Technical Report

UCAM-CL-TR-856
ISSN 1476-2986

Number 856

UNIVERSITY OF CAMBRIDGE
Computer Laboratory

Sentiment analysis of scientific citations

Awais Athar

June 2014

15 JJ Thomson Avenue
Cambridge CB3 0FD
United Kingdom
phone +44 1223 763500

http://www.cl.cam.ac.uk/

This technical report is based on a dissertation submitted April 2014 by the author for the degree of Doctor of Philosophy to the University of Cambridge, Girton College.

Some figures in this document are best viewed in colour. If you received a black-and-white copy, please consult the online version if necessary.

Technical reports published by the University of Cambridge Computer Laboratory are freely available via the Internet:

http://www.cl.cam.ac.uk/techreports/

ISSN 1476-2986

Summary

While there has been growing interest in the field of sentiment analysis for different text genres in the past few years, relatively less emphasis has been placed on extraction of opinions from scientific literature, more specifically, citations. Citation sentiment detection is an attractive task as it can help researchers in identifying shortcomings and detecting problems in a particular approach, determining the quality of a paper for ranking in citation indexes by including negative citations in the weighting scheme, and recognising issues that have not been addressed as well as possible gaps in current research approaches.

Current approaches assume that the sentiment present in the citation sentence represents the true sentiment of the author towards the cited paper and do not take further informal mentions of the citations elsewhere in the article into account. There have also been no attempts to evaluate citation sentiment on a large corpus.

This dissertation focuses on the detection of sentiment towards the citations in a scientific article. The detection is performed using the textual information from the article. I address three sub-tasks and present new large corpora for each of the tasks.

Firstly, I explore different feature sets for detection of sentiment in explicit citations. For this task, I present a new annotated corpus of more than 8,700 citation sentences which have been labelled as positive, negative or objective towards the cited paper. Experimenting with different feature sets, I show the best result of micro-F score 0.760 is obtained using n-grams of length and dependency relations.

Secondly, I show that the assumption that sentiment is limited only to the explicit citation is incorrect. I present a citation context corpus where more than 200,000 sentences from 1,034 paper–reference pairs have been annotated for sentiment. These sentences contain 1,741 citations towards 20 cited papers. I show that including the citation context in the analysis increases the subjective sentiment by almost 185%. I propose new features which help in extracting the citation context and examine their effect on sentiment analysis.

Thirdly, I tackle the task of identifying significant citations. I propose features which help discriminate such citations from citations in passing, and show that they provide statistically significant improvements over a rule-based baseline.

Acknowledgments

This dissertation would not have been possible without the help of many people; only some of them are mentioned here. Firstly, I would like to thank my supervisor, Simone Teufel, who not only kindled in me the appreciation and enthusiasm for the fascinating field of analysis of scientific literature, but also provided the guidance as well as freedom to pursue my research directions. I can only hope to be as organised, clear and succinct in my thinking and writing as her.

I am grateful to the Association of Commonwealth Universities and the British Council for their support which made this work possible. I would also like to thank Dr. Stephen Clarke and Prof. Anna Korhonen, for their valuable feedback and guidance throughout my research. Many thanks to Diarmuid Ó Séaghdha and Andreas Vlachos for their guidance and suggestions, and Saad Aloteibi for many helpful discussions. Thank you James Jardine, Sandro Bauer and Wenduan Xu; my work desk would have been quite boring without you around.

I would like to thank all my friends in Cambridge, in particular, Safwan, Rehan, Murtaza, Ali, Bilal, Nazneen, Arsalan, Anum, Sohab, Beenish and Maryam. I will always cherish the interesting conversations, fun trips, and most importantly the food I had with you.

I would also like to thank my friends and colleagues back home, especially Abbas, Omer, Salman, Umer Salman and Agha Ali Raza, for constantly pushing me to complete my work. Many thanks to everyone who helped me annotate my data.

I am thankful to my family for their love and support, in particular my brother, who introduced me to the amazing field of Computer Science. Lastly, my deepest thanks go to my wife, Munazzah, for her ceaseless encouragement, continual understanding, unwavering support, wonderful company, and excellent editorial help throughout these years.

In a world where there is a lot that needs to be discovered, created and understood, all of you help me claim my share, for which, I can’t thank you enough.

Contents

List of Figures

Chapter 1

Introduction

The growing availability of textual data on the World Wide Web has triggered an increase in research activities focusing on this area. As a result, research on opinion-centric information retrieval continues to be of interest for many applications, ranging from determining product ratings based on consumer reviews to finding out the political stance of individuals and communities.

A popular example of such an application is trying to predict the sentiment of a movie review. With development of sites such as the Internet Movie Database (IMDB[1]), it is possible to gather information about the sentiment of reviewers on the website towards a movie, from the rating assigned by them. The website allows each reviewer to rate a movie on the scale of 1 to 10 (10 being highest), which can be viewed by the later visitors to the website in an aggregated form. While the exact algorithm to calculate this aggregate score is not disclosed in order to avoid any malicious attempts to temper the rating score, the website administrators have acknowledged that the aggregation is performed as a weighted vote average[2]. It is thus trivial to calculate this aggregate score given all ratings by individual reviewers once we have the appropriate weights as it is a function of those ratings.

However, in the absence of any ratings provided by individual reviewers, calculating the aggregate score becomes a problem which is difficult to tackle. This problem is formally known as *sentiment analysis*. Sentiment analysis, also called opinion mining, is the task of identifying positive and negative opinions, sentiments, emotions and attitudes expressed in text. In my dissertation, I focus on the problem of extracting opinions from scientific text, more specifically, scientific citations.

In this dissertation, I define a *citation* as any mention of another paper in the text of a given research paper. The citing paper is often referred to as the *source paper*, and the cited one is referred to as the *target paper*. In scientific text, it is customary to provide a link to the target paper using various standards consisting of either the author names and the year of publication, or only bracketed numbers in various styles. One such standard is the *Harvard style* of citation, which uses the last name of the author followed by the year of publication. The year of publication is either written in parentheses or delimited

[1] `http://www.imdb.com`

[2] `http://www.imdb.com/help/show_leaf?ratingsexplanation`

by a comma, depending on the usage of the author name, for example, Bird (2008) vs (Bird, 2008). I use the term *formal citation* to refer to any Harvard style citation of the target paper. I restrict myself to analysis of citations in the field of computational linguistics[3], an area in which the prevalent style of citation is the Harvard style. Sentences containing formal citations in the Harvard style are referred to as *citation sentences* in this dissertation.

There has been consistent growth in work focusing on the task of extracting opinion from text, and researchers report around 85-90% accuracy (c.f. Appendix A) for sentiment detection in different genres of text (Nakagawa et al., 2010; Yessenalina et al., 2010; Täckström and McDonald, 2011). Given such good results, one might think that a sentence-based sentiment detection system trained on a different genre could be used equally well to classify citations. However, I argue that this might not be the case as sentiment in scientific citations tend to behave differently.

Firstly, sentiment in citations is often hidden. One reason is the general strategy of avoiding overt criticism because of the sociological aspect of citing (MacRoberts and MacRoberts, 1984; Thompson and Yiyun, 1991). Ziman (1968) states that many works are cited out of "politeness, policy or piety". Negative sentiment, while still present and detectable for humans, is expressed in subtle ways and might be hedged, especially when it cannot be quantitatively justified (Hyland, 1994). An example of such a citation is given in the grey box below[4].

> *While SCL has been successfully applied to POS tagging and Sentiment Analysis (Blitzer et al., 2006), its effectiveness for parsing was* ***rather unexplored****.*

Negative polarity is often expressed in contrastive terms, e.g. in evaluation sections. Although the sentiment is indirect in these cases, its negativity is implied by the fact that the authors' own work is clearly evaluated positively in comparison.

> *This method was shown to* ***outperform*** *the class based model proposed in (Brown et al., 1992) ...*

Secondly, citation sentences are often neutral with respect to sentiment, either because they describe an algorithm, approach or methodology objectively, or because they are used to support a fact or statement (Spiegel-Rosing, 1977; Teufel et al., 2006b). This phenomenon is unlike sentiment present in other genres such as movie reviews, and gives rise to a far higher proportion of objective sentences than in other genres.

> *There are five different IBM translation models (Brown et al., 1993).*

Thirdly, there is much variation between scientific texts and other genres concerning the lexical items chosen to convey sentiment. Such lexicon and terms play a large role overall in scientific text (Justeson and Katz, 1995). Some of these carry sentiment as well, an example of which are given in the citations below.

[3] For reasons of choosing this field, see Section 3.1.

[4] These grey boxes are used throughout this dissertation to quote citation text.

> *Similarity-based smoothing (Dagan, Lee, and Pereira 1999) provides an **intuitively appealing** approach to language modeling.*

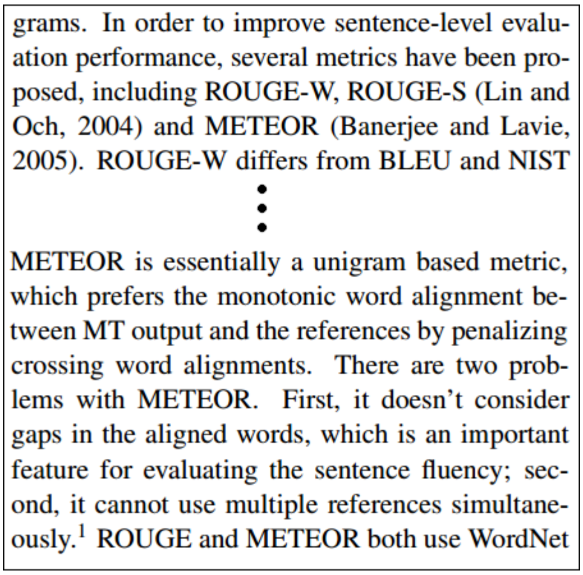
> *Current **state of the art** machine translation systems (Och, 2003) use phrasal (n-gram) features ...*

Lastly, the scope of influence of citations varies widely from a single clause (as in the example below) to several paragraphs:

> *As reported in Table 3, small increases in METEOR (**Banerjee and Lavie, 2005**), BLEU (Papineni et al., 2002) and NIST scores (Doddington, 2002) suggest that ...*

This affects lexical features directly since there could be "sentiment overlap" associated with neighbouring citations.

In this dissertation, I describe my work on the task of detecting sentiment in scientific citations. To address this task, I not only use sentences containing citations, but also the textual passage that contains the citation and its surrounding sentences. In this dissertation, I refer to these sentence as the *citation context* (O'Connor, 1982). Consider Figure 1.1, which illustrates a typical case.



grams. In order to improve sentence-level evaluation performance, several metrics have been proposed, including ROUGE-W, ROUGE-S (Lin and Och, 2004) and METEOR (Banerjee and Lavie, 2005). ROUGE-W differs from BLEU and NIST

⋮

METEOR is essentially a unigram based metric, which prefers the monotonic word alignment between MT output and the references by penalizing crossing word alignments. There are two problems with METEOR. First, it doesn't consider gaps in the aligned words, which is an important feature for evaluating the sentence fluency; second, it cannot use multiple references simultaneously.[1] ROUGE and METEOR both use WordNet

Figure 1.1: Example of the use of anaphora in citation context.

While the first sentence cites the target paper formally, the remaining sentences mentioning the same paper appear after a gap and contain an indirect and informal references to that paper and list its shortcomings. Current techniques are not able to detect linguistic mentions of the citations in such forms. It is clear that criticism is the intended sentiment, but if we define what should count as the right answer only by looking at the citation sentence, we undoubtedly lose a significant amount of sentiment hidden in the text. Given that overall most citations in a text are neutral with respect to sentiment, this makes it even more important to recover what explicit sentiment there is from the context of the citation.

In this dissertation, I examine methods to extract as many sentences from research papers as possible which mention a given paper in as many forms as we can identify (not just as formal citations) in order to identify as much sentiment about the target paper as possible. This context and sentiment information can help researchers in many practical applications, some of which include detecting shortcomings and detecting problems in a particular approach, identifying contributions of a given paper in the research domain, and determining the quality of a paper for ranking in citation indexes by including negative citations in the weighting scheme.

One particular application that I focus on is a new task of identifying importance of a citation in the citing paper. Traditionally, popularity of a paper has been linked to the number of papers citing it. These raw citation counts as well as other more complex bibliometric measures (Garfield et al., 1964; Hirsch, 2005) treat all citations as equal. However, current research shows that the percentage of citations which are important to the content of the citing paper is only 9% (Hanney et al., 2005). I propose features to discriminate these important and significant citations from citations which have been mentioned in passing only, and show that informal citations from the citation context help in getting better results in identifying these *in-passing* citations.

This dissertation is structured as follows: Chapter 2 describes the background literature and existing work in the domain of sentiment analysis as well as citation classification. Chapter 3 explains the construction of three new large citation corpora, along with the process of their annotation. Chapter 4 describes the task for sentence-based citation sentiment classification as well as the features proposed by myself for this task. The task of detecting informal citations is discussed in Chapter 5, where I propose new features for this task, and also show that using these informal citations in sentiment analysis has a positive overall effect. Chapter 6 focuses on the new task I propose for deciding whether or not a cited paper is significant to the citing paper. I conclude the dissertation with a summary in Chapter 7 and also describe future directions of my work.

Chapter 2

Background

This chapter describes previous work on sentiment-based classification of text. Over the past few years, there has been growing interest in the field of sentiment analysis for different text genres such as newspaper text, reviews, narrative text, blogs and microblogs. While relatively less emphasis has been placed on the extraction of opinions from scientific literature, there is at least research work on citation sentiment analysis in the context of related fields such as bibliometrics and summarisation.

Sentiment can be classified using various categorization schemes and can be expressed at multiple levels of granularity of text, some of which are discussed in Section 2.1. The most popular approach for building a sentiment classifier is using supervised methods in a machine learning framework. Some of these methods are discussed in Section 2.2. However, supervised methods require the availability of labelled data. Labelling this data is a time consuming effort and calls for human annotators, who need to examine and label each data instance. Researchers have thus also explored the use of unsupervised methods for classification of sentiment in different text genres. Most of these unsupervised methods rely on a sentiment lexicon to assign a sentiment score to the text. I will describe these methods in Section 2.3.

The sentiment lexicons are largely dependent on the genre of the text being classified. Using the same general purpose lexicon for sentiment analysis in different domains of text is a hard problem. Because there is a marked difference between genres with respect to sentiment detection, most methods are tailored to one particular text type. Section 2.4 describes these differences in genre and some approaches taken by researchers to address this issue.

My approach is targeted towards sentiment detection in citations in scientific papers. This can be represented in the form of a graph, with papers as nodes and citations as edges. Section 2.5 describes some of the exiting graph-based approaches for analysing such networks.

If labelled data is used, the text that is labelled can be a phrase, a sentence or a document. In the case of sentences and phrases, the context in which the text occurs can affect the sentiment expressed in that text and thus plays an important role in determining the intended polarity. Approaches which exploit the text context will be discussed in Section 2.6.

2.1 Categories and Granularity of Sentiment

Various researchers have addressed the problem of classifying sentiment into different categories. The simplest example is the work by Pang et al. (2002), who proposed methods to automatically classify the movie reviews into two categories: *positive* and *negative*. This categorisation is also known as *sentiment polarity detection* or simply *polarity detection*.

While Pang et al. (2002) restricted themselves to only two categories, such a categorisation is not suitable for a review which does not carry a distinctly negative or positive attitude towards a movie. This issue can be solved by introducing a *neutral* category to label text that does not carry any polarity, expanding the classification scheme to three basic categories: *positive*, *negative* and *neutral*. I use this three-class scheme in my research because it is simple as well as sufficiently general to cover most cases of expression of sentiment in scientific citations.

Another line of research revolves around distinguishing between polar and neutral sentiment. This task is often referred to as *subjectivity detection*, and a subjective sentiment can either be positive of negative. Text with a neutral sentiment is thus also called *objective* text. An example of this task is the work by Wiebe et al. (1999), who used this *subjective* vs *objective* categorisation scheme to automatically classify newspaper text.

As in the case of movie reviews, annotation is generally performed by assigning a label from a list of labels to each instance of the training data. However, utilising any hierarchical relation which may be present amongst the labels may lead to improved performance. Pang and Lee (2004) provide an example of this work when they used a two-step hierarchical approach for sentiment classification of movie reviews. In the first step, the system labelled text as either objective or subjective, and in the second step, the extracted subjective text was used to detect polarity.

Some categorisation schemes use finer divisions for assigning subjectivity or sentiment. For example, Wilson et al. (2004) used a four-class scheme for detecting the strength of sentiment in subjective clauses. Their proposed categories were *neutral*, *low*, *medium*, and *high*. More recently, Thelwall et al. (2011) used a scale of one to five for determining positive and negative sentiment in short informal text.

Sentiment categories in text are closely linked to the granularity of the observed text. By granularity, I refer to the different units of length of text to be examined, which can range from a word or a phrase to a sentence or even a whole document. For instance, the word-level method presented by Hatzivassiloglou and McKeown (1997) retrieves positive or negative semantic orientation information only for adjectives. Wiebe and Mihalcea (2006) addressed the task of automatic assignment of subjectivity labels to word senses. For phrase-level categorisation, Turney (2002) worked on predicting the polarity of phrases containing adjectives and adverbs which were extracted from text using POS tag patterns. Wilson et al. (2009) proposed a system for automatically distinguishing between prior and contextual polarity of phrases. At a sentence level, Pang and Lee (2004) used classifiers to detect subjective sentences from movie reviews. The subjective sentences were then used to calculate the overall sentiment of a whole movie review document. Yu and Hatzivassiloglou (2003) presented a solution for three subtasks of sentiment detection at different levels of granularity: detecting subjectivity of documents, detecting subjectivity of sentences, and assigning polarity to sentences. Keeping the methods described above for sentiment classification of citations in mind, labelling phrases would be too fine-grained to provide any useful information about the target citations. On the other hand, labelling the whole citing document does not capture much of the sentiment present in the paper because of the inherent objective nature of scientific writing. Therefore, I use a middle approach by choosing sentences as units of classification.

A whole body of work exists on sentence based sentiment classification in social media, in particular, the micro-blogging platform Twitter [1]. Twitter users post sentences, also called tweets, which consist of at most 140 characters. Such short sentences are mostly about a particular subject or event and applications of detecting sentiment from these sentences range from predicting elections (Tumasjan et al., 2010) results to forecasting stock market trends (Bollen et al., 2011). For this level of granularity, certain features become very important in predicting the sentiment, for example, Read (2005) and Go et al. (2009) explore the use of emoticons for sentiment classification of tweets. Similarly, Brody and Diakopoulos (2011) leverage word lengthening of polar words (such as *cooooooooooooooollllllllllllll!!*) to build a Twitter-specific sentiment classifier.

Granularity becomes particularly important in the case of product reviews, where different features of a product have to be examined. In such cases, sentiment is determined with respect to the target objects. Product features can be thought of as individual target objects towards which the sentiment is to be detected. For example, a camera with an overall positive review may have low battery life. This task has been tackled by many researchers with varying degrees of success (Hu and Liu, 2004; Liu et al., 2005; Wu et al., 2009; Dasgupta and Ng, 2009). This can be mapped to my research problem, where a single sentence can contain citations to many different papers. In such cases, the sentiment toward only the target paper is to be determined while the rest of the sentiment, however strongly expressed, needs to be ignored.

In the next section, I describe various supervised learning methods for the prediction of sentiment, i.e. those methods which use data that has been labelled with target categories, in order to predict sentiment.

2.2 Supervised Methods

The increasing availability of labelled data has played an important role in the application of supervised machine learning methods to sentiment analysis. These methods represent the labelled data in the form of a set of features. The features are then used to learn a function for classification of unseen data. In this dissertation, I approach the problem of sentiment analysis as a classification task.

The foremost example of an approach utilising labelled data is the work by Pang et al. (2002). The authors used Naive Bayes (NB), Maximum Entropy (MaxEnt) and Support Vector Machines (SVMs) classifiers (Berger et al., 1996; Joachims, 1998) for determining if a movie review is positive or negative. The IMDB archive of `rec.arts.movies.reviews` newsgroup[2] was used as the source of data, and only those reviews were selected which

[1] `http://twitter.com`

[2] `http://www.imdb.com/reviews/`

had a rating expressed as a number or as stars. These ratings were used as an automatic indication of the sentiment associated with the review text, and thus converted to two labels: *positive* and *negative*. For obtaining a balanced corpus, 700 documents of each label were selected. N-grams, part-of-speech (POS) tags and their combinations were used as features, and three-fold cross validation was used. Their best system achieved an accuracy of 82.9% when using the unigram presence feature set with SVMs. It should be noted that the corpus used in this work was balanced artificially, thus avoiding the problem of data sparsity for under-represented classes. However, it is possible to learn more from real-life balance between classes, which is what I do in my research.

Supervised learning can also be performed using multiple classifiers, particularly if the labelling scheme allows for hierarchical relations. As described earlier, one example of this is the work by Pang and Lee (2004). They represented sentences in the given document as graph nodes and calculated the minimal cut on that graph to identify the subjective sentences. Afterwards, standard machine learning classification algorithms (NB and SVMs) were applied only on the extracted subjective sentences to predict their polarity. On a balanced polarity corpus of 2,000 reviews, the minimum-cut framework resulted in an accuracy of 86.4%, which represents a statistically significant improvement in polarity classification accuracy over previous attempts.

Another example of supervised sentiment classification at a finer granularity is the work by Wilson et al. (2009), who proposed a system for automatically distinguishing between prior and contextual polarity at the sentence level. The key idea is that the polarity of words can change on the basis of other words present in their context. For instance, in the phrase *National Environment Trust*, the word *trust* has positive priors but is used here in a neutral context. Starting with a collection of clues marked with prior polarity, the so-called contextual polarity in the corpus was identified as *positive*, *negative*, *neutral* or *both*. A two-step approach was taken by first classifying each phrase containing a clue as either neutral or polar. In the second step, only the phrases marked as polar were considered further. A corpus of 8,984 sentences was constructed from 425 documents. Unlike movie reviews, these documents did not contain any information for the automatic extraction of labels. Therefore, this corpus was annotated manually by examining each sentence. As a result, 15,991 subjective expressions were found in the 8,984 sentences. 66 documents (1,373 sentences/2,808 expressions) were separated as the development test and the remaining data was used as the test set. For the task of neutral-polar classification, 28 different features based on word context, dependency parse tree, structural relationship, sentence subjectivity and document topic we used in a machine learning framework. For the task of polarity classification, ten features based on word tokens, word prior polarity, as well as presence of negation, modification relationships and polarity shifters were used. The BoosTexter AdaBoost.HM (Schapire and Singer, 2000) achieved 75.9% accuracy on the former tasks and 65.7% on the latter.

When the units of classification, whether words, phrases or sentences, are present in a sequence, identification of sentiment can also be viewed as a tagging task. Breck et al. (2007) used a simple tagging scheme which tagged each term in a sentence as either being 'in' a polar expression (/I), or 'out' of it (/O). They used various lexical, syntactic and dictionary-based features to identify both /I and /O types of subjective expressions. The MPQA data corpus of 535 newswire articles was used, 135 of which were put aside for parameter tuning and the rest were kept for evaluation. Using 10-fold cross validation and CRFs, the system achieved an F-measure score 70.6%.

Supervised learning with dependency trees was also used by Joshi and Penstein-Rose (2009), who worked on solving the problem of identifying opinions from product reviews. Their method was to transform syntactic dependency relation triplets into features for classification. The motivation was to capture the general relationship between opinionated phrases by 'backing off' to the head word in the triplet. For instance, consider the phrases *a great camera* and a *great mp3 player* with the relations $\{amod, camera, great\}$ and $\{amod, player, great\}$. Here, backing off the head words (*camera* and *player*) to their POS tags results in a more generalised form $\{amod, NN, great\}$, which makes a better indicator for opinion extraction. A collection of 2,200 reviews from the extended version of the `Amazon.com/CNet.com` product review corpus[3] was used, 1,053 of which were subjective. With 11-fold cross-validation in an SVM learner, their method of backing off the head word to the POS achieved approximately 68% accuracy.

In this thesis, I will also use supervised learning methods for sentiment classification of text. This is motivated by the good performance of the methods reviewed in this section. However, there is a second area of research in sentiment analysis which is lexicon-based and uses unsupervised learning. The following section describes this research.

2.3 Unsupervised and Semi-supervised Methods

Another common approach for the detection of sentiment is to develop a sentiment lexicon in an unsupervised way, and then to classify the input text as being positive, negative or neutral using some scoring function based on that lexicon.

The simplest way to explore the sentiment attributed to an object is perhaps to examine the adjectives used to describe that object. Hatzivassiloglou and McKeown (1997) presented an automatic method to classify positive or negative semantic orientation information for adjectives. They proposed that conjunctions between adjectives provide indirect information about orientation. For instance, most adjectives joined by the word *and* have similar orientation (*fair **and** legitimate*) and the ones joined by *but* have different orientation (*simple **but** popular*). They extracted conjunctions of adjectives from the Wall Street Journal corpus along with their morphological relations. This information was used in a log-linear regression model to determine if each pair of conjoined adjectives was of the same or different orientation.

A graph was then obtained where the vertices were the adjectives and the edges represented the orientation link between them. This graph was partitioned into two subsets using clustering. Since these clusters were not labelled, some criteria was needed to distinguish between the positive and negative one. Using the premises that the unmarked member in a pair of antonyms is almost always positive (Lehrer, 1985), and that the unmarked member is likely to be the most frequent term (Hatzivassiloglou and McKeown, 1995), the average frequencies in each subset were compared and the subset with the higher value was labelled as positive. Their best configuration achieved a classification

[3] `http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html`

accuracy of 92% for dense graphs. For sparser graphs, the accuracy was reported to be 78%.

Another way to exploit the information provided by adjectives is to examine the context in which they are used. This idea was exploited by Turney (2002), who used the hypothesis that in a document, words of the same orientation should occur close to each other. This co-occurrence can be estimated by calculating the Pointwise Mutual Information (PMI) between the words. Formally, the PMI between two words w_1 and w_2 is defined as:

$$PMI(w_1, w_2) = \log_2 \left(\frac{p(w_1 \,\&\, w_2)}{p(w_1)p(w_2)} \right) \tag{2.1}$$

Turney extracted phrases containing adjectives and adverbs by using POS tag patterns. He then estimated the PMI of each phrase combined with one of the two target phrases, *poor* and *excellent*, using the hit count reported by the AltaVista search engine[4] with the NEAR operator. The semantic orientation of a phrase was calculated by subtracting the PMI between that phrase and the word *poor* from the PMI between that phrase and the word *excellent*.

$$\begin{aligned} SO(\mathit{phrase}) &= PMI(\mathit{phrase}, \textbf{excellent}) - PMI(\mathit{phrase}, \textbf{poor}) \\ &= \log_2 \left(\frac{\mathit{hits}(\mathit{phrase}\ \text{NEAR}\ \textbf{excellent})\ \mathit{hits}(\textbf{poor})}{\mathit{hits}(\mathit{phrase}\ \text{NEAR}\ \textbf{poor})\ \mathit{hits}(\textbf{excellent})} \right) \end{aligned}$$

Each review was then classified as positive or negative based on the sign of the average semantic orientation. The system was tested on 410 reviews collected from epinions.com on four different domains. The accuracy achieved ranged from 66% to 84% over these domains, with an average of 74%.

A similar approach was taken by Taboada and Grieve (2004), who used information about the position of the text in a document to expand the lexicon. They assigned more weight to certain parts of the text where they believed most subjective content was concentrated. Using the AltaVista search engine, this method achieved an overall accuracy of 65%. On exploring the technique later (Taboada et al., 2006), they found that this method does not perform as well if the Google search engine[5] is used.

In the work above, the seeds of polarity information were only two words, *excellent* and *poor*. However, using a wider range of polar phrases may lead to improvements in polarity classification. Yu and Hatzivassiloglou (2003) used this idea of expanding the list of words for the task of assigning polarity to the sentences. They measured the co-occurrence of words in each sentence with a known seed set of 1,336 semantically oriented words. An average semantic orientation was then calculated for each sentence and cut-off thresholds for assigning positive and negative polarities were applied. These thresholds were obtained from the training data through density estimation using a small hand-labelled set of sentences. This system was tested using different POS filters on the semantically oriented words; the best feature set (adjectives, adverbs and verbs) achieved an accuracy of 90%.

[4] `http://www.altavista.com/`

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

Another way to increase the number of items in the lexicon is to use bootstrapping. The general approach is to start with a limited number of polar phrases in the lexicon and to extract similar phrases from unlabelled data. These extracted phrases are then added to the polar lexicon and the process is repeated until a stopping criterion is reached. Riloff and Wiebe's (2003) bootstrapping algorithm learns linguistically rich patterns for subjective expressions. For example, the pattern *<subj> was satisfied* will match all sentences with the passive form of the verb satisfied. High precision classifiers trained on known subjective vocabulary were used to automatically identify objective and subjective sentences in unannotated text. The labelled sentences from these classifiers were fed to an extraction pattern learner. The remaining unlabelled sentences were filtered through a separate pattern-based subjective sentence classifier that used the extraction pattern previously learned. To close the bootstrap loop, the output of the pattern-based classifier was returned back to the extraction pattern learner and the extracted patterns were used again in the initial high-precision classifiers. This system was used on a balanced corpus of roughly 34,000 sentences and achieved a precision of 0.902 and a recall of 0.401.

Exploring more lexical features in a later work, Wiebe and Riloff (2005) developed a Naive Bayes (NB) classifier using data extracted by a pattern learner. This pattern learner was seeded with known subjective data. Additional features for this NB classifier included strong and weak subjective clues from a pre-existing rule-based system, POS tags, pronouns, modals, adjectives, cardinal number data and adjectives. The classifier was used to classify the unlabelled text corpus, and the most confidently classified sentences were added to the training set for another cycle. They trained the system for two cycles. Using test a corpus of 9,289 sentences, 5,104 of which were subjective, they reported up to 0.863 subjective recall with a subjective precision of 0.713 (corresponding to 73.4% accuracy).

While the bootstrapping methods described above determine the similarity between phrases statistically, manually compiled dictionaries and thesauruses can also act as a source of additional lexical items. One such resource is WordNet (Miller, 1995), a large lexical dictionary which provides sets of nouns, verbs, adjectives and adverbs. Each set in WordNet corresponds to a word sense, not a word string. Each set represents a unique concept and members of each set, each of which is a sense, are synonyms of each other. These sets are also called synsets. WordNet also provides information about lexical and semantic relations present between these synsets. Each synset also contains a brief definition of the sense of the words in the synset, also known as 'gloss'. Due to the availability of this resource, many researchers have based their work of sentiment analysis on WordNet.

For instance, Kim and Hovy (2004) classified word polarities by using a manually scored seed word list and by adding related antonyms and synonyms from WordNet. The system calculated the sentiment score for a sentence by taking the product of all word polarities in a region window starting from the mention of the opinion holder to the end of the sentence. For a test set of 100 sentences, this method achieved an accuracy of 81%.

Another WordNet-based approach was put forward by Esuli and Sebastiani (2006a). To obtain the training set, they used bootstrapping on WordNet glosses, synonyms and antonyms of three seed sets (positive, negative and objective). This training set was fed to three different learning approaches. The first approach used a binary classifier to separate subjective and objective terms, and another binary classifier to separate positive and negative terms from the subjective ones. The second approach used a binary classifier to separate negative from non-negative terms, and another binary classifier to separate positive from non-positive terms. The non-positive negative terms were classified as negative, the non-negative positive terms as positive, and the rest were classified as objective. The third approach used a simple ternary classifier. They experimented with these three approaches using different learning algorithms (Naive Bayes, SVMs, Rocchio, PrTFIDF) and mutual information based feature selection cut off points. For the best parameter combination, their system achieved an accuracy of 67.6% for subjective/objective classification and 66% for positive/negative/objective classification.

The authors also made available a lexical resource called SentiWordNet. In SentiWordNet, each WordNet synset(s) is associated to three numerical scores describing how positive, negative or objective ($Pos(s)$, $Neg(s)$, $Obj(s)$) are the terms contained in the synset. This resource provides a general purpose sentiment lexicon; however, it has not yet been explicitly evaluated.

A major issue in lexicon-based methods for sentiment analysis is sentiment analysis across multiple domains. This is due to the subjective nature of domain dependent features, such as adjectives, when they are moved across disciplines. For example, when moved from movie reviews to product reviews, an *unpredictable movie* is considered good but an *unpredictable steering* is not (Turney, 2002). Consequently, current research approaches to sentiment analysis in multiple domains are bifurcated. While some researchers limit themselves to a specific domain (which, in my case, is scientific citations), others aim to port existing labelled data from one domain to other domains. Some of these approaches are described in the next section.

2.4 Genre Differences

Since my work involves sentiment analysis of scientific text, the problem of classifier dependence on the genre of text becomes particularly relevant for the following reasons.

- The style of writing for science is prescriptive and mostly objective. Sentiment and opinions are not expressed directly in the text. Prescriptive writing guidelines advice against the presence of too much opinion or sentiment in text.
- Due to the social aspects of citations, much of the negative sentiment is covert and where present, is mostly hedged. Therefore, lexically observable sentiment is hard to detect.
- Negative sentiment is often present only indirectly in terms of problems and thus has to be inferred.

Detection of sentiment (particularly criticism) is thus a difficult problem in scientific text and not much research on science-specific sentiment detection exists. In the absence of a such methods, it might be possible to use a general classifier and adapt it to scientific text.

However, work in the area of domain and topic independent sentiment classification suggests that existing supervised and unsupervised classification approaches are highly topic dependent and creating a general classifier is a difficult task (Engström, 2004) . One such example is the research by Choi and Cardie (2009), who proposed a method to adapt an existing general-purpose lexicon to newswire domain using Integer Linear Programming (Wiebe and Riloff, 2005; Wilson et al., 2005). The constraints were based on word and expression level polarities as well as *content-word negators*. Content-word negators are words such as *prevent* and *eliminate*, which act semantically as negators but are not functional words. Extrinsic evaluation on an expression level polarity task was used. Using the adapted polarity lexicon improved the accuracy from 70.4% to 71.2%.

Some efforts (Aue and Gamon, 2005; Blitzer et al., 2007) have been made to address these difficulties and adapt a sentiment classification system to another domain, but they do not seem to work equally well across multiple domains. Recent work by Taboada et al. (2011) on construction of such a lexicon reports 78.74% accuracy over four different review datasets, but domain specific classifiers have reported results up to 90% accuracy when the datasets are analysed in isolation (Maas et al., 2011). Therefore, I do not use domain adaptation in my research and limit it to analysis of sentiment only in scientific text.

2.5 Graph-Based Approaches

Citations to scientific articles can be represented as a graph such that each article is a node and each citation is an edge from the citing paper to the cited paper. While analysis of such graphs or network has a rich history (Garfield et al., 1964; Yu and Van de Sompel, 1965; Small, 1982), most of the past work focused on analysis of only small manually annotated sets of citations. However, with the increase in availability of machine readable text through optical character recognition (OCR) and digital libraries, much work has been performed recently on automatic extraction and of citations from scientific articles and their indexing.

An early example or such work is by Giles et al. (1998), an improved version of which is now available as CiteSeerX (Li et al., 2006). Tang et al. (2008) presented a system ArnetMiner[6] to automatically extract author profiles and citation network by mining existing digital libraries. They also proposed a unified model for topical aspects of papers, authors and publication venues. However, their work differs with mine as they do not focus on detection of sentiment in citations.

Microsoft[7] and Google[8] also offer their academic search engines, which have triggered recent work on not only different information retrieval related task faced by them but also their comparative evaluation (Meho and Yang, 2007; Bar-Ilan, 2008; Roy et al., 2013). However, both products use propriety algorithms which are neither publicly available nor use any information about citation sentiment.

Nevertheless, the algorithm behind the popular Google web search, PageRank by Page et al. (1999), has been the topic of much research. The authors modelled each web pages as nodes and links between them as edges. They calculated the rank of each web

[6] http://arnetminer.org/

[7] http://academic.research.microsoft.com/

[8] http://scholar.google.com

page with the help of a random surfer model, which assigns higher rank to nodes which are encountered more in a random walk. In this way, PageRank is a general algorithm for ranking nodes in a graph and has consequently been applied to citation networks of various fields and has been reported to yield interesting results (Chen et al., 2007; Ma et al., 2008b; Radev et al., 2009). One salient feature of PageRank is the assumption that edges from important nodes should carry more weight than edges from unimportant ones. However, it also assumes that these edge weights are always positive. Therefore, PageRank may not be applied to networks which carry negatively weighed edges and model polar relationships such as praise/criticism and trust/distrust. Many researchers have tried to overcome this shortcoming. For example, Guha et al. (2004) approached this problem by first calculating the rank with only positive edges of trust relations and then adding them incrementally. De Kerchove and Dooren (2008) approached the problem by introducing a model based on a memory-capable random surfer which 'remembers' the distrusted nodes and avoids them when walking the edges of the graph. Kunegis et al. (2009) present a PageRank-based measure called Negative Rank in order to identify malicious users (also called *trolls*) in a social network.

All the techniques described above assume the presence of negatively weighed edges and in most cases, these weights as assigned by the users of the networks. However, such weights do not exist for citation networks. In this dissertation, I present techniques to automatically assign positive and negative weights to the edges in a citation network.

The next section describes some general approaches which take the context into account while analysing sentiment of words, phrases, and sentences in a particular domain. I also describe methods for identifying sentences in the context of a scientific citation.

2.6 Analysis of Context

In comparison to other domain-based text classification tasks, sentiment detection is affected by discourse context in particular. Polanyi and Zaenen (2006) argue that the attitude of the author of the given text cannot be calculated on the basis of individual terms alone. They describe how this polarity, or valence as they call it, can be "shifted" by the context and is thus critically affected by discourse structures. For example, consider the following sentence.

> *Although Boris is **brilliant** at math, he is a **horrible** teacher.*

The authors state that if we ignore the valence shifters and assign the lexical level score of +2 to *brilliant* and -2 to *horrible*, the sum would be 0 and the sentence would be considered neutral. However, we can see that the intention behind this sentence is to highlight Boris' ineptitude of teaching. Assigning a neutral score to this sentence would thus be incorrect. In order to address this issue, the authors provide a fine-grained calculation scheme for assigning valence value to some given text, on the basis of sentence-based contextual shifters like negatives, intensifiers and modals, as well as discourse-based contextual shifters like connecters, reported speech and subtopics. Using the scheme proposed, the valence shifter *although* nullifies the effect of the positive term and the total score for this sentence is evaluated to be -2.

Polanyi and Zaenen (2006) conclude that while valance calculation for documents describing multiple entities can be performed with respect to each entity separately, encompassing factors which influence the document structure (like genre) must also be taken into account. For this reason, defining a general-purpose sentiment lexicon is only part of the job, as the sentiment of a particular phrase may change with the context as well as the document topic.

One important issue when addressing the changes introduced by the context is how negation is handled. Presence of a negation word can flip the polarity of a sentence. For instance, the positive sentiment in the sentence "*This camera is good*" would become negative when the word *not* is added i.e. "*This camera is not good*". Das and Chen (2001) proposed addressing this at a lexical level by appending a special token to all words near the negation term. The sentence in the example above would become "*This camera is not good_NOT*". Thus for a supervised classifier, the term *good* would be different from the term *good_NOT*, which may help in classifying the intended sentiment correctly. However, this technique does not work for more complicated sentences such as "*The camera was not the best but still good*". Here, the intended sentiment is not negative but if the 'scope' of negation is taken to be the entire sentence, an incorrect sentiment label would be attached to this sentence.

To address the problem of determining the scope of negation, Khan (2007) explored grammatical relations in sentences. He showed that such relations can improve the accuracy of sentiment detection in movie reviews from 87.3% to 88.4%. Jia et al. (2009) also proposed a rule-based scope identification procedure in which parse trees and typed dependencies are used to identify the scope of each negation term present in the sentence. Polarity of each sentiment term inside the scope of a negation is also flipped in accordance with the corresponding nesting level of negation terms. In other words, if a term is inside the scope of n negation terms, its sentiment is flipped n times. The procedure, evaluated on a review corpus of 1,000 sentences, achieved an accuracy of 88.4%. This scope identification was also shown to improve performance of an opinion retrieval system by 11.3%. Detecting negation scope in science-related text is however a non-trivial task and is the focus of much current research (Councill et al., 2010; Sohn et al., 2012; Morante and Blanco, 2012).

A similar phenomenon is observed by Hornsey et al. (2008) who report that authors want to 'sweeten' the negative sentiment toward a cited paper by mentioning a positive attribute of that paper before mentioning the shortcoming. This strategy serves to soften the effects of the intended criticism among their peers. Such rhetorical strategies are easily detectable by humans but are difficult for a bag-of-features to capture.

Sentence position also plays a part in the context of summarizing the overall document sentiment. As argued by Pang and Lee (2004), the last few sentences of a review in particular serve as a better source for summarizing the overall review sentiment. Mao and Lebanon (2006) tried to capture the global sentiment of movie reviews through local sentiments at sentence level. They proposed a sequential flow model to represent a document where each sentence is scored locally using a predictor based on Conditional Random Fields (CRFs)(Lafferty et al., 2001). A nearest neighbour classifier was used to predict the global sentiment of the document from local scores. This method produced better results than using a bag-of-words model, but the maximum accuracy achieved was only 36%.

Since the focus of this thesis is citation text, it is imperative to examine work on discourse structures with respect to citations as well. While the efforts on automatic extraction of citations are relatively fewer (Giles et al., 1998; Ritchie, 2008; Councill et al., 2008; Radev et al., 2009), classification of citations has a rich history in the literature which can be traced back to the work of Garfield et al. (1965), who proposed a 15-class scheme for classifying citations on the basis of the reason the referenced papers were cited in the paper. Many different schemes have been proposed since then for annotating citations according to their function, ranging from the three-class scheme to one with 35 categories (Cole, 1975; Spiegel-Rosing, 1977; Finney, 1979; Peritz, 1983; Garzone, 1997; Nanba and Okumura, 1999; Teufel et al., 2006b). I use the state-of-the-art work of Radev et al. (2009) in order to obtain a list of papers which cite a given paper, the details of which are given in Section 3.1. However, this research does not take into account the textual passage that contains the citation and its surrounding sentences. As mentioned in Chapter 1, these sentences form the citation context, and a contribution of my work is to use this context in sentiment analysis of citations.

The task of determining a citation's context has a long tradition in library sciences (O'Connor, 1982). Nanba and Okumura (1999) focused on the tasks of extracting areas of influence of a citation and classifying sentences in this area using a 3-class annotation scheme: *type B* for citations to base on other work, *type C* for comparison with other work, and *type O* for any other citation. Using a training set of 100 citations they extracted cue phrases to identify the citation area and trained a rule-based system. On a test set of 50 citations, they reported an F score of 0.779. For the task of classification of citations, they trained a rule-based system on 282 manually identified citation contexts. The accuracy for a test set of 100 citations was 83%.

One of the recent works which deals with scientific text and its rhetorical analysis is by Teufel et al. (2006b). The authors used cue phrases along with other features to classify citations based on the author's reason for citing a given paper. They worked on a 2,829 sentence citation corpus manually annotated using the class labels from a 12-category annotation scheme for citation function (Teufel et al., 2006a). Using the machine learning IBk algorithm, the method achieved a macro-F score of 0.68. Since this approach is similar in some aspects to my work, I compare my results with this score in Chapter 4.

Connection of citations with anaphora has also been noted by Kim and Webber (2006). The authors addressed the task of resolving the pronoun *they*, and linking it to its corresponding citation in scientific papers from the field of Astronomy. The method used a maximum entropy classifier with features based on the distance to the nearest preceding citation and the categories of verb associated with the pronoun. On a test set of 377 examples of *they*, 81 of which referred to a citation, the method achieved a precision of 0.90 and a recall of 0.73.

Exploring co-reference chains for citation extraction, Kaplan et al. (2009) used a combination of co-reference resolution techniques. Their corpus consisted of 94 sentences of citations to 4 papers, which is likely to be too small to be representative. While they reported a macro-average F_1 score of 85% for their method, the micro-average F_1 score was only 69%.

Most recently, Qazvinian and Radev (2010) proposed a framework of Markov Random Fields (MRF). The task is to classify whether or not a sentence contains an *implicit citation* to a target paper. The authors define an implicit citation as any sentence which contains information about a given paper, without citing it explicitly. Therefore, according to their definition, it is not necessary for an implicit citation to contain any linguistic reference to a formal (or *explicit*) citation, for instance anaphora or acronyms. In contrast, my work focuses on detection of sentences which contain such mentions.

In their model, each sentence is taken as a node in a graph. Links are created between neighbouring sentences using a sigmoid function based on the cosine similarity between the sentences. The weight S_{ij} for the link between sentence i and j is calculated by:

$$S_{ij} = \frac{1}{1 + e^{-cosine(i,j)}} \quad , \tag{2.2}$$

where $cosine(i, j)$ gives the cosine similarity score between the two sentences. The neighbourhood varies from using only one sentence in the immediate proximity to all sentences present in the paper. The individual observable potential of a sentence containing an implicit citation is based on multiple features. One of these features is a binary flag which is active when the sentence contains an explicit citation. Other features include similarity with the reference and presence of lexical phrases such as this method and these techniques. The authors used Belief Propagation (Yedidia et al., 2003) to calculate the marginal probabilities of the constructed network using different neighbourhood radii. They used 10 papers and 203 annotation instances, each one corresponding to a single paper–reference pair. Their best system achieves an average $F_{\beta=3}$ score of 0.54.

However, none of the approaches discussed above focus on sentiment analysis of scientific citations. There is also a lack of research which takes the context of a citation into account when determining the sentiment. In this dissertation, I attempt to fill this void and present a large annotated corpus of citation sentiment which takes the citation context into account as well.

2.7 Summary

This chapter describes current approaches to sentiment analysis for different genres of text. While the state-of-the-art work on sentiment analysis uses supervised methods in a machine learning framework, unsupervised approaches to the detection of sentiment have also been explored. Some work on topic dependencies in sentiment analysis exists, but developing a ‘general’ sentiment analyser still remains a hard task. Many different approaches have been described in this chapter which use various evaluation measures as well as datasets. For this reason, a direct comparison of these approaches is not feasible.

While much work has been done on sentiment analysis, annotation schemes and the classification of scientific text in general and citations in particular, not much of the previous work uses a simple three-way classification. Similarly while there exists a body of work which analyses the context of a citation, no approach addresses the task of including the context in citation sentiment analysis. My research is a synthesis of both these areas and I present my approach to context-based sentiment detection in scientific citations in the following chapters.

Chapter 3

Corpus Construction and Annotation

This chapter describes the corpora that I have created for training and testing of detection of citation sentiment, citation context, and reference significance detection. For these tasks, I needed an annotation of each citation and the area around it with the sentiment towards the cited paper.

I could not find any publicly available data source with this type of annotation. However, there exist possible starting points for the development of my own annotated corpus, which I will discuss in Section 3.1. In Section 3.2, I will describe the corpus of citation sentences which I use, along with my selection criteria. My annotation of this corpus is described in Section 3.3. This annotation is based on a set of target papers only, and does not provide full coverage of all citations in the citing paper text. My reason for doing so is that I am interested in aggregation of opinion towards a target paper. Section 3.4 describes the classification scheme of a subset of these citation sentences, which includes the context as well. Section 3.5 explains how this subset was annotated. Lastly, section 3.6 describes a paper-level annotation scheme which classifies whether or not a cited paper is significant to the content of the citing paper.

3.1 ACL Anthology and Related Corpora

I restricted myself to the field of Computational Linguistics (CL) for various reasons. Most of the scientific publications in the field are available electronically in the form of machine-readable text. This leads to a relatively complete coverage of citations within the corpus. Moreover, since annotation is involved, my own training in this field is of significant help to decide which classes to assign. While there exist other publicly available collections of open access articles (such as PLoS[1]), these datasets are less suitable for my requirements as it would require availability of annotators which are experts in the field of medical science. Thus, as far as the corpus is concerned, I analyse sentiment in scientific literature in the field of CL only and the resulting classifier may thus be suitable for this

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

field only. However, we will see in Section 4.2 that my proposed method does not include any domain-specific information. While its application to literature from other branches of science might be possible, I focus only on CL in this dissertation.

I use the ACL (Association for Computational Linguistics) Anthology[2] (Bird et al., 2008). The ACL anthology is a digital archive which contains conference and journal papers in natural language processing (NLP) and CL, since the beginning of the field in 1965, in the PDF format (Bienz et al., 1993). It is maintained and curated by researchers in the field on a voluntary basis, currently by Min-Yen Kan[3]. At the time of writing, the anthology hosts over 21,800 papers from various conferences and workshops and the *Computational Linguistics* journal. Each paper is assigned a unique ID, which I will refer to as the *ACL-ID* hereafter. The anthology is regularly updated with data from new conferences.

As mentioned in Chapter 1, the prevalent style in the field of Computational Linguistics for citing a paper formally is the Harvard style. This style is used to mention the *cited paper* (or *target paper*) within the text of the citing paper. Other information about the paper, such as the journal it appears in or the conference it was presented in, is listed as one of the reference items at the end of the citing paper in the section titled "References". I will refer to this section as the *reference section* in this dissertation.

For my research on sentiment detection, the mere availability of paper text on its own is not sufficient; I additionally need a markup of citation in running text. As mentioned earlier, in this dissertation, I use the term *formal citation* to refer to any Harvard style citation of the target paper. The running text of the formal citation is necessary to identify the sentences which may carry sentiment towards the target paper. The ACL anthology provides neither fully machine-readable text nor citation information. However, the following corpora for the ACL anthology are available with annotated citation sentences:

1. ACL Anthology Network
2. PTX/Ritchie Citation Corpus
3. Citation Function Corpus

I will now describe these corpora in detail.

3.1.1 The ACL Anthology Network

One of the much used resources based on the ACL anthology is the ACL Anthology Network (AAN) (Radev et al., 2009). The AAN corpus has been created from the ACL Anthology and published by the University of Michigan CLAIR Group[4]. New releases of the corpus are provided at regular intervals.

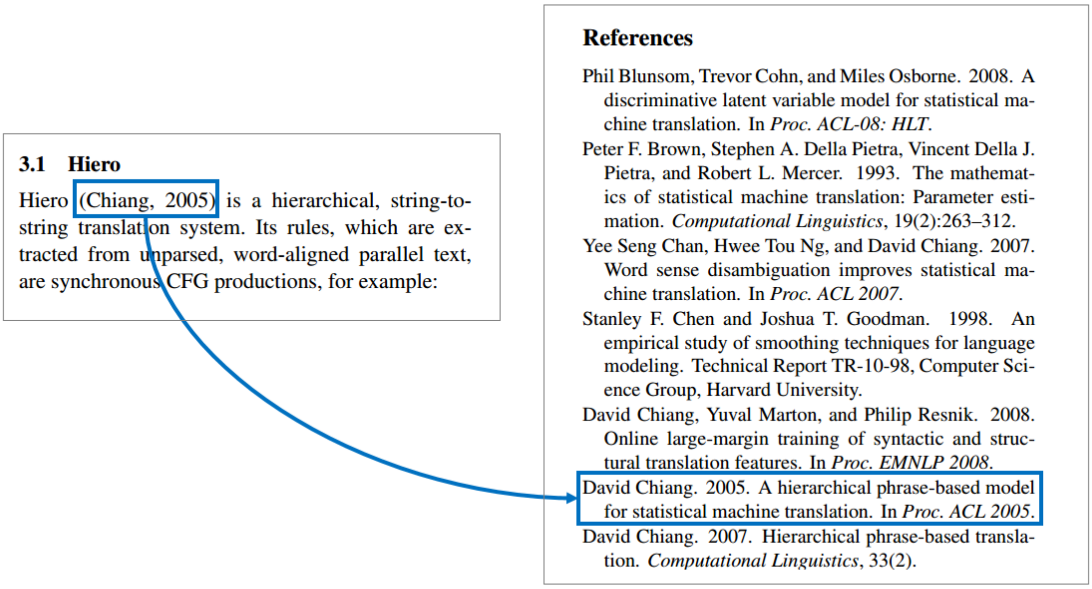
The AAN corpus consists of paper text converted from PDFs from the ACL Anthology using automated tools. However, as like the ACL Anthology, AAN is a dynamic resource,

[2] http://aclweb.org/anthology-new/

[3] http://www.comp.nus.edu.sg/~kanmy/

[4] http://aan.eecs.umich.edu/homepage/

and the number of paper it includes has risen to 19,647 (as of February 2013). The experiments performed in this dissertation are based on the release of December 2009, which was the latest version available at the time of my initial data collection. This release contains only 15,160 papers. All references to the AAN in this thesis refer to this version of the network.



3.1 Hiero

Hiero (Chiang, 2005) is a hierarchical, string-to-string translation system. Its rules, which are extracted from unparsed, word-aligned parallel text, are synchronous CFG productions, for example:

References

Phil Blunsom, Trevor Cohn, and Miles Osborne. 2008. A discriminative latent variable model for statistical machine translation. In *Proc. ACL-08: HLT*.

Peter F. Brown, Stephen A. Della Pietra, Vincent Della J. Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–312.

Yee Seng Chan, Hwee Tou Ng, and David Chiang. 2007. Word sense disambiguation improves statistical machine translation. In *Proc. ACL 2007*.

Stanley F. Chen and Joshua T. Goodman. 1998. An empirical study of smoothing techniques for language modeling. Technical Report TR-10-98, Computer Science Group, Harvard University.

David Chiang, Yuval Marton, and Philip Resnik. 2008. Online large-margin training of syntactic and structural translation features. In *Proc. EMNLP 2008*.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proc. ACL 2005*.

David Chiang. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2).

Figure 3.1: Example of a citation and its reference section entry.

In scientific articles, it is customary to list every information source which has been cited in the article, at the end of the article text in a section titled *References*. An example of the text in the reference section corresponding to a citation is given in Figure 3.1. In the ANN, the reference sections of the papers have been separated automatically from the paper text. Each reference item has been matched manually with the corresponding ACL anthology paper. The remaining text is processed automatically using string-based heuristics to split sentences. Each sentence is assigned an ID, which consists of the sentence number and total sentences in the paper, separated by a colon. I will now describe the information present in the corpus which is of particular relevance to my work.

Metadata: Paper metadata is available for each paper, containing title, authors' names, year of publication, publication venue, etc. Figure 3.2 shows an example of this metadata for the paper with ACL-ID P04-1035.

```
id = {P04-1035}
author = {Pang, Bo; Lee, Lillian}
title = {A Sentimental Education: Sentiment Analysis Using
Subjectivity Summarization Based On Minimum Cuts}
venue = {Annual Meeting Of The Association For Computational
Linguistics}
year = {2004}
```

Figure 3.2: Paper metadata.

```
C08-1104 ==> P04-1035
D07-1035 ==> P04-1035
P04-1035 ==> J94-2004
P04-1035 ==> P02-1053
P04-1035 ==> P97-1023
```

Figure 3.3: Sample paper–reference pairs from the citation graph file.

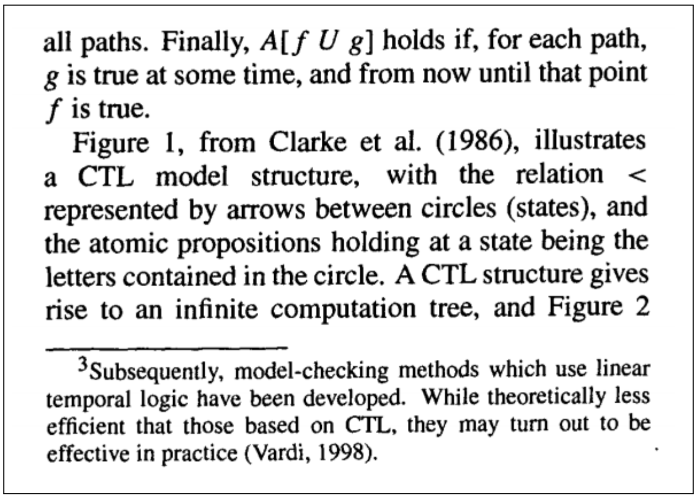
Citation Graph: A graph of citation links is also provided. Each node is an ACL paper – the source node is the citing paper, and the target node is the cited paper. This mapping has been curated manually and is thus reliable. An extract from this data is shown in Figure 3.3, where the direction of the arrow in the pair indicates that the target paper on the right has been cited one or more times in the source paper on the left. I will refer to these pairs as *paper–reference pairs* in this dissertation.

Paper Text: Raw paper text is in a form where each line represents one sentence in the paper. Since this text is extracted from the PDF files automatically, it is noisy and contains various errors such as incorrect spellings and tables merged in text. Sentence boundary recognition is sometimes erroneous as well.

Figure 3.4 illustrates the quality of AAN text conversion, where 3.4a shows a snippet of the PDF and 3.4b shows the corresponding AAN text. We can see that the paragraph break is not represented in the converted text. The text of sentence 31 contains extra characters in the formula-to-text conversion. Sentences 32 and 33 should ideally be merged but the text processing pipeline has split the sentence incorrectly. Footnote 3 has been incorrectly merged with sentence 34. Moreover, the word *logic* is also incorrectly merged with the word *have*. Such conversion errors may have knock-on effects in tasks further downstream, such as the research performed in this thesis. However, I do not correct any errors and use the text *as-is* because I want my research to be comparable to previous work in the field.

Citation Sentences: In the AAN terminology, the *citation summary* of a target paper is defined as the set of all sentences from all corpus papers citing that target paper. These sentences have been extracted automatically for all papers in the corpus using string-based heuristics by matching the citation pattern, author names and publication year within the sentences. Each line starts with the ACL-ID of the citing paper and a citation serial number, followed by sentence ID and text from that paper. An example is shown in Figure 3.5.

The availability of the citation sentences is particularly important to my research as they contain the exact positional mapping to the sentence in the source text where the citation occurs, which is required additionally when examining the context of citations (as I will do in Section 3.4). The availability of positional information was the main reason why I decided to use this data set.



all paths. Finally, $A[f\ U\ g]$ holds if, for each path, g is true at some time, and from now until that point f is true.

Figure 1, from Clarke et al. (1986), illustrates a CTL model structure, with the relation < represented by arrows between circles (states), and the atomic propositions holding at a state being the letters contained in the circle. A CTL structure gives rise to an infinite computation tree, and Figure 2

[3] Subsequently, model-checking methods which use linear temporal logic have been developed. While theoretically less efficient that those based on CTL, they may turn out to be effective in practice (Vardi, 1998).

(a) Actual paper text (PDF).

```
31:134  Finally, A\[f U g\] holds if, for each path , g is true at
    some time, and from now until that point f is true.
32:134  Figure I, from Clarke et al.
33:134  (1986), illustrates a CTL model structure, with the relation <
     represented by arrows between circles (states), and the atomic
    propositions holding at a state being the letters contained in the
     circle.
34:134  A CTL structure gives rise to an infinite  computation tree,
    and Figure 2 3 Subsequently, model-checking methods which use
    linear temporal logichave been developed.
35:134  While theoretically less efficient that those based on CTL,
    they may turn out to be effective in practice (Vardi, 1998).
```

(b) Automatically extracted text.

Figure 3.4: Comparison of original paper with corresponding AAN text.

```
C08-1104:89     104:203 Movie-domainSubjectivityDataSet(Movie): Pang and Lee (2004) used
    a collection of labeled subjective and objective sentences in their work on review
    classification.5 The data set contains 5000 subjective sentences, extracted from
    movie reviews collected from the Rotten Tomatoes web formed best.
--------------------------------------------------
C08-1104:90     26:203  2 Related Work There has been extensive research in opinion
    mining at the document level, for example on product and movie reviews (Pang et al.,
    2002; Pang and Lee, 2004; Dave et al., 2003; Popescu and Etzioni, 2005).
--------------------------------------------------
D07-1035:91     39:276  (2003), Pang and Lee (2004, 2005).
--------------------------------------------------
P06-2079:92     125:254 4.1 Experimental Setup Like several previous work (e.g. , Mullen
    and Collier (2004), Pang and Lee (2004), Whitelaw et al.
```

Figure 3.5: Sample data for the citation summary of paper P04-1035.

3.1.2 Alternate Citation Corpora

Two other corpora based on the ACL Anthology are available which I do not use. The following sections describe these corpora as well as the reasons why they are unsuitable for my work.

3.1.2.1 PTX/Ritchie Citation Corpus

The PTX corpus is an earlier snapshot of about 10,000 ACL documents created in the CITRAZ project[5], which took place in the Computer Lab at the University of Cambridge (2003-2006). It exists as an XML-based representation of papers and was created by a specialised PDF-to-SciXML converter (Hollingsworth et al., 2005). This corpus has been citation-parsed by Ritchie (2008) using regular expressions. The citation-parsed corpus contains metadata about ACL papers and one-sentence and three-sentence windows around each citation of that paper.

While it is a relatively large corpus, it has some shortcomings. Since the corpus consists of processed PDFs from an earlier snapshot of the ACL anthology, the corpus does not contain papers published later than 2005. There are also some papers which have a placeholder name instead of their ACL-IDs assigned to them as they were not published at the time of collection but needed to be included in the PTX corpus for external experimental reasons. Therefore, it would be necessary to resolve these papers to their ACL-IDs, which is a non-trivial task. Moreover, while the evaluation results of citation discovery for this corpus show a high precision (86%), the recall is relatively low (32%). It should however be noted that the corresponding precision and recall figures for AAN are not known as no such evaluation has been performed.

3.1.2.2 The Citation Function Corpus

The Citation Function Corpus[6] (CFC) is the scientific text corpus which was created by Teufel et al. (2006b) in SciXML, the same converted XML format as PTX. The corpus consists of 116 articles with 2,829 citations. However, I had access to only a subset of 86 articles comprising of 12,388 sentences, 1,942 of which contain one or more citations. This is of course a substantially smaller corpus than AAN.

Moreover, this corpus consists of papers which do not necessarily cite each other, and it does not contain all incoming citations. However, having access to all incoming citations is essential for my work, in order to perform accumulative sentiment analysis for specific target papers. For these reasons, I did not use this corpus for my research. I will, however, experimentally evaluate my methods on this corpus for comparison in Chapter 4.

[5] `http://www.cl.cam.ac.uk/~sht25/Project_Index/Citraz_Index.html`

[6] The citation function is defined as the "author's reason for citing a given paper"

3.2 The Citation Sentence Sentiment Corpus

The 2009 version of the AAN corpus, that I use, consists of 15,160 papers containing 72,122 citations to 6,519 AAN-internal papers. Annotating sentiment in all these citations manually would be an impossible task. A smaller set of target papers was therefore required for annotation. As a starting point, I chose papers which were mentioned in computational linguistics related blogs on the Internet. There were two reasons for doing so.

Firstly, blogs provide a list of papers which have been mentioned by researchers in an informal setting and are thus of high interest to the research community. Researchers maintaining these personal blogs often post about interesting papers which introduce new and interesting problems in emerging fields as well as papers with an established position in the field. If the researcher finds a paper interesting enough to blog about, they are more likely to cite it in their formal research work later, and other researchers reading the blog might do the same. Selecting such papers would therefore make sure that my dataset is not sparse.

Secondly, it is reasonable to assume that personal blogs are opinionated in nature. Such blogs can act as a good source for additional sentiment above and beyond the citation sentiment occurring in scientific publications, which I have discussed so far. This would lead to a corpus which is more balanced with respect to sentiment, and is not biased towards citations without sentiment.

While this approach might make more citations available which carry sentiment, it is not a random sample and the resulting dataset may not be representative of the literature in the field of Computational Linguistics and Natural Language Processing in general.

To obtain a list of papers in the ACL anthology which have been mentioned in blogs, I used a three-step approach.

- I used the list of blogs related to CL available on the ACL Anthology website[7]. This list is publicly editable and at the time of the data collection (2009), it contained 23 entries.
- If a blog post contained a link to another CL or NLP related blog, I added it recursively to the list. At the end of this step, the total number of blogs in my list was 42.
- I then used the Google Blogs Search Engine[8] to search for various queries related to CL and NLP. Additionally, I searched for CL conferences by name and acronyms. These acronyms were obtained from the ACL anthology website and are listed in Appendix B. Only the top 100 search results for each query were examined. I also searched for all blog posts which contained any link to the ACL anthology website using the `site:` operator[9].

[7] `http://www.aclweb.org/aclwiki/index.php?title=Blogs`

[8] `http://www.google.com/blogsearch`

[9] Adding `site:example.com` shows search results linking only to `example.com`

The final list consists of 62 blogs, which are given in Appendix D. At the time of collection, these blogs contained a total of 12,575 posts. Within these posts, I wanted to mark all mentions of any ACL paper. For the purpose of this work, I define a *paper mention* in a blog post as one of the following:

Direct Any HTML link which points to a paper on the ACL anthology website.

Indirect Any HTML link which points to the authors' personal page, any bibliographic index (Citeseer, ACM etc.) or university-hosted research group site. This includes links pointing to a paper not hosted on the ACL site.

Text Any text which refers to a paper but not in the form of a hyperlink. This may consist of the a paper title, method name or other similar referring expression such as anaphora.

This annotation was performed using a custom-developed annotation tool. I chose to treat blogs with more that 500 posts (there were two of these) differently from those (60) with fewer than 500 posts. The latter set of blogs had a total of 6,357 posts. Each of these posts was manually inspected for mentions of any paper in the ACL anthology. Blogs with more than 500 posts were searched using regular expressions in order to save time.

This resulted in a total of 115 blog posts which contained a mention to a paper in the ACL anthology recognised by this procedure. These posts contained a total of 676 paper mentions (184 direct, 93 indirect, and 341 text), referring to 310 unique papers. However, out of these 310, 116 were not contained in the AAN corpus. The remaining 194 papers are present in the AAN corpus will from now on be called *blog papers*.

I wanted to collect citation sentences from all papers within the ACL network which cite blog papers. For this task, I used the citation data in the AAN. 8,736 sentences in the AAN citation data cite one of the blog papers. These sentences come from 2,992 unique citing papers. A representation of the corpora discussed is shown in Figure 3.6, which shows a blog post that mentions an ACL paper cited by three other ACL papers. The annotation of these citations is described in the next section.

3.3 Annotation of Citation Sentiment

The process of citation sentiment annotation consisted of me examining each citing sentence manually and assigning it to a class indicating whether it is positive (p), negative (n), or neutral and objective (o). The guidelines for this annotation are as follows.

3.3.1 Annotation Guidelines

For a target paper P, the criteria for marking negative sentences is presence of either of the following:

- Direct mention of a problem or shortcoming in P.

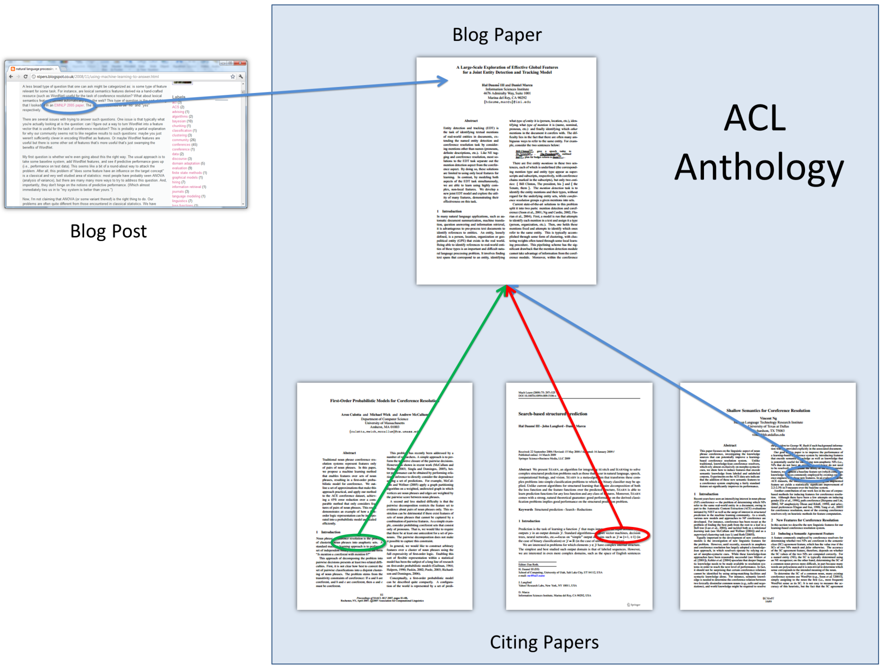


Figure 3.6: Incoming links to a blog paper.

- Comparison of P with some other paper Q, where Q can be the citing paper, such that Q improves upon P by using a better approach.
- Comparison of P with some other paper Q, where Q can be the citing paper, such that Q outperforms P in some objective evaluation.

Some examples of negative citations are given below. Citations to the target papers are shown in bold face.

> The morpological [sic] processing in PairClass (Minnen et al., 2001) is more sophisticated than in **Turney (2006)**.

> Poon and Domingos (2008) outperformed **Haghighi and Klein (2007)**.

> Finally, we show that our contextually richer rules provide a 3.63 BLEU point increase over those of **(Galley et al., 2004)**.

> While the model of **(Matsuzaki et al., 2005)** significantly outperforms the constrained model of (Prescher, 2005), they both are well below the state-of-the-art in constituent parsing.

There are cases where more than one sentiment is present towards the target citation in a single sentence, such as in the 4th and the 5th example above. The use of connectives such as *however*, *although* and *while* in these sentences is of particular importance. Polanyi and

Zaenen (2006) call them discourse-based contextual valence shifters and state that such valence shifters are used for neutralising the effect of the sentiment mentioned earlier in the discourse. The intended sentiment in such cases is more likely to be the last mentioned one (MacRoberts and MacRoberts, 1984). For sentences containing contradictory sentiments separated by a valence shifter, I therefore decided to annotate the last sentiment mentioned in the sentence.

The criteria for marking positive sentences for a target paper P is presence of the either of the following:

- Direct mention of a positive attribute or advantage of P. This includes references to methods and techniques with attributes like *efficiency* and *success*.
- Comparison of P with some other paper Q, where Q can be the citing paper, such that P improves upon Q by using a better approach.
- Comparison of P with some other paper Q, where Q can be the citing paper, such that P outperforms Q in some objective evaluation measure.

Some examples of positively annotated citation sentences are given below.

> An efficient algorithm for performing this tuning for a larger number of model parameters can be found in **Och (2003)**.

> **Dasgupta and Ng (2007)** improves over (Creutz, 2003) by suggesting a simpler approach.

> Here we choose to work with stupid backoff smoothing (**Brants et al., 2007**) since this is significantly more efficient to train and deploy in a distributed framework than a contextdependent [sic] smoothing scheme such as Kneser-Ney.

> State-of-the-art measures such as BLEU (**Papineni et al., 2002**) or NIST (Doddington, 2002) aim at measuring the translation quality rather on the document level1 [sic] than on the level of single sentences.

> Decision lists have already been successfully applied to lexical ambiguity resolution by (**Yarowsky, 1995**) where they perfromed [sic] well.

Citation sentences which contained no positive or negative sentiment towards the target paper were annotated as objective. Some examples of these objective or neutral citations are given below.

> Statistical machine translation (SMT) was originally focused on word to word translation and was based on the noisy channel approach (**Brown et al., 1993**).

> Exploratory work in this vein was described by **Hajic et al. (1999)**.

> Discriminative parsing has been investigated before, such as in Johnson (2001), **Clark and Curran (2004)**, Henderson (2004), Koo and Collins (2005), Turian et al.

> Finally, our newly constructed parser, like that of (**Collins 1997**), was based on a generative statistical model.

> Although we have argued (section 2) that this is unlikely to succeed, to our knowledge, we are the first to investigate the matter empirically.11 [sic] The best-known MT aligner is undoubtedly GIZA++ (Och and Ney, 2003), which contains implementations of various IBM models (**Brown et al., 1993**), as well as the HMM model of Vogel et al.

It should be noted that while the 5th example contains positive sentiment towards GIZA++, which is referred to as the 'best-known' aligner, this sentiment is not directed at the target paper (i.e. *Brown et al., 1993*). The sentence is therefore annotated as objective.

Using the guidelines described above, I annotated all sentences containing citations to the 194 blog papers. Some properties the resulting corpus are described in the following section.

3.3.2 Corpus Statistics

The citation sentiment corpus contains the aforementioned 8,736 sentences. There are 829 positive citation sentences, 280 negative ones, and the remaining 7,627 are objective. Table 3.1 gives the distribution of classes in the corpus, which is heavily skewed, with 87% being objective and only around 13% carrying any sentiment. This is in line with earlier work (Spiegel-Rosing, 1977; Teufel et al., 2006b) which shows that citations are mostly neutral in sentiment.

Class	Count	Percentage
o	7,627	87.3%
n	280	3.2%
p	829	9.5%

Table 3.1: Distribution of classes.

Assigning sentiment to a citation is a subjective task and that annotators can be expected to be inconsistent in attaching sentiments to the same sentence at different times. To measure the consistency of annotation, the commonly used method is to ask multiple annotators to annotate the corpus. Their inter-annotator agreement can then be calculated using measures like Cohen's κ (Cohen et al., 1960).

Carletta (1996) discusses why κ is a good measure for determining inter-annotator agreement. This value of κ can be calculated by

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)}, \tag{3.1}$$

where $P(A)$ is the observed agreement. If an annotation with c classes with a total of n observations by two annotators, it is defined as

$$P(A) = \frac{1}{n} \sum_{i=1}^{c} x_{ii} \tag{3.2}$$

and $P(E)$ is the expected agreement which can be calculated as

$$P(E) = \frac{1}{n^2} \sum_{i=1}^{c} x_{i+} x_{+i} \tag{3.3}$$

where x_{i+} is the marginal for row i and x_{+i} is the marginal for column i (McHugh, 2012).

Taking a two-class, two-annotator problem as an example, consider the contingency table for annotators *Alice* and *Bob* given in Table 3.2.

Alice ↓	Class 1	Class 2	← *Bob*
Class 1	x_{11}	x_{12}	
Class 2	x_{21}	x_{22}	

Table 3.2: Contingency table for calculating κ.

Here, $P(A)$ will be the number of times *Alice* and *Bob* agree, divided by the total number of annotations. It is given by:

$$P(A) = \frac{x_{11} + x_{22}}{x_{11} + x_{12} + x_{21} + x_{22}} \tag{3.4}$$

Similarly, $P(E)$ is modelled by calculating the probabilities of each annotator randomly choosing a class. $P(E)$ will thus be the probability of both annotators agreeing by chance. It can be calculating using:

$$P(E) = \frac{(x_{11} + x_{12})(x_{11} + x_{21}) + (x_{12} + x_{22})(x_{21} + x_{22})}{(x_{11} + x_{12} + x_{21} + x_{22})^2} \tag{3.5}$$

Values of κ ranges between -1 and 1. A κ value of 0 means that the agreement is only as expected by chance, a value of 1 means perfect agreement, and a value of -1 means perfect disagreement. Various scales have been proposed for interpreting the κ values (Krippendorff, 1980; Rietveld and Van Hout, 1993; Green, 1997; Eugenio and Glass, 2004), the strictest one being Krippendorff's. According to Krippendorff, an annotation is only considered to be reliable when κ value is greater than or equal to 0.8. For $0.67 < \kappa < 0.8$, the annotation agreement is sufficient for tentative conclusions to be drawn.

When more than one person performs the annotation, the agreement measure is called *inter-annotator agreement*. It is generally preferable to have multiple annotators perform the annotation. However, annotation of the complete corpus by more than one annotator was too expensive for me due to the large size of my corpora. For this reason, I use the *intra-annotator agreement*.

The intra-annotator agreement is the agreement of the annotations of a single annotator after a period of time long enough for them to have forgotten the original annotation (Carletta, 1996). It measures the stability of a corpus and a high intra-annotator agreement indicates that the annotation can be reproduced almost identically from scratch each time. This supports the claim that the annotation is well-defined and repeatable.

Therefore, I report the intra-annotator agreement for both the Citation Sentence Sentiment and the Citation Context corpora as they are larger in size. However, I calculate inter-annotator agreement on the smaller Citation Significance corpus as well as for a small subset of the Citation Sentence Sentiment corpus.

I re-annotated the Citation Sentence Sentiment corpus after 4 months and found the agreement with my earlier annotation to be $\kappa = 0.89$ ($N = 8736; k = 1; n = 3$), where N is the total number of annotations, n is the number of classes and k is the number of annotators. This level of intra-annotator agreement indicates a stable annotation according to Krippendorff (1980).

To calculate the inter-annotator agreement, I separated a sub-set of 100 randomly selected citation sentences and asked two experts and two non-experts to assign labels to each of these sentences. All annotators were Computer Science or Economics graduates, and I considered anyone who had been trained in NLP and/or Computational Linguistics as an expert. All annotators were given the same instructions on how to assign labels to the sentences (c.f. Appendix E). The agreement between the annotators (including myself) was found to be $\kappa = 0.675$ ($N = 100; k = 5; n = 3$), which is acceptable for tentative conclusions to be drawn according to Krippendorff's stricter interpretation.

Since the non-experts had not been trained to evaluate the domain under consideration, the agreement of my annotation with that of the experts ($\kappa = 0.786; N = 100; k = 3; n = 3$) was found to be greater than the agreement with the non-experts ($\kappa = 0.649; N = 100; k = 3; n = 3$). The pair-wise results of the inter-annotators agreements are available in Table 3.3. Upon analysis, it was found that my agreement with *Expert 2* was low because they considered the 'use' of algorithms and techniques presented in a cited paper as positive. A similar problem has been discussed by Teufel et al. (2006b) where the annotators had difficulty in distinguishing whether a tool or technique in the cited paper is merely being 'used', or has provided a deeper intellectual basis for the citing author. This ambiguity was one of the major sources of the low agreement between myself and *Non-Expert 2* as well. Another source was the phrase *state-of-the-art*, which was considered by *Expert 2* to be neutral rather that positive.

The Citation Sentence Sentiment corpus is publicly available at `http://www.cl.cam.ac.uk/~aa496/citation-sentiment-corpus/`.

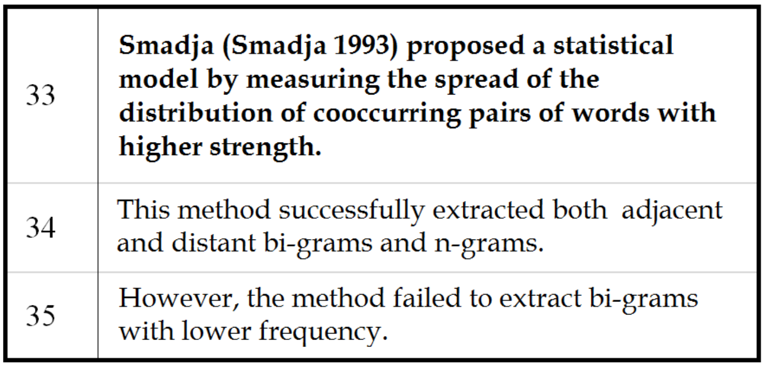
	Author	*Expert 1*	*Expert 2*	*Non-Expert 1*
Expert 1	0.90			
Expert 2	0.75	0.71		
Non-Expert 1	0.82	0.77	0.68	
Non-Expert 2	0.64	0.55	0.49	0.58

Table 3.3: Pairwise inter-annotator agreement κ ($N = 100; k = 2; n = 3$).

3.4 The Citation Context Corpus

We saw in the previous section that most citation sentences are neutral with respect to sentiment. The dominant assumption in current citation identification methods (Councill et al., 2008; Radev et al., 2009) is that the information present in the citation sentence represents the true sentiment of the author towards the cited paper. One of the possible reasons for this assumption might be that it is difficult to determine the relevant context of a citation, whereas identification of the citation sentence is substantially easier.

As argued by Teufel (2010), much sentiment about a citation is to be found in the discourse context. If only the citation sentence is used, sentiment contained in that part of the citation context which is outside the citation sentence, is lost. This is particularly true for criticism, as researches tend to hedge their opinions due to social motivations (Ziman, 1968).



33	**Smadja (Smadja 1993) proposed a statistical model by measuring the spread of the distribution of cooccurring pairs of words with higher strength.**
34	This method successfully extracted both adjacent and distant bi-grams and n-grams.
35	However, the method failed to extract bi-grams with lower frequency.

Figure 3.7: Example annotation of a citation context.

An example of a snippet of a target paper is given in Figure 3.7, where the first column shows the sentence number and the second one shows the text. In the example, we can see that the citation sentence which contains the formal citation *Smadja 1993* is neutral, i.e., it carries no sentiment towards the target citation. However, things are different if we consider the context. The sentence following the formal citation (34) describes the method of the target paper as *successful*. Furthermore, sentence 35 specifies a shortcoming of the method by stating that it *failed* a certain task. While the sentence with the formal citation is objective, the sentences in its context carry both positive and negative sentiment respectively. One possible reason for this sentiment shift may be the intended 'sweetening' of criticism by the author. In other words, while there is no sentiment present towards the paper in the formal citation sentence, the text carries sentiment in the succeeding sentences (MacRoberts and MacRoberts, 1984). I hypothesise that, in general, much sentiment lies outside the sentence with the formal citation and in the citation context.

To provide empirical evidence for this hypothesis, I decided to annotate *all* sentences in the entire paper which carry sentiment towards the target citation. This was done by examining each sentence of the citing paper and assigning a sentiment class to those sentences which carry any mention of the target paper. These mentions are referring expressions of various forms listed below.

Anaphora Grammatical substitute used in place of the cited paper such as *this*, *these*, *he*, *she* and *they*.

Author names Last name of the primary author, sometimes found with an apostrophe. Existing citation detection mechanisms tend to overlook these names Pham and Hoffmann (2004); Radev et al. (2009); Qazvinian and Radev (2010).

Acronyms Condensed versions of methods from the target paper such as SCL for *Structural Correspondence Learning.*

"Work nouns" Words used to refer to the target work such as *approach*, *system*, *method*, *technique*, etc.

Method names Frequently occurring method names such as *Xerox tagger*, *Collin's parser*, etc.

This means that the sentences of a citing paper can be divided into three categories.

1. Sentences which mention the target paper using a formal Harvard style citation.
2. Sentences which mention the target paper in any form other than as a formal citation (c.f. list above).
3. Sentences which do not mention the target paper at all.

My definition of *context of a citation* includes the first and second category, and excludes the third one[10]. In comparison to previous work in citation sentiment detection, the novelty of my work concerns the additional treatment of sentences of category 2. If my hypothesis is correct, sentiment detection should be improved by the discovery of *all* existing sentiment towards the target paper that is expressed in the citing paper.

In order to incorporate the citation context into the analysis of citation sentiment, I first needed to identify the sentences in category 2. The citation sentiment dataset was too large to allow for manual annotation of context for each citation to each blog paper due to time constraints. Therefore, I selected a subset of 20 target papers with the the highest number of citations. These 20 target papers correspond to approximately 20% of citations in the original dataset. 852 unique citing papers mention the 20 target papers in a total of

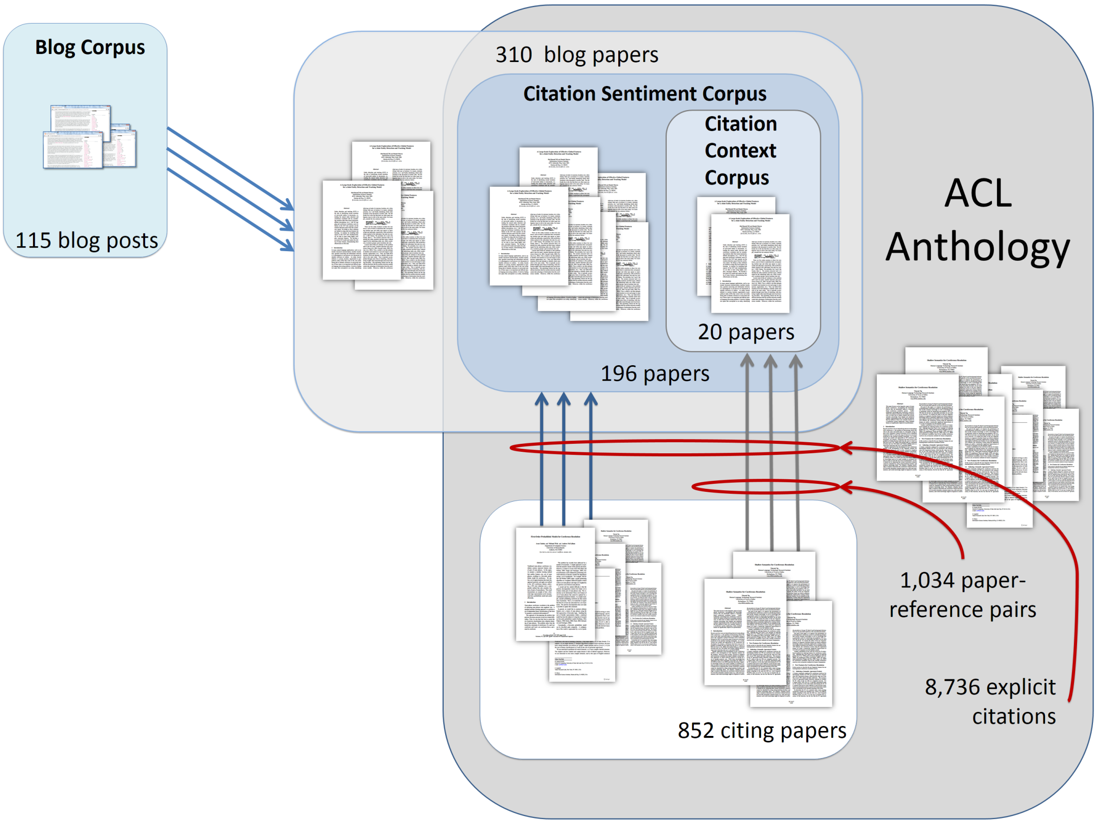


Figure 3.8: Relationship between blogs and papers in ACL anthology

1,741 sentences. Some citing papers cite more than one target paper, so the total number of paper–reference pairs is 1,034.

An overview of the set overlap relations between the papers can be viewed in Figure 3.8. We can see that the blog corpus contains 115 blog posts. These posts cite 310 papers, i.e. the blog papers. 196 of these blog papers are in the ACL anthology. The AAN contains 8,736 formal citations to the 196 blog papers. These formal citations form the Citation Sentiment Corpus. A subset of 20 target papers has been selected from the blog papers for the Citation Context corpus. The figure does not show all existing citations links; ACL internal citations not involving these 852 citing papers have been omitted.

The target papers are listed in Table 3.4 along with the number of citation sentences mentioning them as well as the number of unique paper–reference pairs. In the next section, I describe the annotation of these papers in detail.

3.5 Annotation of Citation Context

All sentences in each of the 852 citing papers are annotated with the sentiment towards each one of the 20 target papers. This annotation was performed by one annotator only

[10] A similar scheme has been proposed by Qazvinian and Radev (2010), but they do not address the problem of citation sentiment analysis and use a much smaller dataset.

ACL-ID	Mentions	P–R Pairs	ACL-ID	Mentions	P–R Pairs
P02-1053	170	95	J96-2004	155	124
J90-1003	136	102	W04-1013	118	50
J93-1007	112	70	W02-1011	107	68
P04-1035	95	62	P90-1034	95	66
N03-1003	90	51	A92-1018	83	62
C98-2122	79	36	W05-0909	74	58
P04-1015	69	33	N06-1020	60	28
W06-1615	53	16	N04-1035	53	31
D07-1031	50	13	P04-1041	47	21
P07-1033	47	18	P05-1045	46	30
			Total	1,739	∗1,034

∗ Corresponding to 852 unique citing papers.

Table 3.4: List of papers for citation context annotation with the frequency of mentions and unique paper–reference (P–R) pairs.

(me).

In Section 3.2, I used a 3-class scheme for annotating citation sentences with sentiment. However, this scheme is unable to identify sentences which are excluded from the context. Therefore, I extended this scheme by adding a fourth class, *x*. This class was used to assign a class to those sentences which do not contain any direct or indirect mention of the target citation. The classes for citation context annotation are thus as follows.

p: Refers to citation and is *p*ositive

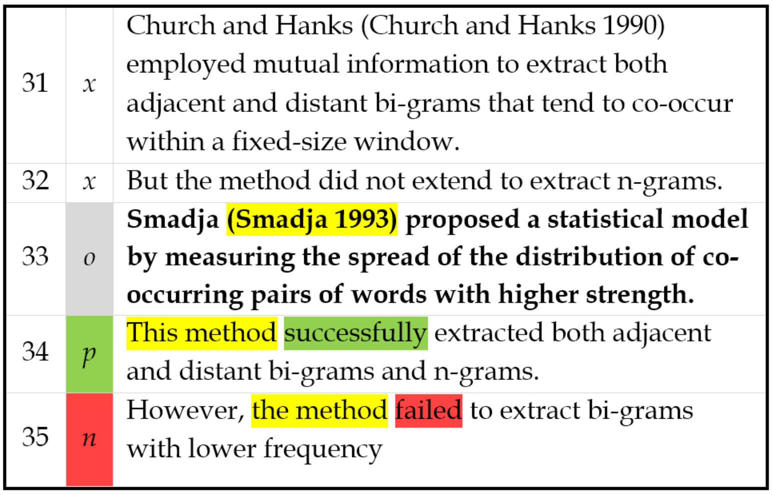
n: Refers to citation and is *n*egative

o: Refers to citation but is *o*bjective or neutral

x: Does not refer to citation (e*x*cluded from the context)

If we take a look at the example given in Figure 3.9 (which we have encountered before on Page 39 as Figure 3.7), sentence 33 uses a formal citation to mention the target paper but carries no sentiment towards it. It is thus annotated as objective. The rest of the sentences use other linguistic expressions to refer to the citation, marked in yellow. Sentence 34 uses the phrase "*This method*" to refer to *Smadja (1993)* in a positive manner and is thus annotated as postivie. Sentence 35 describes a shortcoming of "*the method*" and is thus marked as negative.

Note that as a consequence of my decision to perform target-paper annotation, sentences which carry sentiment but refer to a different citation from the respective target papers were marked as excluded from the context. In the example, sentence 32 specifies a shortcoming of *Church and Hanks (1990)*, which was annotated *x* as it does not refer to the target citation.



31	x	Church and Hanks (Church and Hanks 1990) employed mutual information to extract both adjacent and distant bi-grams that tend to co-occur within a fixed-size window.
32	x	But the method did not extend to extract n-grams.
33	o	**Smadja (Smadja 1993) proposed a statistical model by measuring the spread of the distribution of co-occurring pairs of words with higher strength.**
34	p	This method successfully extracted both adjacent and distant bi-grams and n-grams.
35	n	However, the method failed to extract bi-grams with lower frequency

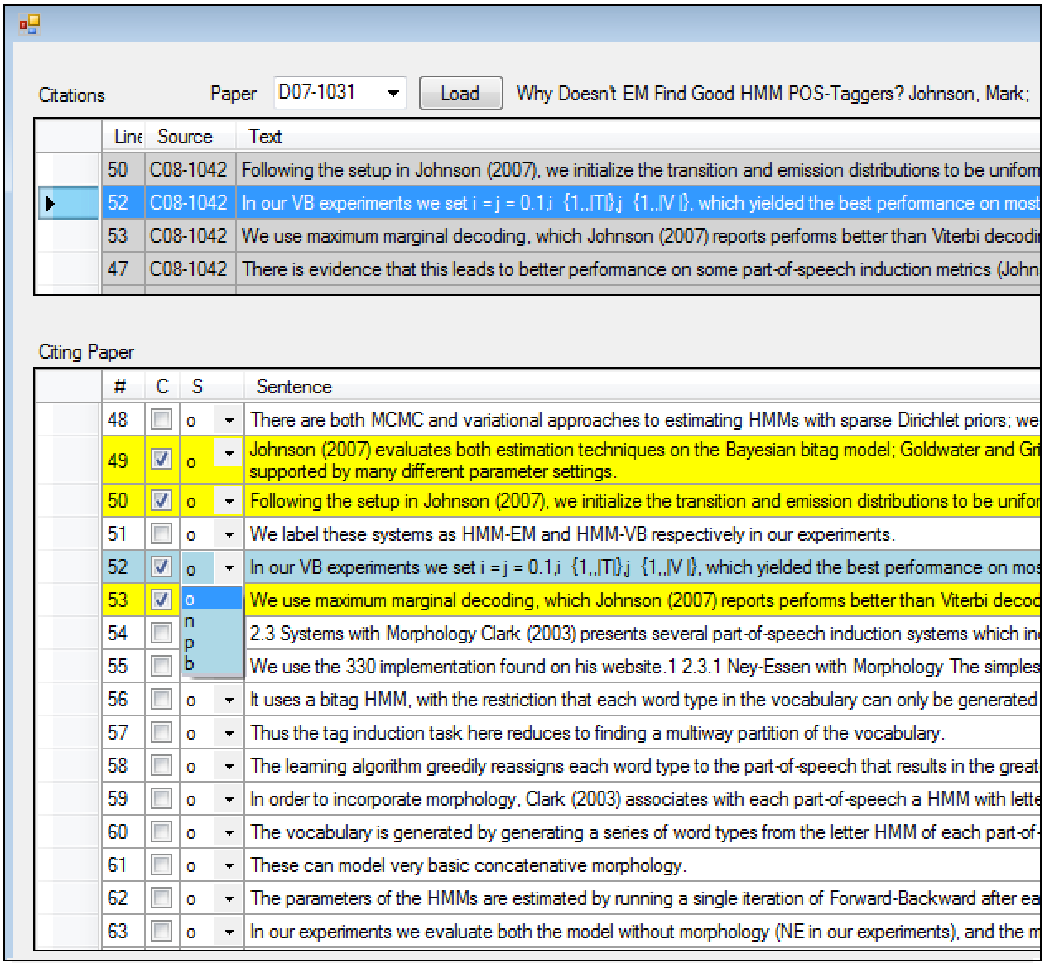
Figure 3.9: Example annotation of a citation context.

To speed up the annotation process, I again developed a customised annotation tool. A screenshot of the interface is given in Figure 3.10. The interface of the tool consists of two main panels. The upper panel displays a list of citations to a specified paper. Selecting the citation loads the text of the citing paper as a list of sentences in the lower panel. The lower panel shows the line number and a checkbox for each line, which allows the user to annotate whether that line contains a mention of the cited paper. The user assigns a class to each sentence through a listbox.

With this tool, it was possible for myself to complete the annotation process in approximately 60 hours over a period of one month. A total of 203,803 sentences from 852 unique papers were annotated. To measure intra-annotator agreement, two target papers (W06-1615 and D07-1031) were randomly chosen for reannotation. These target papers incur incoming citations from 29 papers. These paper–reference pairs comprised $6,051$ sentences in total and I performed the reannotation after 12 months. The intra-annotator agreement was $\kappa = 0.85$ $(N = 6051; k = 1; n = 4)$. This indicates that the annotation is stable according to the strict interpretation by Krippendorff.

I was also interested in measuring whether or not the sentences included in the citation context were stable. This can be measured by ignoring the sentiment information from the classes. For this purpose, I collapsed the o, n and p classes into a single class o_n_p. This reduces the classes to a binary classification scheme $(x$ OR $o\_n\_p)$. This resulted in agreement of $\kappa = 0.87$ $(N = 6051; k = 1; n = 3)$, showing that this annotation is also stable.

The resulting annotated corpus contains 3,760 sentences which mention one of the target papers, only 1,739 of which contain formal citations. Let us now consider how the total amount of sentiment changes when the context is taken into account. We can do this for the 1,739 citations for which we have annotations available in both the citation sentiment and the citation context corpus. Table 3.5 compares the sentiment classes of the formal citation sentences with classes of all sentences in the citation context. The figures show that by including the citation context, the number of sentences containing negative sentiment which is recoverable increases from 86 to 368, which corresponds to an increase of 328% per citation context. Similarly, the positive sentiment increases from


Citations Paper D07-1031 Load Why Doesn't EM Find Good HMM POS-Taggers? Johnson, Mark;

Line	Source	Text
50	C08-1042	Following the setup in Johnson (2007), we initialize the transition and emission distributions to be uniform
52	C08-1042	In our VB experiments we set i = j = 0.1,i {1,.|T|},j {1,.|V |}, which yielded the best performance on most
53	C08-1042	We use maximum marginal decoding, which Johnson (2007) reports performs better than Viterbi decodi
47	C08-1042	There is evidence that this leads to better performance on some part-of-speech induction metrics (John

Citing Paper

#	C	S	Sentence
48	□	o	There are both MCMC and variational approaches to estimating HMMs with sparse Dirichlet priors; we
49	☑	o	Johnson (2007) evaluates both estimation techniques on the Bayesian bitag model; Goldwater and Gri supported by many different parameter settings.
50	☑	o	Following the setup in Johnson (2007), we initialize the transition and emission distributions to be unifor
51	□	o	We label these systems as HMM-EM and HMM-VB respectively in our experiments.
52	☑	o	In our VB experiments we set i = j = 0.1,i {1,.|T|},j {1,.|V |}, which yielded the best performance on mos
53	☑	o	We use maximum marginal decoding, which Johnson (2007) reports performs better than Viterbi decod
54	□	n	2.3 Systems with Morphology Clark (2003) presents several part-of-speech induction systems which in
55	□	p b	We use the 330 implementation found on his website.1 2.3.1 Ney-Essen with Morphology The simples
56	□	o	It uses a bitag HMM, with the restriction that each word type in the vocabulary can only be generated
57	□	o	Thus the tag induction task here reduces to finding a multiway partition of the vocabulary.
58	□	o	The learning algorithm greedily reassigns each word type to the part-of-speech that results in the great
59	□	o	In order to incorporate morphology, Clark (2003) associates with each part-of-speech a HMM with lette
60	□	o	The vocabulary is generated by generating a series of word types from the letter HMM of each part-of-
61	□	o	These can model very basic concatenative morphology.
62	□	o	The parameters of the HMMs are estimated by running a single iteration of Forward-Backward after ea
63	□	o	In our experiments we evaluate both the model without morphology (NE in our experiments), and the m

Figure 3.10: Screenshot of the annotation tool.

146 to 292, which corresponds to an increase of 100%. However, the total number of sentences has also increased from 1,739 to 3,760. Another indication of change is thus change with respect to sentences considered. This shows an increase of 100% in the number of sentences with negative sentiment but no increase with respect to the positive sentiment per sentence considered.

	Formal Only	With Context	Std. Dev. (σ)	Change per Sentence Considered	Change per Citation Context
o	1,509	3,100	795.5	-6%	$+105\%$
n	86	368	166	$+100\%$	$+328\%$
p	146	292	73	$\pm0\%$	$+100\%$
$Total$	1,739	3,760	1010.5	-	-

Table 3.5: Distribution of classes.

These numbers provide evidence for the hypothesis that much sentiment towards a citation is present in the citation context outside the immediate citation sentence. Ignoring this context would therefore lead to incomplete coverage of the true sentiment. Therefore, an analysis of the entire citation context is necessary for the task of detection of sentiment in citations.

After the annotation process, the user may profit from seeing an overview of the distribution of sentiment for each citing paper in one line. This representation can be thought of as the sentiment fingerprint for the target paper. I therefore developed a viewing mode in the annotation tool using HTML and JavaScript. In this mode, one page corresponds to one target paper and lists citing papers line by line.

Figure 3.11 shows the annotation for the target paper with ACL-ID J93-1007. Here, each line corresponds to a citing paper and each square in that line is a sentence in the citing paper. Rows are sorted by increasing publication date. The color of each square represents the sentiment of the sentence. A legend of these colors is visible below the title. Hovering the mouse over the first square in the line displays the paper metadata for that line. Hovering the mouse over any other square displays the text content of the corresponding sentence. Clicking on the checkboxes at the top hides or shows squares of the corresponding sentiment. This data is publicly available at `http://www.cl.cam.ac.uk/~aa496/citation-context-corpus/`.

3.6 Annotation of Citation Significance

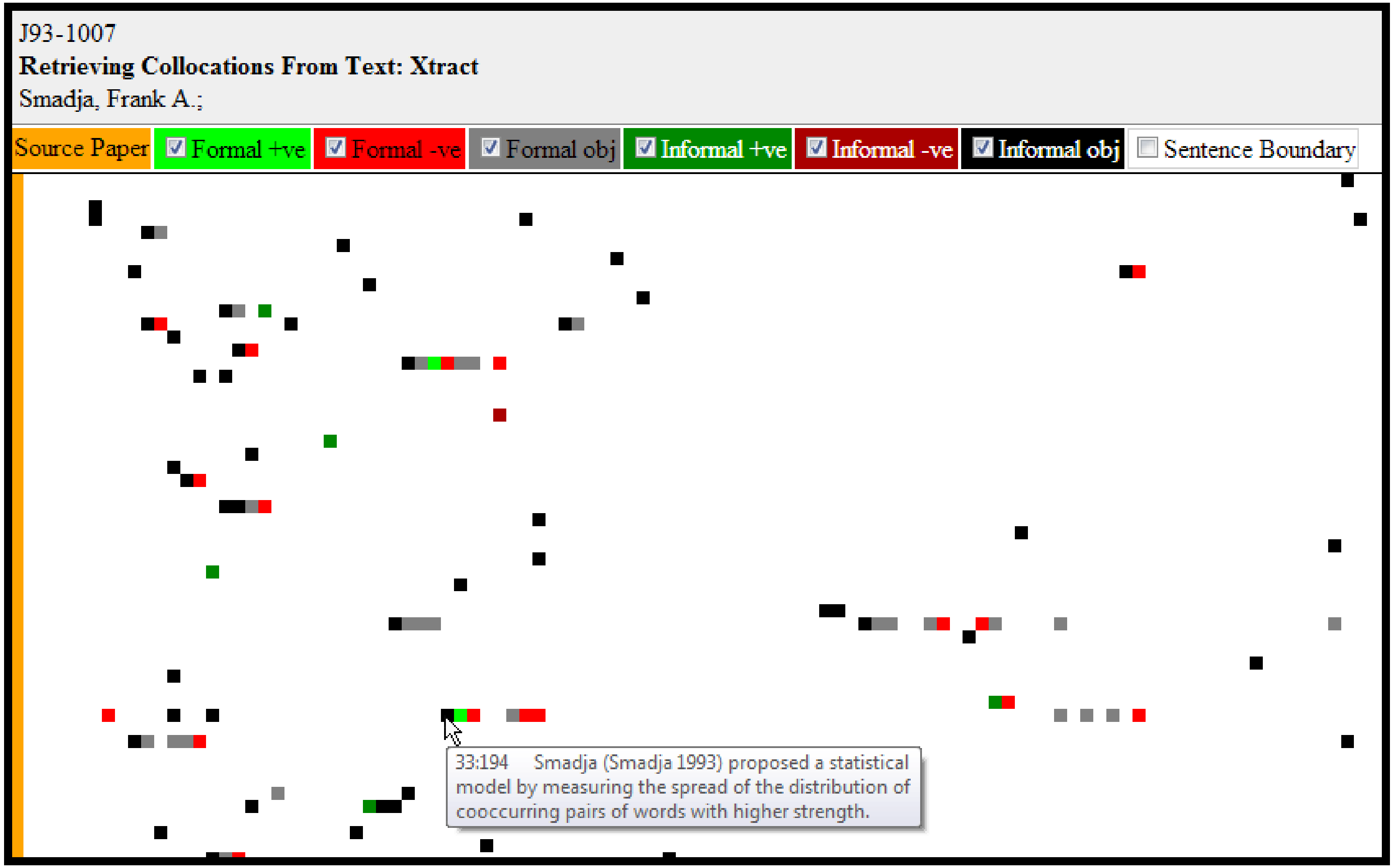
As mentioned earlier in Chapter 1, raw citation counts and other more complex bibliometric measures such as h-Index (Hirsch, 2005) and Impact Factor (Garfield et al., 1964) treat all citations as equal. In these algorithms, no distinction is made between citations to a paper central to the citing text, and citations which have been mentioned in passing only. This contradicts Ziman's (1968) observation about the prevalence of polite but insignificant citations in the scientific community.

The new task I propose here is to distinguish between citations made only in passing and citations which are significant to the content discussed in the cited (target) paper. I approach this as a binary classification problem where the significance of a cited paper is predicted by examining the text of the cited paper. I call this task *Citation Significance Detection*.

Since the task is new, no annotated corpus exists and again I had to construct my own corpus. For this purpose, I reused the citation significance corpus, as the text needed for this task was already available in the corpus. I manually judged the total contribution of the target paper reference for all paper–reference pairs and assigned one of the following classes to each pair.

Y means that the contributions of the reference were significant in the citing paper.

N means that the contributions of the reference were not significant in the citing paper.



J93-1007

Retrieving Collocations From Text: Xtract

Smadja, Frank A.;

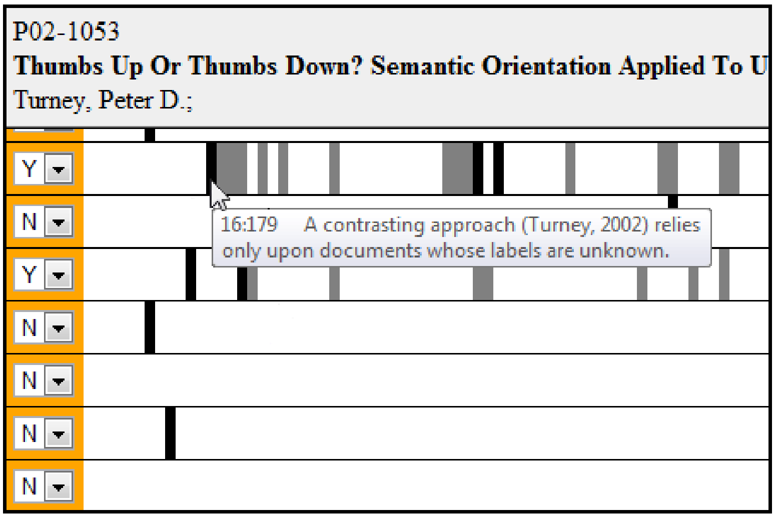
33:194 Smadja (Smadja 1993) proposed a statistical model by measuring the spread of the distribution of cooccurring pairs of words with higher strength.

Figure 3.11: Screenshot of annotation viewing tool.

This required re-reading all sentences in the citing paper that mention the cited paper (i.e. all formal and informal citations) and assigning a class to the citing paper. The annotation unit therefore corresponds to the 1,034 paper–reference pairs mentioned in the previous section. In detail, the decision was based on the following criteria:

- Is the cited paper mentioned only in the literature review? If yes, then assign class N.
- How many times is the cited paper mentioned? The higher this number, the more I was inclined towards class Y.
- How are these mentions distributed in the citing text? The more even the distribution, the more I was inclined towards class Y.
- Would removing the citation sentences result in the loss of significant information from the citing paper? If so, assign class Y.

To simplify the annotation process, I adapted the existing annotation viewer tool described on Page 46. Sentiment information available in the original tool was removed in order to exclude any possible bias this information might introduce to my new annotation task. A view of the tool is given in Figure 3.12. Grey cells represent sentences containing an informal citation to the target paper, whereas black cells show sentences containing formal citations.



P02-1053
Thumbs Up Or Thumbs Down? Semantic Orientation Applied To U
Turney, Peter D.;

Y
N
Y
N
N
N
N

16:179 A contrasting approach (Turney, 2002) relies only upon documents whose labels are unknown.

Figure 3.12: Annotation of significance of seven papers citing target paper *Turney (2002)*.

Out of the 1,034 paper–reference pairs, only (16%) 162 were found to be significant. This seems to confirm Ziman's (1968) view about prevalence of political citations.

For measuring the inter-annotator agreement of the annotation of significance of a reference, three referenced papers were examined by another annotator trained in computational linguistics, who performed the annotation independently from myself. Unlike the citation context corpus, this task was tractable, since only the sentences related to the citation needed to be examined. Inter-annotator agreement was found to be $\kappa = 0.82 (N = 47; k = 2; n = 2)$, showing that the annotation is reliable by Krippendorff's (1980) standard.

3.7 Chapter Summary

In this chapter, I describe the collection and annotation of two corpora with three different annotations. I selected the ACL Anthology Network as my data source as it contains complete paper text as well as manually curated citation links between papers of the ACL Anthology.

The first corpus I created is the *citation sentence sentiment corpus* (c.f. Sections 3.2 and 3.3), which consists of 8,736 sentences containing formal citations to 194 target papers. These target papers are papers in the ACL anthology which are mentioned in 115 CL-related blog posts collected from the internet. The 8,736 sentences were manually annotated using a 3-class (*n*egative/*p*ositive/*o*bjective) annotation scheme. The corpus will be used in the next chapter for the task of sentiment detection in citation sentences.

The second corpus is the *citation context corpus* (c.f. Section 3.4), which consists of the full text of 852 papers which cite the top 20 target papers in the citation sentiment corpus with the highest number of objective citations. The corpus contains 1,034 paper–reference pairs and 203,803 sentences. I examined all 203,803 sentences manually and identified the sentences in the citation context. These context sentences contained formal or informal citations to the target paper and were assigned classes according to their sentiment (n, p, o). The remaining sentences, i.e., those that did not refer to the target papers were classified as being excluded (x) from the context. The annotation confirmed my hypothesis that much sentiment is present in the citation context outside the immediate citation sentence, particularly so for negative sentiment. There was an absolute increase of 328% in the number of sentences containing negative sentiment per citation context, which even if we take into consideration the increase in the overall annotated sentences, still corresponds to a 100% increase for negative sentiment. For positive sentiment, we observe an absolute increase of 100% per citation context, but 0% relative change per sentence considered. This is extremely relevant for the overall task addressed in this dissertation, in light of the sparse nature of non-neutral citation sentiment. This corpus will be used later in Chapter 5 for the task of detecting informal citations and examining their effect on the downstream task of detecting citation sentiment.

I also define the new task of detecting the significance of a target paper. This is a paper-level task where each paper–reference pair is assigned a class Y if the contribution of the target paper in the citing paper in significant, and a class N otherwise. For this task, I reannotated the citation context corpus at a paper-level. This means that annotations of the 1,034 paper–reference pairs were provided. This corpus will be used later in Chapter 6.

For the citation sentiment and citation context annotations, I measured the intra-annotator agreement using Cohen's κ. The annotation was found to be stable/reliable for all three corpora according to Krippendorff's (1980) interpretation, which is the strictest one currently in use in the annotation community. Similarly, for the smaller citation significance annotation, I found that the inter-annotator agreement indicates that the annotation is reliable in Krippendorff's interpretation. The following chapters will describe the experiments I performed using these annotation.

Chapter 4

Sentence-Based Citation Sentiment Classification

This chapter describes my implementation of the automatic detection of sentiment in citation sentences in scientific text. I present a method to predict the label of each citation sentence as either positive (p), negative (n) or objective/neutral (o). I approach this problem as a classification task. In the field of machine learning, classification is the problem of assigning a label, from an existing set of labels, to a new observation. In supervised machine learning, this is achieved by analysing a training set of data for which the labels are already known and creating a statistical model from this information. The works described in this chapter has been published as Athar (2011).

This chapter is organized as follows. Section 4.1 describes the preprocessing performed on the corpus described in Section 3.2. Section 4.2 lists the different feature sets used in the classifier. The classification algorithm I use for learning a model from these features is described in Section 4.3. The evaluation metrics are described in Section 4.4 and the results are given in Section 4.5.

4.1 Data Preprocessing

I use the citation sentiment corpus described earlier in Section 3.2. It contains 8,736 citation sentences, which have been labelled as positive, negative or objective. I separated 736 sentences for use as a training set, and an equal number of sentences for a development-test set to achieve an approximate 1:10 partitioning of the corpus. The remaining corpus of 7,264 citations was used as the test set. This data contains citations to 196 target papers by 150 unique first authors. With respect to sentiment, it contains 244 negative, 743 positive and 6,277 objective citations.

Some of the authors cited in my corpus are well-established in their field. Such individuals have a record of publishing seminal works which is widely used and consequently, cited positively. Since citations in Computational Linguistics follow the Harvard style, the last names of these authors becomes a part of the sentence text. When using the traditional bag-of-word model on such text, these names are tokenised along with other words in the sentences. In this case, such names trivially correlate with the sentiment of citations and thus, these names may induce a bias in the classifier. In other words, if the classifier learns that citation sentences containing a particular name are mostly positive, it will predict a positive label for *every* work by the same author, which will lead to the wrong results in those cases where it is not cited positively.

In order to remove this lexical bias, I identified the text of the formal citations using regular expressions, and replaced it with a special token (<*CIT*>). For example, the citation

> **Turney (2002)** starts from a small (2 word) set of terms with known orientation (excellent and poor).

was changed to

> <**CIT**> starts from a small (2 word) set of terms with known orientation (excellent and poor).

A similar approach is used by Read (2005) and Go et al. (2009) in their work on sentiment analysis of tweets. They remove emoticons from their data in order to force their classifier to learn from other features. In the next section, I describe different features that I use in my classifier.

4.2 Classifier Features

For developing a sentence-based sentiment classifier, I follow the state-of-the-art sentiment classification methods in using a machine learning framework (Pang et al., 2002; Wilson et al., 2009; Maas et al., 2011). This section describes different features I use in the framework. Their performance is then evaluated in Section 4.5.

4.2.1 *n*-grams

An *n*-gram is a contiguous sequence of words in a given text. The character n refers to the length of this sequence. If $n = 1$, the sequence is called a *unigram*, if $n = 2$ or 3, it is called a *bigram* or *trigram*, respectively. For the simple sentence *"The results were good"*, the *n*-grams are as follows:

unigrams The – results – were – good

bigrams The results – results were – were good

trigrams The results were – results were good

4-grams The results were good

N-grams of length 1 and 2 have been shown to perform well in existing sentiment classification tasks for movie reviews (Pang et al., 2002). For my experiments, I restrict myself to n-grams of length 1 to 3. The reason for including trigrams is to capture (parts of) technical terms, which play a large role overall in scientific text Justeson and Katz (1995). Many technical terms, such as *Structural Correspondence Learning*, consist of three words and would be covered by trigrams. While many longer technical terms exist, they can be captured by overlapping trigrams. The n-grams are extracted from the citation sentence text with the help of the WEKA (Hall et al., 2008) toolkit. I use the default parameters of the toolkit for tokenisation and word selection with an IDF (Inverse Document Frequency) weighting scheme. The IDF weight is used in Information Retrieval, and measures how common a term is in the corpus. It is obtained by the following formula:

$$IDF_w = log(\frac{N}{n_w}) \tag{4.1}$$

Here N is the total number of sentences, and n_w is the number of sentences which contain the word w. I chose to use IDF as it produced the best results on the development set. One possible explanation for this might be the fact that IDF diminishes the weights of the n-grams that occur with a higher frequency across all classes in the corpus. Since I do not use a stop-word list, using IDF results in reduction of noise in this feature.

4.2.2 Part-of-Speech Tags

Part-of-speech (POS) tags are commonly used in text classification tasks as they provide a rough-grained mechanism for word sense disambiguation (Wilks and Stevenson, 1998). Theoretically, using POS tags as features should help the classifier distinguish between word forms with different parts-of-speech. In some cases, this is correlated with the word form having different senses. For instance, the word *lead* can be tagged as a noun (*NN*) or a verb (*VBZ*). The POS tag feature would help in distinguishing between the 'noun' sense, which might refer to the metal, from the 'verb' sense as in the following sentence:

> This lead to good results.

Various researchers have examined the effect of using different POS features for sentiment analysis. Hatzivassiloglou and McKeown (1997) work on predicting the semantic orientation of adjectives. Turney (2002) explore extracting POS patterns for sentiment analysis of reviews. Pang et al. (2002) appended the corresponding POS tags to every word and used the resulting structures as features. Wiebe and Riloff (2005) based their features on the existence of modals, adjectives, cardinal numbers, and adverbs.

Another method to exploit the POS information is to use POS n-grams as features (Gamon, 2004; Bekkerman et al., 2007; Kouloumpis et al., 2011). These n-grams may help in providing information about recurring sequences of the quasi-syntactic POS patterns to the classifier (Argamon-Engelson et al., 1998). For instance, POS 2-grams in the sentence *"This lead to good results"* would be "*DT VBP*", "*VBP TO*", "*TO JJ*", and "*JJ NNS*", where presence of *JJ NNS* can indicate subjectivity.

I use both these approaches for adding POS information to my feature set. For my first approach, I follow Pang et al. (2002) and transform each word into a new token by appending its POS tag to the end, separated by a forward slash. For instance, the example sentence discussed above would be converted to:

> This/DT lead/VBP to/TO good/JJ results/NNS ./.

This does not increase the number of tokens per sentence. For my second approach, I add the POS unigrams, bigrams and trigams to my feature set.

4.2.3 Science-Specific Sentiment Lexicon

The idea of using a sentiment lexicon to predict the polarity of words has been explored extensively (Baccianella et al., 2010; Esuli and Sebastiani, 2006a,b; Riloff and Wiebe, 2003; Wiebe and Riloff, 2005; Wilson et al., 2005). It has been found that creating a general purpose lexicon is a difficult task (c.f. Section 2.4) and that existing approaches to sentiment detection are highly topic dependent (Engström, 2004; Blitzer et al., 2007). This means that a sentiment lexicon specialised for a domain might be more helpful than a general lexicon.

For this reason, I include a science-specific sentiment lexicon in the feature set. This lexicon consists of 83 polar phrases which I have extracted manually from a development set of 736 citations which does not overlap the test set or training set. A list of these phrases is given in Appendix F.1. The question is now *how* to combine the information from the lexicon with the other features. Since individual phrases would already be covered by the n-grams described earlier, I follow the sentence-based scoring approach used by Wilson et al. (2009), which collapses all occurrences of negative and positive lexical entries in a sentence into just two features. The value of the first feature is set to *true* if any positive phrase in the list is present in the text to be classified, and to *false* otherwise. Similarly, the value of the second feature is set to *true* if any negative phrase in the list is present in the text to be classified, and to *false* otherwise.

4.2.4 Other Features from Wilson et al. (2009)

To take advantage of the existing work on sentiment analysis, I include some other features from the state-of-the-art system by Wilson et al. (2009). For the task of determining the polarity of phrases, the authors use features which range from word-level to document-level in granularity. However, since my task is the sentence-level classification of citations, I include only the sentence-level features in my classifier. A list of these features is as follows:

Adjectives,
Adverbs: These two features count the number of adjectives and adverbs in the sentence respectively. Adjectives and adverbs have been shown to be very good indicators of subjectivity Hatzivassiloglou and McKeown (1997); Turney (2002); Wiebe and

Riloff (2005).

Subjectivity Clues,
Strong Subjectivity Clues,
Weak Subjectivity Clues: These are three related features dependent on a list of 8,000 subjective words and phrases which have been manually annotated by Riloff and Wiebe (2003) from newspaper articles. They use a binary classification scheme and label a clue as *strongsubj* if it is subjective in most contexts, and *weaksubj* otherwise. The first feature is binary, which is set to `true` if a subjectivity clue is present in a sentences and `false` otherwise. The other features are numeric and their value is set to the number of strong or weak subjective clues present in the sentence.

Cardinal Numbers: This is a binary feature, which is set to `true` if a cardinal number is present in a sentences and `false` otherwise. In earlier work (Wiebe et al., 1999), cardinal numbers have been found to be useful for the task of subjectivity classification as their presence might indicate an objective fact or statement, thus avoiding sentiment.

Modal Auxiliary Verbs: This is a binary feature which is set to `true` if a modal auxiliary verb is present in a sentences and `false` otherwise. Such verbs are normally used for hedging or to assume a stance, which may be an indicator of subjectivity.

Negation: Negation has the effect of reversing the polarity of any subjective phrase that succeeds it. This feature is set to `true` if a negation word is found within the citation sentence, and `false` otherwise.

General Polarity Shifters,
Positive Polarity Shifters,
Negative Polarity Shifters: These three features indicate the presence of particular types of phrases which can shift the polarity of a sentence. General polarity shifters, such as the word *little* in the phrase '*little work*', reverse polarity. Negative polarity shifters, such as *so-called* in '*so-called effort*', make the polarity negative of the negative. Similarly, positive polarity shifters change the polarity of a phrase to positive, for instance, *abate* in '*abate the damage*'.

The lexicons for polarity shifters, subjectivity clues, and negation words are publicly available with their earlier work on the OpinionFinder system (Wilson et al., 2005).

4.2.5 Dependency Structures

Dependency structures describe the grammatical relationships between words. Each structure represents a binary relation between a *governor* (or head) word and a *dependent* word. Dependency structures (or dependencies) are usually presented as triples of the form *relation*(*governor*, *dependent*). Consider the following sentence as an example:

> Our system outperforms competing approaches.

This sentence has 5 tokens which correspond the following triplets.

1. `poss(system, Our)`
2. `nsubj(outperforms, system)`
3. `root(ROOT, outperforms)`
4. `amod(approaches, competing)`
5. `dobj(outperforms, approaches)`

Each triple can also be thought of as a labelled edge from the governor to the dependent where the relation name is the edge label. The root node is identified by the relation *root*. In the example sentence above, the word *outperforms* is the root node, which does not have any incoming edges. The sentence can thus be represented by the dependency graph shown in Figure 4.1.

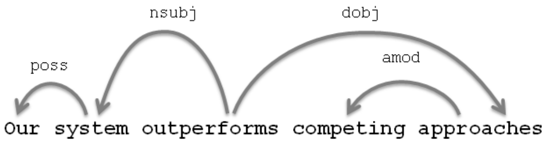


Figure 4.1: An example of a dependency tree.

Dependency relations have been shown to be useful for sentiment analysis and various researchers have focused on using the *adjective-noun* and *modifier-noun* information in their systems (Wilson et al., 2009; Nakagawa et al., 2010; Joshi and Penstein-Rose, 2009).

Following these researchers, my motivation for using dependency structures is to capture the long distance relationships between words. An example of these relations is given in the following citation:

> <CIT> showed that the results for French-English were competitive to state-of-the-art alignment systems.

The dependency graph corresponding to this sentences is shown in Figure 4.2. I use the state-of-art Stanford Typed Dependency Parser (de Marneffe and Manning, 2008) for my implementation, who use a set of 53 grammatical relations to represent a sentence. For example, the relation between the words `showed` and `competitive` in the sentence above is denoted by `ccomp`, the *clausal complement*. The clausal complement of a verb is defined by de Marneffe and Manning (2008) as 'a dependent clause with an internal subject which functions like an object of the verb, or adjective'. In this instance, the target citation placeholder, `<CIT>`, is the nominal subject of the independent clause, shown by the relation `nsubj` between the verb `showed` and `<CIT>`. Similarly, the word `results` is the nominal subject of the subordinate clause. This relation has been highlighted

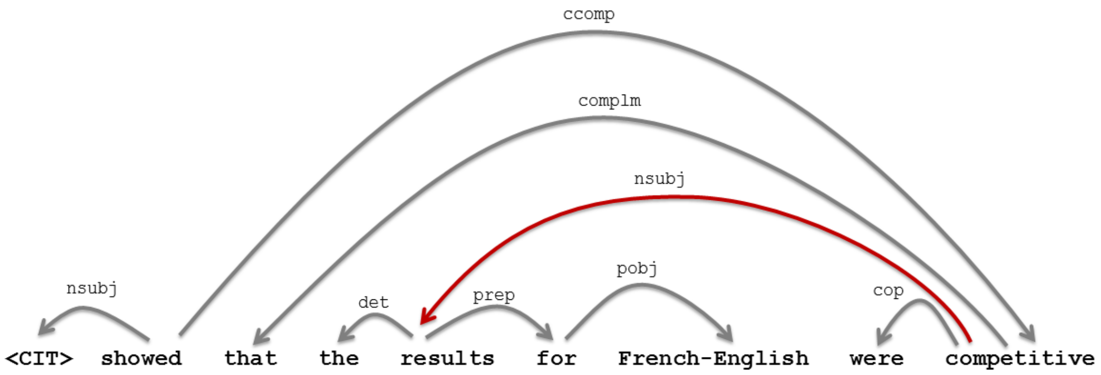


Figure 4.2: An example of long-distance dependency relation.

in the figure as a red arrow between *results* and *competitive*. While this relationship would be missed by trigram features, the dependency triplet captures it in the feature `nsubj_competitive_results`.

To include dependencies in my processing pipeline, I parse all citation sentences and convert each dependency triplet to a single feature. Following Joshi and Penstein-Rose (2009), I concatenate the relation, governor and dependent of each triplet together using an underscore as a delimiter. For example, the triplet `nsubj(competitive, results)` is converted to the feature `nsubj_competitive_results`.

4.2.6 Sentence Splitting

I also test the hypothesis that removing irrelevant polar phrases around a citation improves results. To understand the problem posed by irrelevant polar phrase, consider the following citation sentence:

> This new model leads to significant improvements in quality as measured by <CIT>

We can see that this sentence contains a positive phrase *significant improvements* in the neighbourhood of citation <CIT>. This positive attribute is however not related to the citation and its target is what is described in the sentence as the *new model*. The cited paper is mentioned here as the source of the evaluation metric only and the true sentiment towards the citation is neutral. However, the classifier may label this citation as positive because of the presence of the positive phrase.

This problem can be overcome by splitting the sentence so that textual material logically belonging to one citation is separated from the textual material in its vicinity. In the example above, the clause *measured by <CIT>* should thus be separated from the sentence. The rest of the sentence should be ignored as it does not relate to the citation. This should in principle help the classifier to associate only the correct indicators (features) with the citation.

While there exists some work on detecting the influence of a citation across sentences (Nanba and Okumura, 1999; Kaplan et al., 2009; Qazvinian and Radev, 2010), the problem of detecting it inside a sentence has been largely ignored. At the time of designing my experiments, there was no work which addressed this problem. However, a later paper by Abu-Jbara and Radev (2011) developed the same approach as mine for the task of coherent citation-based summarisation.

I use the parse tree of each sentence and keep only the deepest clause in the subtree of which the citation is a part. All other clauses in the sentence are discarded. In order to find the deepest embedded clause, we note that it is rooted at the lowest S node, i.e., the closest S from the leaf node (<CIT>) representing the citation.

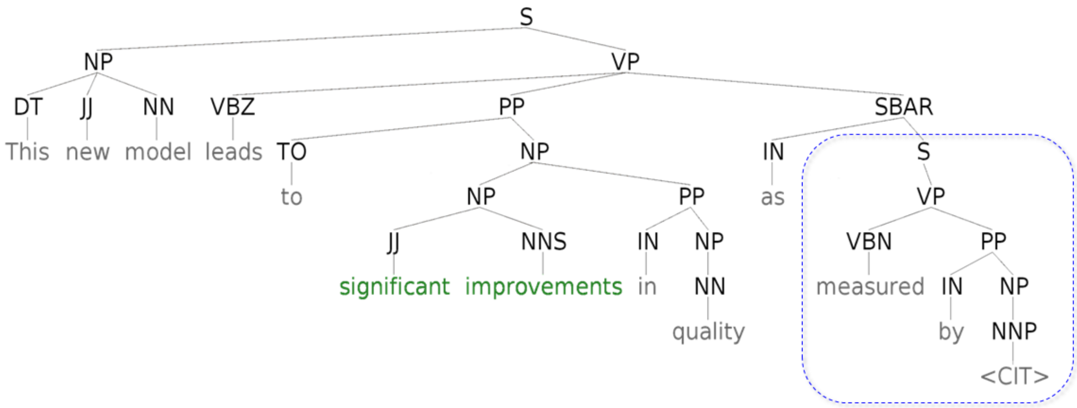


Figure 4.3: An example of parse tree trimming

Figure 4.3 shows the tree structure returned by parsing the example citation using the Stanford Typed Dependency Parser. The trimmed tree in this case, surrounded the blue rounded rectangle, contains only the phrase *"measured by <CIT>"*. The phrase with positive connotations, shown in green, would be discarded after the split and should thus not unduely influence the classification of <CIT>. The trimmed tree is extracted by walking from the citation leaf node (*<CIT>*) towards the root and selecting the subtree rooted at the first sentence node (S) encountered. The pseudocode for this procedure is given in Listing 1.

```
Tree function trim(Tree tree) {
  Node leaf = tree.find("<CIT>");
  Node n = leaf.parent;
  while ( !n.tag.equals("S")) {
      n = n.parent;
  }
  return n.subTree();
}
```

Listing 1: Pseudocode for trimming a parse tree.

While this treatment clearly solves the problem in the example, it assumes that the influence of the citation constitutes a single grammatical fragment. Experimentation is therefore needed to establish whether this fact holds for a majority of actual cases.

4.2.7 Window-Based Negation

There has been much work in handling negation and its scope in the context of sentiment classification (Polanyi and Zaenen, 2006; Moilanen and Pulman, 2007). Detection of both negation and its scope are non-trivial tasks on their own. Das and Chen (2001) use a window-based approach, where they orthographically modify all words within a fixed window which follow a negation word. Councill et al. (2010) use features from a dependency parser in a CRF framework to detect the scope of negation. More recently, Abu Jbara and Radev (2012) propose a similar framework, but with lexical, structural, and syntactic features while solving a shared task for resolving the scope and focus of negation.

Resolving the scope of negation is thus a nontrivial task, which is out of the scope of this dissertation. Hence, I follow the simpler window-based approach used by Pang et al. (2002) (initially proposed by Das and Chen (2001)). All words inside a k-word window of any negation term are suffixed with a token *_neg* to distinguish them from their non-polar versions. I experiment with different values of k in my experiments ranging from 1 to 15. The negation list contains 15 terms (*no,not without, etc.*), and is taken from Wilson et al. (2005). This list is available in Appendix F.2.

An example of a 2-word negation window applied to a sentence is given below.

> Turney's method did not **work_neg well_neg** although they reported 80% accuracy in <CIT>.

In this case, the lexical features of the phrase *work well* can signal the classifier to label the citation as positive. However, if we use the negation window after the term *not*, the lexical features undergo an orthographical change which can help the classifier distinguish between the phrase with negation and the phrase without negation. This may provide a deeper model of negation than presence of negation clues, as described earlier in Section 4.2.4.

The next section describes the classifier I decide to train for detection of sentiment in citation sentences using the features described here.

4.3 Classification

While many classification algorithms exist in the literature, I use the Naive Bayes (NB) and the Support Vector Machine (SVM) (Cortes and Vapnik, 1995) frameworks in my work. The Naive Bayes classifier uses Bayes' rule to estimate the probability a class c given a document d.

$$P(c|d) = \frac{P(c)P(d|c)}{P(d)}, \tag{4.2}$$

where $P(c)$ is the prior probability of any document occurring in class c, $P(d|c)$ is the probability of a document d given it is in class c, and $P(d)$ is the probability of the document d. According to the NB classifier, the document is predicted as belonging to the class c^* which maximises the posterior probability $P(c|d)$. Thus, $P(d)$ does not play any role in predicting the class and Equation 4.2 can be rewritten as:

$$P(c|d) \propto P(c)P(d|c) \tag{4.3}$$

The NB estimates the term $P(d|c)$ by assuming that the document d can be represented as a set of independent features f_1, f_2, ... , f_n. This makes the predicted class c^* to be:

$$c^* = \underset{c \in \mathbb{C}}{\arg\max}\, P(c|d) = \underset{c \in \mathbb{C}}{\arg\max}\ P(c) \prod_{i=1}^{n} P(f_n|c) \tag{4.4}$$

NB classifiers have been shown to perform well for various classification tasks (Wang and Manning, 2012), in particular, when the features used are equally important and jointly contribute to the prediction. However, more sophisticated methods may be helpful where NB classifier fail. One such algorithm is SVMs, which have traditionally been used for text classification and have shown to achieve substantial improvements on a variety of tasks (Joachims, 1998). In particular, they have been shown to produce good results for the task of sentiment classification (Pang et al., 2002; Wilson et al., 2009; Maas et al., 2011).

Like all supervised machine learning methods, SVMs produce a model based based on the training data, which can then be used to predict the class labels of the test data. Each instance of the training and test data is first converted into a set of points in a feature space. For example, for unigram-based text classification, each sentence can be thought of as a point in a high-dimension feature space with each dimension representing a particular unigram in that sentence.

Consider a binary classification scheme with l training instances in an n-dimensional feature space, such that each instance $\mathbf{x}_i \in R^n, i = 1, 2, ..., l$ has a class label $y_i \in \{1, -1\}$. Figure 4.4 shows such a problem where the blue dots ($y_i = -1$) need to be separated from the red ones ($y_i = 1$).

Such a set of instances is said to be linearly separable if we can draw a line such that it all instances of the two classes are on the opposite sides of that line. More formally, it means that there exists a vector $\mathbf{w}$ and a scalar b, for which the following inequality is valid for all instances–label pairs $(\mathbf{x}_i, y_i)$:

$$y_i \left(\mathbf{w} \cdot \mathbf{x}_i + b\right) \ \geq \ 1 \tag{4.5}$$

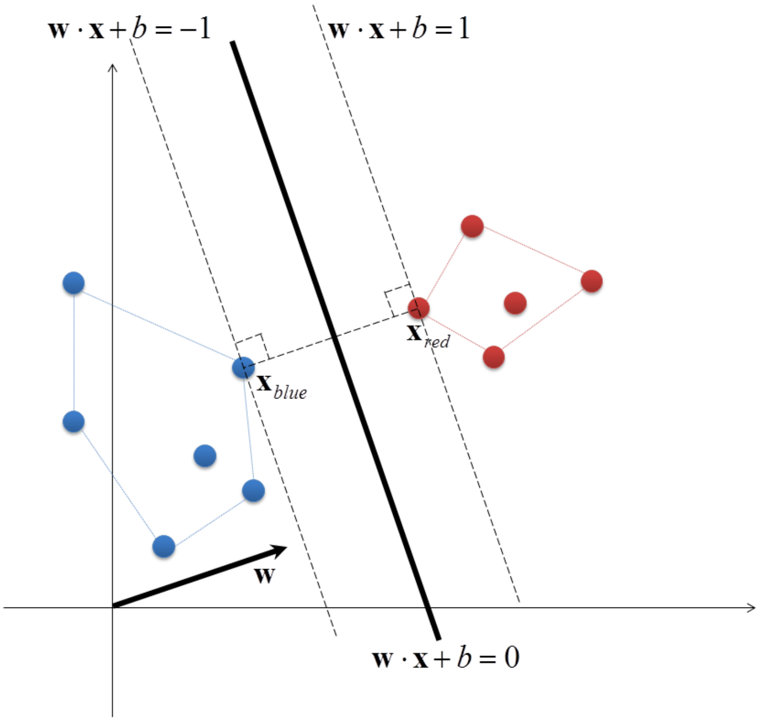

$\mathbf{w}\cdot\mathbf{x}+b=-1$
$\mathbf{w}\cdot\mathbf{x}+b=1$
$\mathbf{x}_{red}$
$\mathbf{x}_{blue}$
$\mathbf{w}$
$\mathbf{w}\cdot\mathbf{x}+b=0$

Figure 4.4: Example of an SVM classifier

The decision function which assigns a label to an unseen vector $\mathbf{x}$ is given by:

$$f(\mathbf{x}) = \text{sign}(\mathbf{w}\cdot\mathbf{x}+b) \tag{4.6}$$

While many such vectors exist, it can be shown that the *optimal* vector would result in maximising the margin of separation between the two classes. This vector would be perpendicular to the shortest line connecting the convex hulls of the two classes, and would also bisect it. In the figure above, the shortest line is the line between points $\mathbf{x}_{red}$ and $\mathbf{x}_{blue}$. For the weight vector $\mathbf{w}$, these points lie on the following parallel lines.

$$\mathbf{w}\cdot\mathbf{x}_{red}+b = 1 \tag{4.7}$$
$$\mathbf{w}\cdot\mathbf{x}_{blue}+b = -1 \tag{4.8}$$

These lines are called the *support vectors* and are shown as dashed lines in the figure. Subtracting these equations, we get:

$$\begin{aligned}\mathbf{w}\cdot(\mathbf{x}_{red}-\mathbf{x}_{blue}) &= 2\\ \left(\frac{\mathbf{w}}{\|\mathbf{w}\|}\cdot(\mathbf{x}_{red}-\mathbf{x}_{blue})\right) &= \frac{2}{\|\mathbf{w}\|}\end{aligned}$$

Therefore, in order to maximise the margin, we need to minimise $\|\mathbf{w}\|$ subjected to the constraint given in Equation 4.5. Cortes and Vapnik (1995) show that the solution to this constrained optimisation problem has an expansion $\mathbf{w} = \sum_i v_i \mathbf{x_i}$, where v_i are training instances that lie on the lines given by Equation 4.7 and Equation 4.8. The decision function can thus be revised to:

$$\begin{aligned} f(\mathbf{x}) &= \text{sign}(\mathbf{w} \cdot \mathbf{x} + b) \\ &= \text{sign}(\sum_i v_i \, (\mathbf{x} \cdot \mathbf{x}_i) + b) \qquad (4.9) \end{aligned}$$

We can see that the solution to the optimisation problem as well as the final decision boundary depend only on the dot products between these instances. This fact is important as it allows nonlinear classification problems to be solved using SVMs by mapping them to linear problems in a higher dimensional space.

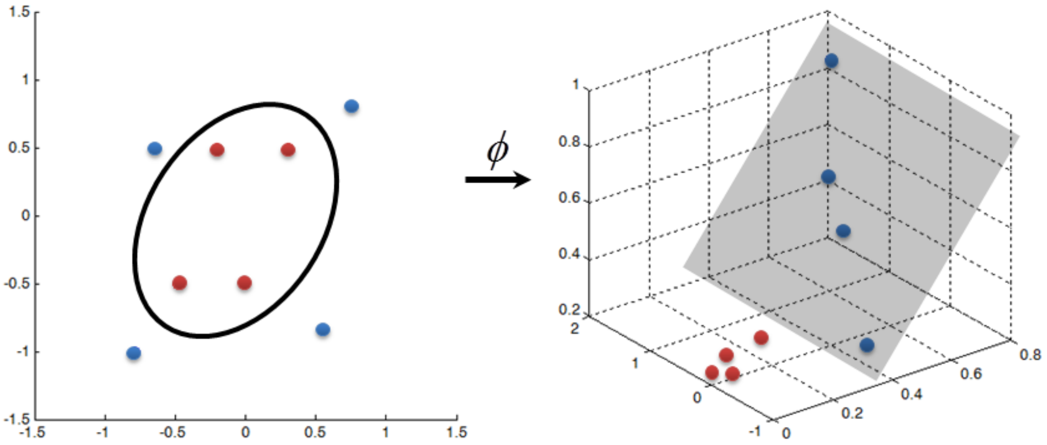


Figure 4.5: Using a Kernel Function

As an example, consider the problem in Figure 4.5. The data is linearly inseparable in a two-dimensional vector space. However, using a function $\phi : \mathbb{R}^2 \rightarrow \mathbb{R}^3$ converts the problem to a linearly separable one in a higher-dimensional space. Here, each data point is mapped to another point in a three-dimensional vector space. The decision function for an unseen vector $\mathbf{x}$ in Equation 4.9 can thus be updated to:

$$f(\mathbf{x}) \;=\; \text{sign}(\sum_i v_i \, (\phi(\mathbf{x}) \cdot \phi(\mathbf{x}_i)) + b) \qquad (4.10)$$

As stated earlier, the decision function requires calculation of dot products only. Cortes and Vapnik (1995) suggested using the general form of dot products in a Hilbert space (Anderson and Bahadur, 1962) for this purpose. This general form is given by

$$k(\mathbf{u}, \mathbf{v}) = \langle \, \phi(\mathbf{u}) \cdot \phi(\mathbf{v}) \, \rangle$$

Using a simple *kernel* k, the dot product can be calculated efficiently in the higher dimensional space as well. For instance, for the problem in Figure 4.5 with two points $\mathbf{u} = (u_1, u_2)$ and $\mathbf{v} = (v_1, v_2)$, let us consider the following polynomial kernel:

$$\begin{aligned} k(\mathbf{u}, \mathbf{v}) &= \langle\, \mathbf{u} \cdot \mathbf{v} \,\rangle^2 \\ &= (u_1 v_1 + u_2 v_2)^2 \\ &= (u_1^2 v_1^2 + 2 u_1 v_1 u_2 v_2 + u_2^2 v_2^2) \\ &= \left\langle\, (u_1^2, \sqrt{2} u_1 u_2, u_2^2) \cdot (v_1^2, \sqrt{2} v_1 v_2, v_2^2) \,\right\rangle \\ &= \langle\, \phi(\mathbf{u}) \cdot \phi(\mathbf{v}) \,\rangle \end{aligned}$$

This implies that the mapping function is given by:

$$\phi(\mathbf{x}) \rightarrow (x_1^2, \sqrt{2} x_1 x_2, x_2^2)$$

Therefore, the final decision function for an unseen data point (x) using a kernel k is given by:

$$f(\mathbf{x}) \quad = \quad \mathrm{sign}(\sum_i v_i \cdot k(\mathbf{x}, \mathbf{x}_i) + b) \tag{4.11}$$

The values for v_i are calculated by solving the function as a constrained programming problem. Many implementations of solving this problem are available publicly. For my experiments, I use the LIBSVM library (Chang and Lin, 2001; EL-Manzalawy and Honavar, 2005) implementation available through the WEKA toolkit (Hall et al., 2009). WEKA provides excellent resources for text processing as well as implementation and evaluation bindings of various machine learning algorithms.

The LIBSVM library also provides a selection of kernels. In the literature, the most popular kernels for text classification are the linear kernel and Radial Basis Function (RBF) kernel. These kernels are defined as follows:

$$\begin{aligned} k_{LINEAR}(\mathbf{u}, \mathbf{v}) &= \mathbf{u}^T \mathbf{v} \\ k_{RBF}(\mathbf{u}, \mathbf{v}) &= e^{-\gamma \| u - v \|} \end{aligned}$$

The linear kernel has a better running time and requires fewer memory resources as well. However, Keerthi and Lin (2003) show that the RBF kernel is more general. For this reason, I decided to use the RBF kernel for my classifier.

Model selection requires searching for optimal values of parameters. The RBF kernel has two cost parameters: C and γ. C is a cost parameter in the constrained programming solution and is used to reduce error. γ controls the sphere of influence of a data point. As recommended by Hsu et al. (2009), I used grid-search and 10-fold cross-validation to calculate the optimal parameters over the development set in my experiments.

To validate the performance of my classifier pipeline, I compare it with the work of Pang et al. (2002). Pang et al. (2002) examine movie reviews on a balanced corpus of $2,000$

positive and $2,000$ negative documents. They use 3-fold cross-validation with SVMs and report that their best system produces an accuracy of 83% using unigram features. On the same movie review corpus and feature set processed through my machine learning pipeline, I achieved an accuracy of 84%. This validates that my machine learning pipeline is comparable to their pipeline.

The work described by Pang et al. (2002), which uses a balanced corpus, measures performance using accuracy as the evaluation metrics. But accuracy may not be the most suitable metric to measure the performance in case of an imbalanced corpus. The following section describes my choice of evaluation metrics.

4.4 Evaluation Metrics

Evaluation toolkits, such as the one distributed with WEKA, offer an implementation of multiple evaluation metrics such as precision, recall and F score. These toolkits usually use a weighted average of F scores for each class to calculate the final F score. This weighted average is also called the micro-F score. This metric, however, becomes problematic when the corpus is not balanced. For a binary classification problem, if the instances of one class far outnumber the instances of the other class, the weight for the class with more members increases in tune with the degree of skew between the classes, and the resulting weighted average of the F score would be biased towards this class.

We can adjust for this bias by using the macro-F score as the evaluation metric. The macro-F score is another commonly used metric in text categorization (Lewis, 1991). It is the average F score over all classes. As it does not depend on the number of instances of a class, it gives relatively more importance to classes with lower representation than micro-F does. In fact, macro-F is a stricter evaluation metric because classifiers generally do not perform well on classes with lower data instances. Therefore, results measured using macro-F are generally lower than those measured using micro-F. Since all corpora I use have a skewed class distribution that reflects real-world occurrences, I report both the micro-F and the macro-F scores for the classifier evaluations. I perform 10-fold cross-validation to address the data sparsity problem.

In the next section, I describe the results of my experiments using the features and classifier discussed above.

4.5 Results

For the task of predicting sentiment of citation sentences, I use the corpus described in Section 3.2, which is labelled using a 3-class annotation scheme which labels each sentence as positive (p), negative (n) or objective (o). The classifier used is the SVM classifier discussed in Section 4.3.

As a baseline method to compare against, I use n-grams of length 1 to 3 as features. The baseline SVM classigied performs at a macro-F score of 0.597. I experiment with different combinations of the features described in Section 4.2. The results are given in Table 4.1.

	NB		**SVM**	
Features	**macro-F**	**micro-F**	**macro-F**	**micro-F**
1 gram	0.482	0.776	0.581	0.863
1-2 grams	0.473	0.764	0.592	0.864
1-3 grams $*$	0.474	0.764	0.597	0.862
Word/POS $1-3$-grams	0.476	0.764	0.535	0.859
1-3 grams + POS 3-grams	0.474	0.764	0.596	0.859
1-3 grams + Science Lexicon	0.474	0.764	0.597	0.860
1-3 grams + POS 3-grams + Science Lexicon	0.474	0.764	0.535	0.859
1-3 grams + Wilson et al. (2009) Features	0.478	0.766	0.418	0.859
FS1 (1-3 grams + Dependency Features) †	0.469	0.755	0.760	0.897
FS1 + Sentence Splitting + Window-Based Negation	0.449	0.754	0.683	0.872
FS1 + Sentence Splitting	0.450	0.751	0.642	0.866
FS2 (FS1 + Window-Based Negation) ‡	0.471	0.755	**0.764**	**0.898**

$*$ Baseline
† Feature Set 1 (FS1). Improvement statistically significant over Baseline
‡ Feature Set 2 (FS2). Improvement *not* statistically significant over FS1

Table 4.1: Results for Citation Sentiment Classification.

It should be noted that the Naive Bayes classifier does not perform well in as compared to the SVM classifier. One explanation for this behaviour might be the fact that while NB classifiers are efficient and robust with respect to irrelevant features and noisy data, their efficiency lies in the assumption of independence between the features (Kohavi, 1996; Lewis, 1998). This assumption may not hold for my corpus, an evidence of which is available in the results above which show that dependency features achieving the best scores. This may be attributed to the difficulty of the problem of sentiment detection (Pang et al., 2002). For this reason, I discuss the results of only the SVM classifier.

As shown in Table 4.1, trigrams seemingly work better than unigrams and bigrams for detecting sentiment in citations (macro-F = 0.597). This might indicate that the structure of scientific text is indeed lexically complex, as predicted in Section 4.2.1. Unlike the domain of movie reviews (Pang et al., 2002), in scientific text individual polar words by themselves are not enough to interpret the intention of their usage, and consequently, the sentiment. However, this improvement in results might be due to chance alone. The traditional approach to address this issue is to perform a test for statistical significance. In this dissertation, I use the pairwise Wilcoxon rank-sum test at 0.05 significance level, where indicated. All my significance tests are performed on the stricter macro-F scores.

Appending part-of-speech tags to each word in the trigrams does not achieve good results either. Adding POS trigrams produces F-scores in the same range as the ones produced appending a POS tag to each word (macro-F = 0.596 vs macro-F = 0.535). This might be because scientific text, after going through several revisions by the authors as well as the reviewers, is generally written clearly and without any ambiguities, thus reducing the need for any ambiguity resolution by the POS features.

We can also see that contextual polarity features proposed by Wilson et al. (2009) do not work well on citation text (macro-F = 0.418). Adding a science-specific lexicon instead also does not help much (macro-F = 0.597). The low F-scores of the Wilson et al. (2009) features, which includes a lexicon from the newswire domain, may indicate that n-grams are sufficient to capture the lexical structures that help the classifier in discriminating between two classes, and thus a specialised lexicon does not add much value. The results also show that lexical and contextual polarity features are surpassed by dependency features in improving the classification performance. Dependency triplets are thus critical as adding them in the feature set, along with 1-3 grams, increases the macro-F score from 0.597 to 0.760. This is shown as *Feature Set 1* (FS1) in the table on the preceding page. Adding the sentence splitting feature to the pipeline does not help (macro-F = 0.642). One possible cause of this might be that most citation scopes occurring in the text are larger than a single sentence.

To examine the effect of the negation window, I experimented with the test corpus using different lengths of the window ($0 <= k < 20$). The best score of macro-F = 0.764 was found at a window length of $k = 15$. The results are shown as *Feature Set 2* (FS2) in Table 4.1. While adding the negation window improved the system performance by an absolute 0.4%, this improvement was not found to be statistically significant. A graph of the system performance against a varying negation window is shown in Figure 4.6, where the error bars show standard error.

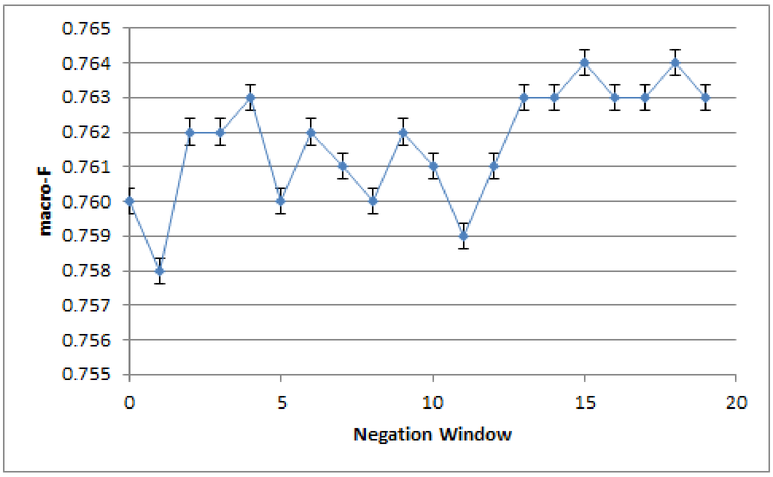


Figure 4.6: System performance on negation windows of different lengths.

I also compare my results with Teufel et al. (2006b). They group the 12 categories into three meta-categories in an attempt to perform a rough approximation of sentiment analysis over the classifications and report a 0.710 macro-F score. Unfortunately, I have access to only a subset[1] of this citation function corpus. I extracted 1-3 grams, dependencies and negation features from the reduced citation function dataset and used them in my classifier with 10-fold cross-validation. This results in an improved macro-F score of 0.797 achieved by my system for the available subset of the citation function corpus used by Teufel et al. (2006b). This shows that my approach is comparable to theirs. However, when this subset is used to test the approach trained on the citation sentiment corpus, a low macro-F score of 0.484 is achieved by my system. The most likely explanation is that there is a mismatch in the annotated class labels, i.e., citation sentiment classification is indeed different from citation function classification.

We can see that the overall results for sentiment detection in scientific citations are not as high as those reported in other domains, such as movie reviews (c.f. Section 2.4). One

[1] This subset contains 591 positive, 59 negative and 1259 objective citations.

reason might be that sentiment detection in scientific text is a difficult problem due to the presence of covert sentiment in the prevalent style of scientific writing. In science, direct criticism is avoided and opinions are generally expressed indirectly, thus making them harder to detect.

Another possible reason is that the corpus used in the experiments is limited to only the sentences containing the citation. As discussed in Section 3.4, much sentiment is present in the citation context which is being ignored in this corpus. The next chapter describes my work on the task of detecting context-sensitive citation sentiment.

4.6 Summary

This chapter explains the construction of a sentence-based sentiment classifier for scientific citations. I use the newly constructed citation sentiment corpus for this purpose under a machine learning framework. I explain the features which are used to describe each citation sentence. As far as lexical features are concerned, I experiment with n-grams, POS tags, general and scientific sentiment lexicons. I also include features from a state-of-the-art sentiment detection system built for contextual polarity of phrases. Next, I explore methods for extracting features based on the structure of the citation sentence: using dependency triplets as features, and a method of excluding the subtree which does not fall in the scope of the citation under review and using a k-word window for modifying the lexical features so as to include the scope of negation. The best results (macro-F = 0.764) were obtained using n-grams, dependency triplets and window-based negation as features. The same features are used in an SVM classifier with an RBF kernel, the results of which are reported in Chapter 6. The next chapter will describe classifiers for the tasks of detecting sentiment in citation context and significance of a citation in the citing paper.

Chapter 5

Detecting Citation Context

This chapter describes my experiments on context-sensitive citation sentiment detection. I decouple the tasks of identifying sentences in the context of a citation and examining their effect on citation sentiment analysis. First, I explore methods to automatically identify all mentions of a paper, as defined in Chapter 3, in a supervised manner. Section 5.1 lists different feature sets used to automatically identify these citation context sentences. Section 5.2 describes the classifier used for this task. The results of the classification are described in 5.3. Next, I combine the context determination with sentiment analysis, presenting the final result of context-sensitive sentiment detection. This work was published as Athar and Teufel (2012a) and Athar and Teufel (2012b).

5.1 Features for Detecting Citation Context

For the task of identifying informal citations that occur in the citation context, I use the citation context corpus described earlier in Section 3.5, with its four-class sentence-based scheme, *o/n/p/x*. In this scheme, class x means that the sentence contains no mention of the given citation, whereas the remaining three classes label the sentiment polarity of a citation in the obvious way.

Since the task at hand concerns context determination only rather than sentiment detection, I combine sentences of any polarity into a single class label *o_n_p*. This class label covers sentences with all formal and informal citations. However, it is trivial to detect sentences with formal citations as all of them contain the token <CIT>, which was inserted in the pre-processing step in Section 4.1. A classifier using lexical features would be able to trivially predict such sentences, resulting in an artificial increase in the evaluation results. In order to make sure that the results reflect the real difficulty of the task, all such sentences are excluded from the *o_n_p* class and assigned the x class instead[1].

I represent each citation as a set of features, each of which is a binary variable which is set to `true` or `false` based on presence of certain words and phrases in that sentence. The features are as follows:

[1] One may argue that by doing so, I have now artificially made the task *harder* for myself because I may have detected such sentences by other features that I *do* cover.

- Formal citation
- Author's name
- Acronyms
- Work nouns (e.g. *method*, *technique*, *algorithm*)
- Pronouns
- Connector
- Section markers
- Citation lists
- Lexical hooks (e.g. *Xerox tagger*)
- *n*-grams

These feature would be used in a machine learning framework in Section 5.3. A description of the feature follows.

5.1.1 Formal Citation

Since sentences with formal citations are easy to identify, including them in the evaluation metric would result in an artificial improvement in the final score. While such sentences have been excluded in our annotation, we need a feature that should help the classifier in identifying them as belonging to this excluded class. This is why I have included the following feature in my feature set: For the i^{th} sentence S_i in the citing paper,

f_1: S_i contains a formal citation to the target paper.

Let us now turn to the next feature, f_2. A sentence immediately succeeding a sentence with a formal citation is more likely to continue talking about the cited work. For this reason, I also have included the following feature in my feature set.

f_2: S_{i-1} contains a formal citation to the target paper.

As an example for this discourse property of citation contexts, consider the following sentences from the paper with ACL-ID W03-1017, which cites *Turney (2002).* Here, the paper ID and sentence numbers are shown at the start of each sentence.

> *W03-1017:31* **Turney (2002)** showed that it is possible to use only a few of those semantically oriented words (namely, "excellent" and "poor") to label other phrases co-occuring with them as positive or negative.

W03-1017:32 He then used these phrases to automatically separate positive and negative movie and product reviews, with accuracy of 66-84%.

Sentence 31 contains a formal citation to the *Turney (2002)* paper and sentence 32 continues to talk about the same paper. For sentence 32, the value of the f_1 will be `false` but the value of f_2 will be `true`. This value of f_2 may help the classifier in predicting the correct label for sentence 32. I extract this feature using a simple rule which searches for the presence of the last name of the primary author of the target paper followed by its year of publication.

As mentioned in Section 2.6, presence of a formal citation has also been included as a feature by Qazvinian and Radev (2010) to detect implicit citations. They use it for calculating the dependence between two consecutive sentences in a CRF-based model.

5.1.2 Author's Name

Tools, methods and algorithms are sometimes named after their inventors. Moreover, it is also a common practice to use the author's name instead of citing them formally after the initial mention of their work. An example of the use of an author's name without a formal citation is given below.

W05-0408:7 We show that our approach outperforms **Turney's** approach.

The phrase *Turney's approach* refers back to the citation *Turney (2000)* which occurred textually earlier in the article. For this reason, I have included the following feature for capturing such mentions.

f_3: S_i contains the last name of the primary author of the target paper.

The feature f_3 would help the classifier in capturing such informal citations. I use regular expressions to search each sentence for the last name of the primary author as given in the ACL metadata described in Chapter 3. This feature has been used in earlier work Giles et al. (1998); Day et al. (2007); Councill et al. (2008) on formal citation extraction, but has not yet been applied to the problem of identifying informal ones, as is done here. This feature has also been used for argumentative zoning (Teufel, 1999) and citation function classification (Teufel, 2006).

5.1.3 Acronyms

Acronyms are used in the scientific literature to refer to techniques, algorithms, theories, and evaluation metrics. Once introduced and connected to the paper, acronyms can be used in the place of a formal citation to refer to the paper. For example, the word *METEOR* in Figure 1.1 is an acronym that replaces a formal citation in subsequent sentences to refer to the paper that introduced METEOR. To capture such instances, I use the presence of an acronym within a four-word window of a formal citation in the citation sentence as a feature.

f_4: S_i contains an acronym co-occurring with a citation.

For each cited paper, I collect a list of acronyms from all sentences containing formal citations to that paper. I use regular expressions to identify words with only uppercase alphabets and digits within a four-word window of the formal citation. For instance, consider the following sentences.

> *D09-1136:62* This algorithm is referred to as **GHKM (Galley et al., 2004)** and is widely used in SSMT systems (Galley et al., 2006; Liu et al., 2006; Huang et al., 2006).
>
> *D09-1136:63* The word alignment used in **GHKM** is usually computed independent of the syntactic structure...

> *D09-1136:62* N06-2006:30 Version 1.5.5 of the **ROUGE** scoring algorithm **(Lin, 2004)** is also used for evaluating results.

In the first sentence, the target paper is *Galley et al. (2004)*. The acronym *GHKM* is present within a four-word window of the citation text and is thus added to the set of acronyms for this target paper. When the value of feature f_4 is calculated for sentence 63, it is set to `true` because of the presence of the acronym *GHKM* in this sentence near *Galley et al. (2004)*. The choice of the window was decided by experimenting with various values on the development set and selecting the value which yielded the best results.

Similarly, in the second example, there are two words between the citation to *Lin (2004)* and the acronym ROUGE, but the acronym would still be captured in the feature.

Existing work on acronyms focuses mostly on extracting a mapping between an acronym and its definition or expansion, and does not address the problem of citation detection (Chang et al., 2002; Nadeau and Turney, 2005; Larkey et al., 2000; Taghva and Gilbreth, 1999).

5.1.4 Work Nouns

Work nouns (Siddharthan and Teufel, 2007) are words which are used to refer to other work by other authors e.g. *study, technique, etc.* Work nouns usually signal a continuation of a topic related to a target citation in earlier sentences. For example:

> *W04-1113:31* Church and Hanks (**Church and Hanks 1990**) employed mutual information to extract both adjacent and distant bi-grams that tend to co-occur within a fixed-size window.
>
> *W04-1113:32* But **the method** did not extend to extract n-grams.

Here, sentence 32 contains the phrase *the method* which refers to *Church and Hanks (1990)* mentioned in sentence 31.

Teufel (1999) used work nouns as a feature in argumentative zoning, and Siddharthan and Teufel (2007) use them for the task of scientific attribution. Qazvinian and Radev (2010) also use work nouns in combination with determiners to identify informal citations. They construct a list of bigram patterns in the form dw, where d is a determiner or possessive pronoun from the set $D = \{this, that, those, these, his, her, their, such, previous\}$, and w is a word from the set $W = \{work, approach, system, method, technique, result, example\}$. They set the value of the feature to `true` if a sentence contains any member of the bigram set.

I follow a similar approach and use the same bigram pattern. However, I expand the set of determiner and possessive pronouns to include the word *other*. I also expand the work noun set by adding the words used by Siddharthan and Teufel (2007), as given in Teufel (2010). Most of these work nouns are however singular as Teufel had a mechanism to produce plurals on the fly, so I add the corresponding word plurals to the set. My final work noun list contains 79 words and is available in Appendix F.3. The feature I use is thus:

f_5: S_i contains a determiner followed by a work noun.

In the example above, the value of feature f_5 for sentence 32 will thus be set to `true` because of the presence of the phrase *the method*.

5.1.5 Pronouns

Using a pronoun to refer to a particular author is common in scientific text. Nanba and Okumura (1999) use first and third person pronouns as cue phrases for citation area extraction. Qazvinian and Radev (2010) use possessive pronouns in connection with work nouns to detect informal citations. Kim and Webber (2006) focus on detecting informal citations using the personal pronoun *they*.

My feature, f_6, is similar to these approaches. It is set to `true` if the first word of a sentence is in the set of possessive and demonstrative pronouns $P = \{he, she, his, her, they, their, this, these, such$

f_6: S_i starts with a third person pronoun.

An example of a pronoun used to refer to a preceding citation is given below.

> *N09-1068:103* Because **Daume III (2007)** views the adaptation as merely augmenting the feature space, each of his features has the same prior mean and variance, regardless of whether it is domain specific or independent.
>
> *N09-1068:104* **He** could have set these parameters differently, but he did not.

The value of feature f_6 will be set to `true` for sentence 104 because it starts with the pronoun *he*. Here, *he* refers to the author of the target citation in the previous sentence. I limit the position of the pronoun in the sentence to only the first word to ensure that the pronoun is referring to a noun in the previous sentence, and not in the current one.

5.1.6 Connectors

Connectors are words and phrases that signal a shift or continuation in the discourse structure (Trivedi and Eisenstein, 2013). They are indicators of the continuity of a topic previously discussed, particularly when used at the start of the sentence (Howe et al., 2010). To model this continuity, I use the presence of a connector at the beginning of a sentence as a feature.

> f_7: S_i starts with a connector.

A list of 23 connectors (e.g. *however, although, moreover, etc.*) has been compiled by examining the high frequency connectors from a development set of papers from the same domain. The complete list of these connectors is given in Appendix F.4. Since this list was created using citation data, each connector represents an instance of sentence linkage between a formal and an informal citation. Therefore, I collapse the presence of any of these connectors into a single feature. An example of the usage of connectors is given below.

> *P06-2054:14* An additional consistent edge of a linear-chain conditional random field (CRF) explicitly models the dependencies between distant occurrences of similar words (Sutton and McCallum, 2004; Finkel et al. , 2005).
>
> *P06-2054:15* **However**, this approach requires additional time complexity in inference/learning time and it is only suitable for representing constraints by enforcing label consistency.

Here, sentence 15 mentions some problems about the papers cited in the preceding neutral sentence. Since sentence 15 starts with the connector *however*, the value for feature f_7 will be `true`.

Polanyi and Zaenen (2006) discuss the influence that connectors such as *however, although* etc. have on information elsewhere in the text: they reduce or nullify the effect of a polar word present in the context. Connectors have also been shown to be effective in predicting the semantic orientation of adjectives (Hatzivassiloglou and McKeown, 1997) as well as creating interest in the reader towards the presented research (Bondi, 2004).

5.1.7 Section Markers

In scientific text, section boundaries are to used to signal the start of a new rhetorical structure (Nanba and Okumura, 1999; Teufel, 2006). New sections are more likely to indicate a topic shift as well as be the end of an existing discourse. Therefore, I use the following three binary features which help capture section boundaries.

f_8: S_i starts with a (sub)section heading.
f_9: S_{i-1} starts with a (sub)section heading.
f_{10}: S_{i+1} starts with a (sub)section heading.

Features f_8–f_{10} are set or unset to record the presence of a section boundary in the current, preceding or succeeding sentence.

Unfortunately, the AAN corpus I use does not contain any information about the structure of the paper text except sentence boundaries. Due to processing problems, the section numbers and headings are typically merged to the subsequent text. Consequently, I identify this information myself in order to incorporate section boundaries into my feature set. Consider the following example:

> *P06-2079:67* For instance, instead of representing the polarity of a term using a binary value, Mullen and Collier (2004) use **Turneys (2002)** method to assign a real value to represent term polarity and introduce a variety of numerical features that are aggregate measures of the polarity values of terms selected from the document under consideration.
>
> *P06-2079:68* **3 Review Identification** Recall that the goal of review identification is to determine whether a given document is a review or not.

Here, the section boundary starts at sentence 68 and the context of *Turney (2002)* ends at sentence 67. Therefore, if the first word is a number followed by a capitalised phrase, it should indicate the start of a section or a sub-section. In this case, the heading shown in bold signals the start of section 3. This is done by examining the start of each sentence for a section number followed by multiple words in title case. If found, the last word in title case is included in the sentence and the remaining words are taken to be the section title. For example, the word *Recall* in the text above would not be part of the section title. Thus, for sentence 67 in the example, feature f_8 will be `true`; where as for sentence 68, feature f_{10} will be `true`.

5.1.8 Citation Lists

Authors sometimes mention together citations referring to similar or related approaches in a citation list[2]. This usually occurs in sentences located in the background literature review section. In the following example, the target citation (i.e. *Church and Hanks, 1990*) is listed together with seven other citations.

> Concrete similarity measures compare a pair of weighted context feature vectors that characterize two words (**Church and Hanks, 1990**; Ruge, 1992; Pereira et al., 1993; Grefenstette, 1994; Lee, 1997; Lin, 1998; Pantel

[2] In the Harvard citation style, citations in a citation list are separated by a semicolon

> and Lin, 2002; Weeds and Weir, 2003)

For sentences like these, it is more likely that the context of a citation is limited to this sentence only when other citations are mentioned in the same list. Teufel (2010) argues that there is an inverse relationship between the number of citations in a sentence and its importance. Indeed, the target citation, *Church and Hanks, 1990*, is part of a list of with seven other citations, and we can see that the next line does not contain any mention of the target citation. For this reason, I include the presence of two or more citation in the sentence as a feature.

> f_{11}: S_i contains a citation other than the one under review.

In the example above, the value of f_{11} for line 10 would thus be `true`.

Another such case is when a citation does not occur as part of a list, but is nevertheless associated with one textual description in an enumeration of methods or other scientific entities, ad in the following example:

> *W03-0404:214* **(Turney, 2002)** used patterns representing part-of-speech sequences, (Hatzivassiloglou and McKeown, 1997) recognized adjectival phrases, and (Wiebe et al. , 2001) learned N-grams.
>
> *W03-0404:215* The extraction patterns used in our research are linguistically richer patterns, requiring shallow parsing and syntactic role assignment.

Here, the target citation is *Turney (2002)*, which is included as one of the examples among a list of methods in sentence 214. In accordance with my prediction, the next sentence does not mention this target citation, so the value of `true` for f_{11} should help in identification of a citation's relative insignificance.

5.1.9 Lexical Hooks

Authors sometimes use lexical substitutes other than acronyms, pronouns and work nouns to refer to a citation. I call such mentions *lexical hooks*. For example, the following sentences have been taken from two different papers, but cite the same target paper (i.e. *Cutting et al., 1992*).

> *E95-1014:88* This text was part-of-speech tagged using the **Xerox** HMM tagger (**Cutting et al., 1992**).
>
> *J97-3003:23* The **Xerox** tagger (**Cutting et al., 1992**) comes with a set of rules that assign an unknown word a set of possible pos-tags (i.e. , POS-class) on the basis of its ending segment.

While the acronym *HMM* will be captured by the feature f_3 described earlier, the word *Xerox* however will be missed. It is obvious from the examples that the tagger concerned is often described as the *Xerox tagger*. To capture such instances, I include the following feature in my set for each sentence.

> f_{12}: S_i contains a lexical hook.

I define a lexical hook as the most frequent capitalized phrase in all citation sentences to the target paper. In the example above, *Xerox* would be included in the list of lexical hooks. The value of the f_{12} will thus be set to `true` for all those sentences which contain the word *Xerox*.

It should be noted that all features described until now were obtained by examining the sentences of the citing paper only. In the case of this feature however, the formal citations from which the lexical hooks are extracted are gathered from *all* citing papers in the corpus. This 'domain level' feature thus makes it possible to extract the commonly used name for techniques and methods which may have been missed by the acronym feature.

Such phrases might also be associated with acronyms, even if the acronym is not always mentioned together with the phrase. I therefore generate the most likely acronym associated with a lexical hook and include it in the feature set. For example, *SCL* in the citation below is also included as a lexical hook along with *Structural Correspondence Learning*.

> *W09-2205:115* The paper compares **Structural Correspondence Learning** (**Blitzer et al., 2006**) with (various instances of) self-training (Abney, 2007; McClosky et al., 2006) for the adaptation of a parse selection model to Wikipedia domains.
>
> *W09-2205:117* The more indirect exploitation of unlabeled data through **SCL** is more fruitful than pure self-training..

The value of feature f_{12} will thus be `true` for sentence 117 as well, since it contains the lexical hook *SCL*.

5.1.10 *n*-Grams

I also add *n*-grams of length 1 to 3 to the feature set. *n*-grams have been shown to perform consistently well in various NLP tasks (Bergsma et al., 2010). I introduce *n*-grams as a fall-back feature set that captures those lexical phenomena which have been overlooked by other features described in this section.

> f_{ng}: n grams of length 1 to 3 extracted from S_i.

Following my earlier experiments, the *n*-grams are extracted from the citation text with the help of the WEKA (Hall et al., 2008) toolkit and an *IDF* weighting scheme (c.f. Section 4.2.1).

5.2 Classification

As discussed in Section 4.3, SVMs have been shown to achieve good results for the task of sentiment classification (Pang et al., 2002; Wilson et al., 2009; Maas et al., 2011). For this reason, I decided to use an SVM-based classifier for the tasks of detecting citation context as well. Following the setup used in detecting citation sentiment, I use the SVM implementation provided by the LIBSVM library (Chang and Lin, 2001; EL-Manzalawy and Honavar, 2005) available through the WEKA toolkit (Hall et al., 2009). The n-gram features were also obtained using the WEKA library functions. Implementation for calculating the remaining features was provided using Java. In the next section, I report the results of my experiments using the features described above.

5.3 Results

Since n-grams have been shown to perform consistently well in various NLP tasks (Bergsma et al., 2010), I use an n-gram only baseline. I compare the results using precision (P), recall (R) and F-scores as evaluation metrics. While I report the scores for both classes, score for only the *o_n_p* class are relevant as my corpus is not balanced (c.f. Section 4.4). A comparison of the confusion matrices and the results obtained using all features described in the previous section is given in Table 5.1.

	Baseline		My System		
predicted ↓	*o_n_p*	x	*o_n_p*	x	← truth
o_n_p	518	1,215	862	871	
x	641	201,429	766	201,314	
P	0.299	0.997	0.497	0.996	
R	0.447	0.994	0.529	0.996	
F	0.358	0.995	0.513	0.996	

Table 5.1: Confusion matrices and results for citation context detection.

For the task of detecting citation mentions, the F-score for an n-gram baseline is only 0.358. By adding the new features listed above, the performance of my system increases by almost 0.16 absolute to a F-score of 0.513, which corresponds to an improvement of 43%. Using the pairwise Wilcoxon rank-sum test at 0.05 significance level, I found that the difference between the baseline and my system is statistically significant. Despite the improvement, the score is low which shows that detection of informal citations is a hard task; a better solution may require a deeper analysis of the context.

We can also see in Table 5.1 that the precision has improved over the baseline from 0.299 to 0.497 by an absolute 0.198, which corresponds to an increase of 66%. The recall has also improved from 0.447 to 0.529 by absolute 0.082, corresponding to an 18% increase.

To analyse the contribution of individual features in my feature set, I used an implementation of the information gain feature evaluation method (Gupta, 2006) provided in WEKA. Information gain measures how good a feature is for predicting each label with respect to a class. For a class C, the information gain ratio IG for an attribute f is given by:

$$IG(C, f) = H(C) - H(C|f),$$

where H is the *entropy* which is defined for a class C by:

$$H(C) = -\sum_{c} \frac{n_c}{N} \log_b \frac{n_c}{N},$$

where n_c is the number of instances belonging to the class c and N is the total number of instances. Calculating the information gain (IG) for each feature, it is possible to rank them according to the information they provide for correct identification of the label. The sorted ranking for the new features proposed by me is given in Table 5.2. I do not list the n-gram feature as it is part of the baseline.

Feature	Description	IG×10^2
f_{12}	S_i contains lexical hooks	20.98
f_4	S_i contains acronym	14.84
f_2	S_{i-1} contains formal citation	7.02
f_3	S_i contains author's name	0.98
f_5	S_i starts with a pronoun	0.39
f_7	S_i contains work nouns	0.20
f_1	S_i contains formal citation	0.11
f_{10}	S_{i+1} starts with a section marker	0.04
f_{11}	S_i contains citation lists	0.03
f_8	S_i starts with a section marker	0.02
f_9	S_{i-1} starts with a section marker	0.00
f_6	S_i starts with a connector	0.00

Table 5.2: Ranked information gain for features.

We can see that presence of lexical hooks and acronyms are ranked the highest, confirming my hypothesis that these features are important for identification of an informal citation in a sentence. Lexical hooks contribute the most with an IG ratio of 20.98. Acronyms produce an IG ratio of 14.84, followed by the feature for presence of an informal citation in the previous sentence. In contrast, connectors and section markers do not perform very well for this task. This can be further illustrated by Figure 5.1 which shows the percentage class distribution of each feature in the corpus. We can see that lexical hooks are better at predicting the informal citations as their corresponding feature (f_{12}) is true for more than 50% of the sentences containing informal citations and less than 1% for the ones containing formal citations. Similarly, the acronym feature f_4 is true for more than 30% of informal citations and is thus one of the important features in my classifier.

I now explore the effect of context detection on citation sentiment detection.

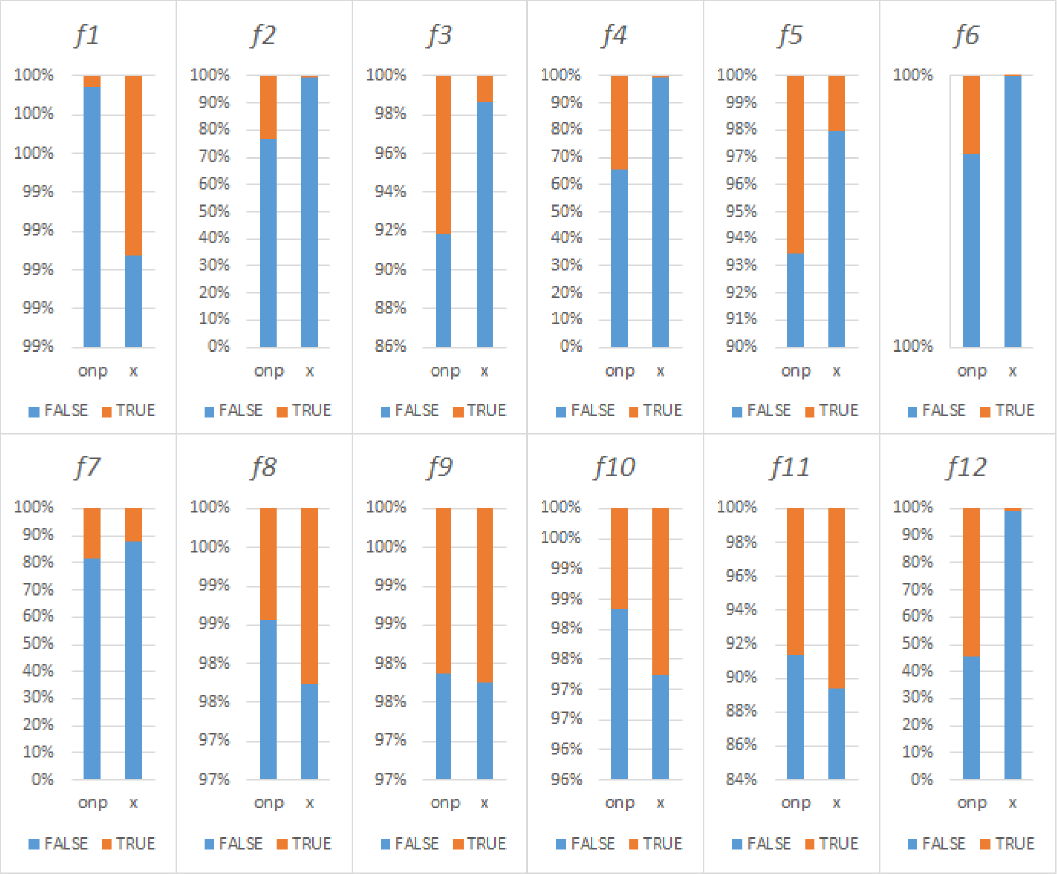


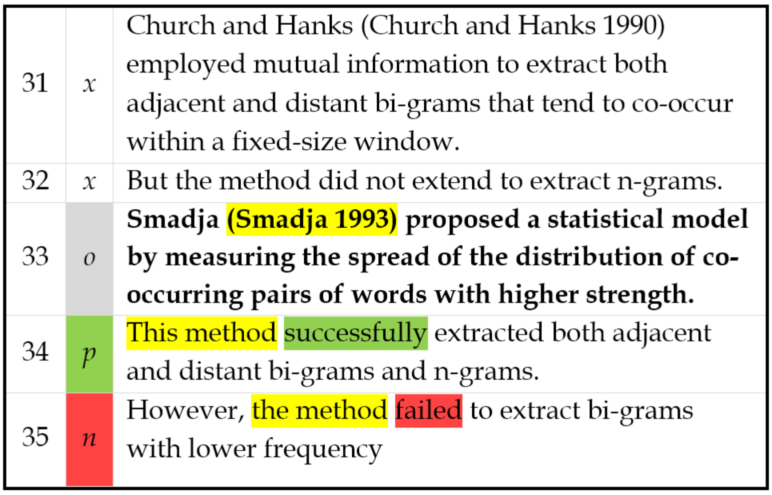
Figure 5.1: Distribution of features.

5.4 Impact on Citation Sentiment Detection

This section describes my experiment for determining whether or not including the informal citations in the context of a formal citation improve sentiment analysis. In order to do so, the first step is to perform sentiment analysis on formal citations only. While AAN contains a list of sentences containing formal citations to a given target paper, this list has been extracted automatically using regular expressions and has not been formally evaluated (c.f. Section 3.1.1). For this reason, I use the same method used in feature f_1 to identify formal citations. This resulted in identification of 1,874 sentences, which were then classified for sentiment using the classifier described in Chapter 4, which uses n-grams and dependency relation as features. I use this single-sentence system as the baseline for my experiment.

My new context-enhanced annotation is however not restricted to only formal citations, and there may exist more than one sentiment per explicit citation when informal citations are taken into account. For instance, let us re-examine the example from Figure 3.9, reproduced here as Figure 5.2.

The citation sentence sentiment detection system from Chapter 4 is restricted to analysing sentence 33 only. However, more sentiment towards this formal citation exists in the succeeding sentences. While sentence 34 in Figure 5.2 is positive towards the cited paper, the next sentence criticises it. Thus for this formal citation with a class label o, there are



31	x	Church and Hanks (Church and Hanks 1990) employed mutual information to extract both adjacent and distant bi-grams that tend to co-occur within a fixed-size window.
32	x	But the method did not extend to extract n-grams.
33	o	**Smadja (Smadja 1993) proposed a statistical model by measuring the spread of the distribution of co-occurring pairs of words with higher strength.**
34	p	This method successfully extracted both adjacent and distant bi-grams and n-grams.
35	n	However, the method failed to extract bi-grams with lower frequency

Figure 5.2: Example annotation of a citation context (reproduced).

two sentences containing informal citations with sentiment. However, to make sure the annotations are comparable to the baseline, I need to produce a single sentiment label per citation.

For this purpose, I mark the true citation sentiment to be the last sentiment mentioned in a 4-sentence context window of the formal citation, as this is pragmatically most likely to be the real intention (MacRoberts and MacRoberts, 1984). The window length is motivated by recent research (Qazvinian and Radev, 2010), which favours a four-sentence boundary for detecting non-explicit citations. Analysis of my corpus shows that more than 60% of the subjective citations lie in this window. Figure 5.3 shows a graphical representation of this labelling, where the grey, green and red circled represent sentences $33, 34$ and 35 respectively.

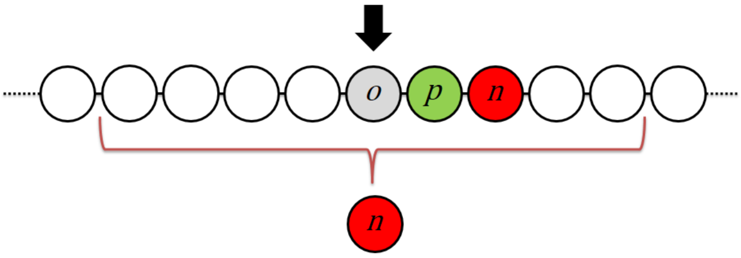


Figure 5.3: Extraction of a single sentiment label from a 4-sentence window.

The new class label for this example would this be n. In order to make my classifier aware of the sentences containing informal citations, I append them to the sentence containing the formal citation, thus making a larger sentence. This new sentence is also processed using the same classifier as the baseline. Table 5.3 compares the results of using this context-enhanced text with the formal citation sentence baseline.

The results show that my system outperforms the baseline in almost all evaluation criteria. We can see that the precision has improved for all classes, the largest one being the

	Baseline			My System			
predicted ↓	*o*	*n*	*p*	*o*	*n*	*p*	← truth
o	1323	270	120	1238	128	78	
n	5	25	8	73	182	13	
p	31	30	62	31	30	62	
P	0.772	0.658	0.504	0.857	0.679	0.611	
R	0.974	0.077	0.326	0.911	0.56	0.521	
F	0.861	0.138	0.396	0.883	0.614	0.563	
micro-F		0.689			0.804		
macro-F		0.465			0.687		

Table 5.3: Results for citation-enhanced sentiment detection.

positive sentiment (p) changing from 0.504 to 0.611, an increase of absolute 0.10, which corresponds to an improvement of 21%. The recall scores are more interesting and we see a 6% decrease in the recall for class o from 0.974 to 0.911. However, at the cost of this small decrease, the recall for sentences with negative sentiment (class n) increase by an absolute of 0.483 from 0.077 to 0.560, corresponding to an improvement of 627%. The recall of sentences with positive sentiment (class p) also improves by an absolute 0.20 from 0.326 to 0.521, corresponding to a 60% increase. These results are in line with my hypothesis that much negative sentiment is hidden in sentences containing informal citation, and including such sentences in sentiment analysis leads to a better coverage of sentiment towards a target paper.

In the aggregated metrics, the micro-F score increases from 0.689 to 0.804, an absolute improvement of absolute 0.12, which corresponds to a 17% improvement over the baseline. The macro-F score increases from 0.465 to 0.687, corresponding to an absolute improvement of 0.23, which corresponds to a 48% improvement.

Performing the pairwise Wilcoxon rank-sum test at 0.05 significance level, I found the improvement to be statistically significant. The baseline system does not use any context and thus misses out on all the sentiment information contained within. While this window-based representation does not capture all the sentiment towards a citation perfectly, it is closer to the truth than a system based on single sentence analysis and is able to detect more sentiment.

5.5 Summary

This chapter describes the feature sets that I use in the tasks of detecting citation context. I describe the features for detecting if a sentence is in the context of the citation or not. These binary features are based on presence of formal citations, author name, acronyms, work nouns, section markers, and lexical hooks in the sentence. I use an SVM-based classifier with these features and report an improvement of 8% absolute over an n-gram baseline. I also examine the effect of using the sentences in the citation context for the eventual task of sentiment analysis. I show that including these sentences results in an improvement of 48% over a single-sentence baseline. The next chapter describes my experiments on detection of citation significance using the features described above.

Chapter 6

Detecting Significance of a Cited Paper

This chapter describes my experiments on the new task of detecting whether or not a referenced paper is significant to the content of the citing paper. As stated earlier, Ziman (1968) argues that most papers are cited out of "politeness, policy or piety". This phenomenon has been confirmed empirically by Spiegel-Rosing (1977), who analysed 2,300 citations and found that 80% are cited just to point to further information. More recently, Hanney et al. (2005) manually annotated 623 references and found that only 9% were of essential importance to the citing paper. These informal citations also play a role in assessing the significance of a cited paper in the citing paper. Traditionally, the popularity of a paper has been linked to the number of papers citing it. In the existing literature, this number is a key feature for determining the significance of a paper, and is known as the citation count for that paper (Garfield, 1979). If the function $citers(P)$ gives the set of papers that cite the paper P, the total citations CT of an author with n papers can be found by:

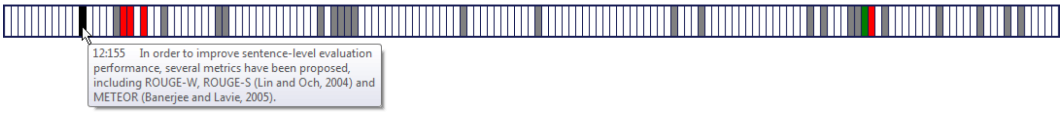
$$CT = \sum_{i=1}^{n} \|citers(P_i)\|$$

CT is used in bibliometrics as a measure of the influence and impact of individuals and journals. Other researcher have experimented with random-walk models for the task of identification of "good" papers in a network (Chen et al., 2007; Ma et al., 2008a).

CT, as well as other existing measures such as h-Index (Hirsch, 2005), treat all citations as equal. In these algorithms, no distinction is made between citations that are central to a citing paper and those which have been mentioned just in passing. These in-passing citations do not contribute much to the content of the citing paper and are consequently likely to be less significant to someone who is looking for more information related to the cited paper.

Another way to estimate the significance of a citation to the citing paper is to count the number of times the cited paper has been mentioned in the text of the citing paper. A higher count should indicate that the cited paper is being discussed in more detail and this would imply that it has a higher significance to the citing paper. I will use this feature as a baseline in the experiments that follow. This method would be even more effective if we count *all* linguistic mentions of a particular paper and not just the formal citations. This becomes particularly important when an acronym is commonly used in a field as a referent expression to the cited paper.

An example is given in Figure 6.1, which represents the sentiment of an article (P06-2070) towards the *METEOR* paper by *Banerjee and Lee (2005)*, a text extract of which has already been shown in Figure 1.1. This *sentiment fingerprint* is interpreted as follows. Each rectangle represents a single sentence (155 in this example). The black rectangle representing sentence 12 corresponds to the formal citation. The grey rectangles represent sentences containing informal citations which are neutral in sentiment. Similarly, the red and green rectangles correspond to sentences containing negative and positive informal citations respectively.



12:155 In order to improve sentence-level evaluation performance, several metrics have been proposed, including ROUGE-W, ROUGE-S (Lin and Och, 2004) and METEOR (Banerjee and Lavie, 2005).

Figure 6.1: Sentiment fingerprint of a paper.

As is apparent from the fingerprint, there is only one formal citation in the citing paper, but 27 informal ones. If we consider only the formal citation, the contribution of the *METEOR* paper does not seem significant, but in reality, the content of the citing paper revolves around this cited paper. My method for detecting the significance of a paper in the citing paper will attempt to cover as many mentions of a paper as possible. In the next section, I describe features for building a classifier for this task.

6.1 Features for Detecting Significance of a Target Paper

For detecting the significance of a target paper T in the citing paper, I use the following features from the citation context feature set from Chapter 5:

f_1: **Formal Citation -** Number of sentences which contain a formal citation to the paper T.

f_2: **Author's Name -** Number of sentences which contain the name of the author of paper T.

f_3: **Acronyms -** Number of sentences containing acronyms extracted from sentences with formal citations to T.

f_4: **Lexical Hooks -** Number of sentences containing a lexical hook to T (determined globally).

f_5: **Pronouns -** Number of sentences immediately following a sentence with a formal citation to T that start with a third person pronoun.

f_6: **Connectors -** Number of sentences immediately following a sentence with a formal citation to T which start with a connector.

f_7: **Citation Lists -** Number of sentences which contain an formal citation to T and two or more citations to other papers.

f_8: **Work Nouns -** Number of sentences immediately following a sentence with a formal citation to T which contain a determiner followed by a work noun.

The principles behind using these features have already been described in Section 5.1. However, here the feature value types have been changed from binary to nominal by counting all sentences which have the value of the respective feature set to `true`. For example, if sentences 4, 5, 9 and 17 of a paper contain the name of the author of the target paper, the value of feature f_3 would be set to 4.

The two novel features that I use for citation significance detection (f_9 and f_{10}) are described below.

6.1.1 Average Sentence Similarity

Cosine similarity is commonly used in information retrieval as a measure of document similarity (Manning et al., 2008). It assumes that a document can be represented by a vector in a W-dimensional vector space, where W is the number of unique words in the document set. The similarity between two sentences d_1 and d_2 represented by vectors $\vec{V}(d_1)$ and $\vec{V}(d_2)$ is thus the cosine of the angle between these two vectors. If $\vec{V}(d_1) = (d_{11}\ d_{12}\ ...\ d_{1W})$ and $\vec{V}(d_2) = (d_{21}\ d_{22}\ ...\ d_{2W})$, it can be calculated by the following formula:

$$\begin{aligned} sim(d_1, d_2) &= \frac{\vec{V}(d_1) \cdot \vec{V}(d_2)}{|\vec{V}(d_1)||\vec{V}(d_2)|} \\ &= \frac{d_{11}d_{21} + d_{12}d_{22} + ... + d_{1W}d_{2W}}{\sqrt{d_{11}^2 + d_{12}^2 + ... + d_{1W}^2}\sqrt{d_{21}^2 + d_{22}^2 + ... + d_{2W}^2}} \end{aligned}$$

where the numerator represents the dot product of the vectors and denominator is the product of their Euclidean lengths.

The similarity score should aid the classifier in identifying sentences which discuss the topic mentioned in the title of the target paper. Consider the following sentences from a paper which cites a paper titled *Effective Self-Training For Parsing.*

> *C08-1071:23* While the recent success of **self-training** has demonstrated its merit, it remains unclear why **self-training** helps in some cases but not others.
>
> *C08-1071:24* Our goal is to better understand when and why **self-training**

is beneficial.

Here, the lexical overlap of the phrase *self-training* with the title of the cited paper would indicate the subject matter is being discussed in detail in the citing paper. In fact, 57 out of 236 sentences of the citation paper (C08-1071) contain this phrase. The feature is thus:

f_9: Average similarity of sentences with the title of paper T.

The value for feature f_9 for this paper would thus be far greater than that of a paper which does not mention *self-training* at all. I calculate this value by taking the average of the cosine similarity of each sentence with the title of the target paper. The cosine similarity is calculated after removing stop words. The stop word list has been taken from the Snowball project and is publicly available[1]. Ritchie (2008) uses cosine similarity for information retrieval of citations. Qazvinian and Radev (2008) employ cosine measure on the task of scientific paper summarization using citations. Qazvinian and Radev (2010) detect informal citations by using cosine measure to model the similarity between two sentences in a CRF framework.

6.1.2 Citation Location

It is common in scientific text to cite a number of papers in the introduction and related work sections. If a paper is cited in passing, it is more likely that it will be mentioned in these sections. Similarly, in-passing citations also occur when authors discuss their results or refer to future work at the end of the paper. Therefore, I have added the following feature to represent the formal citations to the target paper.

f_{10}: Number of sentences in the middle 50% of the paper which contain a formal citation to T.

It is assumed here that the middle 50% sentences in the paper, i.e. the sentences that occur between 25% and 75% of the total sentences in the paper, contain the 'meat' or actual substance of the paper. A paper central to the citing paper would thus have many citations in the middle 50%. Therefore, this feature which captures the number of citations in that area of text, should help model importance.

In the next section, I briefly discuss the classifier that I use for the task of automatic detection of significance of a reference.

6.2 Classification

As discussed in Section 4.3, SVMs have been shown to achieve good results for the task of sentiment classification (Pang et al., 2002; Wilson et al., 2009; Maas et al., 2011). Naive Bayes is also another common classification method which has been shown to perform well

[1] http://snowball.tartarus.org/algorithms/english/stop.txt

for text-based classification tasks (Wang and Manning, 2012). For this reason, I decided to use both NB and SVM classifiers for the task of detecting significance of a referenced paper. Following the setup used for detecting citation sentiment (c.f. Chapter 4), I use the SVM implementation provided by the LIBSVM library (Chang and Lin, 2001; EL-Manzalawy and Honavar, 2005) available through the WEKA toolkit (Hall et al., 2009). The WEKA implementation is also used for the NB classifier.

As a baseline, I use a simple rule-based system which determines the number of formal citations in running text for each reference in the citing paper. It uses the first feature discussed above, i.e., the number of sentences containing the last name of the author of the reference and its publication date. If this count amounts to three or more sentences, the paper is considered to be significant. Otherwise, it is assumed that the citation was in passing only. The cut-off point of three was chosen because it gave the best results for the range 1 to 15 on a development corpus.

I now present the results of my experiments using the described features.

6.3 Results

As evaluation metrics, I use precision (P), recall (R) and F-scores of only the significant citations, i.e., class Y. A comparison for NB and SVM classifiers using different sets of features is given in Table 6.1. The best configuration of my system, shown in bold, performs significantly better than the baseline and improves the F-score by 0.08 absolute, which corresponds to an improvement of 17%. Using the Wilcoxon signed-rank test Wilcoxon (1945) at 0.05 significance level, I found this improvement over the baseline to be statistically significant.

	NB			**SVM**		
Features	P	R	F	P	R	F
f_1 (baseline)	0.791	0.327	0.463	0.791	0.327	0.463
f_{1-2}	0.744	0.377	0.500	0.725	0.309	0.433
f_{1-4}	0.660	0.420	0.513	0.485	0.290	0.363
$f_{1-2,5-6}$	0.711	0.364	0.482	0.616	0.327	0.427
f_{1-8}	0.640	0.451	0.529	0.456	0.352	0.397
$f_{1-3,5-9}$	**0.655**	**0.469**	**0.547**	0.376	0.309	0.339
$f_{1-3,5-7,9}$	0.664	0.451	0.537	0.426	0.389	0.406
f_{1-10}	0.619	0.451	0.521	0.284	0.204	0.237

Table 6.1: Comparison of F-scores for significant citation detection.

It should be noted that the performance of the NB classifier is significantly better than that of the SVM classifier. In fact, the best results for the SVM classifier are for the baseline. One reason for this might be the sparseness and skewness of the training data. As argued by Jordan (2002), generative classification approaches such as NB perform better than discriminative methods for smaller datasets, where the number of training examples is relatively low. This also indicates that the SVM results may improve if more labelled data is used for training.

Another interesting aspect is the decline in precision as more features are introduced. We can see that two feature sets produce the same F-score of 0.543, but the feature set with f_8 (work nouns) has a lower precision ($P = 0.658$) and higher recall ($R = 0.463$), where as the one without it has higher precision ($P = 0.682$) but lower recall ($R = 0.451$). This behaviour is typical for Information Retrieval tasks where a tradeoff between precision and recall is usually a factor in tuning system performance as required by the application.

As described earlier in Section 5.2, I calculate the information gain (IG) for each feature in order to identify their importance in the classification task. The sorted ranking is given in Table 6.2

Feature	Description	IG$\times 10^2$
f_1	Number of sentences with formal citations	15.44
f_2	Number of sentences with author's name	11.23
f_7	Number of sentences with citation lists	4.84
f_9	Avg. similarity of sentences with the title	4.50
f_4	Number of sentences with lexical hooks	4.39
f_3	Number of sentences with acronyms	3.61
f_{10}	Citation location	3.15
f_8	Number of sentences with work nouns	2.65
f_5	Number of sentences starting with a pronoun	1.37
f_6	Number of sentences starting with a connector	0.00

Table 6.2: Ranked information gain for features.

We can see that the best feature is that of the baseline, i.e., number of sentences with formal citations. It is followed closely by the number of sentences with the name of the primary author of the target paper in the text. However, the rest of the features do not contribute as much to the classifier and their respective gains are thus low. Out of the two newly introduced features, average similarity with the title f_9 performs better than the citation location indicator f_{10}.

Figure 6.2 shows the distribution of classes for each feature in the dataset. The y-axis displays the frequency of data instances which belong to the class Y or N, where as the x-axis shows the corresponding feature value. In this case, no single feature is exclusively representative of any class and this leaves much room for improvement by introduction of better features in our system. Having access to more data may also help in increasing the performance of the classifiers.

6.4 Summary

This chapter describes my method for detecting citation context and significance. I describe the features used for detecting if a citation is significant to a cited paper or not.

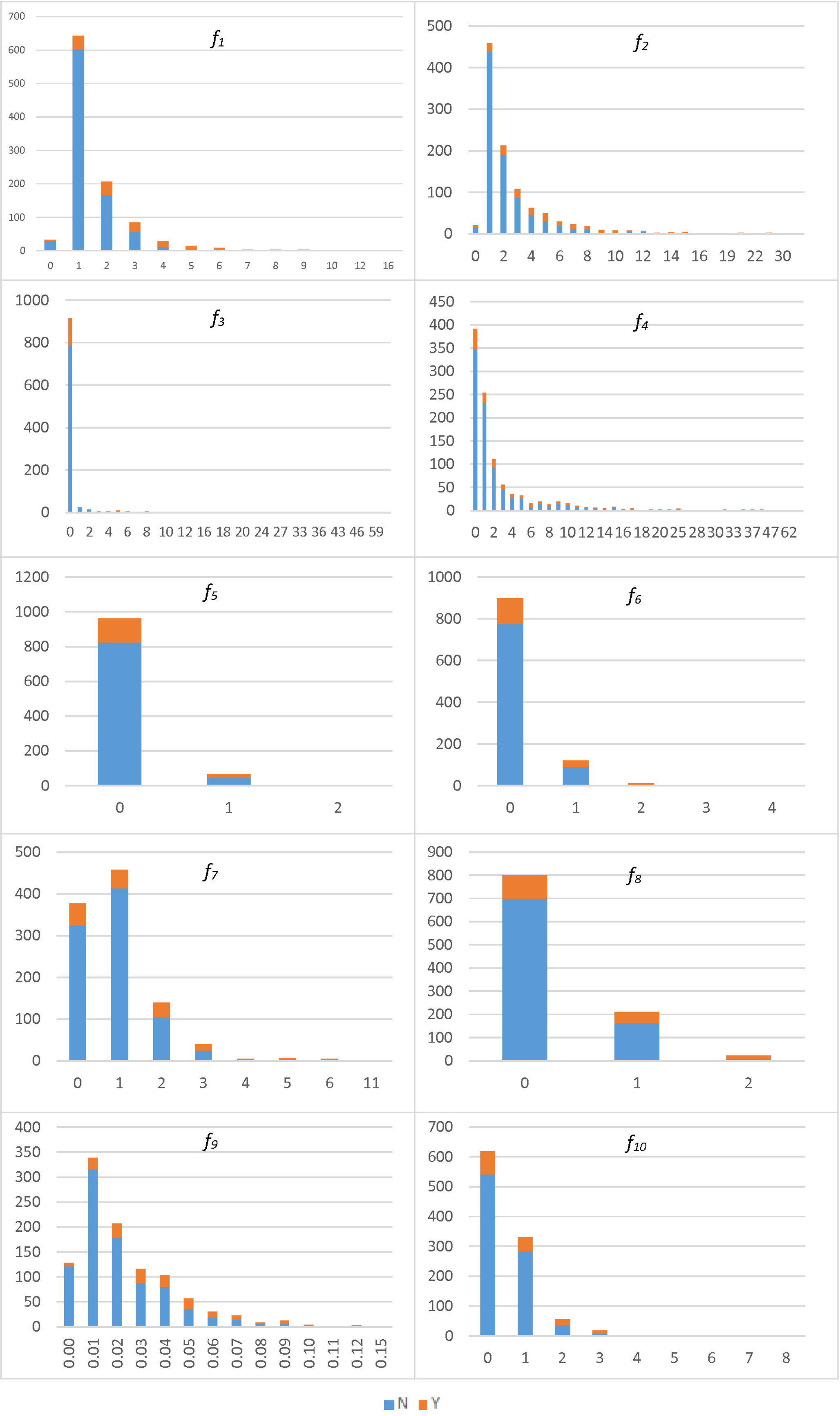


Figure 6.2: Distribution of features in classes.

In contrast to the work described in Chapter 4 and Chapter 5, I use a set of features to describe each *paper* instead of each *sentence*. This set includes summation of the sentence-based features described earlier. I also add two more features – average similarity of all sentences of the citing paper with the title of the cited paper as well as the number of formal citations in the middle of the citing paper. Using these features in a NB classifier, I show a 0.05 absolute increase in the macro-F score, which corresponds to an improvement of 6.8%.

In the next chapter, I summarize all the work described in this dissertation and propose future directions for research.

Chapter 7

Conclusion

> *'Would you tell me, please, which way I ought to go from here?'*
> *'That depends a good deal on where you want to get to,' said the Cat.*
>
> *Alice in Wonderland – Lewis Carrol*

In this chapter, I summarise the contributions of this dissertation and also give future directions of my work and possible frontiers for its application.

This dissertation focuses on the task of detecting sentiment in scientific citations. To address this task, I have used information from informal citations as well as the citation context. While there has been some work in recent years on the analysis of sentiment present in citations, my approach differs from the existing research in the several ways.

Current work on citation sentiment detection assumes that the sentiment present in the citation sentence represents the true sentiment of the author towards the cited paper and does not deal with informal citations. I was able to quantify how much sentiment information towards the cited paper would be lost if we disregarded the sentences surrounding the formal citations.

There have been no attempts on citation sentiment detection on a large corpus. Since most existing work utilises labelled data in a machine learning framework, lack of a large amount of data should make it probable for the learning algorithms to be less general and be more inclined towards over-fitting the resulting model.

Current techniques for informal citation detection do not focus on identifying linguistic mentions of citations which are present in the form of acronyms and other lexical hooks. Being a conference-driven field, computational linguistics articles usually have a page limit in submissions. Using acronyms and other co-reference related linguistic tools is therefore a common method to satisfy this constraint. Ignoring this phenomenon leads to additional loss of sentiment information about a citation.

In my research, I overcome these deficiencies by detecting informal citations and using their context in order to identify opinions, using a large annotated corpus. The next section describes my contributions to the field and gives an overview of the dissertation.

7.1 Overview and Contributions

In this dissertation, I have explored three main directions for identification of sentiment in scientific citations.

Firstly, I explore different feature sets for detection of sentiment in formal citations and show that the best results (macro-$F = 0.764$) are obtained using n-grams and dependency triplets. These results, given in Chapter 4, have been published in Athar (2011). I also constructed a new citation corpus for these experiments, which is publicly available at `http://www.cl.cam.ac.uk/~aa496/citation-sentiment-corpus/`.

Secondly, I show that the assumption that sentiment should be harvested only from citation sentence, is incorrect. Detecting the citation context helps in extracting more sentiment from the indirect and informal citation. I present a citation sentiment detection approach which takes into account all mentions of a given citation in any form, including indirect mentions and acronyms. I propose new features which help in extracting such citations and achieve a macro-F score of 0.754, an 11% improvement over an n-gram based baseline. This makes it possible to capture more sentiment about each cited paper as the total number of sentences increases from 1,739 to 3,760. Furthermore, I examine the effect of using the citation context in sentiment analysis and show that upon using informal citations, the precision increases from 0.299 to 0.497, an improvement of 66%. These methods have been described in Chapter 5 and published as Athar and Teufel (2012a) and Athar and Teufel (2012b).

To perform these experiments, I present a new large corpus which contains more than 200,000 sentences that have been annotated for sentiment. Construction of this corpus is described in Chapter 3, and I have made it publicly available for research at `http://www.cl.cam.ac.uk/~aa496/citation-context-corpus/`.

Lastly, I define a new task of detecting citations which have been mentioned in passing only. In chapter 6, I present a new annotation for this problem using the corpus described above. This corpus is publicly available at `http://www.cl.cam.ac.uk/~aa496/citation-significance-corpus/`. I propose features which help discriminate such citations from citations in passing, and also show that they result in an improvement of 17% over the baseline.

7.2 Applications and Future Work

Citation sentiment detection is an attractive task. This section describes some of the ways in which it can help researchers.

Citation sentiment can be used for determining the quality of a paper for ranking in citation indexes by including negative citations in the weighting scheme. Ranking algorithms, such as PageRank, have already been extended to include negative edge weights for different domains (Kamvar et al., 2003; De Kerchove and Dooren, 2008; Kunegis et al., 2009). For bibliometrics, however, the effects of these negative edges have yet to be analysed. This information can be used to not only rank articles on a global level for determining their importance, but also provide personalised ranking and recommendations of articles which have been tailored to a particular researcher.

A related application is assessing the relative impact of a paper by discounting 'in passing' citations. If citation is not significant to the citing paper, its contributions towards the field might be overestimated if we include such mentions in the citation count of the cited paper. As discussed above, in case of a graphical ranking algorithm, such citations can be assigned a weight which is minimal, arguably zero. This will make sure that the citation count is not biased due to citations which are less central and should result in a better estimate of the impact of the cited paper.

Analysing the citation sentences which are positive about a paper can help one identifying contributions of that research work in the domain. Such sentences can be a good source for detecting the improvements and innovations that a given paper has introduced in the field.

Similarly, identifying shortcomings and detecting problems in a particular approach is possible by analysing negative citations about a paper. Negation citation sentences can be used for recognising unaddressed issues and possible gaps in current research approaches. This may also prove useful for novice researchers who are searching for a problem to target their work on. Such researchers might profit from knowing about research gaps that have already been mentioned with a negative sentiment in a relevant citing paper.

Another application is the identification of personal bias of an author by observing their criticism and appraisal trends. For example, if a given author cites a particular technique positively, they might be more likely to be interested in relevant research. This can be beneficial for organisers of conferences who need to identify referees for a particular topic.

All these applications are based on the problems faced by researchers in practice. Development of any system around these applications can thus be of great use for the research community, especially when they are used with a citation indexing system such as CiteSeer or Google Scholar.

Future research directions for citation sentiment analysis include the use of a more sophisticated approach for modelling negations. New techniques for detection of the negation scope such as the one proposed by Councill et al. (2010) might be helpful in this respect.

Another direction is the identification of better features for detecting coherent citation blocks. I addressed the coherence of citation contexts using certain referring expressions and coherence-indicating conjunctions. The reintroduction of citation contexts was addressed via lexical hooks. A more fine-grained model of coherence might include linguistically deeper anaphora resolution (Lee et al., 2011), which is still an unsolved task for scientific texts. Other models of lexical coherence such as lexical chains (Barzilay and Elhadad, 1997) and entity coherence (Barzilay and Lapata, 2008) are also likely to be of interest for improving citation context determination.

Lastly, the most immediate research directions that I am currently following are practical applications which exclude citations that have been mentioned in passing only, from sentiment analysis of citation sentences. Such citations are mostly objective and ignoring such citation might help overcome the bias that these neutral citations introduce in the corpus.

Appendix A

Evaluation Metrics

To evaluate the effectiveness of a classifier, the most commonly adapted metrics are accuracy, precision and recall. Consider a two-class classifier where the label for an instance can take a value of either `Y` or `N`. Generally, the *truth* value for each instance in the data is determined during annotation. The classifier produces a *predicted* value for each instance which can be either `Y` or `N` as well. Therefore, there are four possible combinations which need to be examined to calculate the effectiveness of a classifier. These combinations are given below in the form of a confusion matrix or a contingency table.

predicted ↓	`Y`	`N`	← truth
`Y`	true positives (*tp*)	false positives (*fp*)	
`N`	false negatives (*fn*)	true negatives (*tn*)	

The accuracy A, precision P and recall R are defined as:

$$A = \frac{tp + tn}{tp + fp + tn + fn}$$

$$P = \frac{tp}{tp + fp}$$

$$R = \frac{tp}{tp + fn}$$

Accuracy is not a good measure for when the class distribution is not even. For this reason, precision and recall are ubiquitous. However, there are usually tradeoffs between precision and recall when a system is tuned. A combined measure is thus required to capture this tradeoff as a single value. For this reason, the F measure is defined which is the weighted harmonic mean of precision and recall. it is calculated as:

$$\begin{aligned} F &= \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}} \\ &= \frac{(\beta^2+1)PR}{\beta^2 P + R}, \quad where \quad \beta^2 = \frac{1-\alpha}{\alpha} \end{aligned}$$

It is common to use equally weighted precision and recall scores as balanced metric. This balanced score is written as $F_{\beta=1}$ or F_1. It can be calculated as:

$$F_1 = \frac{2PR}{P + R}$$

Appendix B

Acronyms for Conference Titles

ACL EACL NAACL SemEval ANLP
EMNLP SIGAAN SIGBIOMED SIGDAT SIGDIAL
SIGFSM SIGGEN SIGHAN SIGLEX SIGMEDIA
SIGMOL SIGMT SIGNLL SIGPARSE SIGMORPHON
SIGSEM SIGSEMITIC SIGSLPAT SIGWAC COLING
HLT IJCNLP LREC PACLIC Rocling
TINLAP ALTA RANLP MUC Tipster

Appendix C

Citation Function Corpus Sample

```
<PAPER>
 <METADATA>
   <FILENO>0001012</FILENO>
   <APPEARED>ACL 1999</APPEARED>
 </METADATA>
 <TITLE> Measures of Distributional Similarity</TITLE>
 <AUTHORLIST>
   <AUTHOR>Lillian Lee</AUTHOR>
 </AUTHORLIST>
 <ABSTRACT>
   <A-S ID="A-0">We study distributional similarity measures for the
   purpose of improving probability estimation for unseen
   cooccurrences .</A-S> <A-S ID="A-1">Our contributions are three-fold : an empirical
   comparison of a broad range of measures ; a classification of
   similarity functions based on the information that they incorporate
   ; and the introduction of a novel function that is superior at
   evaluating potential proxy distributions .</A-S>
 </ABSTRACT>
 <BODY>
   <DIV DEPTH="1">
     <HEADER ID="H-0"> Introduction </HEADER>
     <P>
       <S ID="S-0">An inherent problem for statistical methods in
       natural language processing is that of sparse data -- the
       inaccurate representation in any training corpus of the
       probability of low frequency events .</S>
       <S ID="S-1">In
       particular , reasonable events that happen to not occur in the
       training set may mistakenly be assigned a probability of zero
       .</S>
       ...
```

Appendix D

Blog data

D.1 Blog URLs

Title	Author	URL	Count
?- true	Aleks Dimit	`http://a-dimit.blogspot.com/`	21
Alphabit Glottophilia	Isabella Chiari	`http://www.alphabit.net/blog/blogger.html`	0
Amy Iris	Amy Iris (bot)	`http://blog.amyiris.com/`	25
AndyHickl.com	Andy Hickl	`http://andyhickl.com/`	0
Apperceptual	Peter Turney	`http://apperceptual.wordpress.com/`	65
Attempted Axiomatisation	David Petar Novakovic	`http://dpn.name/`	0
Automatic Mind	Niels Ott	`http://www.drni.de/niels/cl/blog/`	8
Computational Linguistics	Ali Reza Ebadat	`http://computationallinguistic.blogspot.com/`	153
Computational Linguistics	Alexander Valet	`http://www.alexandervalet.com/blog/`	0
Data Mining: Natural Language Processing	Matthew Hurst	`http://datamining.typepad.com/`	50
David R. MacIver	David R. MacIver	`http://www.drmaciver.com/`	239
Earning My Turns	Fernando Pereira	`http://earningmyturns.blogspot.com/`	306
Emerging Computational Linguistics	Bob New	`http://emergingcl.com/`	38

Evri	Evri	`http://blog.evri.com/`	37
Hacklog: Blogamundo	Patrick Hall	`http://blogamundo.net/dev/`	241
hakia	Hakia	`http://blog.hakia.com/`	38
Information Engineering	Dragomir Radev	`http://lada.si.umich.edu:8080/wordpress/`	61
Information Retrieval on the Live Web	Paul Ogilvie	`http://livewebir.com/blog/`	11
Jeff's Search Engine Caffe	Jeff Dalton	`http://www.searchenginecaffe.com/`	392
Language Log	Mark Liberman et al.	`http://languagelog.ldc.upenn.edu/nll/`	5504
Language Wrong	Roddy Lindsay	`http://languagewrong.tumblr.com/`	12
LarKC	Large Knowledge Collider	`http://blog.larkc.eu/`	115
LexaBlog	Lexalytics	`http://www.lexalytics.com/lexablog/`	114
Lingformant: The Science of Linguistics in the News	Vili Maunula	`http://lingformant.vertebratesilence.com/`	684
LingPipe	Bob Carpenter	`http://lingpipe-blog.com/`	246
Manos Tsagkias	Manos Tsagkias	`http://staff.science.uva.nl/~tsagias/`	18
Matthew L. Jockers	Matthew L. Jockers	`https://www.stanford.edu/~mjockers/cgi-bin/drupal/?q=blog`	7
Mendicant Bug	Jason Adams	`http://mendicantbug.com/`	517
Misc Research Stuff	Delip Rao	`http://resnotebook.blogspot.com/`	59
mSpoke	mSpoke	`http://www.mspoke.com/blog/`	26
Nathan Sanders : Journal	Nathan Sanders	`http://sandersn.com/blog/index.php`	272
Natural Language Processing	Nisha	`http://khassanali-nlp-research.blogspot.com/`	14
Natural Language Processing Blog	Hal Daume	`http://nlpers.blogspot.com/`	201
Nerd Industries	Stuart Robinson	`http://prospero.bluescarf.net/stuart/`	48

NLPadLog		http://www.nlpado.de/blog/	154
Office Natural Language Team Blog	Microsoft	http://blogs.msdn.com/naturallanguage/default.aspx	69
Official Google Research Blog	Google	http://googleresearch.blogspot.com/	73
OpenCalais	OpenCalais	http://opencalais.com/blog	43
Outer Thoughts: Computational Linguistics	Alexandre Rafalovitch	http://blog.outerthoughts.com/category/computational-linguistics/	37
Powerset	Powerset	http://www.powerset.com/blog/	48
Probably Irrelevant	Fernando Diaz et al.	http://probablyirrelevant.org/	10
Ramifications of a Linguist's Life	bloggy007	http://ramslifeofalinguist.blogspot.com/	73
Research Log		http://topicmodels.wordpress.com/	20
Science for SEO	Marie-Claire Jenkins	http://scienceforseo.blogspot.com/	281
Semantic Hacker	textwise	http://blog.semantichacker.com/	62
Semantics etc.	Kai von Fintel	http://semantics-online.org/	160
Shahzad Khan	Shahzad Khan	http://blog.whyztech.com/	38
Similarity Blog	Krzysztof Janowicz	http://www.similarity-blog.de/	42
STRUCTURED LEARNING	Fernando Pereira	http://structlearn.blogspot.com/	8
Surprise and Coincidence - musings from the long tail	Ted Dunning	http://tdunning.blogspot.com/	20
Synthse	Andre Vellino	http://synthese.wordpress.com/	95
Technologies du langage	Jean Vronis	http://aixtal.blogspot.com/	0
Text and Artificial Intelligence	Shahzad Khan	http://textai.blogspot.com/	5
Text Technologies	Monash Research	http://www.texttechnologies.com/	287
The JamiQ Report	jamiq	http://blog.jamiq.com/	196

The Lousy Linguist	Chris	http://thelousylinguist.blogspot.com/	213
The Noisy Channel	Daniel Tunkelang	http://thenoisychannel.com/	714
The science of searching	Tristan Thomas Teunissen	http://www.w3lab.nl/blog/	13
thought process	Ken Reisman	http://kreisman.wordpress.com/	4
window office	Jon Elsas	http://windowoffice.tumblr.com/	249
	Anil Eklayva	http://anileklavya.wordpress.com/	139

D.2 Annotated Blog XML (Example Extract)

```
<?xml version="1.0" encoding="utf-8"?> <post>
  <url>http://datamining.typepad.com/data_mining/2009/04/acl-2009-accepted-papers.html</url>
  <title>ACL 2009 Accepted Papers</title>
  <text>
    <s>The list is up.</s>
    <s>There are a few papers on sentiment mining. </s>
    <s type="p">Among them, I find this title to be the most
    interesting: <link type="text" aclid="P09-1079">Mine the Easy and
    Classify t        he Hard: Experiments with Automatic Sentiment
    Classification</link>, by Sajib Dasgupta and Vincent Ng</s>.</text>
</post>
```

Appendix E

Annotation Instructions

Authors of scientific articles and research papers use citations to refer to existing related work. While citation sentences are inherently objective and do not carry any sentiment towards the cited article, there still exist some instances with a positive or negative sentiment. Given below are about 100 sentences and you just need to mark them as being Neutral, Positive or Negative towards the cited work (shown in the ⟦brackets⟧). For example, take a look at the following sentences.

1. As a learning algorithm for our classification model, we used Maximum Entropy (⟦Berger⟧et al. , 1996).
2. This is in contrast to purely statistical systems (e.g. , ⟦Brown⟧et al. , 1992), which are difficult to inspect and modify.
3. The Penn Treebank (⟦Marcus⟧et al. , 1993) is perhaps the most influential resource in Natural Language Processing (NLP).

Here, the first sentence is 'Neutral' about the paper by Berger et al. However the second sentence describes a shortcoming or problem in the cited paper (by Brown et al.) and thus carries 'Negative' sentiment towards it. Similarly, the third sentence describes a good trait of the cited paper (Marcus et al.) and is thus 'Positive'.

It is okay if you don't understand the text at times. There may be errors in the text due to imperfect scanning, so just answer what seems best to you. It shouldn't take more than 20 minutes and would be of great help to my dissertation.

Appendix F

Lexical Resources

F.1 My Science Specific Lexicon

Negative:- burden, complicated, daunting, deficiencies, degrade, difficult, inability, lack, poor, restrict, unexplored, worse

Positive:- acceptance, accurately, adequately, aided, appealing, bestperforming, better, central, closely, competitive, considerable, convenient, de facto, dominant, dramatically, easier, easy, effective, effectiveness, efficient, efficiently, excellent, extremely fast, faster, favorably, good, high, highly correlates, high-quality, important, improve, improve the performance, improvements, inexpensive, infuential, intensively, interesting, most important, outperforms, overcome, pioneered, popular, power, predominantly, preferable, preferred, quite accurate, reasonable, reduces overfitting, robust, satisfactory, shown to, significant increases, simpler, state of the art, state-of-art, state-of-theart [sic], state-of-the-art, straightforward, substantially, success, successful, successfully, suitable, well, well known, well-founded, well-known, widely known, widely used

F.2 List of Negation terms by Wilson et al. (2005)

no, not, n't, never, neither, nor, none, nobody, nowhere, nothing, cannot, can not, without, no one, no way

F.3 List of Work Nouns by Teufel (2010)

account, algorithm, analysis, analyses, approach, approaches, application, applications, architecture, architectures, characterization, characterisation, component, components, corpus, corpora, design, designs, evaluation, evaluations, example, examples, experiment, experiments, extension, extensions, evaluation, formalism, formalisms, formalization, formalizations, formalization, formalizations, formulation, formulations, framework, frameworks, implementation, implementations, investigation, investigations, machinery, machineries, method, methods, methodology, methodologies, model, models, module, modules, paper, papers, process, processes, procedure, procedures, program, programs, prototype, prototypes, research, researches, result, results, strategy, strategies, system, systems, technique, techniques, theory, theories, tool, tools, treatment, treatments, work, works

F.4 My List of Connectors

also, although, besides, but, despite, even though, furthermore, however, in addition, in spite of, instead, instead of, moreover, nonetheless, on the contrary, on the other hand, regardless of, still, then, though, whereas, while, yet

Bibliography

A. Abu-Jbara and D. Radev. Coherent citation-based summarization of scientific papers. In *Proc. of ACL*, 2011.

A. Abu Jbara and D. Radev. Umichigan: A conditional random field model for resolving the scope of negation. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 328–334, Montréal, Canada, 7-8 June 2012. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/S12-1043`.

T.W. Anderson and R.R. Bahadur. Classification into two multivariate normal distributions with different covariance matrices. *Annals of Mathematical Statistics*, 33(2): 420–431, 1962.

S. Argamon-Engelson, M. Koppel, and G. Avneri. Style-based text categorization: What newspaper am i reading. In *Proc. of the AAAI Workshop on Text Categorization*, pages 1–4, 1998.

A. Athar. Sentiment analysis of citations using sentence structure-based features. *ACL HLT 2011*, page 81, 2011.

A. Athar and S. Teufel. Context-enhanced citation sentiment detection. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 597–601, Montréal, Canada, June 2012a. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/N12-1073`.

A. Athar and S. Teufel. Detection of implicit citations for sentiment detection. In *Proceedings of the Workshop on Detecting Structure in Scholarly Discourse*, pages 18–26, Jeju Island, Korea, July 2012b. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W12-4303`.

A. Aue and M. Gamon. Customizing sentiment classifiers to new domains: a case study. In *Proceedings of RANLP*, volume 49. Citeseer, 2005.

S. Baccianella, A. Esuli, and F. Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Seventh conference on International Language Resources and Evaluation, Malta. Retrieved May*, volume 25, page 2010, 2010.

J. Bar-Ilan. Which h-index? - a comparison of wos, scopus and google scholar. *Scientometrics*, 74(2):257–271, 2008.

R. Barzilay and M. Elhadad. Using lexical chains for text summarization. In Inderjeet Mani and Mark T. Maybury, editors, *Proceedings of the ACL/EACL-97 Workshop on Intelligent Scalable Text Summarization*, 1997.

R. Barzilay and M. Lapata. Modeling local coherence: An entity-based approach. *Computational Linguistics*, (1):1–34, 2008.

R. Bekkerman, H. Raghavan, J. Allan, and K. Eguchi. Interactive clustering of text collections according to a user-specified criterion. In *Proceedings of IJCAI*, volume 20, 2007.

A.L. Berger, V.J.D. Pietra, and S.A.D. Pietra. A maximum entropy approach to natural language processing. *Computational linguistics*, 22(1):39–71, 1996.

S. Bergsma, E. Pitler, and D. Lin. Creating robust supervised classifiers via web-scale n-gram data. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 865–874, Uppsala, Sweden, July 2010. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P10-1089`.

T. Bienz, R. Cohn, and J. Meehan. *Portable document format reference manual.* Addison-Wesley, 1993.

S. Bird, R. Dale, B.J. Dorr, B. Gibson, M. Joseph, M.Y. Kan, D. Lee, B. Powley, D.R. Radev, and Y.F. Tan. The acl anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In *Proc. of the 6th International Conference on Language Resources and Evaluation Conference (LREC08)*, pages 1755–1759, 2008.

J. Blitzer, M. Dredze, and F. Pereira. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Annual Meeting-Association For Computational Linguistics*, volume 45, page 440, 2007.

J. Bollen, H. Mao, and X. Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011.

M. Bondi. The discourse function of contrastive connectors in academic abstracts. *PRAGMATICS AND BEYOND NEW SERIES*, pages 139–156, 2004.

E. Breck, Y. Choi, and C. Cardie. Identifying expressions of opinion in context. In *Twentieth International Joint Conference on Artificial Intelligence (IJCAI)*, 2007.

S. Brody and N. Diakopoulos. Cooooooooooooooollllllllllllll!!!!!!!!!!!!!!: using word lengthening to detect sentiment in microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 562–570. Association for Computational Linguistics, 2011.

J. Carletta. Assessing agreement on classification tasks: the kappa statistic. *Computational linguistics*, 22(2):249–254, 1996.

C.C. Chang and C.J. Lin. LIBSVM: a library for support vector machines, 2001. *Software available at http://www.csie.ntu.edu.tw/cjlin/libsvm*, 2001.

J.T. Chang, H. Schütze, and R.B. Altman. Creating an online dictionary of abbreviations from medline. *Journal of the American Medical Informatics Association*, 9(6):612–620, 2002.

P. Chen, H. Xie, S. Maslov, and S. Redner. Finding scientific gems with google's pagerank algorithm. *Journal of Infometrics*, 1(1):8–15, 2007.

Y. Choi and C. Cardie. Adapting a polarity lexicon using integer linear programming for domain-specific sentiment classification. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*, pages 590–598. Association for Computational Linguistics, 2009.

J. Cohen et al. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

S. Cole. The growth of scientific knowledge: Theories of deviance as a case study. *The idea of social structure: Papers in honor of Robert K. Merton*, pages 175–220, 1975.

C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

I.G. Councill, C.L. Giles, and M.Y. Kan. Parscit: An open-source crf reference string parsing package. In *Proceedings of LREC*, volume 2008. Citeseer, 2008.

I.G. Councill, R. McDonald, and L. Velikovich. What's great and what's not: learning to classify the scope of negation for improved sentiment analysis. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, pages 51–59. Association for Computational Linguistics, 2010.

S. Das and M. 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.

S. Dasgupta and V. Ng. Mine the easy, classify the hard: a semi-supervised approach to automatic sentiment classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 701–709. Association for Computational Linguistics, 2009.

M. Day, R.T. Tsai, C. Sung, C. Hsieh, C. Lee, S. Wu, K. Wu, C. Ong, and W. Hsu. Reference metadata extraction using a hierarchical knowledge representation framework. *Decision Support Systems*, 43(1):152–167, 2007.

C. De Kerchove and P.V. Dooren. The pagetrust algorithm: how to rank web pages when negative links are allowed. In *Proc. SIAM Int. Conf. on Data Mining*, pages 346–352, 2008.

M.C. de Marneffe and C.D. Manning. The stanford typed dependencies representation. In *Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation*, pages 1–8. Association for Computational Linguistics, 2008.

Y. EL-Manzalawy and V. Honavar. *WLSVM: Integrating LibSVM into Weka Environment*, 2005. Software available at `http://www.cs.iastate.edu/~yasser/wlsvm`.

C. Engström. Topic dependence in sentiment classification. University of Cambridge, 2004.

A. Esuli and F. Sebastiani. Determining term subjectivity and term orientation for opinion mining. In *Proceedings the 11th Meeting of the European Chapter of the Association for Computational Linguistics (EACL-2006)*, pages 193–200, 2006a.

A. Esuli and F. Sebastiani. SentiWordNet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC*, volume 6, pages 417–422. Citeseer, 2006b.

B.D. Eugenio and M. Glass. The kappa statistic: A second look. *Computational linguistics*, 30(1):95–101, 2004.

B. Finney. *The reference characteristics of scientific texts.* PhD thesis, City University (London, England), 1979.

M. Gamon. Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. In *Proceedings of the 20th international conference on Computational Linguistics*, page 841. Association for Computational Linguistics, 2004.

E. Garfield. Is citation analysis a legitimate evaluation tool? *Scientometrics*, 1(4):359–375, 1979.

E. Garfield et al. Science citation index: A new dimension in indexing. *Science*, 144 (3619):649–654, 1964.

E. Garfield et al. Can citation indexing be automated. In *Statistical association methods for mechanized documentation, symposium proceedings*, pages 189–192, 1965.

M.A. Garzone. *Automated classification of citations using linguistic semantic grammars.* PhD thesis, The University of Western Ontario, 1997.

C.L. Giles, K.D. Bollacker, and S. Lawrence. Citeseer: An automatic citation indexing system. In *Proceedings of the third ACM conference on Digital libraries*, pages 89–98. ACM, 1998.

A. Go, R. Bhayani, and L. Huang. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, pages 1–12, 2009.

A.M. Green. Kappa statistics for multiple raters using categorical classifications. In *Proceedings of the 22nd annual SAS User Group International conference*, pages 1110–1115, 1997.

R. Guha, R. Kumar, P. Raghavan, and A. Tomkins. Propagation of trust and distrust. In *Proceedings of the 13th international conference on World Wide Web*, pages 403–412. ACM, 2004.

G.K. Gupta. *Introduction to Data Mining with Case Studies.* PHI Learning Pvt. Ltd., 2006.

D. Hall, D. Jurafsky, and C.D. Manning. Studying the history of ideas using topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 363–371. Association for Computational Linguistics, 2008.

M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten. The weka data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, 2009.

S. Hanney, I. Frame, J. Grant, M. Buxton, T. Young, and G. Lewison. Using categorisations of citations when assessing the outcomes from health research. *Scientometrics*, 65(3):357–379, 2005.

V. Hatzivassiloglou and K. McKeown. A quantitative evaluation of linguistic tests for the automatic prediction of semantic markedness. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pages 197–204, Cambridge, Massachusetts, USA, June 1995. Association for Computational Linguistics. doi: 10.3115/981658.981685. URL `http://www.aclweb.org/anthology/P95-1027`.

V. Hatzivassiloglou and K.R. McKeown. Predicting the semantic orientation of adjectives. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, page 181. Association for Computational Linguistics, 1997.

J.E. Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, 102(46):16569, 2005.

B. Hollingsworth, I. Lewin, and D. Tidhar. Retrieving hierarchical text structure from typeset scientific articles–a prerequisite for e-science text mining. In *Proc. of the 4th UK E-Science All Hands Meeting*, pages 267–273, 2005.

M.J. Hornsey, E. Robson, J. Smith, S. Esposo, and R.M. Sutton. Sugaring the pill: Assessing rhetorical strategies designed to minimize defensive reactions to group criticism. *Human Communication Research*, 34(1):70–98, 2008.

B.M. Howe et al. *Introduction to Academic English Writing (Paperback).* Ewha Womans University Press, 2010.

C.W. Hsu, C.C. Chang, and C.J. Lin. A practical guide to support vector classification, 2003. `http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf`, 2009.

M. Hu and B. Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM, 2004.

K. Hyland. Hedging in academic writing and eap textbooks. *English for Specific Purposes*, 13(3):239–256, 1994.

L. Jia, C. Yu, and W. Meng. The effect of negation on sentiment analysis and retrieval effectiveness. In *Proceeding of the 18th ACM conference on Information and knowledge management*, pages 1827–1830. ACM, 2009.

T. Joachims. Text categorization with support vector machines: Learning with many relevant features. *Machine Learning: ECML-98*, pages 137–142, 1998.

A. Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Advances in neural information processing systems*, 14:841, 2002.

M. Joshi and Penstein-Rose. Generalizing dependency features for opinion mining. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 313–316. Association for Computational Linguistics, 2009.

J.S. Justeson and S.M. Katz. Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural language engineering*, 1(01):9–27, 1995. ISSN 1351-3249.

S.D. Kamvar, M.T. Schlosser, and H. Garcia-Molina. The eigentrust algorithm for reputation management in p2p networks. In *Proceedings of the 12th international conference on World Wide Web*, pages 640–651. ACM, 2003.

D. Kaplan, R. Iida, and T. Tokunaga. Automatic extraction of citation contexts for research paper summarization: A coreference-chain based approach. In *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries*, pages 88–95. Association for Computational Linguistics, 2009.

S.S. Keerthi and C.J. Lin. Asymptotic behaviors of support vector machines with gaussian kernel. *Neural computation*, 15(7):1667–1689, 2003.

S. Khan. *Negation and Antonymy in Sentiment*. PhD thesis, University of Cambridge, 2007.

D.Y. Kim and P.B. Webber. Implicit references to citations: A study of astronomy papers. 2006.

S.M. Kim and E. Hovy. Determining the sentiment of opinions. In *Proceedings of the 20th international conference on Computational Linguistics*, page 1367. Association for Computational Linguistics, 2004.

R. Kohavi. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *KDD*, pages 202–207, 1996.

E. Kouloumpis, T. Wilson, and J. Moore. Twitter sentiment analysis: The good the bad and the omg! In *Fifth International AAAI Conference on Weblogs and Social Media*, 2011.

K. Krippendorff. *Content analysis: An introduction to its methodology.* Sage Publications, Inc, 1980.

J. Kunegis, A. Lommatzsch, and C. Bauckhage. The slashdot zoo: mining a social network with negative edges. In *Proceedings of the 18th international conference on World wide web*, pages 741–750. ACM, 2009.

J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of ICML*, pages 282–289. Citeseer, 2001.

L.S. Larkey, P. Ogilvie, M.A. Price, and B. Tamilio. Acrophile: an automated acronym extractor and server. In *Proceedings of the fifth ACM conference on Digital libraries*, pages 205–214. ACM, 2000.

H. Lee, Y. Peirsman, A. Chang, N. Chambers, M. Surdeanu, and D. Jurafsky. Stanford's multi-pass sieve coreference resolution system at the conll-2011 shared task. *ACL HLT 2011*, 2011.

A. Lehrer. Markedness and antonymy. *Journal of Linguistics*, 21(2):397–429, 1985.

D.D. Lewis. Evaluating text categorization. In *Proceedings of Speech and Natural Language Workshop*, pages 312–318, 1991.

D.D. Lewis. Naive (bayes) at forty: The independence assumption in information retrieval. In *Machine learning: ECML-98*, pages 4–15. Springer, 1998.

Huajing Li, Isaac Councill, Wang-Chien Lee, and C Lee Giles. Citeseerx: an architecture and web service design for an academic document search engine. In *Proceedings of the 15th international conference on World Wide Web*, pages 883–884. ACM, 2006.

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

N. Ma, J. Guan, and Y. Zhao. Bringing pagerank to the citation analysis. *Information Processing and Management*, 44(2):800–810, 2008a.

N. Ma, J. Guan, and Y. Zhao. Bringing pagerank to the citation analysis. *Information Processing & Management*, 44(2):800–810, 2008b.

A.L. Maas, R.E. Daly, P.T. Pham, D. Huang, A.Y. Ng, and C. Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P11-1015`.

M.H. MacRoberts and B.R. MacRoberts. The negational reference: Or the art of dissembling. *Social Studies of Science*, 14(1):91–94, 1984. ISSN 0306-3127.

C.D. Manning, P. Raghavan, and H. Schutze. *Introduction to information retrieval*, volume 1. Cambridge University Press Cambridge, 2008.

Y. Mao and G. Lebanon. Sequential models for sentiment prediction. In *ICML Workshop on Learning in Structured Output Spaces*. Citeseer, 2006.

M.L. McHugh. Interrater reliability: the kappa statistic. *Biochemia Medica*, 22(3):276–282, 2012.

L.I. Meho and K. Yang. Impact of data sources on citation counts and rankings of lis faculty: Web of science versus scopus and google scholar. *Journal of the american society for information science and technology*, 58(13):2105–2125, 2007.

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

K. Moilanen and S. Pulman. Sentiment composition. In *Proceedings of the Recent Advances in Natural Language Processing International Conference (RANLP-2007)*, pages 378–382, Borovets, Bulgaria, September 27-29 2007. URL `http://users.ox.ac.uk/~wolf2244/sentCompRANLP07Final.pdf`.

R. Morante and E. Blanco. Sem 2012 shared task: Resolving the scope and focus of negation. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics*, 2012.

D. Nadeau and P. Turney. A supervised learning approach to acronym identification. *Advances in Artificial Intelligence*, pages 79–93, 2005.

T. Nakagawa, K. Inui, and S. Kurohashi. Dependency tree-based sentiment classification using CRFs with hidden variables. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 786–794. Association for Computational Linguistics, 2010. ISBN 1932432655.

H. Nanba and M. Okumura. Towards multi-paper summarization using reference information. In *IJCAI*, volume 16, pages 926–931. Citeseer, 1999.

J. O'Connor. Citing statements: Computer recognition and use to improve retrieval. *Information Processing & Management*, 18(3):125–131, 1982.

L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. 1999.

B. Pang and L. Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 271. Association for Computational Linguistics, 2004.

B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics, 2002.

B.C. Peritz. A classification of citation roles for the social sciences and related fields. *Scientometrics*, 5(5):303–312, 1983.

S.B. Pham and A. Hoffmann. Extracting positive attributions from scientific papers. In *Discovery Science*, pages 39–45. Springer, 2004.

L. Polanyi and A. Zaenen. Contextual valence shifters. *Computing attitude and affect in text: Theory and applications*, pages 1–10, 2006.

V. Qazvinian and D.R. Radev. Scientific paper summarization using citation summary networks. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 689–696. Association for Computational Linguistics, 2008.

V. Qazvinian and D.R. Radev. Identifying non-explicit citing sentences for citation-based summarization. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 555–564. Association for Computational Linguistics, 2010.

D. R. Radev, P. Muthukrishnan, and V. Qazvinian. The ACL anthology network corpus. In *Proceedings, ACL Workshop on Natural Language Processing and Information Retrieval for Digital Libraries*, Singapore, 2009.

J. Read. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In *Proceedings of the ACL Student Research Workshop*, pages 43–48. Association for Computational Linguistics, 2005.

T. Rietveld and R. Van Hout. *Statistical techniques for the study of language behaviour*. Berlijn: Mouton de Gruyter, 1993.

E. Riloff and J. Wiebe. Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 conference on Empirical methods in natural language processing-Volume 10*, pages 105–112. Association for Computational Linguistics, 2003.

A. Ritchie. *Citation context analysis for information retrieval*. PhD thesis, PhD thesis, University of Cambridge, 2008.

S.B. Roy, M. De Cock, V. Mandava, S. Savanna, B. Dalessandro, C. Perlich, W. Cukierski, and B. Hamner. The microsoft academic search dataset and kdd cup 2013. In *Proceedings of the 2013 KDD Cup 2013 Workshop*, page 1. ACM, 2013.

R.E. Schapire and Y. Singer. Boostexter: A boosting-based system for text categorization. *Machine learning*, 39(2):135–168, 2000.

A. Siddharthan and S. Teufel. Whose idea was this, and why does it matter? attributing scientific work to citations. *Proceedings of NAACL/HLT-07*, 2007.

H. Small. Citation context analysis. *Progress in communication sciences*, 3:287–310, 1982.

S. Sohn, S. Wu, and C.G. Chute. Dependency parser-based negation detection in clinical narratives. *AMIA Summits on Translational Science Proceedings*, 2012:1, 2012.

I. Spiegel-Rosing. Science studies: Bibliometric and content analysis. *Social Studies of Science*, pages 97–113, 1977.

M. Taboada and J. Grieve. Analyzing appraisal automatically. In *Proceedings of AAAI Spring Symposium on Exploring Attitude and Affect in Text (AAAI Technical Report SS-04-07)*, pages 158–161, 2004.

M. Taboada, C. Anthony, and K. Voll. Methods for creating semantic orientation dictionaries. In *Conference on Language Resources and Evaluation (LREC)*, pages 427–432. Citeseer, 2006.

M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307, 2011.

O. Täckström and R. McDonald. Discovering fine-grained sentiment with latent variable structured prediction models. In *Proceedings of the 33rd European Conference on Information Retrieval (ECIR 2011), Dublin, Ireland*, 2011.

K. Taghva and J. Gilbreth. Recognizing acronyms and their definitions. *International Journal on Document Analysis and Recognition*, 1(4):191–198, 1999.

Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. Arnetminer: extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 990–998. ACM, 2008.

S. Teufel. *Argumentative zoning: Information extraction from scientific text*. PhD thesis, Citeseer, 1999.

S. Teufel. Argumentative zoning for improved citation indexing. *Computing Attitude and Affect in Text: Theory and Applications*, pages 159–169, 2006.

S. Teufel. *The Structure of Scientific Articles: Applications to Citation Indexing and Summarization*. Csli Studies in Computational Linguistics. Center for the Study of Language and Inf, 2010. ISBN 9781575865553. URL `http://books.google.co.uk/books?id=j1qyGAAACAAJ`.

S. Teufel, A. Siddharthan, and D. Tidhar. An annotation scheme for citation function. In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*. Association for Computational Linguistics, 2006a.

S. Teufel, A. Siddharthan, and D. Tidhar. Automatic classification of citation function. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 103–110. Association for Computational Linguistics, 2006b. ISBN 1932432736.

M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, 62(2):419, 2011.

G. Thompson and Y. Yiyun. Evaluation in the reporting verbs used in academic papers. *Applied linguistics*, 12(4):365, 1991. ISSN 0142-6001.

Rakshit Trivedi and Jacob Eisenstein. Discourse connectors for latent subjectivity in sentiment analysis. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 808–813, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/N13-1100`.

A. Tumasjan, T.O. Sprenger, P.G. Sandner, and I.M. Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. *ICWSM*, 10:178–185, 2010.

P.D. Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 417–424. Association for Computational Linguistics, 2002.

S. Wang and C. Manning. Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)*, 2012. URL `pubs/simple_sentiment.pdf`.

J. Wiebe and R. Mihalcea. Word sense and subjectivity. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 1065–1072. Association for Computational Linguistics, 2006.

J. Wiebe and E. Riloff. Creating subjective and objective sentence classifiers from unannotated texts. *Computational Linguistics and Intelligent Text Processing*, pages 486–497, 2005.

J.M. Wiebe, R.F. Bruce, and T.P. O'Hara. Development and use of a gold-standard data set for subjectivity classifications. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 246–253. Association for Computational Linguistics, 1999.

F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6): 80–83, 1945.

Y. Wilks and M. Stevenson. The grammar of sense: Using part-of-speech tags as a first step in semantic disambiguation. *Natural Language Engineering*, 4(02):135–143, 1998.

T. Wilson, J. Wiebe, and R. Hwa. Just how mad are you? finding strong and weak opinion clauses. In *Proceedings of the National Conference on Artificial Intelligence*, pages 761–769. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2004.

T. Wilson, P. Hoffmann, S. Somasundaran, J. Kessler, J. Wiebe, Y. Choi, C. Cardie, E. Riloff, and S. Patwardhan. Opinionfinder: A system for subjectivity analysis. In *Proceedings of HLT/EMNLP on Interactive Demonstrations*, pages 34–35. Association for Computational Linguistics, 2005.

T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing Contextual Polarity: an exploration of features for phrase-level sentiment analysis. *Computational Linguistics*, 35(3):399–433, 2009. ISSN 0891-2017.

Y. Wu, Q. Zhang, X. Huang, and L. Wu. Phrase dependency parsing for opinion mining. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3*, pages 1533–1541. Association for Computational Linguistics, 2009.

J.S. Yedidia, W.T. Freeman, and Y. Weiss. Understanding belief propagation and its generalizations. *Exploring artificial intelligence in the new millennium*, 8:236–239, 2003.

A. Yessenalina, Y. Choi, and C. Cardie. Automatically generating annotator rationales to improve sentiment classification. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 336–341. Association for Computational Linguistics, 2010.

H. Yu and V. Hatzivassiloglou. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 conference on Empirical methods in natural language processing-Volume 10*, page 136. Association for Computational Linguistics, 2003.

P Yu and H Van de Sompel. Networks of scientific papers. *Science*, 169:510–515, 1965.

J.M. Ziman. *Public Knowledge: An essay concerning the social dimension of science.* Cambridge Univ. Press, College Station, Texas, 1968.