and Mitsuru Ishizuka (2008). "Identifying sections in scientific abstracts using conditional random fields". In: *In Proc. of the IJCNLP 2008*.
- Hirschman, L. and Rob Gaizauskas (2001). "Natural Language Question Answering: The View from Here". In: *Nat. Lang. Eng.* 7.4, pages 275–300.
- Hobbs, Jerry (2002). "Information Extraction from Biomedical Text". In: *Biomedical Informatics* 35.4, pages 260–264. ISSN: 1532-0464.
- Horwitz, R. I. (1987). "Complexity and contradiction in clinical trial research." In: *Am J Med* 82.3, pages 498–510.
- Hou, Yongshuai, Cong Tan, Xiaolong Wang, Yaoyun Zhang, Jun Xu, and Qingcai Chen (2015). "HITSZ-ICRC: Exploiting Classification Approach for Answer Selection in Community Question Answering". In: *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. Denver, Colorado: Association for Computational Linguistics, pages 196–202.
- Huang, Xiaoli, Jimmy Lin, and Dina Demner-Fushman (2006). "Evaluation of PICO as a knowledge representation for clinical questions." In: *The American Medical Informatics Association Annual Symposium Proceedings*, pages 359–363.
- Hyland, Ken (1995). "The Author in the Text: Hedging Scientific Writing". In: *Hong kong Papers in Linguistics and Language Teaching* 18, pages 33–42.
- Iftene, Adrian and Alexandra Balahur-Dobrescu (2007). "Hypothesis Transformation and Semantic Variability Rules Used in Recognizing Textual Entailment". In: *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*. RTE '07. Stroudsburg, PA, USA: Association for Computational Linguistics, pages 125–130.
- Iida, Ryu and Massimo Poesio (2011). "A Cross-lingual ILP Solution to Zero Anaphora Resolution". In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*. HLT '11. Stroudsburg, PA, USA: Association for Computational Linguistics, pages 804–813.
- Ioannidis, John (2005). "Contradicted and initially stronger effects in highly cited clinical research." In: *American Medical Association* 294.2, pages 218–228.
- Ioannidis, John and Thomas A. Trikalinos (2005). "Early extreme contradictory estimates may appear in published research: The Proteus phenomenon in molecular genetics research and randomized trials". In: *Journal of Clinical Epidemiology* 58.6, pages 543–549.

- Jimeno Yepes, Antonio, James Mork, and Alan Aronson (2013). "Using the Argumentative Structure of Scientific Literature to Improve Information Access". In: *Proceedings of the 2013 Workshop on Biomedical Natural Language Processing*. Sofia, Bulgaria: Association for Computational Linguistics, pages 102–110.
- Jimeno-Yepes, Antonio, Caitlin Sticco, James Mork, and Alan Aronson (2013). "GeneRIF indexing: sentence selection based on machine learning." In: *BMC Bioinformatics* 14, page 171.
- Jindal, Nitin and Bing Liu (2006). "Identifying comparative sentences in text documents". In: *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. SIGIR '06. New York, NY, USA: ACM, pages 244–251.
- Kamp, Hans and Uwe Reyle (1993). *From Discourse to Logic: Introduction to Model-theoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*. Volume 42. Studies in Linguistics and Philosophy. Dordrecht: Kluwer.
- Kawahara, Daisuke, Kentaro Inui, and Sadao Kurohashi (2010). "Identifying Contradictory and Contrastive Relations Between Statements to Outline Web Information on a Given Topic". In: *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*. Stroudsburg, PA, USA: Association for Computational Linguistics, pages 534–542.
- Kilicoglu, Halil, Dongwook Shin, Marcelo Fiszman, Graciela Rosembat, and Thomas C. Rindflesch (2012). "SemMedDB: a PubMed-scale repository of biomedical semantic predications." In: *Bioinformatics* 28.23, pages 3158–3160.
- Kim, Jin-Dong, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun'ichi Tsujii (2009). "Overview of BioNLP'09 Shared Task on Event Extraction". In: *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*. Association for Computational Linguistics. Boulder, Colorado: Association for Computational Linguistics, 1–9.
- Kim, Seonho, Juntae Yoon, and Jihoon Yang (2008). "Kernel Approaches for Genic Interaction Extraction". In: *Bioinformatics* 24.1, pages 118–126.
- Koike, Asako, Yoshiki Niwa, and Toshihisa Takagi (2005). "Automatic extraction of gene/protein biological functions from biomedical text". In: *Bioinformatics* 21.7, pages 1227–1236.
- Kronmal, R. A., C. W. Whitney, and S. D. Mumford (1991). "The intrauterine device and pelvic inflammatory disease: the Women's Health Study reanalyzed." eng. In: *Clinical Epidemiology* 44.2, pages 109–122.

- Lafferty, John D., Andrew McCallum, and Fernando C. N. Pereira (2001). "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data". In: *Proceedings of the Eighteenth International Conference on Machine Learning*. ICML '01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., pages 282–289.
- Leacock, Claudia, George A. Miller, and Martin Chodorow (1998). "Using Corpus Statistics and WordNet Relations for Sense Identification". In: *Computational Linguistics* 24.1, pages 147–165.
- Levenshtein, VI (1966). "Binary Codes Capable of Correcting Deletions, Insertions and Reversals". In: *Soviet Physics Doklady* 10, page 707.
- Lewis, S. and M. Clarke (2001). "Forest plots: trying to see the wood and the trees". In: 322.7300, pages 1479–80+.
- Liakata, Maria (2010). "Zones of Conceptualisation in Scientific Papers: A Window to Negative and Speculative Statements". In: *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*. NeSp-NLP '10. Stroudsburg, PA, USA: Association for Computational Linguistics, pages 1–4.
- Liakata, Maria, Paul Thompson, Anita de Waard, Raheel Nawaz, Henk Pander Maat, and Sophia Ananiadou (2012). "A Three-way Perspective on Scientific Discourse Annotation for Knowledge Extraction". In: *Proceedings of the Workshop on Detecting Structure in Scholarly Discourse*. ACL '12. Stroudsburg, PA, USA: Association for Computational Linguistics, pages 37–46.
- Light, Marc, Xin Ying Qiu, and Padmini Srinivasan (2004). "The Language of Bioscience: Facts, Speculations, and Statements In Between". In: *HLT-NAACL 2004 Workshop: BioLINK 2004, Linking Biological Literature, Ontologies and Databases*. Edited by Lynette Hirschman and James Pustejovsky. Boston, Massachusetts, USA: Association for Computational Linguistics, pages 17–24.
- Lin, Dekang (1998a). "An Information-Theoretic Definition of Similarity". In: *Proceedings of the Fifteenth International Conference on Machine Learning*. ICML '98. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., pages 296–304.
- Lin, Dekang (1998b). "Dependency-based Evaluation of MINIPAR". In: *Proceedings of Workshop on the Evaluation of Parsing Systems*. Granada.
- Lin, Dekang and Patrick Pantel (2001). "DIRT: Discovery of Inference Rules from Text". In: *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'01)*. New York, NY, USA: ACM Press, pages 323–328.

- Lin, Ryan, Hong-Jei Dai, Yue-Yang Bow, Min-Yuh Day, Richard Tzong-Han Tsai, and Wen-Lian Hsu (2008a). "Result identification for biomedical abstracts using Conditional Random Fields". In: *Information Reuse and Integration, 2008. IRI 2008. IEEE International Conference on*, pages 122–126.
- Lin, Ryan, Hong-Jie Dai, Yue-Yang Bow, Justing Lian-Te Chiu, and Richardg Tzon-Han Tsai (2009). "Using Conditional Random Fields for Result Identification in Biomedical Abstracts". In: *Integrated Computer Aided Engineering* 16.4, pages 339–352.
- Lin, Ryan T.K., Justin Liang-Te Chiu, Hong-Jei Dai, Min-Yuh Day, Richard Tzong-Han Tsai, and Wen-Lian Hsu (2008b). "Biological question answering with syntactic and semantic feature matching and an improved mean reciprocal ranking measurement". In: *Information Reuse and Integration, 2008. IRI 2008. IEEE International Conference on*, pages 184–189.
- Lippi, Marco and Paolo Torroni (2015). "Argumentation mining: a machine learning perspective". In: *International Workshop on Theory and Applications of Formal Argument (TAFA)*. Buenos Aires, Argentina.
- Lodhi, Huma, Craig Saunders, John Shawe-Taylor, Nello Cristianini, and Chris Watkins (2002). "Text Classification Using String Kernels". In: *Machine Learning Research* 2, pages 419–444.
- MacCartney, Bill, Trond Grenager, Marie-Catherine de Marneffe, Daniel Cer, and Christopher D. Manning (2006). "Learning to recognize features of valid textual entailments". In: *Proceedings of the North American Association of Computational Linguistics*. The Stanford Natural Language Processing Group.
- Malakasiotis, Prodromos (2009). "Paraphrase Recognition Using Machine Learning to Combine Similarity Measures". In: *Proceedings of the ACL-IJCNLP 2009 Student Research Workshop*. ACLstudent '09. Stroudsburg, PA, USA: Association for Computational Linguistics, pages 27–35.
- Malakasiotis, Prodromos and Ion Androutsopoulos (2007). "Learning Textual Entailment Using SVMs and String Similarity Measures". In: *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*. RTE '07. Stroudsburg, PA, USA: Association for Computational Linguistics, pages 42–47.
- Mann, WilliamC. and SandraA. Thompson (1987). "Rhetorical Structure Theory: Description and Construction of Text Structures". In: *Natural Language Generation*. Edited by Gerard Kempen. Volume 135. NATO ASI Series. Springer Netherlands, pages 85–95.
- Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze (2008). *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press.

- Marneffe, Marie-Catherine de and Christopher Manning (2008). "The Stanford typed dependencies representation". In: *Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation*. CrossParser '08. Stroudsburg, PA, USA: Association for Computational Linguistics, pages 1–8.
- Marneffe, Marie-Catherine De, Anna Rafferty, and Christopher Manning (2008). "Finding contradictions in text". In: *In ACL 2008*.
- Mausam, Michael Schmitz, Robert Bart, Stephen Soderland, and Oren Etzioni (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*. EMNLP-CoNLL '12. Stroudsburg, PA, USA: Association for Computational Linguistics, pages 523–534.
- Mayberry, Katherine J. (2009). *Everyday Arguments. A guide to Writing and Reading Effective Arguments*. Houghton Mifflin Company.
- Mayor, S. (1999). "Swedish study questions mammography screening programmes." In: *BMJ* 318.7184, page 621.
- McClosky, David, Mihai Surdeanu, and Christopher D. Manning (2011). "Event Extraction As Dependency Parsing". In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*. HLT '11. Stroudsburg, PA, USA: Association for Computational Linguistics, pages 1626–1635.
- McCray, A. T., A. Burgun, and O. Bodenreider (2001). "Aggregating UMLS semantic types for reducing conceptual complexity." In: *Studies in Health Technology and Informatics* 84.Pt 1, pages 216–220.
- McCray, A. T., S. Srinivasan, and A. C. Browne (1994). "Lexical methods for managing variation in biomedical terminologies." In: *Symposium on Computer Applications in Medical Care*, pages 235–239.
- McInnes, Bridget., Ted Pedersen, and Serguei Pakhomov (2009). "UMLS-Interface and UMLS-Similarity : open source software for measuring paths and semantic similarity." In: *The American Medical Informatics Association Annual Symposium Proceedings 2009*, pages 431–435.
- Mehdad, Yashar (2009). "Automatic Cost Estimation for Tree Edit Distance Using Particle Swarm Optimization". In: *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*. ACLShort '09. Stroudsburg, PA, USA: Association for Computational Linguistics, pages 289–292.
- Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean (2013). "Efficient Estimation of Word Representations in Vector Space". In: *CoRR* abs/1301.3781.

- Miwa, Makoto, Rune Sætre, Yusuke Miyao, and Jun'ichi Tsujii (2009). "A Rich Feature Vector for Protein-protein Interaction Extraction from Multiple Corpora". In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*. EMNLP '09. Stroudsburg, PA, USA: Association for Computational Linguistics, pages 121–130.
- Mizuta, Yoko and Nigel Collier (2004). "Zone Identification in Biology Articles As a Basis for Information Extraction". In: *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and Its Applications*. JNLPBA '04. Stroudsburg, PA, USA: Association for Computational Linguistics, pages 29–35.
- Mochales, Raquel and Marie-Francine Moens (2011). "Argumentation mining". In: *Artificial Intelligence and Law 19.1*, pages 1–22.
- Moens, Marie-Francine (2006). *Information Extraction: Algorithms and Prospects in a Retrieval Context (The Information Retrieval Series)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc. ISBN: 1402049870.
- Moens, Marie-Francine, Erik Boiy, Raquel Mochales Palau, and Chris Reed (2007). "Automatic Detection of Arguments in Legal Texts". In: *Proceedings of the 11th International Conference on Artificial Intelligence and Law*. ICAIL '07. New York, NY, USA: ACM, pages 225–230.
- Mohammad, Saif, Bonnie J. Dorr, and Graeme Hirst (2008). "Computing Word-Pair Antonymy." In: *EMNLP*. ACL, pages 982–991.
- Mollá, Diego and Ben Hutchinson (2003). "Intrinsic Versus Extrinsic Evaluations of Parsing Systems". In: *Proceedings of the EACL 2003 Workshop on Evaluation Initiatives in Natural Language Processing: Are Evaluation Methods, Metrics and Resources Reusable?* Evalinitatives '03. Budapest, Hungary: Association for Computational Linguistics, pages 43–50.
- Mooney, Raymond J. and Razvan C. Bunescu (2006). "Subsequence Kernels for Relation Extraction". In: *Advances in Neural Information Processing Systems 18*. Edited by Y. Weiss, B. Schölkopf, and J.C. Platt. MIT Press, pages 171–178.
- Mumford, SD and E Kessel (1992). "Was the Dalkon Shield a safe and effective intrauterine device? The conflict between case-control and clinical trial study findings". In: *Fertility and sterility* 57.6, 11511176. ISSN: 0015-0282.
- Mutalik, Pradeep G., Aniruddha Deshpande, and Prakash M. Nadkarni (2001). "Use of General-purpose Negation Detection to Augment Concept Indexing of Medical Documents". In: *The American Medical Informatics Association* 8.6, pages 598–609. ISSN: 1067-5027.

- Nakov, Preslav, Lluís Màrquez, Walid Magdy, Alessandro Moschitti, Jim Glass, and Bilal Randeree (2015). "SemEval-2015 Task 3: Answer Selection in Community Question Answering". In: *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. Denver, Colorado: Association for Computational Linguistics, pages 269–281.
- Neustein, Amy, editor (2014s). *Text Mining of Web-Based Medical Content*. Walter de Gruyter.
- Nicosia, Massimo, Simone Filice, Alberto Barrón-Cedeño, Iman Saleh, Hamdy Mubarak, Wei Gao, Preslav Nakov, Giovanni Da San Martino, Alessandro Moschitti, Kareem Darwish, Lluís Màrquez, Shafiq Joty, and Walid Magdy (2015). "QCRI: Answer Selection for Community Question Answering - Experiments for Arabic and English". In: *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. Denver, Colorado: Association for Computational Linguistics, pages 203–209.
- Niu, Yun, Graeme Hirst, Gregory McArthur, and Patricia Rodriguez-Gianolli (2003). "Answering Clinical Questions with Role Identification". In: *Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine - Volume 13. BioMed '03*. Stroudsburg, PA, USA: Association for Computational Linguistics, pages 73–80.
- Niu, Yun, Xiaodan Zhu, Jianhua Li, and Graeme Hirst (2005). "Analysis of polarity information in medical text". In: *Proceedings of the American Medical Informatics Association*.
- Nyström, L., S. Wall, L.E. Rutqvist, A. Lindgren, M. Lindqvist, S. Rydén, J. Andersson, N. Bjurström, G. Fagerberg, J. Frisell, L. Taber, and L.-G. Larsson (1993). "Breast cancer screening with mammography: overview of Swedish randomised trials". In: *The Lancet* 341.8851. Originally published as Volume 1, Issue 8851, pages 973–978.
- Ohta, T., Y. Tateisi, and J.D. Kim (2002). "The GENIA corpus: An annotated research abstract corpus in molecular biology domain". In: *the Human Language Technology Conference*.
- Okazaki, Naoaki (2007). *CRFsuite: a fast implementation of Conditional Random Fields (CRFs)*. <http://www.chokkan.org/software/crfsuite/>.
- Okazaki, Naoaki, Sophia Ananiadou, and Jun'ichi Tsujii (2010). "Building a high-quality sense inventory for improved abbreviation disambiguation". In: *Bioinformatics* 26.9, pages 1246–1253.
- O'Mara-Eves, Alison, James Thomas, John McNaught, Makoto Miwa, and Sophia Ananiadou (2015). "Using text mining for study identification in systematic reviews: a systematic review of current approaches". In: *Systematic Reviews* 4.1.

- Ono, Toshihide, Haretsugu Hishigaki, Akira Tanigami, and Toshihisa Takagi (2001). "Automated extraction of information on protein-protein interactions from the biological literature." In: *Bioinformatics* 17.1, pages 155–161.
- Otani, S. and Y. Tomiura (2014). "Extraction of Key Expressions Indicating the Important Sentence from Article Abstracts". In: *Advanced Applied Informatics (IIAIAAI), 2014 IIAI 3rd International Conference on*, pages 216–219.
- Ounis, I., G. Amati, Plachouras V., B. He, C. Macdonald, and Johnson (2005). "Terrier Information Retrieval Platform". In: *Proceedings of the 27th European Conference on IR Research (ECIR 2005)*. Volume 3408. Lecture Notes in Computer Science. Springer, pages 517–519. ISBN: 3-540-25295-9.
- Oxford-Dictionary (2003). *Oxford English Dictionary Online, 2nd edition*. <http://www.oed.com/>.
- Paganini-Hill, A., A. Chao, R. K. Ross, and B. E. Henderson (1989). "Aspirin use and chronic diseases: a cohort study of the elderly." In: *BMJ* 299.6710, pages 1247–1250. ISSN: 0959-8138.
- Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan (2002). "Thumbs Up? Sentiment Classification Using Machine Learning Techniques". In: *Proceedings of EMNLP*, pages 79–86.
- Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu (2002). "BLEU: A Method for Automatic Evaluation of Machine Translation". In: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. ACL '02. Stroudsburg, PA, USA: Association for Computational Linguistics, pages 311–318.
- Park, Dae Hoon and Catherine Blake (2012). "Identifying comparative claim sentences in full-text scientific articles". In: *Proceedings of the Workshop on Detecting Structure in Scholarly Discourse*. ACL '12. Stroudsburg, PA, USA: Association for Computational Linguistics, pages 1–9.
- Pedersen, Ted and Siddharth Patwardhan (2004). "Wordnet::similarity - Measuring the Relatedness of Concepts". In: pages 1024–1025.
- Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay (2011). "Scikit-learn: Machine Learning in Python". In: *Machine Learning Research* 12, pages 2825–2830.
- Peldszus, Andreas and Manfred Stede (2013). "From Argument Diagrams to Argumentation Mining in Texts: A Survey". In: *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)* 7.1, pages 1–31.

- Pham, Minh Quang Nhat, Minh Le Nguyen, and Akira Shimazu (2013). *Using Shallow Semantic Parsing and Relation Extraction for Finding Contradiction in Text*. Technical Report, pages 1–10.
- Proux, Denys, Francois Rechenmann, Laurent Julliard, and Key Words (2000). "A Pragmatic Information Extraction Strategy for Gathering Data on Genetic Interactions". In: *In Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, pages 279–285.
- Punyakanok, V., D. Roth, and W. Yih (2008). "The Importance of Syntactic Parsing and Inference in Semantic Role Labeling". In: *Computational Linguistics* 34.2.
- Ramani, Arun, Razvan Bunescu, Raymond Mooney, and Edward Marcotte (2005). "Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome". In: *Genome Biology* 6.5, R40.
- Reed C., Raquel Mochales Palau Glenn Rowe and Marie-Francine Moens (2008). "Language Resources for Studying Argument". In: *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco: European Language Resources Association (ELRA).
- Resnik, Philip (1999). *Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language*.
- Riazanov, Alexandre and Andrei Voronkov (2002). "The design and implementation of VAMPIRE". In: *Artificial Intelligence Communications* 15.2,3, pages 91–110.
- Riedel, Sebastian and Andrew McCallum (2011). "Robust Biomedical Event Extraction with Dual Decomposition and Minimal Domain Adaptation". In: *Proceedings of the BioNLP Shared Task 2011 Workshop*. BioNLP Shared Task '11. Stroudsburg, PA, USA: Association for Computational Linguistics, pages 46–50.
- Rinaldi, Fabio, James Dowdall, Kaarel Kaljurand, Michael Hess, and Diego Mollá (2003). "Exploiting Paraphrases in a Question Answering System". In: *Proceedings of the Second International Workshop on Paraphrasing - Volume 16*. PARAPHRASE '03. Sapporo, Japan: Association for Computational Linguistics, pages 25–32.
- Rindflesch, Thomas C. and Marcelo Fiszman (2003). "The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text." In: *Biomedical Informatics* 36.6, pages 462–477.
- Rindflesch, Thomas C., Lorraine Tanabe, John N. Weinstein, and Lawrence Hunter (2000). "Edgar: Extraction of drugs, genes and relations from the biomedical literature". In: pages 517–528.

- Rindflesch, Thomas C., Bisharah Libbus, Dimitar Hristovski, Alan R. Aronson, and Halil Kilicoglu (2003). "Semantic relations asserting the etiology of genetic diseases." In: *The American Medical Informatics Association Annual Symposium Proceedings*, pages 554–558.
- Ripple, Anna M., James G. Mork, John M. Rozier, and Lou S. Knecht (2012). "Structured Abstracts in MEDLINE: Twenty-Five Years Later". In:
- Ritter, Alan, Doug Downey, Stephen Soderland, and Oren Etzioni (2008). "It's a contradiction—no, it's not: a case study using functional relations". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing. EMNLP '08*. Stroudsburg, PA, USA: Association for Computational Linguistics, pages 11–20.
- Ruch, Patrick, Clia Boyer, Christine Chichester, Imad Tbahriti, Antoine Geissbhlér, Paul Fabry, Julien Gobeill, Violaine Pillet, Dietrich Rebholz-Schuhmann, Christian Lovis, and Anne-Lise Veuthey (2007). "Using argumentation to extract key sentences from biomedical abstracts." In: *Medical Informatics 76.2-3*, pages 195–200.
- Ryan, Matthew S. and Graham R. Nudd (1993). *The Viterbi Algorithm*. Technical report. Coventry, UK.
- Sanchez-Graillet, Olivia and Massimo Poesio (2007). "Discovering Contradicting Protein-protein Interactions in Text". In: *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing. BioNLP '07*. Prague, Czech Republic: Association for Computational Linguistics, pages 195–196.
- Sarafraz, Farzaneh (2011). "Finding Conflicting Statements in the Biomedical Literature". PhD thesis. University of Manchester.
- Schwartz, A.S. and M.A. Hearst (2003). "A Simple Algorithm for Identifying Abbreviation Definitions in Biomedical Text". In: *Proceedings of Pacific Symposium on Biocomputing*. Volume 4, pages 451–462.
- Shatkay, Hagit and M.Mark Craven (2012). *Mining the Biomedical Literature*. The MIT Press.
- Smith, L., T. Rindflesch, and W. J. Wilbur (2004). "MedPost: A Part-of-Speech Tagger for Biomedical Text." In: *Bioinformatics* 20.14, pages 2320–2321.
- Somasundaran, Swapna and Janyce Wiebe (2010). "Recognizing Stances in Ideological On-line Debates". In: *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text. CAAGET '10*. Stroudsburg, PA, USA: Association for Computational Linguistics, pages 116–124.
- Stab, Christian and Iryna Gurevych (2014). "Annotating Argument Components and Relations in Persuasive Essays". In: *Proceedings of the 25th International Conference on Computational Linguistics (COLING 2014)*. Edited by Junichi Tsujii and Jan Hajic.

- Dublin, Ireland: Dublin City University and Association for Computational Linguistics, pages 1501–1510.
- Stevenson, M., Y. Guo, A. Alamri, and R. Gaizauskas (2009). “Disambiguation of Biomedical Abbreviations”. In: *Proceedings of the BioNLP 2009 Workshop*. Boulder, Colorado: Association for Computational Linguistics, pages 71–79.
- Strax, Philip, Louis Venet, Sam Shapiro, and Stanley Gross (1967). “Mammography and clinical examination in mass screening for cancer of the breast”. In: *Cancer* 20.12, pages 2184–2188.
- Sutton, Charles and Andrew McCallum (2006). “Introduction to Conditional Random Fields for Relational Learning”. In: edited by Lise Getoor and Ben Taskar. MIT Press.
- Szarvas, György, Veronika Vincze, Richárd Farkas, and János Csirik (2008). “The BioScope Corpus: Annotation for Negation, Uncertainty and Their Scope in Biomedical Texts”. In: *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*. BioNLP '08. Stroudsburg, PA, USA: Association for Computational Linguistics, pages 38–45.
- Taboada, Maite, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede (2011). “Lexicon-based Methods for Sentiment Analysis”. In: *Computational Linguistics* 37.2, pages 267–307.
- Takahashi, Kouji, Asako Koike, and Toshihisa Takagi (2004). “Question answering system in biomedical domain”. In: *Proceedings of the 15th International Conference on Genome Informatics*, pages 161–162.
- Teufel, Simone (2010). *The Structure of Scientific Articles: Applications to Citation Indexing and Summarization*. CSLI Publications.
- Teufel, Simone and Marc Moens (2002). “Summarizing Scientific Articles: Experiments with Relevance and Rhetorical Status”. In: *Computational Linguistics* 28.4, pages 409–445.
- Teufel, Simone, Advait Siddharthan, and Colin Batchelor (2009). “Towards Discipline-independent Argumentative Zoning: Evidence from Chemistry and Computational Linguistics”. In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3 - Volume 3*. EMNLP '09. Stroudsburg, PA, USA: Association for Computational Linguistics, pages 1493–1502.
- Thomas, James, John McNaught, and Sophia Ananiadou (2011). “Applications of text mining within systematic reviews”. In: *Research Synthesis Methods* 2.1, pages 1–14.
- Toulmin, Stephen E. (2003). *The Uses of Argument*. Cambridge University Press.

- Tran, Quan Hung, Vu Tran, Tu Vu, Minh Nguyen, and Son Bao Pham (2015). "JAIST: Combining multiple features for Answer Selection in Community Question Answering". In: *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. Denver, Colorado: Association for Computational Linguistics, pages 215–219.
- Tsafnat, Guy, Adam Dunn, Paul Glasziou, and Enrico Coiera (2013). "The automation of systematic reviews". In: *BMJ* 346.
- Turney, Peter D. (2002). "Thumbs Up or Thumbs Down?: Semantic Orientation Applied to Unsupervised Classification of Reviews". In: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. ACL '02. Stroudsburg, PA, USA: Association for Computational Linguistics, pages 417–424.
- UpToDate. <http://www.uptodate.com/home>. Accessed: 2016-03-01.
- Usami, Y., H.C. Cho, N. Okazaki, and J. Tsujii (2011). "Automatic acquisition of huge training data for bio-medical named entity recognition". In: *ACL HLT 2011*, page 65.
- Vapnik, Vladimir (1982). *Estimation of Dependences Based on Empirical Data: Springer Series in Statistics (Springer Series in Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc.
- Viechtbauer, Wolfgang (2010). "Conducting meta-analyses in R with the metafor package". In: *Statistical Software* 36.3, pages 1–48.
- Walton, Douglas (2009). "Argumentation Theory: A Very Short Introduction". In: *Argumentation in Artificial Intelligence*. Edited by Guillermo Simari and Iyad Rahwan. Springer US, pages 1–22.
- Wan, Stephen, Mark Dras, Robert Dale, and Cecile Paris (2006). "Using Dependency-based Features to Take the "Para-farce" out of Paraphrase". In: *Australasian Language Technology Workshop 2006 (ALTW 2006)*, pages 131–138.
- Wang, Rui and Günter Neumann (2007). "Recognizing Textual Entailment Using Sentence Similarity Based on Dependency Tree Skeletons". In: *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*. RTE '07. Stroudsburg, PA, USA: Association for Computational Linguistics, pages 36–41.
- Wonnacott, Thomas H. and Ronald J. Wonnacott (1990). *Introductory Statistics*. Fifth Edition. John Wiley and Sons.
- Wu, Fei and Daniel S. Weld (2010). "Open Information Extraction Using Wikipedia". In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. ACL '10. Stroudsburg, PA, USA: Association for Computational Linguistics, pages 118–127.

- Yates, Alexander, Michael Cafarella, Michele Banko, Oren Etzioni, Matthew Broadhead, and Stephen Soderland (2007). "TextRunner: open information extraction on the web". In: *NAACL '07*. Morristown, NJ, USA: Association for Computational Linguistics, pages 25–26.
- Yu, Hong and Vasileios Hatzivassiloglou (2003). "Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences". In: *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*. EMNLP '03. Stroudsburg, PA, USA: Association for Computational Linguistics, pages 129–136.
- Yu, Hong, Minsuk Lee, David Kaufman, John Ely, Jerome A. Osheroff, George Hripisak, and James Cimino (2007). "Development, Implementation, and a Cognitive Evaluation of a Definitional Question Answering System for Physicians". In: *Biomedical Informatics* 40.3, pages 236–251.
- Zhang, K. and D. Shasha (1989). "Simple Fast Algorithms for the Editing Distance Between Trees and Related Problems". In: *SIAM Journal on Computing* 18.6, pages 1245–1262.
- Zhou, Deyu, Dayou Zhong, and Yulan He (2014). "Biomedical relation extraction: from binary to complex." In: *Computational and Mathematical Methods* 2014, 298473:1–298473:18.
- Zhou, Guangyou, Li Cai, Jun Zhao, and Kang Liu (2011). "Phrase-based Translation Model for Question Retrieval in Community Question Answer Archives". In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*. HLT '11. Stroudsburg, PA, USA: Association for Computational Linguistics, pages 653–662.