May 2018

Citation Function and Polarity Classification in Biomedical Papers

Meng Jia
The University of Western Ontario

Supervisor
Mercer, Robert E.
The University of Western Ontario

Graduate Program in Computer Science

A thesis submitted in partial fulfillment of the requirements for the degree in Master of Science

Abstract

The traditional reference evaluation method treats all citations equally. However, a citation can serve various functions. It may reflect the citing paper author's motivation as well as his/her true attitude towards the cited paper. Investigating such information can be achieved through citation content analysis.

This thesis develops an 8-category classification scheme on citation function and polarity to help understand what role a citation played in scientific papers. A biomedical citation corpus is annotated with this scheme and experimented with supervised machine learning methods. Several types of features that capture the characteristics of citation sentences are extracted by natural language processing techniques to serve as the inputs of automatic classifiers. The importance of cue phrases in citation classification is also addressed and discussed.

Keywords: citation classification, citation function categorization, sentiment analysis

Acknowledgements

I would like to express my sincere gratitude to my supervisor Dr. Robert E. Mercer for his continuous support of my Master study during the last two years. Without his insightful guidance and great patience, I could not finish this thesis. I would also like to thank the Department of Computer Science for the financial support on this research opportunity. Finally, I would like to thank my parents and friends for their endless care and encouragement of all the time.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

A citation is a reference or link between the current research work and previous studies. In general, it serves several important purposes: first, it can check an author's academic honesty, that is, to avoid plagiarism; second, authors use citations to show research background and the preparation they made for the current work, or to validate or dispute the previous research; third, the later researchers or readers can learn about the evolution of a specific research field, or get inspired for their own works.

The importance of an article is usually evaluated by the number of times it is cited, thus the early studies of citation analysis focused more on citing frequency and other citing statistics, and treated all citations equally [42]. However, with the fast development of Internet and digital technology, more and more academic documents become available in electronic format and easy for people to access. Retrieving an expected document based on citing frequency and standard citation indexes seems not as efficient and accurate as before. Moreover, having lots of retrieved documents will cost readers much time to find the expected information. For these reasons, it is necessary to filter the documents according to the functions of citations, and link the related articles that meet a reader's specific needs by building a citation-based network.

A series of citation classification schemes have been proposed and developed since the last century. Garfield [19] stated 15 reasons for researchers citing other people's work. His work laid the foundations for citation motivation and function studies. Later, Lipetz [27] developed a scheme of 29 categories to describe the relations between citing and cited documents. To improve the efficiency of classifying massive data with computer technology, more and more researchers focus on developing a scheme that can be easily adapted by automatic classifiers.

In recent years, citation polarity classification (or citation sentiment analysis) gradually became a popular topic in citation classification. Analyzing the sentiment of a citation may reveal an author's true attitude towards a cited paper, giving readers a more intuitive judgment and helping them to filter out the articles they don't need. In the biomedical domain, citation

sentiment analysis has much potential "to detect non-reproducible studies" [44]. For example, it can help prevent wasting resources on expensive but unnecessary medical tests.

However, detecting the sentiment of a citation is relatively difficult, compared with sentiment analysis on movie or product reviews. The main reason is that citation sentiment is often hidden due to academic writing style which restricts authors to express obvious personal opinions in order to prevent bias and stay objective. A comparison of a movie review and a scientific citation is given below:

> Movie review:
> The film contains no good jokes, no good scenes, barely a moment when Carvey's Saturday Night Live-honed mimicry rises above the level of embarrassment.
> Citation:
> Lyapina et al. recently demonstrated that partially purified CSN promotes the cleavage of Pcu1p-Ned8p conjugates in vitro [10], but it **remained unclear** whether the enzymatic activity is contained within one of the CSN subunits or a tightly associated peptidase.

Both the movie review and the citation demonstrate negative polarity, but the sentiment in the movie review is more straightforward and stronger, while the citation shows this negativity in a subtle way. This is because in scientific writing, authors tend to use rhetorical techniques to cover such negative opinions. The highlighted phrase "remain unclear" in the citation sentence gives readers a hint about the author's true attitude towards the cited paper, thus we can treat such a phrase as a "cue phrase".

In this thesis, I define the words or phrases that express the polarity of a citation or imply the relationship between citing and cited works as cues. The term "cue phrase" represents both a single word and a word phrase in most chapters. The main tasks of this thesis are to propose and test an automated classification scheme of citation polarity and function, as well as to investigate the importance of fine-grained cue phrases.

The terms "polarity" and "sentiment" have almost the same meaning, except that "sentiment" implies the object of discussion is the author's opinion and contains more emotional reactions, while the object of "polarity" is usually the texts. Therefore, I will use "polarity" when describing the citation content and "sentiment" for mentioning the authors' opinions.

This thesis is structured as follows: Chapter 2 reviews the background literature in citation polarity and function classification. Chapter 3 compares several previous classification schemes and proposes a new categorization scheme. Chapter 4 describes the biomedical corpus used in this research as well as the automatic extraction method for cue phrases. In Chapter

5, I will extract features and apply machine learning methods on the annotated corpus. The experimental results will be presented showing the importance of cue phrases in automatic citation classification. The final chapter summarizes the research work that has been done for this thesis as well as providing some conclusions. It also discusses some potential future work.

Chapter 2

Literature Review

In traditional citation analysis studies, researchers tend to focus only on citation statistics and assign the same weights to all citations. However, with the greatly increased number of academic publications over the past few years, this method has becs ome quite limited. Investigating what role a citation plays in research output evaluation provides a more comprehensive scope, and gradually has become main stream in citation analysis.

To begin with, Section 2.1 reviews the foundations of syntactic and semantic citation analysis in early studies, and discusses several general citation classification schemes. Then, Section 2.2 explores different schemes of citation function/motivation classification developed using computational linguistics approaches, and compares them with the manual approach to identify the advantages from both. Section 2.3 describes some of the semi-automatic citation polarity classifiers built with machine learning algorithms and the features extracted from citation content. Section 2.4 discusses citation polarity classification in the biomedical context together with other linguistic-related citation analysis.

2.1 Foundations and General Classifications

Through the manual examination of a small set of papers, Voos and Dagaev (1976) [42] found there were more highly cited papers in the introduction section than other parts of the citing work. They came to the conclusion that it is possible to evaluate the importance of a citation to the citing work by using both citing frequency and the location information. This is the first finding that addressed the problem of treating all citations equally.

Later syntactic research works, such as Maričić et al. (1998) [30], used a larger corpus to conduct location-based frequency calculation, and found that the introduction section contains less meaningful citations than the methods, results and discussion/conclusion sections. Similarly, Suppe (1998) [38] claimed that the sections about methods and data are more im-

portant because the explanations inside these sections demonstrate whether a new finding can be integrated into the knowledge base and promote the evolution of a specific field.

The 1960's saw the beginning of semantic analysis which gradually became the dominant methodology in citation content studies. Garfield (1964) [19] was the first to suggest further investigation in citation motivation. He listed 15 reasons why authors cite previous works, which can be seen as a generalized classification scheme. These reasons were frequently used by later researchers to identify semantic citation characteristics, along with Lipetz (1965)'s [27] 29 categories on relationships between citing and cited articles.

Moravcsik and Murugesan (1975) [33] explores quantitative measurements of the values of cited works. They proposed a citation motivation scheme that divides citations along four dimensions: conceptual or operational, organic or perfunctory, evolutionary or juxtapositional, confirmative or negative. Moreover, during their exploration, Moravcsik and Murugesan found 40% of citations are perfunctory, "which casts further doubt on the citation-counting approach" [39].

A more recent work, Jochim and Schütze (2012) [24], adopted Moravcsik and Murugesan's scheme and developed a four-faceted set composed of CONC_OP, ORG_PERF, EVOL_JUX and CONF_NEG. It is noted that this set has no undefined facet, in other words, it avoids a neutral class in classification and makes full use of neutral citations. Combined with fine-grained features and machine learning algorithms, they conducted an automatic classification on a large citation corpus and obtained state-of-the-art performance.

Another noticeable classification was from Spiegel-Rösing (1977) [37], which addresses citation evaluative use and citation content. They used a 13-class scheme to manually classify 2309 citations extracted from science articles. The results indicate 80% of citations substantiate a statement or point to further information, and only 0.8% contains criticism.

In summary, the early studies of syntactic and semantic citation content analysis mainly relied on manual examination of a small corpus or conducting surveys, and the classification schemes proposed during this time are mostly generalized. Although the findings were limited and the methods were not very efficient, these studies provide theoretical foundations in citation content analysis and shed light on later more fine-grained classification.

2.2 Citation Function and Polarity Classification

Citation function classification has been extensively explored since the last century, researchers from various scientific domains have proposed different approaches to investigate and describe the nuances among citation functions. Some of these nuances differ on subtle emotions that may represent authors' true opinions towards the cited works. Therefore, such emotions are

identified as citation polarity and were also categorized by researchers when classifying citation functions, although more in an auxiliary way.

According to Ding et al. (2014) [16], there are two major approaches to map the proposed classification schemes to citation corpora. The first approach, which was used more frequently by earlier studies, extracts fine-grained cue words or phrases based on linguistic rules and the classification scheme then makes a decision tree of these cues to classify citations. The second focuses on applying machine learning methods, such as Support Vector Machine [13] or Naïve Bayes [32], to build different classifiers and classify citations based on extracted features. Though the second approach can deal with larger corpora, it also integrates linguistic knowledge and hand-crafted rules to reduce the generated noise during the feature extraction process.

Garzone (1997) [20] introduced a cue phrase based citation function classification scheme in which he broke down citation content into 35 categories. This scheme amalgamated several other schemes and gave special attention to Finney's because of its applicability and insights. However, Garzone also identified two limitations of Finney's scheme, that some citation functions were not covered at all, while many existing categories of this scheme are too broad to capture the nuances inside citation functions. In consideration of these shortcomings, he further divided citation functions into more categories, which resulted in ten top-level types: negational, affirmational, assumptive, tentative, methodological, interpretational/developmental, contrastive, future research, use of conceptual material and reader alert. It is noted that this original proposed scheme contained 34 categories instead of 35, and it was later modified to match semantic parsing rules for cue phrase extraction.

After the original citation function scheme was developed, Garzone implemented an automatic citation classifier using computational linguistics techniques. The most important component of this classifier is a semantic parser, which consists of 195 lexical matching rules and 14 parsing rules and was used to find cue phrases in citations. Other main components are a tagger and a syntactic parser, working together to find the parts of speech for each citation sentence. Eight physics articles and six biochemistry articles were randomly chosen from scientific journals which contain a total of 547 citations from 419 citation sentences and served as the development data set; three physics articles and three biochemistry articles were randomly picked from the same pool, and served as the test data set. Then the classifier was tested on both development and test data sets, and was tested by locations in the article separately. The classification results showed it has good performance on the previously seen data set but fair performance on the previously unseen biochemistry data set. The physics articles got poor performance because of their less well-defined structure. Thus, Garzone suggested several ways to improve automatic classfication, including using appropriate machine learning

techniques to augment the lexical and grammar rules, and improving the section determination algorithm. Other similar studies, such as Nanba et al. (2000) [34] also conducted cue phrase and location-based citation function classification, and the results were not quite satisfactory as well.

The automatic approach and a scheme for citation function classification were also well-studied by another researcher, S. Teufel. In her early works, Teufel found the diverse writing styles of authors from different fields were related to different article sections, and she divided the sentences where authors' arguments appeared into 7 categories: background (generally accepted background knowledge), other (specific other word), own (own work method, results, future work), aim (specific research goal), textual (textual section structure), contrast (contrast, comparison, weakness of other solution) and basis (other work provides basis for own work). This classification scheme was further studied and extended to 12 categories in her later work.

In Teufel et al. (2006a) [39], which is one of her most prominent works in citation content analysis, she and her colleagues compared several citation function schemes from the last century, and argued that most of them are too sociologically oriented thus hard to operationalize without expert knowledge of sociology and apply in other fields. Then, Teufel adapted Spiegel-Rösing's 13-category scheme as it is more flexible and generalized on most articles, and proposed her own citation function annotation scheme, which is designed for information retrieval applications and consists of 4 top-level types of 12 categories.

Though this scheme is detailed and intuitive, Teufel mentioned several potential problems for annotating citations. Firstly, it may be difficult for annotators to interpret authors' intentions on citations. To deal with this, Teufel encouraged annotators to understand citation texts at a general level instead of using further knowledge of a specific field or of the authors, and assigning a particular function only when there is textual evidence found. Secondly, in general authors do not state their purpose clearly and express their opinions, especially negative ones, with hedges. Moreover, it is also particularly hard to distinguish the usage of a method in the citations that state a similarity between a method and the author's own method, and hard to distinguish between the continuation of a previous research and simply referring to it, as well. These difficulties were given special attention later when I annotated my own citation corpus.

This annotation scheme was preliminarily tested using inter-annotator agreement, that is, three annotators manually annotated a corpus of 26 articles with 548 citations extracted from computational linguistics journals. The results showed good performance on overall distinction, which implies this scheme is well-defined and reliable, although some semantic categories, especially *PSup* (cited work supports or is supported by citing work) and *PBas* (citing work is basis for citing work), were less well-understood by annotators due to different subjective judgments on citations. For the machine learning test, a larger corpus of 360 articles from

the same domain was prepared. Similar to Garzone's research, Teufel also used cue phrases in citation classification, but as sentence features. Other features are verb tense and voice, modality of main verbs, locations of citation sentences and self-citations. A k-nearest neighbor classifier with 10-fold cross-validation was employed and tested on extracted citation features. The classification results scored at 79% overall accuracy for four top-level citation function classes and 83% for three sentiment classes (weakness, positive, neutral), demonstrating a strong relationship between citation function and sentiment classification.

Compared with previous citation function and polarity studies that focused on finding a well-defined classification scheme, Abu-Jbara et al. (2013) [1] is concerned more about classification approaches. In this research, Abu-Jbara and his colleagues developed a six-category citation function classification scheme, which was mainly chosen from Spiegel-Rösing's 13 categories, Teufel's 12 categories and Nanba's 3 categories (*Basis*, *Comparison*, *Other*), in order to better serve bibliometric measures and applications. The six categories are criticism, comparison, use, substantiation, basis and neutral. The data sets used for experiments are composed of 3271 citations extracted from the ACL Anthology Network corpus, and were annotated with respect to polarity and purpose. To prepare the data, they firstly used regular expressions to find references and replaced them with placeholders, then identified grouped references, and applied a rule-based algorithm to remove non-syntactic references.

In addition, Abu-Jbara and his colleagues took citation context into consideration to improve classification accuracy. Before classifying citations, they employed a Conditional Random Fields (CRFs) model, which was trained on structural and lexical features of citation sentences, to sequence-label a window of four sentences as citation context. Then several classifiers including SVM, Logistic Regression (LR) and Naïve Bayes were built to classify citation polarity and functions separately. It is noted that the authors used a binary classification scheme for citation polarity, that is, citations were classified as Polarized (Subjective) or Neutral (Objective) at first, then subjective citations were classified as positive or negative. According to the classification results, lexical features that characterize the words surrounding the citation are more important, and identifying citation context enhances the detection of polarized citations as well as classification accuracy.

Similarly, Hernández-Alvarez et al. (2017) [22] also applied citation context to citation classification. They defined the citation context for a citation as the sentence that contains the reference together with its adjacent sentences. Unlike Abu-Jbara's automatic detection method using CRF, Hernández-Alvarez and her colleagues instructed several annotators to manually detect the citation context. Through several experiments, they determined the suitable window size of context is one, two or three sentences, with one-sentence contexts being the majority window size in their corpus. Therefore, I have set the window size of citation context as one

sentence in this thesis, that is, the sentence that contains the reference.

A novel aspect of the research in Hernández-Alvarez et al. (2017) is their incorporation of INFLUENCE as a third dimension to their citation classification scheme in addition to citation function and polarity. They proposed three categories for INFLUENCE: perfunctory, significant positive, and significant negative. Their citation function and polarity classification scheme consists of 4 top-levels with 8 categories, which are USE (based on, supply, useful), COMPARISON (contrast — comparison results are positive, negative and neutral), CRITIQUE (weakness, hedges) and BACKGROUND (acknowledge, corroboration).

To improve the consistency and accuracy of annotating with this scheme, they arranged a pre-annotation of cue words and phrases to help annotators assign a particular function and polarity category to a citation. This pre-annotation step is useful to some extent for solving the annotation problems previously mentioned in Teufel et al. (2006a). After obtaining the manual function and polarity classification, Hernández-Alvarez and her colleagues also processed the corpus with other citation information, such as frequency of citations found in each location and an influence classification measure obtained from a survey conducted on some authors. In this way, the whole corpus was deeply annotated in three dimensions and well-prepared for later automatic influence classification.

With more and more fine-grained citation polarity and function classification schemes being further examined and well studied, there has been growing interests in citation polarity (originally referred to as "sentiment") classification and it gradually became an independent topic in citation analysis research. In the next section, I will describe several prominent studies on citation sentiment analysis conducted with machine learning methods, and compare their approaches on detecting and extracting sentiment features from citation texts.

2.3 Citation Sentiment Analysis

In contrast with the prevailing opinion that all citations should be evaluated with the same weights, Bonzi (1982) [8] argued that if a cited article was criticized by an author in his/her own work, as a result, there should be allowed lower or negative weights for the criticized article during bibliometric measurement. This modified evaluation could be achieved by detecting citation sentiment manually, or better, with automatic methods.

Furthermore, citation sentiment detection provides researchers a particular approach to detect the potential problems inside academic papers. It could also be used as a reference in scientific summarization, recognizing the hidden issues and gaps of current works, thus helping people get better research directions. However, as previously mentioned, detecting the sentiment of a citation is a challenging task since the sentiment is often hidden. Citation sen-

tences are often objective and neutral, and authors are especially cautious about criticism and hedge negative sentiment within contrastive terms.

The traditional and common approach for classifying sentiment is to score the sentences with a labeled lexicon. However, this approach was considered highly topic-related and cannot provide a generalized sentiment classifier that is applicable in different domains. To deal with this narrow scope problem, researchers who work on movie review sentiment analysis built several machine learning classifiers that take sentence structure-based features as inputs, and achieved good performance on automatic sentiment classification.

Athar (2011) [3] focuses on identifying citation sentiments with automatic methods. In this research, the author claims that although the good classification results of classic sentiment analysis based on movie or product reviews seem promising, a well-defined automatic sentiment detection system developed from this genre might not perform well in the scientific domain. This is because sentiments in scientific articles are often hidden deeper, and the science-specific terms and technical terms, which play a major role in scientific writing, carry sentiments as well. Moreover, citations have a wider range of influence that may vary from a single sentence to several paragraphs, thus it's more difficult to capture all of the sentiments for a specific citation.

Considering these potential problems, Athar conducted experiments on various sentence features. He firstly extracted 8736 citations from 310 research papers in the natural language processing domain to create a new sentence-based corpus, and processed it with regular expressions to replace the citation text that contains authors' proper names with a special token, in order to remove any lexical bias. After labeling the corpus as positive, negative or objective, he found it is heavily skewed as subjective citations only occupied 14% of the corpus. This unbalanced sentiment ratio problem was later alleviated by sentiment context detection methods proposed in Athar and Teufel (2012) [5]. In the next step, he represented each citation with a feature set as input for a Support Vector Machine (SVM) system. This feature set consists of word level features, contextual polarity features and sentence structure features.

His classification results showed that only n-grams and dependency features have an obvious effect on citation text, and the negation window improved the performance but not in a significant way, which is possibly due to the skewed sentiment class distribution. Later, this classification system was tested on a subset of the data from the citation function corpus used in Teufel et al. (2006a, 2006b). The not satisfying results indicated that citation sentiment classification is different from citation function classification.

Not long after Athar's original research, Athar and Teufel (2012) [5] claimed that sentiment analysis should take citation context into consideration. They employed a four-class scheme for annotating the corpus, in which every sentence that is in a 4-sentence window of the citation

and does not mention the citation directly or indirectly was labeled as *x*, and the rest of the sentences in that window were labeled as positive *p*, negative *n* or objective/neutral *o*, respectively. Moreover, if a sentence contains multiple sentiments, then it will be labeled with the class of the last mentioned sentiment. With this scheme, the number of subjective sentiment instances, especially negative sentiment, in the corpus, which is a subset of the data in Athar (2011), is greatly increased. However, their annotated dataset is still inevitably skewed as there are many more objective sentences than subjective ones in general.

In the next stage, Athar and Teufel took ten different context-related binary features as a feature set for each citation. These binary features were later summarized in Athar's PhD thesis [4] as formal citation, author's name, acronyms, work nouns, pronouns, connector, section markers, citation lists, lexical hooks and n-grams. Then they input these features to a SVM classifier and compared it with an n-gram only baseline system. They also built another baseline system using n-gram and dependency features, which are proved to be the most useful features in Athar (2011), to explore the effect of this context detection scheme. The results showed their SVM classifier outperformed the baseline systems in all evaluation aspects of citation sentiment classification. Therefore, we may infer that ignoring the citation context would lead to a loss of sentiment in the citation corpus, especially for the negative ones.

Most automatic citation sentiment analyses in the literature were conducted on a corpus extracted from the natural language processing domain. Since this thesis uses a biomedical corpus for citation analysis, it is necessary to address some biomedical specific approaches. However, there are very few studies about citation sentiment analysis in the biomedical domain, and most of them used a manual and exhaustive approach to analyze sentiment in the biomedical citations, such as Yu (2013) [45], which I will mention in the next section.

One of the first studies that applied automatic sentiment classification on biomedical citations is Xu et al. (2015) [44], in which the authors created and annotated a citation corpus composed of clinical trial papers. The authors state that analyzing biomedical citation sentiment may provide a potential application to detect non-reproducible studies thus avoid wasting resources. They constructed a biomedical corpus containing 4182 citations extracted from the discussion section in 285 randomly selected clinical trial articles, since the citations that contain author's opinions or sentiment were mainly found in the discussion section, according to their examination. For corpus annotation, they proposed a decision tree strategy of a series of binary questions for annotators to answer, in which will result a sentiment label assigned for a single citation.

To improve the classification accuracy, the authors employed citation context detection both in corpus annotation and sentence feature extraction phases with a rule-based method and a set of cue phrases. There were three categories of features extracted from citation contexts:

word n-gram features, sentiment lexicons and sentence structural features. These features were merged into a feature vector for each citation context, as the input for an optimized SVM classifier. During the experiment, different combinations of features were tested, and the classification results showed the combination of all features reached the best overall performance and scored the highest F-value for each sentiment class. This may indicate that using automatic methods to analyze citation sentiment in biomedical domain is plausible and promising. However, the skewed corpus and fair inter-annotator agreement imply that this task still remains challenging, even for domain experts.

2.4 Other Related Studies

Besides the previously described studies on citation polarity and function classification, there are some other research works that are worth mentioning.

In Mercer and Di Marco (2003) [31], the authors extended the ideas in Garzone (1997) by claiming that the fine-grained cue phrases within citation sentences play a crucial role in citation function categorization. They reviewed Garzone and Mercer's "pragmatic grammar", which aims to represent the characteristic structural patterns in each citation function category, by using cue phrase-based lexical rules and grammar-like rules that can handle more complicated patterns. As a direct contrast to Garzone and Mercer, according to the authors' view, Teufel casted doubts on the existence of fine-grained cue phrases in citation contexts, and further questioned the applicability of automatic methods for detecting these cue phrases if they do exist.

Though Teufel held an opposite opinion, Mercer and Di Marco thought her claims may imply the importance of cue phrases and the possibility of detecting them by automated means. Taking her claims as a starting point, the authors first reviewed previous studies of discourse analysis and rhetorical relations, to get theoretical support for cue phrase identification. Then they created a corpus consisting of 24 scientific articles for analyzing the frequency of cue phrases in three components, which are full text body, citation sentences and citation window. The results showed cue phrases do exist in citation contexts and the distributions vary among locations. Moreover, the automatic detection of these cue phrases have been previously documented in some studies, which supports the authors' assumption that cue phrases can be extracted by computational methods. This discourse analysis confirms the significance of cue phrases and their applicability in citation function classification.

A more-recent study Bertin et al. (2016) [7] investigates the linguistic patterns and rhetorical structure of citation contexts by applying n-grams. Their n-gram extraction results provide evidence for Mercer and Di Marco (2003)'s statement that cue phrases exist in citation context

and can be extracted automatically. However, the authors also identified several limitations of automatic classification, such as difficulties in distinguishing citation functions and establishing a one-to-one relationship between trigram-patterns and common-word classes, which might be solved by detecting more significant surface patterns.

Although the automatic classification of citation function and polarity is dominant in current citation analysis studies, some researchers argue that the present automatic approach may not reflect the true behavioral patterns of authors in citing articles, thus not meeting users' unique needs in assessing citations. To amend this gap, Yu (2013) [45] reviewed several publications on manual approaches to detect citation bias in the biomedical domain, and compared the methodological differences between automatic analysis and biomedical researchers' methods. He mentioned that current automatic classification mainly focuses on creating a typology of citation functions, while no one investigated whether these schemes are really needed to assist researchers in literature review. After examining all citations in six papers, Yu found there were linguistic cues existing that are helpful for classifying citations and could be identified by computer. However, some citations do not contain any explicit cues or require domain knowledge to make decisions, which makes it a challenging task for computers in the automatic reasoning process. Yu concluded that citation sentiment strength and validity should be given more attention during analysis. This may indicate that the future works on citation function and polarity should integrate more human efforts and domain knowledge to help improve the accuracy of automatic citation classification and fulfill researchers' specific needs in obtaining comprehensive information from prior literature.

2.5 Summary

This chapter firstly describes the theoretical foundations of syntactic and semantic citation content analysis, and introduces several general citation classification schemes from early research. Then the studies on citation function and polarity classification are highlighted, in which both rule-based and automatic approaches for mapping the schemes to corpora are compared and discussed. Spiegel-Rösing's, Teufel's and Garzone's classification schemes are given special attention since they inspired the scheme proposed in this thesis. The citation sentiment analysis section mainly investigates Athar's automatic classification on citation sentiment, which involves sentence-structure feature extraction and supervised machine learning methods. The disadvantages in Athar's experiment were later amended by his co-works with Teufel, which improves the classification accuracy with citation context detection. A citation sentiment study in the biomedical domain was examined closely to meet the specific needs for this thesis. Lastly, another two studies provide evidence for the importance of fine-grained cue phrases

in citation classification, and demonstrate possible automatic methods for cue phrase extraction. A comparison between manual and automatic citation sentiment analysis points out the shortage in current research and suggests what researchers should emphasize in future works.

Chapter 3

Design of A Citation Function and Polarity Scheme

This chapter briefly discusses several citation function and polarity classification schemes from previous studies, and proposes a new 3-dimensional scheme that combines advantages of other schemes. Garzone's [20] and Teufel's [39]schemes are given special attention since Garzone's is the most comprehensive and fine-grained, and Teufel's is well-adapted from Spiegel-Rösing's [37] scheme and augmented for automatic classification.

3.1 Previous Schemes and Their Limitations

Spiegel-Rösing's 13-category scheme is widely adopted by many later researchers due to its reasonable categorization and easy operationalization. It addresses citation evaluative use and citation content, thus several categories are defined together for concepts, methods and data from the cited source that are applied in the citing article. However, these categories, such as "cited source contains the data which are used sporadically in the article" and "cited source contains data and material which is used sporadically in the citing text, in tables or statistics" have a large overlapped portion and could be grouped into one summarized category. In addition, a citation may be assigned more than one category according to the categories' descriptions. For example, "cited source is positively evaluated" and "cited source is negatively evaluated" examine citations' polarity which could be on top of other citation function categories.

Teufel modified Spiegel-Rösing's scheme and proposed 4 top-level groups composed of 12 categories in total. She inherited the two polarity categories, changing the negative label to weakness, and also added two more classes: contrast and neutral. The contrast top level divides the comparative part from Spiegel-Rösing's scheme into four categories, which not only

capture the distribution of comparison/contrast in different article sections, but also distinguish compare/contrast explicitly or implicitly. The categories in Spiegel-Rösing's scheme that classify the usage of data, methods, concepts of the cited article are merged and redesigned as six categories in Teufel's scheme. However, some of them caused much confusion during the annotation process since understanding an authors' interpretation of source content is subjective thus different annotators may have a variety of judgments. For instance, it is hard to distinguish whether the author just simply uses data or methods from previous research, which is defined as the *PUse* category, or takes cited content as starting point, which is defined as *PBas*. These categories are merged into one category in my own scheme.

Garzone's 35 categories provide a comprehensive scope of citation classification. This fine-grained scheme covers almost all possible citation functions and gives a clear description for each category. Some categories also define the degrees of an author's sentiment towards the cited source, such as "citing work *totally* confirms cited work" and "citing work *partially* confirms cited work". However, this is also a disadvantage for corpus annotation since the sentiment degrees mainly depend on the annotators' subjective judgments thus are difficult to be measured quantitatively. Another novel design of this scheme is it implies citing directions between citing and cited works, such as "citing work is totally *supported by* cited work", in which the direction points from the cited work to the citing work. Such direction information is particularly useful for identifying criticism from the citing work's author, as a cited article is not able to criticize an unfinished or unpublished paper (the citing work) due to the logic of time. In this thesis, the citing directions are adopted as one dimension of my own classification scheme.

Although Garzone's scheme elaborately describes the relations between citing and cited work, it is not flexible for automatic classification. Many of the categories have small differences that are difficult even for human annotators to identify, such as "general background" and "specific background", thus such nuances are extremely hard to be recognized by current computational linguistics techniques. Furthermore, some categories are unnecessary for investigating citation relations and could be merged into one class. A proof for this statement is cue phrases are missing in many categories during Garzone's extraction process on test dataset.

To meet the specific needs of automatic citation classification, Abu-Jabara summarized the characteristics of previous schemes and reduced the number of citation function categories to six [1]. These categories are selected for improving bibliometric measures and generalized for a computer program to recognize sentence features. The citation polarity annotation scheme is a two-step method, in which it firstly distinguishes neutral or subjective citations and then classifies subjective citations as positive or negative. Though this method helps improve sentiment annotation accuracy, citation function and polarity are not cross-classified thus correlation

information is lost.

3.2 The Proposed Citation Classification Scheme

As mentioned in previous section, the scheme proposed in this thesis is adapted from former empirical works in content-based citation analysis. This scheme is designed in three dimensions: citation function, citation polarity and citing directions. In line with most of the previous schemes, the three polarity classes are on top of 8 citation function categories, as listed below:

- Positive: author of the citing paper agrees with or makes use of opinions/theories/data in the cited paper (2 categories)
- Negative: author of the citing paper disputes opinions/theories/data or pointed out a weakness the in cited paper (2 categories)
- Neutral: author of the citing paper shows neither positive nor negative sentiment towards the cited paper, or the cited content functions differently from the two classes stated above (4 categories)

The citing direction dimension is inspired by Garzone's scheme and is designed to be intertwined with the other two dimensions in this thesis. It provides directional information about the relationship between the citing and cited paper as well as a particular method to recognize the criticism hidden in the citing paper author's rhetorical hedges. Similar to citation polarity, the citing direction dimension also has three classes, which are shown as followings:

- citing-to-cited: the citing work refers to/confirms/disputes cited work (4 categories)
- cited-to-citing: Data/results from cited work supports/proves the citing work (1 category)
- no direction: the cited work states facts/problems, or compares with other cited works, thus has no interaction with the citing work (3 categories)

The function dimension is mainly governed by polarity and direction dimensions, although citing directions do not exist in every function category. Each category is presented below with an example citation sentence serving as a further clarification for category description.

Neutral Type Categories

1. Perfunctory/Background

This category is merged from the categories that introduce background knowledge in Teufel's and Garzone's schemes. It describes the situation in which the citing article refers to methods/data/theories/statements of the cited article as general introduction, and the cited content is

not analyzed or compared with other studies or the citing article. The citing direction for this category is citing-to-cited. Example:

> **As previously reported** [11], the H. pylori arginase in the E. coli model (pBS-rocF) displayed optimal catalysis with cobalt at pH 6.0.

2. Statement

This category is newly designed in this scheme. It describes the situation in which the cited work states results/phenomena/data and has no interaction with the citing work. There is no citing direction in this category. Example:

> Snyder et al. [29] concluded from electron density profiles of LPS R60 that its charges are located mainly in two distinct planes which are separated by a distance of 1.1 nm.

3. Comparison

This category is a merger of the comparison categories in Teufel's and Garzone's schemes. It describes the situation in which the citing work compares methods/experimental results from its own research with those from cited work. It it noted that this comparison does not contain any affirmative or negative sentiment towards the results. The citing direction for this category is citing-to-cited. Example:

> The Km glyoxylate (70 μM), Km acetyl CoA (12 μM) and Vmax (16.5 μmol/mg MSG) of the P. aeruginosa PAO1 MSG are **comparable to** those of other malate synthases available from literature (Table 1, [16-21]).

4. Multi-comparison

This category is newly designed in this scheme. It describes the situation in which the citing work demonstrates a comparison of results/data among several cited works, in which results from its own research is not involved. There is no citing direction in this category. Example:

> This is a departure from **several** earlier studies relating to HSF1 phosphorylation, including one from our own group [28], in which studies exogenous HSF1 forms were substantially overexpressed.

Positive Type Categories

5. Confirmation

This category is adapted from the positive type in Teufel's scheme. It describes the situation in which the citing work confirms or extends statements/results/theories from the cited work. The citing direction for this category is citing-to-cited. Example:

> The overall affinity determined in the present study for Cry1Aa to BtR175 (2.6 nM) **agrees well with** the findings of Ihara et al. [16] by a different assay (0.8 nM).

6. Being-confirmed

This category combines the strongly and weakly affirmative categories in Garzone's scheme. It describes the situation in which the citing work is confirmed/supported/boosted by data/theories/statements from the cited work. The citing direction for this category is cited-to-citing. Example:

> **In support of** our interpretation, Rilling et al. [57] reported that protein prenylation in Chinese hamster ovary cells can vary as a function of the extracellular mevalonate concentration.

Negative Type Categories

7. Contrast/Conflict

This category is adapted from the contrastive type in Teufel's scheme. It describes the situation in which the citing work has different results/opinions or disputes the cited work. The citing direction for this category is citing-to-cited. Example:

> This finding **contradicts** a previous study showing direct binding of SET/TAF-Iβ to the H3 N-terminal tail, which is disrupted when the tail is modified [34].

8. Unsolved

This category is newly designed in this scheme. It describes the situation in which the cited work has unclear results or statements that remained controversial even after the citing work is finished or published. This category has no citing direction. Example:

> Lyapina et al. recently demonstrated that partially purified CSN promotes the cleavage of Pcu1p-Ned8p conjugates in vitro [10], but it **remained unclear** whether the enzymatic activity is contained within one of the CSN subunits or a tightly associated peptidase.

3.3 Summary

In this chapter, I compare and discuss several citation classification schemes of Spiegel-Rösing, Teufel, Garzone and Abu-Jabara. Both advantages and disadvantages of their schemes are identified, which have inspired my own classification scheme. Some of the categories from previous

schemes, such as those dealing with data and concepts, are amalgamated and redesigned to fit the automatic classification approach. Besides citation function and polarity dimensions, my scheme takes citing direction as an extra dimension to give a finer granularity on relationships between cited object and citing subject. As a result, my proposed scheme is defined by 3 dimensions, with 3 top-level polarity classes of 8 function categories. The application of this scheme on corpus annotation will be discussed in the next chapter.

Chapter 4

Corpus Construction and Annotation

This chapter describes the corpus I constructed for training and testing the classifiers of citation function and polarity. Since the machine learning methods applied in this experiment are supervised, I need to annotate each citation sentence according to its function and polarity as the gold standard for calculating classification accuracy.

At present, there does not exist any publicly available biomedical corpora with the expected annotation. Most of the citation classification research with semi-automatic approaches were conducted in the natural language processing domain, thus the corpora used in previous studies mainly consists of papers extracted from the ACL (Association for Computational Linguistics) Anthology, which have different characteristics from biomedical papers. However, with the help from my supervisor Dr. Robert Mercer, I was able to obtain the necessary biomedical data for developing my own annotated corpus, which I'll introduce in Section 4.1. The cue phrases proved to be useful for citation classification, and they do exist in citation context and could be extracted automatically (Mercer and Di Marco, 2003) [31]. Therefore, in Section 4.2, I will discuss the importance of cue phrases and how I used them in annotating this corpus. Section 4.3 describes a linguistic rule-based method of extracting cue phrases from citation sentences by trimming sentence parse trees. Section 4.4 explains how to assign a particular function category from classification scheme to each citation sentence, and decide a citation's polarity by ranking its possible labels. Both annotations are mainly based on cue phrases. Lastly, Section 4.5 shows some statistics of this annotated corpus, including the distribution of each classification category and the number of cue phrases in different article sections.

4.1 Corpus Construction

I chose citations from biomedical research papers to build my own corpus for the following reasons. As previously mentioned, Garzone's citation function classification scheme was pro-

posed and used to annotate some biomedical papers, and his semantic rule-based parser for cue phrase extraction was also developed from this small corpus. Therefore, building a larger biomedical corpus could further investigate the use of cue phrases as well as compare with his research results. Secondly, as far as I am concerned, there is no publicly available annotated citation corpus in the biomedical domain at the moment, so creating and annotating such a corpus may contribute to later biomedical citation studies.

The original biomedical papers in full body texts were firstly downloaded with the help of PubMed. PubMed is a free resource of over 22 million citations and abstracts for biomedical literature with the National Center for Biotechnology Information (NCBI)'s search and retrieval system integrated. It does not include the full text for the biomedical publications. However, a small portion of full length articles are available and can be obtained from PubMed. Each downloaded paper is originally one huge line comprised of all of the sentences in all article sections as well as the section titles, and written in XML format. Following the citing customs, each citation in a sentence was given an index number which points to the cited article listed in the *References* section at the very end of the current paper. To make the whole article more readable, the huge line was split by sentence scope and section scope using some programming tools, thus resulting in one sentence or one section title per line in the newly saved file. This task was mainly done by my supervisor Dr. Robert Mercer.

As I am only interested in citations and its related location information, I wrote a Python script to extract citation contexts from the IMRaD (Introduction, Method, Results and Discussion/Conclusion) sections, and saved them in a new file for each biomedical paper. The citation context window size in this thesis was set to 1, that is, the sentence that contains the citation. This decision was based on the findings from Hernández-Alvarez et al. (2017) [22], in which they show one-sentence contexts take a large majority of the citation corpus. The citation sentence was identified by regular expressions according to its citation index, which can be a single number representing one cited article such as "[9]", or multiple numbers representing grouped articles such as "[3-5]" or "[4, 6-9]".

As shown in Figure 4.1, citations from the same IMRaD section were grouped together, and each section was divided by a long dash line followed by section title. This group method was adapted from Garzone's data cleaning process, since it is convenient for computer programs to recognize and deal with each paper section independently. The leftmost column of numbers stands for the line numbers of citations in the current file, and the number at the beginning of each citation sentence represents its line number in the original article file. When a citation sentence cannot provide enough evidence for assigning a particular function or polarity category to the citation inside, this location reference will be helpful for tracing back to the original paper and identifying adjacent texts of the current citation sentence, where may exist hidden

information that is in need of category assignment.

```
74  The same occurred with PR of which the cervix and
oviduct had lower concentrations ( 167 ± 71 and 308 ± 117 )
than the uterus ( 1479 ± 153 ) [ 17 ] .

------------------------------------------------------------
Discussion
91  [ 37 ] , reported few ER immunopositive cells in the
cervix of two prepubertal ewes .

92  In contrast , in this study , high levels of ER and PR
were found in both cervix and oviduct , although less than
in the uterus of the same animals [ 17 ] .

94  The presence of ER was expected since it is known that
ER are needed for normal development of the reproductive
tract [ 16 ] , but it is not clear why PR levels were so
high at this stage of development .
```

Figure 4.1: Sample data for the extracted citations in a section

4.2 Use of Linguistic Cues

As previously mentioned in the literature review, Mercer and Di Marco (2003) [31] claims exploring the fine-grained rhetorical structure of a scientific article may greatly assist in citation classification. This statement implies two things: on the one hand, the rhetorical relations that address discourse structure in scientific articles could match with citations; on the other hand, rhetorical relations usually indicate the author's purpose and attitude on using citations referred to a certain article, which may in turn provide much help in classifying a citation.

Several discourse analysis studies have demonstrated that fine-grained linguistic cues play a crucial role in scientific rhetorical structure. According to Knott's definition [25], a cue phrase is a linguistic conjunction or connective for maintaining the cohesion and coherence in general texts, and indicates the semantic relationship between two sentences or clauses. Through the precise test on cue phrases, Knott proposed a five-category classification scheme of linguistic cues, which was later adopted in Mercer and Di Marco's work. More significantly, Knott combined two methods for dealing with cue phrases in discourse analysis, thus providing a solid foundation and evidence for associating rhetorical relations with cue phrases in citations.

Furthermore, as discussed by Mercer and Di Marco, another discourse analysis researcher, Daniel Marcu, extended the Rhetorical Structure Theory (RST) to a rhetorical parsing algorithm, and implemented this algorithm in a rhetorical parser which uses cue phrases to capture the "hypothesized rhetorical relations", showing the possibilities of obtaining rhetorical

structure information through automatic means [29]. Similarly, Garzone applied a semantic grammar in his proposed automatic citation classifier, which is able to capture the rhetorical features and extract cue phrases from citations automatically [20]. These cue phrases and their related citations were later classified according to Garzone's 35-category scheme and obtained good classification performance on previously seen data.

Through the analysis conducted on full text body, citation sentences, and citation windows of several articles, Mercer and Di Marco argued that cue phrases do exist in citation contexts, and the distribution of their frequency in citations is analogous to that in full article texts. This conclusion was accepted by more and more researchers in later citation analysis studies. Teufel, who questioned the existence of fine-grained linguistic cues in citations and casted doubt on the applicability of automatically extracting cues in her early studies, took cue phrases as one of the features in her more-recent citation function classification work. In Teufel et al. (2006a) [39], she and her colleagues implemented two POS-based mechanisms for modeling cue phrases, one used a finite grammar to extract string-based cue phrases, the other integrated recognizers for agents and actions to identify the similar cue phrases that were clustered around main verbs. In addition, they encouraged annotators to record cue phrases that helped assign a category to a citation, and included 12 features of these citation function-related cue phrases in their classifier. The classification results demonstrated good overall performance, from which we may infer cue phrases do assist in automatically classifying citation function and polarity.

Besides improving classification accuracy, cue phrases also greatly contribute to corpus annotation. As described in Hernández-Alvarez et al. (2017) [22], the annotators' categorization opinions on the same citation text varied quite differently at the beginning of the annotation process. To improve the inter-annotator agreement (IAA), the authors employed a pre-annotation step that requires annotators to mark the cue phrases, which are basically keywords and tags, in order to clarify the meanings of citation contexts. One possible explanation for the significant improvement on IAA is this pre-annotation of cue phrases set up matching mental models among different annotators. Moreover, the identified cue phrases contain rhetorical information of citations, and could serve as input features of a classifier.

In summary, the rhetorical nature of cue phrases decides their important role in examining relationships between citing and cited works within a citation. In this thesis, I will use cue phrases as an intermediate agent to map the classification scheme to my corpus. Since the biomedical citation texts have a more complicated structure than those from the natural language processing domain, I developed a mechanism to help identify the cue phrases from surface structure, which I will give details in the next section. It is noted that the characteristics of cue phrases in biomedical contexts are slightly different from both the general interpretation and those in other science domains. Therefore, it requires more human effort to decide the real

cue phrases that correctly reflect relationships between citing object and cited content.

4.3 A Method for Cue Phrase Extraction

To extract the cue phrases from citations, I firstly used a toolkit named the BLLIP parser to process each citation sentence and got its syntactic structure, which can be represented in a tree form. The BLLIP parser[10], also known as the Charniak-Johnson parser or Brown Reranking Parser, is a statistical natural language parser consisting of a generative constituent parser and a discriminative maximum entropy reranker. The coarse-to-fine generative parser constructs 50-best sentence parses that are of substantially high quality based on a dynamic programming n-best parsing algorithm. Then the MaxEnt discriminative reranker takes this set of 50-best parses as input and selects the best parse by examining a wide variety of features. This system outperforms most of the present publicly available parsers. I use this reranking parser mainly because it chooses the best sentence parse from a high quality parse pool, which may give better performance on parsing more complicated sentence structure in biomedical texts.

The parsed sentences are arranged in a flat form by phrase markers, which are basically round brackets, and phrasal category labels (S, VP, NP, etc.). Although they can be transformed to syntactic parse tree form for a more intuitive and hierarchical look, this graphical tree-representation does not help in data processing and programming. Therefore, I employed a bracket counting algorithm to catch different syntactic levels in shallow form. A code snippet of this algorithm is given in Listing 4.1.

Listing 4.1: Code snippet for bracket counting algorithm

```
def label_bracket(parsed_sentence):
    parsed_sentence[2] = ``#''  # replace S1 with S# to avoid confusion
        with bracket label
    counter = 0
    for index, item in enumerate(parsed_sentence):
        if item == ``('':
            counter += 1
            parsed_sentence[i] = str(counter)
        elif item == ``)'':
            parsed_sentence[i] = str(counter)
            counter -= 1
    labeled_sentence = ``''.join(parsed_sentence)  # combine every item
        to a sentence
    return labeled_sentence
```

As shown in Listing 4.1, this algorithm reads a parsed sentence in left-to-right order and

checks each character in the sentence including punctuation. A counter is set to zero at the beginning of the checking process. It increases by 1 when a left round bracket appears, and decreases by 1 when meeting a right round bracket. A number that represents the current value of counter replaces the round bracket as soon as the counter value gets updated, thus the phrases or clauses that are in the same syntactic level will be surrounded by the same pair of numbers. The number "1" in the tree root "S1" is substituted by "#" to avoid confusion with the bracket label. Although the sentence still remains in a flat form, the numbers on round brackets' positions could pass the hierarchical information of syntactic structure to a computer program. For example, the parsed citation sentence

> This is in disagreement with [17, 12] and [14].
>
> (S1 (S (S (NP (DT This)) (VP (VBZ is) (PP (IN in) (NP (NN disagreement))) (PP (IN with) (S (PRN (-LRB- -LSB-) (NP (NP (CD 17)) (, ,) (NP (CD 12))) (-RRB- -RSB-)) (CC and) (LST (-LRB- -LSB-) (CD 14) (-RRB- -RSB-)))))) (. .)))

will change to: (the underscore "_" highlights the bracket number)

> 1S# 2S 3S 4NP 5DT This5 4 4VP 5VBZ is5 5PP 6IN in6 6NP 7NN disagreement7 6 5 5PP 6IN with6 6S 7PRN 8-LRB- -LSB-8 8NP 9NP 10CD 1710 9 9, ,9 9NP 10CD 1210 9 8 8-RRB- -RSB-8 7 7CC and7 7LST 8-LRB- -LSB-8 8CD 148 8-RRB- -RSB-8 7 6 5 4 3 3. .3 2 1

In the next step, I developed a sentence splitting (or parse tree trimming) mechanism based on linguistic rules to extract cue phrases that are in different locations of a syntactic structure. This method is mainly inspired by Athar (2011) [3], where the author tried to remove irrelevant polarity phrases around a citation to improve the classification results. He parsed each citation sentence into a tree form, then kept only the deepest clause inside the subtree which contains the citation phrase, and discarded all other clauses of the citation sentence. In this thesis, I applied similar trimming rules but kept all possible cue phrases that are close to a citation. This is because biomedical texts have longer and messier clauses, which causes the related cue phrases to possibly be distant from citations. Moreover, the "meaningless" cue phrases that exist in citation sentences from ACL articles may have different meanings and sentiments in biomedical contexts. Therefore, keeping only the cue phrases that are instantly followed by citations might lose rhetorical information about citing and cited articles.

During the development of this cue phrase extraction mechanism, I always tried to find a balance between generalized and specific syntactic rules. If the rules catch too many details about sentence structure, they may not be flexible and applicable on new data. However, if they

are too generalized and broad, many irrelevant words will be included and give much noise. Through the close examination on citation data, I came up with four top levels to capture characteristics of different parts of sentences that may contain cue phrases.

NP-VP Type

This is the most common and prevalent structure in sentences. In general, NP (noun phrase) and VP (verb phrase) are directly under the root node "S", and the related cue phrase is inside the VP. Since cue phrases are usually word phrases that contain main verbs, the extraction begins from the current highest VP level and stops at the next-level NP.

> All sequences and positions **were referenced with** the wild-type HLA-G sequence of **Geragthy et al.** [16]

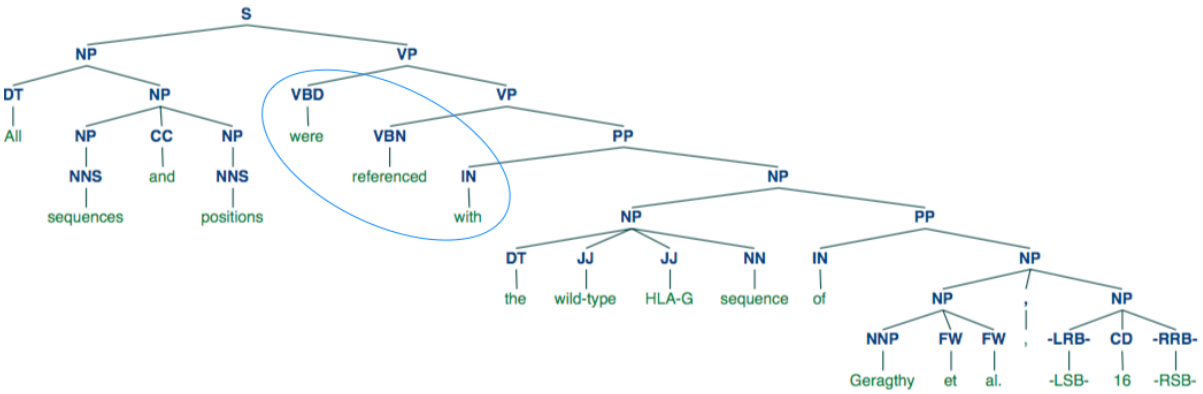

Figure 4.2: A cue phrase inside NP-VP structure

As shown in Figure 4.2, the circled part is where the cue phrase rests, and its extraction starts from the highest VP level, also known as the main VP, and ends before the one-level lower NP (*the wild-type HLA-G sequence of* . . .). The citation *Geragthy et al.* represents the only cited paper in this sentence and is located in the deepest NP level of the main VP.

SBAR Type

According to the Penn TreeBank's definition of part-of-speech tags, the label SBAR is for a clause introduced by a subordinating conjunction (also known as a complementizer), such as "that", "whether", or "if", which has a prepositional tag IN and can possibly be absent from the sentence. The cue word is usually the verb located before this complementizer. A citation sentence may have an SBAR clause at the beginning to serve as a context or general acknowledgment, thus this clause is parallel to the main NP and VP. More frequently, the SBAR appears in the middle of a sentence to explain the author's claim or findings.

[15], however, **found that** D-dimer was included in the final logistic model together with the variables heart rate and chloride.

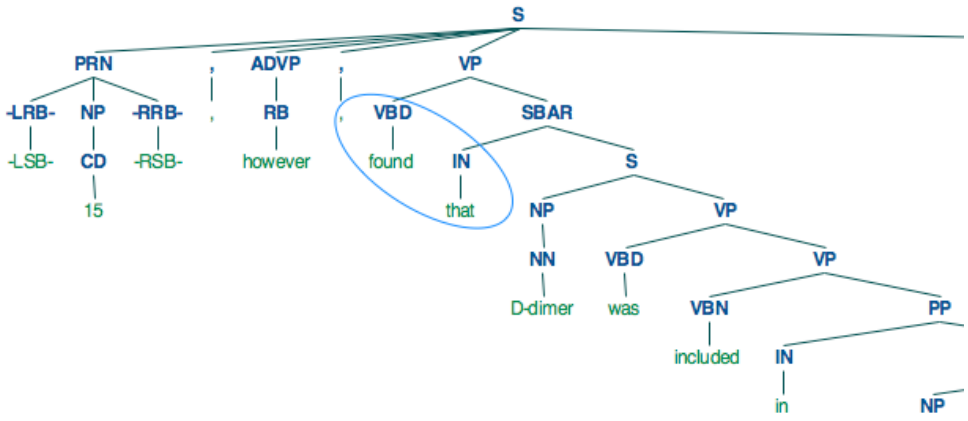

Figure 4.3: A cue phrase before SBAR clause

As shown by Pasquale et al [31], the planes of shear can be calculated from measured ζ-potentials and calculated surface potentials.

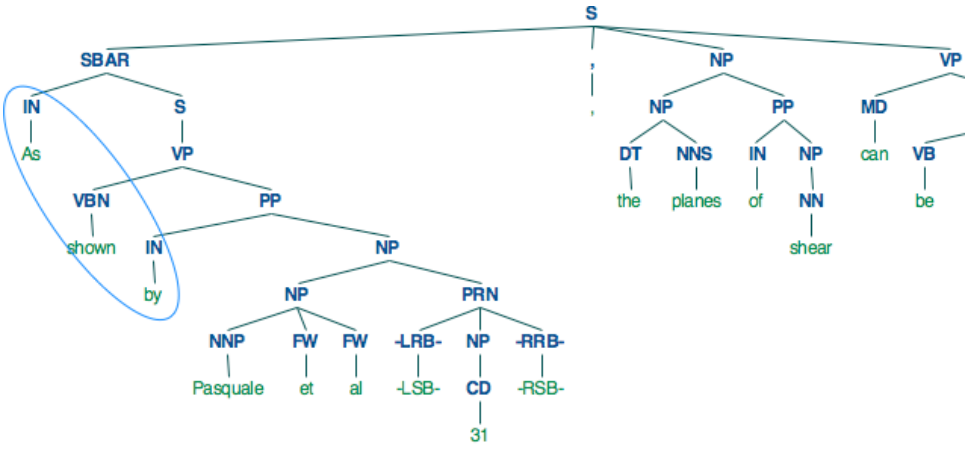

Figure 4.4: A sentence begins with SBAR clause

In Figure 4.3, the word "that" is the complementizer and a child of the SBAR clause. It is preceded by the cue word "found", which is one-level higher than the IN tag but the same level as the SBAR tag. The cue word in this structure usually has a VBD or VBN tag. In Figure 4.4, SBAR occurs at the beginning of the citation sentence as context or a general introduction, and it has the same level as the main NP and VP. Therefore, the extraction process starts from the SBAR level until the next-level NP inside the child node S, during which the words are all extracted, resulting in the cue phrase "as shown by".

ADVP/ADJP/PP Type

In most cases, ADVP/ADJP/PP is located under an S node and precedent to the same level NP and VP. It could be a single word implying a polarity, transition or further investigation, such as "however", "furthermore", etc., or it may serve as general acknowledgment in a phrasal or clausal form to provide context for the later cited content, such as "Based on Xxx method/data in Yyy et al., . . . ".

> **Recently**, a nuclear export signal (NES) was **identified** in the N-terminus of annexin II [40].

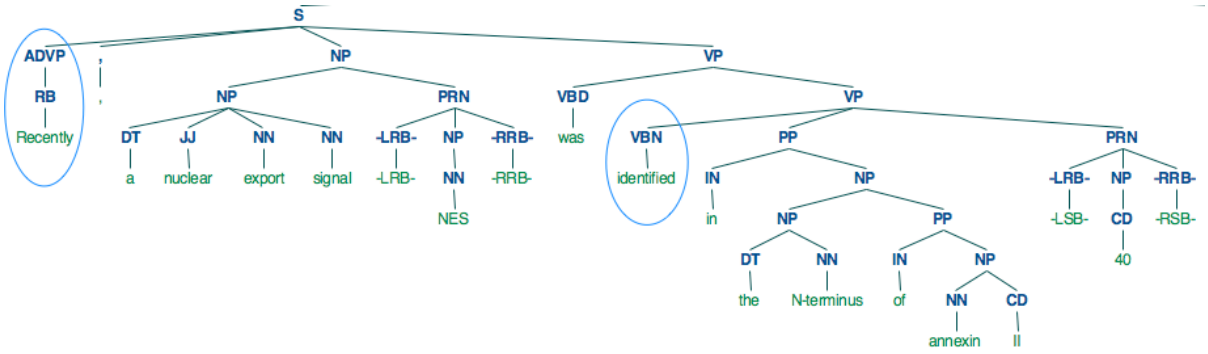

Figure 4.5: A sentence begins with ADVP

As shown in Figure 4.5, the word "recently" implies there will appear a citation in a later part of the sentence, thus words in this position are usually cue words and will be extracted. If there is a clause or word phrase appearing under the ADVP label, then the phrase that has this main adverb is a cue phrase and will be extracted instead. In general, the NP that is parallel to the ADVP contains an object from or related to the cited content, which shows a fact or makes a statement. Therefore, the main verb from the VP that is parallel to the ADVP and the NP is also a cue word, and will be extracted using its VBN label. These rules are applicable to ADJP/PP conditions as well.

Negational Type

The negational cues may imply criticism or disputation towards the cited content, or indicate the author of the citing paper has obtained different experimental results compared with the cited paper, which is more commonly seen in citation sentences. There are two types of negations: one is the word "no" followed by a noun, such as "no results" and "no association", as shown in Figure 4.6; the other is the word "not" followed by a verb, such as "not find" and "not observe", as shown in Figure 4.7.

We identified 1 unpublished and unreported study [21], from which however **no results** could be obtained.

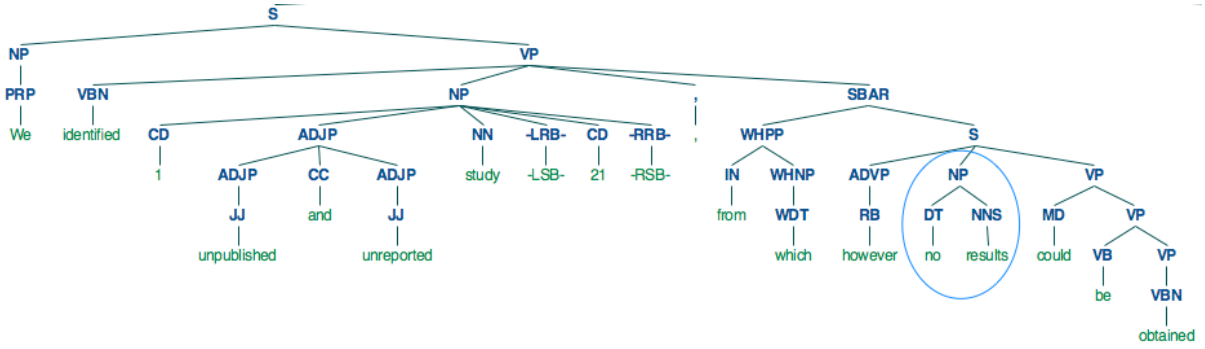

Figure 4.6: The “no” type of negation

Surprisingly, we did **not observe** any previously described polymorphisms by Ober [22] and Matte [23].

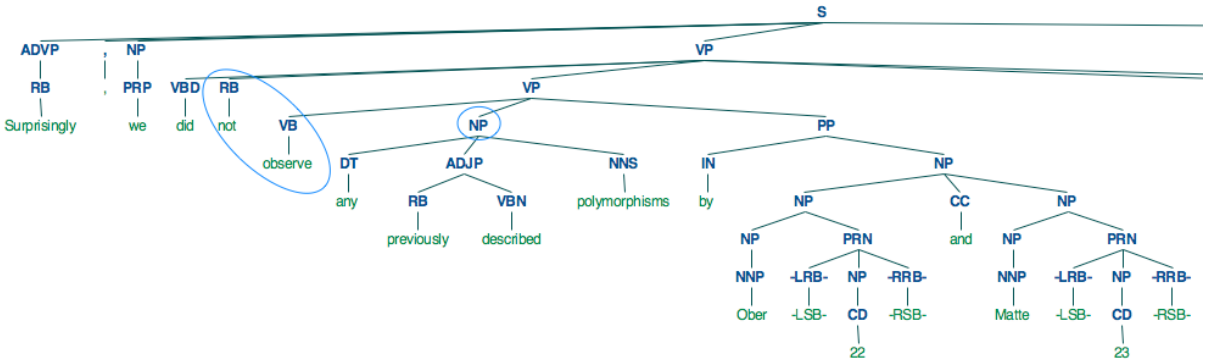

Figure 4.7: The “not” type of negation

The label for the word “no” is always DT, and the following noun is labeled as NNS in most cases and has the same level as “no”. Similarly, the word “not” is always labeled as RB and is followed by a verb in its original form. The extraction of the “not” type starts from the RB level and ends before the next-level NP, during which all words are extracted to make up a cue phrase.

As previously described, each round bracket in the sentence parse is replaced with a number. This mechanism was programmed to extract cue words or phrases by recognizing different syntactic levels through these bracket labels. As a result, hundreds of potential cue phrases were extracted for each paper section. However, there are also many irrelevant word phrases that exist in the extracted data set since this mechanism is not fine-grained enough and does

not have as many parsing rules as those from previous studies. In addition, Figure 4.7 shows an overlap between two extraction types, that is, "did not observe" extracted by the NP-VP type and "not observe" extracted by the negation type have the overlapped phrase "not observe", which may indicate it is a real cue phrase for the current citation. The auxiliary verbs like "did" from the previous example and other meaningless words need to be pruned to keep a succinct format of cue phrases. In summary, this mechanism does help in extracting cue phrases from citation sentences, but identifying the real cue phrases that reflect relationships as well as pruning unnecessary words are still needed to be done manually, which are limitations of this mechanism.

4.4 Corpus Annotation

Citation content is difficult to annotate, as mentioned in Teufel et al. (2006b) [40], because it requires annotators to interpret authors' intentions and sentiments. Moreover, authors of citing articles do not state their purposes and do not show their attitudes towards cited content explicitly in most cases. This phenomenon is especially common in the biomedical context since authors are not willing to express their true opinions when the experiments and data from previous research are non-reproducible. To improve the accuracy and consistency of annotation, I took the pre-annotation step from Hernández-Alvarez et al. (2017) [22], in which the keywords and cue phrases are firstly identified and highlighted in citation sentences, as shown in the following example:

> In the present study there was **no association** between age and gender and outcome **while** some studies have **reported** that older horses have a higher risk of non-survival [9, 13].

The cue phrases are recognized by the extraction method described previously, and collected as a rich feature set for the automatic classifiers that will be introduced in the next chapter. Similar to Garzone (1997) [20] and Agarwal et al. (2010) [2], each function category is assessed based on the cue phrases and can be assigned to a citation sentence directly, except for the categories from the neutral class. This is because the citation sentences that belong to positive or negative classes always have the same function and polarity as the cue phrases inside.

Moreover, I employ a labeling system for dealing with sentences that have multiple cue phrases, and to make the annotation machine-readable. It is shown below:

- Neutral: Perfunctory/Background 1, Statement 2, Comparison 3, Multi-comparison 4

- Positive: Confirmation 5, Being-confirmed 6
- Negative: Contrast/Conflict 7, Unsolved 8

Each cue phrase is labeled with a number that represents the related category. The larger the number is, the more priority the category has. The cue phrases that belong to the neutral class especially the *Perfunctory* category take the majority of the corpus but are less meaningful, thus they are given less weight and represented by smaller numbers. The positive class demonstrates sentimental information about cited content, so the categories in this class are given larger numbers than those of the neutral class. The cue phrases related to the citing paper author's criticism or disputation are the least found in the corpus yet they are the most important. Therefore, the categories from the negative class are given much weight and are represented by the largest numbers. Take the following sentence as an example:

> One polymorphism is **reported by** Hviid [8] at 714insG in the portion of the studied sequence of the untranslated regulatory and full length of exon 1 regions of the HLA-G gene, **however**, we did **not observe** any **previously described** polymorphisms by Ober [22] and Matte [23].

Four cue phrases are identified, and their labels are 1, 7, 7, 1 respectively. The larger the number is, the more importance the cue phrase has. As a result, this citation sentence is assigned as *Contrast/Conflict* category. In addition, the label 7 will be written to the corpus and put at the very beginning of this citation sentence.

In some cases, assigning a category from the neutral class needs more examination of other aspects of the citation content. Accordingly, I developed a strategy that contains several criteria, which are adapted from the binary questions of the annotation decision tree in Xu et al. (2015) [44], to help identify the characteristics of a citation sentence. The criteria are given below:

- The citation sentence refers to both the citing and the cited work or only the cited work is mentioned. This determines the citing direction of the citation sentence and helps distinguish between the *statement* category and the (*background, comparison, multi-comparison*) categories.
- The citation sentence mentions the cited work only as a general reference or compares it with own work. This helps distinguish between the *background* category and the (*comparison, multi-comparison*) categories.
- The comparison is only for the citing and cited works or among several cited works. This helps distinguish between the *comparison* and the *multi-comparison* category.

With the guidelines described above, I annotated all of the citation sentences extracted from different article sections. The distribution of categories and other properties of the annotated corpus are given in the following section.

4.5 Corpus Statistics

The biomedical corpus developed for this thesis is composed of 4950 citation sentences from 640 articles. Since this thesis focuses on the importance of cue phrases in citation classification, the citation sentences that do not contain cue phrases or do not indicate any relation information between citing and cited works are discarded, thus the final number of selected citation sentences is 1823. After annotation, the corpus is heavily skewed with about 69.9% of the citations being neutral (34.1% being *Perfunctory*) and only 30.1% carrying positive or negative sentiment. This result is in line with those from Spiegel-Rösing (1977), Teufel et al. (2006a) and Athar (2011), in which neutral and perfunctory citations take a large portion of the datasets. Table 4.1 shows the distribution of each citation function category within its related polarity class.

Neutral	Perfunctory/Background Statement Comparison Multi-comparison	34.1% 30.5% 4.9% 0.4%	69.9%
Positive	Confirmation Being-confirmed	21.5% 1.4%	22.9%
Negative	Contrast/Conflict Unsolved	6.7% 0.5%	7.2%

Table 4.1: Distribution of citation function categories in polarity classes

The total number of cue phrases extracted from IMRaD is 598 (see Appendix A). Table 4.2 shows the distribution of these cue phrases in each article section. The Results section occupies the majority of the cue phrases and has overlapped parts with those from other sections.

section	Introduction	Method	Results	Discussion/Conclusion
num. of cue phrases	141	53	291	113
%	23.6	8.9	48.7	18.8

Table 4.2: Distribution of cue phrases in article sections

Assigning a citation function and polarity category to a citation sentences is a subjective task and needs to be consistent. Therefore, many previous studies applied Cohen's Kappa

coefficient ($Kappa$) [12] to measure the inter-annotator agreement, which is defined as follows:

$$Kappa = \frac{P(A) - P(E)}{1 - P(E)} \tag{4.1}$$

where $P(A)$ is the relative observed agreement among annotators and $P(E)$ is the hypothetical probability of chance agreement, using observed data to calculate the probabilities of each annotator randomly seeing each category. Since I am the only annotator for this biomedical corpus, I applied the *intra-annotator agreement* method from Athar's research [4], which uses the same formula of Cohen's Kappa but is for a single annotator. Following Athar's description, I firstly annotated the whole corpus, and after one month I re-annotated 800 citation sentences randomly picked from the same corpus without remembering the original annotations. The agreement between the two annotation results on citation function and polarity is $Kappa = 0.71$, which falls in the *substantial agreement* range and indicates the annotated corpus is reliable, according to the interpretation of Kappa values given by Landis and Koch (1977) [26].

4.6 Summary

In this chapter, I firstly describe the corpus construction and the processing procedures performed on the raw data. PubMed was chosen as the data source since it is one of the biggest digital biomedical publication databases that provide complete texts and link referenced papers together to provide a more related source. The corpus I constructed contains 1823 meaningful citation sentences selected from 640 biomedical research papers.

The importance of cue phrases in citation classification is discussed and highlighted with the evidence from previous citation content analysis studies. In addition, I present a cue phrase extraction mechanism that integrates linguistic rules to capture the rhetorical relationships between citing and cited works.

In the following step, I manually annotated my biomedical corpus with the classification scheme described in the previous chapter. To improve the accuracy and consistency of this corpus annotation, I implemented a pre-annotation step which identifies the keywords and cue phrases in sentences by using the cue phrase extraction mechanism mentioned earlier which was used to help decide a particular function for a citation. Furthermore, a strategy of binary questions and a labeling system are proposed to distinguish categories of the neutral class and to generate machine-readable category labels, respectively. The annotated corpus is skewed with 69.9% of the citations being objective and only 30.1% carrying positive or negative sentiment. The distribution of each classification category and the existence of cue phrases in IMRaD are

presented.

The measured Kappa value for intra-annotation agreement on citation function and polarity category assignment is $Kappa = 0.71$, which is found to be substantial according to Landis and Koch's (1977) interpretation of Kappa values. This measurement proves the annotated corpus is stable and reliable, moreover, it is ready for use in the automatic classification experiments that I will describe in the next chapter.

Chapter 5

The Automatic Classification of Citation Function and Polarity

This chapter describes an automatic approach that combines machine learning techniques to classify citation sentences into different citation function and polarity categories. In general, this automatic approach consists of the following procedures: firstly, an annotated corpus is split into training and testing sets along with their related category labels (also known as the gold standard); then different types of features that describe semantic or syntactic characteristics of sentences are collected as inputs; a statistical model will be trained on the features from the training dataset, and return new observations for the testing dataset based on its sentence features; finally the predicted labels are compared with the gold standard to evaluate classification performance of this trained statistical model. This is known as the supervised learning method in the machine learning field, and was implemented in my experiments serving for the automatic classification task on biomedical texts.

This chapter is structured similarly to the supervised learning process mentioned above. A biomedical corpus has been annotated in a previous chapter, thus section 5.1 describes other preparations made on this corpus before the feature extraction procedure, such as stop-word removal and citation index replacement. Section 5.2 demonstrates a series of features extracted from citation sentences with Natural Language Toolkit (NLTK) [28], including lexical features and sentence structure features. Two different machine learning algorithms, which are Support Vector Machine (SVM) [13] and Maximum Entropy [6], are applied to build statistical models for comparison. Section 5.3 gives a brief introduction of these two algorithms and Section 5.4 describes the experimental setup. The automatic classification results and evaluation of the two classifiers are presented and discussed in Section 5.5.

5.1 Data Preprocessing

The goal of data preprocessing is to split the corpus, clean the noise in citation texts and prepare the data for feature extraction. Therefore, as the first step, the annotated biomedical corpus that contains 1823 citation sentences is divided 80/20 using the prevailing rule in machine learning, that is, 80% of sentences compose the training dataset for detecting features and tuning the statistical models, and the remaining 20% of the corpus make up the test dataset used for evaluating the feature-based classification. The citation sentences were randomly allocated to the two sets according to the index of each sentence in the corpus.

Citation texts are different from other normal texts not only in content but also in structure. Generally, there is at least one reference mentioned in citation text, and sentences are composed of more clauses in citation text than in normal text. The references are usually written in two formats: one is the reference indices that point to the related cited articles in the Bibliography surrounded by square brackets; the other is author names followed by publication year written in parentheses. These special formats of references are useless for sentence structure analysis and may even cause errors during part-of-speech tagging. Moreover, the references that contain author names are also a component of citation texts. When a word vector transformer is applied on the tokenized text, the occurrence of author names will also be counted and merged into word vectors. The more the paper is cited by others, the more frequently the related author name will appear in citation texts, which results in more weights in word vectors. As a consequence, such author names bring lexical bias to the classifier and may lead to a wrong prediction for a new citation sentence.

To prevent potential tagging errors and reduce lexical bias, I use two types of placeholders to deal with different reference markers. For citation indices in square brackets, such as "[4]", "[8, 9]" and "[3-5]", the reference markers are replaced with the tokens "[SC]" and "[GC]", which represent single reference and group references respectively. This replacement is inspired by Xu et al. (2015) [44]. For example, the citation sentence shown below

> To detect F-ATPase activity we used its specific inhibitors, azide or oligomycin [19], and ouabain or vanadate were used as specific inhibitors of the Na+/K+-ATPase [20, 21].

will change to

> To detect F-ATPase activity we used its specific inhibitors, azide or oligomycin [SC], and ouabain or vanadate were used as specific inhibitors of the Na+/K+-ATPase [GC].

Similarly, the references represented in a single author name or a combination of author name and parenthesized publication year, such as "Xxx et al." or "Yyy (2006)", are identified by regular expressions then replaced with a single token "[CIT]". This representation is inspired by Athar (2011) [3]. The citation sentence in the following example

> This value appeared lower than that reported by **Skabkin et al.** [22] (about 4nm for monomeric YB-1 in solution of high ionic strength) and probably resulted from YB-1 flattening on the mica surface.

will change to

> This value appeared lower than that reported by **[CIT]** [SC] (about 4nm for monomeric YB-1 in solution of high ionic strength) and probably resulted from YB-1 flattening on the mica surface.

Some words in texts occur very frequently but have no contribution to sentence context or content, so they are referred to as "stop-words". Removing such stop-words can reduce the size of datasets and word vector dimensions, thus improving the classifier's performance. To fit the biomedical content in this corpus, I merged two biomedical-specific stop-word lists, which are obtained from a GitHub page [21] and a PubMed help web page [36], into a machine-readable dictionary for the stop-word removal program.

However, not all stop-words from the two lists are included in my dictionary. The verbs such as "show", "use", "obtain" and the adverbs such as "significantly", "therefore", "especially" are excluded since they indicate citation functions and are cue words or components of cue phrases in my extracted cue lexicon. Therefore, only the tensed auxiliary verbs and some abbreviations are reserved for my stop-word dictionary. In addition, a binary option consisting of "true" or "false" choices is employed on top of the dictionary, thus the stop-word removal process exists only in some types of feature extraction while not in others.

5.2 Features

Considering the characteristics of an automatic framework for citation classification, I extract several types of features from citation sentences at both the lexical and the sentence structural level, including part-of-speech tags, word n-grams, dependency information and other content specific features. The collected feature sets are later converted into feature vectors, and serve as numerical inputs for training and testing statistical classifiers integrated with machine learning algorithms, for which I will give details in the next section.

5.2.1 Part-of-Speech Tags

Part-of-Speech (POS) tagging is widely applied not only in citation function classification but also in sentiment analysis, and has proved to be particularly useful for associating discrete terms. The POS tag reflects a related token's functional role inside a sentence since it encodes a semantic relationship information between the target token and other tokens. For example, a word usually has various meanings and may serve as different roles under different contexts, as shown below:

1. Most of the gallery's contents were damaged in the fire.
2. The boss is content with his employee's work.

The word "content" in the first sentence is a noun and has a "NNS" POS tag, while it functions as an adjective and will be tagged as "JJ" in the second sentence. Moreover, the adjective "content" has higher probability to be associated with the verb "be" and the preposition "with" in most cases. With POS tags as features, a statistical classifier will be able to learn such fixed linguistic patterns, and recognize similar patterns in previously unseen data.

The POS tag features have been extensively applied and examined in many state-of-the-art citation classification frameworks. Teufel et al. (2006b) [40] implemented a POS-based recognizer of agents to help a classifier detect the grammatical subject of a citation sentence according to POS patterns. Athar (2011) [3] manually attached the POS tag to its related word by a delimiter, and applied this method to each single word in the sentence in order to distinguish homonyms and signal the presence of adjectives. Jochim and Schütze (2012) [24] used POS tags to identify 1st and 3rd person pronouns as well as comparatives and superlatives in their feature extraction process. Xu et al. (2015) [44] built their own POS tagging system, which is rule-based and fits the characteristics of citation sentences, to construct structural feature sets.

Most present POS taggers are designed for natural language processing publications and trained on the Wall Street Journal (WSJ) Corpus or the Brown Corpus, which consist of general literature that have a very different structure from scientific citation texts. Therefore, I have used the GENIA tagger [41], which is specially tuned for biomedical documents, as my processing tool. The tagged outputs not only contain POS tags, but also have word base forms, chunk tags and named entity tags, as shown in Figure 5.1. To make the tagged sentences more readable, I extract only original words and their POS tags from the output, and connect each word with its related POS tag with an underscore. For example, the following citation sentence

```
word1   base1   POStag1 chunktag1 NEtag1
word2   base2   POStag2 chunktag2 NEtag2
  :       :       :       :        :
```

Figure 5.1: The output format of the GENIA tagger

> Detailed information on changes in leukocyte counts in milk, lymph and blood is reported by [CIT] [SC].

will be tagged as

> Detailed_JJ information_NN on_IN changes_NNS in_IN leukocyte_NN counts_NNS in_IN milk_NN ,_, lymph_NN and_CC blood_NN is_VBZ reported_VBN by_IN [_(SC_NN]_) [_(CIT_NN]_) ._.

In this way, each word-tag pair becomes a new token in the citation sentence while the length of the sentence does not change and the syntactic structure remains as before.

5.2.2 N-grams

N-gram features have been applied extensively in previous text classification studies and were proved to be useful for classifiers especially SVM [18]. Abu-Jbara et al. (2013) [1] implemented the first bigram and trigram, such as “This approach” and “One problem with”, extracted from the beginning of sentences as features. Different length n-grams were also used by Athar (2011), and the classification results indicate the combined features of unigram, bigram and trigram have better performance than only unigrams and unigrams plus bigrams.

An n-gram is a contiguous sequence of n items from a given text. When n, which refers to the size of sequence, is equal to 1, this sequence is called “unigram”; size 2 and 3 are “bigram” and “trigram”, respectively. Larger sizes of sequences also exist and their names are referred to by the value of n, such as “4-gram” and “5-gram”. Take the following sentence as an example:

> The dog jumps over the fence.

The n-grams of this sentences are:

> **unigrams** The, dog, jumps, over, the, fence, .
>
> **bigrams** The dog, dog jumps, jumps over, over the, the fence, fence .
>
> **trigrams** The dog jumps, dog jumps over, jumps over the, over the fence, the fence .

4-grams The dog jumps over, dog jumps over the, jumps over the fence, over the fence .

5-gram The dog jumps over the, dog jumps over the fence, jumps over the fence .

Suppose X is number of tokens in a given sentence S, then the number of n-grams with size N for sentence S will be:

$$n\text{-}grams_S = X - (N - 1) \tag{5.1}$$

Although n-grams provide abundant information about texts, taking a large n results in a huge data size as calculated by the formula 5.1, and causes data sparsity since a large portion of n-grams appear only once. Therefore, various experiments have been conducted in previous works to find a proper n. Athar (2011) [3] and Bertin et al. (2016) [7] claim that $n = 3$ gives the best classification results, Fürnkranz (1998) [18] also demonstrated $n > 3$ is not useful and may even decrease the performance. In addition, many scientific terms exist in citation sentences, and most of them are 3-words long. Based on these findings, I limit n to 3, that is, I construct unigrams, bigrams and trigrams for each citation sentence. Moreover, I take only the top 300 n-grams as features according to Cavnar and Trenkle (1994)'s [9] observation on frequency of n-grams in Zipf's law distribution.

I also combine POS tags and n-grams to generate POS n-gram features. This may provide classifiers with more information about lexical patterns within sentences. For example, the POS trigrams of the above sentence would be <DT NN VBZ>, <NN VBZ IN>, <VBZ IN DT>, <IN DT NN> and <DT NN .>. The <VBZ IN DT> and <IN DT NN> appear more frequently in the corpus than other POS trigrams due to linguistic rules, which could be learned by the classifier through training.

5.2.3 Dependency Relations

The cue phrase extraction method described previously in Chapter 4 is based on phrase structure constituency relations. The tree forms shown in the examples indicate constituency is a one-to-one-or-more correspondence, which means each element in the sentence is mapped to one or more nodes in the structure.

As opposed to constituency, dependency is a one-to-one correspondence, that is, for each item in a sentence, there is only one node in the sentence structure corresponding to that item. For this reason, the dependency relation is a binary asymmetrical relation between the words of a sentence and it helps identify the semantically related concepts. The two words that form a dependency relation are called *governor* (the head word) and *dependent*, respectively. Similar

to the constituency structure tree form, the dependency structure of a sentence can also be represented as a syntactic tree or a graph consisting of edges and nodes. For instance, the sentence from the previous section

> The dog jumps over the fence.

can be represented as the dependency graph shown in Figure 5.2:

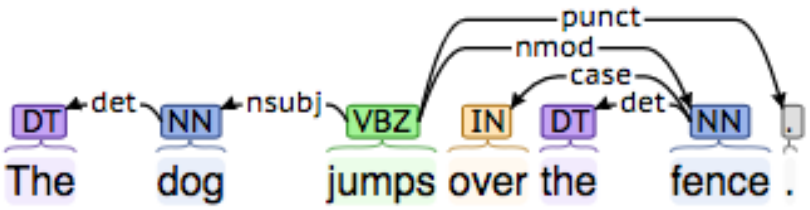

Figure 5.2: The dependency graph for an example sentence

There is a single designated root node "jumps" with the relation `root` that has no incoming arcs. With the exception of the root node, each dependency relation is an arc pointed from the *governer* to the *dependent*, in which the arc label indicates the relation name. The dependency relations of the above sentence could be summarized as follows:

1. `det (dog, The)`
2. `nsubj (jumps, dog)`
3. `nmod (jumps, fence)`
4. `case (fence, over)`
5. `det (fence, the)`
6. `punct (jumps, .)`
7. `root (ROOT, jumps)`

Compared to the part-of-speech tags and n-grams features, the dependency relation concept is a relatively new feature type in citation classification studies but has proved to be quite effective for classifiers making a prediction on new observations. Athar (2011) [3] used dependency structures to capture the long distance relationships between words, and the classification results showed an outstanding improvement over the baseline classifier. Instead of extracting the whole dependency relation, Jochim and Schütze (2012) [24] included only the dependency root node as a component of the word-level feature set.

As mentioned previously, the biomedical citation sentences prefer using clauses and long phrases to demonstrate a finding or concept. Such complicated structures are difficult for classifiers to recognize the semantically related but remote entities, which are the cited content, the linguistic cues and the cited paper. An example is given below:

> **Regulation** of the proteasome complex by phosphorylation of the α7-subunit in COS-7 cells has been **demonstrated** by **[CIT]** [SC].

With dependency parsing, such distant semantic relations could be represented by the dependency relations (see Figure 5.3). This method is in line with those from Athar (2011) and his technical report [4]. To fit the page width, I replaced the long noun phrase of the word "Regulation" with a token "PHRASE" in order to display the dependency structure of the whole sentence here.

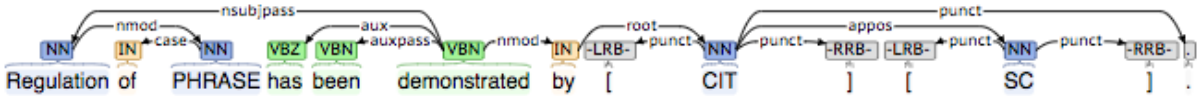

Figure 5.3: The dependency graph for a citation sentence

As shown in the above graph, a dependency relation `nsubjpass` is established between the cited content "regulation" and the cue phrase "demonstrated" even if there is a long distance separating them. I implemented the dependency parsing for all of the citation sentences with spaCy [23], which is a Python library that includes a dependency parser. I chose this toolkit since there is no biomedical-specific trained dependency parser available at the moment, and it proved to be faster and more accurate than the Standford dependency parser [11] if combined with an appropriate POS tagger. Similar to the concatenation of POS tags and tokens, I convert each dependency relation to the format `relation_governor_dependent`. In this way, the above "regulation" example can be represented as the `nsubjpass_demonstrated_regulation` feature.

5.2.4 Lexicon of Linguistic Cues

In previous sentiment detection tasks, using a labeled lexicon to score sentences is a common approach. Athar (2011) includes a science-specific sentiment lexicon that consists of 83 polarity phrases extracted from a development dataset, and combines the information from this lexicon with other features by applying the binary labeling approach from Wilson et al. (2009) [43]. Jochim and Schütze (2012) extracted the citation context words that appear in a polarity lexicon as features, which are later represented as a bag of words and divided into two polarity

categories. Xu et al. (2015) tagged sentiment tokens and generated n-grams from them as a category of structure features. Though the authors from Athar and Teufel (2012) argue that such an approach is highly topic-related thus not useful to build a generalized classifier, this argument implies a detailed lexicon for a specialized domain might provide rich information in classification tasks.

Therefore, I took the cue phrase set extracted during the corpus pre-annotation step that is described in Chapter 3 as my lexicon (see Appendix A). Since each cue phrase has already been manually assigned a citation function and polarity category, this lexicon not only contains sentiment information but also indicates citation relationships. Then I developed a set of tags based on my proposed classification scheme to encode this lexicon with other features. The tags in this set are PERF (Perfunctory), STMT (Statement), COMP (Comparison), MULCOMP (Multi-comparison), CONF (Confirmation), BECONF (Being-confirmed), CTRST (Contrast) and NSLVD (Unsolved). Once a cue phrase that exists in my lexicon is identified within a sentence, each word inside the cue phrase will be prefixed by the tag representing the category to which that cue phrase belongs. In addition, I pruned long cue phrases and limited the length of each cue phrase to 3 words in order to match n-gram features on cue phrase detection. For instance, the example sentence discussed in the corpus annotation chapter will be converted to:

> One polymorphism is **PERF_reported PERF_by** [CIT] [SC] at 714insG in the portion of the studied sequence of the untranslated regulatory and full length of exon 1 regions of the HLA-G gene, **CTRST_however**, we did **CTRST_not CTRST_observe** any **PERF_previously PERF_described** polymorphisms by [CIT] [SC] and [CIT] [SC].

Furthermore, I calculate polarity score as one feature for each citation sentence. The cue phrase that belongs to the positive categories has 1 point, the neutral phrase has 0 points, and the negative phrase has -1 point. In the above sentence, two identified cue phrases are from the neural categories while the other two belong to the negative class, thus the total polarity score is 0 + (-1) + (-1) + 0 = -2 which is later added to the feature vector.

5.2.5 Other Features

Besides the general features described previously, there are some other types of features that capture the characteristics of certain parts of citation sentences, or designed specifically for biomedical texts. These are now described.

Location information

This feature describes the location information of a citation sentence from two aspects. In Teufel et al. (2006b) and Dong and Schäfer (2011) [17], the authors identified the section of the paper where the citation sentence is located. I follow this method and create a denotation `section_LABEL`, in which the label could be INTRO, MTHD, RSLT, or DISS referring to the IMRaD article structure. In addition, I include the estimated position of a citation found in a sentence, which is in line with Jochim and Schütze (2012), and encode this information as `sentence_pos` where `pos` could be BEG (beginning), MID (middle) or END (end). Suppose a citation sentence extracted from the Introduction section has two citations, which are located in the beginning and middle of the sentence, respectively. The location feature of this citation sentence would be `(section_INTRO, sentence_BEG, sentence_MID)`.

Word negation

This feature is partially overlapped with my negational cue phrases. However, there were only two negation types, which are "no" and "not", applied during the extraction process in order to fit the syntactic structure of sentences. As a lexical feature, I use a negation scope and more negational words to capture the surface characteristics of citation sentences.

The implementation of negation features was initially proposed by Das and Chen (2001) [14] and later extended by Pang et al. (2002) [35]. I follow the window-based approach described in the latter paper and set the window size to 2. The tokens that come after a negational word and fall in the 2-word negation scope within a citation sentence would be prefixed with a label `NOT_`. In this way, the negation feature is integrated into the sentence without changing the sentence length, as shown below:

> We have already demonstrated that these regions are not **NOT_essential NOT_for** the catalytic activity [SC], which retain their catalytic activity upon modification.

The word phrase "essential for" is a positive expression, however, the negation "not" causes the meaning of a citation sentence to be the opposite of that word phrase. Other negational words used for this feature are *no, nothing, without, none, neither, never, n't*, which are partially taken from the negation list in Athar's technical report.

Citation count

This feature describes the number of citations that appear in a sentence, thus it is numerical. It may highlight a comparison or contrast category since the citation sentences that belong to such categories usually have two or more cited papers compared together. The count of citations are

simply calculated by the occurrence of placeholders `[CIT]`, `[SC]` and `[GC]`. It is noted that if a `[CIT]` is closely followed by a `[SC]`, then they are counted together as one. This is because the `[SC]` generally refers to the same cited paper as `[CIT]`, which is originally written as *Xxx et al.* with or without published year.

1st/3rd-person pronouns

This feature describes whether a citation sentence contains 1st and 3rd person pronouns by using the Boolean values *True* and *False*. Since the opinion target is the cited paper instead of a scientific claim, the person pronouns help distinguish between a general statement and a confirmation. For instance, the following citation sentence

> **They** confirm similar results obtained by [CIT] using the affinity cleavage method for a different TFO sequence [SC].

belongs to the *background* category rather than the *confirmation* category. Though the cue words "confirm" and "similar" express strong positive polarity, the 3rd-person pronoun "they" indicates the subject of confirmation is actually from another cited article, not own work. Therefore, it was the citing work stating a fact that the results from a cited work was confirmed by the authors of another cited article, which could be seen as background acknowledgment.

I use *I, we, our, ours, we, us* for 1st-person pronouns and *she, he, it, his, hers, him, her, they, them, their* for 3rd-person pronouns. The denotation for this feature is a tuple of values written as `(BOOL, BOOL)`, where the two `BOOL` represent whether 1st and 3rd person pronouns are detected in a citation sentence. Thus, the above sentence would have the feature `(F, T)`.

5.3 Classifiers

Since there are 8 categories in my citation function and polarity scheme, I chose Maximum Entropy and Support Vector Machine (SVM) algorithms to deal with this multi-class classification task. Although the philosophies behind these two mechanisms are quite different, each has proved to be effective in previous text categorization studies.

5.3.1 Maximum Entropy

Maximum entropy is a probability distribution estimation technique (Berger et al., 1996 [6]) widely applied in various natural language processing tasks. It could be seen as a generalization of logistic regression for multi-class problems.

Suppose f_i is a feature from the feature vector $\{f_1, f_2, f_3, \ldots, f_k\}$ that contains k features extracted from document d. The estimation of $P(c \mid d)$, which represents the probability that class c appears in document d, takes the following exponential expression:

$$P(c \mid d) = \frac{1}{Z(d)} \exp\left(\sum_i \lambda_i f_i(d, c)\right) \tag{5.2}$$

where λ_i is a weight to be estimated. The larger that λ is, the more information the feature contains. $Z(d)$ is a normalization function to ensure a proper probability distribution, and has the following form:

$$Z(d) = \sum_c \exp\left(\sum_i \lambda_i f_i(d, c)\right) \tag{5.3}$$

For text classification with maximum entropy, the scaled word counts are used as features. Thus, each word-class combination is instantiated by a feature defined as:

$$f_{w,c'}(d, c) = \begin{cases} 0, & \text{if } c \neq c' \\ \frac{n(d,w)}{n(d)}, & \text{otherwise} \end{cases} \tag{5.4}$$

where $n(d, w)$ is the occurrence of word w in document d, and $n(d)$ is the number of words in the document. In most maximum entropy classification tasks, the features are naturally binary features representing real-valued functions of document d and class c, so $f_i(d, c)$ from Equation 5.2 is a binary feature that could be defined as follows:

$$f_i(d, c') = \begin{cases} 1, & \text{if } c' = c \text{ and } n_i(d) > 0 \\ 0, & \text{otherwise} \end{cases} \tag{5.5}$$

An advantage of maximum entropy is that it does not make independence assumptions about the features. For instance, the two words in phrase "Monte Carlo" rarely appear by themselves, thus the maximum entropy algorithm would reduce the weights λ of theses two terms by half during classification, while the algorithms that contain a prior, such as Naïve Bayes, might double the count of this phrase. Moreover, maximum entropy could achieve better classification performance than Naïve Bayes when conditional independence assumptions are not met.

I used the maximum entropy package from NLTK for my implementation. Since the typical iterative scaling techniques are exhaustive and caused the training process to be very slow, I integrated my maximum entropy classifier with an optimization package named `Megam` [15], which greatly boosted the training and testing speed.

5.3.2 Support Vector Machine

Support Vector Machines (SVMs) has been shown to achieve good performance in a variety of state-of-the-art citation classification studies. This is because the SVM is able to select useful features effectively from a huge number of features, which could potentially be an issue in many NLP tasks, for a specific classification problem by assigning different weights.

In general, an NLP problem can be treated as a multi-class classification problem, which is later transformed into multiple binary classification cases. As a binary classier, the intuition behind an SVM is to find a proper decision boundary that maximizes the distance to the nearest sample from either the positive or negative class. Such kinds of decision boundaries are represented by the SVM discriminant function that has the following form (in 2 dimensions):

$$f(\vec{x}) = \vec{w}^T \vec{x} + b \tag{5.6}$$

where w is the weight vector and b is the bias. The learning function is $\text{sign}(f(\vec{x}))$, and the linear decision boundary is specified by $f(\vec{x}) = 0$. The correct class c_j of document d_j, is defined as y_i in this binary case, and $y_i \in \{-1, 1\}$.

Given a training dataset $\{(\vec{x}_i, y_i)_{1 \le i \le n}\}$, we need to find a decision boundary $\vec{w}^T, b$ that maximizes the Euclidean distance of the closest points, expressed as below:

$$\max_{\vec{w},b} \min_{i=1}^{n} \frac{y_i \left(\vec{w}^T \vec{x}_i + b\right)}{\|\vec{w}\|} \tag{5.7}$$

Since it is difficult to optimize the above object directly, this problem is converted to:

$$\begin{aligned} \underset{\vec{w},b}{\text{minimize}} \quad & \frac{1}{2}\|\vec{w}\|^2 \\ \text{subject to} \quad & y_i \left(\vec{w}^T \vec{x}_i + b\right) \ge 1 \qquad \forall i \end{aligned} \tag{5.8}$$

Now the problem 5.8 has become a *quadratic* problem, where the object is a quadratic function subject to linear constraints. By applying Lagrange multipliers α, this quadratic problem could be transformed to its dual formulation, which is to be maximized with the following constraints:

$$\begin{aligned} \text{maximize:} \quad & L(\alpha) = -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \vec{x}_i^T \vec{x}_j + \sum_{i=1}^{n} \alpha_i \\ \text{subject to} \quad & \alpha_i \ge 0, \ \forall i \\ & \sum_{i=1}^{n} \alpha_i y_i = 0 \end{aligned} \tag{5.9}$$

of which the solution is :

$$\begin{aligned} \vec{w} &= \sum_{i=1}^{n} \alpha_i y_i \vec{x}_i \\ b &= y_k - \vec{w}^T \vec{x}_k \quad \forall \vec{x}_k; \alpha_k \neq 0 \end{aligned} \tag{5.10}$$

where most of the α_i are zeros, and each non-zero α_i indicates that the related $\vec{x}_i$ is a *support vector*. Therefore, the learning function could be re-written as:

$$f(\vec{x}; \vec{w}, b) = \text{sign}\left(\vec{w}^T \vec{x} + b\right) = \text{sign}\left(\sum_{i=1}^{n} \alpha_i y_i \vec{x}_i^T \vec{x} + b\right) \tag{5.11}$$

I used the SVM package from the `scikit-learn` library for testing and training, with linear kernels and all parameters set to default values. Moreover, I integrated a one-versus-all (OVA) framework with the SVM classifier for dealing with the multi-class problem. This strategy builds k classifiers F_i (k is the number of classes and $1 \leq i \leq k$), where class i is positive and all other classes are negative for F_i. The maximum value i of $F_i(x)$ at a data point x would be chosen for the class after comparing all classifiers.

5.4 Experimental Setup

In previous citation classification studies, regardless of how well the scheme is defined, one can always observe that the class-imbalance issue existed in the dataset due to the large amount of perfunctory/neutral citations. Different from most works that usually ignored this fact, Athar and Teufel (2012) [5] applied the 4-sentence long context detection during corpus annotation. Though this approach greatly improved the number of subjective citations, detecting context is a non-trivial task and requires much human effort, which is beyond the scope of this thesis.

Dong and Schäfer (2011) [17] claims that decreasing the number of instances in big classes and slightly reducing the intensity of the class imbalance might avoid misclassifying too much data from small classes. I followed their method and condensed my neutral class, especially the *Background* and *Statement* categories. For each of these two categories, I took 30% of the sentences and combined them with the rest of the sentences from same class, that is, two sentences that have the same label are merged into one line. In this way, the class-imbalance problem was alleviated to some degree.

All of the citation sentences integrated with text features in the training and testing datasets were converted into numerical feature vectors with a blackbox transformer from the `scikit-learn` library. The experiments were all conducted with 10-fold cross validation and built as a pipeline. A bag-of-words model, which transforms raw text data into numerical feature vectors

in terms of the occurrence or frequency of each token, was built with unigrams as a baseline system to compare with maximum entropy and SVM classifiers.

Both micro-F and macro-F were applied to evaluate the classification performance. The difference between these two evaluation metrics is that micro-F uses different weights for classes of various sizes to calculate the final F score, while macro-F has a stricter standard that averages F scores over all classes with the same weight. Since my corpus has a skewed class distribution, the averaged macro-F for all classes might be lower than micro-F, but more reflects the real-world problem.

5.5 Results

I combined different types of features described in Section 5.2, and tested them with the maximum entropy and SVM classifiers, respectively. As shown in Table 5.1, the combinations that use POS tags have higher F scores than those without POS tags. With the cue phrase lexicon, both the unigram and unigram + POS tags combinations have an obvious increase on F scores, which proves the previous statement that cue phrases are crucial in citation classification. It should be noted that the cue lexicon could not be added to feature combinations that include bigram and trigram features because the results were negatively affected by a data sparseness problem in the small corpus. The metadata features, which are citation counts, location information and pronouns, do not have a significant impact on classification. Through this comparison of all feature types, we may infer that both structural and lexical features are important for capturing sentence characteristics.

Features	SVM		Maximum Entropy	
	micro-F	macro-F	micro-F	macro-F
unigram (baseline)	0.484	0.42	0.46	0.409
unigram + bigram	0.506	0.47	0.499	0.452
unigram + bigram + trigram	0.66	0.618	0.647	0.593
POS tags + unigram	0.614	0.593	0.601	0.576
POS tags + 1-3 grams	0.713	0.681	0.671	0.62
POS tags + 1-3 grams + dependencies	**0.731**	0.68	0.7	0.64
unigram + cue lexicon	0.573	0.516	0.552	0.501
POS tags + unigram + cue lexicon	0.694	0.613	0.671	0.6
POS tags + unigram + cue lexicon + dependencies	0.723	0.652	0.691	0.63
POS tags + unigram + cue lexicon + other features	0.70	0.62	0.672	0.602

Table 5.1: Results for Different Feature Combinations on Citation Function Classification

Since the SVM classifier with POS tags + 1-3 grams + dependencies features achieved the

best performance, which is 0.731 as micro-F, I take this combination and compute the detailed evaluation for each citation function category.

Category	Precision	Recall	$F1$
Perfunctory/Background	0.687	0.792	0.736
Statement	0.802	0.581	0.674
Comparison	0.557	0.788	0.653
Multi-comparison	0.552	0.431	0.484
Confirmation	0.822	0.638	0.719
Being-confirmed	0.77	0.42	0.54
Contrast/Conflict	0.77	0.52	0.62
Unsolved	0.554	0.463	0.504

Table 5.2: Citation Function Classification with the SVM Classifier

The results given in Table 5.2 are not quite satisfying. This is because my corpus is still heavily skewed even after I condensed the neutral class. The low $F1$ scores of *Multi-comparison* and *Unsolved* are mainly caused by the very small number of samples in these two categories. Furthermore, the sharp contrast between the $F1$ scores of *Confirmation* and *Being-confirmed* also questions whether it is applicable and reasonable to take citing direction as an extra dimension in the classification scheme, as most citations have the same citing direction, which might be a potential factor that gives rise to the imbalance problem.

Class	Precision	Recall	$F1$
Neutral	0.806	0.931	0.838
Positive	0.825	0.630	0.714
Negative	0.959	0.66	0.782

Table 5.3: Citation Polarity Classification with the SVM Classifier

The overall results for the citation polarity classification in Table 5.3 are better than those of the citation function categorization. This performance is in line with Xu et at. (2015), but not as high as Athar (2011) and Hernández-Alvarez et al. (2017). One possible reason might be the citation sentiment detection in biomedical domain is more difficult because of the complicated sentence structures and prevalent hedges applied in biomedical writing.

5.6 Summary

This chapter illustrates the implementation of automatic classification on citation function and polarity. In data preparation, I replace citation indices and references with two placeholders to

avoid part-of-speech tagging errors and lexical bias, respectively. In addition, some stop-words are removed from the corpus to reduce feature vector dimensions.

I extract part-of-speech tags and n-grams as lexical features, and dependency relations as structural features. A lexicon of cue phrases, which are identified during corpus annotation, is included to capture contextual characteristics of sentences. Some metadata features, such as citation location and citation count are also extracted as supplements.

To alleviate the imbalanced corpus issue, I condensed the neutral class. Different types of features are combined and experimented separately with the two classifiers, which are maximum entropy and SVM. The best classification performance was obtained with POS tags + 1-3 grams + dependencies features using the SVM classifier. This combination was later applied for computing the precision, recall and $F1$ scores of each citation function category and sentiment class. The results indicate it is difficult to classify citation function and detect sentence sentiment in the biomedical domain.

Chapter 6

Conclusions and Future Work

This chapter concludes the research work that has been done for this thesis, as well as discusses the possible future improvements in biomedical citation classification.

6.1 Conclusions

In this thesis, I firstly reviewed the literature in citation classification. Various categorization schemes have been closely examined, and their automatic classification experiments combined with machine learning algorithms are also well studied. The research that addressed the important role of cue phrases in citation classification were given special attention.

In terms of the classification problems in previous studies and limited works in biomedical citation research, the following results have been obtained in this work:

- A new citation classification scheme, which investigates citation function and polarity, is proposed for the purpose of automation. This new scheme consists of 3 top-levels of 8 categories, and is defined by 3 dimensions to fit the biomedical citation content.

- A biomedical citation corpus is constructed and annotated. To improve the annotation accuracy, I developed a cue phrase extraction mechanism to identify the linguistic cues that are crucial in deciding a particular category for a citation. The extracted cue phrases compose a biomedical-specific lexicon, which may contribute to the future citation research in the biomedical domain.

- A series of automatic classification experiments are conducted with lexical, structural and metadata features as inputs for two statistical models: maximum entropy and SVM. Good performance was obtained in citation polarity classification, while fair performance was achieved on citation function categorization due to the imbalanced corpus. The fact

that linguistic cue features significantly improved the classification results indicates that cue phrases do play an important role in citation content analysis.

6.2 Future Work

A common problem in present citation classification research is the imbalanced corpus. One possible solution is to take the context sentences into consideration. However, the context length of a biomedical citation varies widely, from an adjacent sentence to the whole paragraph. This is because biomedical authors tend to use more comparative terms and cite much data to hedge their own true opinions. Defining a proper context scope could be one meaningful aspect of future citation classification studies.

In recent years, neural networks have been employed in more and more natural language processing tasks. A multi-layer neural network is a combination of various statistical classifiers, and is able to learn sentence characteristics by itself instead of through hand-crafted features. Such implementations have proved to be effective in some latest citation function and polarity classification studies. Therefore, in my future research works, I plan to apply a neural network system to detect more useful information in perfunctory or neutral citations.

Bibliography

[1] Amjad Abu-Jbara, Jefferson Ezra, and Dragomir Radev. Purpose and polarity of citation: Towards nlp-based bibliometrics. In *NAACL HLT 2013 - 2013 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 596–606. Association for Computational Linguistics (ACL), 2013.

[2] Shashank Agarwal, Lisha Choubey, and Hong Yu. Automatically classifying the role of citations in biomedical articles. In *AMIA Annual Symposium Proceedings*, 1942-597X, pages 11–15, 2010.

[3] Awais Athar. Sentiment analysis of citations using sentence structure-based features. In *Proceedings of the ACL 2011 Student Session*, pages 81–87. Association for Computational Linguistics, 2011.

[4] Awais Athar. Sentiment analysis of scientific citations. Technical Report UCAM-CL-TR-856, University of Cambridge, Computer Laboratory, 2014.

[5] Awais Athar and Simone Teufel. Detection of implicit citations for sentiment detection. In *Proceedings of the Workshop on Detecting Structure in Scholarly Discourse*, pages 18–26. Association for Computational Linguistics, 2012.

[6] Adam L. Berger, Vincent J. Della Pietra, and Stephen A. Della Pietra. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71, 1996.

[7] Marc Bertin, Iana Atanassova, Cassidy R. Sugimoto, and Vincent Lariviere. The linguistic patterns and rhetorical structure of citation context: an approach using n-grams. *Scientometrics*, 109(3):1417–1434, 2016.

[8] Susan Bonzi. Characteristics of a literature as predictors of relatedness between cited and citing works. *Journal of the American Society for Information Science*, 33(4):208–216, 1982.

[9] William B. Cavnar and John M. Trenkle. N-gram-based text categorization. In *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175, 1994.

[10] Eugene Charniak and Mark Johnson. Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 173–180. Association for Computational Linguistics, 2005.

[11] Danqi Chen and Christopher D. Manning. A fast and accurate dependency parser using neural networks. In *Proceedings of EMNLP 2014*, pages 740–750, 2014.

[12] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

[13] C. Cortez and V. Vapnik. Support-vector networks. *Machine Learning*, 20:273–297, 1995.

[14] Sanjiv Das and Mike Chen. Yahoo! for Amazon: Extracting market sentiment from stock message boards. In *Proceedings of the Asia Pacific Finance Association Annual Conference (APFA)*, volume 35, page 43, 2001.

[15] Hal Daumé III. Notes on CG and LM-BFGS optimization of logistic regression. Paper available at `http://pub.hal3.name#daume04cg-bfgs`, implementation available at `http://hal3.name/megam/`, August 2004.

[16] Ying Ding, Guo Zhang, Tamy Chambers, Min Song, Xiaolong Wang, and Chengxiang Zhai. Content-based citation analysis: The next generation of citation analysis. *Journal of the Association for Information Science and Technology*, 65(9):1820–1833, 2014.

[17] Cailing Dong and Ulrich Schäfer. Ensemble-style self-training on citation classification. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 623–631. Asian Federation of Natural Language Processing, 2011.

[18] Johannes Fürnkranz. A study using n-gram features for text categorization. Technical Report OEFAI-TR-98-30, Austrian Research Institute for Artificial Intelligence, 1998.

[19] Eugene Garfield. Can citation indexing be automated? *Symposium Proceedings Washington 1964*, pages 189–192, 1964.

[20] Mark A. Garzone. Automated classification of citations using linguistic semantic grammars. Master's thesis, University of Western Ontario, 1997.

[21] Stopwords from Ovid (medical information services). https://github.com/igorbrigadir/stopwords/blob/master/en/ovid.txt, 2016.

[22] Myriam Hernández-Alvarex, José M. Gomez Soriano, and Patricio Martínez-Barco. Citation function, polarity and influence classification. *Natural Language Engineering*, 23(4):561–588, 2017.

[23] Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017.

[24] Charles Jochim and Hinrich Schütze. Towards a generic and flexible citation classifier based on a faceted classification scheme. In *24th International Conference on Computational Linguistics - Proceedings of COLING 2012: Technical Papers*, pages 1343–1358, 2012.

[25] Alistair Knott. *A Data-Driven Methodology for Motivating a Set of Coherence Relations*. PhD thesis, University of Edinburgh, 1996.

[26] J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.

[27] Ben-Ami Lipetz. Improvement of the selectivity of citation indexes to science literature through inclusion of citation relationship indicators. *Journal of the American Society for Information Science and Technology*, 16(2):81–90, 1965.

[28] Edward Loper and Steven Bird. NLTK: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*, ETMTNLP '02, pages 63–70. Association for Computational Linguistics, 2002.

[29] Daniel Marcu. *The Theory and Practice of Discourse Parsing and Summarization*. MIT Press, Cambridge, MA, USA, 2000.

[30] Siniša Maričić, Jagoda Spaventi, Leo Pavičić, and Greta Pifat-Mrzlijak. Citation context versus the frequency counts of citation histories. *Journal of the American Society for Information Science and Technology*, 49(6):530–540, 1998.

[31] Robert E. Mercer and Chrysanne Di Marco. The importance of fine-grained cue phrases in scientific citations. In *Advances in Artificial Intelligence*, pages 550–556. Springer Berlin Heidelberg, 2003.

[32] Tom M. Mitchell. *Machine Learning*. McGraw-Hill, 1997.

[33] Michael J. Moravcsik and Poovanalingam Murugesan. Some results on the function and quality of citations. *Social Studies of Science*, 5(1):86–92, 1975.

[34] Hidetsugu Nanba, Noriko Kando, and Manabu Okumura. Classification of research papers using citation links and citation types: Towards automatic review article generation. *Advances in Classification Research Online*, 11(1), 2000.

[35] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of EMNLP*, pages 79–86, 2002.

[36] PubMed Help [Internet]. Bethesda (MD): National Center for Biotechnology Information (US). `https://www.ncbi.nlm.nih.gov/books/NBK3827/table/pubmedhelp.T.stopwords/`, 2005.

[37] Ina Spiegel-Rösing. Science studies: Bibliometric and content analysis. *Social Studies of Science*, 7(1):97–113, 1977.

[38] Frederick Suppe. The structure of a scientific paper. *Philosophy of Science*, 65(3):381–405, 1998.

[39] Simone Teufel, Advaith Siddharthan, and Dan Tidhar. An annotation scheme for citation function. In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, SigDIAL '06, pages 80–87, 2006.

[40] Simone Teufel, Advaith Siddharthan, and Dan Tidhar. Automatic classification of citation function. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pages 103–110, 2006.

[41] Yoshimasa Tsuruoka, Yuka Tateishi, Jin-Dong Kim, Tomoko Ohta, John McNaught, Sophia Ananiadou, and Jun'ichi Tsujii. Developing a robust part-of-speech tagger for biomedical text. In *Advances in Informatics - 10th Panhellenic Conference on Informatics*, LNCS 3746, pages 382–392, 2005.

[42] Henry Voos and Katherine S. Dagaev. Are all citations equal? or, did we op. cit. your idem? *Journal of Academic Librarianship*, 1(6):19–21, 1976.

[43] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational Linguistics*, 35(3):399–433, 2009.

[44] Jun Xu, Yaoyun Zhang, Yonghui Wu, Jingqi Wang, Xiao Dong, and Hua Xu. Citation sentiment analysis in clinical trial papers. In *AMIA Annual Symposium Proceedings 2015*, pages 1334–1341, 2015.

[45] Bei Yu. Automated citation sentiment analysis: What can we learn from biomedical researchers. In *Proceedings of the 76th ASIS&T Annual Meeting: Beyond the Cloud: Rethinking Information Boundaries*, ASIST '13, pages 1–9. American Society for Information Science, 2013.

Appendix A

Cue Phrase Lexicon

A regular cue phrase pattern in BACKGROUND			
position 1 (optional)	position 2 (optional)	position 3	position 4 (optional)
as	previously	calculated	by
been	recently	described	in
also	already	demonstrated	previously
		developed	before
		discussed	earlier
		done	extensively
		found	in
		identified	for
		measured	recently
		noted	to
		observed	
		performed	
		postulated	
		proposed	
		presented	
		published	
		recommended	
		reported	
		shown	
		studied	
		suggested	
		used	

The words on position 1, 2, 3 and 4 can make various combinations thus take majority of cue phrase lexicon. For example, both “as previously described” and “as previously described by” exist in lexicon. The rest of cue phrases of BACKGROUND category are shown in the following table.

BACKGROUND		
according to	analyzed	articulated in
as with	aware of	because of
commonly employed in	considered data from	defined by
demonstrate	determined by	depicted in
developed by	documented	e.g.
employed (with)	encountered by	follow(s/ed/ing)
for example	for review	from a previous study
refer to	has/have been shown	given by
generally considered	(also) known as	less likely
introduced	known (from/to/about)	it is known
investigated by	may explain	in a previous report
in the earlier study of	no...has yet been reported	reflect
see	note(d) (previously)	prepared previously by
reflected by	reproduce	well-founded
well-known	widely known	widely used
such as	shown to (be)	since
(previously) thought to	using	utilize(d/s)
referenced with		

STATEMENT		
assert	claimed	calculate (using)
calculated (as/from)	concluded from	demonstrated (to/in)
detected	established	evaluated (as)
examined	explain	found in/to
followed by	identified (from)	illustrates
implicated	indicate(d/ing/s)	included (in)
including	mapped	previously
previously pointed out	proposed (as)	replicated
reported	revealed	show(ed/n/s)
speculated	suggest(ed/ing/s)	underline

COMPASIRON		
also a result of	analogous to	as follows
comparable to/with	compare	deviate normally from
equivalent to	fits...to some degree	for comparison
found	highly homologous to	high similarity with
in response to	in addition	obtain
obtained (from/by/using)	resemble(s/d) (highly)	similar result(s)
similarly	subjected to	the same level as
very close to		

MULTI-COMPASIRON		
a departure from	besides	furthermore
in addition	moreover	needs to be pointed out
reviewed in	takes a different approach to	in previous studies
state of the art	state-of-art	state-of-the-art

COMFIRMATION		
acceptance	accurately	adequately
appealing	bestperforming	better
agree(s) (well) with	as expected	based on/upon
close to	confirm(ed/ing/s)	consistent with
closely	competitive	considerable
correspond(ing) to	correspond well with	corroborate
correlates	high-quality	important
does match	except	extend(ed)
dominant	dramatically	easier
faster	favorably	high
further investigate	good	in accord(ance) with
improve	improve the performance	improvements
influential	intensively	interesting
in (close) agreement with	in congruence/line with	matched with
may hypothesize	met	not affect
no discrepancies between...and	not surprising	particular useful
not statistically significantly different from	remarkably	similar (to)
offer an explanation for	similarly to	similarity
outperforms	overcome	pioneered
predominantly	preferable	preferred
sought to confirm	support	supported by
used successfully	significantly enhance	substantially
success	successful	successfully

BEING-COMFIRMAED		
aided	as suggested by	effective
efficient	efficiently	excellent
effectiveness	supported by	confirmed by
enhanced by	evidenced by	fortunately
in support of	it is conceivable	on the basis of
same finding	was achieved using	quite accurate
reasonable	reduces overfitting	robust
satisfactory	significant increases	simple

CONTRAST/CONFLICT		
appears different with	burden	but/indicate
complicated	contradictory	daunting
contradicts	contrary to	contrasting with
difficult	differ(s) (significantly) from	difference
different from	discrepancies between	striking difference between
lack	higher	poor
distinct from	however	not discover
it is unlikely	argues strongly	in disagreement with
in (sharp/striking) contrast (to/with)		not appear
in contrast	shown...not	no association
no...were observed	not addressed	not demonstrate
not be relevant to	not find	not have
not described before	not match	not observe
not possible	not show	no results
on the contrary	rule out	surprisingly
unlike	very different	while
yet	restrict	worse

UNSOLVED		
deficiencies	degrade	inability
far from clear	not shown	not clear
remained (unclear)	presently unknown	so far...limited to
still unknown	limitations	no solutions were given
unexplored		

Curriculum Vitae

Name:	Meng Jia
Post-Secondary Education and Degrees:	University of Western Ontario London, ON 2016 – 2018 M.Sc. University of Arizona Tucson, AZ, USA 2014 – 2016 M.Sc. Handan College Handan, Hebei, China 2010 – 2014 B.A.
Honours and Awards:	Western Graduate Research Scholarship 2016-2017
Related Work Experience:	Teaching Assistant University of Western Ontario 2016 – 2018 Research Assistant University of Western Ontario 2016 – 2018