Domain-Focused Summarization of Polarized Debates

Nattapong Sanchan

A thesis submitted in partial fulfilment of the requirements for the degree of
Doctor of Philosophy

The University of Sheffield
Faculty of Engineering
Department of Computer Science

May, 2018

ABSTRACT

Due to the exponential growth of Internet use, textual content is increasingly published in online media. In everyday, more and more news content, blog posts, and scientific articles are published to the online volumes and thus open doors for the text summarization research community to conduct research on those areas. Whilst there are freely accessible repositories for such content, online debates which have recently become popular have remained largely unexplored. This thesis addresses the challenge in applying text summarization to summarize online debates. We view that the task of summarizing online debates should not only focus on summarization techniques but also should look further on presenting the summaries into the formats favored by users.

In this thesis, we present how a summarization system is developed to generate online debate summaries in accordance with a designed output, called the *Combination 2*. It is the combination of two summaries. The primary objective of the first summary, Chart Summary, is to visualize the debate summary as a bar chart in high-level view. The chart consists of the bars conveying clusters of the salient sentences, labels showing short descriptions of the bars, and numbers of salient sentences conversed in the two opposing sides. The other part, Side-By-Side Summary, linked to the Chart Summary, shows a more detailed summary of an online debate related to a bar clicked by a user. The development of the summarization system is divided into three processes.

In the first process, we create a gold standard dataset of online debates. The dataset contains a collection of debate comments that have been subjectively annotated by 5 judgments. We develop a summarization system with key features to help identify salient sentences in the comments. The sentences selected by the system are evaluated against the annotation results. We found that the system performance outperforms the baseline.

The second process begins with the generation of Chart Summary from the salient sentences selected by the system. We propose a framework with two branches where each branch presents either a term-based clustering and the term-based labeling method or X-means based clustering and the MI labeling strategy. Our evaluation results indicate that the X-means clustering approach is a better alternative for clustering.

In the last process, we view the generation of Side-By-Side Summary as a contradiction detection task. We create two debate entailment datasets derived from the two clustering approaches and annotate them with the *Contradiction* and *Non-Contradiction* relations. We develop a classifier and investigate combinations of features that maximize the F1 scores. Based on the proposed features, we discovered that the combinations of at least two features to the maximum of eight features yield good results.

ACKNOWLEDGMENTS

I wish to express my deepest gratitude to Prof. Kalina Bontcheva and Dr. Ahmet Aker for their patience, insightful guidance, the devotion of time and energy, and for seeing every aspect of this thesis through from beginning to end. Without their supervision, this thesis would not have been possible.

I am also profoundly grateful to Dr. Josiah Wang for tips and techniques on thesis writing, thinking in research perspectives, and for other intuitive guidance on a user study and machine learning problems.

I would especially like to thank Mr. Mark Tice for programming questions I had. He always gave me feedback and useful techniques to achieve my programming tasks. My sincere thanks are also dedicated to dear classmates and colleagues at The University of Sheffield for the assistance and the guidance on research work, providing facilities, and giving advice on mathematical problems.

Heartfelt thanks also go to friends, brothers, and sisters in Sheffield, Leeds, and Surrey for sharing happiness during my Ph.D. and giving encouragement when the work seemed too difficult to be completed. I would have probably given it up but the power of friendship helped me through the hard days.

I gratefully acknowledge the funding source, Bangkok University, that brought me to Sheffield to meet all the amazing people, sparkled my researching experience, and made this long journey possible.

Finally, with boundless love and appreciation, I thank my family who have always inspired and endlessly supported me throughout my life.

DECLARATION

I hereby declare that this thesis is of my own composition, and that it contains no material previously submitted for the award of any other degree. The work reported in this thesis has been executed by myself, except where due acknowledgment is made in the text.

Nattapong Sanchan

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

SSSD. Salient Sentence Selection Dataset. This set of data is used in the Automatic Salient Sentence Selection process for training and evaluating tasks. In this dataset, salient sentences in each comment were manually selected based on a compression rate of 20%.

DEDT. Debate Entailment Dataset from the Term-based clustering approach. This is a dataset used in the contradiction detection task. It is created on the results derived from the term-based clustering approach.

DEDX. Debate Entailment Dataset from the X-means clustering approach. This is a dataset used in the contradiction detection task. It is created on the results derived from the X-means clustering approach.

GLOSSARY

Agree side. An opposing side that agrees with or supports the argument in an online debate.

Annotator. *Annotator* is noun form of *annotate.* According to the definition given by the English Oxford Dictionary (EOD), *annotator* refers to “add note (a text or diagram) giving explanation or comment” (Definition of annotate in English: Annotate, 2014). In this thesis, the *annotator* is used as a generic term for the individuals who assign, annotate, or judge their perspective on a given context. For instance, this term is used in a sentence: an annotator was asked to create a reference summary for a given task.

Debate comment. A debate comment refers to a post that answers to a debate. A debate comment reflects users’ opinions towards debates or argues to comments of the opposing side. A shorter term, **comment**, is also used interchangeably in this thesis.

Debate. The definition of debate given by EOD is as “an argument about a particular subject, especially one in which many people have involved: the national debate on abortion” (Oxford, 2014a).

Debate title. This term is used to represent the argumentation of a debate (i.e. “Does global warming exist?”).

Disagree side. An opposing side that disagrees with the argument in an online debate.

Opposing sides. According to EOD, a term *opposing* is defined as “in conflict or competition with someone or something” (Oxford, 2014b). The term *opposing side* is used by Somasundaran and Wiebe (2009, 2010) to represent a side of debates that contain different perspective on a topic. An example of using this word in a sentence is “while debating, participants often refer to and acknowledge the viewpoints of the opposing side” (Somasundaran and Wiebe, 2010). Moreover, the term *opposing sides* is used to refer to more than one opposing side.

Rebuttals. Rebuttal is defined by English Oxford Dictionary as an “instance of rebutting evidence or an accusation.” (Oxford, 2014c). Therefore, according to this definition, rebuttal is appropriate when opposing sides have a contradictory opinion toward a topic. In other words, rebuttal should be mentioned in the two opposing sides.

Salient sentences. Salient sentences are those represent a summary of a debate comment. The number of salient sentences to be selected from each debate comment is depended on a defined compression rate. For instance, with a compression rate of 20%, there are two sentences to be selected from a comment containing 10 sentences.

Stance. Stance refers to “the attitude of a person or organization towards something” (Definition of stance in English: Stance, 2014). In this thesis, we use this term to represent a side that a user takes which also refers to the *agree side* or *disagree side.*

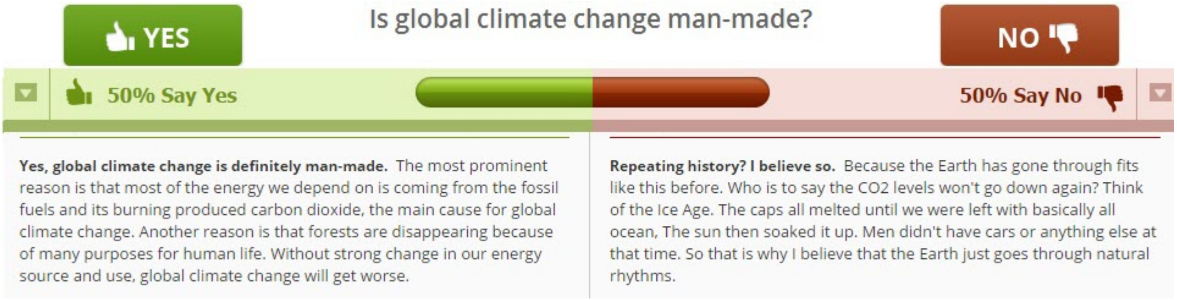
Chapter 1

INTRODUCTION

The exceptional growth of internet use has changed the way people communicate and share their opinions in online media. For example, the micro-blogging website Twitter, allows users to post their content in 280-characters length. A popular social networking site like Facebook allows users to interact and share content with their communities of friends. An electronic commercial website, Amazon, allows users to ask questions on items that they are interested in and give reviews on their purchased products. Whilst these textual datasets are available extensively, creating the abridged versions of the data is necessary for quick and easy access. Automatic Text Summarization has emerged to help readers digest content expressed in such textual data.

Automatic Text Summarization is "the process of distilling the most important information from a text to produce an abridged version for a particular task and user" (Jurafsky and Martin, 2000). Since the emerging of automatic text summarization, a variety of domains has been investigated by the research community. For instance, Ganesan et al. (2010) worked on generating abstractive summaries on product reviews data related to cars, hotels, and electronic products. Additionally, another product review dataset was proposed by Ganesan et al. (2012) which explored reviews on televisions, mobile phones, and GPS devices. Zhuang et al. (2006) focused on summarizing movie reviews. Morales et al. (2008) summarized bio-medical literature with an adoption of a medical ontology. Galgani et al. (2012) developed a summarization system to summarize legal documents. Banerjee et al. (2015) and Wang and Cardie (2012) focused on the summarization of meeting texts.

Whilst these domains have been studied widely by text summarization researchers, online debates, which recently became popular among Internet users, are yet largely unexplored. Online debate data is different from these domains. For instance, as shown



Yes, global climate change is definitely man-made. The most prominent reason is that most of the energy we depend on is coming from the fossil fuels and its burning produced carbon dioxide, the main cause for global climate change. Another reason is that forests are disappearing because of many purposes for human life. Without strong change in our energy source and use, global climate change will get worse.

Repeating history? I believe so. Because the Earth has gone through fits like this before. Who is to say the CO2 levels won't go down again? Think of the Ice Age. The caps all melted until we were left with basically all ocean, The sun then soaked it up. Men didn't have cars or anything else at that time. So that is why I believe that the Earth just goes through natural rhythms.

Figure 1.1: An example of comments in a climate change debate

in Figure 1.1, on the topic *Is global climate change man-made?* people have different viewpoints about the debate topic. People take a stance, express their opinions to agree or disagree with the debate topic, and oppose the other stance. As more content is published, they tend to add more evidence to support their opinions or even oppose other stance. This leads to the arising of contradictory arguments in online debates. However, such contradictory viewpoints are not caused by people's incorrect judgment, but due to their personal opinions, experience, interpretation of events, facts, scientific reasons, or situations taken to make the judgments. This makes the debate data more complex than other domains. Therefore, the summarization of online debates is an important and challenging task and thus motivates us to carry out this research in online debate summarization.

This thesis focuses on a domain-dependent summarization of online debates. Data related to global warming will be explored and used to generate the summaries. At the beginning stage, we define the summary representation of our online debate summarization system and introduce a system architecture for the summary generation. The architecture is used to obtain the results are domain dependent. The next section discusses the aims and scope of this research.

1.1 Aims and Scope

This research aims to address the problem of online debate summarization through the following research questions:

1.1.1 ***How do humans prefer online debates to be summarized?***

When readers face a large number of contrastive viewpoints expressed in online debates, they may need to see the primary direction of the viewpoints; what do other people primarily think about such topic, what are the main positive or negative opinions, why people have different opinions, and how they interact with other people having contrastive ideas. As a software developer, we ask ourselves a question, *How this information should be presented to readers?* We thus conducted an empirical study to investigate how debate summaries should be presented to readers.

In this thesis, we explored seven summary representations of online debates, called *summary designs*. The first summary design is a *Chart Summary*, consisting of bars representing groups of related sentences for the two opposing sides, labels showing short descriptions of the bars, and the counts indicating a number of sentences expressed on a certain topic (Liu et al., 2005). The objective of the Chart Summary is to give an overview of online debate summaries. The second one is a *Table Summary* presenting a table on a certain topic which is split into two separate columns for the two opposing sides. Each side contains sentences related to the topic. The third summary design is a *Side-By-Side Summary*. It is similar to the second one but only shows pairs of sentences in side-by-side views (Paul et al., 2010). The pairs show the contradictions between sentences. The fourth one is a *Conceptual Map* showing a tree of online debate related to a certain topic. The other three summary designs are Combination 1, Combination 2, and Combination 3 which are created by combining a Chat Summary and a Table Summary, a Chart Summary and a Side-By-Side Summary, and a Chart Summary and a Conceptual Map respectively. The objective of creating the combination types is to visualize the summary in both a high-level view and a detailed view. When a user clicks on a bar in a Chart Summary, a detailed summary design related to the clicked bar is generated.

After the design was completed, we conducted a user study to investigate which summary design is the most preferred one. We recruited a group of volunteers to give preference scores and feedback on each summary design. We quantitatively and qualitatively evaluated the study results, which showed that the Combination 2, the combination of Chart Summary and Side-By-Side Summary is the most preferred one. More detail of the empirical study of summary designs can be found in Chapter 3.

1.1.2 ***How the Combination 2 summaries are generated from online debates?***

The target output of our online debate summarization system is Combination 2, the combination of a Chart Summary and a Side-By-Side Summary. We divide the generation of such a summary into three main processes.

1.1.2.1 ***Automatic Salient Sentence Selection***

Salient sentences refer to those which contain the most important pieces of information in debate comments. In short, they are the summaries of the comments. We pave the way to the generation of Combination 2 by developing a summarization system to automatically select salient sentences from debate comments. For instance, an example below shows a debate comment extracted from a debate. The system determines which sentences should be considered as salient.

> Yes, Global Warming is very Real! And if you would look around you, you would see it. Ice burgs are melting in Antarctica and are causing water to rise 7 inches in the last ten years. There are more wildfires, extremer whether. Violent storms etc and it's only getting worst. Temperatures are heating up. Dangerous heat waves are becoming more. And just look at the effect of climate change. I think it's very real and it's just gonna keep getting worst. So yes Global Warming is very real.

In order to determine which sentences in debate comments are salient, we defined a dominant set of features to advocate the selection. A support vector regression model uses the features to score each sentence in the comments. The sentences with the highest scores are automatically selected by the system. Additionally, to determine the number of salient sentences to be selected from debate comments, we applied a compression rate of 20%. To illustrate, from the example above, there are 10 sentences in the comment. Thus, there are two sentences must be selected from this comment. The result of the automatic selection is shown as below.

> And just look at the effect of climate change. I think it's very real and it's just gonna keep getting worst.

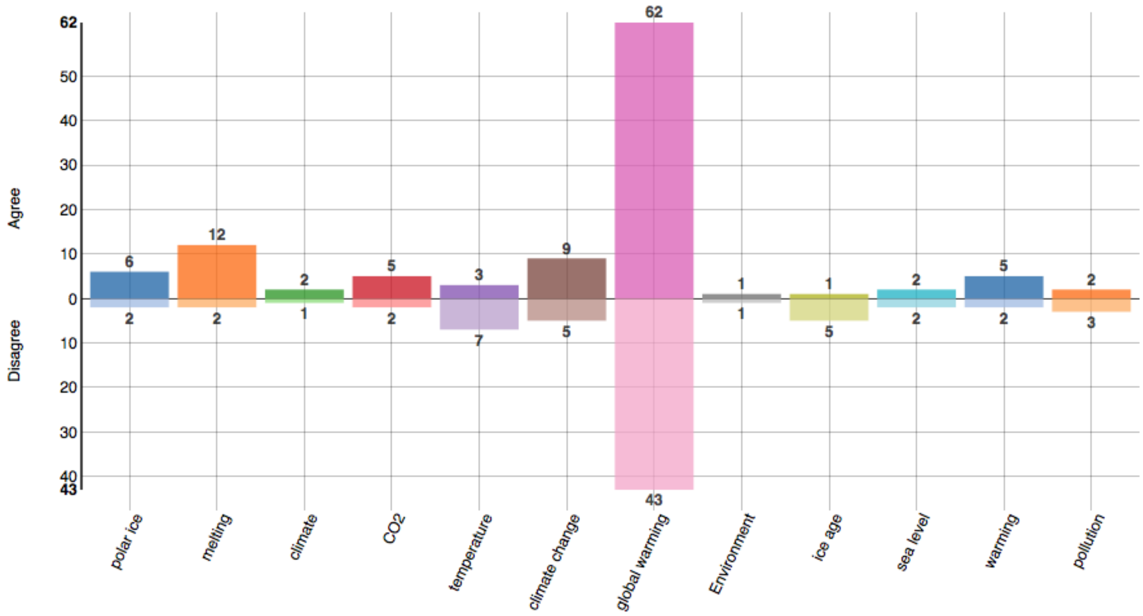


Figure 1.2: An example of the Chart Summary

1.1.2.2 ***Chart Summary Generation***

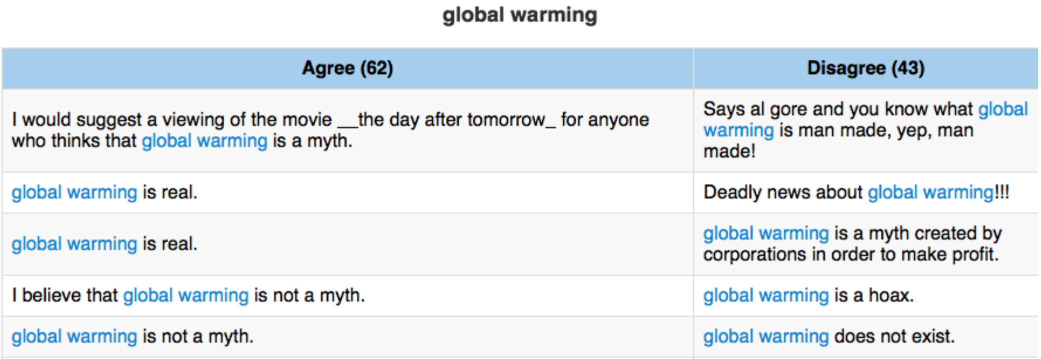
In the second process, we regard Chart Summary generation as clustering tasks. The salient sentences previously selected by the system will be clustered in this process. As shown in Figure 1.2, the bars represent clusters of the salient sentences. We created the bars by exploring two clustering approaches: a term-based approach and an X-means clustering approach. In the term-based approach, we employed ontologies to simply cluster the salient sentences based on the climate change terms shared in the sentences. The sentences containing the similar terms are placed in the same cluster. In the X-means clustering approach, we used X-means clustering algorithm to automatically detect a number of clusters (Pelleg and Moore, 2000). The automatic selected salient sentences are transformed into vectors using the Vector Space Model (VSM) (Salton et al., 1975). In the document indexing stage, we employed the ontologies to automatically annotate key climate change terms in the sentences. The employment of ontologies benefits the transformation of words to vectors by help capturing relevance of specific topics. After the clustering is complete, the number of the salient sentences in the clusters is counted and represented as the frequencies on the bars.

In the next step, we extracted labels from each cluster to present brief descriptions of the clusters. In the term-based approach, we basically consider that the terms which are shared in the clusters are the cluster labels. This is based on an empirical assumption that the terms already illustrate the central meaning of the clusters. In the X-means clustering approach, we applied a Mutual Information approach to generate labels (Manning et al., 2008). The approach calculates a score for each candidate term. The terms having the highest score is chosen as the cluster labels. Once all components are generated, they are combined and visualized as Chart Summary as shown in Figure 1.2. More detail of Chart Summary generation is discussed later in Chapter 5.

1.1.2.3 ***Side-By-Side Summary Generation***

In the last process, we view the generation of Side-By-Side Summary as a contradiction detection task. In general, a contradiction detection task is a subtask in classifying sentences in textual entailment which can be two-way and three-way tasks. In the three-way tasks, a system determines whether a *hypothesis* is entailed, contradictory, or unknown to a *text*. These classes are defined depending on the task's objectives.

In order to generate the Side-By-Side Summary, we created two debate entailment datasets derived from the two clustering approaches and annotated them with one of the two relations: *Contradiction* or *Non-contradiction*. The datasets contain pairs of *hypothesis* and *text* sentences which will be used in a contradiction detection task. The system determines whether pieces of information in the hypothesis are contradictory with those in the text. We developed a classifier together with a set of key features to automatically classify the sentence pairs according to the two relations. The target output for this classification is to obtain the contradictory pairs in which will be the main component in the Side-By-Side Summary. Figure 1.3 illustrates an example of a Side-By-Side Summary. It shows the summary of a topic, *global warming*. The table is split into two opposing sides, Agree and Disagree. Each row represents the sentence pairs which is contradictory. The details of how the Side-By-Side Summary is generated can be found in Chapter 6.



global warming

Agree (62)	Disagree (43)
I would suggest a viewing of the movie __the day after tomorrow_ for anyone who thinks that global warming is a myth.	Says al gore and you know what global warming is man made, yep, man made!
global warming is real.	Deadly news about global warming!!!
global warming is real.	global warming is a myth created by corporations in order to make profit.
I believe that global warming is not a myth.	global warming is a hoax.
global warming is not a myth.	global warming does not exist.

Figure 1.3: An example of the Side-By-Side Summary

1.1.3 *What resources can be used for evaluating the Combination 2 summaries?*

As discussed in Chapter 4, there is no collection of online debates available for the purpose of evaluating debate summarization. Therefore, it is necessary to create a gold standard dataset of online debates summaries for such purpose. Figure 1.4 gives an overview of how online debates are annotated. In the annotation, a group of annotators annotated a set of sentences from each debate comment. The number of the sentences to be annotated is calculated based on a compression rate of 20% of the total sentences, meaning that 2 sentences will be selected from a comment containing 10 sentences. These 20% of the sentences are treated as the summaries of the comments. In this annotation, there are 11 debates[1] annotated based on 5 judgments. In total, 55 annotation sets were derived: 11 debates and each with 5 annotation sets. Each annotation set consists of 341 comments with total 519 annotated salient sentences. We refer to this dataset as the Salient Sentence Selection Dataset (SSSD). This dataset will be used to evaluate the quality of the salient sentences selected by the system, in the Automatic Salient Sentence Selection process. ROUGE evaluation metrics are used to determine n-grams overlaps between the salient sentences selected by the system and the SSSD. More detail of this dataset will be discussed in Chapter 4.

Furthermore, we also created and annotated two datasets for the generation of Side-By-Side Summary. These datasets are derived from the generation of Chart Summary.

[1] A debate contains several debate comments. The statistical information of the comments is discussed in Chapter 4.

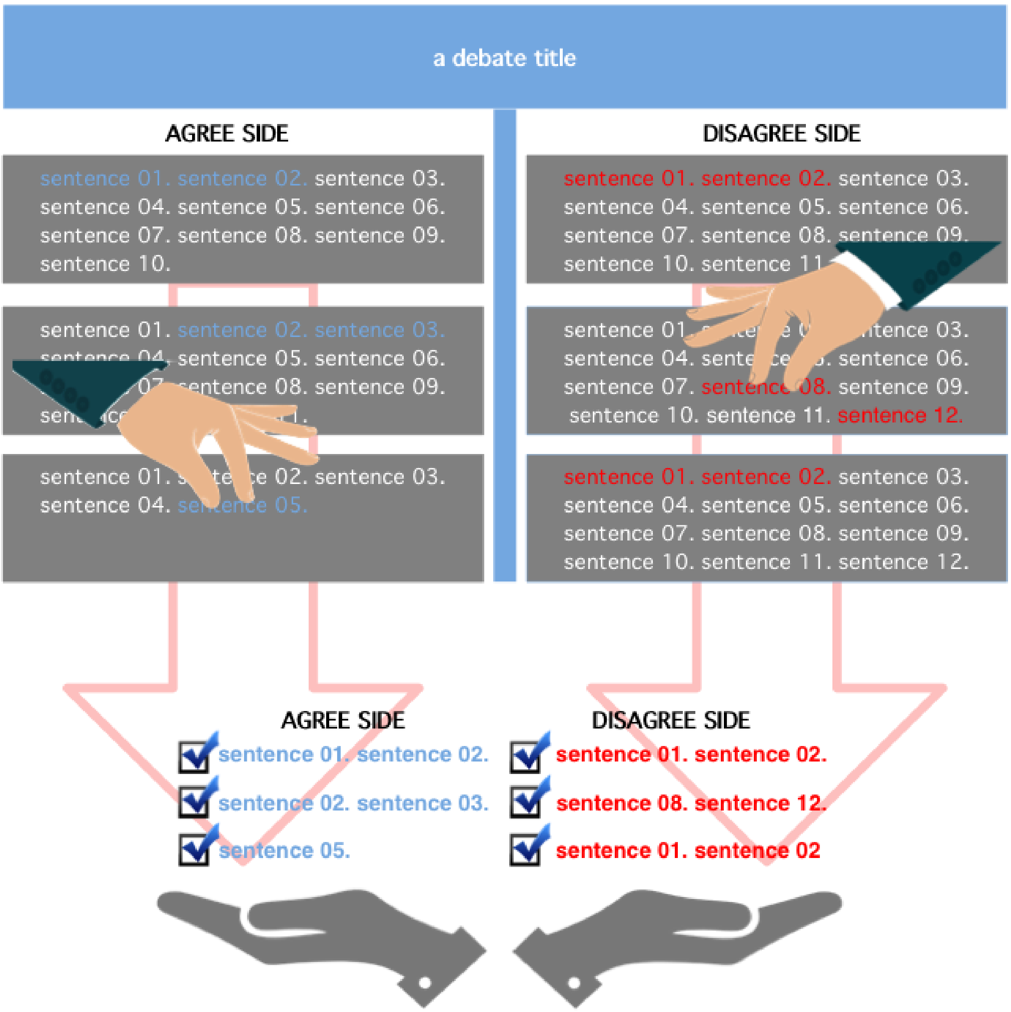

a debate title
AGREE SIDE
DISAGREE SIDE
sentence 01. sentence 02. sentence 03. sentence 04. sentence 05. sentence 06. sentence 07. sentence 08. sentence 09. sentence 10.
sentence 01. sentence 02. sentence 03. sentence 04. sentence 05. sentence 06. sentence 07. sentence 08. sentence 09. sentence 10. sentence 11
sentence 01. sentence 02. sentence 03. sentence 04. sentence 05. sentence 06. 07. sentence 08. sentence 09.
sentence 01. tence 03. sentence 04. sente . sentence 06. sentence 07. sentence 08. sentence 09. sentence 10. sentence 11. sentence 12.
sentence 01. te ce 02. sentence 03. sentence 04. sentence 05.
sentence 01. sentence 02. sentence 03. sentence 04. sentence 05. sentence 06. sentence 07. sentence 08. sentence 09. sentence 10. sentence 11. sentence 12.
AGREE SIDE
DISAGREE SIDE
sentence 01. sentence 02.
sentence 02. sentence 03.
sentence 05.
sentence 01. sentence 02.
sentence 08. sentence 12.
sentence 01. sentence 02

Figure 1.4: A procedure showing how annotators annotate sentences in debate comments

The first dataset was annotated from the term-based clustering results and the second one was from the X-means clustering approach. The datasets were prepared by creating pairs of RTE sentences, text (T) and the hypothesis (H), in each cluster. Sentences from one opposing side were paired with those on the other side. The longer sentences were chosen to be the text and the shorter sentences were chosen to be the hypothesis (Lendvai et al., 2016). Once the data preparation was completed, the pairs of H-T sentences were annotated with one of the two entailment relations: contradiction and non-contradiction. More detail of these datasets will be discussed in Chapter 6.

To conclude, in this thesis, we view that the task of summarizing online debates should not only focus on summarization techniques but also should look further on presenting the summaries into the formats favored by users. We present the generation of the summaries can be achieved by automatic salient sentence selection, salient sentence clustering, and contradiction detection tasks. We also created the necessary evaluation datasets. Some of them are available publicly and we are planning to release the other datasets in a later occasion. More detail of the datasets can be found in Chapter 4 and Chapter 6.

1.2 Thesis Structure

The remaining chapters of this thesis are organized as follows.

- **Chapter 2:** This chapter presents an overview of automatic text summarization according to the approaches and purposes for generating textual summaries. The summarization of online debates falls into the purpose category.

- **Chapter 3:** We conduct an empirical study to investigate how human prefer online debate to be summarized. Seven different summary representations are presented. The result derived from this user study is a target output for our debate summarization system.

- **Chapter 4:** This chapter begins with the introduction of a system architecture and the internal processes for the summary generation. In addition, we also introduce a system for collecting reference summaries and how the reference summaries are annotated. Examples of online debates, the statistical information of the collected data, and the inter-annotator agreement are discussed in this section.

- **Chapter 5:** In the first part of the chapter, we introduce how the summarization system automatically selects salient sentences from debate comments. We create a model and a set of dominant features for the selection. We evaluate the selected sentences against ROUGE evaluation metrics. The next section paves the way for the generation of a high-level view summary, Chart Summary. The chart is generated by applying two clustering approaches, extracting cluster labels, and counting the present of sentences in clusters. We report the clustering results with mean silhouette coefficient.

- **Chapter 6:** This chapter presents how Side-By-Side Summary is generated. It elaborates how debate entailment datasets are created and annotated. Moreover, in the next section, we introduce how the system classifies whether sentence pairs are contradictory. Different combinations of features that maximize the classification results are investigated.

- **Chapter 7:** The key results and contributions from this thesis are concluded. The conclusion looks towards future work, including approaches to enhance the system performance and future research directions in online debate summarization.

1.3 Publications

- **Chapter 3**: Published in the Computational Linguistics and Natural Language Processing (CICLing 2016): Sanchan, N., Bontcheva, K., and Arker, A. (2016). *Understanding Human Preferences for Summary Designs in Online Debates Domain.* Polibits, 54:79–85.

- **Chapter 5**, Section 5.1: Published in the Computational Linguistics and Natural Language Processing (CICLing 2017): Sanchan, N., Arker, A., Bontcheva, K. (2017). *Gold Standard Online Debates Summaries and First Experiments Towards Automatic Summarization of Online Debate Data.* Lecture Notes in Computer Science, 2017, ***Best Paper Award***.

- **Chapter 5**, Section 5.2: Published in the Recent Advances in Natural Language Processing, RANLP 2017, Natural Language Processing and Information Retrieval Workshop: Sanchan, N., Aker, A., Bontcheva, K. (2017). *Automatic Summarization of Online Debates.* Proceedings of Natural Language Processing and Information Retrieval Workshop associated with Recent Advances in Natural Language Processing, RANLP 2017, 2-8 September, Varna, Bulgaria.

- **Chapter 6**: Aim to submit this chapter and the datasets to a conference in later occasion.

Chapter 2

GENERAL OVERVIEW OF AUTOMATIC TEXT SUMMARIZATION

This chapter presents an overview of text summarization. It begins with the introduction of basic summarization methods together with some of about them. In the next section, we review and classify research in text summarization into two groups, including methods for summary generation and the purposes of summary usage. The final section discusses the overview of text summarization evaluation metrics.

2.1 Introduction to Automatic Text Summarization

In order to manually create a text summary, we have to read and understand the original document and then specify important aspects to meet the objective of the summary. As we need to compress the original document (based on the specified text-compression ratios) and to form a summary, the summary may not contain all pieces of information of the original document, but only the information considered as significant. Traditionally, text summarization is described as *extractive summarization* and *abstractive summarization* (Jurafsky and Martin, 2000; Nenkova and McKeown, 2011). In the former category of summarization, the summary is created by selecting and concatenating sentences found in the original document. In contrast, the summary in the latter category contains different words which are not presented in the source. Listing 2.1 to 2.3 illustrate examples of extractive and abstractive summaries. It is a well-known speech by Abraham Lincoln, the Gettysburg address.

Original Text of Gettysburg Address by Abraham Lincoln

Fourscore and seven years ago our fathers brought forth on this continent a new nation, conceived in liberty, and dedicated to the proposition that all men are created equal. Now we are engaged in a great civil war, testing whether that nation, or any nation so conceived any so dedicated, can long endure. We are met on a great battle-field of that war. We have come to dedicate a portion of that field as a final resting-place for those who here gave their lives that this nation might live. It is altogether fitting and proper that we should do this. But, in a larger sense, we cannot dedicate... we cannot consecrate... we cannot hallow... this ground. The brave men, living and dead, who struggled here, have consecrated it for above our poor power to add or detract. The world will little not nor long remember what we say here, but it can never forget what they did here. It is for use, the living, rather, to be dedicated here to the unfinished work which they who fought here have thus far so nobly advanced. It is rather for us to be here dedicated to the great task remaining before us...that from these honored dead we take increased devotion to that cause for which they gave the last full measure of devotion; that we here highly resolve that these dead shall not have died in vain; that this nation, under God, shall have a new birth of freedom; and that government of the people, by the people, for the people, shall not perish from the earth.

Listing 2.1: The Gettysburg Address by Abraham Lincoln. Adapted from *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (p. 823), by Jurafsky, D., & Martin, J. H., 2000: Prentice Hall.

Extractive Summary

Four score and seven years ago our fathers brought forth upon this continent a new nation, conceived in liberty, and dedicated to the proposition that all men are created equal. Now we are engaged in a great civil war, testing whether that nation can long endure. We are met on a great battle-field of that war. We have come to dedicate a portion of that field. But that brave man, living and dead, who struggled here, have consecrated it far above our poor power to add or detract. From these honored dead we take increased devotion to that caused for which they gave the last full measure of devotion – that government of the people, by the people, for the people, shall not perish from the earth.

Listing 2.2: Extractive summary of the Gettysburg Address by Abraham Lincoln.Adapted from *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (p. 823), by Jurafsky, D., & Martin, J. H., 2000: Prentice Hall.

Abstractive Summary

This speech by Abraham Lincoln commemorates soldiers who laid down their lives in the Battle of Gettysburg. It reminds the troops that it is the future of freedom in American that they are fighting for.

Listing 2.3: Abstractive summary of the Gettysburg Address by Abraham Lincoln.Adapted from *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (p. 823), by Jurafsky, D., & Martin, J. H., 2000: Prentice Hall.

Moreover, text summarization can be single-document summarization whereas the summary is derived from one document. Producing the headline of a document is an example. Another type is multi-document summarization in which a summary is produced from a group of documents. Multi-document summarization can be applied in summarizing web content (which contains several pages or sections) on the same topic.

Furthermore, there are three more types that text summarization can be. In generic summarization, the summary is aimed to be created for the general audience — not considering specific users or information need. This summary allows readers to quickly determine what the document is about. In query-focused summarization, the summary contains only the information that answers the user query. For instance, a user issues a query to search for the relevant document. A search engine returns a summary of each document which helps the user to easily determine which document is needed. The final type is update summarization. It is multi-document summarization that summarizes "a set of recent documents relatively to another set of earlier documents" (Alfonesca and Delort, 2012).

2.2 Related Work in Text Summarization

In this thesis, we classify research in automatic text summarization into two groups. In the first group, text summarization is classified according to approaches that are used to process the input text. Single-feature scoring, multi-feature scoring, topic signatures, cluster-based approach, graph-based approach, lexical chain approach, knowledge-based approach, and latent semantic analysis are its division. In the second group, we classify text summarization by the purpose of generating the automatic summary for certain tasks. For example, update summarization aims to generate new information, assuming that readers have previously seen information on this topic. Figure 2.1 outlines these classifications.

2.2.1 *Text Summarization Classified by Approaches*

Regardless of the type of automatic text summarization, an extractive summary is generally produced according to two steps.

1. Textual Unit Scoring. Textual unit to be included in a summary can be words,

phrases, sentences or entire paragraphs which will be scored. The objective of textual unit scoring is to determine the important sentences to be included in the final summary.

2. Summary Generation. The summary is generated by choosing high scoring textual units until a defined summary compression rate is achieved.

The next section will discuss different approaches that have been investigated in text summarization.

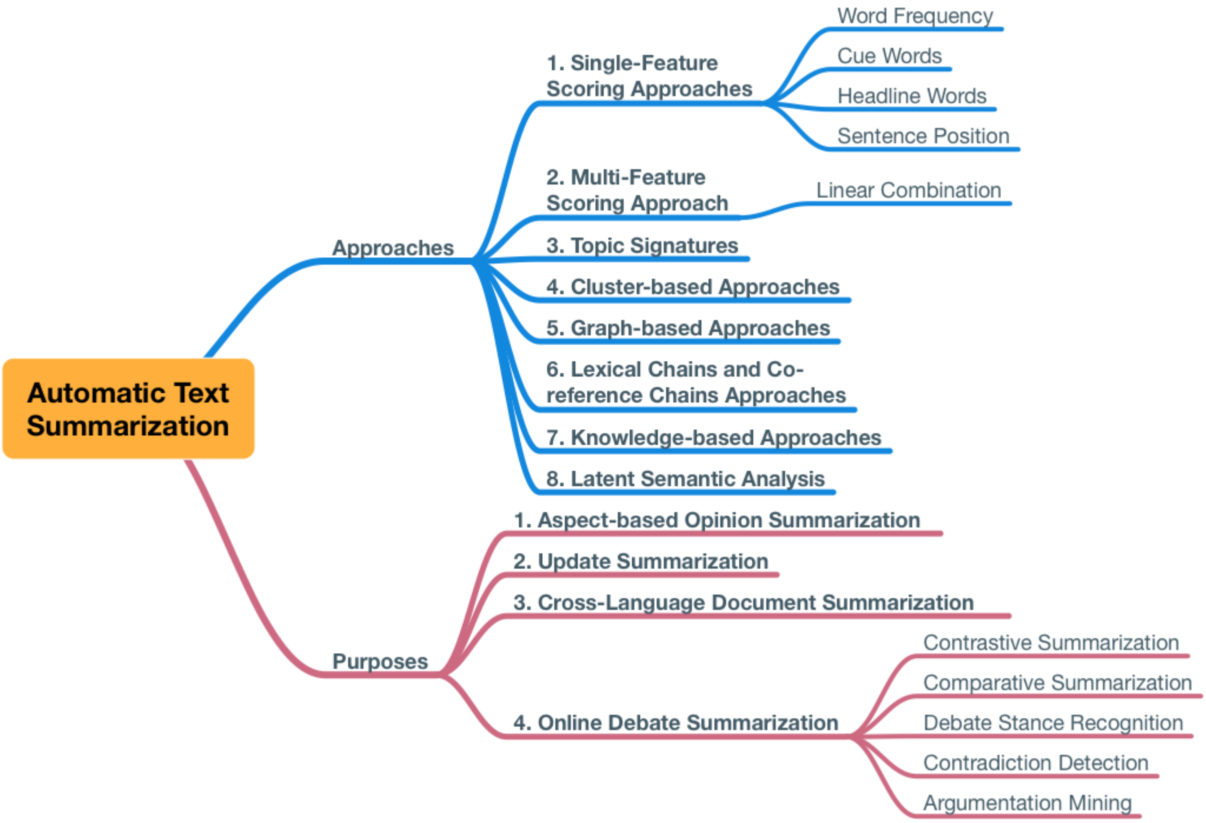

Automatic Text Summarization
Approaches
1. Single-Feature Scoring Approaches
Word Frequency
Cue Words
Headline Words
Sentence Position
2. Multi-Feature Scoring Approach
Linear Combination
3. Topic Signatures
4. Cluster-based Approaches
5. Graph-based Approaches
6. Lexical Chains and Co-reference Chains Approaches
7. Knowledge-based Approaches
8. Latent Semantic Analysis
Purposes
1. Aspect-based Opinion Summarization
2. Update Summarization
3. Cross-Language Document Summarization
4. Online Debate Summarization
Contrastive Summarization
Comparative Summarization
Debate Stance Recognition
Contradiction Detection
Argumentation Mining

Figure 2.1: Classification of Automatic Text Summarization

*2.2.1.1 **Single-Feature Scoring Approaches***

2.2.1.1.1 Word Frequency. The core concept of word frequency employs statistical counts of words in the input. This approach assumes that important sentences are those containing words that occur frequently. The score of a sentence increases if it contains more frequent words. Luhn (1958) explored automatic methods to obtain abstracts by applying word frequency, based on the assumption that a writer generally repeats certain words when he expresses on an aspect of a subject. Such words are considered as *significant words*. Luhn (1958) defined a high-frequency line and a low-frequency line as a boundary of significant and non-significant words. Outbound words will be ignored. Inbound words will be used for sentence scoring. Sentences with high-frequency words are more significant to be included in the summary, than those with low-frequency words.

2.2.1.1.2 Cue Words. Cue words or phrases in a sentence indicate important information. An example of work that applied this approach is Edmundson (1969). The author applied cue words in his work based on the assumption that relevant sentences can be found based on the presence of cue words. Examples of those words are “impossible”, “significant”, and “hardly”.

2.2.1.1.3 Title Words. Title words are also known as headline words and heading words. The presence of words in document title or headline indicates the importance of information. This is based on an assumption that a writer tends to repeat the title words in documents. Thereby, a sentence containing words presented in a document title or a headline is likely to contain important information. As shown in Equation 2.1, Suanmali et al. (2009) applied a heading word approach to score sentences.

$$Title\,Word, T = \frac{number\;of\;title\;words\;in\;a\;sentence}{number\;of\;words\;in\;the\;title} \tag{2.1}$$

2.2.1.1.4 Sentence Position. The position of sentences in the text also indicates the salient parts of documents. Baxendale (1958) conducted an experiment to discover which position in the text the important sentences are mostly found. In this experiment, he conducted a test on 200 paragraphs and concluded that about 85% of paragraphs the important information is placed in the first paragraph and about 7% in the final paragraph. Edmundson (1969) also studied the sentence position. He additionally described that heading and sub-heading are also the significant positions where salient sentences can be found. However, this is not a standard for text in every domain. Lin and Hovy

(1997) found that text in different genres can have different structures so that salient sentences are located at different positions, depending on text genre.

$$\begin{aligned} Sentence\ Position, P = \ & 5/5\ for\ 1^{st}, 4/5\ for\ 2^{nd}, \\ & 3/5\ for\ 3^{rd}, 2/5\ for\ 4^{th}, \\ & 1/5\ for\ 5^{th}, 0/5\ for\ other\ sentences \end{aligned} \tag{2.2}$$

2.2.1.1.5 Sentence Length. Length of the sentences can be also used to determine which sentences should be included in the summary. The assumption behind this is either a very short or a very long sentence is unlikely to be included in the summary. Equation 2.3 is a formula to acquire a score for sentence length.

$$Sentence\ Length, L = \frac{number\ of\ words\ in\ a\ sentence}{number\ of\ words\ in\ the\ longest\ sentence} \tag{2.3}$$

2.2.1.2 ***Multi-Feature Scoring Approaches***

In the selection of sentences in the input text for the extractive summarization, a general approach is to score all sentences in the input. High scoring sentences will be consequentially selected and included in the summary. In the previous section, we discussed single features such as *word frequency*, *cue words*, *headline words*, and *sentence position*. In this section, we introduce the utilization of multi-feature scoring approaches which use a combination of features to score sentences.

2.2.1.2.1 Linear Combination. The concept of linear combination is expressed in the filed of linear algebra and other related areas in mathematics. A linear combination formula is defined by multiplying all numbers and constant values (or weights) and adding all of them up (Strang, 2006; Lay, 2006). The linear combination of features has been broadly used by in text summarization. One of the work utilizing this approach was by Suanmali et al. (2009). They generated the summaries by extracting sentences from the input text based on 8 feature, including *Title Feature*, *Sentence length*, *Term Weight*, *Sentence Position*, *Sentence to Sentence Similarity*, *Proper Noun*, *Thematic*

Word, and *Numerical Data*. In this work, each pre-processed sentence is transformed into a vector of eight features. Later, a score of a sentence is calculated from the features. A sentence weighting equation was defined to calculate weights for the features. Equation 2.4 shows how the definition of the weight function for the eight features is defined. $Score(S)$ refers to the score of the sentence S and $S\_Fk(S)$ refers to the score of the feature k.

$$Score(S) = \sum_{k=1}^{8} S_{Fk(S)} \tag{2.4}$$

To score a sentence, values derived from the features are multiplied with those derived from the weight function. Then, all the values are added up. A set of highest scoring sentences is then extracted and used for generating the summary. Equations 2.1 to 2.3 show some features used by Suanmali et al. (2009). Note that in equation 2.2, the first 5 sentences in the paragraphs are considered as the significant sentences.

$$Score\ S_i = w_1 T_i + w_2 L_i + w_3 P_i \tag{2.5}$$

After calculating the score of each feature, the total score of a sentence is represented by Equation 2.5. $Score\ S_i$ is the total score of a sentence S_i. The variables T_i, L_i, and P_i indicate feature scores of sentence S_i. w_1, w_2, and w_3 refer to the weights of the linear combination of the title, sentence length, and sentence position features.

Another example of applying linear combination was proposed by Saggion (2008). They implemented a toolkit on GATE (Cunningham et al., 2011) for summarizing text, called. SUMMA. The tool allows users to combine language resources and components in the application for creating different summarization pipelines. Examples of the features used for producing scores and generating summaries are sentence position, term frequency (TF-IDF), cue words, title words, etc.

2.2.1.3 ***Topic Signatures***

In the summary generation, one important key is to identify concepts in the input text to be included in the summary. Topic signatures are the important concepts automatically extracted from the input text. The intuition of topic signatures was introduced by Lin and Hovy (2000). They are generated by comparing words in two sets of text using the concept of the likelihood ratio. Words appearing occasionally in the input but rarely in other text are considered the topic signatures. Lin and Hovy (2000) generated summaries by extracting sentences containing unigram, bigram, and trigram topic signatures. They reported that the best summaries are obtained by the utilization of bigram topic signatures.

2.2.1.4 ***Cluster-based Approaches***

The intuition of the cluster-based approach is to group similar sentences into the same clusters. Similar sentences are usually measured by applying the concept of similarity in which highly similar sentences are grouped into the same cluster. The most common approach is the utilization of cosine similarity to define which sentences are close to the other sentences. Equation 2.6 illustrates the equation of cosine similarity measure.

$$cos(\vec{q},\vec{d}) = \frac{\sum\limits_{i=1}^{n} q_i d_i}{\sqrt{\sum\limits_{i=1}^{n} q_i^2}\sqrt{\sum\limits_{i=1}^{n} d_i^2}} \tag{2.6}$$

Generally, clustering algorithms can be classified as agglomerative and partitional. Agglomerative clustering is an iterative clustering algorithm in which each sentence is considered as an individual cluster at the first stage. Later on, all individual clusters will be iteratively merged into larger clusters when some conditions are met (Tunggal, 2012). In contrast, in partitional clustering approach, all sentences are placed in a single large cluster. This cluster will be iteratively divided into smaller clusters. These clusters will consequently contain only high similarity sentences. The K-Means algorithm is an example of partitional clustering algorithm (Tunggal, 2012).

Saggion and Gaizauskas (2004) worked on a multi-document summarization by determining sentences at the cluster centroid. A cluster centroid represents a group of related sentenced. Sentences in the clusters are scored by combining a set of similarity features, an absolute document position feature, and the adjusted weights. The final scores are used for ranking the sentences.

Radev et al. (2004) presented an extractive multi-document summarizer, MEAD, which creates a summary of news documents based on document cluster centroids. In this work, related documents are grouped together into clusters. Each document is represented as a vector of TF-IDF (Term Frequency Inverse Document Frequency). Centroids of documents are pseudo-documents which compose of words that have TF-IDF scores above a predefined threshold. These centroids are used to determine sentences in each cluster that are most similar to the centroid. As the centroid contains highly ranked sentences, sentences containing words from a centroid are considered as significant and included in the summary. Centroid values, position values, and first sentence overlap are the key features that are used by the MEAD algorithm to extract sentences.

2.2.1.5 ***Graph-based Approaches***

Graph-based approaches are used in automatic text summarization to map text as a graph. The representation of the graph is the connection or linkage between nodes and edges. While nodes represent textual units in documents, edges represent different types of relations between the connected nodes.

Erkan and Radev (2004) applied a graph-based approach to summarize the DUC 2004 data, newswire articles. In the graph, each node is represented by a sentence and edges are similarity relations between sentences. Sentences are transformed into Bag-of-Word representation. Words in sentences are used to compute TF-IDF values. These values are later used to calculate cosine similarity scores, indicating the similarity between two nodes. If the similarity is higher than a predefined threshold, a relation connected between the two nodes is drawn between them. Sentences that are strongly connected to many other sentences are considered as salient sentences and should be included in the summary.

Plaza et al. (2010) worked on a summarization of biomedical text as a graph-based approach. In the graph, they present the nodes as concepts derived from the UMLS Metathesaurus. The links among the nodes represent different semantic relationships. In this work, the researchers point out that the performance of the system improved with the use of word sense disambiguation.

2.2.1.6 ***Lexical Chains and Co-reference Chains Approaches***

The notion of *cohesion* was initially presented by Halliday and Hasan (1976), elaborating how words attached together to form text. Text formation is based on the use of *grammatical cohesion* and *lexical cohesion*. Lexical cohesion is a cohesion that results from semantic relations between terms. Morris and Hirst (1991) described an example of a semantic relation in sentences as:

> 1. Mary likes *green* apples.
> 2. She does not like *red* ones (Morris and Hirst, 1991).

Morris and Hirst (1991) classified lexical cohesion into two classes: *reiteration* and *collocation*. The first class covers repetition terms, synonyms, and hyponyms. The second class covers words that tend to occur together (i.e. teacher and school). However, Morris and Hirst (1991) mentioned that lexical cohesion does not appear only between two terms, but also over the sequences of nearby related terms. The sequence of related terms is called *lexical chains*. It contributes to the continuity of lexical meaning. In other words, lexical chains go beyond sentence boundaries and can connect to other terms over the entire text. Morris and Hirst (1991) expressed this example as below. The terms: *virgin*, *pine*, *bush*, *trees*, *trunks*, and *trees* are lexical chains spanned in the following example:

> In front of me lay a virgin crescent cut out of pine bush. A dozen houses were going up, in various stages of construction, surrounded by hummocks of dry earth and stands of precariously tall trees nude halfway up their trunks. They were the kind of trees you might see in the mountains (Morris and Hirst, 1991).

Lexical chain approaches heavily rely on WordNet, an online thesaurus source providing access to word senses, synonym, antonym, general meaning, and specific meaning (Nenkova and McKeown, 2011). For instance, Barzilay and Elhadad (1997) used the lexical chains technique integrated with WordNet to construct a summary, based on the concept that a generated summary for "network" has to reflect the occurrences of chained terms "network", "net", and "system". Otherwise, the summary would extract information separately for every term. The authors chose candidate nouns and noun compounds (e.g "digital computer") from the input text and derive chains (related words) of these candidate words from WordNet. Groups of lexical chains were constructed. Each group had a different meaning. Lexical chains are scored by utilizing Equation 2.7 and the strength of the chain is determined according to the criterion shown in Equation 2.8. The strong chains indicate important sentences which will be extracted.

$$Score(Chain) = Length * Homogeneity \tag{2.7}$$

where:

Length: Number of occurrences chain members

Homogeneity index: 1 - the number of distinct occurrences divided by the length

$$Score(Chain) > Average(Scores) + 2 * StandardDeviation(Scores) \tag{2.8}$$

To extract sentences by utilizing the selected strong chains, three alternatives were investigated: for each chain, (1) extract the sentence that contains the first appearance of a chain member in the text, (2) extract the sentence that contains the first appearance of a representative chain member in the text, and (3) extract the sentence matching several chain members. Representative chain members refer to terms that represent topics more than other terms in the chain. Those terms occur more frequently than other terms in the chain are regarded as representative chain members. In this work, the authors reported that the second alternative outperformed the others.

Brunn et al. (2001) also applied lexical chains to increase the summary coherent. In the preprocessing step, the input text was divided into segments that express the same topics. Two-phases sentence selection approach was used. Segments are ranked with scores of a lexical chain. The best scoring segments are used to select the most salient and the best-connected sentences.

Azzam et al. (1999) focused on generic text summarization. The system produces summaries from text by identifying the best chain that represents the core topic of a text. The author defines some criteria in the selection of the best chain. First, the chains should contain the instances which are frequently mentioned in a text. If multiple chains exist with the same length, only a single chain is selected. Second, the candidate chains which span similarity to the original text are more flavored. Third, the chains that contain instances appearing in earlier paragraphs or in the headlines are preferred. This is based on the assumption that terms appearing in the titles and earlier sentences are significant. In the summary generation, several heuristic rules are defined to extract the best coreference chain from the set of coreference chains. The best chain is used to extract sentences to be included in the summary.

Saggion and Gaizauskas (2005) worked on a multi-document summarization on biographical text using a pattern-based approach. The system generates a summary of a target based on the given cluster of documents related to the target. An example of a summary includes aspects of the target i.e. name, occupational background, age, and health condition. These are identified by co-referencing algorithms. To generate a summary, the system identifies a set of representative sentences from the input using a pattern-matching algorithm. Then, the redundant information in the representative sentences is reduced until meeting a defined compression rate.

2.2.1.7 ***Knowledge-based Approaches***

Tunggal (2012) mentioned that content in documents typically related to certain topics or events. In general, they are a member of particular domains which uniquely have their own knowledge structure. One example of knowledge that is frequently used in automatic text summarization is ontology. “Ontology is a collection of key concepts and their inter-relationship collectively providing an abstract view of an application domain” (Lee et al., 2003). The utilization of ontology in text summarization enhances summarization process. For example, Morales et al. (2008) applied UMLS[1], a medical ontology, to help the summarization of the bio-medical literature. The authors applied the medical ontology to capture concepts in data and presented them as nodes in the graph-based representation.

[1] `http://www.nlm.nih.gov/research/umls/`

2.2.1.8 *Latent Semantic Analysis*

In Latent Semantic Analysis (LSA), the input document is transformed into a word by sentence matrix A. The rows of matrix A represent words appearing in the input and the columns are the sentences. In a matrix, an entry a_{ij} corresponded to the weight of word i in sentence j. The weight is derived from the calculation of TF-IDF. The sentence without words indicates that the weight is zero. The size of the matrix depends on the size of the input document. Singular Value Decomposition (SVD) is a standard technique applied to matrix A, as the product of three matrices $U\Sigma V^T$. The complete SVD formula is shown in equation 2.9 (Gong and Liu, 2001; Steinberger et al., 2007).

$$A = U\Sigma V^T \tag{2.9}$$

Gong and Liu (2001) proposed a generic summarizer that generates a summary by ranking and extracting sentences from the input text based on two methods. The first method applies standard IR methods to measure sentence relevance and the second method uses LSA to semantically determine salient sentences. By applying the LSA approach, Gong and Liu (2001) found that matrix V^T indicates salient information (i.e. topic words) discussed in the input document. In order to generate a summary, each row of matrix V^T is determined and the highest-value sentences are selected until the summary compression rate is reached.

2.2.2 *Text Summarization Classified by Purposes*

In this section we classify text summarization by the purpose of generating automatic summaries for certain tasks. For example, update summarization aims to generate new information assuming that readers have previously read some information on this topic. Moreover, the problem of generating summaries for online debates is also grouped in this section, as we aim to generate the summaries of contradictory in debate data.

2.2.2.1 *Aspect-Based Opinion Summarization*

Aspect-based opinion summarization focuses on capturing the opinionated aspects in content and extracting sentiment on those aspects. Hu and Liu (2004) worked on the summarization of product reviews, categorized by sentiment orientations. They began with mining product aspects from customers' opinions, determining opinion orientations, and then summarizing the overall results in a table format. Another interesting work was done by Zhuang et al. (2006) on mining and summarizing movie reviews. They extracted features from the reviewed opinions and classified whether the opinions are positive or negative. A multi-knowledge based approach integrated WordNet, a statistical analysis, and a movie knowledge (i.e. movie names, names of characters, etc.) was proposed to achieve the summarization task.

2.2.2.2 ***Update Summarization***

Another research trend in text summarization is *update summarization.* The objective of generating summaries in update summarization task is to inform the readers about new information from the previous ones they had read. Du et al. (2010) leveraged a manifold ranking with sink points. The sink point refers to sentences having a minimum score. Other sentences with the scores closed to the sink points (e.g sentence sharing similar information with the sink points) will be penalized. It helps capture redundant information in newer information. Moreover, as considering that documents may arrive sequentially, Wang and Li (2010) proposed a new summarization method, an incremental hierarchical clustering framework, to update summary in real time when new documents arrive.

2.2.2.3 ***Cross-Language Document Summarization***

Given a document or a set of documents in one source language, *cross-language document summarization* aims to create a summary in a specific language. Several works, such as Wan et al. (2010), employed machine translation technique to translate documents to the target language before producing a summary. Another popular approach produces a summary in the contrast way. The summary will be initially extracted and then translated to a target language. In this work, we are not emphasizing on this genre of summarization as it is different from the summarization of online debates.

2.2.2.4 ***Online Debate Summarization***

2.2.2.4.1 Contrastive Summarization. Contrastive Summarization is the study of generating the summary for two entities and finding the difference of sentiments among them (Lerman and McDonald, 2009). This type of summarization requires the classification of polarity in order to "contrast" opinions expressed in different sentiments (Campr and Jezek, 2012; Paul et al., 2010). Several researchers have been studying the problem of contrastive summarization. One interesting work focuses on summarizing contrastive sentence pairs by aligning positive and negative opinions on the same aspect (Kim and Zhai, 2009). In this work, contrastive sentence pairs were constructed based on two criteria: 1) choose sentences that represent a major sentiment orientation; and 2) the two sentences should have opposite opinions on the same aspect. Similarity functions were used for determining contrastive sentence pairs. Then sentence pairs were used as input for generating a contrastive summary. The summary was aimed to help readers compare the pros and cons of mixed opinions.

Another similar work was done by Paul et al. (2010). They attempted to summarize contrastive perspectives in the opinionated text by constructing two types of summary, a macro multi-view summary and a micro multi-view summary. The former type contains multiple sets of sentences in which each sentence has different perspectives. These sets can be compared to understand the different perspectives. The latter type contains a set of contrastive sentence pairs. Each pair has two sentences (different perspectives) for a better understanding of the differences between two perspectives. They assumed that input documents have a common opinion target. To determine the sentiment orientation, they used the Subjectivity Clues lexicons presented by Wilson et al. (2005). In addition, the researchers applied the Topic-Aspect Model (TAM), an extension of the Latent Dirichlet Allocation (LDA) model, to extract aspects and opinions in the opinionated text in the first step. In the second step, a random walk formulation was used to score sentences and pairs of sentences.

Witte and Bergler (2007) proposed a *Topic Clusters* approach to generate extractive summaries for contrastive, focused, and update summarization. A topic cluster is an abstract representation of topics occurring in the whole collection of documents. Topic clusters are generated in three steps 1) the extraction of noun phrases (NP); 2) the production of co-reference chains from the generated noun phrases; and 3) the generation of clusters using a fuzzy algorithm and the determination of clusters' size (i.e. a number of NP it contains). Focus on the contrastive summarization problem. The authors defined the contrastive summary as the composition of two parts of the summary: common-theme summary and contrastive theme summary. The first part is generated by obtaining sentences that share most common topics in the collection. As topics are identified by clusters, the larger clusters indicate the more important topics. A highest-rank candidate noun phrase is selected from each cluster. The candidate noun phrases are used for choosing the sentences they appear in. In another part of the summary, the contrastive theme summary is generated by using a defined threshold. For example, common topics are identified if more than 90% of the topics are shared in the collection. If less than 5% of the topics are shared, these indicate unique or distinguishing topics. The authors sorted the distinguishing clusters by the size to obtain a list of topics that are the most important for a document but not mentioned in any other documents. Likewise, to the first part summary, highest-rank candidate noun phrases are identified and used for selecting sentences in the contrastive summary.

Lerman and McDonald (2009) also investigated contrastive summarization for pairs of entities in consumer reviews. The researchers aimed to highlight differences between two products where, for example; a person who is making a decision to purchase a product wants to see the differences between the top candidates without reading reviews for each product. However, this work only focused on generating summaries for two entities in order to highlight their differences. It does not summarize key opinions in text on the same topic.

Another related work was presented by Fang et al. (2012). The authors did not fully work on text summarization. Instead, they worked on the Information Retrieval problem of mining contrastive opinions in political texts. Given a user query, the system finds opinions of multiple aspects respected to the query and quantifies their differences. For example, to answer a question “what are the respective opinions of U.S., China, and India on *Dalai Lama* and how much the difference among them?”, the opinions words are returned as “nonviolent” for U.S., “rebellious” for China, and “Holy” for India. Additionally, the system also reports a score showing contrastive opinions among aspects. The authors proposed a Cross-Perspective Topic (CPT) model to tackle the problem of contrastive opinion modeling. The model helps stimulate how opinions are generated in documents of different aspects. To extract opinion words, the authors utilized opinion clues, which were presented by Furuse et al. (2007), as criteria to judge whether a sentence expressed an opinion. Example categories of opinion clues are thought (I think), intensifier (extremely), impression (confusing), emotion (glad), etc. Then the authors used a scoring function to find statements which best express opinionists' perspectives.

2.2.2.4.2 Comparative Summarization. Comparative Summarization is the study of finding differences between two comparable topics. Sentiment classification may not be required for this type of summarization (Campr and Jezek, 2012). Zhai et al. (2004) worked on comparative text mining problem which aimed to discover common topics in news articles and laptop reviews and to summarize commonalities and differences in a given set of comparable text collections. A probabilistic mixture model was proposed. It generates clusters of topics across all collections and in each collection of documents. The model generates *k* collections of specific topics for each collection and *k* common topics across all collections. Each topic is characterized by multinomial word distribution (also called a unigram language model). High probability words are used as representatives of each cluster and are also included in the summary.

Huang et al. (2011) worked on comparative news summarization. To solve an optimization problem, they used a linear programming approach to select the most appropriate sentences that are able to maximize the comparatives in the summary and the representativeness in news topics. A sentence is considered as *comparative* or *representative* depending on the sharing of comparative concepts or the expression of important concepts about the news topics in the sentence. To identify concepts, the authors determine named entities and bigrams which appear frequently in the documents. The concepts are later used for detecting comparable sentences. The example below shows that the phrase "*FIFA World Player of the Year*" appears in both sentences making the two sentences comparable.

> *Lionel Messi named FIFA World Player of the Year 2010.*
>
> *Cristiano Ronalo FIFA World Player of the Year 2009.*

As the authors assume that the same phrases appearing in the same sentences allowing two sentences to be comparable, this assumption cannot be applied to other pairs of sentences. To illustrate, although a phrase appears in both sentences, the other sentence may focus on other information in that sentence, not to the phrase. This case makes the two sentences incomparable.

Another related work was proposed by Wang et al. (2012). They studied the summarization of differences in groups of comparable documents by proposing a discriminative sentence selection method to extract the most discriminative sentences which best describe specific characteristics of each group of documents.

Most of the work has been focusing on the summarization of negative and positive aspects and making the comparison of those entities side-by-side. It will be more interesting and challenging to summarize contradictory opinions of entities. The next section will discuss a debate stance recognition problem which considers contradictory arguments in the text.

2.2.2.4.3 Debate Stance Recognition. Debate Stance Recognition is also relevant to the problem of online debate summarization that we are interested in. Stance refers to "the attitude of a person or organization towards something" (Definition of stance in English: Stance, 2014). For example, in a debate "*Do you believe in the existence of Global Warming?*", there are two stances or sides which a person may either agree or disagree with the existence of global warming.

A number of research have been studied a recognition of stance in debate text. Somasundaran and Wiebe (2009) noticed that in online debate posts, people debate issues, express their favorites, oppose other stances, and argue why their thought is correct. To determine a positive sentiment about one target, expressing a negative sentiment about the other side is a key target. For instance, in a debate "*which mobile phone is better: iPhone VS Blackberry*", people supporting iPhone may give reasons to affirm why iPhone is better. In addition, they may also express why Blackberry is not. On the Blackberry side, people may also find reasons to support their opinions and argue why the phone is unfavorable. However, to identify stance in this work, it is important to not only consider positive and negative sentiment but also consider which target an opinion refers to.

Somasundaran and Wiebe (2009) presented an unsupervised opinion analysis method for debate-side classification. Their work emphasized on dual-sided, dual-topic debates — there are two sides of a debate. The debate data about named entities, *iPhone vs. Blackberry*, was used in this study. In addition, the subjectivity lexicon was used to determine opinions. The overall processes can be briefly summarized as:

1. Finding and pairing opinions with targets. Opinions are detected by using the identification of subjectivity lexicon words and are paired by using a rule-based system, based on a dependency parse information.

2. Learning aspects. Find probabilities of how frequent an aspect is followed by an opinion expressed on each opposing side. These probabilities are generated from a new set of data collected from an online discussion forum. The author made an assumption that the aspects may be associated with debates.

3. Apply Integer Linear Programming (ILP) to probabilities in (2) to score and classify debate side.

However, in this work, there are some important concerns that should be stressed. First, as some lexicon words contain both objective and subjective senses, the system made a wrong interpretation in sentences. Second, as the debates contain few lexicons, it is difficult to identify opinions in sentences by applying only lexicons.

Another work from Somasundaran and Wiebe (2010) also explored dual-side debate in different domains: *the existence of God*, *health care*, *gun rights*, *gay rights*, *abortion*, and *creationism*. The opinion-target pair, sentiment, and arguing lexicons were used to recognize stance of debates. MPQA corpus which was annotated with arguing subjectivity was used to generate unigram, bigram, and trigram arguing lexicons. For example, "insist" in a sentence, "*Iran insists its nuclear program is purely peaceful purposes.*" indicates the speaker is arguing. Support Vector Machine algorithm integrated with an arguing-based feature, an arguing-lexicon feature, and a sentiment-based feature is used to classify stances of debate posts.

Anand et al. (2011) also worked on the dual-sided debate classification problem. They used a rule base classifier, JRip to identify whether a post is contradictory. A supervised method based on Naive Bayes is utilized for classifying stance. Several linguistic and structural features including unigram, bigram, cue words, repeated punctuation, and opinion dependencies were utilized to construct a stance classification model. Two sets of contradiction data used in this study are *capital punishment* and *cats VS dogs*. This similar approach was also applied in Abbott et al. (2011) for distinguishing agreement and disagreement in political data.

2.2.2.4.4 Contradiction Detection. Contradiction occurs when information is expressed differently in two given texts (Harabagiu et al., 2006). Contradiction is frequently found in online debate data and thus is related to the problem of debate summarization. Researchers proposed various approaches to help identify a contradiction in text. The contradiction detection task is viewed as 1) it is one of the subtasks in classifying relations in textual entailment; and 2) it is predominantly related to sentence similarity and sentence relatedness. Text entailment can be two-way and three-way tasks. In two-way text entailment task, a system determine whether a *Hypothesis (H)* is entailed by *Text (T)* (Dagan et al., 2013). The classification can be "YES" if H is entailed and "NO" otherwise. In the three-way task, the classification can be "YES" if H is entailed, "NO" if texts are contradictory, and "UNKNOWN" if the texts are neither entailed or contradictory. Text entailment beneficially allows us to recognize whether the same meaning is inferred by different sentences. The following elaborates examples of the related work for detecting the contradiction in text.

One of the traditional approaches to detect the contradiction in text was presented by Harabagiu et al. (2006). Their framework has primarily relied on negation and antonym which helped identify the contradiction in text. However, the application of these features only achieves the accuracy of 62%.

Nguyen and Shirai (2013) proposed two classifiers in the detection of agreement and contradiction in English news articles. The first classifier is a rule-based approach which employs a lexical matching to evaluate whether words in sentence pairs have equal meaning and employ negation clues to determine negation in the pairs. Another classifier is a bootstrapping-based classifier which calculates a polarity score of each pair. If the two sentences agree to a defined condition, add 1 to the polarity score. Otherwise, the polarity score is subtracted by 1. The results of this work reveal that their proposed work can efficiently classify agreement but not the contradiction due to the complexity and various kinds of controversy in text. A possible reason is the usage of the cosine similarity measure for word overlaps in the alignment process. It is only able to detect sentence similarity and relatedness, not contradictory text.

Another interesting work was proposed by de Marneffe et al. (2008). They primarily worked on a contradiction detection task. Logistic regression with polarity, number difference, date and time difference, antonym features was also used to classify if sentence pairs are contradictory. The authors pointed out that for two sentences to be contradictory, they must express the same event. Thus, event co-reference method was included to handle such task.

Marneffe et al. (2009) expanded their work by constructing the alignment of hypothesis and passage (text) from a defined function. To determine whether the hypothesis is entailed by the passage text, dot product scores are calculated from a set of features. To classify the contradiction class, Marneffe et al. (2009) utilized a logistic regression approach to classify whether or not sentences pairs are contradictory. The authors reported their precision and recall of the contradiction detection as 28% and 8% respectively.

2.2.2.4.5 Argumentation Mining. Argumentation mining is a new research area which recently attracts the attention of research communities. When people involve in argumentation, they try to understand the stated problems, make a scientific judgment, explain, and defend their opinions (Palau and Moens, 2009). A task of argumentation mining is to detect arguments and their relations presented in text (Šnajder, 2017). The term *relations* in this context refers to a representation of arguments which is constituted by pieces of evidence. In addition, arguments are also presented as *claims*, premise supporting people' opinions which can be true or false (Palau and Moens, 2009). To extract arguments in text, a computer system determines the boundary of text spanning in documents to make a judgment whether it is contributing to an argument, justifies a piece of text is claim or premise, and finally concludes which statement in text is an argument. For these reasons, the task of argument mining is difficult and becomes a challenging task.

Researchers have proposed various methods to detect arguments in textual corpora. Cabrio and Villata (2012) aimed to detect arguments which users used to support their propositions in text. The authors combined the textual entailment framework and the argumentation theory to automatically extract arguments from online debates. In the creation of H-T pairs, they extracted opinions expressed on a topic (referred as an argument) and pair this argument with other arguments. Next, to extract arguments, elements in contradiction and entailment were mapped to those attack and support relations respectively.

Boltužić and Šnajder (2014) collected argumentative text from online discussion forums and train their classifiers to detect arguments in the text. They matched the annotated text to a set of predefined topic-based arguments which can be either attack or support relation. Textual entailment features, semantic similarity features, and a stance alignment features were utilized in the extraction.

2.2.2.5 ***Other Related Summarization Problems***

Park et al. (2011) proposed a method for classifying news articles with different views on contentious issues. This work did not consider the polarities of the article. Instead, they focused on identifying two group of disputants in contentious issues as important features for understanding the discourse. They utilized quotes appearing in news articles to extract disputants before partitioning disputants into two groups. However, the major problem with this approach is that it can be only applied to data containing a great number of quotations.

To sum up, researchers have proposed different approaches to handle the problem of stance identification and stance classification (classification of debate sides), contradiction detection, and argumentation mining. The work that we are interested in, automatic summarization on debate text is rarely found. The next section will discuss the work of Ranade et al. (2013) which is the pioneer in online debate summarization and discuss one related work that grounds argumentation mining work for the summarization of online debates.

2.2.3 *Online Debate Summarization*

Since the work related to online debate text summarization is a novel research area, its related work is rarely found. Ranade et al. (2013) worked on the problem of extractive summarization in online debates. System summaries are generated by ranking the smallest units of debates, called Dialogue Acts (DAs). The most highly ranked DAs are chosen until the desired summary length is reached. A linear combination equation for this ranking uses four different features to calculate DAs.

1. Topic Relevance. Sentences having information or expressing opinions about debate topics (topic-related sentiment sentences) are most important in debate summarization. Topic Directed Sentiment Score feature and Topic Co-occurrence are used to capture topic relevance of DAs.

   (a) Topic Directed Sentiment Score Feature. Topic-related sentiment sentences are scored using a dependency parse of the DAs and the sentiment lexicon, SentiWordNet.

   (b) Topic Co-occurrence Feature captures DAs that contain words which highly co-occur with the debate topic.

2. Document Relevance Feature. TF-IDF and sentiment scores of words are used to calculate the document relevance of DAs as shown in Equation 2.10.

$$tf - idf_{DA} = \sum_{w \in DA} (tf - idf(w) * sentiScore(w)) \tag{2.10}$$

3. Sentiment Relevance Feature is used to count a number of sentiment words and to determine a sentiment score of each DA word.

4. Document Context Relevance focuses on two textual units.

   (a) Sentence position considers the initial and end of debate post are important.

   (b) Sentence length considers long sentences are more meaningful to be included in summary than the short ones.

In this work, extractive gold standard summaries were constructed by two people. ROUGE-1, ROUGE-2, and ROUGE-L (Lin, 2004) were used to evaluate the system against the gold standard summaries. The main concern in this work is that they focused on the application of sentiment words in debate text. However, in some debate text, for example; climate change topics, the data mostly lacks sentiment words. Other opinion clues might be more useful for detecting opinions in this case. Another important concern is that some data may contain multiple debate issues where people argue about the issues without positive or negative expressions. Thus, the opinion that should be included in the summary may hide in text. In the worst case, Park et al. (2011) also noticed that people may ignore the opponent's argument and emphasize a different discussion point instead. For this reason, the debate topic will appear only one opposing side. These are the important concerns.

Another relevant research is by Trabelsi and Zaiane (2014). The authors proposed a Latent Dirichlet Allocation (LDA) model for mining arguing expressions in online debates. The concept of *arguing expression* in this work refers to sentences containing opinions (viewpoints) expressed over K possible topics. Arguing expressions to be mined should express the same topics and viewpoints but converse different lexicons. The authors viewed this as an extraction of textual units in documents by modeling a document as a pair of mixtures of topics and viewpoints. The corpus used in this work consists of three different online debates in which each of them expresses one or more viewpoints and each viewpoint contains one or more arguments. The model was tested on these datasets and evaluated with a perplexity criterion and the Kullback Leibler Divergence. Perplexity measures the generalization of the model to unseen datasets. The lower the perplexity, the better generalization. The latter was used to assess degrees of separation between the probability distributions of their model and a model proposed by other researchers. However, in this work, Trabelsi and Zaiane (2014) did not explicitly create side-by-side pairs of arguing expressions. Only the mining task was performed.

2.3 Evaluation Metrics

Evaluation is one of the most crucial tasks in automatic text summarization. It helps assess the quality of the summaries generated by the system. Evaluation task can be divided into *intrinsic evaluation metrics* and *extrinsic evaluation metrics*. Intrinsic evaluation metrics focus on the assessment of coherence and informativeness of summaries, whereas extrinsic evaluation metrics evaluate how useful is a summary of a given task (Mani, 2001b). As its purpose indicates, it is also called task-based evaluation. This metric is time-consuming, costly, and requires an amount of considerable well-planned processes. Thus, it is not appropriate for system comparison and evaluation during development (Nenkova and McKeown, 2011).

2.3.1 ***Intrinsic Evaluation Metrics***

Two types of summaries that are commonly used in intrinsic evaluation metrics is *reference* or *model* summaries and *system*, *peer*, or *automated* summaries. Reference summaries refer to those manually constructed or annotated by human subjects (annotators) for the purpose of testing and training the system. System summaries are those automatically generated by a computer system. In intrinsic evaluation metrics, system summaries are compared with reference summaries to assess their quality.

2.3.1.1 ***Readability Evaluation***

In intrinsic evaluation metrics, efforts to assess the quality of summaries have been attempted to cover the evaluation of the *readability* and the *informativeness* of system summaries. The assessment of system summaries in the readability evaluation usually covers the quality of text coherence, how the summary is read. One traditional approach to evaluating the coherence of summary is to ask annotators to manually rate the summary based on specific criteria. For instance, Minel et al. (1997) invited a set of participants to manually score the readability levels of a summary based on the presence of *dangling anaphors*, repeating the redundancies of concepts, missing specific content in the structure of summary, etc. In Saggion and Lapalme (2000), participants were asked to rate the acceptability score based on criteria such as spelling and structure, concepts presented in the source documents, conciseness, and the full description of acronyms in text. Moreover, a possible approach to check the coherence is to use a specific software such as grammar checking software Mani (2001a). However, the manual evaluation requires time, labor efforts, and costs. It is therefore not effectively practical for evaluation.

2.3.1.2 ***Informativeness Evaluation***

The evaluation of readability is not sufficient to measure the quality of system summary. To illustrate, very beautiful, cohesive sentences many not cover all necessary information as in the reference summary. Thus, the evaluation of informativeness is required. The term *informativeness* is used to present how much information that system summary can cover compared to the reference summary, in a different compression rate (Mani, 2001a). To evaluate the informativeness, the general evaluation process is to measure how much information is presented in the system summary compared to the reference summary. Edmundson (1969) applied a Subjective Similarity Rating approach to evaluating the informativeness. In this work, system summaries were automatically extracted by using cue words, title words, and sentence position methods. Then, individual judges were invited to rate the similarity score between system summaries and reference summaries on a five-point scales of similarity. However, due to this subjective judgment, it seems to be that the evaluation might not be consistent.

2.3.1.3 ***Agreement Among Annotators***

In the evaluation, it is essential to ensure that the reference summaries are in a good standard for measuring the quality of the summaries generated by the system. Especially, when the selecting reference summaries judged by humans, an issue of whether the summaries are created in the defined criteria or are in a consistent manner arises. For instance, it is probable that annotators disagree and annotate different sentences since the same content can be described in different ways. If this cases frequently occur, the summaries may not be a valuable standard for the evaluation. For this reason, it is important to assess the agreement between the annotators.

2.3.1.3.1 Percentage Agreement is one of the basic measures for determining agreement between two annotators. From Equation 2.11, percentage agreement is a proportion of a number of items in which two annotators mutually agree on to the total number of items that are observed (Scott, 1955). When two annotators agree on an item, $item_{agree}$ will be assigned to *1*. Otherwise, the assignment will be *0*, as shown in Equation 2.12 (Artstein and Poesio, 2008).

Table 2.1: An example of agreement by two annotators

		ANNOTATOR 01		
		ITEM 01	ITEM 02	TOTAL
ANNOTATOR 02	ITEM 01	**20**	20	40
	ITEM 02	10	**50**	60
	TOTAL	30	70	100

$$Percent\ Agreement = \frac{\sum item_{agree}}{total_{observed}} \tag{2.11}$$

where:

$item_{agree}$ = number of agreed items

$total_{observed}$ = total number of observed items

$$item_{agree} = \begin{cases} 1, & \text{if } \textit{the two annotators have a mutual agreement} \\ 0, & \text{if } \textit{the two annotators does not agree} \end{cases} \tag{2.12}$$

Table 2.1 adapted from Artstein and Poesio (2008) summarizes the agreement of both annotators on the two items. From this table, the items at the diagonal are the number in which both annotators have mutually agreed. Thus, the percentage agreement is calculated by summing the number of the two items and dividing it by the total number of items measured. The result is shown below.

$$\begin{aligned} Percentage\ Agreement &= \frac{20 + 50}{100} \\ &= 0.7 \end{aligned}$$

Percent agreement was used in Jing et al. (1998). They found that out of 5 annotators the agreement among 3 annotators or more is considered as the majority opinion. They generated two sets of system summaries, 10% and 20% length summaries. The percentage of the average agreement of those summaries are 96% and 90% respectively. In another work, Passonneau and Litman (1993) had 7 annotators to annotate documents and they found that the boundary of 4 annotators is an effective number for annotating their datasets. However, the drawback of percentage agreement is it does not consider the agreement that would be occurred by chance. This might cause the overestimation in the level of agreement (Hallgren, 2012).

2.3.1.3.2 Cohen's Kappa is one of the most popular approaches for measuring inter-rater agreement between two raters which measure the amount of agreement that could have occurred by chance (Pallant, 2013). Equation 2.13 illustrates the equation of Kappa. From the equation P(A) refers to the times that annotators agree and P(E) is the ratio of times expecting that the annotators to agree. Kappa values range from 1 to 0. The value of 1 indicates there is a completed agreement among the annotators and the value of 0 indicates no agreement other than what is expected by chance (Mani, 2001a). Landis and Koch (1977) broke down the strength of agreement to different levels. For example, the values of K between 0.00 - 0.20 indicate *slight agreement*, 0.21 - 0.40 indicate a *fair agreement*, 0.41 - 0.60 indicate *moderate agreement*, 0.61 - 0.80 indicate *substantial agreement*, and 0.81 - 1.00 indicates *almost perfect agreement*. However, the definition of these ranges of K values was arbitrarily defined.

$$K = \frac{P(A) - P(E)}{1 - P(E)} \tag{2.13}$$

Light (1971) proposed an approach for computing kappa value for 3 or more annotators. The calculation for more than two coders can be performed by computing kappa values for all pairs of annotators and then calculating the arithmetic mean from these values. The agreement measurement still follows the scale defined by Landis and Koch (1977).

2.3.1.3.3 Krippendorff's Alpha can be used to measure agreement in studies with more than two annotators. Equation 2.14 shows the general form of Krippendorff's Alpha in which D_o represents observed agreement and D_e refers to the disagreement that can be expected when chance dominates (Krippendorff, 2004). The value of alpha is ranged between 0 and 1. The value of 1 indicating perfect agreement in which the observed agreement is perfect and disagreement is absent. Thus, D_o will be 0 according to the calculation of the equation. According to the equation, D_e is equal to D_o shaping the calculation of alpha to be 0. Krippendorff (2004) suggests that the alpha value of 0.80 indicates a good reliability of agreement. The value of 0.67 to 0.80 can be used only for the cautious conclusion case.

Krippendorff's Alpha is more generalized than Cohen's Kappa since it supports distant metrics which can be applied to different kinds of variables including, ordinal, interval, and ratio (Hallgren, 2012; Artstein and Poesio, 2008). The advantage of applying Krippendorff's alpha with a distance metric is we can measure the agreement which is completely or partially agreed among the annotators (Passonneau, 2004).

$$\alpha = 1 - \frac{D_o}{D_e} \tag{2.14}$$

2.3.1.4 ***Utility-Based Measure***

The utility-based measure is another approach that compares system summaries to reference summaries. Each sentence in reference summaries is considered as unequal. Unlike the boolean annotations, an annotator needs to make more fine-grained decisions on assigning a score, called a *utility point*, to a sentence which can be ranged from 1-10 depending on the size or the number of sentences in a reference summary. The intensity of labor-effort in the annotation process is the drawback of this approach (Radev et al., 2000).

2.3.1.5 ***Pyramid Method***

Another interesting summary evaluation metric is the Pyramid method. Multiple human summaries are manually analyzed to form a gold standard summary for evaluation. Chunks of information having the same meaning are grouped into a summary content unit (SCU). The pyramid method assigns each SCU a weight which is reflected by the number of human summarizers that have highlighted the SCU in the text. Each SCU is then put into a pyramid where each pyramid layer represents how many summarizers have suggested the SCU. Finally, the number of pyramid layers is therefore equal to the number of summarizers – the higher the more important SCU (Nenkova and McKeown, 2011; Hobson, 2007). The drawback of the pyramid method is that it is human labor and time-consuming since it requires significant effort for the annotation tasks.

2.3.1.6 ***ROUGE***

ROUGE, Recall-Oriented Understudy for Gisting Evaluation, is one of most common evaluation metric in text summarization as it is recall-oriented. ROUGE is an automatic approach which considers n–grams as units for comparing system and gold standard summaries (Lin, 2004). The ROUGE-N formula is illustrated in equation 2.15.

$$ROUGE\text{-}N = \frac{\sum\limits_{S\in\{GoldStandard\ Summaries\}} \sum\limits_{gram_n\in S} Count_{match}(gram_n)}{\sum\limits_{S\in\{GoldStandard\ Summaries\}} \sum\limits_{gram_n\in S} Count(gram_n)} \tag{2.15}$$

where:

n is the length of n-gram, $gram_n$. $Count_{match}(gram_n)$ is the maximum number of n-gram co-occurring in the system summary and a set of gold standard summaries.

As shown in equation 2.15, ROUGE-N is recall-oriented since the denominator is the total summation of the number of n-grams derived from the gold standard summaries. By adding more gold standard summaries, the number of n-grams in the denominator of the ROUGE-N formula will increase (Lin, 2004).

2.3.2 ***Extrinsic Evaluation Metrics***

The objective of the extrinsic evaluation is to measure the usefulness of a summary toward a particular task (Juan-Manuel, 2014; Steinberger and Jezek, 2009). An example of applying extrinsic evaluation in text summarization is by McKeown et al. (2005). They investigated whether muti-document summaries generated by a system would benefit users in a task. The authors compare four groups of subjects either with the original text, one-sentence summaries, system summaries, and human summaries. The results show that participants given summaries produce better quality reports than those without summaries.

2.4 Summary

In this chapter, we presented an introduction of automatic text summarization and broke down the related work into two groups, including methods for summary generation and the purposes of summary usage. We also discussed the evaluation metrics that have been used to evaluate summaries in automatic text summarization research.

Having been conducting research on automatic text summarization in various domains, researchers have not yet widely explored the work in online debate summarization. The related work only focuses on summarizing contrastive text, summarizing the difference and commonality among sets of text, detecting debate stances, finding the contradiction in text, and extracting arguments in text. Additionally, other relevant work only aims to create summaries for online debates. They neither explicitly highlight what is the key content to be summarized nor present the contradictory summaries side-by-side (Ranade et al., 2013; Trabelsi and Zaiane, 2014). These leave a research gap and inspire us to explore the work in online debate summarization.

The next chapter paves the way to the summarization of online debates. We will discuss how the output of our online debate summarization is defined. We present different summary representations for online debates and highlight their advantages for presenting debate content. Later on, we conduct an empirical study to investigate which summary representation is the most preferred one. The most prefer summary representation will be considered as the target output for our online debate summarization system.

Chapter 3

A STUDY OF HUMAN PREFERENCES FOR SUMMARY DESIGNS

Research on automatic text summarization has primarily focused on summarizing news, web pages, scientific papers, etc. While in some of these text genres, it is intuitively clear what constitutes a good summary, the issue is much less clear-cut in social media scenarios like online debates, product reviews, etc., where summaries can be presented in many ways. As yet, there is no analysis about which summary representation is favored by readers. In this work, we empirically analyze this question and elicit readers' preferences for the different designs of summaries for online debates. Seven possible summary designs in total were presented to 60 participants via an online study. Participants were asked to read and assign preference scores to each summary design. The results indicated that the combination of Chart Summary and Side-By-Side Summary is the most preferred summary design. This finding is important for future work in automatic text summarization of online debates.

3.1 Related Work in Summary Representations

Due to the availability of social media sites and the exponential growth of Internet use, online users communicate and share their opinions in textual form in online media. Debate forums are one example of the media in which users express their opinions about their favorite debates. As more and more content is published it becomes increasingly difficult for readers and potential debate participants to easily or quickly digest and understand the overall details of controversial discussions. Automatic text summarization can be used to overcome this problem by helping users digest the information on web forums.

Related work has investigated different summarization approaches such as aspect-based

(Hu and Liu, 2004; Zhuang et al., 2006; Lu et al., 2009), meeting (Banerjee et al., 2015; Wang and Cardie, 2012), contrastive, (Lerman and McDonald, 2009; Paul et al., 2010; Kim and Zhai, 2009; Campr and Jezek, 2012) and comparative summarization (Zhai et al., 2004; Huang et al., 2011; Wang et al., 2012; Witte and Bergler, 2007). The summary either contains statistics about negative and positive opinions provided for each aspect (Liu et al., 2005), lists most frequent positive and negative opinionated sentences (Hu and Liu, 2004) or contains positive and negative sentences side-by-side so that they are contrastive to each other (Paul et al., 2010). Some studies claim that one of these outputs is preferred to the another (e.g. Kim and Zhai (2009)). However, there is no empirical evidence establishing which summary output is favored by human readers. This lack of evidence requires an empirical study in order to acquire appropriate information about user preferences and summary outputs for a specific purpose.

In this chapter, we present an empirical study that investigates different types of summary outputs, called *summary designs*, for debate discussion. We aim to answer the research question: "Which summary design is the most preferred for presenting the abridged version of debate content?". To answer this question, we collected opinionated comments about climate change from the Debate discussion forum[1] and manually constructed the following summary designs: a Chart Summary, a Table Summary, a Side-By-Side Summary, and a Conceptual Map. The first three designs were informed by prior research (i.e. Hu and Liu (2004); Paul et al. (2010); Liu et al. (2005)) and the latter was proposed in this study. In addition, we also manually constructed three combinations of those summary designs. In total, there are 7 summary designs used in this study. Next, 60 participants were recruited to an online study. The study asked the participants to give preference scores to each summary design. We found that the combination of the Chart Summary and the Side-By-Side Summary is the most preferred summary design. To the best of our knowledge, this is the first empirical study conducted to understand which type of summarization outputs is favored by humans, and we think that our results are a valuable contribution to future studies that aim to summarize online debates.

The rest of this chapter is organized as follows: first, we briefly describe the climate

[1] http://www.debate.org

change data and our approach to select salient sentences from it to construct our summaries in Section 3.2. Section 3.3 introduces 7 different summary designs and the methodology we used to manually construct them. We discuss the empirical study in Section 3.4 and analyze the results in Section 3.6. Section 3.7 is the conclusion.

3.2 Debate Data and Salient Sentence Selection

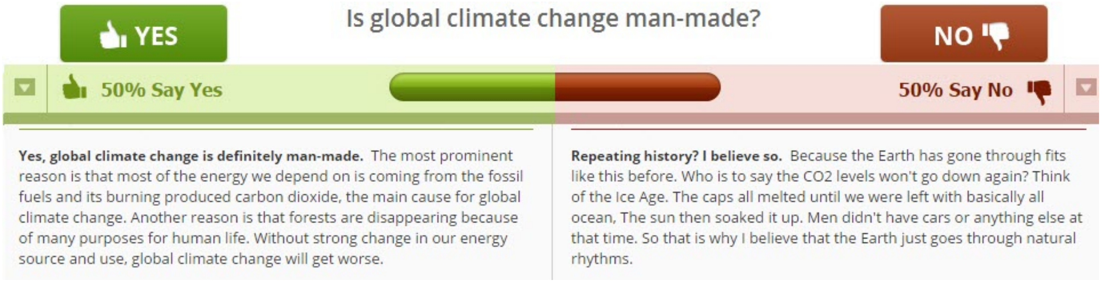
3.2.1 Data

Previous research has focused on summarizing documents in news articles, product reviews, movie reviews, medical data, and other related domains. Our aim is to investigate how to summarize debates on the highly discussed topic of global warming or climate change[2].

Within the Debate discussion forum, people position themselves differently in the debate on the existence of global warming. This leads to debates, in which proponents and opponents of the global warming phenomenon controversially express their sentiments and opinions on diverse global warming topics. Contradictory opinions are voiced on many topics of global warming such as its characteristics, causes, consequences, and its existence. Due to a high volume of contributions, reading and digesting all these discussions are not possible for readers. A summary covering the different topics as well as the different opinions on each topic would help the reader digest the overall discussion. However, it is not clear at present what such a summary should look like. Therefore, we empirically investigate how to best present such a summary to the readers.

The data that we used to construct the summary designs were collected from the Debate discussion forum. Overall, 259 debates with total 1600 comments were collected. Examples of the debates are “Is global warming a myth?”, “Is global warming fictitious?”, “Is global warming true?”, etc. The comment’s length varies between 16 and 385 words, averaging at 91 words. Figure 3.1 shows an extract from the debate “Is global climate

[2] We use the term “global warming” and “climate change” interchangeably. In the scientific context, climate change has a broader meaning: the changes in climate characteristics. The earth’s average temperature change, the flow of ocean current that causes the decrease and increase of temperate in some areas, rainfall, and snow falling are examples of climate change. Global warming has more specific meaning in which the temperature increases over the time (Boykoff and Boykoff, 2007; Markner-Jäger, 2008).



Is global climate change man-made?

Yes, global climate change is definitely man-made. The most prominent reason is that most of the energy we depend on is coming from the fossil fuels and its burning produced carbon dioxide, the main cause for global climate change. Another reason is that forests are disappearing because of many purposes for human life. Without strong change in our energy source and use, global climate change will get worse.

Repeating history? I believe so. Because the Earth has gone through fits like this before. Who is to say the CO2 levels won't go down again? Think of the Ice Age. The caps all melted until we were left with basically all ocean, The sun then soaked it up. Men didn't have cars or anything else at that time. So that is why I believe that the Earth just goes through natural rhythms.

Figure 3.1: An example of comments in a climate change debate

change man-made?". From the figure, we see that the debate contains two opposing sides, *Agree* and *Disagree*, which are originally divided by the forum. As shown in the figure, one side argues that climate change is man-made and the other side thinks that it is not the case. Both opposing sides also provide evidence for their propositions about the existence of global warming. We stored the data for each opposing side separately.

Table 3.1: The distribution of salient sentences containing each frequent topic

Frequent Topics	**Agree Side**	**Disagree Side**	**Total**
gas	5	3	8
plant	15	6	21
carbon dioxide	38	14	52
climate change	17	7	24
global warming	6	6	12
government	10	5	15
science	13	6	19
Total	104	47	151

3.2.2 Salient Sentence Selection

In order to manually select the salient sentences, we explored the debate "Is global climate change man- made?"[3] since it is one of the longest debates and covers diverse topics compared to the other debates in our data. From the debate, we counted a number of topics appearing in each sentence. We then manually extracted the top 7 frequent

[3] http://www.debate.org/opinions/is-global-climate-change-man-made

topics, which are mentioned in opinions expressed by global warming proponents. Those topics include *gas*, *plant*, *carbon dioxide*, *climate change*, *global warming*, *government*, and *science*. For each of these topics, we manually selected salient sentences expressing the topics. Our selection process was guided by the following aspects:

1. **Topic Filter.** For each opposing side, the sentences should contain or mention one of the frequent topics. Otherwise, they were ignored.

2. **One Topic Expression.** In the manual salient sentence selection, sentences are chosen based on the assumption that a sentence refers to only one primary topic.

In this selection, we derived 151 salient sentences in total. Table 3.1 demonstrates the distribution of these sentences across the 7 frequent topics. The stance of the sentences is derived from the stance of the original comments, from which these sentences were extracted. After the selection process, we manually presented them in the summary designs described in the next section.

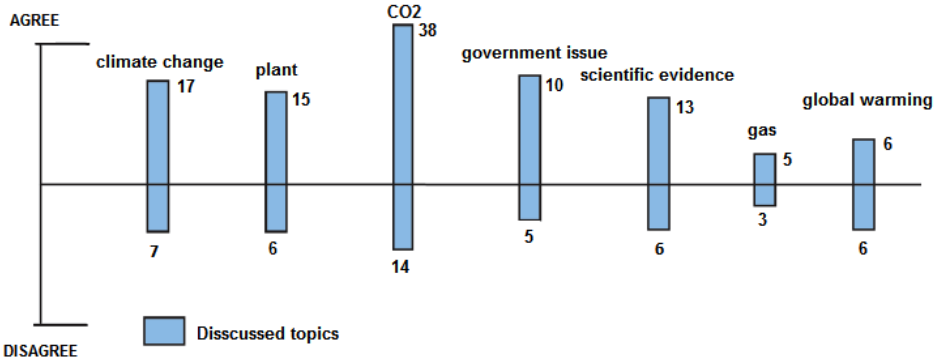


Figure 3.2: Chart Summary

3.3 Summary Designs

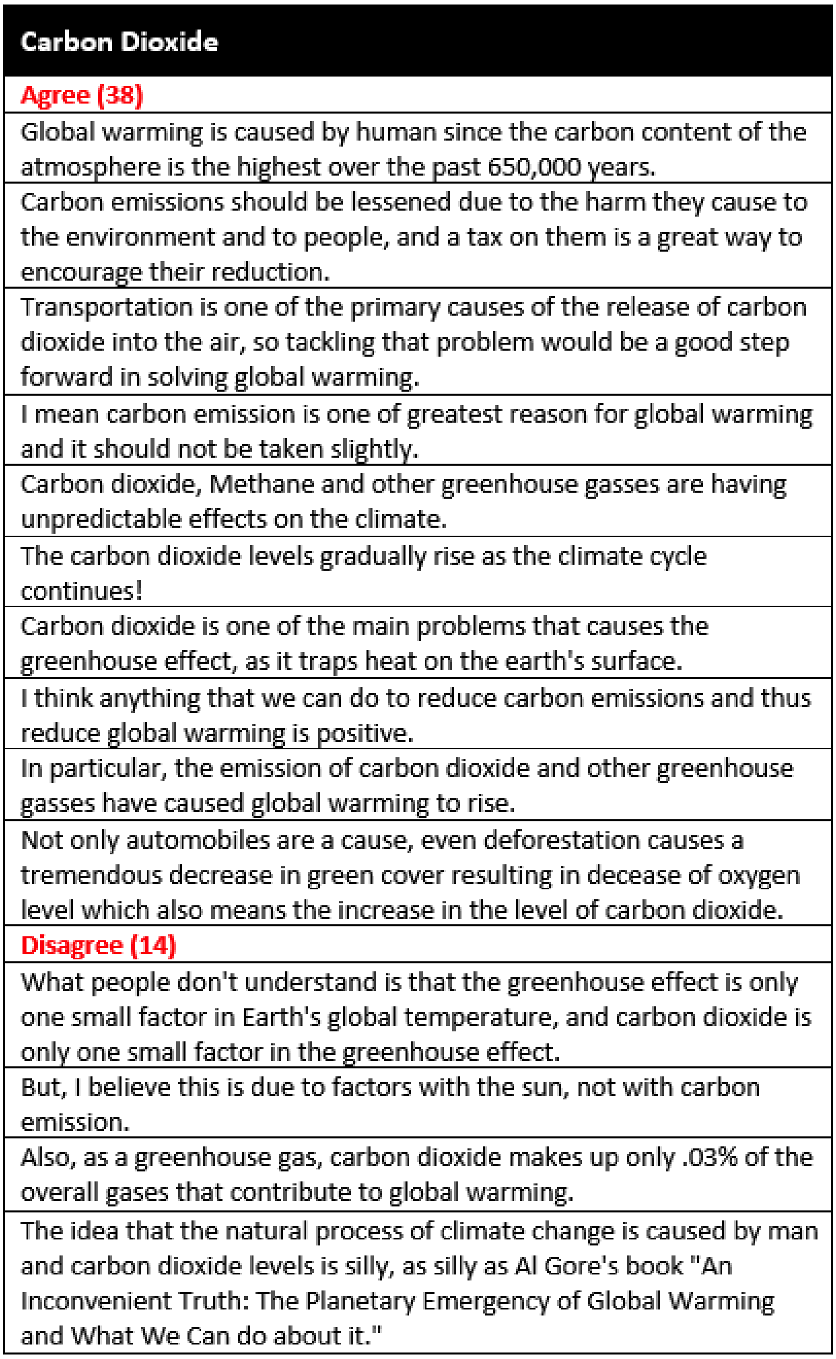
From the data described in the previous section, we manually extracted salient sentences by using the frequent topics as the keywords. Once the sentences from each opposing side were selected they were mapped to the different summary designs. We constructed four summary designs: a Chart Summary, a Table Summary, a Side-By-Side Summary and a Conceptual Map. We also constructed the combined versions of those summary designs. In total, there are 7 summary designs used in this study.

3.3.1 ***Chart Summary***

The Chart Summary is shown in Figure 3.2. It was first reported by Liu et al. (2005). From the figure, we can see that it shows the frequent topics that are discussed in debate data, in high level. The numbers indicate the frequency of the salient sentences that agrees or disagrees with particular frequent topics (see Section 3.2.2). The labels on the bars in the chart are the names of groups of salient sentences which indicate the central meaning of the groups.

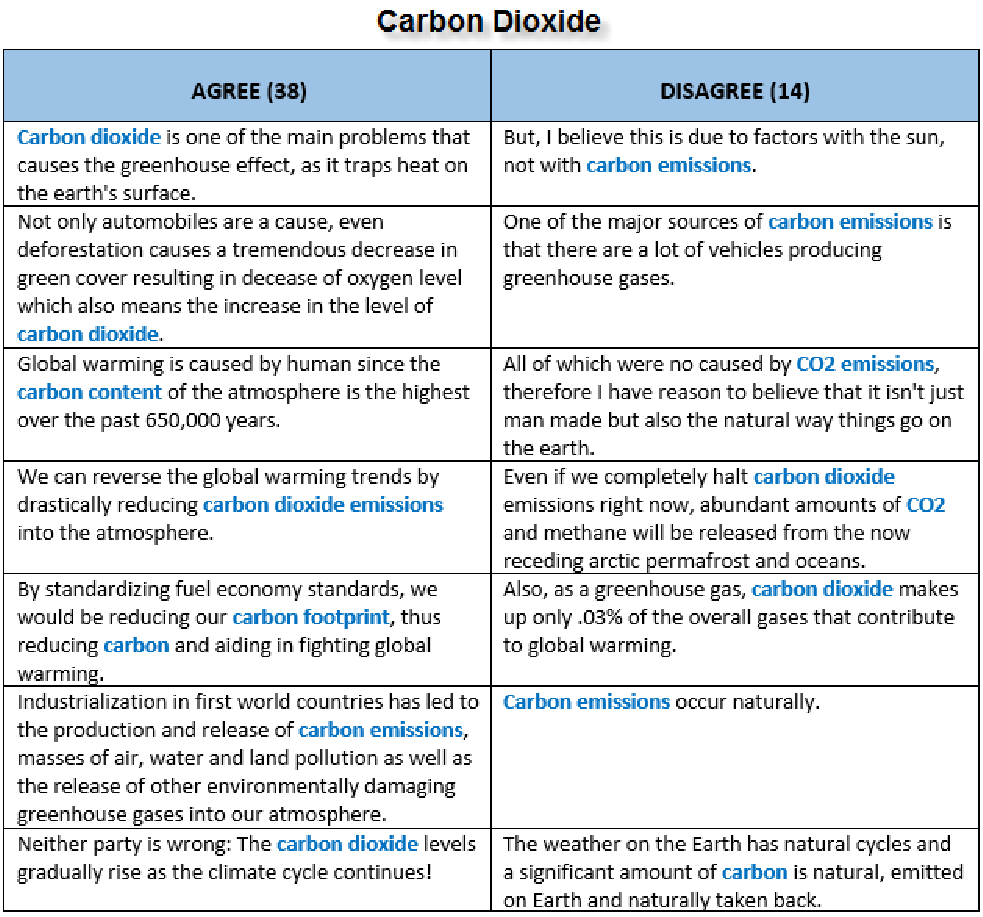
3.3.2 ***Table Summary***

Several systems present their summaries in a table format such as in Lin (1999). In our work, we adopt it to represent summaries for climate change debates and call it a *Table Summary*. A Table Summary mentions only one primary topic. The rows in the table are the salient sentences expressing different opinions about a frequent topic from both opposing sides, Agree and Disagree. As shown in Figure 3.3, the table shows an example of a *Carbon Dioxide* topic. The numbers indicate the frequencies of the salient sentences that support the topics expressed on each opposing side.



Carbon Dioxide
Agree (38)
Global warming is caused by human since the carbon content of the atmosphere is the highest over the past 650,000 years.
Carbon emissions should be lessened due to the harm they cause to the environment and to people, and a tax on them is a great way to encourage their reduction.
Transportation is one of the primary causes of the release of carbon dioxide into the air, so tackling that problem would be a good step forward in solving global warming.
I mean carbon emission is one of greatest reason for global warming and it should not be taken slightly.
Carbon dioxide, Methane and other greenhouse gasses are having unpredictable effects on the climate.
The carbon dioxide levels gradually rise as the climate cycle continues!
Carbon dioxide is one of the main problems that causes the greenhouse effect, as it traps heat on the earth's surface.
I think anything that we can do to reduce carbon emissions and thus reduce global warming is positive.
In particular, the emission of carbon dioxide and other greenhouse gasses have caused global warming to rise.
Not only automobiles are a cause, even deforestation causes a tremendous decrease in green cover resulting in decease of oxygen level which also means the increase in the level of carbon dioxide.
Disagree (14)
What people don't understand is that the greenhouse effect is only one small factor in Earth's global temperature, and carbon dioxide is only one small factor in the greenhouse effect.
But, I believe this is due to factors with the sun, not with carbon emission.
Also, as a greenhouse gas, carbon dioxide makes up only .03% of the overall gases that contribute to global warming.
The idea that the natural process of climate change is caused by man and carbon dioxide levels is silly, as silly as Al Gore's book "An Inconvenient Truth: The Planetary Emergency of Global Warming and What We Can do about it."

Figure 3.3: Table Summary



Carbon Dioxide

AGREE (38)	DISAGREE (14)
Carbon dioxide is one of the main problems that causes the greenhouse effect, as it traps heat on the earth's surface.	But, I believe this is due to factors with the sun, not with carbon emissions.
Not only automobiles are a cause, even deforestation causes a tremendous decrease in green cover resulting in decease of oxygen level which also means the increase in the level of carbon dioxide.	One of the major sources of carbon emissions is that there are a lot of vehicles producing greenhouse gases.
Global warming is caused by human since the carbon content of the atmosphere is the highest over the past 650,000 years.	All of which were no caused by CO2 emissions, therefore I have reason to believe that it isn't just man made but also the natural way things go on the earth.
We can reverse the global warming trends by drastically reducing carbon dioxide emissions into the atmosphere.	Even if we completely halt carbon dioxide emissions right now, abundant amounts of CO2 and methane will be released from the now receding arctic permafrost and oceans.
By standardizing fuel economy standards, we would be reducing our carbon footprint, thus reducing carbon and aiding in fighting global warming.	Also, as a greenhouse gas, carbon dioxide makes up only .03% of the overall gases that contribute to global warming.
Industrialization in first world countries has led to the production and release of carbon emissions, masses of air, water and land pollution as well as the release of other environmentally damaging greenhouse gases into our atmosphere.	Carbon emissions occur naturally.
Neither party is wrong: The carbon dioxide levels gradually rise as the climate cycle continues!	The weather on the Earth has natural cycles and a significant amount of carbon is natural, emitted on Earth and naturally taken back.

Figure 3.4: Side-By-Side Summary

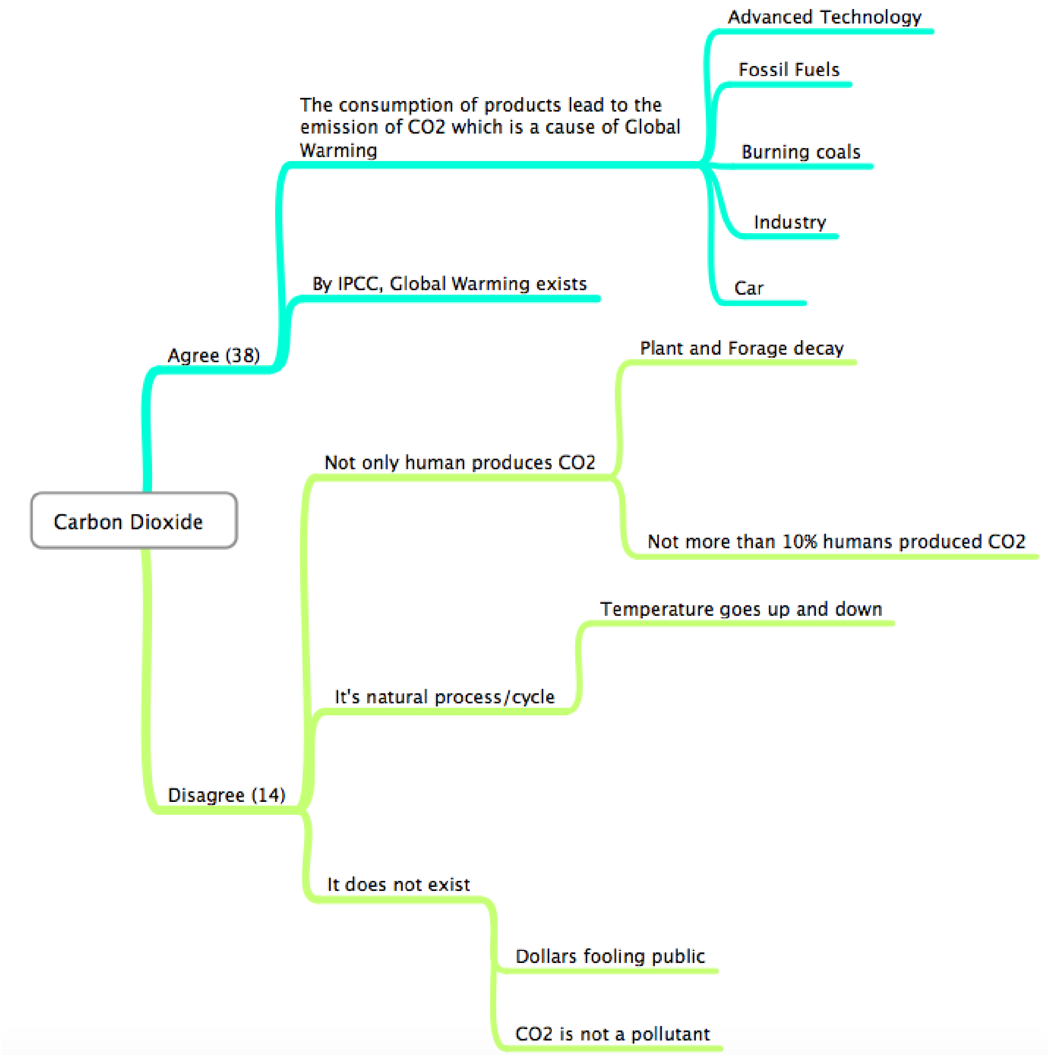

Carbon Dioxide
Agree (38)
The consumption of products lead to the emission of CO2 which is a cause of Global Warming
Advanced Technology
Fossil Fuels
Burning coals
Industry
Car
By IPCC, Global Warming exists
Disagree (14)
Not only human produces CO2
Plant and Forage decay
Not more than 10% humans produced CO2
It's natural process/cycle
Temperature goes up and down
It does not exist
Dollars fooling public
CO2 is not a pollutant

Figure 3.5: Conceptual Map

3.3.3 ***Side-By-Side Summary***

Another summary design is a Side-By-Side Summary. It is adopted from the work presented by Paul et al. (2010). Similar to the Table Summary, the Side-By-Side Summary only shows one topic at a time. As shown in Figure 3.4, the Side-By-Side Summary contains pairs of Agree and Disagree sentences in which each pair mentions the same topic (i.e. Carbon Dioxide) – one sentence is from the Agree side and the other is from the Disagree side. A pair is called *rebuttal.* The numbers in the brackets show the frequency of the salient sentences that have been mentioned on each opposing side. The content shown in the table is only a list of rebuttals.

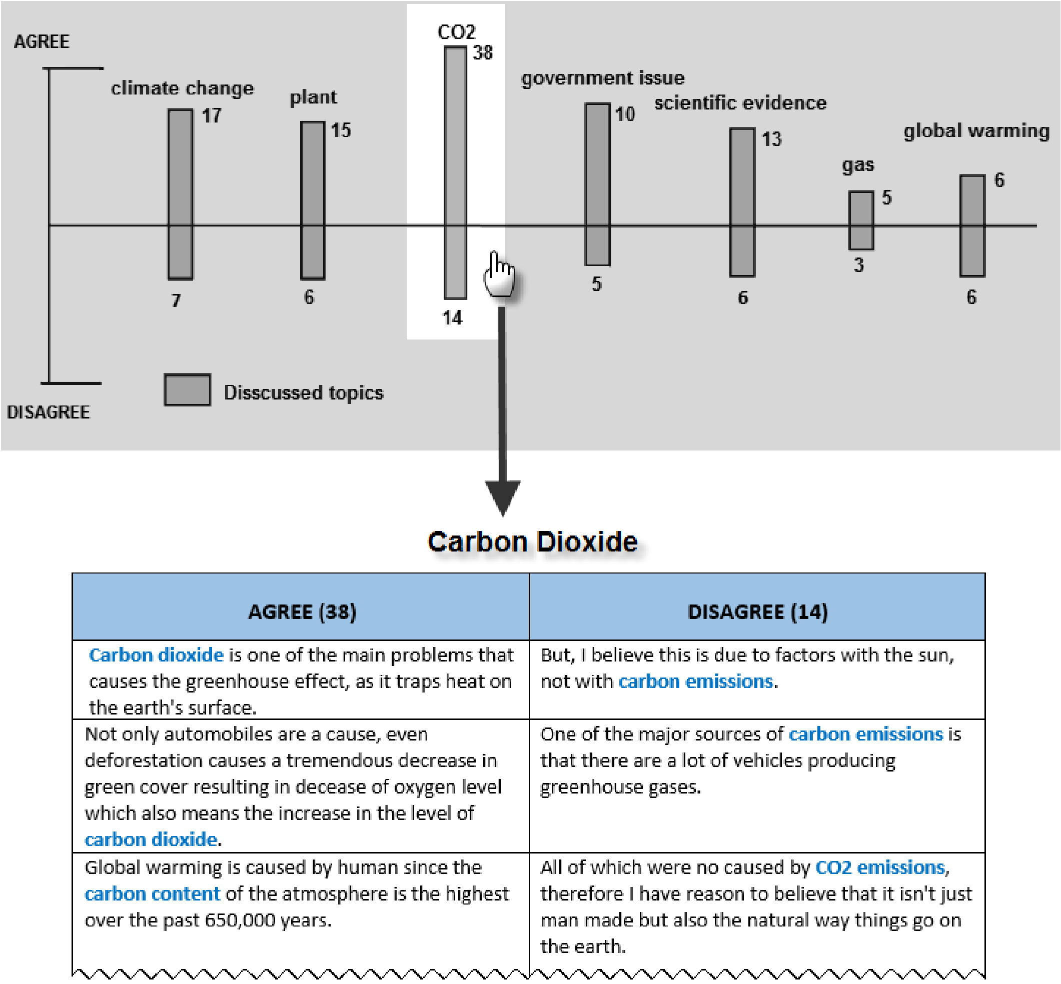
To construct a rebuttal, we manually matched two salient sentences from each opposing side which have the closest meaning, but opposite direction of the opinions. In other words, the two sentences are contradictory. For instance, in the Side-By-Side Summary shown in Figure 3.4, one sentence mentions that carbon dioxide is the main problem that causes global warming, but the other sentence argues that it is because of the sun.

3.3.4 ***Conceptual Map***

A Conceptual Map is a graphical representation of ideas, usually enclosed in circles or boxes. A connection of circles or boxes is drawn by a line or an arrow, which presents the relationship between ideas (Novak and Cañas, 2006). We applied this concept and redesigned a Conceptual Map to represent a summary of the existence of global warming issue. Similar to the Table Summary and the Side-By-Side Summary, the Conceptual Map only presents one topic at a time.

As shown in Figure 3.5, the opinions of public responses, regarding a *Carbon Dioxide* topic causing the global warming, are separated into two opposing sides, Agree and Disagree. On both opposing sides, people mention arguments to support their opinions about carbon dioxide. Each branch of the side shows the main category of a topic. The subordinated branches contain additional arguments to support the main category.

In the design, we only show how an annotator understands the content regarding a particular topic and manually summarize the content a conceptual map. As shown in Figure 3.5, the Conceptual Map was manually constructed by determining the information expressed on *Carbon Dioxide*. The node in the graph is split when additional information is elaborated[4]. From the figure, the creation of sub-branches is to give an additional information about the Carbon Dioxide topic expressed in the debate. When additional detail of Carbon Dioxide is found, a sub-branch is created (i.e. the sub-branch "the consumption of products leading to the emission of Carbon Dioxide"). Deeper sub-branches which elaborate the previous sub-branch are constructed until no elaboration is found.



Carbon Dioxide

AGREE (38)	DISAGREE (14)
Carbon dioxide is one of the main problems that causes the greenhouse effect, as it traps heat on the earth's surface.	But, I believe this is due to factors with the sun, not with carbon emissions.
Not only automobiles are a cause, even deforestation causes a tremendous decrease in green cover resulting in decease of oxygen level which also means the increase in the level of carbon dioxide.	One of the major sources of carbon emissions is that there are a lot of vehicles producing greenhouse gases.
Global warming is caused by human since the carbon content of the atmosphere is the highest over the past 650,000 years.	All of which were no caused by CO2 emissions, therefore I have reason to believe that it isn't just man made but also the natural way things go on the earth.

Figure 3.6: The combination of a Chart Summary and a Side-By-Side Summary

[4] Note that in the design, we did not carefully define the conditions of when and how many sub nodes should be split. This is an important concern in future work.

3.3.5 ***Combination of Summary Designs***

The Chart Summary as shown in Figure 3.2 is an abstract representation of topics. It does not provide full details of opinions expressed on topics whereas the other three summary designs provide evidential sentences about different opinions. Therefore, one possible way to present summaries is to combine the abstract chart with a more detailed summary. For instance, a combination of a Chart Summary and another detailed summary design would benefit readers to have a high-level summary and a detailed summary. If a reader is interested in further details, he can click on one of the chart bars (topics) to obtain more details. The detailed summary can be displayed as one of the other three summary designs. Figure 3.6 illustrates a combination of summary designs, namely the Chart Summary combined with the Side-By-Side Summary. In the figure, the topic CO_2 is highlighted (simulating the case where a user has clicked that topic). This activates the Side-By-Side Summary and shows rebuttals for the activated topic. The idea of the combination is also applied to the Table Summary and the Conceptual Map. The combination of the Chart Summary and the Table Summary, the Chart Summary and the Side-By-Side Summary, and the Chart Summary and the Conceptual Map are called Combination 1, Combination 2 and Combination 3 respectively.

3.4 The Empirical Study

To collect user preferences for the seven different summary designs we recruited 60 participants to an online questionnaire advertised via Facebook, Twitter, and the Pantip discussion forum[5]. This work is volunteering and without pay. Table 3.2 - 3.6 show the demographic data of the participants. Of these, female participants completed 51.7% and male participants completed 48.3%. In approximates, 18.3% of the participants were in the age ranges between 18-24, 60% of the participants ranged between 25 and 34 years of age, 13.3% of the participants ranged between 35 and 44 years of age, 3.3% were in the range between 45 and 55 years old, and 1.7% were above 55 years old. The demographic data on educational levels reported that the majority of the participants had post-graduate university degrees and undergraduate university degrees with the approximation of 61.7% and 23.3% respectively. In addition, the majority of the ethnic group had Asian (Other) backgrounds which amount to 75%. The information on job

[5] `http://www.pantip.com/`

functions reported that about30% of the participants were students and approximately 13.3% worked in academic areas and other related fields.

In the questionnaire, the participants were asked to read a portion of a debate article similar to Figure 3.1, which contains two sets of comments with opposing opinions on the existence of global warming. Next, the seven different summary designs and their descriptions were shown to the participants. The participants were asked to read and understand each summary design. Then, each summary design along with a list of questions was shown. They were asked to give opinions, answer questions, and specify preference scores to rate each summary design. Five-point Likert scales were used: excellence (5), good (4), fair (3), poor (2) and very poor (1). The questions below illustrate example questions used in the study. The first three questions are Likert-Scale questions and the last two questions are the open-ended questions.

1. By reading the summary in the XXX[a], is it easy to follow ideas in debate article?

2. How much is the XXX suitable for debate data?

3. Overall, please specify your preference on the XXX.

4. What do you think is the best part of the XXX?

5. What do you think is the worst part of the XXX?

[a] XXX refers to the name of summary design.

Table 3.2: Demographic data on genders

Genders	**Frequency**	**Percent**
Female	31	51.7
Male	29	48.3
Total	60	100.0

Table 3.3: Demographic data on age ranges

Age Ranges	**Frequency**	**Percent**
Prefer not to say	2	3.3
18-24 years old	11	18.3
25-34 years old	36	60.0
35-44 years old	8	13.3
45-54 years old	2	3.3
55 or above	1	1.7
Total	60	100.0

Table 3.4: Demographic data on educational levels

Educational Levels	**Frequency**	**Percent**
Prefer not to say	2	3.3
GCSE or equivalent qualification	1	1.7
A-Level or equivalent qualification	4	6.7
NVQ and/or other professional qualifications	2	3.3
Undergraduate university degree	14	23.3
Post-graduate university degree	37	61.7
Total	60	100.0

Table 3.5: Demographic data on ethnicity

Ethnic Groups	**Frequency**	**Percent**
Prefer not to say	1	1.7
White (British)	3	5.0
Asian (Other)	45	75.0
White (European)	6	10.0
Asian (British)	1	1.7
Asian (Chinese)	4	6.7
Total	60	100.0

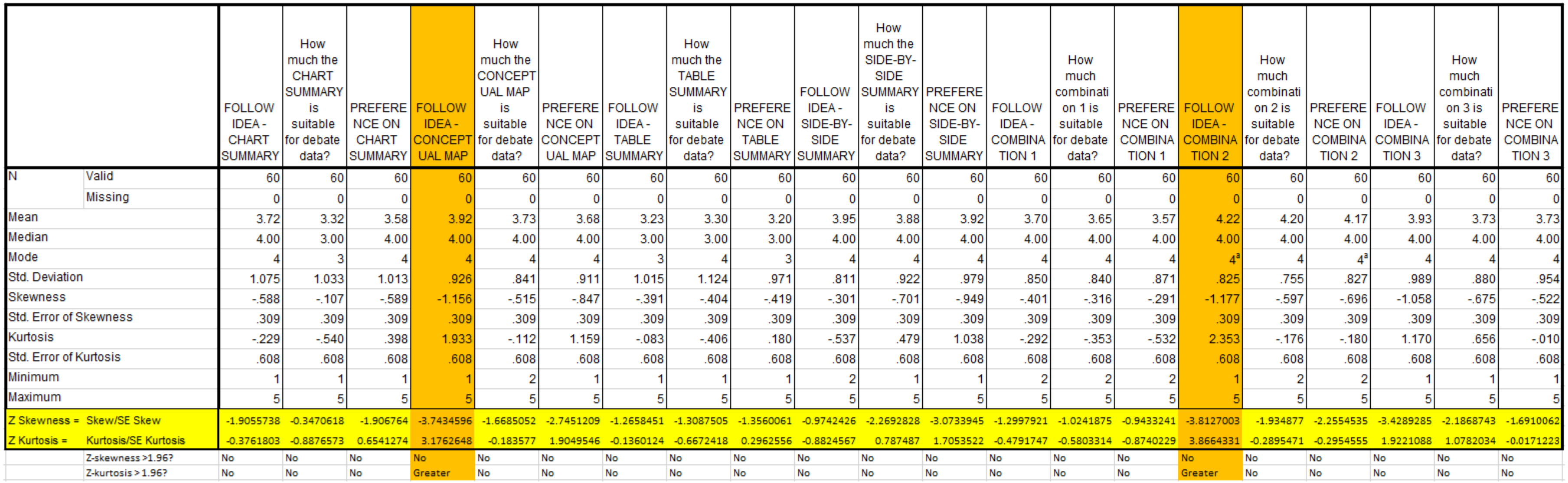
Table 3.6: Demographic data on job functions

Job Functions	**Frequency**	**Percent**
Accounting	1	1.7
Banking / Finance	1	1.7
Design	2	3.3
Education (Lecturer, researcher, etc)	8	13.3
Engineering	3	5.0
Information Technology (IT)	2	3.3
Insurance	3	5.0
Management	2	3.3
Manufacturing	1	1.7
Marketing / Public Relations	1	1.7
Out of work and looking for work	1	1.7
Prefer not to say	2	3.3
Professional Services	2	3.3
Public / Civil	1	1.7
Student	18	30.0
Transportation and Logistics	1	1.7
Unable to work	1	1.7
Total	60	100.0

3.5 Pre-Data Analysis

3.5.1 Correlation Coefficient Selection

There were 60 participants who answered the online questionnaire. We used IBM SPSS Statistics for the data analysis. To prevent misleading results, we examined how our data is distributed and then chose an appropriate correlation coefficient. Bachman (2004) explained that skewness and kurtosis values can be used to determine data distributions. While Pearson correlation is suitable for normal data distributions, Spearman correlation works appropriately for non-normal data distributions. Figure 3.7 illustrates the data distribution test using Z-score for skewness and kurtosis Z-score for skewness. Z-score for skewness is the proportion of skewness and the error of skewness. Z-score for kurtosis is the ratio of kurtosis to the error of kurtosis. A Z-score value of a variable above 1.96 indicates non-normal distribution (Ghasemi and Zahediasl, 2012). On this basis, we conclude that Spearman correlation is more suitable for our data analysis.



		FOLLOW IDEA - CHART SUMMARY	How much the CHART SUMMARY is suitable for debate data?	PREFERENCE ON CHART SUMMARY	FOLLOW IDEA - CONCEPTUAL MAP	How much the CONCEPTUAL MAP is suitable for debate data?	PREFERENCE ON CONCEPTUAL MAP	FOLLOW IDEA - TABLE SUMMARY	How much the TABLE SUMMARY is suitable for debate data?	PREFERENCE ON TABLE SUMMARY	FOLLOW IDEA - SIDE-BY-SIDE SUMMARY	How much the SIDE-BY-SIDE SUMMARY is suitable for debate data?	PREFERENCE ON SIDE-BY-SIDE SUMMARY	FOLLOW IDEA - COMBINATION 1	How much combination 1 is suitable for debate data?	PREFERENCE ON COMBINATION 1	FOLLOW IDEA - COMBINATION 2	How much combination 2 is suitable for debate data?	PREFERENCE ON COMBINATION 2	FOLLOW IDEA - COMBINATION 3	How much combination 3 is suitable for debate data?	PREFERENCE ON COMBINATION 3
N	Valid	60	60	60	60	60	60	60	60	60	60	60	60	60	60	60	60	60	60	60	60	60
	Missing	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Mean		3.72	3.32	3.58	3.92	3.73	3.68	3.23	3.30	3.20	3.95	3.88	3.92	3.70	3.65	3.57	4.22	4.20	4.17	3.93	3.73	3.73
Median		4.00	3.00	4.00	4.00	4.00	4.00	3.00	3.00	3.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00
Mode		4	3	4	4	4	4	3	4	3	4	4	4	4	4	4	4^a	4	4^a	4	4	4
Std. Deviation		1.075	1.033	1.013	.926	.841	.911	1.015	1.124	.971	.811	.922	.979	.850	.840	.871	.825	.755	.827	.989	.880	.954
Skewness		-.588	-.107	-.589	-1.156	-.515	-.847	-.391	-.404	-.419	-.301	-.701	-.949	-.401	-.316	-.291	-1.177	-.597	-.696	-1.058	-.675	-.522
Std. Error of Skewness		.309	.309	.309	.309	.309	.309	.309	.309	.309	.309	.309	.309	.309	.309	.309	.309	.309	.309	.309	.309	.309
Kurtosis		-.229	-.540	.398	1.933	-.112	1.159	-.083	-.406	.180	-.537	.479	1.038	-.292	-.353	-.532	2.353	-.176	-.180	1.170	.656	-.010
Std. Error of Kurtosis		.608	.608	.608	.608	.608	.608	.608	.608	.608	.608	.608	.608	.608	.608	.608	.608	.608	.608	.608	.608	.608
Minimum		1	1	1	1	2	1	1	1	1	2	1	1	2	2	2	1	2	2	1	1	1
Maximum		5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
Z Skewness =	Skew/SE Skew	-1.9055738	-0.3470618	-1.906764	-3.7434596	-1.6685052	-2.7451209	-1.2658451	-1.3087505	-1.3560061	-0.9742426	-2.2692828	-3.0733945	-1.2997921	-1.0241875	-0.9433241	-3.8127003	-1.934877	-2.2554535	-3.4289285	-2.1868743	-1.6910062
Z Kurtosis =	Kurtosis/SE Kurtosis	-0.3761803	-0.8876573	0.6541274	3.1762648	-0.183577	1.9049546	-0.1360124	-0.6672418	0.2962556	-0.8824567	0.787487	1.7053522	-0.4791747	-0.5803314	-0.8740229	3.8664331	-0.2895471	-0.2954555	1.9221088	1.0782034	-0.0171223
	Z-skewness >1.96?	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
	Z-kurtosis > 1.96?	No	No	No	Greater	No	No	No	No	No	No	No	No	No	No	No	Greater	No	No	No	No	No

Figure 3.7: Data distribution test with Z-Score

Table 3.7: Descriptive statistics of the questions toward each summary design

By reading the summary in the summary design, is it easy to follow ideas in debate article?

Summary Designs	Mean	Median	Mode	SD.	Min	Max
Chart Summary	3.72	4	4	1.075	1	5
Conceptual Map	3.92	4	4	.926	1	5
Table Summary	3.23	3	3	1.015	1	5
Side-By-Side Summary	3.95	4	4	.811	2	5
Combination 1	3.70	4	4	.850	2	5
Combination 2	**4.22**	4	4[a]	.825	1	5
Combination 3	3.93	4	4	.989	1	5

How much the summary design is suitable for debate data?

Summary Designs	Mean	Median	Mode	SD.	Min	Max
Chart Summary	3.32	3	3	1.033	1	5
Conceptual Map	3.73	4	4	.841	2	5
Table Summary	3.30	3	4	1.124	1	5
Side-By-Side Summary	3.88	4	4	.922	1	5
Combination 1	3.65	4	4	.840	2	5
Combination 2	**4.20**	4	4	.755	2	5
Combination 3	3.73	4	4	.880	1	5

Overall, please specify your preference on the summary design.

Summary Designs	Mean	Median	Mode	SD.	Min	Max
Chart Summary	3.58	4	4	1.013	1	5
Conceptual Map	3.68	4	4	.911	1	5
Table Summary	3.20	3	3	.971	1	5
Side-By-Side Summary	3.92	4	4	.979	1	5
Combination 1	3.57	4	4	.871	2	5
Combination 2	**4.17**	4	4[a]	.827	2	5
Combination 3	3.73	4	4	.954	1	5

a. Multiple modes exist. The smallest value is shown

3.6 Results and Analysis

3.6.1 Quantitative Results

The descriptive statistics of the empirical study shown in Table 3.7 justify the conclusion that, the Combination 2, the combination of the Chart Summary and the Side-By-Side Summary, is the best one in representing the idea in the debate article, the most suitable one for representing debate content, and the most preferred summary design. For instance, the statistical information for the third question shows that the Combination 2 is the most preferred summary design. It has the highest means score of 4.22. This is further supported by the standard deviation. It has a lower value than of the other summary designs (0.825) showing that individual responses are closer to the mean. This also applies to other questions.

Moreover, we also conducted statistical tests using the Kruskal-Wallis tests to determine if there is any statistical difference between the Combination 2 and the other summary designs. We conducted the tests for the first three questions. In the first question, the Kruskal-Wallis test indicates that there is a statistical difference between the Combination 2 and the other summary designs, χ^2 (6, n = 60) = 51.453, p < .001. Also in the second question, χ^2 (6, n = 60) = 41.094, p < .001, reveals a statistical difference. Similarly, in the last question, χ^2 (6, n = 60) = 37.039, p < .001 indicates there is a statistical difference as well. For these reasons, there is a statistical difference between the Combination 2 and the other designs.

According to the descriptive statistics evidence and the results of the statistical test, we therefore conclude that the Combination 2, the combination of the Chart Summary and the Side-By-Side Summary is the most preferred output for representing the abridged version of debate content.

3.6.2 Qualitative Results

The qualitative comments that participants were asked to provide along with the Likert scores reflect the quantitative results. Participants were asked to elaborate the least and the most advantages of each summary design.

Positive feedback for the Chart Summary primarily focused on the concise information that the chart provides. Participants can see a clear summary at the first glance. Some points of views from our participants were “The chart can represent the overall picture of the debate topic very well.”, “Picture: easy to understand and eliminate a lot of texts”, and “It is an option to see the content of an article at a glance”. However, we found that due to its conciseness the Chart Summary cannot provide enough information. It is unable to identify subordinated topics mentioned in debates. Readers may instantly jump to the conclusion without reading the content behind. Some participants mentioned in the study that “The chart does not provide any detail why they agree or disagree.”, “Lack of details. The presenter cannot identify the sub-debated topics under each issue.”, and “Opinions and argumentation are not shown”.

Participants praised the Table Summary as giving detailed summaries of the debate and showing a clear division between Agree and Disagree information. “Full of details from each side.” and “The augmentations are split up into two categories, it’s very clear and easy to use.” were the opinions from our participants. Conversely, the Table Summary is too deep in detail which takes time for readers to make comparisons for each argument. Some examples of the opinions are that “Too much data. It couldn’t count as the summary. It is an essay.”. Another viewpoint is “It’s a bit slow to read and hard to make a comparison on each. It’s too much wording and difficult to follow.”.

In general, the advantages of the Conceptual Map focused on its readability. Participants viewed that “Key points of the topic are shown in a very easy to read and tidy way.”, “Readers might want to know details briefly but not too big paragraph”. In contrast, the disadvantages are “It is not so clear to a quick look. If I did not know what was this article about, I would need more time to get the correct picture.”, “Might be hard to read when there are more branches in the map.”, and “It’s not so immediate for the comparison between each argumentation.”.

The positive feedback on the Side-By-Side Summary focused on the comparison between issues and readability. The example standpoints of participants are “Easy comparison, quite concise, points laid out in a logical order” and “Compare to the previous summary. It is easy to follow agree/disagree opinion as I can see it side by side. This is the most useful summary for me and this is well-arranged". Participants rarely provided negative feedback for this summary. Few comments mentioned that the Side-By-Side Summary contains a long list of rebuttals which takes time to read.

Participants argued that the Combination 1 (the combination of the Chart Summary and the Table summary) is better than just the chart itself. For example, one feedback mentioned that "It is good to have details on the chart". Still, the deep details and long representation of the Table Summary are the drawbacks of this combination. A participant said that "Still too long to be called a summary".

The positive feedback on the Combination 3 (the combination of the Chart Summary and the Conceptual Map) was similar to the feedback on the Chart Summary only. The participants commented that it is simple and concise to read. However, it is less informative compared to other summary designs. The participants indicated that the Conceptual Map is limited in providing details and thus combining it with the abstract Chart Summary does not make the Combination 3 detailed enough. For instance, participants commented that "Sometimes the conceptual map is complex, especially, when the sub-issues are varied. Lacking in detail compared to previous combinations.", "Less informative than previous ones overall.", and "Not easy to read and understand".

In general, participants agreed that Combination 2 (the combination of the Chart with the Side-By-Side Summary) provides a good insight into topics and is a helpful alternative to follow the discussion of debates line by line. This side-by-side visualization helps readers compare the logic and the fact in each debate. Another qualitative feedback is that Combination 2 also provides the high-level summary and the detailed summary of each debate which provides readers clear discussion and simplicity to follow the discussion. For example, participants mentioned that "It is better arranged than combination 1, but still requires more action to see details (need to click to see the detailed summary). However, it is good option to have a chart and details as well.", "Contains high-level summary and details highlighted by keywords.", and "Easy to follow, the logical order of points.". Negative feedback on the Side-By-Side Summary was rarely found. Only a few comments mentioned that a long list of rebuttals takes a long time to read.

3.7 Summary

Currently, there is no analysis about which summary representation for debate summaries is preferred by human readers. We have empirically investigated which summary designs humans prefer, an important question for automatically generated summaries of debates in online forums. To answer our research question, *Which summary design is the most preferred for presenting the abridged version of debate content?*, we conducted an empirical study by recruiting 60 participants to give preference scores for each summary design. Our results indicated that the Chart Summary combined with the Side-By-Side Summary is the most preferred summary design for presenting the summary of debate content. Our hypothesis test indicated that there is a statistical difference in the user preferences among the summary designs. Moreover, in this study, we proposed a novel summary representation that represents the summary of debate contents in a Conceptual Map. Even though it is not the most favored one, it has received some positive feedback from the participants.

As Combination 2 is the most preferred summary design, it will be the output of our debate summarization system. To generate this summary, in the next chapter, we introduce a system architecture of our debate secularization system. Additionally, we also introduce a system for collecting reference summaries and how the reference summaries are annotated. Examples of online debates, the statistical information of the collected data, and the inter-annotator agreement are discussed.

Chapter 4

SYSTEM ARCHITECTURE

The previous chapter discussed an empirical study of human preferences for summary designs. The results indicated that the most preferred summary design for summarizing online debates is Combination 2. This chapter paves the way for the generation of Combination 2 summaries by introducing the system architecture. The structure of the system begins with the discussion of the automatic selection of salient sentences in debate comments. Then they will be the input for the generation of the Chart Summary element of the Combination 2 summary. In Chapter 5, we describe how each component of the Chart Summary is generated and combined as a Chart Summary. The Chart Summary consists of bars which represent clusters of salient sentences, labels of the bars which give short descriptions of the bars, and the frequencies of the bars showing the numbers of particular topics discussed in each bar. The combination of these components constitutes a Chart Summary. In the final process, we define the generation of Side-By-Side Summary as a contradiction detection task. The system will classify whether the given pairs of sentences contradict each other. Only pairs considered by the system as contradictory will be included in the Side-By-Side Summary. The generation of Side-By-Side Summary will be introduced in Chapter 6.

4.1 Debate Summarization System Architecture

Debate summarization is one of the novel research areas in automatic text summarization which has been largely unexplored. The summary of the related work includes *Contrastive Summarization*, *Comparative Summarization*, and *Debate Stance Recognition*. Contrastive Summarization is the study of generating the summary for two entities and finding the difference in sentiments among them (Lerman and McDonald, 2009). This kind of summarization requires the classification of polarity in order to "contrast" opinions expressed in different sentiments. Comparative Summarization aims to find the difference between two comparable entities so that sentiment classification may not be required (Campr and Jezek, 2012). Debate Stance Recognition aims to detect stance of opinions' holders in the text (Somasundaran and Wiebe, 2009). For instance, in a debate topic of the existence of global warming issue in which people could agree or disagree with the issue. Thus, the stance of this debate can be either *agree* or *disagree*.

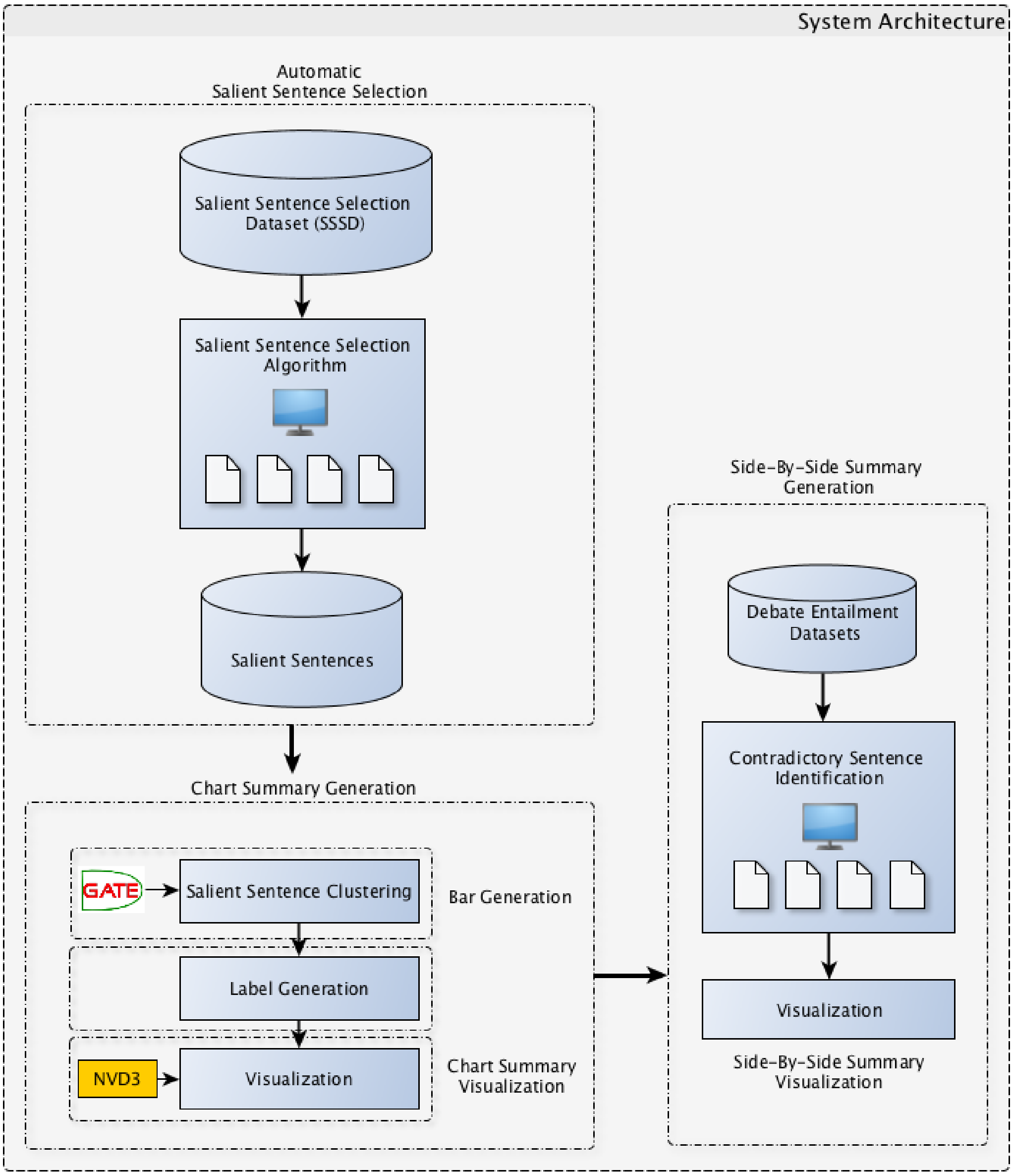

System Architecture
Automatic Salient Sentence Selection
Salient Sentence Selection Dataset (SSSD)
Salient Sentence Selection Algorithm
Salient Sentences
Chart Summary Generation
GATE
Salient Sentence Clustering
Bar Generation
Label Generation
NVD3
Visualization
Chart Summary Visualization
Side-By-Side Summary Generation
Debate Entailment Datasets
Contradictory Sentence Identification
Visualization
Side-By-Side Summary Visualization

Figure 4.1: The system architecture blueprint for developing the debate summarization system

Currently, there is only one work on debate text summarization and another relevant work exploring the detection of arguments, aiming for generating extractive summaries in their future work. As aforementioned, in Ranade et al. (2013), system summaries are extracted by ranking the smallest units of debates, called Dialogue Acts (DAs). The ranking of sentences is based on some features including, words in DAs that is co-occurring in debate topic, topics with opinions expressed on it, sentence position, and sentence length features. However, this work does not explicitly highlight what is the key content to be summarized and how the debate summary is presented and visualized. This is different to our work. On the other hand, in our work, we highlight the summarization of key content in debate and visualize them to be easily accessed by users. As well as the work presented by Trabelsi and Zaiane (2014), they explore the mining of arguments and will explore clustering approaches for generating extractive summaries for their future work. They did not explicitly determine contradicted arguments in a side-by-side comparison. This leaves the gap for us to propose a novel system architecture for tackling the problem of online debate summarization.

To tackle the problem of debate summarization, we propose a novel system architecture as shown in Figure 4.1. It consists of three primary processes which are Automatic Salient Sentence Selection, Chart Summary Generation, and Side-By-Side Summary Generation. The following sections discuss those processes in detail.

4.1.1 ***Automatic Salient Sentence Selection***

The aim of this process is to automatically select salient sentences from debate comments. The salient sentences are considered as containing the most important information in the comments. In short, they are the summaries of the debate comments. In this section, we begin with how a compression rate is defined for the purpose of summarizing online debates. In a later section, we introduce how a gold standard dataset of online debates is created based on the defined compression rate. Later on, we explain how the salient sentences will be automatically selected from debate comments. We present how a model together with a set of dominant features that help extract salient sentences is defined. After the salient sentences are extracted, they will be the input for the next process.

4.1.1.1 Compression Rate

The compression rate is an important part of text summarization which indicates the amount of summary to be generated from the original text. In other words, it is the proportion between the length of the summary and the original context. Compression rate influences the quality of the generated summary as it leads to the amount of information included in the generated summary. Researchers have discovered the ideal compression rate that shapes the summary to best covers all necessary information as mentioned in the original context.

Morris et al. (1992) generated a set of summaries, called *extracts*, with compression rates of 20% and 30% from a set of sample Graduate Management Aptitude Test (GMAT). The summaries were evaluated by having annotators read the summaries and answer multiple choice questions relating to the information presented in the original documents. Their answers were assessed based on how well they understand the original content in term of reading comprehension. Overall, the results indicated that the summarization of 20% of sentences is informative as the original documents.

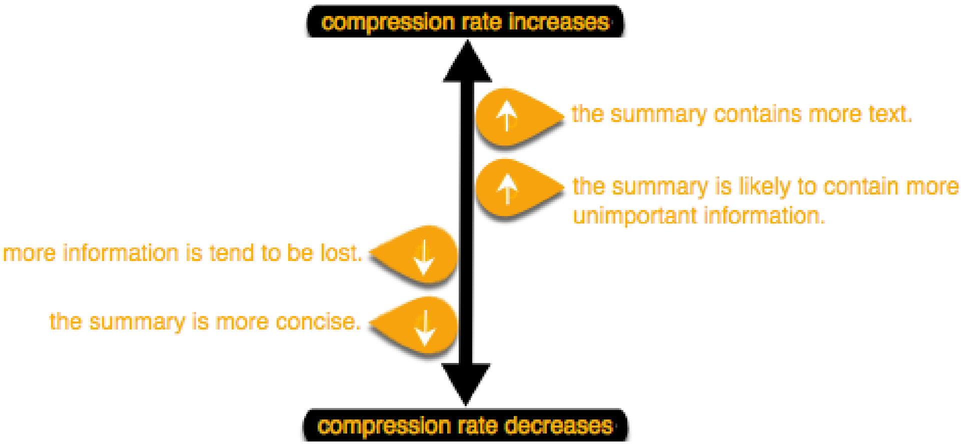

compression rate increases
the summary contains more text.
the summary is likely to contain more unimportant information.
more information is tend to be lost.
the summary is more concise.
compression rate decreases

Figure 4.2: Description of Increased and Decreased Compression Rate

Neto et al. (2002) generated extractive summaries as a classification task. The Naive Bayes algorithm and the C4.5 decision tree algorithm were predominantly used for the classification. In this work, the compression rate of 10% and 20% of the length of the original content were applied. The quality of the summaries was measured in terms of precision and recall. The authors reported that the classification with the Naive Bays classifier yielded greater results for both compression rates. From this work, the precision and recall of the 20% compression rate were higher than the one with the rate of 10%. This is common for a higher compression rate scheme as there is more possibility that a larger number of sentences is likely to match with the reference summaries.

Figure 4.2 summarizes details of the decreasing and the increasing compression rate according to the explanation of Yeh et al. (2005). The basic idea is that when the compression rate is increased, it is likely that the summary will contain more text and the summary may also contain more insignificant information. In contrast, if the compression rate is decreased, the summary is more concise and more information is tend to be lost.

According to the literature in the previous work combined with the explanation shown in Figure 4.2, a compression rate of 20% is suitable and can necessarily cover the important information expressed in the source text. We therefore focus on the summarization of online debate with a compression rate of 20%.

4.1.1.2 ***Web-Based System for Salient Sentence Annotation***

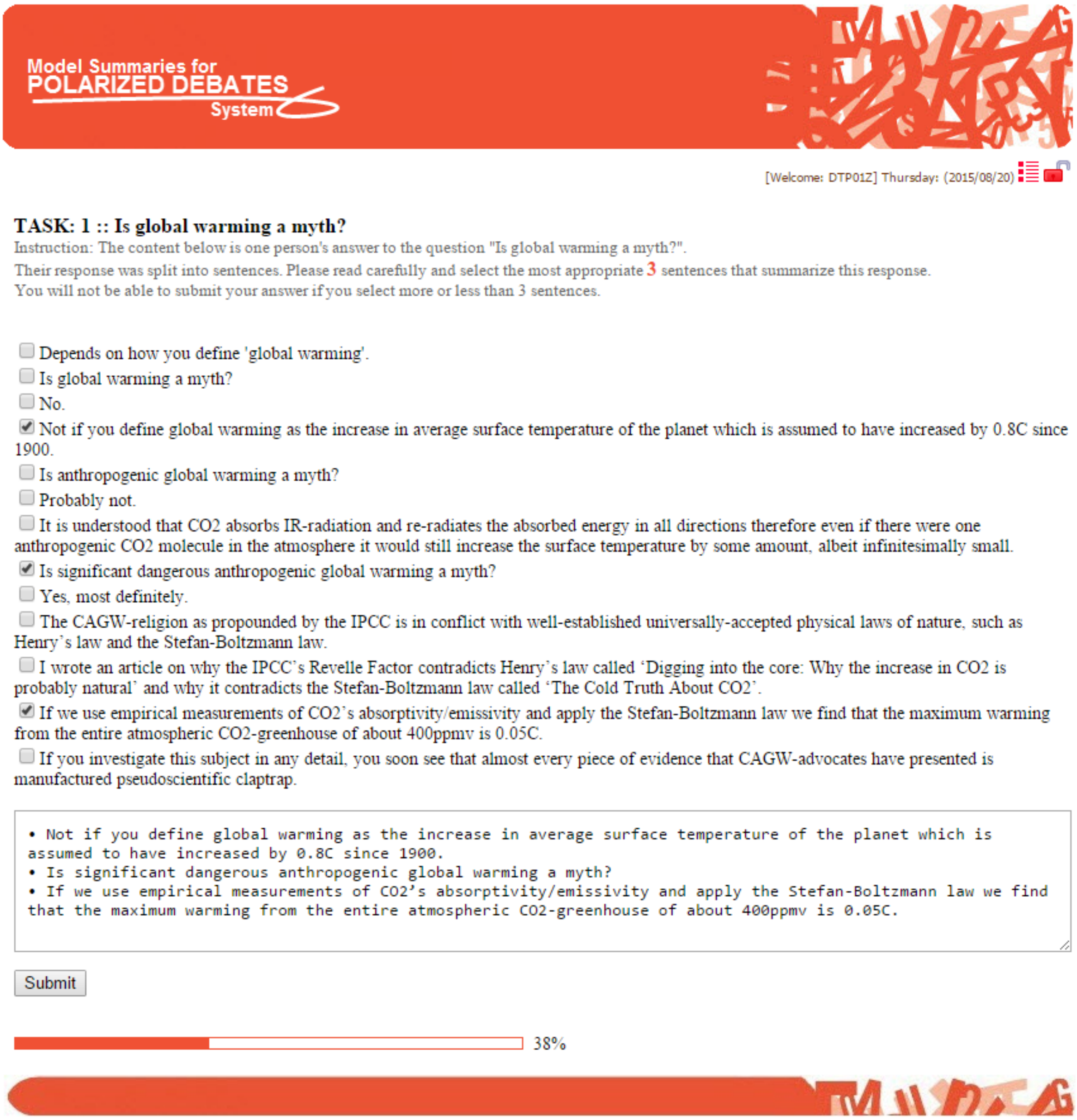
Many researchers have proposed different techniques to effectively collect data for annotation. Orăsan (2002) pointed out that a low quality of annotation results is generally caused by annotators, especially when annotators work on complicated tasks or on unfamiliar domains. They may get exhausted. To address this problem, Orăsan (2002) implemented a user-friendly annotation system integrated with semi-automatic features to help annotators analyze important sentences. Examples of the features are based on clue words, TF-IDF, and detecting similar textual units. In contrast, the drawback of the system is that annotators excessively rely on the system's suggestion so the quality of the annotations drops.

Stenetorp et al. (2012) introduced BRAT, a web-based annotation system, integrated with a machine learning-based disambiguation system to help annotators make the judgment on annotation tasks. BRAT was used for annotating data in various domains, such as cancer, Japanese verb schemes, and gene data. In the annotation task, users are required to select a span of text or double-click on a term.

Knowtator was developed to serve general-purpose text annotation schemes. A key feature of this system is ontology-based which allows to efficiently capture name entities of the domains it provides. Another highlighted feature is the report of an inter-annotator agreement which summarizes the mutual agreement of annotation results (Ogren, 2006).

Moreover, the ease of use of annotation systems also enhances annotation tasks. Orăsan (2002) stated that a good user interface allows annotators to conveniently interact with annotation systems. For example, Stenetorp et al. (2012) used drag-and-drop, text highlight, and double-clicking functions in their annotation task. Bontcheva et al. (2013) developed GATE Teamware on top of GATE (Cunningham et al., 2011) which allows project managers to create annotation projects, monitor annotation tasks, and manage different user privileges for the annotation tasks.

As aforementioned, the annotation tools above do not fulfill the requirements for our annotation objectives. For instance, BRAT requires users to select a span of text and choose on certain terms. In this circumstance, such tool is not practical as there are a large number of sentences in the comments to be shown to users. Therefore, in this thesis, we developed an annotation system, especially for annotating salient sentences in online debates. The system splits each debate comment into a list of sentences. When the sentences are selected, they will be appeared in the box below, confirming the summary of a comment created by the users. If the users want to make changes to the summary, they just simply un-select the unwanted sentences and the sentences will be removed from the box. Figure 4.3 illustrates the interface of the Salient Sentence Annotation system.



[Welcome: DTP01Z] Thursday: (2015/08/20)

TASK: 1 :: Is global warming a myth?

Instruction: The content below is one person's answer to the question "Is global warming a myth?".
Their response was split into sentences. Please read carefully and select the most appropriate 3 sentences that summarize this response.
You will not be able to submit your answer if you select more or less than 3 sentences.

☐ Depends on how you define 'global warming'.

☐ Is global warming a myth?

☐ No.

☑ Not if you define global warming as the increase in average surface temperature of the planet which is assumed to have increased by 0.8C since 1900.

☐ Is anthropogenic global warming a myth?

☐ Probably not.

☐ It is understood that CO2 absorbs IR-radiation and re-radiates the absorbed energy in all directions therefore even if there were one anthropogenic CO2 molecule in the atmosphere it would still increase the surface temperature by some amount, albeit infinitesimally small.

☑ Is significant dangerous anthropogenic global warming a myth?

☐ Yes, most definitely.

☐ The CAGW-religion as propounded by the IPCC is in conflict with well-established universally-accepted physical laws of nature, such as Henry's law and the Stefan-Boltzmann law.

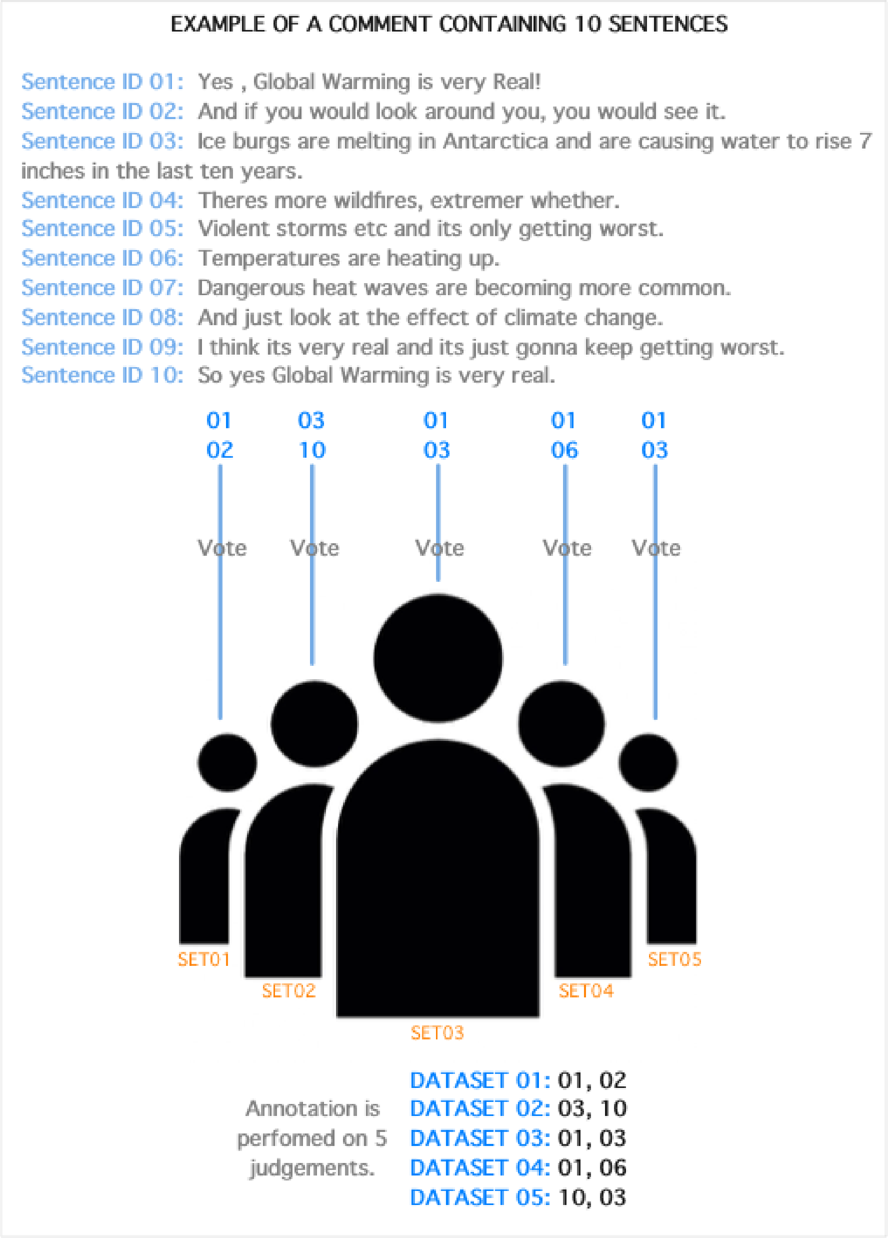
☐ I wrote an article on why the IPCC's Revelle Factor contradicts Henry's law called 'Digging into the core: Why the increase in CO2 is probably natural' and why it contradicts the Stefan-Boltzmann law called 'The Cold Truth About CO2'.

☑ If we use empirical measurements of CO2's absorptivity/emissivity and apply the Stefan-Boltzmann law we find that the maximum warming from the entire atmospheric CO2-greenhouse of about 400ppmv is 0.05C.

☐ If you investigate this subject in any detail, you soon see that almost every piece of evidence that CAGW-advocates have presented is manufactured pseudoscientific claptrap.

```
• Not if you define global warming as the increase in average surface temperature of the planet which is
assumed to have increased by 0.8C since 1900.
• Is significant dangerous anthropogenic global warming a myth?
• If we use empirical measurements of CO2's absorptivity/emissivity and apply the Stefan-Boltzmann law we find
that the maximum warming from the entire atmospheric CO2-greenhouse of about 400ppmv is 0.05C.
```

Figure 4.3: The interface of the Salient Sentence Annotation system



EXAMPLE OF A COMMENT CONTAINING 10 SENTENCES

Sentence ID 01: Yes , Global Warming is very Real!
Sentence ID 02: And if you would look around you, you would see it.
Sentence ID 03: Ice burgs are melting in Antarctica and are causing water to rise 7 inches in the last ten years.
Sentence ID 04: Theres more wildfires, extremer whether.
Sentence ID 05: Violent storms etc and its only getting worst.
Sentence ID 06: Temperatures are heating up.
Sentence ID 07: Dangerous heat waves are becoming more common.
Sentence ID 08: And just look at the effect of climate change.
Sentence ID 09: I think its very real and its just gonna keep getting worst.
Sentence ID 10: So yes Global Warming is very real.

Annotation is perfomed on 5 judgements.

DATASET 01: 01, 02
DATASET 02: 03, 10
DATASET 03: 01, 03
DATASET 04: 01, 06
DATASET 05: 10, 03

Figure 4.4: Example of data annotation for the Salient Sentence Selection dataset (SSSD)

In the annotation task, we recruited 22 participants: 10 males and 12 females to annotate salient sentences. The participants were fluent in English and in the age range between 24 and 45 years old. When participants logged in to the system, a debate topic and a comment which is split into a list of sentences were shown. The annotators were given a guideline to read and select salient sentences that summarize the comments. From each comment we allowed the annotators to select only 20% of the comments sentences. This proportion is treated as the summary of the shown comment. Figure 4.4 illustrates an example of annotators annotating a comment containing 10 sentences. Based on the compression rate, the annotators are only allowed to select 2 sentences from the original comments. The system disallows users to submit the answers unless they choose all the sentences according to the compression rate.

Table 4.1: Statistical information of the online debate corpus

Topic ID	**Debate Topics**	**Comments**	**Sentences**	**Words**
01	Is global warming a myth?	18	128	2,701
02	Is global warming fictitious?	28	173	3,346
03	Is the global climate change man-made?	10	47	1,112
04	Is global climate change man-made?	103	665	12,054
05	Is climate change man-made?	9	46	773
06	Do you believe in global warming?	21	224	3,538
07	Does global warming exist?	68	534	9,178
08	Can someone prove that climate change is real (yes) or fake (no)?	8	49	1,127
09	Is global warming real?	51	434	6,749
10	Is global warming true?	5	26	375
11	Is global warming real (yes) or just a bunch of scientist going to extremes (no)?	20	192	2,988
	Average	**31**	**229**	**3,995**
	Total	**341**	**2,518**	**43,941**

To create the dataset, we aimed to have 5 annotations for each debate topic. Due to a limited number of annotators and a long list of comments to be annotated in each debate topic, 11 participants were asked to complete more than one debate topic but were not allowed to annotate the same debate topics. In total, 55 annotation sets were derived: 11 debate topics and each with 5 annotation sets. Each annotation set consists of 341 comments with total 519 annotated salient sentences. To conclude, we derived 5 X 519 annotated salient sentences based on five annotation sets and named this dataset as the Salient Sentence Selection Dataset (SSSD)[1]. Table 4.1 illustrates the statistical information of the SSSD.

4.1.1.3 ***Inter-Annotator Agreement***

In order to compute inter-annotator agreement, we calculated the averaged Cohen's Kappa and Krippendorff's alpha with a distant metric, Measuring Agreement on Set-valued Items metric (MASI)[2]. The scores of averaged Cohen's Kappa and Krippendorff's alpha are 0.28 and 0.27 respectively. According to the scale of Krippendorff (2004), our alpha did neither accomplish the reliability scale of 0.80, nor the marginal scales between 0.667 and 0.80. Likewise, our Cohen's Kappa only achieved the agreement level of *fair agreement*, as defined by Landis and Koch (1977). However, such low agreement scores are also reported by others who aimed to create gold standard summaries from news texts or conversational data (Mitra et al., 1997; Liu and Liu, 2008).

Our analysis shows that the low agreement is caused by the subjective judgments of annotators in the selection of salient sentences. As shown in Listing 4.1 the sentences are syntactically different but bear the same semantic meaning. In a summarization task with a compression threshold applied, such situation limits the annotators to select one of the sentences but not all. Depending on each annotator's subjectivity, the selection leads to the results of a different set of salient sentences. To address this we relaxed the agreement computation by treating sentences equal when they are semantically similar.

4.1.1.4 ***Relaxed Inter-Annotator Agreement***

When an annotator selects a sentence, other annotators might select other sentences expressing similar meaning. In this experiment, we aim to detect sentences that are semantically similar by applying Doc2Vec from the Gensim package (Řehůřek and Sojka, 2010). Doc2Vec model simultaneously learns the representation of words in sentences and the labels of the sentences. The labels are numbers or chunks of text which are

[1] This dataset can be downloaded at https://goo.gl/3aicDN.

[2] In the calculation, we applied a package available in Python, called *nltk.metrics package* (NLTK Metrics, 2015).

Paraphrased Arguments

Example 1: Propositions from the proponents

- Global warming is real.
- Global warming is an undisputed scientific fact.
- Global warming is most definitely not a figment of anyone's imagination because the proof is all around us.
- I believe that global warming is not fictitious, based on the observational and comparative evidence that is currently presented to us.

Example 2: Propositions from the opponents

- Global warming is bull crap.
- Global Warming isn't a problem at all.
- Just a way for the government to tax people on more things by saying they are trying to save energy.
- Yes, global warming is a myth, because they have not really proven the science behind it.

Listing 4.1: Examples of Paraphrased Arguments

used to uniquely identify sentences. We used the debate data and a richer collection of sentences related to climate change to train the Doc2Vec model. In total, there are 10,920 sentences used as the training set.

To measure how two sentences are semantically referring to the same content, we calculated the cosine similarity scores among sentences. A cosine similarity score of 1 means that the two sentences are semantically equal and 0 is when the opposite is the case. In the experiment, we manually investigated pairs of sentences at different threshold values and found that at the threshold of 0.44 and above the sentence pairs have the most similarity of semantic meaning. The example below shows a pair of sentences obtained at 0.44 level.

S1: *Humans are emitting carbon from our cars, planes, and factories, which is a heat-trapping particle.*

S2: *So there is no doubt that carbon is a heat-trapping particle, there is no doubt*

Table 4.2: Inter-Annotator Agreement before and after applying the semantic similarity approach.

Trial	**Threshold** ($\geq$)	κ	α
Before		0.28	0.27
After	0.00	0.81	0.83
	0.10	0.62	0.65
	0.20	0.46	0.50
	0.30	0.40	0.43
	0.40	0.39	0.41
	0.42	0.38	0.41
	0.44	**0.38**	**0.40**
	0.46	0.38	0.40
	0.48	0.38	0.40
	0.50	0.38	0.40
	0.60	0.38	0.40
	0.70	0.38	0.40
	0.80	0.38	0.40
	0.90	0.38	0.40
	1.00	0.38	0.40

that our actions are emitting carbon into the air, and there is no doubt that the amount of carbon is increasing.

In the pair, the two sentences mention the same topic (i.e. *carbon emission*) and express the idea in the same context. We used the threshold 0.44 to re-compute the agreement scores. By applying the semantic approach, the inter-annotator agreement scores of Cohen's Kappa and Krippendorff's alpha increase from 28% to 35.71% and from 27% to 48.15% respectively. The inter-annotator agreement results are illustrated in Table 4.2. Note that, in the calculation of the agreement, we incremented the threshold by 0.02. Only particular thresholds are shown in the table due to the limited space.

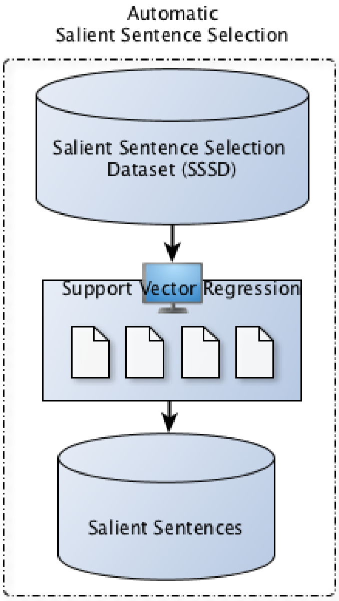

Automatic
Salient Sentence Selection
Salient Sentence Selection
Dataset (SSSD)
Support Vector Regression
Salient Sentences

Figure 4.5: The first process of the system architecture, Automatic Salient Sentence Selection process

4.1.1.5 ***Automatic Salient Sentence Selection Methodology***

Automatic Salient Sentence Selection is the first process in the system architecture of our online debate summarization system. Figure 4.5 illustrates the activities occurred in this process. Automatic salient sentence selection begins with the input of the SSSD to the debate summarization system. The system assumes an input of n comments from the *Agree* and *Disagree* sides. Each comment consists of several sentences. Salient sentences refer to those which are the most meaningful content or the summaries of the comment. The system automatically extracts the most salient sentences from each comment, based on a compressed rate of 20%. Listing 4.2 shows an example of a comment from an online debate and Listing 4.3 is the salient sentences automatically selected by the system. The original comment contains 10 sentences and the system select 10 X 20 / 100 $=$ 2 sentences from the comment.

We view that the automatic selection of salient sentences can be achieved by a regression task. To train the model, a regression score for a sentence is defined between 1 to 5. It is derived from the number of annotators voted for that sentences divided by

Original Comment

Yes, Global Warming is very Real! And if you would look around you, you would see it. Ice burgs are melting in Antarctica and are causing water to rise 7 inches in the last ten years. There are more wildfires, extremer whether. Violent storms etc and its only getting worst. Temperatures are heating up. Dangerous heat waves are becoming more. And just look at the effect of climate change. I think it's very real and it's just gonna keep getting worst. So yes Global Warming is very real.

Listing 4.2: Example of an original comment from an online debate

Salient Sentences

And just look at the effect of climate change. I think it's very real and it's just gonna keep getting worst.

Listing 4.3: The salient sentences automatically selected by the system

the number of all annotators. In order for the system to determine which sentences are salient, we build a support vector regression model together with a set of dominant features to score each sentence. The model combines the features for scoring sentences in debate comments. Based on the compression rate, for each comment, the sentences with the highest regression scores are considered the most salient ones. These activities are repeated for all comments. Once this process is complete, a list of salient sentences is derived and will be used as the input in the next process. More detail can be found in Chapter 5.

In the evaluation of this process, ROUGE evaluation metric is used to determine n-grams overlap between the automatic selected salient sentences and the SSSD.

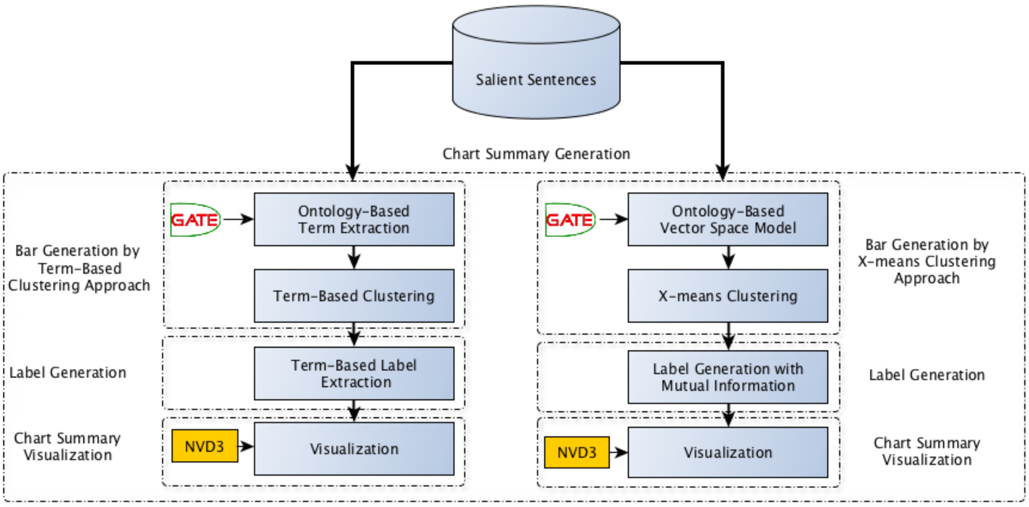

Salient Sentences
Chart Summary Generation
Bar Generation by Term-Based Clustering Approach
GATE
Ontology-Based Term Extraction
Term-Based Clustering
Label Generation
Term-Based Label Extraction
Chart Summary Visualization
NVD3
Visualization
GATE
Ontology-Based Vector Space Model
X-means Clustering
Bar Generation by X-means Clustering Approach
Label Generation with Mutual Information
Label Generation
NVD3
Visualization
Chart Summary Visualization

Figure 4.6: Two methodologies investigated for the generation of a Chart Summary

4.1.2 Chart Summary Generation

The objective of this process is to generate a Chart Summary. Chart Summary is constituted by combining three components: bars, labels, and figures. The bars are the clusters of related salient sentences which were automatically selected by the system in the previous process. Labels indicate a brief description of the bars. Figures represent the number of salient sentences in the bars. The salient sentences previously selected by the system is the input for this process.

In this thesis, we view that the generation of Chart Summary can be achieved by two methodologies. Figure 4.6 summarizes the two methodologies for constructing a Chart Summary. The two methodologies are discussed in the following sections.

4.1.2.1 Term-Based Clustering

The first methodology that we explore in the creation of Chart Summary is a term-based clustering approach. In this approach, we use a list of key terms to cluster salient sentences into the same groups. The salient sentences that share the same terms are placed in the same clusters. These terms are derived from an ontology service. As the terms already elaborate the central meaning of the clusters, we regard that they are the cluster labels. In the next step, salient sentences in all clusters are counted to represented as the figures of the bars.

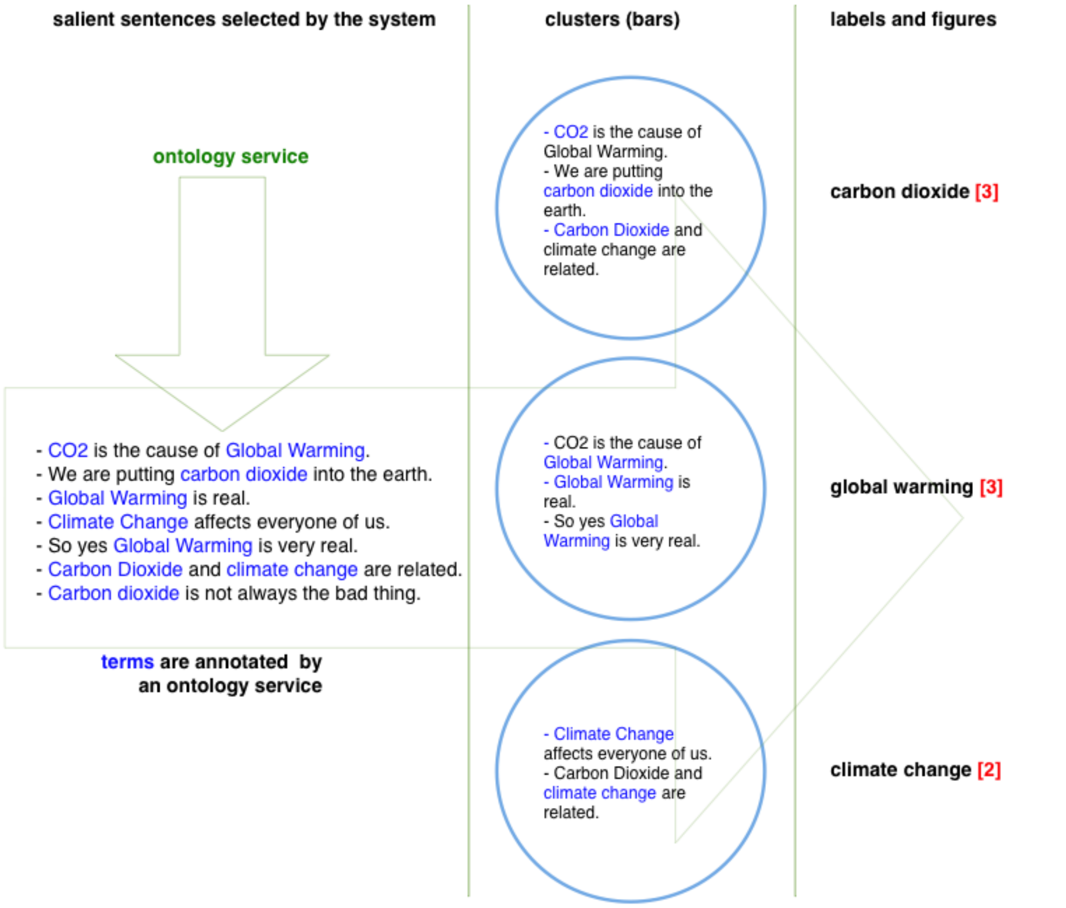

salient sentences selected by the system
clusters (bars)
labels and figures
ontology service
- CO2 is the cause of Global Warming.
- We are putting carbon dioxide into the earth.
- Global Warming is real.
- Climate Change affects everyone of us.
- So yes Global Warming is very real.
- Carbon Dioxide and climate change are related.
- Carbon dioxide is not always the bad thing.
terms are annotated by an ontology service
- CO2 is the cause of Global Warming.
- We are putting carbon dioxide into the earth.
- Carbon Dioxide and climate change are related.
- CO2 is the cause of Global Warming.
- Global Warming is real.
- So yes Global Warming is very real.
- Climate Change affects everyone of us.
- Carbon Dioxide and climate change are related.
carbon dioxide [3]
global warming [3]
climate change [2]

Figure 4.7: An example of salient sentence clustering by a term-based clustering approach

Figure 4.7 summarizes the generation of bars, labels, and figures by the term-based clustering approach. In the figure, assume the salient sentences on the left-hand side are those selected by the system. They are fetched to the ontology service and later the terms in the sentences are captured by the service. Those terms are considered as the labels, indicating groups of the related salient sentences. The final step is the count of a total number of the sentences in clusters. More details of the term-based clustering approach will be discussed in Chapter 5.

4.1.2.2 ***X-means Clustering***

The other clustering methodology that we investigate for the generation of Chart Summary is an X-means clustering approach. X-means is a clustering algorithm which is an extended version of K-means. It allows us to automatically detect a number of clusters in text (Pelleg and Moore, 2000). More details of this algorithm will be discussed in Chapter 5.

In this methodology, we also employ ontologies as the background knowledge for capturing important terms in the text. The terms are transformed into similarity vectors and then the vectors are clustered using the X-means algorithm. In the generation of labels, we apply a Mutual Information approach to score all candidate terms in clusters. More details of Mutual Information will be discussed in Chapter 5.

In this process, we measure the quality of the clustering results by calculating mean silhouette scores, indicating how well the sentences in clusters coherently connect. For the evaluation of cluster labels, we follow the manual evaluation method presented by Aker et al. (2016). A set of Likert-scale questions is given to subjects to evaluate the quality of labels.

To sum up, at the end of the second process, we derived two sets of clusters generated by the term-based clustering approach and the X-means clustering approach. The clusters contain several related sentences. These clusters will be used to create new datasets for a contradiction detection task which will be discussed in the next section.

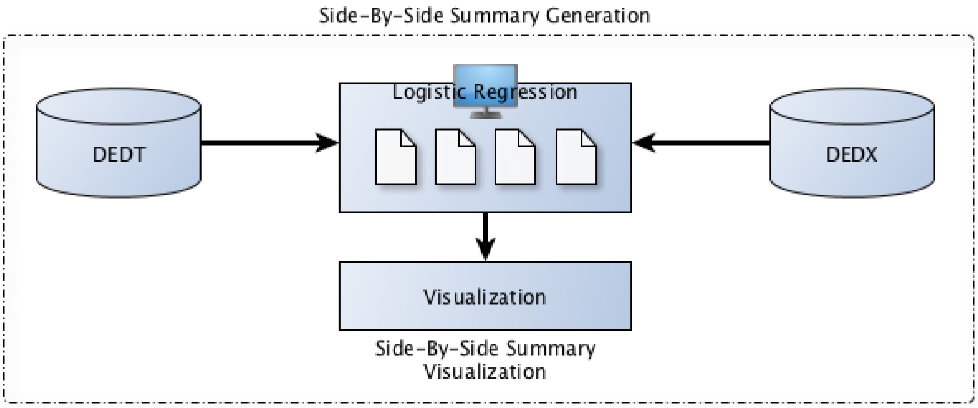

Side-By-Side Summary Generation
DEDT
Logistic Regression
DEDX
Visualization
Side-By-Side Summary Visualization

Figure 4.8: The last process of the system architecture, Contradiction Detection process

4.1.3 Side-By-Side Summary Generation

The main objective of this process is to generate a Side-By-Side Summary. The summary is visualized as a table of a topic. Each row in the table consists of pairs of contradictory sentences related to that topic. This section begins with a brief discussion of how datasets used for the generation of Side-By-Side Summary are created and annotated. Next, we discuss how we classify examples in the datasets and visualize the classification results as a Side-By-Side Summary. Figure 4.8 illustrates the activities occurred in this process.

4.1.3.1 Data Annotation for the Contradiction Detection Task

The clustering results derived from the previous process are further annotated for the contradiction detection task. As the clustering results are derived from the two clustering approaches, we create and annotate two debate entailment datasets. The datasets are prepared by creating pairs of RTE sentences: text (T) and the hypothesis (H). In each cluster, sentences in clusters from the Agree side will be matched with those on the Disagree side. The longer sentences are considered to be the text and the shorter sentences are chosen to be the hypothesis (Lendvai et al., 2016). In a later step, the pairs of H-T sentences are annotated with one of the two entailment relations: *Contradiction* or *Non-Contradiction*. In total, we derive two datasets, Debate Entailment Dataset from the Term-based clustering approach (DEDT) and Debate Entailment Dataset from the X-means clustering approach (DEDX). More details of the annotation methodology are discussed in Chapter 6.

4.1.3.2 Logistic Regression Model

In order to generate a Side-By-Side Summary, we classify the sentences pairs whether they are contradiction or non-contradiction. Figure 4.9 shows an example of the classification. The two datasets will be the input to the system. We apply the logistic regression algorithm to create a classifier. Additionally, we also define key features to help the classification of sentence pairs. The evaluation approach for the contradiction detection task from de Marneffe et al. (2008) is followed and reported as Precision, Recall, and F1 scores. More details of the classification and evaluation are discussed in

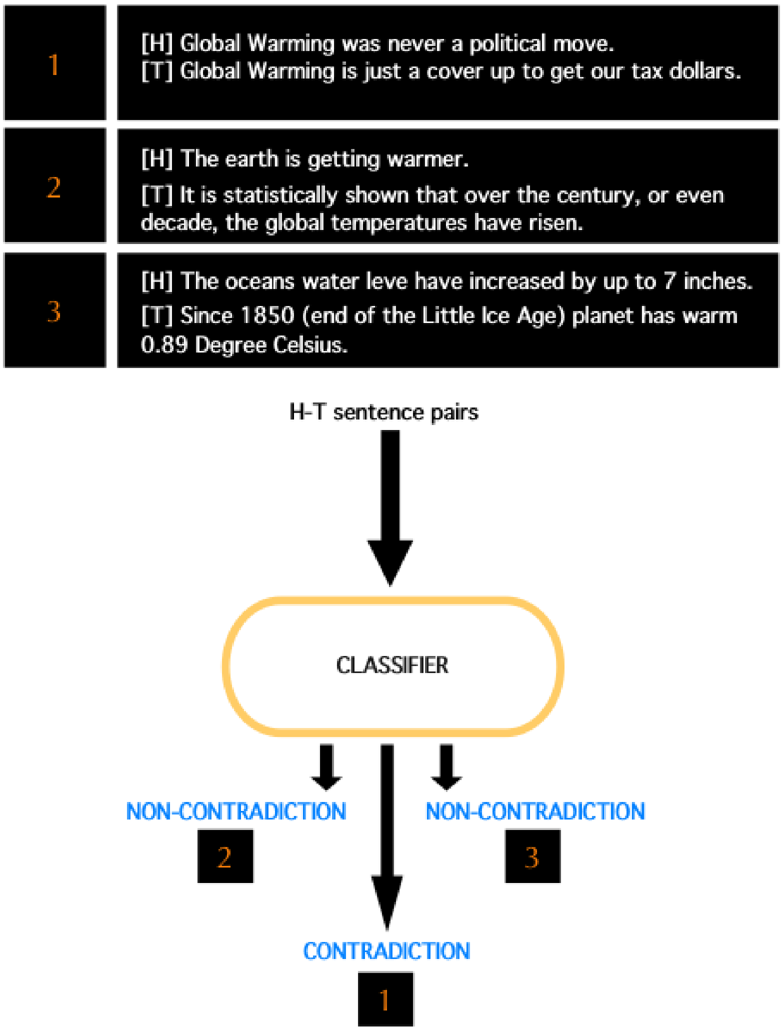

1
[H] Global Warming was never a political move.
[T] Global Warming is just a cover up to get our tax dollars.
2
[H] The earth is getting warmer.
[T] It is statistically shown that over the century, or even decade, the global temperatures have risen.
3
[H] The oceans water leve have increased by up to 7 inches.
[T] Since 1850 (end of the Little Ice Age) planet has warm 0.89 Degree Celsius.
H-T sentence pairs
CLASSIFIER
NON-CONTRADICTION
2
NON-CONTRADICTION
3
CONTRADICTION
1

Figure 4.9: An example of the classification in the contradiction detection task

the next chapter. Once the classification is complete, we export contradictory sentence pairs and visualize them as a Side-By-Side Summary in HTML pages.

4.2 Summary

To summarize, aside from the work done by Ranade et al. (2013), we proposed a novel system architecture which is not explored before in debate summarization. Our architecture for debate summarization is specially designed for generating Combination 2 summaries, presenting both a chart-based overview and addition of in-depth side-by-side comparison of opposing debate stances. Our system structure is therefore significantly unique compared to the previous work. For instance, in the generation of a Chart

Summary, labels are required to fulfill the requirements of the chart. In addition, we need a mechanism to detect whether a topic is mentioned on both opposing sides are rebuttal since it is one of the major requirements for the generation of Side-By-Side Summary. For these reasons, the combination of these crucial requirements makes our system distinctive and cannot be found in the system architectures of the related work in debate summarization (Witte and Bergler, 2007; Kim and Zhai, 2009; Huang et al., 2011; Campr and Jezek, 2012; Lerman and McDonald, 2009; Ranade et al., 2013).

The next chapter will discuss the first and the second stages of the system architecture in detail. It elaborates how each component of a Chart Summary is generated and combined into a Chart Summary.

Chapter 5

CHART SUMMARY GENERATION

This chapter paves the way for the generation of a Chart Summary. The Chart Summary consists of bars which represent clusters of salient sentences, labels of the bars which give short descriptions of the bars, and the frequencies of the bars showing the numbers of particular salient sentences discussed in each bar. These components are developed according to the processes shown in the blueprint of our system architecture. This chapter discusses two main processes of the system architecture of online debate summarization system. In the first process, we discuss how salient sentences are automatically selected by the system. Then in the next process, we cluster the salient sentences, extract labels, and count the frequencies. These components are combined and then visualized as a Chart Summary.

5.1 *Automatic Salient Sentence Selection*

In this process, we aim to select sentences that are deemed important or that summarize the information mentioned in the comments. We call these sentences, the *salient sentences*. The number of salient sentences selected from each comment is based on the compression rate of 20%. In order to select salient sentences, different features are defined. A Support Vector Regression model combines the features for scoring sentences in each debate comment. The following sections discuss the salient sentence selection process in detail, including the regression model, features, experiments, baseline, results and evaluation metric.

5.1.1 ***Support Vector Regression Model***

In this experiment, we work on the extractive summarization problem and aim to select sentences that are deemed important or that summarize the information expressed in debate comments. Additionally, we aim to investigate the keys features which play the important roles in the summarization of the debate data. We view this salient sentence selection as a regression task. Hirao et al. (2002), Li et al. (2007), and Hong et al. (2015) report that Support Vector Machine (SVM) is an efficient approach for sentence extraction. For this reason, we use a popular machine learning package which is available in Python, called Scikit-learn (Pedregosa et al., 2011) to build our support vector regression model. In order to train the model, a regression score for a sentence is defined between 1 to 5. It is derived from the number annotators selected that sentence divided by the number of all annotators. In this experiment, we defined 8 different features and the support vector regression model combines the features for scoring sentences in each debate comment. From each comment, sentences with the highest regression scores are considered the most salient ones.

5.1.2 ***Feature Selection***

The following features were experimented for automatically extracting salient sentences in debate comments. In total, there are 9 features including the combination one.

1. **Sentence Position (SP).** Sentence position correlates with the important information in the text (Baxendale, 1958; Edmundson, 1969; Goldstein et al., 1999). In general, humans are likely to mention the first topic in the earlier sentence and they express more information about it in the later sentences. We prove this claim by conducting a small experiment to investigate which sentence positions frequently contain salient sentences. We processed the annotated SSSD, kept records of the positions, and illustrated the statistical information of the sentence positions selected by the annotators in Figure 5.1. From our data annotation, 60 percent of salient sentences locate at the first three positions of the comments, shaping the assumption that the first three sentences are considered as containing salient pieces of information. Equation 5.1 shows the calculation of the score for the sentence position features.

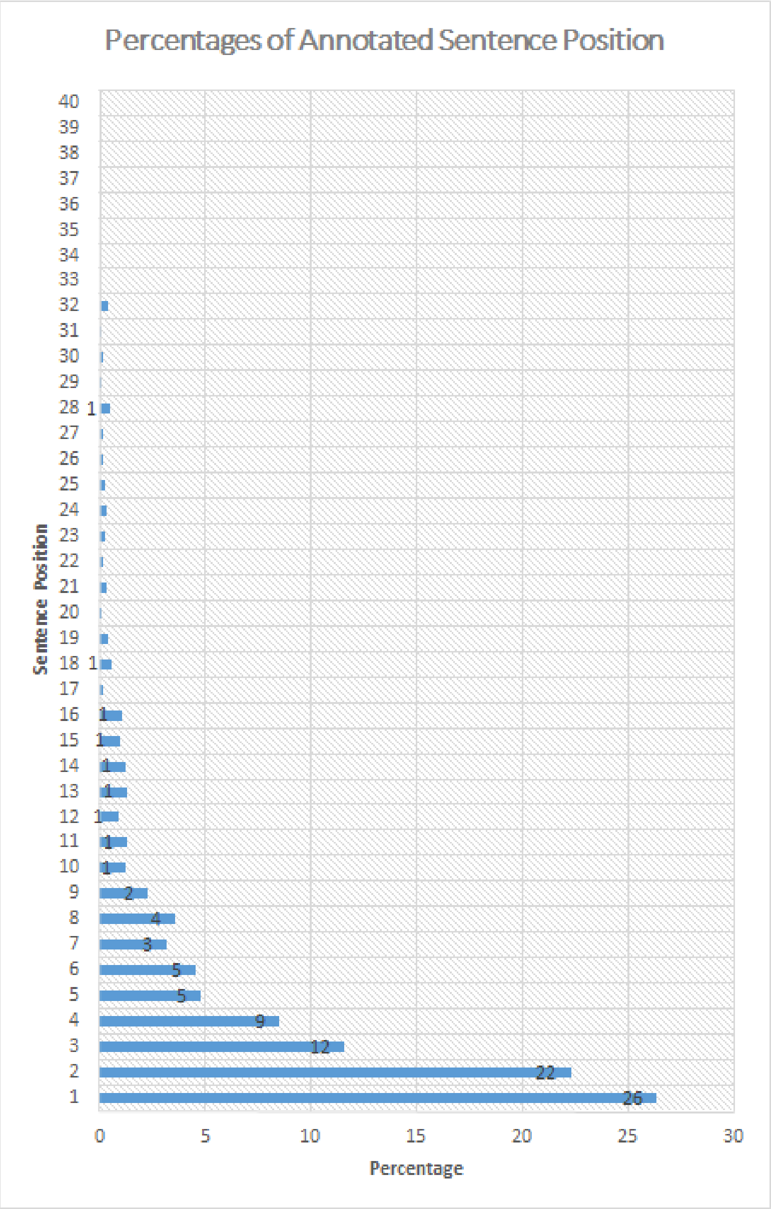


Figure 5.1: The percentage of annotated sentence position

$$Score = \begin{cases} \frac{1}{sentence\ position}, & \text{if } position < 4 \\ 0, & \text{otherwise} \end{cases} \tag{5.1}$$

2. **Debate Titles (TT).** In writing, a writer tends to repeat the title words in a document. For this reason, a sentence containing title words is likely to contain important information. We collected 11 debate titles as shown in Table 4.1. In our experiment, a sentence is considered as important when it contains mutual words as in debate titles. Equation 5.2 shows the calculation of the score for this feature.

$$Score = \frac{number\ of\ title\ words\ in\ sentence}{number\ of\ words\ in\ debate\ titles} \tag{5.2}$$

3. **Sentence Length (SL).** Sentence length also indicates the importance of sentence based on the assumption that either very short or very long sentences are unlikely to be included in the summary. Equation 5.3 is used in the process of extracting salient sentences from debate comments.

$$Score = \frac{number\ of\ words\ in\ a\ sentence}{number\ of\ words\ in\ the\ longest\ sentence} \tag{5.3}$$

4. **Conjunctive Adverbs (CJ).** One possible feature that helps identify salient sentence is to determine conjunctive adverbs in sentences. Conjunctive adverbs were proved that they support cohesive structure of writing. For instance, "the conjunctive adverb *moreover* has been used mostly in the essays which lead to a conclusion that it is one of the best-accepted linkers in the academic writing process."(Januliené and Dziedravičius, 2015). The NLTK POS Tagger[1] was used to determine conjunctive adverbs in our data.

5. **Cosine Similarity.** Cosine similarity has been used extensively in Information Retrieval, especially in the vector space model (Salton et al., 1975). Documents

[1] `http://www.nltk.org/api/nltk.tag.html`

are ranked according to the similarity of the given query. Equation 5.4 illustrates the equation of cosine similarity where $\vec{q}$ and $\vec{d}$ are n-dimensional vectors (Manning and Schütze, 1999). Cosine similarity is one of our features that is used to find similarity between two textual units. The following pairs of textual units are used to compute the score of cosine similarity.

$$cos(\vec{q}, \vec{d}) = \frac{\sum_{i=1}^{n} q_i d_i}{\sqrt{\sum_{i=1}^{n} q_i^2} \sqrt{\sum_{i=1}^{n} d_i^2}} \tag{5.4}$$

(a) **Cosine similarity of debate title words and sentences (COS_TTS).** For each sentence in debate comments, we compute its cosine similarity score with the title words. This is based on the assumption that a sentence containing title words is deemed as important.

(b) **Cosine similarity of climate change terms and sentences (COS_CCTS).** A list of climate change terms was collected from news media about climate change. We calculate cosine similarity scores between the terms and sentences. In total, there are 300 most frequent terms relating to location, person, organization, and chemical compounds.

(c) **Cosine similarity of topic signatures and sentences (COS_TPS).** Topic signatures play an important role in automatic text summarization and information retrieval. It helps identify the presence of complex concepts or the importance in text. In a process of determining topic signatures, words appearing occasionally in the input text but rarely in other text are considered as topic signatures. They are determined by an automatic predefined threshold which indicates descriptive information. Topic signatures are generated by comparing words in two sets of text using using a concept of the likelihood ratio (Nenkova and McKeown, 2011; Lin and Hovy, 2000), λ presented by Dunning (1993). It is a statistical approach which calculates a likelihood of a word. For each word in the input, the likelihood of word occurrence is calculated in a pre-classified text collection. Another likelihood value of the same word is calculated and compared in another out-of-topic collection. The word, on the topic-text collection that has higher likelihood value than the out-of-topic collection, is regarded as the topic signature of a topic. Otherwise, the word is ignored.

6. **Semantic Similarity of Sentence and Debate Titles (COS_STT).** Since the aforementioned features do not semantically capture the meaning of context, we create this feature for such purpose. We compare each sentence to the list of debate titles based on the assumption that forum users are likely to repeat debate titles in their comments. Thus, we compare each sentence to the titles and then calculate the semantic similarity score by using *Doc2Vec* (Řehůřek and Sojka, 2010).

5.1.3 ***Baseline***

MEAD is a multi-document summarization system that extracts sentences based on a linear combination of features. The key features used in MEAD are 1) centroid which centers terms in documents in the clusters; 2) position of sentences in the documents; and 3) the similarity of sentences overlapping with the first sentence in documents (Radev et al., 2000). Debate comments from each opposing side were fed to MEAD. Then, the sentences in the comments were scored based on the three features and afterward ascendingly ranked based on the highest score. Table 5.1 illustrates ROUGE scores derived from the automatic salient sentence selections performed by MEAD.

Table 5.1: The results of ROUGE scores for the automatic salient sentence selection performed by MEAD

Evaluation Metrics	**ROUGE-1**	**ROUGE-2**	**ROUGE-SU4**
ROUGE Scores	0.4579	0.4011	0.4029

Automatic Selected Salient Sentences

- Global Warming isn’t a problem at all.

- In my opinion Global Warming doesn’t even exist.

- Global warming is a myth created by corporations in order to make profit.

- Why did we all hop on board the global warming bandwagon started by politicians when the scientific community didn’t back it?

- Global warming does not exist..

- Cars, factories, etc. Earth has its own phases, and just because “the temp is increasing over time,” doesn’t mean its global warming..

- No, I do not think that global warming is true.

- If global warming was real, Perhaps we would be seeing the sea level rising rapidly.

- Yes, global warming is a myth, because they have not really proven the science behind it.

- People are saying any change in temperature or natural disaster is caused by global warming.

- Global Warming has been happening since the Earth was born.

- Says Al Gore and you know what Global warming is Man made, Yep, man made!

- Also, look at the main drivers behind "global warming", “climate change”, and “human induced climate change”.

- The real main factor to global warming is water vapor, so technically not our fault.

- Depends on how you define “global warming”.

- Not if you define global warming as the increase in average surface temperature of the planet which is assumed to have increased by 0.8C since 1900.

- If there was global warming it would just be plain rain.

- I think that the earth has been regulating its own temperature, and humans have had very little if anything at all to do with global warming.

- Global warming is not humans fault as the government would have you believe.

- Global warming is a myth that modern liberals use to push their environment-friendly agenda.

- Global Warming is just a cover up to get our tax dollars.

Listing 5.1: An example of the salient sentences selected by the system.

Table 5.2: ROUGE scores of salient sentences selection derived from different features and the baseline

Features / ROUGE-N (Recall)	**ROUGE-1**	**ROUGE-2**	**ROUGE-SU4**
Sentence Position (SP)	**0.6124**	**0.5375**	**0.4871**
Debate Titles (TT)	0.5407	0.4693	0.4303
Sentence Length (SL)	0.4307	0.3550	0.3335
Conjunctive Adverbs (CJ)	0.4988	0.4346	0.4147
Cosine Similarity of Topic Signatures and Sentences (COS_TPS)	0.3907	0.2986	0.2699
Cosine similarity of Debate Title Words and Sentences (COS_TTS)	0.5630	0.5076	0.4780
Cosine Similarity of Climate Change Terms and Sentences (COS_CCTS)	0.3389	0.2558	0.2340
Semantic Similarity of Sentence and Debate Titles (COS_STT)	0.4304	0.3561	0.3340
Combination of Features	0.4773	0.3981	0.3783
MEAD Baseline	0.4579	0.4011	0.4029

Table 5.3: The statistical information of comparing sentence position and other features after applying Doc2Vec. The table shows the abbreviations of the feature names.

Comparison Pairs	**ROUGE-1**		**ROUGE-2**		**ROUGE SU4**	
	Z	**Asymp. Sig. (2-tailed)**	**Z**	**Asymp. Sig. (2-tailed)**	**Z**	**Asymp. Sig. (2-tailed)**
SP VS CB	-4.246^b	0*	-3.962^b	0*	-3.044^b	0.002
SP VS CJ	-3.570^b	0*	-3.090^b	0.002	-2.192^b	0.028
SP VS COS_CCTS	-6.792^b	0*	-6.511^b	0*	-6.117^b	0*
SP VS COS_TTS	-1.307^b	0.191	$-.789^b$	0.43	$-.215^b$	0.83
SP VS COS_TPS	-6.728^b	0*	-6.663^b	0*	-6.384^b	0*
SP VS SL	-4.958^b	0*	-4.789^b	0*	-4.110^b	0*
SP VS COS_STT	-4.546^c	0*	-4.322^c	0*	-3.671^c	0*
SP VS TT	-3.360^c	0.001*	-2.744^c	0.006	-2.641^c	0.008

a) Wilcoxon Signed Ranks Test.

b) Based on negative ranks.

c) Based on positive ranks.

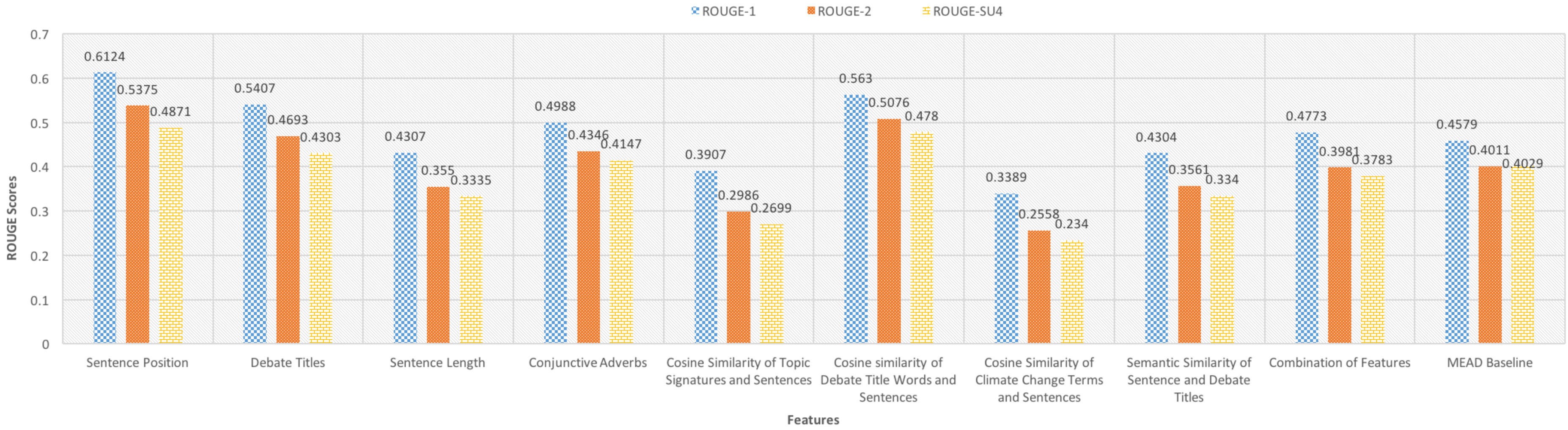


Figure 5.2: The graphical information of ROUGE scores generated by different features and the baseline

5.1.4 ***Results and Discussion***

In this process, we selected salient sentences from the SSSD. Listing 5.1 illustrates an example of salient sentences which were automatically selected by the system. In order to evaluate the performance of our system, we applied ROUGE-N evaluation metrics. We reported ROUGE-1 (unigram), ROUGE-2 (bi-grams) and ROUGE-SU4 (skip-bigram with the maximum gap length of 4). Additionally, we also compared the system's performance to a baseline, MEAD. Table 5.2 illustrates macro average of ROUGE scores (recall) of the salient sentences selected by each feature and those generated by the baseline. Figure 5.2 shows the ROUGE scores in a graphical format.

From the figure, it can be concluded that the *sentence position* feature plays the most important role in the selection of salient sentences. This evidence conforms to our statistical information shown in Figure 5.1 in which approximately 60% salient sentences locate at the earlier sentences. Additionally, the *debate titles* feature is also one of the most important features in the selection of the salient sentences as it also yields a high ROUGE score. This is also implied by the cosine similarity scores of other pairs. For instance, it may be observed that when the similarity of sentences and debate titles is measured, the score of cosine similarity increases. In contrast, the score is lower when the similarity is calculated by measuring sentences against topic signatures and climate change terms. This is possible when the sentences hardly express *person* and *organization* entities as in topic signatures and the climate change terms. For the combination of all features, it yields a satisfying result.

As can be seen from the table and the figure, the performance of our system outperforms the baseline. One primary reason leading a superior performance is the use of suitable features in the regression model. Whereas our features, for example, help identify a location of the sentence in debate comments, MEAD only selects sentences at the centers of the clusters – no useful features are integrated into MEAD to help identify the salient sentences. For this reason, MEAD is not able to identify the salient sentences in our experiment and therefore does not yield higher ROUGE scores than our system.

Moreover, to measure the statistical significance of the ROUGE scores generated by the features, we calculated a pairwise Wilcoxon signed-rank test with Bonferroni correction.

We report the significance p = .0013 level of significance after the correction is applied. Our results indicated that there is statistical significance among the features. Table 5.3 illustrates the statistical information of comparing sentence position and other features. The star indicates that there is a statistically significant difference between each comparison pair.

In our experiment, we conclude that the best results are derived with ROUGE-1. This is possible when the frequent terms in our data are mostly unigram and they are frequently included in the reference summaries.

5.1.5 ***Conclusion***

In this process, we worked on an annotation task for a new annotated dataset, online debate data. We have manually collected reference summaries for comments given to global warming topics. The data consists of 341 comments with total 519 annotated salient sentences. We have performed five annotation sets on this data so that in total we have 5 X 519 annotated salient sentences. In addition, we also implemented an extractive text summarization system on this debate data. Our results reveal that the key feature that plays the most important role in the selection salient sentences is sentence position. Other useful features are debate title words feature, and cosine similarity of debate title words and sentences feature. Therefore, we use the salient sentences which were automatically selected through the sentence position feature in the subsequent experiments.

5.2 *Chart Summary Generation*

Recall the components in a Chart Summary as shown in Figure 3.2. A Chart Summary consists of bars which represent clusters of salient sentences, the labels of the bars providing short descriptions, and the frequencies indicating the numbers of salient sentences in each bar. In our work, we view that the generation of a Chart Summary can be achieved by sentence clustering, cluster labeling, and visualization. These are the main components which will be discussed in this section. In this process, we investigate two different clustering approaches for the generation of Chart Summaries. In the first approach, we generate the chart by applying a term-based clustering approach and a cluster labeling method. The second approach makes use of X-means for clustering and a Mutual Information for labeling the clusters. Both approaches are driven by ontologies. This process is completed with the combination of the components for visualizing the Chart Summary. The following sections discuss the stages occurred in this process in detail.

5.2.1 ***Term-Based Clustering***

In the generation of Chart Summary, we firstly investigate a term-based clustering approach. As its name indicates, clusters are generated based on the terms sharing in the salient sentences. In this approach, we use a list of key terms to cluster sentences into the same group. The salient sentences derived from the previous process are the input.

To perform clustering we used terms extracted from ontologies. We employed the English ClimaPinion service[2] from the DecarboNet project[3] as the background knowledge to capture climate change topics and extract from each salient sentence topical terms. To obtain clusters we grouped sentences containing the same label within the same cluster. If a sentence contained more than one term then it was assigned to several clusters allowing the sentence to be soft-clustered.[4] Also note, terms with the same semantic meaning can be expressed differently. To address this, for each label, we obtained a list of its synonyms from WordNet (Miller, 1995). If the labels shared common syn-

[2] `http://services.gate.ac.uk/decarbonet/sentiment/`

[3] https://www.decarbonet.eu

[4] Within a cluster all sentences must share one particular term but each sentence may contain other terms that are not shared by other sentences within the same cluster.

onyms, we considered them as the same. Consequentially, the sentences automatically annotated with such labels were merged into the same cluster.

5.2.1.1 *Term-based Clustering Evaluation*

The evaluation of the ontology-based term extraction has been already carried out by Maynard and Bontcheva (2015). By consisting of two environmental ontologies, GEMET (GEneral Multilingual Environmental Thesaurus) and Reegle, the ClimaPinion yields great results in recognizing environmental terms in the text, with the precision, recall, and F1 measure of 85.87%, 53.05%, and 65.58% respectively (Maynard and Bontcheva, 2015).

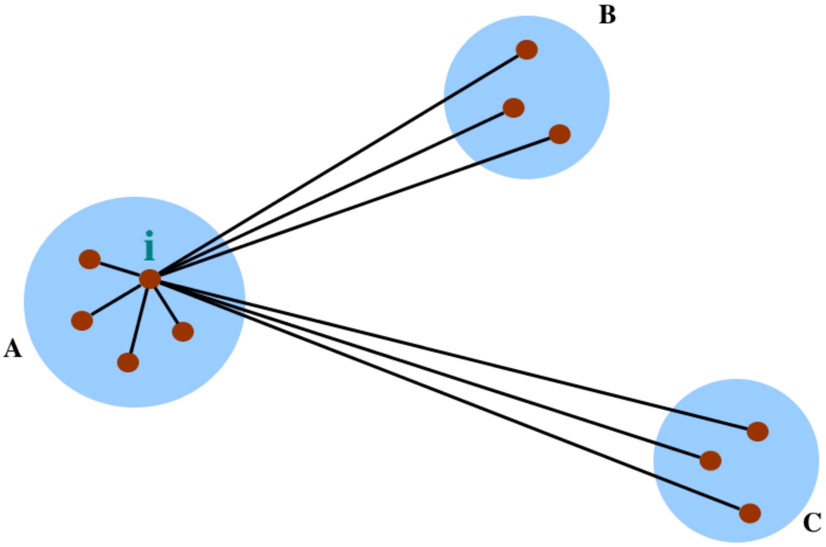


Figure 5.3: An illustration of coordinates and clusters for the calculation of s(i). Adapt from "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis", by P. Rousseeuw, *Journal of Computational and Applied Mathematics*, 53-65, 1987.

The results derived from the term-based sentence clustering are evaluated with the mean silhouette coefficient (Rousseeuw, 1987). The concept of how silhouettes are constructed is illustrated in Figure 5.3. As shown in the figure, to assess how well the coordinate i is well clustered to the cluster A, Rousseeuw (1987) defined how a value of the coordinate i, $s(i)$, is calculated. The algorithm requires two types of average dissimilarities: the average dissimilarities of coordinate i in its own cluster (a(i)) and the average dissimilarities of the coordinate i in other clusters (d(i, C)). Note that C refers to other clusters. By repeating the calculation of d(i, C), the algorithms aims to find the nearest cluster and records the smallest numbers denoted in Equation 5.5.

$$b(i) = \min_{C \neq A} d(i, C) \tag{5.5}$$

Assume that cluster B is the nearest clusters obtained from the calculation, the number of s(i) is derived by determining a(i) and b(i) as shown in Equation 5.6 which can be rewritten as shown in Equation 5.7. The equation shows the calculation for only one coordinate. The overall performance of the clustering results is generally obtained by the calculation of the average of all coordinates in the whole dataset. This is called the *mean silhouette coefficient*.

$$s(i) = \begin{cases} 1 - a(i)/b(i), & \text{if } a(i) < b(i) \\ 0, & \text{if } a(i) = b(i) \\ b(i)/a(i) - 1, & \text{if } a(i) > b(i) \end{cases} \tag{5.6}$$

$$s(i) = \frac{b(i) - a(i)}{max(a(i), (b(i))} \tag{5.7}$$

In summary, as implied by the above equations, the silhouette does not require the gold standard data for calculating the silhouette coefficient. It instead evaluates the clustering performance by determining the cohesion of the documents assigned to a cluster rather than to the other clusters. These documents are represented as coordinates. Silhouette calculates the pairwise difference in both inter-cluster and intra-cluster distance.

In the interpretation of mean silhouette coefficient, the coefficient close to 1.0 indicates a good cohesion and separation of the clustering results, meaning that the average distance from a coordinate in a cluster to the other coordinates within its own cluster is less than the average distance to all coordinates in the nearest cluster. In addition, when the coefficient is close to 0, the coordinates in the clusters are nearly close or on the decision boundary between two neighboring clusters. A negative mean silhouette coefficient is obtained when coordinates might be assigned to wrong clusters. In other words, the coordinates are very close to the neighboring clusters rather than the coordinates in their own clusters (Rousseeuw, 1987).

In this experiment, we derived the mean silhouette score of 0.0000 with a total number of 39 clusters. This is similar to the work presented by Wang and Koopman (2017). The interpretation based on the coefficient is that the data points are assigned near to the decision boundaries of the clusters. Especially, when salient sentences contain multiple climate change topics, clear clustering boundaries are difficult to achieve. This circumstance indicates that such a simple clustering approach is less applicable for grouping semantically similar sentences together and that the task required more sophisticated ways of achieving better performance. We will discuss an alternative solution in Section 5.2.3.

5.2.2 ***Term-Based Label Extraction***

After grouping salient sentences together, the groups or clusters should be given labels which clearly reflect the content in the clusters (Aker et al., 2016). Similar to the clustering approach, where we grouped salient sentences by the ontological term they share, we used the shared term as the label to represent the cluster. This is based on the assumption that the climate change terms which are annotated in the sentences do already elaborate the central meaning of the clusters. Listing 5.2 illustrates an example of the labels extracted by choosing the shared terms.

5.2.2.1 ***Labeling Evaluation for Term-based Clustering Appraoch***

5.2.2.2 Baseline

In the labeling evaluation, we compared the system labels against the baseline labels. We generated the baseline labels by applying *tf*idf*. It is a common approach in most information retrieval systems which consists of two main components, *tf* and *idf* (Ponte and Croft, 1998). In our experiment, *tf* indicates the frequency of terms occurs in a cluster[5]. *idf* presents the number of clusters in which the term occurs in. These components allow us to reduce common terms in the clusters and discover more discriminative

[5] Since sentences can contain more than one term it is likely that a cluster has several climate change terms.

Shared-Term Labels

polar ice
average temperature
habitat loss
mother nature
climate change
industrial revolution
water vapour
greenhouse gases
Environment
pollution
natural disasters
weather
surface temperature
global warming
hurricanes
melting
emissions
carbon emissions
warming
global temperature
climatology
ozone layer
climate
CO2
temperature
greenhouse effect
floods
cleaning
water level
sea level
electric car
ice age
deforestation
scientific evidence
scientific research
forest fires
methane
pollutants
power plants

Listing 5.2: An example of labels generated by choosing the shared terms in clusters

terms having fairly low term frequency in the clusters. To determine the candidate labels, we calculated the score for each term by the multiplication of *tf* and *idf*. The term with the top score was chosen as the cluster label.

5.2.2.3 Manual Labeling Evaluation

In the evaluation of cluster labels, we followed the manual evaluation method presented by Aker et al. (2016). We invited three participants having a background in Computer Science to evaluate the labels. Those are fluent in English and aged above 18 years old.

The evaluation was presented as an online form. The participants were asked to read the sentences in the given clusters and score the labels. The baseline and system labels were shown in random order. For each label, the participants were asked to answer five-point Likert scale questions, ranging from *strongly disagree* (1) to *strongly agree* (5). The questions include i) Question 1: By reading the label, I can understand it, ii) Question 2: This label is a complete phrase, and iii) Question 3: This label precisely reflects the content of the sentences in the cluster. Along with the three questions, we presented 13 clusters with a maximum of 10 salient sentences (so that the participants are able to read the content prior to the labeling evaluation) and a minimum of 2 salient sentences. Figure 5.4 illustrates the results of the labeling evaluation.

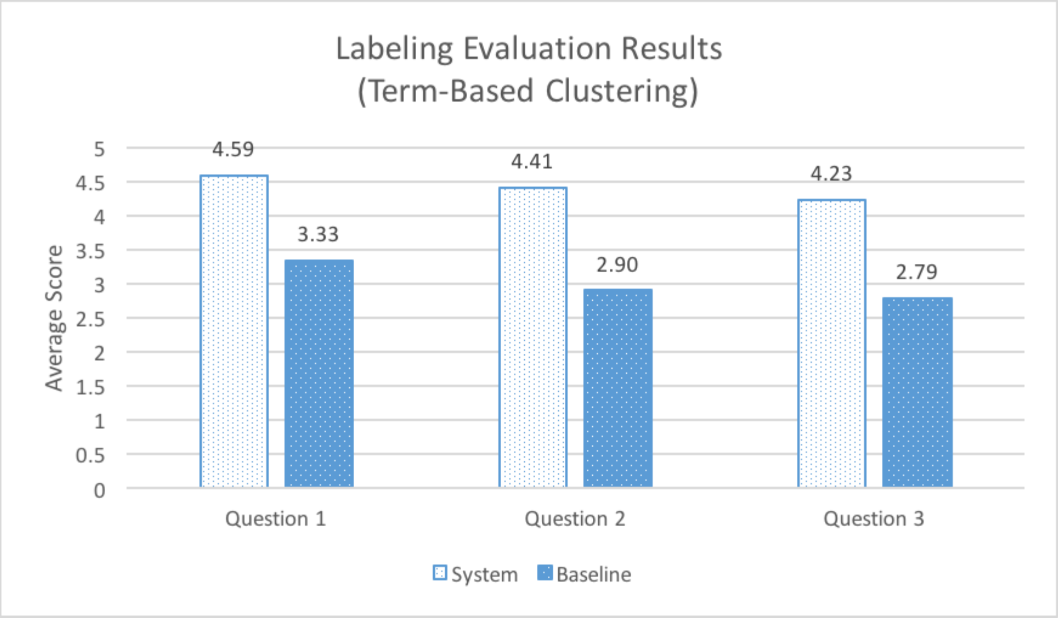


Figure 5.4: The labeling evaluation performed on the term-based clustering approach. The average preference scores generated by 3 participants on a scale 1: strongly disagree to 5: strongly agree

As we can see from the figure, overall, the quality of the system labels is higher than the baseline labels. In Q1, the system labels compared to the baseline labels are more understandable with the average score of 4.59 and 3.33 respectively. Likewise, in Q2, the system labels are more completed phrases than the baseline with a mean difference of 1.51. Lastly, with the average preference scores of 4.23 in Q3, the system labels reflect better the quality of the content in the clusters, than those generated by the baseline having the score of 2.79. Additionally, the quality of the system labels is further confirmed by a statistical significance analysis with Mann-Whitney U Test. The test reveals that significance difference is found in the system labels ($Md_{Q1-Q3} = 5$, $n_{Q1-Q3} = 39$) and the baseline labels ($Md_{Q1} = 4$, $Md_{Q2} = 3$, $Md_{Q3} = 2$), $U_{Q1} = 363$, $U_{Q2} = 343$, $U_{Q3} = 386$, $z_{Q1} = -4.25$, $z_{Q2} = -4.36$, $z_{Q3} = -3.92$, $p < 0.01$, $r_{Q1} = 0.48$, $r_{Q2} = 0.49$, $r_{Q3} = 0.44$. We also measured the inter-annotator agreement using

Krippendorff's alpha coefficient[6]. The agreement in Q1, Q2, and Q3 are 0.31, 0.27, and 0.35 respectively. We consider these agreement scores are sufficient to demonstrate the quality of the system labels compared to the baseline labels.

5.2.3 ***X-means Clustering***

In Section 5.2.1 we have shown that the idea of performing clustering based on shared terms results in poor clustering performance. In this section, we aim to overcome the problem of the poor performance of the term-based clustering approach by using X-means (Pelleg and Moore, 2000) clustering algorithm, an extended version of K-means, to cluster the salient sentences selected by the summarization system. One of the benefits of X-means is that it is able to automatically detect the number of clusters.

In order to automatically detect a number of clusters, the algorithm completes a set of repeated operations searching for the best scoring model, indicating a total number of clusters should be finally created. X-means begins with the running of the conventional K-means and then determines whether centroids should be split. There are two splitting strategies: choosing one centroid and choosing a half number of centroids. By applying a splitting strategy, the algorithm determines whether the model score improves after splitting. If the score improves, the splitting is accepted. Otherwise, the splitting is rejected. (Pelleg and Moore, 2000). The summary of these steps is shown in Figure 5.5. After BIC scores are calculated, the algorithm determines which centroids should be kept.

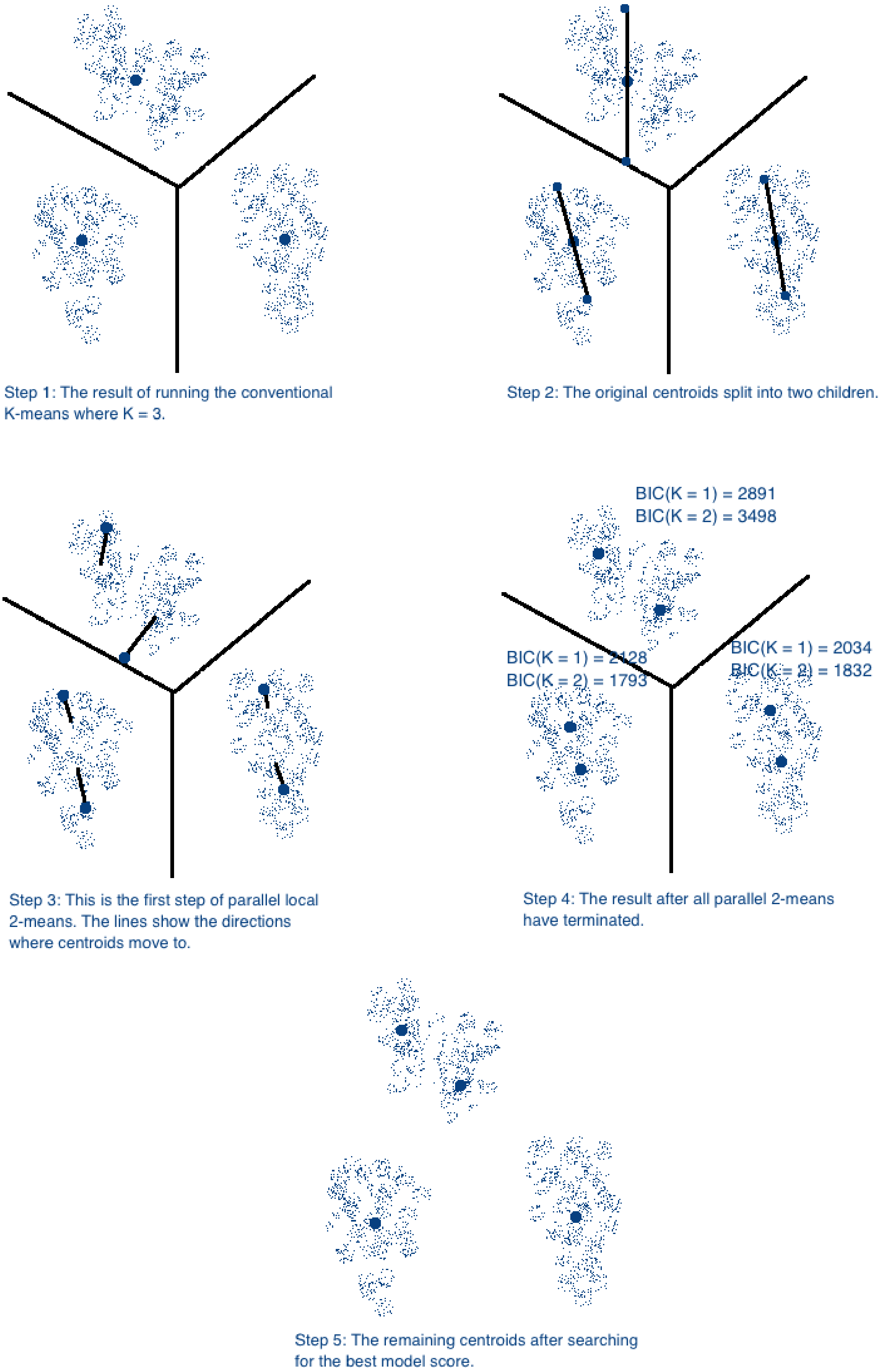
$$Pr[M_j|D] \tag{5.8}$$

where:

D refers to the input set of coordinates.

M_j is a set of alternative models derived from the results after applying different value of K.

[6] *nltk metrics*, `http://www.nltk.org/api/nltk.metrics.html`.



Step 1: The result of running the conventional K-means where K = 3.

Step 2: The original centroids split into two children.

Step 3: This is the first step of parallel local 2-means. The lines show the directions where centroids move to.

Step 4: The result after all parallel 2-means have terminated.

Step 5: The remaining centroids after searching for the best model score.

Figure 5.5: The steps of searching for the best model score. Adapt from "X-means: Extending K-means with Efficient Estimation of the Number of Clusters", by D. Pelleg & A. Moore, *Proceedings of the Seventeenth International Conference on Machine Learning*, 2-3, 2000.

$$BIC(M_j) = \hat{l}_j(D) - \frac{p_j}{2} \cdot \log_R \tag{5.9}$$

where:

$\hat{l}_j(D)$ refers to the log-likelihood of the data as of the *j-th* model, at the maximum point.

p_j is the number of parameters in M_j.

R is the number of coordinates.

$$\hat{\sigma} = \frac{1}{R-K} \sum_i (x_i - \mu_{(i)})^2 \tag{5.10}$$

$$\hat{P}(x_i) = \frac{R_{(i)}}{R} \cdot \frac{1}{\sqrt{2\pi}\hat{\sigma}^M} exp(-\frac{1}{2\hat{\sigma}^2}\|x_i - \mu_{(i)}\|^2) \tag{5.11}$$

$$l(D) = log\Pi_i P(x_i) = \Sigma_i(log\frac{1}{\sqrt{2\pi}\sigma^M} - \frac{1}{2\sigma^2}\|x_i - \mu_{(i)}\|^2 + log\frac{R_{(i)}}{R}) \tag{5.12}$$

In the processing of choosing a model score, the posterior probability shown in Equation 5.8 is determined along with the Schwarz criterion presented by Schwarz et al. (1978) shown in Equation 5.9. In addition, the maximum likelihood estimate (MLE), the coordinate probabilities, and the log-likelihood of the data are applied as shown in Equation 5.10 - 5.12 respectively.

By fixing the value of n between 1 and K, $1 \leq n \leq K$, on a set of coordinates D_n in a centroid n with the maximum likelihood estimate applied, the result yields as shown in Equation 5.13. The formula is used to calculate the BIC scores when centrods are split as aforementioned.

$$\hat{l}_j(D_n) = -\frac{R_n}{2}log(2\pi) - \frac{R_n \cdot M}{2}\log(\hat{\sigma}^2) - \frac{R_n - K}{2} + R_n\, logR_n - R_n\, logR \tag{5.13}$$

5.2.3.1 ***Similarity Measurement***

To enable X-means to process the clustering, a similarity needs to be defined to determine which sentences are close to each other. In the definition of our similarity measurement, the automatic selected salient sentences are transformed into vectors using the vector space model. In the document indexing stage, we employed the ontologies to automatically annotate key climate change terms in the SSSD. The employment of ontology-based approach benefits the transformation of words to vectors by capturing relevance of specific topics. We derived 64 significant climate change topics. Term frequency was counted for each term to generate vectors for each sentence. To generate a similarity matrix, cosine similarity measure was used to calculate cosine similarity scores among the vectors. After the similarity matrix was constructed, we applied a Principal Component Analysis (PCA)[7] for dimensionality reduction.

5.2.3.2 ***X-means Clustering Evaluation***

Similar to the term-based clustering we evaluated the results of the X-means clustering using the silhouette. The mean silhouette coefficient is derived from the calculation based on the similarity definition obtained by the ontology-based vector space model. We achieved a high coefficient score of 0.9878, with the total number of 19 generated clusters. As discussed in the previous section, a mean silhouette coefficient close to 1.0 indicates that the average distance from a coordinate in a cluster to the other coordinates within its own cluster is less than the average distance to all coordinates in the nearest cluster (Rousseeuw, 1987). In our experiment, we concluded that the clustering results obtained by the X-means clustering algorithm have a strong clustering structure.

5.2.4 ***Label Generation with Mutual Information***

To generate labels from the X-means clusters we could have followed the same approach as described in Section 5.2.2, namely selecting a term that is shared by all or majority of the salient sentences within a cluster. We tried this. However, to our surprise, the performance was very low compared to what we achieved in Section 5.2.2. Nevertheless, this helped us draw two conclusions. First, the performance in Section 5.2.2 is high because the labels were so selected that all salient sentences within a group shared that label. Second, the size of the clusters was not big so that the label had high chance to

[7] sklearn.decomposition.PCA: https://goo.gl/QqiWec

be representative of the cluster. This simulation changed once the cluster size increased and also the salient sentences covered several different climate change terms. Because of this, selecting a label was not about simply selecting the term that appears in all or in the majority of the salient sentences. Instead, we used Mutual Information (MI) to make this decision for us.

MI is a prevalent feature selection approach that involves the calculation of a utility measure A(t,c). MI quantifies how much information term t is contributing to the correct classification judgment on class c (Manning et al., 2008). The MI formula is shown in Equation 5.14, where U is a random variable that holds the value e_t. If a sentence contains term t, the value of e_t is *1*. Otherwise, the e_t is 0. C is a random variable that holds the value e_c. The value of e_c is 1 indicating that a sentence is in class c and it is 0 if it is not. Table 5.4 and Equation 5.16 illustrate how to calculate a mutual information score for a term *climate* in a class X (Manning et al., 2008). Listing 5.3 shows an example of labels generated by Mutual Information.

$$I(U;C) = \sum_{e_t \in \{1,0\}} \sum_{e_c \in \{1,0\}} P(U = e_t, C = e_c) \log_2 \frac{P(U = e_t, C = e_c)}{P(U = e_t)P(C = e_c)'} \tag{5.14}$$

$$I(U;C) = \frac{N_{11}}{N} \log_2 \frac{NN_{11}}{N_{1.}N_{.1}} + \frac{N_{01}}{N} \log_2 \frac{NN_{01}}{N_{0.}N_{.1}} + \frac{N_{10}}{N} \log_2 \frac{NN_{10}}{N_{1.}N_{.0}} + \frac{N_{00}}{N} \log_2 \frac{NN_{00}}{N_{0.}N_{.0}} \tag{5.15}$$

To calculate the mutual information scores for candidate terms, we applied the maximum likelihood estimation of probability as shown in Equation 5.15 (Manning et al., 2008). From the equation, N refers to the counts of sentences in which their subscripts take the values of e_t and e_c. For instance, N_{01} refers to the number of sentences that do not contain term t ($e_t = 0$) but in class c ($ec = 1$). $N_{1.}$ is derived from the addition of N_{10} and N_{11}. N refers to the total number of sentences. In each cluster, we calculated the score of each candidate term. The term with the highest MI score was selected as the cluster label for that cluster. Note that in this work we only focused on unigram.

Table 5.4: An example of the values for a term *climate* in a class X

	$e_c = e_X = 1$	$e_c = e_X = 0$
$e_t = e_{climate} = 1$	$N_{11} = 37$	$N_{10} = 23{,}512$
$e_t = e_{climate} = 0$	$N_{01} = 129$	$N_{00} = 652{,}015$

$$\begin{aligned} I(U;C) = {} & \frac{37}{67,5693} \log_2 \frac{675,693 \cdot 37}{(37 + 23,512)\ (37 + 129)} \\ & + \frac{129}{675,693} \log_2 \frac{675,693 \cdot 129}{(129 + 652,015)\ (37 + 129)} \\ & + \frac{23512}{675,693} \log_2 \frac{675,693 \cdot 23,512}{(37 + 23,512)\ (23,512 + 652,015)} \\ & + \frac{652,015}{675,693} \log_2 \frac{675,693 \cdot 652,015}{(129 + 652,015)\ (23,512 + 652,015)} \\ & = 0.0000869501 \end{aligned} \tag{5.16}$$

Labels by MI

polar
penance
years
assumed
melting
habitat
taxing
climate
hair
taxing
temperature
reputable
global
greenhouse
cyclic
yes
weather
parties
whose
warming

Listing 5.3: An example of labels generated by the Mutual Information

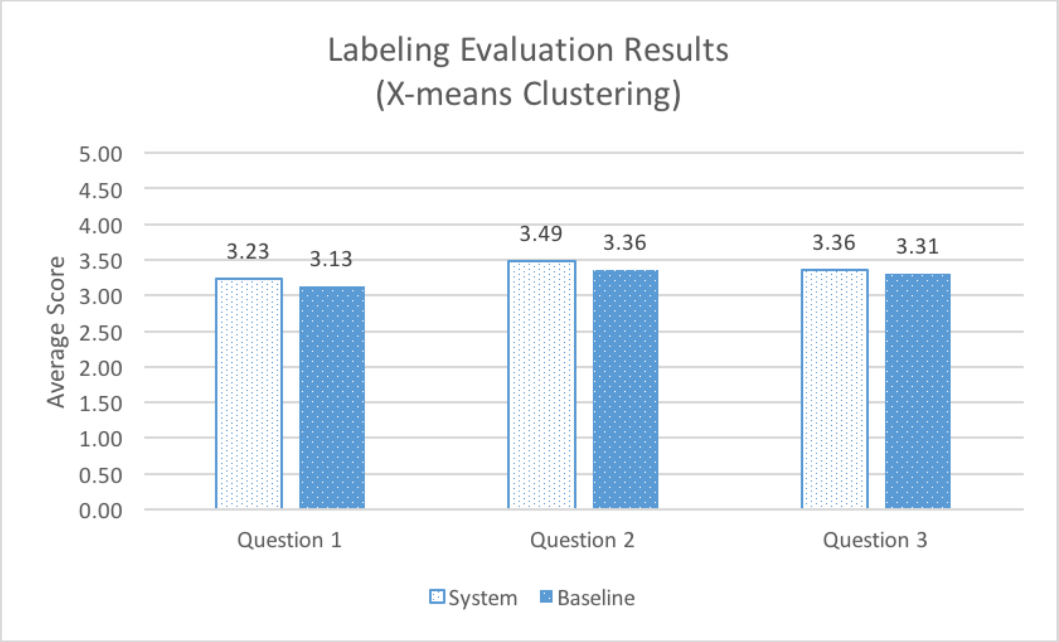


Figure 5.6: The labeling evaluation performed on the X-means clustering approach. The average preference scores generated by 3 participants on a scale 1: strongly disagree to 5: strongly agree.

5.2.4.1 ***Labeling Evaluation for X-means Clustering Approach***

In order to evaluate the system labels generated by the results derived from X-means clustering approach, we applied the same evaluation procedure as well the baseline discussed in Section 5.2.2.1. The results are illustrated in Figure 5.4. As can be seen from the figure, the average preference scores of the system are higher than the baseline. In Q1, the system labels are more understandable than the baseline, with the mean difference of 0.10. In Q2, the system labels more completed phrases than the baseline labels, with a higher mean score of 0.13. Lastly, in Q3, the system labels are still better than the baseline labels with the mean difference of 0.05. The system labels are more meaningful for presenting the central meaning of the content in the clusters. However, as there is a slight difference between the results of the system labels and baseline labels, Mann—Whitney U test reveals no significant difference, with the z values of -0.705, -0.427, and -0.389, with the significance levels of p= 0.481, 0.670, and 0.697 respectively. The values of Krippendorff's alpha, by another three participants, for Q1, Q2, and Q3 are 0.33, 0.44, and 0.56 respectively. We consider these agreement scores are high and demonstrate the quality of the system labels compared to the baseline labels.

5.2.5 *Visualization*

Chapter 3 discussed the investigation of various representation models for displaying or visualizing summaries of online debates. Unlike traditional summaries, the debate extracts have to capture the main concepts discussed on both sides of the arguments and enable the reader to look at those concepts from both the proponent and opponent sides. We proposed the Chart Summary which presents the clusters by bars. Each bar is marked with a cluster label. The previous sections illustrated how components in the Chart Summary are generated. In this section, we combine those components as a Chart Summary.

In the generation of the bars in the Chart Summary, the bars are the clusters that express related content on the two opposing sides. Therefore, it is important to match clusters from the two opposing sides which express the related content. We refer this approach as *alignment*. From the two opposing sides, we align the clusters based on the cluster labels. The clusters sharing mutual labels are aligned. For alignment, we used the cosine similarity over vector spaces representing the labels. The vector also contains semantically related words enriched from WordNet. Clusters which have no pair will not be aligned and thus will not be presented in the Chart Summary. Once the pairs of aligned clusters are derived, we count the number of salient sentences in those clusters, separately in each opposing side. Those numbers represent the frequencies of the bars.

After all components of a Chart Summary are completely generated, they are exported to NVD3 JAVA script[8] for the purpose of visualizing the Chart Summary. Figure 5.7 - 5.8 illustrate a Chart Summaries from the term-based and X-means clusters. The summaries run on a web browser[9].

[8] `http://nvd3.org`

[9] An example of a Chart Summary can be accessed via https://goo.gl/wjBh7V.

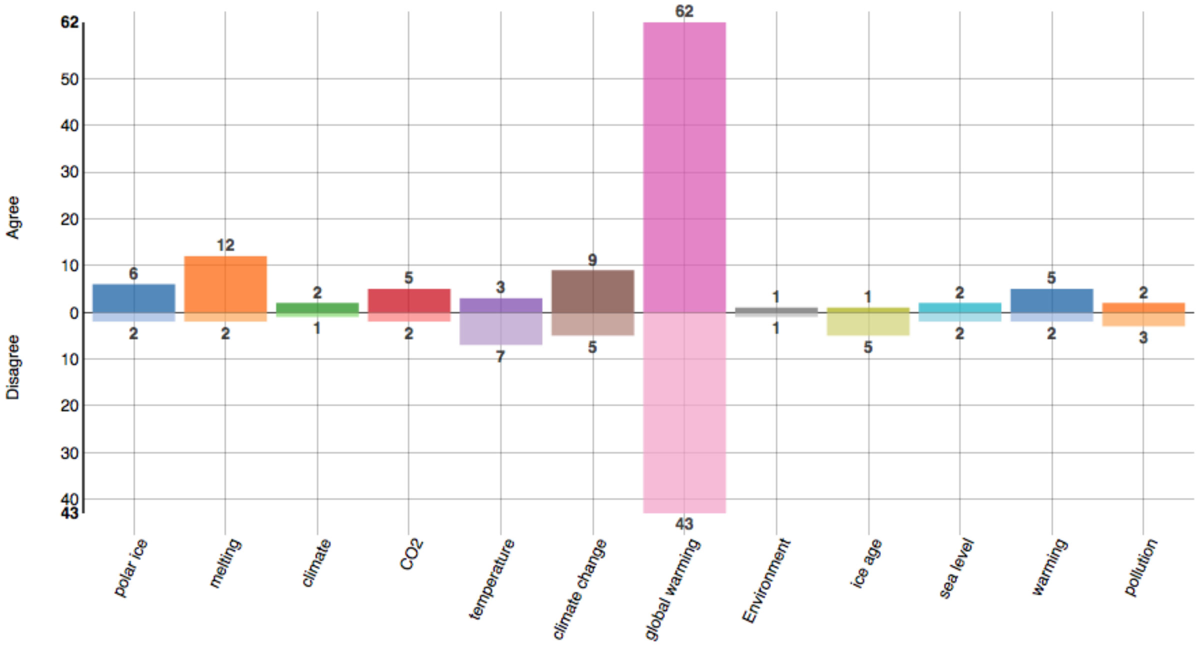


Figure 5.7: Chart Summary for debate data derived from the term-based clusters

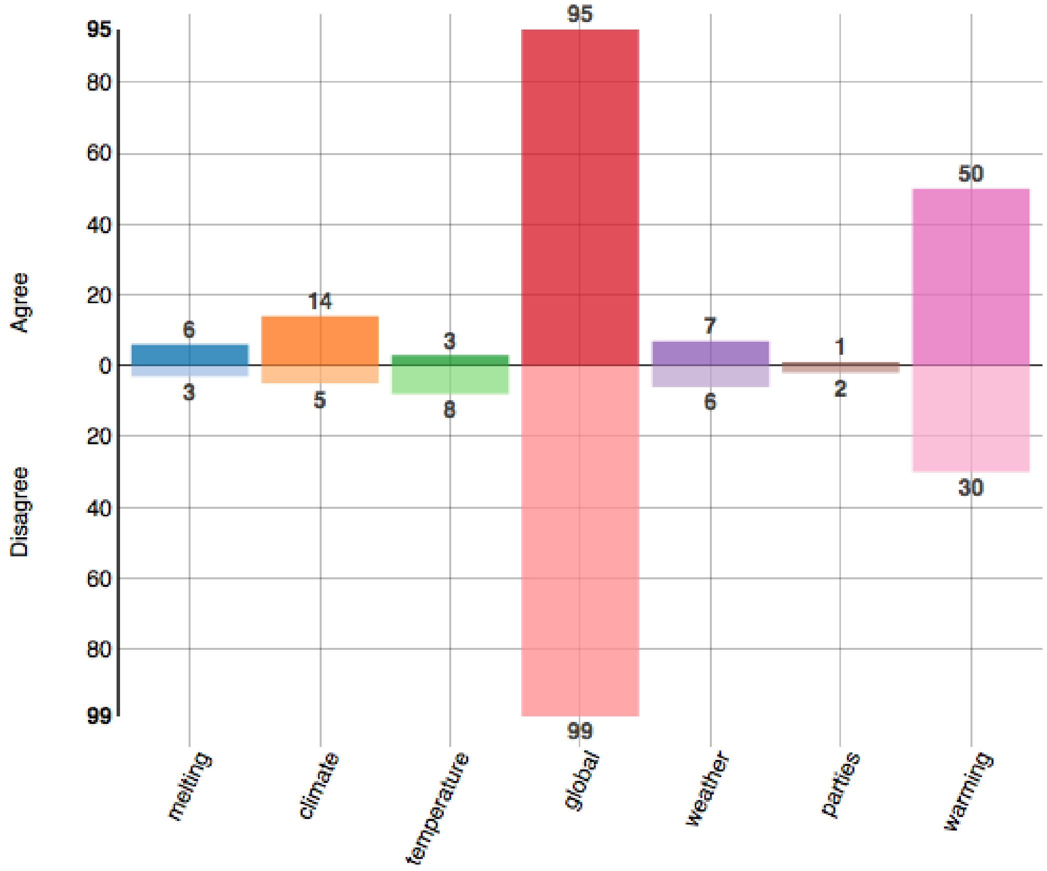


Figure 5.8: Chart Summary for debate data derived from the X-means clusters

5.3 Summary

In this chapter, we discussed the first and the second processes for generating Chart Summaries. The first process began with the collection and annotation of the SSSD. The data was used as the input for the automatic salient sentence selection. In this process, we defined a set of key features to help extract salient sentences from online debate comments. Sentence position yields highest ROUGE score and thus plays the most important role in the selection. The salient sentences which were automatically selected through the sentence position feature will be used in the subsequent experiments.

In the second process, we aimed to generate Chart Summaries which represent the high-level topics of online debates. The Chart Summary is composed of three main components, including the bars, labels, and frequencies of the bars. We proposed a clustering and cluster labeling pipeline to guide the debate summary generation.

In our approach, we used an online service to automatically annotate climate change terms in salient sentences and to group related salient sentences into the same cluster. For the clustering, we investigated two variants both making use of ontological terms. The first, a simple approach, groups salient sentences by shared terms. The second approach applies X-means clustering. The evaluation has shown that the X-means approach is a better choice for clustering.

For the label generation, we created labels to represent each cluster. Again here we investigated two different approaches both making use of ontological terms. The first approach, again a simple one, labels each cluster with the term shared by all members of the cluster. The second approach picks the best term according to Mutual Information (MI). The manual evaluation showed that the simple approach achieves higher results than the MI one. However, as discussed the simple approach achieved high results because of the size of the clusters and led to poor results when the size of the clusters grew which is the case with the X-means clustering. Once the clusters and labels are generated with the alignment of the agree and disagree parts, we visualized the results using NVD3.

In the next chapter, we enrich the Chart Summaries with additional details such as enabling the users to see example debates for each cluster. When a user clicks on a bar in a Chart Summary, a Side-By-Side Summary of a topic of the clicked bar is shown. We view the generation of Side-By-Side Summary as a contradiction detection task.

Chapter 6

SIDE-BY-SIDE SUMMARY GENERATION

The previous chapters presented how salient sentences are selected and clustered for the purpose of Chart Summary generation. In the next process, we focus on the generation of the Side-By-Side Summary, which we regard as a contradiction detection task. The clustering results derived from the previous process are further processed in this task. In this chapter, we begin the discussion with how data is prepared and annotated for the contradiction detection task. We also discuss an experiment to automatically generate the Side-By-Side Summary.

6.1 Side-By-Side Summary Generation as a Contradiction Detection Task

The Side-By-Side Summary consists of a list of sentence pairs. Each pair consists of two sentences from the Agree and Disagree side, showing disputed information, called rebuttal. The generation of rebuttals requires two sentences to contradict each other. For this reason, we view the extraction of rebuttals as a contradiction detection task. This is a subtask of classifying sentences in Recognizing Textual Entailment (RTE), represented as the relationship between hypothesis and text (Marneffe et al., 2009).

6.1.1 Data Annotation for Contradiction Detection Task

In order to create data for the contradiction detection task, we employed the results from the previous process which are the clusters of salient sentences. We created two datasets. One dataset was annotated from the term-based clustering results and the other was from the X-means clustering one. The datasets were prepared by creating pairs of RTE sentences, text (T) and the hypothesis (H), in each cluster. Sentences from one opposing side were paired with those on the other side. The longer sentences were considered to be the text and the shorter sentences were chosen to be the hypothesis (Lendvai et al., 2016). Once the data preparation was completed, the pairs of H-T sentences were annotated with one of the two entailment relations, *Contradiction* or *Non-Contradiction*. The following sections elaborate how each entailment relation is defined.

6.1.1.1 Contradiction Relation

In each cluster, we pair the Agree sentences to those on the Disagree sides. A pair of sentences consists of hypothesis and text. In the justification whether a pair of sentences is contradictory, the information expressed in the hypothesis is focused. If a piece of information in the text compared to that of the hypothesis appears to be false or wrong, the sentence pair is considered as contradictory. The following pair exemplifies a contradiction: the hypothesis expresses that global warming is real but the hypothesis contrarily opposes.

<**h**> Yes global warming is **real**. </**h**>
<**t**> Global warming is a **myth**! </**t**>

Another example for the contradiction relation is shown below. Whereas the hypothesis on global warming is not about a political move, the text expresses that global warming is a way to get tax dollars. The issue expressed in the text may be potentially influenced by a government launching a policy to collect money from its citizen. As the hypothesis states that the information is not related to politics and the text is considered as expressing politics, the information expressed in both sentences cannot be true at the same time. For this reason, this example is annotated as *contradiction*.

> <**h**> Global Warming was **never a political move**. </**h**>
>
> <**t**> Global Warming is just a cover up **to get our tax dollars**. </**t**>

Moreover, sentence pairs are considered as contradictory when they have disagreement on an issue discussed in the hypothesis. For instance, the following pair exemplifies the contradiction based on a science topic. In an issue of the existence of global warming, the main information in the hypothesis is global warming exists due to NASA. In contrast, the information in the text expresses that global warming does not exists because there is no scientific evidence. This sentence pair is therefore annotated as contradiction.

> <**h**> Even **NASA** agrees that global warming exists. </**h**>
>
> <**t**> Yes, global warming is a myth, because they have **not really proven the science behind it**. </**t**>

6.1.1.2 ***Non-Contradiction***

This entailment relation is contrary to the contradiction relation. A pair of sentences is annotated as *Non-Contradiction* when a piece of information in both sentences are not contradictory. For instance, the following pair of sentences does not show the contradiction of the information. They only entail each other' s content – expressing the same directions of agreement on the existence of global warming based on the same topic. The pair exemplifies that both sentences agree that the temperature is rising.

> <**h**> The earth is getting **warmer**. </**h**>
>
> <**t**> It is statistically shown that over the century, or even decade, the global **temperatures have risen**. </**t**>

Furthermore, a pair of sentences is considered in an *Non-Contradiction* when a conclusion cannot be made. The information expressed in the sentences does not indicate the agreement or contradiction in the content. The following example illustrates the text and hypothesis express different topics. One mentions the water level and the other express the temperature on the planet.

<**h**> The oceans **water level** have increased by up to 7 inches. <**h**>

<**t**> Since 1850 (end of the Little Ice Age) planet has **warmed 0.89 Degree Celsius**. </**t**>

The data annotation for the contradiction detection task was partially performed by two annotators. The inter-annotator agreement between the annotators was $\kappa = 0.59$ which is considered as *moderate agreement* according to the scales defined by Landis and Koch (1977). In the experiment, two datasets were annotated. One dataset was annotated from the term-based clustering results and the other was from the X-means clustering one. In the annotation, given a pair of H-T sentences, the pair was manually annotated with one of the two entailment relations guided above. After all the pairs of sentences were annotated, we derived two datasets as shown in Table 6.1. To simply reference the two datasets later, we named the datasets based on the clustering results where they were created from, Debate Entailment Dataset from the Term-based clustering approach (DEDT) and Debate Entailment Dataset from the X-means clustering approach (DEDX)[1]. The percentages of the contradiction and non-contradiction relations in the DEDT are 41.26% and 58.74% respectively. Also in the DEDX, the percentages of the contradiction and non-contradiction relations are 15.28% and 84.72% respectively. Note that the size of the dataset from the term-based approach is smaller because the clustering approach only captures the sentences containing climate change topics and the rest of the sentences are ignored. Thus, the number of RTE pairs is smaller than those of the X-means approach.

Table 6.1: Statistical Information of the RTE Corpora

Entailment Relations	**DEDT**	**DEDX**
Contradiction	966	1,412
Non-Contradiction	1,375	7,827
Total	2,341	9,239

[1] We performed the annotation for two separated datasets with the two main reasons. First, it is possible that, for example, a salient sentence, which should be clustered to a cluster A, is wrongly clustered to another cluster. When the H-T sentence pairs are created and annotated, that sentence may not have a contradiction relation with the other sentences as it should be. This results in the second reason that we need to increase the number of training data as we expected more contradiction relation pairs could be found.

6.2 Contradiction Detection

6.2.1 Logistic Regression Classifier

The aim of this experiment is to generate a summary of online debates, Side-By-Side Summary. In the generation of the summary, it is important to detect rebuttals, issues which are argued in the two opposing sides. We view this task as a binary classification problem – to classify whether the given H-T pairs are *contradiction* or *non-contradiction*. In this experiment, we apply a logistic regression package available in Python *scikit-learn*[2] to create the classifier.

6.2.2 Feature Definition

The following features were experimented in the classification of the sentence pairs. There are 9 features in total.

1. **Alignment Score with Dependency Parsing (AS).** In order to detect contradiction, it is significant for a system to understand the meaning of words which are expressed in hypothesis and text as much as possible. In this stage, we conduct a dependency parsing for text and hypothesis to determine semantics and syntax of the sentences. We employ Spacy[3] to parse the sentences. It is a speedy and an accurate parser which is compatible with Python.

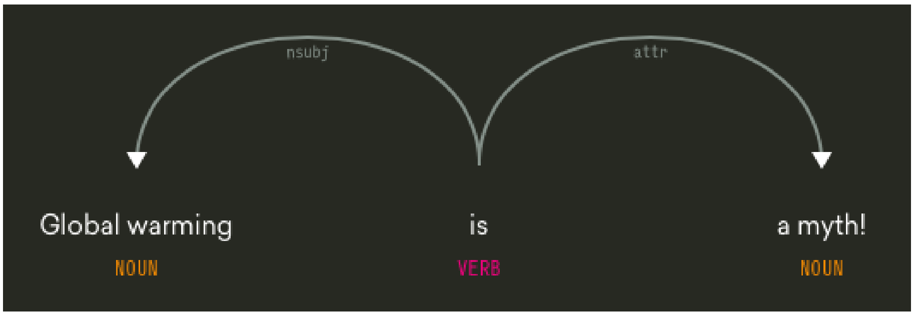


Figure 6.1: An example of the parsing results for a hypothesis sentence

After parsing, a dependency graph is used to represent the parsing results for each sentence. It outlines a semantic structure of a sentence which includes words of sentences are connected with different grammatical relations (e.g *nsubj* indicates

[2] http://scikit-learn.org/stable/

[3] https://spacy.io

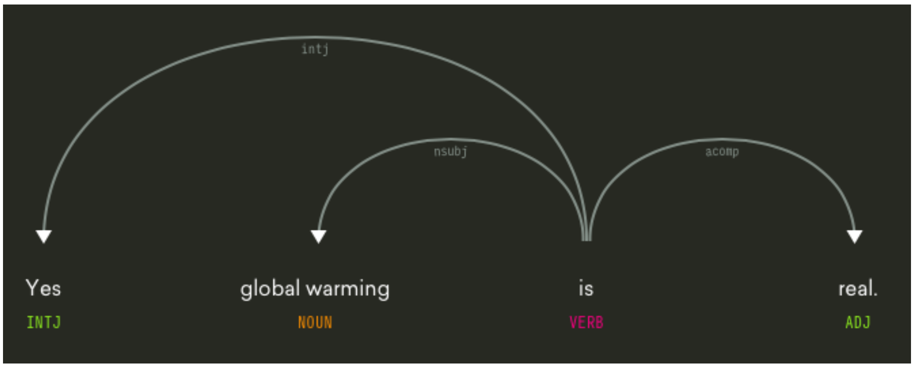


Figure 6.2: An example of the parsing results for a text sentence

nominal subject) (Kübler et al., 2009). Figure 6.1 and Figure 6.2 illustrate examples of the parsing results from a hypothesis and a text. Words in the sentences are represented as nodes linked by dependency relations[4].

After hypothesis and text are transformed into dependency graphs, they are aligned with each other. In the concept of dependency graph alignment, each node in the hypothesis is mapped to a unique node in the text. If a node has no similar pair, it is ignored (de Marneffe et al., 2008). The result of each alignment is represented as an alignment score, indicating the degree of similarity of each hypothesis and text pair. We adopt a graph alignment methodology from Partha Pakray and Bandyopadhyay (2011) to obtain an alignment score for each sentence pair, in the range of [0, 1]. The score of 1 indicates that complete matches are found in the alignment of hypothesis and text nodes. The partial match between the nodes yields the score of 0.5. The score of 0 indicates that no match is found. For example, the subjects ($warming_h$, $warming_t$) and verbs[5] (is_h, is_t) shown in Figure 6.1 and Figure 6.2 are aligned. Other nodes including *myth, yes, and real* which have no pairs are neglected. Based on the graph alignment methodology presented by Partha Pakray and Bandyopadhyay (2011), the alignment score for this pair is 1.0. The following conditions, as proposed by Partha Pakray and Bandyopadhyay (2011), elaborate how alignment scores are defined. The symbols illustrated in Table 6.2 are used in the explanation of the conditions.

[4] We employed *displaCy* to visualize the parsing results, `https://demos.explosion.ai/displacy/`.

[5] A library presented by Schrading (2016) is used to determine parts of speech of the sentences from the dependency parsing results.

Table 6.2: Symbols used in in the description for alignment scores

Symbols	**Explanation**
S_h	hypothesis subject
V_h	hypothesis verb
O_h	hypothesis object
S_t	text subject
V_t	text verb
O_t	text object

- **Subject-Verb Alignment.** The system identifies subjects through the relations *nominal subject* (nsubj), *nominal subject (passive)* (nsubjpass), *clausal subject* (csubj), *clausal subject (passive)* (csubjpass), *noun compound modifier* (agent), and *expletive* (expl). Then the system compares S_h and V_h to S_t and V_t. An alignment score of 1 is assigned in case of the complete match. Otherwise, the following conditions are considered.

- **WordNet Distance for Subject-Verb Alignment.** If S_h and S_t do match in the Subject-Verb Alignment condition but V_h and V_t do not match, we calculate WordNet distance of the two verbs. The lower the distance, the closer the two verbs. If the distance is less than 0.5, the alignment score of 0.5 is assigned. Otherwise, we consider the following Subject-Subject Alignment condition.

- **Subject-Subject Alignment.** The system compares S_h and S_t. A complete math returns the alignment score of 0.5.

- **Object-Verb Alignment.** The system returns the alignment score of 0.5 when O_h and V_h match O_t and V_t respectively.

- **WordNet Distance for Object-Verb Alignment.** The system compares O_h to O_t. If they match, the system further compares the verb, which is related to O_h, to the verb related to O_t. If the verbs do not match, WordNet distance between the verbs are calculated. If the distance is lower than 0.5, the alignment score of 0.5 is returned.

- **Cross Subject-Object Alignment.** In some occasions (i.e. passive form), a user may refer an object in hypothesis as a subject. In the text, another user may mention a subject as an object. We therefore perform a cross check for such circumstance. If S_h and V_h match O_t and V_t or O_h and V_h math S_t and V_t, the score of 0.5 is allocated.

- **Prepositional Phrase Alignment.** The system identifies the prepositional phrases in text and hypothesis. The alignment score of 1 is assigned is the completed match is found.
- **Determiner Alignment.** Determiners in text and hypothesis are checked in the same manner as the prepositional phrases. If the matching is found, the alignment score of 1 is allocated.

2. **Longest Common Subsequence (LCS).** The size of Longest Common Subsequence indicates the longest string which is commonly shared in hypothesis and text. The LCS value is derived from 1) determining the size of longest common string and 2) divide it by the size of the longer sentence (Marques, 2015).

3. **Edit Distance (ED).** The distance between hypothesis and text is defined as the concept of edit distance. By determining the distance, a system is able to recognize textual entailment (Kouylekov and Magnini, 2005). The edit distance between text and hypothesis is defined as the minimum number of characters for inserting, deleting, and editing operations to transform text t to hypothesis h (Manning et al., 2008). The package *editdistance*[6] available in Python is used to create this feature.

4. **Ontological Term Overlap (OTO).** Ontological terms derived from the environmental service (see Section 5.2.1) are also beneficial for the problem of recognizing textual entailment. We determine the ontological term overlaps in hypothesis and text and report the overlap score as Jaccard similarity, ranging between *0* and *1*. As shown in Equation 6.1, Jaccard similarity is measured by the intersection of two sets, A and B divided by the union set of A and B (Manning et al., 2008). Jaccard similarity was used by Marques (2015) in feature definition and it yields great results.

[6] `https://pypi.python.org/pypi/editdistance`

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{6.1}$$

5. **Modal Verb Overlap (MVO).** Inspired by Marques (2015), the identification of modal overlap is explored to detect contradiction in text. Marques (2015) determines Jaccard similarity for the modal verbs overlapping in text and hypothesis. Examples of modal verbs include *can*, *could*, *may*, *might*, *must*, *will*, *should*, *would*, and *ought to*.

6. **Negation Overlap (NO).** Another feature to help identify a contradiction relation is negation. It was viewed as an important feature for detecting contrastive viewpoints by Paul et al. (2010). Negation shifts the direction of a sentence from a positive polarity to a negative polarity (Ikeda et al., 2008). In this feature, negations (i.e. *no*, *not*, *n't*, *never*, *none*, and *nothing*) are extracted. Jaccard similarity shown in Equation 6.1 is used to determine the negation similarity in text and hypothesis.

7. **Antonyms (ANT).** The presence of an antonym in text and hypothesis can also cause the contradiction. A list of antonyms is derived from WordNet. For each word in a hypothesis, the system determine whether an antonym of that word presents in the text. If one or more antonym presents in a pair of H-T sentences, the score of the feature is *1*. Otherwise, the score is assigned to *0*. The calculation of a score for this feature is summarized in Equation 6.2.

$$Score = \begin{cases} 1, & no\_ant >= 1 \\ 0, & no\_ant < 1 \end{cases} \tag{6.2}$$

where:

no_ant refers to a number of antonyms in a pair of H-T sentences.

8. **Negated-Term Parsing (NTP).** This feature makes use the dependency parsing from the previous feature. For each comparable pair of nodes, the system recognizes negated terms and determines whether the terms in the other sentence

are negated. For instance, as shown in Figure 6.3, the system determines a negation which is tied to the verb, *is*, in the text. It then checks whether the verb *is* in the hypothesis are also negated. If the a negated term is found in either a text of hypothesis but not the other, the system returns the value of *1*. Otherwise, the value of *0* is returned. Equation 6.3 summarizes how the score for this feature is calculated.

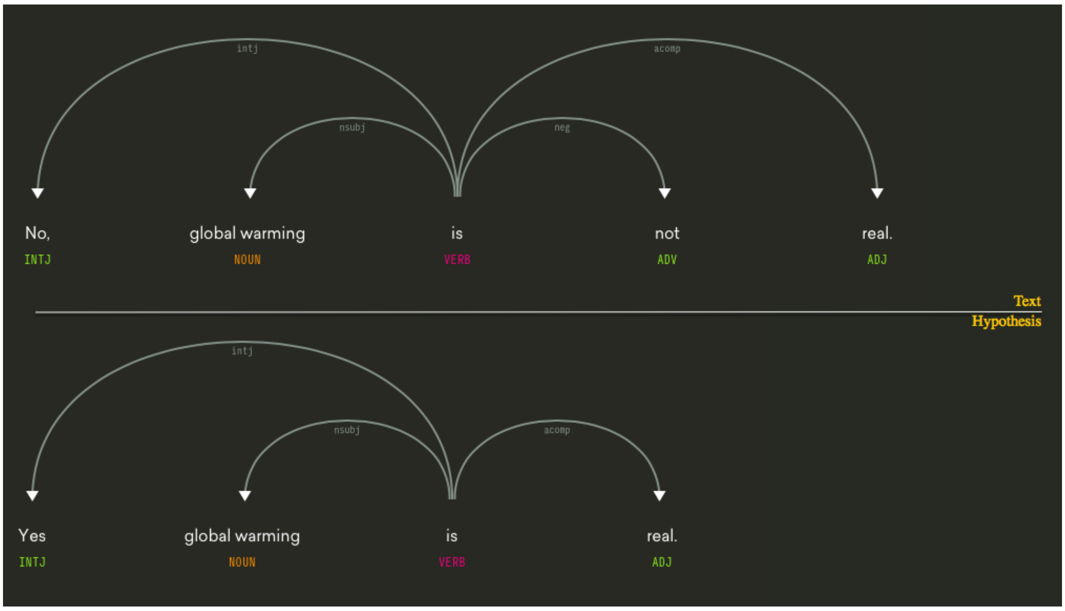


Figure 6.3: The determination of negated terms

$$Score = \begin{cases} 1, & (\neg Term_h) \wedge (Term_t) \mid (\neg Term_t) \wedge (Term_h) \\ 0, & \text{otherwise} \end{cases} \tag{6.3}$$

where:

$\neg Term$ refers to the term recognized as negated.

$Term$ is none negated term that is used to compare to $\neg Term$.

9. **Structural Feature (SF).** This feature was used by de Marneffe et al. (2008) to determine whether syntactic structures in text and hypothesis cause a contradiction relation. This is similar to the *Cross Subject-Object Alignment* defined in the previous section in a condition, if S_h and V_h match O_t and V_t or O_h and V_h math S_t and V_t, the feature score is returned as *1*. Otherwise, the score of *0* is assigned.

6.2.3 Experimental Results and Discussion

To evaluate the performance of the classifier, StratifiedKFold[7], a package available in Python was used. This package employs a k-fold cross validation approach which divides the training and test examples for each fold as shown in Equation 6.4. In the division of examples in each fold, a stratified option is applied – meaning that the folds have nearly the same proportion of each class (StratifiedKFold, 2018).

$$\begin{aligned} training\ size &= total\ examples\ - (\frac{total\ examples}{k}) \\ test\ size &= \frac{total\ examples}{k} \end{aligned} \tag{6.4}$$

Table 6.3: The amount of training and testing examples for each fold, with k = 4

Datasets	Contradiction	Non-Contradiction	Total	Train	Test
RTE1_DEV1	145	142	287	215	71
RTE1_DEV2	140	140	280	210	70
RTE2_DEV	400	400	800	600	200
RTE3_DEV	388	412	800	600	200
DEDX	1412	7827	9239	6929	2309
DEDT	966	1375	2341	1755	585

In the evaluation, 4-fold cross validation was applied. The number of $k = 4$ is used by Lendvai and Reichel (2016) and Lendvai et al. (2016) in the detection of detecting disputed information and contradiction in text. In each fold, the sizes of training and test examples are calculated according to the number of k supplied. The classifier is trained and tested by these examples. The results obtained in this fold are recorded and aggregated for the creation of an aggregated confusion matrix in the final step. This process is repeated until the number of k is satisfied. Figure 6.3 summarizes the sizes of the training and test examples for each fold. To evaluate the performance of the classifier, Stanford's datasets[8] indicating with *RTE* are also used in this experiment. Note that

[7] `http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html`

[8] `https://nlp.stanford.edu/projects/contradiction/`

the Stanford's datasets consist of 3 classes: contradiction, entailment, and unknown. In the purpose of evaluation in the binary classification task, the non-contradiction class is recasted by combining the entailment and the unknown classes.

Table 6.4: Possible combinations of features that maximize F1 scores in each dataset

Datasets	**Total Comb.**	**F1**	**F1 (All)**	**Min**	**Max**	**LCS**	**ED**	**NO**	**MVO**	**OTO**	**NTP**	**ANT**	**SF**	**AS**
RTE1_DEV1	16	0.60	0.45	2	6	0	0	50	50	50	100	100	50	0
RTE1_DEV2	5	0.67	0.55	2	3	0	0	60	60	60	0	0	0	0
RTE2_DEV	7	0.67	0.54	2	4	0	0	57.14	57.14	57.14	100.00	0	0	0
RTE3_DEV	120	0.68	0.61	2	7	0	0	52.50	52.50	52.50	52.50	52.50	52.50	52.50
DEDX	16	0.19	0.18	2	6	100	0	50	50	50	0	50	0	100.00
DEDT	8	0.37	0.36	5	8	100	100	50	50	50	100	100	0	100
				Macro Average		33.33	16.67	53.27	53.27	53.27	58.75	50.42	17.08	42.08

As our objective is to generate Side-By-Side Summary for debate data, we focus on detecting the contradiction class. The system was evaluated with precision, recall, and F1 measures. Precision indicates how many contradiction examples are correctly predicted by giving all classes. Recall refers to from all contradiction examples how many contradiction examples are correctly captured. F1 measures the harmonic mean of precision and recall. We evaluated the system performance against different datasets and aimed to investigate which combinations of features that maximize the F1 scores of the contradiction class. We consider a combination is a feature vector which contains at least two features. In this experiment, we created $2^n - 10$ combinations of features where n indicates the number of actual features in the system. The first feature is excluded because no combination of features is created (zero feature vector). The other nine features indicate the nine single features. The contradiction results are illustrated in Table 6.4. The **Total Comb.** refers to the number of possible combinations of features that maximize the F1 scores. **Min** and **Max** refer to minimum and maximum numbers of features that are combined to obtain the **F1** scores respectively. **F1 (All)** illustrates F1 scores derived by the combination of all nine features. Other columns show the proportions of individual features contributing to the maximization of F1 scores, in all possible combinations.

From the table, it can be clearly seen that the investigation of the possible combinations of features (F1) outperforms the combinations of all features (F1 (All)). The possible combinations of features can be created by two to eight features. According to the experimental results derived from each dataset, the key feature is *Negated-Term Parsing* which contributes to 58.75%. Other important features contributing to nearly the same scores are *Negation Overlap*, *Modal Verb Overlap*, and *Ontological Term Overlap*. The least important features are *Edit Distance* and *Structural Feature* features which contribute to approximately 17%. Moreover, another conclusion that can be made after analyzing the classification results is that the system cannot detect contradiction in the examples containing complex sentences. This results in the higher F-1 scores when the evaluation is performed on the Stanford's datasets. Additionally, the results derived from DEDT are better than those of the DEDX due to the same fact that the majority of the sentences in the DEDT are shorter than those of the DEDX.

6.2.3.1 Error Analysis

In order to measure the performance of the results, we perform an error analysis and report detailed information about classifier performance, in the form of an aggregated confusion matrix as shown in Figure 6.4[9]. The aggregated confusion matrix is created by summing up the individual confusion matrix from each fold. The average F-1 scores are calculated based on the aggregated confusion matrix (Kelleher et al., 2015; Yen et al., 2010). In the matrix, the rows correspond to the actual labels and the columns refer to the predicted labels.

From the figure, it can be clearly seen that the classifier is not efficient to detect contradiction class. For the contradiction class, out of 966 examples, the classifier was able to detect only 262 correct examples, accounting for 27.12 % of the examples. In addition, a majority of 72.88% examples was incorrectly classified as contradiction. This is similar to the classification of the non-contradiction class where the classifier was not successfully detect the class. Only 186 examples, accounting for 13.53 % were correctly classified. The missclassification rate was high, accounting for 86.47 %.

[9] In this experiment, we explored different combinations of potential features and there are more than one combinations that yield the highest F-1 score. In this analysis, based on the highest F-1 score, we created an aggregated confusion matrix from the combination yielding the maximum score on classifying the contradiction class. The confusion matrix here is generated from the DEDT dataset.

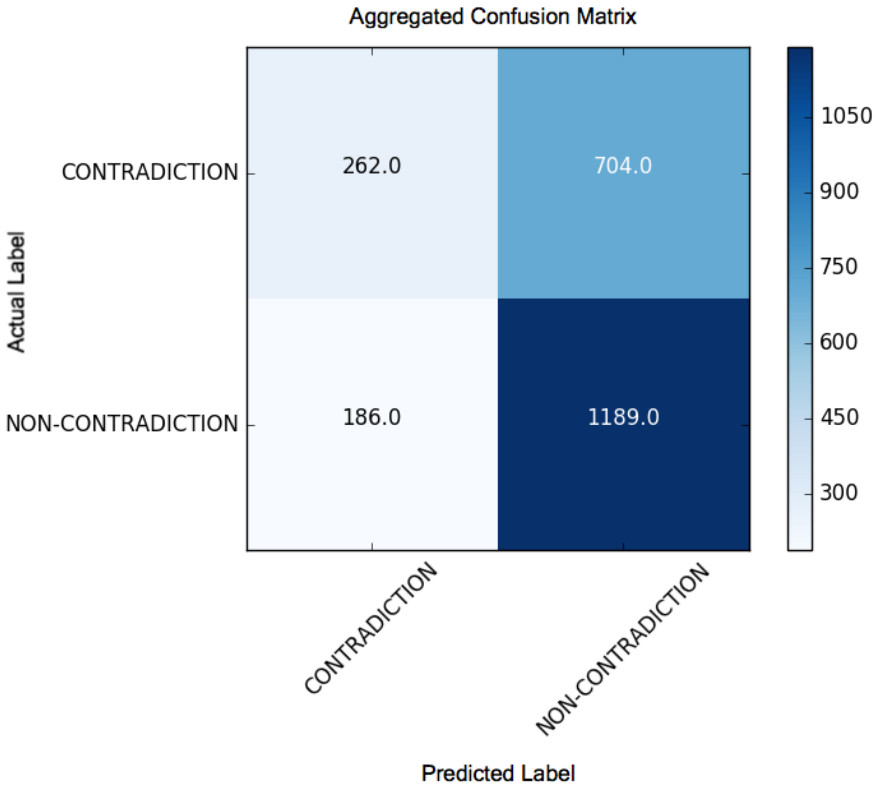


Figure 6.4: The aggregated confusion matrix derived from the evaluation of the DEDT dataset

Unfortunately, as shown in the confusion matrix, the classifier mistakenly classified many examples from the non-contradiction class, as being in the contradiction class. The examples below fall in this category. As the performance of the classifier was low at identifying the correct classes of the examples, potential improvements discussed below would help address the problem.

Example 1:

<**h**> I'm not sure if anyone thinks it doesn't anymore. </**h**>

<**t**> Now it is in the middle of December and, in my opinion, it's quite chilly out. </**t**>

Example 2:

<**h**> I personally don't believe in global warming. </**h**>

<**t**> I'm not sure if anyone thinks it doesn't anymore. </**t**>

1. **Removing non co-reference sentence pairs.** The examples of the H-T pairs above are actually annotated as belonging to the non-contradiction class. This is because the H-T sentences do not contain or express the same topic. For such examples, we could make an assumption that they are in the non-contradiction class since there is no information in the sentence pairs which is related to each other. Therefore, we can filter out such examples before performing the contradiction detection task.

2. **More useful features should be further investigated.** In this experiment, the defined features are not be able to successfully capture most of the examples in the contradiction class. Additionally, some examples may be wrongly classified as non-contradiction. From the Example 2 above, the Negated-Term Parsing may recognize the term *don't believe* and *think* as contradiction. Unless filtering out of such example, other more useful features should be further investigated.

3. **Determine the actual referring arguments.** In our datasets, there are several examples similar to the Example 2. In such examples, it is unclear which argument or content that the H-T sentence pairs are referring to. For instance, as in the hypothesis, a person expresses his opinion that he does not believe in global warming. However, the content in the text may express that some people may believe in the existence of global warming (*it doesn't anymore* might indicate the global warming *doesn't exist*). Additionally, the content in the text may also express something else which is not related to the content in the hypothesis (in this case we do not know what *it doesn't anymore* is referring to). This happens when a compression rate is applied to a summarization task so that not all the informative sentences are extracted. Thus, all important pieces of information cannot be covered. For these reasons, in future work, we aim to investigate an argumentation mining approach which might help identify the correct arguments before performing debate summarization.

4. **World knowledge is required.** Moreover, one drawback of the system is that some contradicting examples are difficult to be detected as world knowledge might be necessary required for the task. For example, a sentence, *The earth is warming, so all of a sudden we're in a crisis and it's our fault?*, indicates that the opinion

> holder implies that he or she thinks the world is warming and believes in the existence of global warming. Another example, a user saying that *No, it is political hype*, wants to point out that the global warming does not exist and it is about the government making up this issue. For this reason, world knowledge is important for the contradiction detection task.

Furthermore, the complexity of the sentence structures is also a major problem in contradiction classification. To illustrate, the sentence pair below has a complex structure. In the hypothesis, the sentence contains the conjunctions, *and* and *but*, leading to the difficulty in determining which piece of information should be processed and matched to that in the text. For this reason, a more efficient parsing tool which can deeply analyze complex sentences should be adopted.

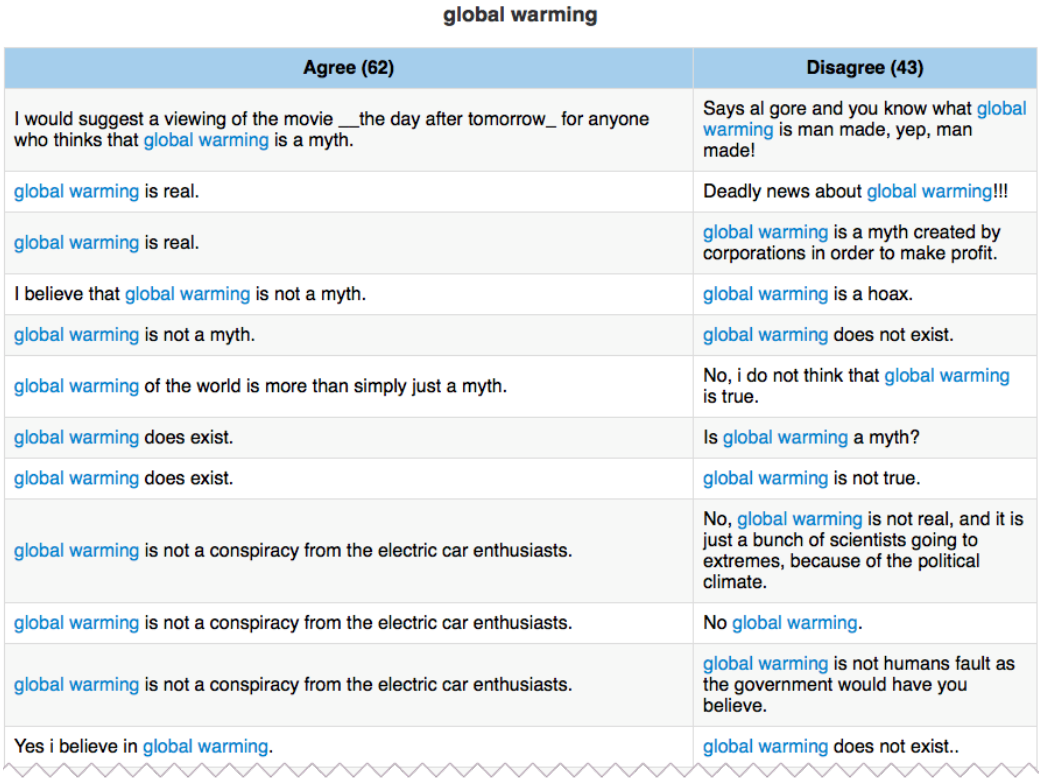
<**h**> I know there is still some debate as to how to stop it, **and** how much Man is contributing, **but** there is no question that the Earth is getting warmer. </**h**>

<**t**> It has been getting warmer for almost 100,000-years as part of a 100,000-year cycle when the earth's orbit becomes slightly more circular (as opposed to elliptical). </**t**>

6.2.4 Side-By-Side Summary Visualization

To visualize the Side-By-Side Summary, we processed the contradiction results derived from the contradiction detection task and exported them as HTML tables. Figure 6.5 to 6.6 illustrate the Side-By-Side Summaries generated on the DEDT and DEDX. The sentence pairs are separately illustrated in the two opposing sides which are derived from the stance of the original comments, from which the sentences were extracted.

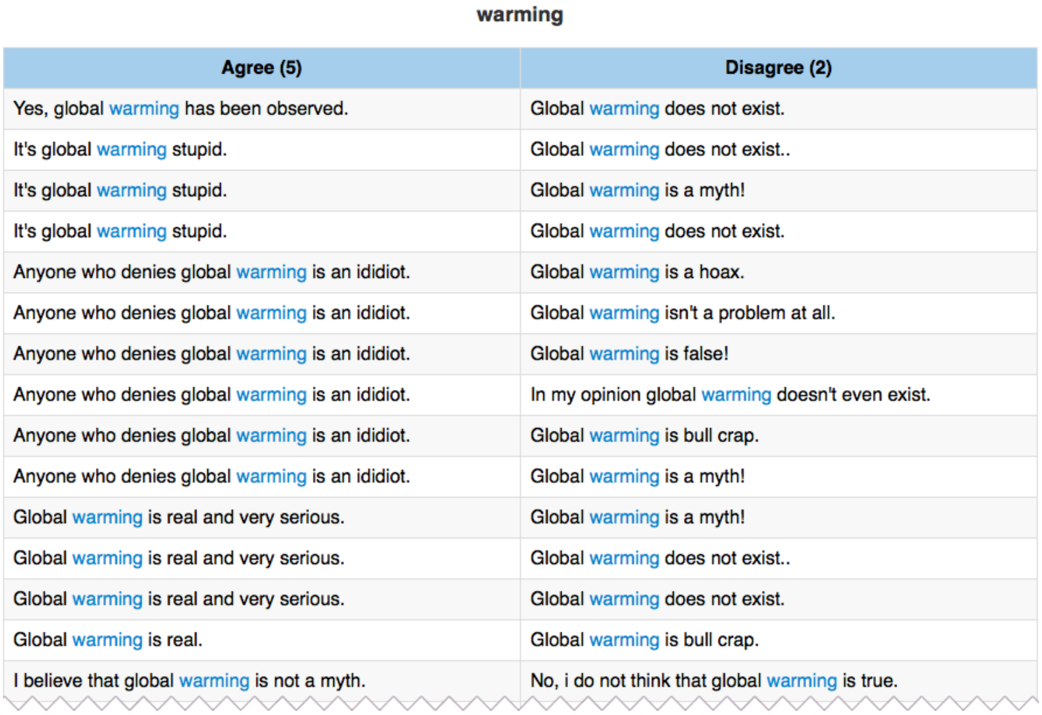
As shown in the figures, the summaries still contain repetitive sentences compared to the ones on the other side. In future work, we aim to investigate an argument mining technique which might help identify arguments in the text and formulate better Side-By-Side Summaries. Moreover, as aforementioned, a better parsing technique is still needed to improve the contradiction results.



global warming

Agree (62)	Disagree (43)
I would suggest a viewing of the movie __the day after tomorrow_ for anyone who thinks that global warming is a myth.	Says al gore and you know what global warming is man made, yep, man made!
global warming is real.	Deadly news about global warming!!!
global warming is real.	global warming is a myth created by corporations in order to make profit.
I believe that global warming is not a myth.	global warming is a hoax.
global warming is not a myth.	global warming does not exist.
global warming of the world is more than simply just a myth.	No, i do not think that global warming is true.
global warming does exist.	Is global warming a myth?
global warming does exist.	global warming is not true.
global warming is not a conspiracy from the electric car enthusiasts.	No, global warming is not real, and it is just a bunch of scientists going to extremes, because of the political climate.
global warming is not a conspiracy from the electric car enthusiasts.	No global warming.
global warming is not a conspiracy from the electric car enthusiasts.	global warming is not humans fault as the government would have you believe.
Yes i believe in global warming.	global warming does not exist..

Figure 6.5: Side-By-Side Summary for Global Warming topic from the DEDT



warming

Agree (5)	Disagree (2)
Yes, global warming has been observed.	Global warming does not exist.
It's global warming stupid.	Global warming does not exist..
It's global warming stupid.	Global warming is a myth!
It's global warming stupid.	Global warming does not exist.
Anyone who denies global warming is an ididiot.	Global warming is a hoax.
Anyone who denies global warming is an ididiot.	Global warming isn't a problem at all.
Anyone who denies global warming is an ididiot.	Global warming is false!
Anyone who denies global warming is an ididiot.	In my opinion global warming doesn't even exist.
Anyone who denies global warming is an ididiot.	Global warming is bull crap.
Anyone who denies global warming is an ididiot.	Global warming is a myth!
Global warming is real and very serious.	Global warming is a myth!
Global warming is real and very serious.	Global warming does not exist..
Global warming is real and very serious.	Global warming does not exist.
Global warming is real.	Global warming is bull crap.
I believe that global warming is not a myth.	No, i do not think that global warming is true.

Figure 6.6: Side-By-Side Summary for Global Warming topic from the DEDX

6.3 Summary

In this chapter, we presented the generation of Side-By-Side Summaries by further processing the results derived from the clustering process. We viewed that the generation of the summary can be achieved by a contradiction detection task. In the experiment, we created and explored a set of feature combinations to detect the contradiction relation. We discovered that it is not necessary to combine all the features that we defined. A combination of at least two features can achieve good results. In addition, we found that the key feature is *Negated-Term Parsing*. Other important features contributing to nearly the same scores are *Negation Overlap*, *Modal Verb Overlap*, and *Ontological Term Overlap*. The least important features are *Edit Distance* and *Structural Feature* features. Additionally, we also conducted an error analysis and reported the results as an aggregated confusion matrix. We discussed some alternative approaches that could help enhance the performance of the classification task. In future work, a more efficient parsing technique together with an argumentation mining approach should be investigated to improve the quality of the Side-By-Side Summary.

Chapter 7

CONCLUSION

While the automatic text summarization community has conducted research on summarizing product reviews, medical, political, legal, and meeting texts, we addressed a new domain, online debates, which recently became popular among Internet users. This thesis focuses on the summarization of online debates in forums. As the ultimate objective is to create a complete end-to-end summarization system, we do not only focus on the individual summarization tasks, but also consider how the summaries should be designed and presented to meet the user requirements. In this thesis, we have thus addressed both the summarization tasks which employ summarization techniques and the visualization tasks which visualize the summary in the preferred representation.

In this chapter, we summarize the research outcomes and define the key contributions made in this thesis. The conclusion also looks forward to future work, including alternative approaches for enhancing system performance and research directions in online debate summarization.

7.1 Summary of Work

In this thesis, contributions were made along with the development of the debate summarization system (as shown as dash lines in Figure 7.1). The following sections summarize the work and key contributions.

7.1.1 Preferred Summary Designs

In general, the outputs of a summarization system subjectively depend on the developers in which there could be generic text compressed by a defined compression rate, table, charts, etc. However, no empirical evidence establishing which summaries output is favored by users exists. This leaves the gap to explore appropriate information about user preferences and summary outputs for online debates.

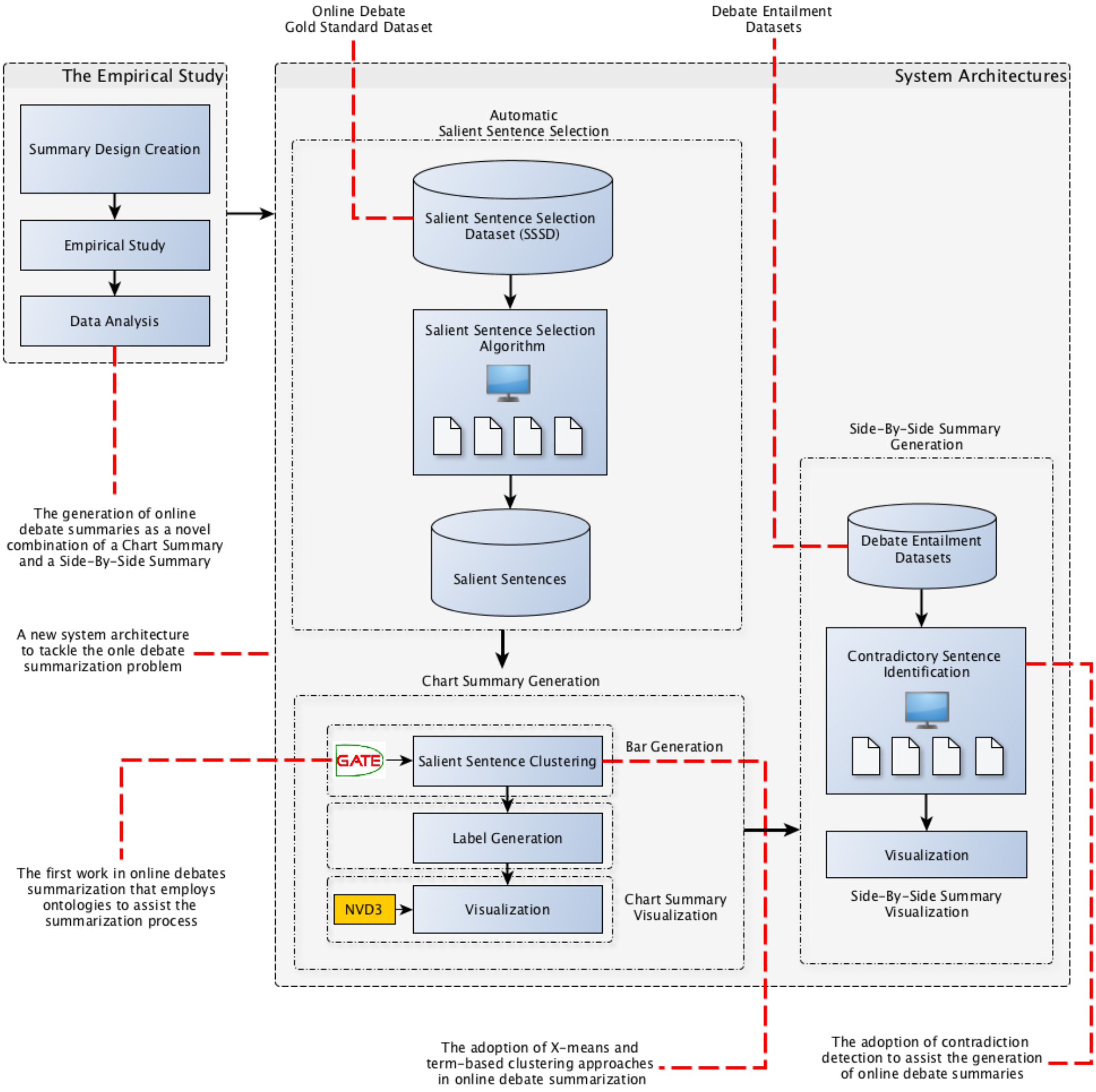


Figure 7.1: The summary of work and key contributions made along in this thesis

We experimented with seven summary representations, called summary designs, including a Chart Summary, a Table Summary, a Side-By-Side Summary, a Conceptual Map, the combination of Chart Summary and Table Summary, the combination of a Chart Summary and a Side-By-Side Summary, and the combination of a Chart Summary and a Conceptual Map. The aim of presenting the combination versions is for the first one to give the overview summaries and for the second one to present the more detailed summaries of the online debate.

We conducted an empirical study to establish which summary design is preferred by readers for summarizing online debates. Participants were recruited to rate the summary designs. The results indicated that the combination of a Chart Summary and a Side-By-Side Summary is the most preferred summary design. A hypothesis test indicated that there is a statistical difference in the user preferences between the summary designs. Therefore, we selected this combined chart and side-by-side summary design as the target output for our debate summarization system. Moreover, in this study, we proposed a novel summary representation that represents summary of debate contents in a Conceptual Map. Even though it is not the most preferred one, it has received some positive feedback by the participants.

7.1.2 ***System Architecture for Online Debate Summarization***

The system architecture of online debate summarization system is constructed based on the target combination of a Chart Summary and a Side-By-Side Summary. It consists of three primary stages.

In the first process, the system selects salient sentences from each debate comment based on nine features. The most important feature is Sentence Position which is significantly able to help identify positions of salient sentences in debate comments. The sentences selected by the system were evaluated against the SSSD and we reported the results with ROUGE-1, ROUGE-2, and ROUGE-SU4. We compared the system results against MEAD and found that they outperform the baseline.

In the second process, the salient sentences previously selected by the system were further clustered for the purpose of generating Chart Summary. We generated Chart Summary with two clustering approaches. This is the first work that applied ontologies to assist the summarization of online debates. In the first clustering approach, term-based clustering approach, we employed ontologies to capture climate change terms in the sentences. Sentences were placed in the same clusters if they share the same climate change terms. In addition, the shared terms were also picked as the cluster labels as they already show the core meaning of the clusters. In the cluster evaluation, we reported the clustering results with mean silhouette coefficient. The results indicated that the coordinates in the clusters are positioned near to the decision boundaries of the clusters and thus no clear-cut separation between clusters was found. For the evaluation of the label, the output was manually evaluated by a group of participants, by rating against a point 5-Likert scale. The results indicated that the labels can efficiently present the content in the clusters.

In the clustering approach, we applied X-means clustering approach to cluster the sentence. X-means is more superior that K-means as it is able to automatically detect the number of clusters. Using X-means, we were able to obtain good cohesion between the coordinates in the clusters, meaning that the salient sentences were well clustered. In the generation of labels, we adopt a Mutual Information approach. However, a better approach should be investigated to acquire better quality labels. By comparing the two clustering approaches, we concluded that the X-means clustering algorithm is the better alternative for clustering debate data. After the clustering is completed, we derived clusters which represent bars in the chart. We simply counted the number of salient sentences in the clusters to derive the counts in the chart. We finally combined the bars, labels, and the counts to visualize the Chart Summary using NVD3.

The main objective of the final process is the generation of the Side-By-Side Summary. We viewed this process a contradiction detection task. Research began with the creation of debate entailment datasets for training and evaluation. We generated RTE sentence pairs from the clustering results. For instance, for each cluster, a sentence from the Agree side was paired with other sentences on the Disagree side. A shorter sentence is considered as Hypothesis and the longer sentence is Text. Each pair was manually annotated with one of the two contradiction relations: *Contradiction* or *Non-contradiction*. As we applied two clustering approaches, we were able to create two debate entailment datasets as shown in Figure 7.1. Later on, we developed a classifier and investigated combinations of feature which maximize the F1 scores. Based on the proposed features, we discovered that the combinations of at least two features to the maximum of eight features yield good results for detecting contradiction relation. The generation of Side-By-Side Summary was completed by exporting the results which were detected as contradiction to be shown in HTML pages.

We finally linked Chart Summary to Side-By-Side Summary. When a user clicks on a bar in Chart Summary, the detailed summary of the clicked bar is visualized in a Side-By-Side Summary.

7.1.3 ***Salient Sentence Selection Dataset***

In order to evaluate online debate summarization, we created a gold standard dataset refereed to as the, *Salient Sentence Selection Dataset* (SSSD). We collected online debates related to the existence of global warming from a debate forum. For each comment, salient sentences were manually annotated by a group of participants. The number of sentences selected from a comment is controlled by a compression rate of 20%. For instance, two sentences were annotated in a comment containing 10 sentences. In total, the gold standard data contains 341 comments with 2595 annotated salient sentences. The dataset was used for the training, testing, and finally creating summaries for online debates.

7.1.4 ***Summary of Key Contributions***

This section summarizes the key contributions made in this thesis. They are presented as follows.

- The generation of online debate summaries as a novel combination of a Chart Summary and Side-By-Side Summary.
- A new system architecture to tackle the online debate summarization problem.
- The adoption of a contradiction detection task to assist the generation of online debate summaries
- Gold standard datasets:
  - Salient sentence selection dataset
  - Debate entailment dataset from the term-based clustering approach (DEDT)
  - Debate entailment dataset from the X-means clustering approach (DEDX)

7.2 Limitations

1. **Portability.** The online debate summarization system is a domain-dependent summarizer that focuses on summarizing debates related to global warming domain. For the system to be ported to a different domain, the following components are required:

    - **A list of key domain terms.** In this work, a list of key domain terms was collected from news media. The terms were used to create a feature in the automatic salient sentence selection task for determining the cosine similarity between the terms in the list and the sentences. To port the system to other domains, it is necessary to collect a list of key terms for those domains.
    - **Ontologies.** The online debate summarization system relies on the ontologies derived from the English ClimaPinion service[1]. They are used as the background knowledge to capture climate change terms in the term-based clustering approach and in the contradiction detection approach.

2. **New Summary Design.** In this work, we conducted a user study on 7 summary designs, including Chart Summary, Table Summary, Side-By-Side Summary, Conceptual Map, the Combination 1, the Combination 2, and the Combination 3. A new concern may arise when new summary designs come. For this reason, a further study may need to be re-conducted on the new summary designs together with the existing ones.

3. **Data Annotation.** In this work, the summarization system requires two kinds of annotated data. The first set of data is for the automatic salient sentence selection task. This set of data is needed to be annotated by selecting the sentences in debate comments indicating the summaries for the comments. The other set of data is for the Side-By-Side Summary generation task. The system generates the summaries by determining whether pairs of sentences are contradiction or non-contradiction. In the annotation, pairs of the sentences from each opposing sides are needed to be created and consequentially annotated with one of the two relations.

[1] http://services.gate.ac.uk/decarbonet/sentiment/

7.3 Future Work

The generation of such combined debate summaries still needs improvement, which we plan to address in future work.

*7.3.1 **Broad Coverage Summarization System***

In this thesis, on online debate summarization, we took a domain-dependent approach. It is therefore yet to be established how easy and generalized is the approach to new domains, especially since it requires domain ontologies. However, one could argue that the approach is likely to be portable, as long as a suitable domain ontology exists. A possible alternative could be to use a name entity recognition approach to help extract key concepts in the data. To evaluate this, online debate summarization in other domains should be investigated as well.

*7.3.2 **Improvement on the Contradiction Detection Task***

In the contradiction detection task, we only investigated one classification algorithm, logistic regression. In future work, we aim to explore other algorithms which might potentially improve the detection of the contradiction relation. Other algorithms which had been explored in prior work are *Random Forest*, *Support Vector Machine*, *Stochastic Gradient Descent*, and *K-Nearest Neighbors* (Marques, 2015).

*7.3.3 **Improvement in the Label Generation***

How to get the best labels for clusters? is still an open research question for the labeling extraction task. In this work, we applied Mutual Information to extract labels from the clusters and evaluated those using a manual approach. The evaluation results indicated that the quality of the labels is near to the TF-IDF baseline. In future work, we could apply an approach that generalizes label extraction. To illustrate, the system might firstly select some candidate labels such as *Carbon Dioxide*, *gas*, and *methane.* Then, generate a single generalized label from those candidates such as *Greenhouse Gases.* This is known as *generic title labeling* as presented by Tseng (2010).

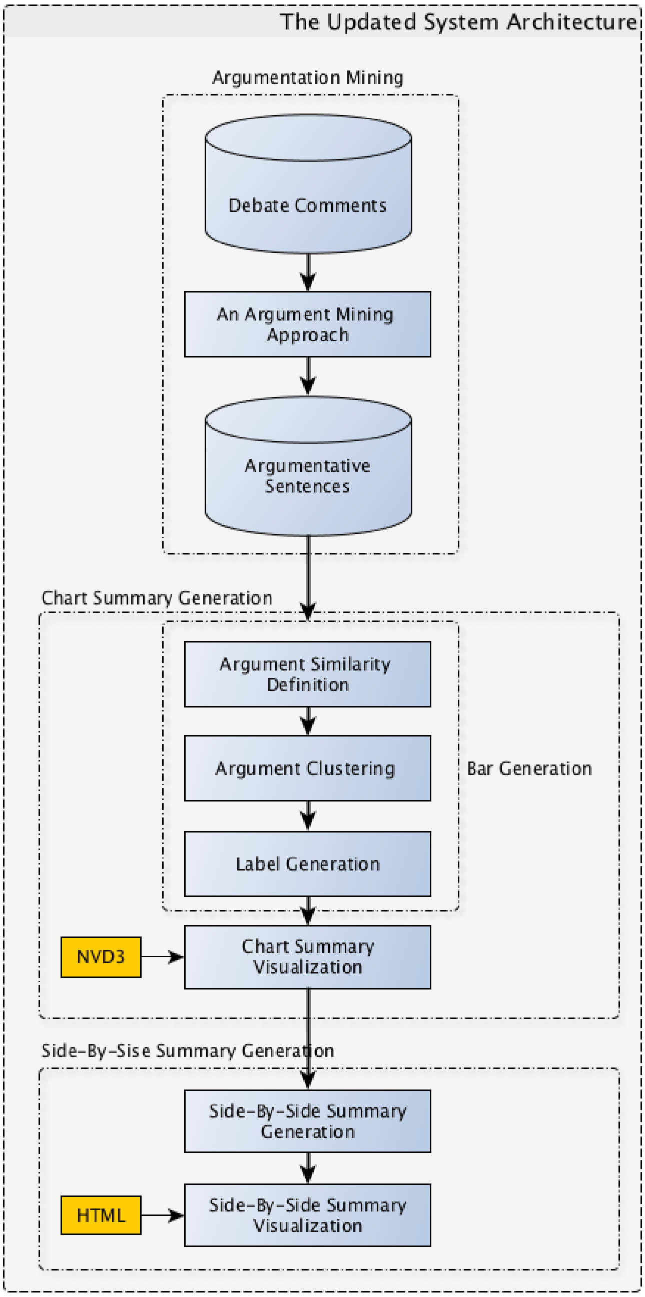

The Updated System Architecture
Argumentation Mining
Debate Comments
An Argument Mining Approach
Argumentative Sentences
Chart Summary Generation
Argument Similarity Definition
Argument Clustering
Bar Generation
Label Generation
NVD3
Chart Summary Visualization
Side-By-Sise Summary Generation
Side-By-Side Summary Generation
HTML
Side-By-Side Summary Visualization

Figure 7.2: An updated system architecture by the replacement of salient sentence selection with an argumentation mining approach

7.3.4 ***Argumentation Mining***

In our work, the system automatically selects important sentences from each debate comment in the first process. However, the selected sentences only convey extractive summaries of the comments. They may not show arguments or contradiction relations between sentences. In other words, arguments may be hidden as other sentences may not be selected by the system. The investigation of argument mining could potentially help detect more arguments and increase the quality of the system summaries. To determine whether this is the case, it will be one of our lines of future work.

Argumentation mining is a new challenge for the text summarization research community. As aforementioned in Chapter 3, researchers proposed different approaches for the detection of arguments in text. By applying argumentation mining, we could amend the system architecture for summarizing online debates as shown in Figure 7.2.

The adoption of an argumentation mining approach could potentially help save a substantial amount of time (i.e. data collection and data annotation in the salient sentence selection and the contradiction detection tasks) and scale down the work in the development of the online debate summarization system, especially in the generation the Combination 2 summaries. The approach could replace the contradiction detection task as the argumentation mining approach could help identify arguments across debates. In the first process, an ideal argumentation approach will automatically extract arguments from debate comments. The arguments derived from this process will be clustered[2] for the purpose of generating bars, and later the labels, and the figures. The bars could show arguments expressed across the two opposing sides. The labels will have the purpose of indicating descriptions for the bars and the figures will report the number of arguments in each cluster pair. After obtaining arguments on the two opposing sides, we can simply visualize the Chart Summary and the Side-By-Side Summary.

7.3.5 ***Evaluation without Reference Summaries***

In general, the annotation task requires a significant amount of time to be completed. It will additionally require extra time if the task contains a great amount of data to be

[2] Note that in future work we could either use the same clustering approaches or explore other clustering algorithms specifically for the argument clustering.

annotated. Moreover, the annotation will even take more time as more human entities are needed in order to ensure that the annotated data is quality and accurate.

In this thesis, a significant amount of time and effort was spent on data collection and annotation. This process is essential, especially for evaluation, as we need to measure the quality of the summaries generated by the system. Therefore, evaluation without reference summaries would be beneficial for the research community. Researchers including Louis and Nenkova (2013) and Saggion et al. (2010) proposed ways of assessing the summaries in such manner. The basic idea is to compare the content-based summaries to the original text (Louis and Nenkova, 2009). Future research on these novel approaches would potentially benefit the automatic summarization of online debates.

BIBLIOGRAPHY

Abbott, R., Walker, M., Anand, P., Fox Tree, J. E., Bowmani, R., and King, J. (2011). How can you say such things?!?: Recognizing disagreement in informal political argument. In *Proceedings of the Workshop on Languages in Social Media*, LSM '11, pages 2–11, Stroudsburg, PA, USA. Association for Computational Linguistics.

Aker, A., Paramita, M. L., Kurtic, E., Funk, A., Barker, E., Hepple, M., and Gaizauskas, R. J. (2016). Automatic label generation for news comment clusters. In *INLG 2016 - Proceedings of the Ninth International Natural Language Generation Conference, September 5-8, 2016, Edinburgh, UK*, pages 61–69.

Alfonesca, E. and Delort, J.-Y. (2012). A topic-model based approach for update summarization. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (ECAL) 2012*, Avignon, France.

Anand, P., Walker, M., Abbott, R., Tree, J. E. F., Bowmani, R., and Minor, M. (2011). Cats rule and dogs drool!: Classifying stance in online debate. In *Proceedings of the 2Nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, WASSA '11, pages 1–9, Stroudsburg, PA, USA. Association for Computational Linguistics.

Artstein, R. and Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Comput. Linguist.*, 34(4):555–596.

Azzam, S., Humphreys, K., and Gaizauskas, R. (1999). Using coreference chains for text summarization. In *Proceedings of the Workshop on Coreference and Its Applications*, CorefApp '99, pages 77–84, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bachman, L. F. (2004). *Statistical Analyses for Language Assessment.* Cambridge University Press. Cambridge Books Online.

Banerjee, S., Mitra, P., and Sugiyama, K. (2015). Abstractive meeting summarization using dependency graph fusion. In *Proceedings of the 24th International Conference on*

World Wide Web, WWW '15 Companion, pages 5–6, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Barzilay, R. and Elhadad, M. (1997). Using lexical chains for text summarization. In *In Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization*, pages 10–17.

Baxendale, P. B. (1958). Machine-made index for technical literature: An experiment. *IBM J. Res. Dev.*, 2(4):354–361.

Bayerl, P. S. and Paul, K. I. (2011a). What determines inter-coder agreement in manual annotations? a meta-analytic investigation. *Comput. Linguist.*, 37(4):699–725.

Bayerl, P. S. and Paul, K. I. (2011b). What determines inter-coder agreement in manual annotations? a meta-analytic investigation. *Comput. Linguist.*, 37(4):699–725.

Boltužić, F. and Šnajder, J. (2014). Back up your stance: Recognizing arguments in online discussions. In *Proceedings of the First Workshop on Argumentation Mining*, pages 49–58, Baltimore, Maryland. Association for Computational Linguistics.

Bontcheva, K., Cunningham, H., Roberts, I., Roberts, A., Tablan, V., Aswani, N., and Gorrell, G. (2013). Gate teamware: A web-based, collaborative text annotation framework. *Lang. Resour. Eval.*, 47(4):1007–1029.

Boykoff, M. T. and Boykoff, J. M. (2007). Climate change and journalistic norms: A case-study of us mass-media coverage. *Geoforum*, 38(6):1190 – 1204. Theme Issue: Geographies of Generosity.

Brunn, M., Chali, Y., and Pinchak, C. (2001). Text summarization using lexical chains. In *in Document Understanding Conference (DUC*, pages 135–140.

Cabrio, E. and Villata, S. (2012). Combining textual entailment and argumentation theory for supporting online debates interactions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2*, ACL '12, pages 208–212, Stroudsburg, PA, USA. Association for Computational Linguistics.

Campr, M. and Jezek, K. (2012). Comparative summarization via latent semantic analysis. In *Lastest Trends in Information Technology;Proceedings of the 1st WSEAS*

International Conference on Information Technology and Computer Networks (ITCN '12), Proceedings of the 1st WSEAS International Conference on Cloud Computing (CLC '12), Proceedings of the 1st WSEAS International Conference on Programming Languages and Compilers (PRLC '12), Recent Advances in Computer Engineering Series 7, pages 279–284, Stroudsburg, PA, USA. WSEAS Press.

Cunningham, H., Maynard, D., and Bontcheva, K. (2011). *Text Processing with GATE.* Gateway Press CA.

Dagan, I., Roth, D., Sammons, M., and Zanzotto, F. M. (2013). *Recognizing Textual Entailment: Models and Applications.* Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

de Marneffe, M., Rafferty, A. N., and Manning, C. D. (2008). Finding contradictions in text. In *ACL 2008, Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, June 15-20, 2008, Columbus, Ohio, USA*, pages 1039–1047.

Debate (2013). *Is global climate change man-made?* `http://www.debate.org/opinions/is-global-climate-change-man-made`.

Definition of annotate in English: Annotate (2014). Retrieved august 25, 2017, from `http://www.oxforddictionaries.com/definition/american_english/annotate`.

Definition of stance in English: Stance (2014). Retrieved april 23, 2014, from `http://oxforddictionaries.com/definition/stance`.

Di Fabbrizio, G., Aker, A., and Gaizauskas, R. (2013). Summarizing online reviews using aspect rating distributions and language modeling. *IEEE Intelligent Systems*, 28(3):28–37.

Dongen, S. (2000). *Graph Clustering by Flow Simulation.* PhD thesis, University of Utrecht.

Du, P., Guo, J., Zhang, J., and Cheng, X. (2010). Manifold ranking with sink points for update summarization. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pages 1757–1760, New York, NY, USA. ACM.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Comput. Linguist.*, 19(1):61–74.

Edmundson, H. P. (1969). New methods in automatic extracting. *J. ACM*, 16(2):264–285.

Elhassan AT., Aljourf M., A.-M. F. and M., S. (2016). Classification of imbalance data using tomek link(t-link) combined with random under-sampling (rus) as a data reduction method. *Journal of Informatics and Data Mining*, 1(2):1–11. ISSN: 2472-1956.

Erkan, G. and Radev, D. R. (2004). Lexrank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Int. Res.*, 22(1):457–479.

Fang, Y., Si, L., Somasundaram, N., and Yu, Z. (2012). Mining contrastive opinions on political texts using cross-perspective topic model. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM '12, pages 63–72, New York, NY, USA. ACM.

Furuse, O., Hiroshima, N., Yamada, S., and Kataoka, R. (2007). Opinion sentence search engine on open-domain blog. In Veloso, M. M., editor, *IJCAI*, pages 2760–2765.

Gale, W., Church, K. W., and Yarowsky, D. (1992). Estimating upper and lower bounds on the performance of word-sense disambiguation programs. In *Proceedings of the 30th Annual Meeting on Association for Computational Linguistics*, ACL '92, pages 249–256, Stroudsburg, PA, USA. Association for Computational Linguistics.

Galgani, F., Compton, P., and Hoffmann, A. (2012). Combining different summarization techniques for legal text. In *Proceedings of the Workshop on Innovative Hybrid Approaches to the Processing of Textual Data*, HYBRID '12, pages 115–123, Stroudsburg, PA, USA. Association for Computational Linguistics.

Gambhir, M. and Gupta, V. (2017). Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review*, 47(1):1–66.

Ganesan, K., Zhai, C., and Han, J. (2010). Opinosis: a graph-based approach to abstractive summarization of highly redundant opinions. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 340–348. Association for Computational Linguistics.

Ganesan, K., Zhai, C., and Viegas, E. (2012). Micropinion generation: An unsupervised approach to generating ultra-concise summaries of opinions. In *Proceedings of the 21st International Conference on World Wide Web 2012 (WWW '12).*

Ghasemi, A. and Zahediasl, S. (2012). Normality tests for statistical analysis: A guide for non-statisticians. *Int J Endocrinol Metab*, 10(2):486–489.

Goldstein, J., Kantrowitz, M., Mittal, V., and Carbonell, J. (1999). Summarizing text documents: Sentence selection and evaluation metrics. In *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, pages 121–128, New York, NY, USA. ACM.

Gong, Y. and Liu, X. (2001). Generic text summarization using relevance measure and latent semantic analysis. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, pages 19–25, New York, NY, USA. ACM.

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology*, 8(1):23–34.

Halliday, M. A. and Hasan, R. (1976). *Cohesion in English.* Longman, London.

Harabagiu, S., Hickl, A., and Lacatusu, F. (2006). Negation, contrast and contradiction in text processing. In *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1*, AAAI'06, pages 755–762. AAAI Press.

Hirao, T., Isozaki, H., and Maeda, E. (2002). Extracting important sentences with support vector machines. In *In Proc. 19th COLING*, pages 342–348.

Hobson, S. F. (2007). *Text Summarization Evaluation: Correlating Human Performance on an Extrinsic Task with Automatic Intrinsic Metrics.* PhD thesis, Institute for Advanced Computer Studies, University of Maryland.

Hong, K., Marcus, M., and Nenkova, A. (2015). System combination for multi-document summarization. In Màrquez, L., Callison-Burch, C., Su, J., Pighin, D., and Marton, Y., editors, *EMNLP*, pages 107–117. The Association for Computational Linguistics.

Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '04, pages 168–177, New York, NY, USA. ACM.

Huang, X., Wan, X., and Xiao, J. (2011). Comparative news summarization using linear programming. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, pages 648–653, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ikeda, D., Takamura, H., Ratinov, L.-A., and Okumura, M. (2008). Learning to shift the polarity of words for sentiment classification. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I*, pages 296–303.

Januliené, A. and Dziedravičius, J. (2015). On the use of conjunctive adverbs in learners'academic essays. *Verbum*, 6:69–83.

Jing, H., Barzilay, R., McKeown, K., and Elhadad, M. (1998). Summarization evaluation methods: Experiments and analysis. In *IN AAAI SYMPOSIUM ON INTELLIGENT SUMMARIZATION*, pages 60–68.

Juan-Manuel, T.-M. (2014). *Automatic text summarization / Juan-Manuel Torres-Moreno.* ISTE Ltd ; John Wiley & Sons, Inc London : Hoboken, NJ.

Jurafsky, D. and Martin, J. H. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition (Prentice Hall Series in Artificial Intelligence).* Prentice Hall, 1 edition.

Kass, R. E. and Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, 91(435):1343–1370.

Kaufman, L. and Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis.* John Wiley.

Kelleher, J. D., Mac Namee, B., and D'Arcy, A. (2015). *Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies.* MIT Press.

Kim, H. D. and Zhai, C. (2009). Generating comparative summaries of contradictory opinions in text. In Cheung, D. W.-L., Song, I.-Y., Chu, W. W., Hu, X., and Lin, J. J., editors, *CIKM*, pages 385–394. ACM.

Kouylekov, M. and Magnini, B. (2005). Recognizing textual entailment with tree edit distance algorithms. In *PASCAL Challenges on RTE*, pages 17–20.

Krippendorff, K. (2004). *Content Analysis: An introduction to its methodology.* Sage Publications, Inc, Thousand Oaks, CA, 2nd edition.

Kübler, S., McDonald, R. T., and Nivre, J. (2009). *Dependency Parsing.* Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1).

Lay, D. (2006). *Linear Algebra and its Applications*, volume 2: MyMathLab. Pearson, Addison Wesley, 3 edition.

Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. In Jebara, T. and Xing, E. P., editors, *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1188–1196. JMLR Workshop and Conference Proceedings.

Lee, C.-S., Chen, Y.-J., and Jian, Z.-W. (2003). Ontology-based fuzzy event extraction agent for chinese e-news summarization. *Expert Systems with Applications*, 25(3):431 – 447.

Lendvai, P., Augenstein, I., Rout, D., Bontcheva, K., and Declerck, T. (2016). Algorithms for detecting disputed information: Final version.

Lendvai, P. and Reichel, U. D. (2016). Contradiction detection for rumorous claims. *arXiv preprint arXiv:1611.02588.*

Lerman, K. and McDonald, R. (2009). Contrastive summarization: An experiment with consumer reviews. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, NAACL-Short '09, pages 113–116, Stroudsburg, PA, USA. Association for Computational Linguistics.

Li, S., Ouyang, Y., Wang, W., and Sun, B. (2007). Multi-document summarization using support vector regression.

Light, R. J. (1971). Measures of response agreement for qualitative data: Some generalizations and alternatives. *Psychological Bulletin*, 76:365–377.

Lin, C.-Y. (1999). Training a selection function for extraction. In *Proceedings of the Eighth International Conference on Information and Knowledge Management*, CIKM '99, pages 55–62, New York, NY, USA. ACM.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In Marie-Francine Moens, S. S., editor, *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Lin, C.-Y. and Hovy, E. (1997). Identifying topics by position. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, ANLC '97, pages 283–290, Stroudsburg, PA, USA. Association for Computational Linguistics.

Lin, C.-Y. and Hovy, E. (2000). The automated acquisition of topic signatures for text summarization. In *Proceedings of the 18th Conference on Computational Linguistics - Volume 1*, COLING '00, pages 495–501, Stroudsburg, PA, USA. Association for Computational Linguistics.

Liu, B., Hu, M., and Cheng, J. (2005). Opinion observer: analyzing and comparing opinions on the web. In *Proceedings of the 14th international conference on World Wide Web*, pages 342–351. ACM.

Liu, C., Chen, M., and Tseng, C. (2015). Incrests: Towards real-time incremental short text summarization on comment streams from social network services. *Knowledge and Data Engineering, IEEE Transactions on*, 27(11):2986–3000.

Liu, F. and Liu, Y. (2008). Correlation between rouge and human evaluation of extractive meeting summaries. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, HLT-Short '08, pages 201–204, Stroudsburg, PA, USA. Association for Computational Linguistics.

Louis, A. and Nenkova, A. (2009). Automatically evaluating content selection in summarization without human models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 306–314, Stroudsburg, PA, USA. Association for Computational Linguistics.

Louis, A. and Nenkova, A. (2013). Automatically assessing machine summary content without a gold standard. *Comput. Linguist.*, 39(2):267–300.

Lu, Y., Zhai, C., and Sundaresan, N. (2009). Rated aspect summarization of short comments. In *Proceedings of the 18th international conference on World wide web*, pages 131–140. ACM.

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM J. Res. Dev.*, 2(2):159–165.

López, V., Fernández, A., García, S., Palade, V., and Herrera, F. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences*, 250:113 – 141.

Magnini, B. and Cabrio, E. (2010). Contradiction-focused qualitative evaluation of textual entailment. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, NeSp-NLP '10, pages 86–94, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mani, I. (2001a). *Automatic Summarization.* John Benjamins Publishing Company.

Mani, I. (2001b). Summarization evaluation: An overview.

Mani, I., House, D., Klein, G., Hirschman, L., Firmin, T., and Sundheim, B. (1999). The tipster summac text summarization evaluation.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval.* Cambridge University Press.

Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing.* MIT Press, Cambridge, MA, USA.

Marcu, D. (1997). The rhetorical parsing of natural language texts. In *Proceedings of the Eighth Conference on European Chapter of the Association for Computational Linguistics*, EACL '97, pages 96–103, Stroudsburg, PA, USA. Association for Computational Linguistics.

Markner-Jäger, B. (2008). *Technical English for Geosciences: A Text/Work Book.* Springer.

Marneffe, D., Maccartney, B., Rafferty, A. N., Yeh, E., and Manning, C. D. (2009). Deciding entailment and contradiction with stochastic and edit distance-based alignment. In *In Proceeding of TAC, 2008 Workshop.*

Marques, R. (2015). Detecting contradictions in news quotations. Master's thesis, Information Systems and Computer Engineering, Instituto Superior Técnico, https://fenix.tecnico.ulisboa.pt/downloadFile/1126295043834330/dissertacao.pdf.

Marsi, E., Krahmer, E., and Bosma, W. (2007). Dependency-based paraphrasing for recognizing textual entailment. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, RTE '07, pages 83–88, Stroudsburg, PA, USA. Association for Computational Linguistics.

Maynard, D. and Bontcheva, K. (2015). Understanding climate change tweets: an open source toolkit for social media analysis. In Vivian Kvist Johannsen, Stefan Jensen, V. W. C. P. E. E., editor, *Atlantis Press*, pages 242–250, Atlantis Press. Morgan Kaufmann Publishers Inc.

McKeown, K., Passonneau, R. J., Elson, D. K., Nenkova, A., and Hirschberg, J. (2005). Do summaries help? In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, pages 210–217, New York, NY, USA. ACM.

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Minel, J.-L., Nugier, S., and Piat, G. (1997). How to appreciate the quality of automatic text summarization? examples of fan and mluce protocols and their results on seraphin. In *Intelligent Scalable Text Summarization*, pages 25–30.

Mitra, M., Singhal, A., and Buckley, C. (1997). Automatic text summarization by paragraph extraction.

Morales, L. P., Esteban, A. D., and Gervás, P. (2008). Concept-graph based biomedical automatic summarization using ontologies. In *Proceedings of the 3rd Textgraphs Workshop on Graph-Based Algorithms for Natural Language Processing*, TextGraphs-3, pages 53–56, Stroudsburg, PA, USA. Association for Computational Linguistics.

Morris, A., Kasper, G., and Adams, D. (1992). The effects and limitations of automated text condensing on reading comprehension performance. *Information Systems Research*, 3(1):17–35. cited By 51.

Morris, J. and Hirst, G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Comput. Linguist.*, 17(1):21–48.

Nenkova, A. and McKeown, K. (2011). Automatic Summarization. *Foundations and Trends in Information Retrieval*, 5(2):103–233.

Neto, J. L., Freitas, A. A., and Kaestner, C. A. A. (2002). Automatic text summarization using a machine learning approach. In *Proceedings of the 16th Brazilian Symposium on Artificial Intelligence: Advances in Artificial Intelligence*, SBIA '02, pages 205–215, London, UK, UK. Springer-Verlag.

Nguyen, H.-M. and Shirai, K. (2013). Recognition of agreement and contradiction between sentences in support-sentence retrieval.

NLTK Metrics (2015). Nltk.metric package.

Novak, J. D. and Cañas, A. J. (2006). The theory underlying concept maps and how to construct them. Technical report, Technical Report IHMC CmapTools 2006-01.

Ogren, P. V. (2006). Knowtator: A protÉgÉ plug-in for annotated corpus construction. In *Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Companion Volume: Demonstrations*, NAACL-Demonstrations '06, pages 273–275, Stroudsburg, PA, USA. Association for Computational Linguistics.

Orăsan, C. (2002). Building annotated resources for automatic text summarisation. In *In Proceedings of the Third International Conference on Language Resources and Evaluation (LREC-2002), Las Palmas de Gran Canaria*, pages 1780–1786.

Oxford, O. D. (2014a). Retrieved november 9, 2014, from `http://www.oxforddictionaries.com/definition/english/debate`.

Oxford, O. D. (2014b). Retrieved november 9, 2014, from `http://www.oxforddictionaries.com/definition/english/opposing`.

Palau, R. M. and Moens, M.-F. (2009). Argumentation mining: The detection, classification and structure of arguments in text. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, ICAIL '09, pages 98–107, New York, NY, USA. ACM.

Pallant, J. (2013). *SPSS Survival Manual: A Step by Step Guide to Data Analysis Using SPSS*. Open University Press, Berkshire, UK, SL6 2QL, 5th edition.

Park, S., Lee, K., and Song, J. (2011). Contrasting opposing views of news articles on contentious issues. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 340–349, Stroudsburg, PA, USA. Association for Computational Linguistics.

Partha Pakray, A. G. and Bandyopadhyay, S. (2011). Textual entailment using lexical and syntactic similarity. *In International Journal of Artificial Intelligence & Applications (IJAIA)*, Vol.2, No.1:43–58. DOI : 10.5121/ijaia.2011.2104.

Passonneau, R. J. (2004). Computing reliability for coreference annotation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC-2004)*, Lisbon, Portugal. European Language Resources Association (ELRA). ACL Anthology Identifier: L04-1486.

Passonneau, R. J. and Litman, D. J. (1993). Intention-based segmentation: Human reliability and correlation with linguistic cues. In *Proceedings of the 31st Annual Meeting on Association for Computational Linguistics*, ACL '93, pages 148–155, Stroudsburg, PA, USA. Association for Computational Linguistics.

Paul, M. J., Zhai, C., and Girju, R. (2010). Summarizing contrastive viewpoints in opinionated text. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 66–76. Association for Computational Linguistics.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.*, 12:2825–2830.

Pelleg, D. and Moore, A. W. (2000). X-means: Extending k-means with efficient estimation of the number of clusters. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, pages 727–734, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Plaza, L., Stevenson, M., and Díaz, A. (2010). Improving summarization of biomedical documents using word sense disambiguation. In *Proceedings of the 2010 Workshop on*

Biomedical Natural Language Processing, pages 55–63. Association for Computational Linguistics.

Ponte, J. M. and Croft, W. B. (1998). A language modeling approach to information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pages 275–281, New York, NY, USA. ACM.

Radev, D. R., Jing, H., and Budzikowska, M. (2000). Centroid-based summarization of multiple documents: Sentence extraction, utility-based evaluation, and user studies. In *Proceedings of the 2000 NAACL-ANLPWorkshop on Automatic Summarization - Volume 4*, NAACL-ANLP-AutoSum '00, pages 21–30, Stroudsburg, PA, USA. Association for Computational Linguistics.

Radev, D. R., Jing, H., Styś, M., and Tam, D. (2004). Centroid-based summarization of multiple documents. *Inf. Process. Manage.*, 40(6):919–938.

Ranade, S., Gupta, J., Varma, V., and Mamidi, R. (2013). Online debate summarization using topic directed sentiment analysis. In *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining*, WISDOM '13, pages 7:1–7:6, New York, NY, USA. ACM.

Rath, G. J., Resnick, A., and Savage, T. R. (1961). The formation of abstracts by the selection of sentences. part i. sentence selection by men and machines. *American Documentation*, 12(2):139–141.

Řehůřek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. http://is.muni.cz/publication/884893/en.

Renals, S., Bourlard, H., Carletta, J., and Popescu-Belis, A., editors (2012). *Multimodal Signal Processing: Human Interactions in Meetings.* Cambridge University Press.

Rousseeuw, P. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, 20(1):53–65.

Saggion, H. (2008). A robust and adaptable summarization tool. *Traitement Automatique des Langues*, 49(2).

Saggion, H. and Gaizauskas, R. (2004). Multi-document summarization by cluster/profile relevance and redundancy removal. In *Proceedings of the Document Understanding Conference*, pages 6–7.

Saggion, H. and Gaizauskas, R. (2005). Experiments on statistical and pattern-based biographical summarization. *Lecture notes in computer science*, 3808:611.

Saggion, H. and Lapalme, G. (2000). Concept identification and presentation in the context of technical text summarization. In *Proceedings of the 2000 NAACL-ANLPWorkshop on Automatic Summarization - Volume 4*, NAACL-ANLP-AutoSum '00, pages 1–10, Stroudsburg, PA, USA. Association for Computational Linguistics.

Saggion, H., Torres-Moreno, J.-M., Cunha, I. d., and SanJuan, E. (2010). Multilingual summarization evaluation without human models. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 1059–1067, Stroudsburg, PA, USA. Association for Computational Linguistics.

Salton, G., Wong, A., and Yang, C.-S. (1975). A vector space model for automatic indexing. *Commun. ACM*, 18:613–620.

Sammons, M., Vydiswaran, V. G. V., Vieira, T., Johri, N., Chang, M.-W., Goldwasser, D., Srikumar, V., Kundu, G., Tu, Y., Small, K., Rule, J., Do, Q., and Roth, D. (2009). Relation alignment for textual entailment recognition. In *TAC*. NIST.

Sanchan, N., Aker, A., and Bontcheva, K. (2017). Gold standard online debates summaries and first experiments towards automatic summarization of online debate data. TBA:TBA.

Sanchan, N., Bontcheva, K., and Aker, A. (2016). Understanding human preferences for summary designs in online debates domain. *Polibits*, 54:79–85.

Schrading, N. (2016). Introduction to spacy for nlp and machine learning. `https://github.com/NSchrading/intro-spacy-nlp`.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

Scott, W. A. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, 19(3):321–325.

Somasundaran, S. and Wiebe, J. (2009). Recognizing stances in one debates. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1*, ACL '09, pages 226–234, Stroudsburg, PA, USA. Association for Computational Linguistics.

Somasundaran, S. and Wiebe, J. (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, pages 116–124, Stroudsburg, PA, USA. Association for Computational Linguistics.

Steinberger, J. and Jezek, K. (2009). Evaluation measures for text summarization. 28:251–275.

Steinberger, J., Poesio, M., Kabadjov, M. A., and Jeek, K. (2007). Two uses of anaphora resolution in summarization. *Inf. Process. Manage.*, 43(6):1663–1680.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., and Tsujii, J. (2012). Brat: A web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 102–107, Stroudsburg, PA, USA. Association for Computational Linguistics.

Strang, G. (2006). *Linear algebra and its applications.* Thomson, Brooks/Cole, Belmont, CA.

StratifiedKFold (2018). Retrieved march 1, 2018, from `http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html`.

Suanmali, L., Salim, N., and Binwahlan, M. S. (2009). Fuzzy Logic Based Method for Improving Text Summarization. *Journal of Computer Science*, 2:6.

Trabelsi, A. and Zaiane, O. R. (2014). Finding arguing expressions of divergent viewpoints in online debates. In *Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM)*, pages 35–43, Gothenburg, Sweden. Association for Computational Linguistics.

Tseng, Y.-H. (2010). Generic title labeling for clustered documents. *Expert Syst. Appl.*, 37(3):2247–2254.

Tunggal, D. (2012). Automatic Multi Document Summarization Approaches Yogan Jaya Kumar and 2 Naomie Salim Faculty of Information and Communication Technology , Faculty of Computer Science and Information Systems ,. 8(1):133–140.

Šnajder, J. (2017). Social media argumentation mining: The quest for deliberateness in raucousness. *CoRR*, abs/1701.00168.

Wan, X., Li, H., and Xiao, J. (2010). Cross-language document summarization based on machine translation quality prediction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 917–926, Stroudsburg, PA, USA. Association for Computational Linguistics.

Wang, D. and Li, T. (2010). Document update summarization using incremental hierarchical clustering. In Huang, J., Koudas, N., Jones, G. J. F., Wu, X., Collins-Thompson, K., and An, A., editors, *CIKM*, pages 279–288. ACM.

Wang, D., Zhu, S., Li, T., and Gong, Y. (2012). Comparative document summarization via discriminative sentence selection. *ACM Trans. Knowl. Discov. Data*, 6(3):12:1–12:18.

Wang, L. and Cardie, C. (2012). Focused meeting summarization via unsupervised relation extraction. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, SIGDIAL '12.

Wang, S. and Koopman, R. (2017). Clustering articles based on semantic similarity. *Scientometrics*, 111(2):1017–1031.

Wilson, T., Wiebe, J., and Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language Technologies Conference/Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*, Vancouver, CA.

Witte, R. and Bergler, S. (2007). Next-Generation Summarization: Contrastive, Focused, and Update Summaries. In *International Conference on Recent Advances in Natural Language Processing (RANLP 2007)*, Borovets, Bulgaria.

Yeh, J.-Y., Ke, H.-R., Yang, W.-P., and Meng, I.-H. (2005). Text summarization using a trainable summarizer and latent semantic analysis. *Inf. Process. Manage.*, 41(1):75–95.

Yen, S.-J., Lee, Y.-S., Wu, Y.-C., Ying, J.-C., and Tseng, V. S. (2010). Automatic chinese text classification using n-gram model. In *Proceedings of the 2010 International Conference on Computational Science and Its Applications - Volume Part III*, ICCSA'10, pages 458–471, Berlin, Heidelberg. Springer-Verlag.

Zhai, C., Velivelli, A., and Yu, B. (2004). A cross-collection mixture model for comparative text mining. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 743–748, New York, NY, USA. ACM.

Zhuang, L., Jing, F., Zhu, X., and Zhang, L. (2006). Movie review mining and summarization. In *Proceedings of the ACM SIGIR Conference on Information and Knowledge Management (CIKM)*.