University of Sussex

Unsupervised and Knowledge-poor Approaches to Sentiment Analysis

Taras Zagibalov

Submitted for the degree of Doctor of Philosophy

University of Sussex

September 2010

Declaration

I hereby declare that this thesis has not been and will not be, submitted in whole or in part to another University for the award of any other degree.

Signature:.....

Taras Zagibalov

UNIVERSITY OF SUSSEX

TARAS ZAGIBALOV (DPHIL)

UNSUPERVISED AND KNOWLEDGE-POOR APPROACHES TO SENTIMENT ANALYSIS

SUMMARY

Sentiment analysis focuses upon automatic classification of a document's sentiment (and more generally extraction of opinion from text). Ways of expressing sentiment have been shown to be dependent on what a document is about (domain-dependency). This complicates supervised methods for sentiment analysis which rely on extensive use of training data or linguistic resources that are usually either domain-specific or generic. Both kinds of resources prevent classifiers from performing well across a range of domains, as this requires appropriate in-domain (domain-specific) data.

This thesis presents a novel unsupervised, knowledge-poor approach to sentiment analysis aimed at creating a domain-independent and multilingual sentiment analysis system. The approach extracts domain-specific resources from documents that are to be processed, and uses them for sentiment analysis. This approach does not require any training corpora, large sets of rules or generic sentiment lexicons, which makes it domain- and language-independent but at the same time able to utilise domain- and language-specific information.

The thesis describes and tests the approach, which is applied to different data, including customer reviews of various types of products, reviews of films and books, and news items; and to four languages: Chinese, English, Russian and Japanese. The approach is applied not only to binary sentiment classification, but also to three-way sentiment classification (positive, negative and neutral), subjectivity classification of documents and sentences, and to the extraction of opinion holders and opinion targets. Experimental results suggest that the approach is often a viable alternative to supervised systems, especially when applied to large document collections.

Acknowledgements

I owe my deepest gratitude to my academic supervisor John Carroll for valuable advice and friendly guidance, for encouragement and support. I am also grateful to Bill Keller and David Weir, my Thesis committee member, for their guidance and suggestions.

I am indebted to my colleagues for their support, especially to Jonathon Read, who was always ready to help and advise me. I would like to deeply thank my friend Martine Self and her family for their help and friendship.

I am grateful to Ford Foundation Fellowship Program who sponsored my research and stay in the UK.

I owe a lot to my parents, Maria and Evgenij, for everything they have done for me, for all their love and care.

This thesis would not have been possible without the love, support and patience of my beloved wife Olesya. Thank you, my dear!

Contents

List of Tables **ix**

List of Figures **xii**

1 Introduction **1**

1.1 Background 1

1.2 Research Overview 3

1.2.1 The Scientific Question 3

1.2.2 Hypotheses 3

1.2.3 Contributions of this Work 4

1.3 Outline of the Thesis 5

1.3.1 Approach and Methodology 5

1.3.2 Structure of the Thesis 6

2 Literature Review **7**

2.1 Study of Affect 7

2.1.1 Private States 7

2.1.2 Categorical and Dimensional Paradigms 8

2.1.3 Affect Across Cultures 9

2.2 Sentiment Analysis 10

2.2.1 Tasks 10

2.2.2 Techniques 18

2.2.3 Features 23

2.2.4 Levels 26

2.2.5 Text Types and Domains 27

2.3 Resource Development 29

2.4 Challenges of Sentiment Analysis 31

2.4.1 Cross-Domain Approaches 31

2.4.2 Cross-Language Approaches 33

3 Features for Chinese Sentiment Classification 35

3.1 The 'Word' in Chinese Language Processing 35

3.1.1 Preliminary Word Segmentation of Chinese Texts 37

3.1.2 Preliminary Segmentation Experiment 38

3.2 Words and Characters as Features for Sentiment Classification 40

3.2.1 Basic Concepts 40

3.2.2 Experimental Data and Classification Algorithm 43

3.2.3 Evaluation Metrics and Statistical Significance Test 43

3.3 Experiments with Classification Units 45

3.3.1 Unigram-Based Classification 46

3.3.2 Zone-Based Classification 48

3.3.3 Sentence-Based Classification 50

3.3.4 Discussion 50

3.4 Sentiment Score Extensions 54

3.4.1 Negation Check 54

3.4.2 Length Ratio 57

3.4.3 Discussion 59

3.5 Summary 60

3.5.1 Accuracy 60

3.5.2 Coverage 61

3.5.3 Skew 61

3.5.4 Precision 61

3.5.5 Conclusion 61

4 Classifier Improvements and Extensions 63

4.1 Dictionary Adjustment 64

4.1.1 Adjustment to Corpus 64

4.1.2 Adjustment to Topic 67

4.1.3 Discussion 68

4.2 Vocabulary Extraction 69

4.2.1 Seed-Based Approach 69

4.2.2 Automatic Seed Word Selection 71

4.2.3 Iterative Approach 80

4.2.4 Discussion . . . 88
4.3 Performance Improvements . . . 89
4.3.1 Score Difference . . . 90
4.3.2 Zone Difference . . . 92
4.3.3 Using Supervised Techniques to Extend Unsupervised Classifier . . . 94
4.3.4 Comparison of Supervised and Unsupervised Classifiers . . . 101
4.4 Discussion . . . 102
4.5 Conclusion . . . 103

5 Multilingual Sentiment Classification 106
5.1 Data . . . 107
5.1.1 Language-Specific Issues . . . 107
5.1.2 Book Review Corpora . . . 109
5.1.3 Issues that may Affect Automatic Processing . . . 117
5.1.4 Movie Review Corpus . . . 119
5.2 Supervised Classification Experiments . . . 120
5.2.1 Lexical Unit Extraction . . . 121
5.2.2 Experimental results . . . 121
5.3 Unsupervised Classification Experiments . . . 123
5.3.1 Seed-Based Classification . . . 123
5.3.2 Classification Results . . . 127
5.3.3 Score Difference . . . 128
5.3.4 Zone Difference for Result Ranking . . . 130
5.3.5 Combining with Supervised Machine Learning Techniques . . . 130
5.4 Discussion . . . 131

6 Multi-Aspect Sentiment Analysis 135
6.1 Three-Way Classification . . . 135
6.1.1 Sentiment Classification . . . 137
6.1.2 Subjectivity Classification . . . 139
6.2 Sentence-Level Subjectivity and Sentiment Classification . . . 141
6.2.1 Data . . . 142
6.2.2 Classification Using an Existing Classifier . . . 142
6.2.3 Discussion . . . 145
6.2.4 Stand-Alone Subjectivity Classification . . . 145

6.2.5 Evaluation Results . . . 148
6.2.6 Discussion . . . 151
6.3 Opinion Holder and Opinion Target Extraction . . . 152
6.3.1 Overview of the Approach . . . 152
6.3.2 Language-specific Adjustment . . . 153
6.3.3 System Summary . . . 155
6.3.4 Experiments . . . 156
6.3.5 Discussion . . . 158
6.4 Conclusion . . . 160

7 Conclusion **161**
7.1 Unsupervised Sentiment Classification . . . 161
7.2 Other Tasks . . . 163
7.3 Cross-domain Sentiment Classification . . . 163
7.4 Multilingual Sentiment Classification . . . 164
7.5 Hypotheses . . . 165
7.6 Future Work . . . 166
7.7 Practical Implementation . . . 167

Bibliography **169**

List of Tables

3.1 Results of sentiment classification of product reviews from the web-site IT168, with and without segmentation. The features are NTU sentiment dictionary items. 40
3.2 Results of unigram-based sentiment classification using different types of features 46
3.3 Results of sentiment classification with the characters present only in a single class 47
3.4 Results of zone-based sentiment classification 50
3.5 Results of sentence-based sentiment classification 51
3.6 Precision of the unigram, zone-based and sentence-based sentiment classifiers 53
3.7 Results of unigram-based sentiment classification with negation 55
3.8 Results of zone-based sentiment classification with negation 56
3.9 Results of sentence-based sentiment classification with negation 57
3.10 Results of unigram-based sentiment classification with length ratio 58
3.11 Results of zone-based sentiment classification with length ratio 58
3.12 Results of sentence-based sentiment classification with length ratio 59
3.13 Results of unigram-based sentiment classification with length ratio and negation check combined 60
3.14 Results of word-based sentiment with different features 62

4.1 List of top 10 words 65
4.2 Results of word-based sentiment classification before and after feature adjustment 66
4.3 Results of combined classifier sentiment classification before and after feature adjustment 66
4.4 Average of the results of five runs on a test corpus of the word classifier sentiment classification before and after feature adjustment 66

4.5 Product types and sizes of the test corpora. 67
4.6 Classification results of different topics with sentiment vocabulary with and without topic-adjusted scores. 68
4.7 Single word seed lists 70
4.8 Multi-word seed lists 70
4.9 Results of the seed list classifier sentiment classification. 71
4.10 Classification results with the seed *good*, and seed lists *allPOS* and *all*. . . . 72
4.11 Seeds automatically identified for each corpus. 77
4.12 Classification results with the *allPOS* seed list and extracted seeds. 78
4.13 Classification results with only positive extracted seeds vs the same seeds augmented with generic negative seeds. 79
4.14 Classification results with only positive extracted seeds, the same seeds augmented with generic negative seeds and with the *all* seed list. 80
4.15 Results of sentiment classification of 10 iterations with seed list *all* applied to two topics *Mobile phones* and *Monitors*. 82
4.16 Results of sentiment classification on completion of iterations. 83
4.17 Top 10 positive lexical units found on completion of iterations. 85
4.18 Top 10 negative lexical units found on completion of iterations. 86
4.19 Classification results with *allPos* seed list and only positive extracted seeds. 87
4.20 Classification results with generic seeds and extracted seeds combined with generic negative seeds. 88
4.21 Classification results with the seed list *all* (*Seeds*) and the vocabulary-based classifier (*Vocabulary*) after a number of iterations. 89
4.22 Classification results with the seed list *all* and the automatically extracted seeds with generic negative lexical units (*ExtractedNeg*). 92
4.23 Supervised classifiers with the three feature sets 99
4.24 The NBm classifier with the three feature sets 100
4.25 The SVM classifier with the three feature sets 101
4.26 Supervised classifiers compared with unsupervised classifiers 102
4.27 Classification results with extracted seeds 103

5.1 Case forms of Russian adjectives 108
5.2 Overall quantitative measures of the English and Russian corpora. 110
5.3 Ways of expressing sentiment in the English Book Review Corpus (numbers of documents). 112

5.4 Ways of expressing sentiment in the Russian Book Review Corpus (numbers of documents). 112
5.5 Supervised classification results (10-fold cross-validation, lexical units). . . . 122
5.6 Supervised classification results (10-fold cross-validation, words). 122
5.7 The manually selected Russian seeds. 124
5.8 Semi-automatically extracted Russian seeds. 125
5.9 Semi-automatically extracted English seeds. 126
5.10 Russian book reviews: results of classification. 127
5.11 English corpora: results of classification. 128
5.12 Russian book reviews: results of classification. 129
5.13 English corpora: results of classification using score difference. 129
5.14 Russian book reviews: results of classification. 131
5.15 English corpora: results of classification using machine learning (NBm). . . 132

6.1 Extracted seeds 143
6.2 Subjectivity and sentiment classification results 143
6.3 Semi-automatically extracted English seeds 144
6.4 Manually-selected opinionated words (all glosses are very approximate). . . 147
6.5 Sizes of the lists of words. 148
6.6 Relevance and opinion results for Chinese (Traditional). 149
6.7 Relevance and opinion results for Chinese (Simplified). 150
6.8 Relevance and opinion results for Japanese. 150
6.9 Relevance and opinion results for English. 151
6.10 Opinion holder and target performance on the NTCIR-7 MOAT test sets. . 158

List of Figures

4.1 Classification results with the seed list *all* with the score difference technique. 90
4.2 Classification results with the seed list *all* and with the zone difference technique. 93
4.3 Classification results with the seed list *all* and the zone distance technique (Topics). 95
4.4 Classification results with extracted seeds and the zone distance technique (Topics). 96
4.5 Information retrieval simulation results with the seed list *all* and the zone distance technique. 97
4.6 Information retrieval simulation results with extracted seeds and the zone distance technique. 98

5.1 Distribution of documents by the number of words contained. 111
5.2 Information retrieval simulation results with the zone distance technique. . 130
5.3 Score difference results for the movie review corpus. 133

6.1 The distribution of Chinese customer reviews with respect to Sentiment Score and Sentiment Density. 139
6.2 The distribution of factual documents with respect to Sentiment Score and Sentiment Density. 140
6.3 The distribution of factual documents with respect to Sentiment Score and Sentiment Density with the NTU Sentiment Dictionary. 141

Chapter 1

Introduction

1.1 Background

This thesis is about the automated analysis of sentiment in written language. Sentiment analysis is concerned not with the topic or factual content in it, but rather with the opinion expressed in a document. Sentiment analysis has often been broken down into a set of subtasks, including subjectivity classification, opinion classification (sentiment classification), opinion holder and opinion target extraction, and feature-based opinion mining.

Opinion classification is usually framed as a two-way classification of positive and negative sentiment, and has been applied at different levels: phrases, sentences, documents and collections of documents. An opinion may have a holder (a person or a group that expresses an opinion) and a target (an object which is being discussed or evaluated). Feature-based opinion mining tries to find opinions about particular features of a product or service (as opposed to an overall opinion about something).

Automatic classification of document sentiment (and more generally extraction of opinion from text) has recently attracted much interest. One of the main reasons for this is the importance of such information to companies, other organizations, and individuals. Applications include marketing research tools that help a company see market or media reaction towards their brands, products or services. Another type of application is search engines that help potential purchasers make an informed choice of a product they want to buy. Such search engines include a sentiment classification subsystem that may not only present to a customer overall sentiment about a product, but also select positive or negative reviews to illustrate advantages and shortcomings of a product.

Automated sentiment analysis provides a range of possibilities for researchers in humanities whose studies involve analysis of large amount of human-generated data. For

example, in media studies one might be interested to see if sentiments regarding the same events are shared in mainstream media and in social media. Analysis of user-generated content may be very helpful in political studies. For example, monitoring of political debates in social media may help to estimates prospects of political candidates in elections or evaluate effectiveness of political campaigns. The study of "the language of hatred" contributes to efforts against political and religious extremism and intolerance. Many aspects of social studies may benefit from automatic analysis of sentiments expressed by people in ever-growing social networks. This approach offers unintrusive and fast access to large amount of data.

In recent white paper addressing the role of sentiment analysis in organisations, Grimes (2010) noted that "one axiom of full-circle sentiment analysis is ability to use all relevant sentiment sources". This obviously includes resources in different languages, of different genres and written in different styles. The most widely used approach to opinion and subjectivity classification is based on supervised machine learning, in which a system learns from human-annotated training data how to classify documents. However, a major obstacle for automatic classification of sentiment and subjectivity is often a lack of training data, which limits the applicability of approaches based on supervised machine learning. With the rapid growth in the amount of textual data and the emergence of new domains of knowledge it is virtually impossible to maintain corpora of annotated data that cover all – or even most – areas of interest. The cost of manual annotation also adds to the problem. Re-using the same corpus for training classifiers for new domains is also not effective: several studies report decreased accuracy in cross-domain classification (Engström, 2004; Read, 2005; Aue and Gamon, 2005). Indeed, a classifier trained in a film review domain might consider word *unpredictable* (e.g. *unpredictable plot*) to be used to express a positive characteristic. However, the same word in an car review might be a marker of a negative sentiment (e.g. *unpredictable steering*) (Turney, 2002). A similar problem has also been observed in classification of documents created over different time periods (Read, 2005). Some words were found to express a certain sentiment only for a definite period of time. The word *ice-axe*, for example, was a strong indicator of positive sentiment because it was frequently used in mostly positive reviews of a film that featured a particularly stirring scene involving this tool.

Rule-based or dictionary-based classifications also have similar limitations and they also rely on a large set of manually created resources used for classification.

A major current challenge, therefore, is to be able to automatically extract sentiment

information from a variety of documents in different languages and from different domains. Most existing solutions are based on adapting systems designed for one language (or domain) to another. Obviously, there are differences between cultures, languages and even within a language (consider the difference in the language used for evaluations of a company financial prospects in a business newspaper and reviews of a hard-rock festival in a participant's blog). Such differences make adaptation problematic. Porting a sentiment analysis system to new languages is even more difficult.

This thesis proposes an approach based on the idea of finding all data needed for classification within the documents to be classified. Domain-specific data is often hard to find, and generic resources, such as for example, sentiment lexicons, often fail to include all relevant markers of opinion. Even well-known and 'obvious' markers of sentiment may demonstrate a sharp twist in their meaning in certain domains. For example, Ghose et al. (2007) found that the word *good* is an indicator of negative sentiment in the domain of eBay customer reviews: to describe something really good customers tend to use *perfect* and *excellent*, reserving *good* for polite expression of negative appraisal (as in *the package is good (but might have been better)*).

To overcome this problem the approach investigated in this thesis is to bootstrap sentiment-related data from documents using a very limited number of seed lexical units. This approach is used across domains, as well as across languages.

1.2 Research Overview

1.2.1 The Scientific Question

The main goal of the research presented in this thesis is to investigate the extent to which it is possible to build an unsupervised domain-independent cross-lingual sentiment analysis system. Such a system could be of great utility due to the ever-growing amount of all kinds of unstructured information in different languages which often contain opinions and evaluations.

1.2.2 Hypotheses

The research explores five main hypotheses:

- Hypothesis 1: Unsupervised systems can be developed for performing sentiment analysis in different domains and in different languages that perform comparably with supervised systems.

- Hypothesis 2: Unsupervised and knowledge-poor sentiment analysis may not require much domain- or language-specific input. Such a system might require only a basic indication of what positive and negative sentiments are, in the form of lexical 'seeds'.
- Hypothesis 3: A sentiment-related vocabulary automatically extracted from a corpus can produce similar or better results compared to a specialised hand-built sentiment vocabulary.
- Hypothesis 4: An automatically acquired training corpus in conjunction with machine learning techniques can produce sentiment classification results similar or close to a standard supervised approach.
- Hypothesis 5: A uniform notion of 'lexical unit' can be used across languages for sentiment analysis tasks.

1.2.3 Contributions of this Work

This thesis presents a number of novel and significant contributions to research in sentiment analysis:

1. An unsupervised knowledge-poor approach to domain-independent sentiment analysis
2. Use of the approach as a means of multilingual sentiment analysis
3. Sentiment zones (sequences of characters between punctuation marks) as units of classification
4. Sentiment score (a score based on the relative frequencies of units in documents of opposite sentiment) as a technique for sentiment classification
5. A score-difference technique for filtering out noise in sentiment classification. The technique is based on calculating the difference between opposite sentiment scores of an item.
6. A zone-difference technique for ranking sentiment classification. Zone-difference is a difference of zones of opposite sentiment in a document.
7. An unsupervised opinion holder and opinion target extraction technique
8. A scale-based sentiment classification, as an alternative to a traditional binary classification

9. A working multilingual system for sentiment analysis

1.3 Outline of the Thesis

1.3.1 Approach and Methodology

This study is concerned with the applicability of an unsupervised approach to sentiment analysis rather than with a single specific technique. The approach is applied not only to binary sentiment classification, but also to three-way sentiment classification (including a neutral class), to subjectivity classification at the document and sentence levels, and to opinion holder and opinion target extraction.

The approach is motivated by concerns related to both basic and applied research. With regard to former, I want to investigate if an unsupervised approach can produce acceptable results and facilitate domain-independent and multilingual sentiment analysis without using many external resources. With regard to the latter, practical applications aimed at on-line tracking sentiments, should be easily adjustable to new domains or languages. They, of course, can be augmented by other techniques that may increase their performance, but they need to be based on an approach that is robust across domains and languages.

The methodology is somewhat unusual, since multilingual issues are investigated first through experiments on the Chinese language. Most research in natural language processing (NLP) is concentrated on the English language and then 'spreads' to other languages. This approach results in an almost mechanical application of 'English-born' techniques to other languages. This occurred, for instance, in Linguistics, in which analysis of the Chinese language was initially based on the European notion of 'word'. It also occurred in NLP with word segmentation being treated as a prerequisite for any kind of language processing task. In contrast, the methodology in the research reported here was to first develop a technique based on the Chinese language and then apply it to other languages, including English.

The choice of the other languages addressed in this thesis can be justified objectively as follows. The English language is a well-studied language with a lot of resources available. The Russian language is very interesting in the context of this research as it is very different from both English and Chinese. Surprisingly, both English and Chinese have much in common (when compared to Russian): predominantly fixed word order and very limited morphology make these two languages very similar in the context of unsuper-

vised processing. The Russian language, however, features free word order and complex morphology. The structural difference of the languages makes unsupervised multi-lingual processing a challenging problem.

1.3.2 Structure of the Thesis

Chapter 2 covers aspects of sentiment analysis relevant to this thesis. It starts with a review of related studies of 'affect' in NLP that sets a general background for the research. Then the review discusses the various aspects of sentiment analysis, including its main tasks, techniques employed, features used, as well as different levels of classification and the different domains in which sentiment analysis is used. Different approaches to resource development are also described. Special attention is paid to outstanding problems and challenges in sentiment analysis.

Chapter 3 covers Chinese NLP in the context of sentiment analysis. It explores different kinds of features for Chinese sentiment classification and proposes an algorithm for sentiment classification based on a novel sentiment score calculation. This chapter investigates the influence of negation and lexical unit length on classification accuracy, and experiments with different units of classification (unigrams, sentences and 'zones').

Chapter 4 introduces an iterative approach to sentiment classification. This approach facilitates almost unsupervised sentiment classification using only a small set of lexical 'seeds' (which themselves could also be found automatically). This chapter also proposes and tests a number of techniques aimed at improving precision of iterative sentiment classification.

Chapter 5 applies the techniques to different languages: Russian and English. This chapter also tests the cross-domain applicability of the approach: the technique is applied to book and film reviews rather than to reviews of consumer electronics as in the previous chapters.

Chapter 6 tests the unsupervised approach on different tasks. The unsupervised classifier is used for three-way sentiment classification that using three classes: positive, negative and neutral. This is extended to a novel, continuous scale-based approach. The unsupervised approach is also applied to subjectivity classification at the document and sentence levels. The chapter ends with a set of experiments on opinion holder and opinion target extraction. This chapter tests the techniques in English, Chinese and Japanese.

Chapter 7 concludes and describes future work.

Chapter 2

Literature Review

This chapter presents an overview of approaches to sentiment analysis and the various research paradigms used. Section 2.1 describes research in 'affect' which sets background for sentiment analysis as part of NLP. The following section (2.2) describes different aspects of sentiment analysis, covering its main tasks, as well as different types of features and techniques used in this research field; the section also surveys domains where sentiment analysis is used. Approaches to resource development are discussed in Section 2.3. Section 2.4 discusses the most significant outstanding challenges in sentiment analysis.

2.1 Study of Affect

This section discusses the theoretical background of sentiment analysis, touching on relevant work in linguistics, psychology and ethnography as these areas provide important foundations for cross-lingual sentiment analysis.

2.1.1 Private States

The linguistic concept of non-factual information expressed in a text is relatively young. Quirk et al. (1985) introduced the linguistic term *private state* that denotes mental or emotional states, hidden from objective observation. Banfield (1982) proposed a term for the linguistic expression of private states: *subjectivity*. Thus *subjectivity analysis* is aimed at identification of attributes of private states: the subject who expresses a private state, the object about whom the state is expressed, the type of the attitude, the intensity of private state etc. In this sense, subjectivity analysis and sentiment analysis are often used interchangeably. Pang and Lee (2008) give a different, more narrow, NLP-specific, definition of subjectivity analysis as classifying a given text (a text or a sentence) into one

of two classes: objective (not expressing any private state) or subjective (expressing one or more private states).

2.1.2 Categorical and Dimensional Paradigms

Most research in sentiment analysis is based on one of two basic approaches: categorical and dimensional. The first approach puts all emotions into a finite number of categories (e.g. anger, fear, sadness, surprise), while the other one delineates emotions according to multiple dimensions rather than discrete categories.

The categorical approach is represented by the Cognitive Structure of Emotions (Ortony et al., 1988) which provides a taxonomy of emotions based on the different conditions that cause them. But since this approach is based on psychological contexts (for example, relations between people) which usually are not represented in the text, it is quite difficult to base any NLP study on it.

Another theory within the categorical paradigm that is derived from psychology is Appraisal Theory. It claims that all emotions are the result of evaluations (appraisals) of events that cause specific reactions in different people (Scherer and Schorr, 2001). Appraisal Theory is applied to language by Systemic Functional Linguistics as a theory of evaluation in text. Appraisal Theory analyses the way opinion is expressed in text and provides taxonomies for systematic identification of expressions of opinions and emotions in context. The taxonomies not only include words related to certain emotions or opinions but also cover the way authors interact with other authors and their audience.

According to Appraisal Theory, appraisal consists of three subsystems that function interactively: attitude, engagement and graduation. Attitude addresses one's feelings (emotional reactions, judgements of people and appreciations of objects); Engagement is concerned with the positioning of oneself with respect to the opinions of others and with the respect to one's own opinions; Graduation considers the ways a language increases or decreases the attitude and engagement in a text. Since this theory describes linguistic means of expression of emotions (lists of words that convey appraisal, for example) it can immediately be applied to NLP studies (for example, Read and Carroll, 2009).

Another way of representing affect is to put it into a multi-dimensional semantic space. For example, a two-factor structure of affect (described by Watson and Tellegen, 1985) puts emotion in two dimensions: Pleasantness (from *happy* to *sad*) and Engagement (from *surprised* to *quiet*).

Osgood et al. (1971) delineates emotions according to multiple dimensions: the two

primary dimensions in this account are along a 'good–bad' axis (the dimension of valence or evaluation) and a 'strong–weak' axis (the dimension of activation or intensity).

The dimensional understanding of affect is very productive for NLP as a basis for sentiment classification studies that also use (a very simplified) scale of sentiments ranging from two-point (positive – negative) to multi-point classifications (the 'five-star' system of Pang and Lee, 2005).

2.1.3 Affect Across Cultures

Since the research presented in this thesis addresses sentiment analysis in a multilingual context, the cross-cultural aspects of affect are also very relevant. Important questions include: Is sentiment universal? Is it expressed in comparable ways and can a unified approach be adopted? Is such an approach potentially applicable to other languages not tested in this research?

Ekman and Friesen (1971) found that particular facial behaviours are universally associated with particular emotions regardless of ethnic or cultural background. The existence of cross-cultural constants in emotional behaviour suggests that similar constants may be found in language. This was studied by Osgood et al. (1975) in 20 different countries with the help of about 80 anthropologists, psychologists and linguists. The study was done in the paradigm of semantic space measurement (Osgood et al., 1971; Osgood, 1976). The authors' general objective was to demonstrate that three affective dimensions of meaning – Evaluation, Potency, and Activity (E-P-A) – are in fact, pancultural. They found in particular found that the two most common modes of affect qualification across the world are GOOD and BIG (or some close synonym). They ranked the qualifiers found in each ethno-linguistic community in terms of both frequency and diversity of usage (i.e. productivity) and then correlated rankings in terms of translation equivalents, and found sizable and significant relationships. Osgood et al. (1975) concluded that "Human beings, no matter where they live or what language they speak, apparently abstract about the same properties of things for making comparisons, and they order these different modes of qualifying in roughly the same way in importance".

These findings suggest that a unified approach to sentiment analysis across multiple languages is in principle well-founded, providing a solid basis for the work presented in this thesis.

2.2 Sentiment Analysis

Sentiment analysis has been a popular research topic in recent years and has evolved into a big and diverse research field. A number of approaches have been used to create new research prototype and applied sentiment analysis systems. This section surveys the various tasks in sentiment analysis and methods utilised to perform them.

2.2.1 Tasks

There are four main tasks that are tackled in present day sentiment analysis research: subjectivity analysis, sentiment classification, opinion summarisation, and opinion extraction and mining.

Subjectivity Analysis

Subjectivity analysis, as indicated in Chapter 1, aims to distinguish subjective text (documents, sentences) from factual text. Subjective texts are those that express private states, which differ them from objective (factual) text that expresses only objective information, or facts.

Subjectivity analysis is a difficult task. The difficulty is mostly caused by the nature of private states that subjectivity analysis deals with. The subjective or objective nature of text is hardly ever stated explicitly (Wiebe, 1994) which complicates automatic processing of information that contains private states. Another challenging aspect of subjectivity analysis is that documents are almost never entirely either objective or subjective. Even a single sentence may contain factual information and some subjective evaluation of it. However a number of studies demonstrate reasonable success in subjectivity analysis.

A widely used technique in NLP, supervised machine learning, is often applied to subjectivity classification. Yu and Hatzivassiloglou (2003) describe document-level classification of news items using a Naïve Bayes classifier. Their research also investigated three approaches to identifying subjective sentences. The first was based on a hypothesis that, within a given topic, opinion sentences will be more similar to other opinion sentences than to factual sentences. The second used a Naïve Bayes classifier trained on documents that were supposed to be subjective (e.g. editorials). The features included words, bigrams, and trigrams, as well as the parts of speech in each sentence. Thirdly, the authors applied an algorithm using multiple classifiers, each relying on a different subset of the features. The study found that the Naïve Bayes classifier proved to be the most effective tool for sentiment classification, multiple classifiers slightly increasing performance. Wilson et al.

(2004) describe experiments on supervised subjectivity classification of the strength of opinions and other types of subjectivity, and classifying the subjectivity of deeply nested clauses. They used different features, including new syntactic features developed for opinion recognition, and support vector regression.

Another technique used in subjectivity classification is knowledge-based processing. This technique relies on resources (lexicons, rules etc.) that help distinguish subjective text. For subjectivity analysis Durbin et al. (2003) used a lexicon of individually rated (in relation to affect) words applied to part-of-speech tagged documents, taking into account modifiers (such as *very* or *slightly*) and negations. They also used syntactic rules to determine whether negation applies to the rated words. All these data were used to calculate an overall affect rating of a document.

Bootstrapping is a technique that allows the 'growing' of data from a limited amount of initial information. Wiebe (2000) used a set of manually annotated seeds for growing a list of strong indicators of subjectivity using the results of clustering words according to distributional similarity. Riloff and Wiebe (2003) used bootstrapping to learn linguistically rich extraction patterns for subjective expressions. First, they used high-precision classifiers to extract a learning set for extracting patterns that were subsequently used for finding further subjective sentences. Wiebe and Riloff (2005) further extended this approach. They started with seed-based extraction of a training corpus which was used to train an extraction pattern learner and a probabilistic classifier. Then the system was extended with a self-training mechanism that improved the coverage of the classifier.

Baroni and Stefano (2004) used a 'web-as-corpus' approach to calculate a subjectivity score for a list of adjectives. They used a list of seeds to calculate the mutual information between each seed and adjective, using frequency and co-occurrence frequency counts on the World Wide Web, collected through queries to the AltaVista search engine.

Some studies have used contextual information to improve subjectivity classification. Wiebe et al. (2004) generated and tested indicators of subjectivity, such as low-frequency words, collocations, and adjectives and verbs, using distributional similarity. The study found that the density of subjectivity indicators in the nearest context helps predict the subjectivity of a word. Pang and Lee (2004) discuss a method for finding subjective portions of a document with techniques for finding minimum cuts in graphs, assuming that sentences occurring near each other may share the same subjectivity status, everything else being equal.

Apart from being an important task in its own right, subjectivity analysis may facil-

itate other tasks, as observed by Wiebe (1994). Subjectivity classification, for instance, can help in information extraction by filtering out subjective clauses and leaving objective ones that should contain more reliable, factual information (Riloff et al., 2005). Separating subjective clauses from opinionated information improves the performance of opinion question answering (Yu and Hatzivassiloglou, 2003). Neutral (objective) information also affects the performance of sentiment classification and finding contextual polarity (sentiment orientation) of a word in text: the best way to improve performance is to improve the system's ability to identify when an instance is neutral (Wilson et al., 2009). Eriksson (2006) proposes objective sentence removal to improve established methods of sentiment analysis of film reviews. Word sense disambiguation may also improve performance if subjectivity annotation is used for learning senses (Wiebe and Mihalcea, 2006). A similar approach is used by Pang and Lee (2004) for sentiment classification. Finding emotional (subjective) information in stories helps increase the quality of text-to-speech (Alm et al., 2005).

However, combining subjectivity analysis with other tasks, even one so close as sentiment classification, may negatively affect performance. Esuli and Sebastiani (2006a) observe that determining subjectivity and orientation is a much harder problem than determining orientation alone. They extended their previous seed-based method (Esuli and Sebastiani, 2005) for word polarity detection to detect a word's subjectivity as well. The system was applied to a three-way classification task: Positive, Negative and Objective. The authors tested three different approaches. Two of them were based on a two-stage classification method and the third one classified words directly into the three categories. The latter system performed significantly worse. This finding shows that subjectivity analysis and sentiment analysis are different tasks and running them in one classifier degrades performance.

Sentiment Classification

The task of sentiment classification is to label text according to its sentiment. There is a diversity of methods and approaches used for sentiment classification and the most significant of these are outlined below.

Sentiment classification is usually regarded as a variant of traditional binary classification with the two classes: positive and negative (e.g., Pang et al. (2002) and many others). But there are exceptions: Pang and Lee (2005) try to determine an author's evaluation with respect to a multi-point scale (e.g., one to five "stars"). A similar approach based

on a three-way classification (positive, negative and neutral) was proposed by Koppel and Schler (2006) who stressed the importance of the neutral class for sentiment classification.

Sentiment and Subjectivity Pang and Lee (2004) propose a supervised machine-learning method of determining polarity that applies text-categorization techniques to subjective portions of a document only. These portions are extracted using minimum cuts in graphs. The idea of minimum cuts is inspired by the observation that text spans occurring near each other (within discourse boundaries) may share the same subjectivity status, other things being equal (Wiebe, 1994). Pang and Lee found that subjectivity detection can compress reviews into much shorter extracts that still retain polarity information at a level comparable to that of the full review. These extracts can be used for polarity classification which improves accuracy (from 82% to 86% for full reviews), suggesting that they are not only shorter, but also "cleaner" representations of document polarity.

The role of neutral (objective) text in sentiment classification was studied by Koppel and Schler (2006). The authors showed that in learning polarity, neutral examples cannot be ignored. Using only negative and positive training examples does not permit accurate classification of neutral examples. Moreover, better distinction between positive and negative examples can be achieved using neutral training examples. Properly combining pairwise learned classifiers leads to extremely significant improvement in overall classification accuracy. But the combination of the classifiers depends on the nature of the corpus, more specifically on the nature of the neutral documents in the corpus – whether they are truly neutral or in fact balanced (containing both sentiments).

Supervised Sentiment Classification Sentiment can be expressed in numerous ways and some studies have investigated what parts of the language are the most important for detecting sentiments. For example, Alm et al. (2005) used 14 kinds of features for supervised machine learning experiments into recognizing emotional passages and on determining their valence (i.e. positive versus negative) with a corpus of children's stories. The authors used a very large set comprising 14 different kinds of features: word lists, syntactic, story-related, orthographic, conjunctions, content BOW ("bag-of-words"), some of which were found automatically, some manually.

Another type of feature was used by Whitelaw et al. (2005b). They used adjectival appraisal groups as features for supervised sentiment classification of film reviews. The appraisal groups, coherent groups of words that express together a particular attitude, are part of a full appraisal expression as defined in Appraisal Theory (Martin and White, 2005).

The list of appraisal groups was produced semi-automatically, and manually modified to filter out noise. In total, 1329 terms were produced from 400 seed terms.

Other studies have experimented not only with different features but also with various machine learning classifiers (most notably Support Vector Machines, Naïve Bayes, and Maximum Entropy) and their combinations. Das and Chen (2007) tried a classifier voting technique for extracting small investor sentiment (buy, sell or hold) from stock message boards. Their approach was based on voting amongst five classifiers: naïve classifier (simply counting words with positive or negative meaning), vector distance classifier (a standard vector-based approach), discriminant-based classifier (counting discriminant scores of each word), adjective-adverb phrase classifier (counting only noun phrases with adjectives or adverbs) and a Naïve Bayes classifier. The features were a hand-picked collection of finance domain words. In particular, they observed that the Naïve Bayes classifier performed quite well, producing fewer false positives.

Sentiment Classification and Linguistics A more linguistic-driven approach was investigated by Eriksson (2006), who explored a linguistic method that facilitates sentiment analysis by using more information from a text than traditional methods based on machine learning. Eriksson's Linguistic Tree Transformation Algorithm is designed to exploit the syntactic dependencies between words in a sentence and to disambiguate word senses. Another technique introduced by Eriksson is an objective sentence removal algorithm. The approach specially addresses two major problems in the area of sentiment analysis, the non-local dependencies problem and the word-sense disambiguation problem. The Linguistic Tree Transform Algorithm uses parsing to find all bigrams (mostly adjective – noun phrases) relevant to the sentiment analysis task, while filtering out all irrelevant ones. Then an Objective Sentence Removal Algorithm filters out all sentences that do not contain topic words of interest (such as for film reviews, the names of the films, directors and screenwriters or some topic-related nouns). The algorithm is based on the assumption that some prior knowledge in this domain is readily available for automatic processing. These two algorithms produce a pruned version of the initial corpus containing only opinionated sentences relevant to the topic (for example, plot descriptions are removed). 100% accuracy is reported for the experiments with a frequency SVM model run on the data produced by the two algorithms.

Linguistically-motivated features help improve existing state-of-the-art sentiment classification results in a task of detecting implicit sentiment, a novel vision of sentiment classification proposed by Greene and Resnik (2009). Obviously implicit sentiment can-

not be detected by traditional indicators, such as words. This enabled the authors to investigate the syntactic "packaging" of ideas, studied previously by Greene (2007).

Opinion Summarisation

Opinion Summarisation aims to aggregate opinions on a given topic from multiple documents (probably from different sources) rather than classifying individual documents. Most approaches start with finding documents relevant to the topic and then classifying retrieved documents according to their sentiment. The topic might be found automatically from a set of documents (Hu and Liu, 2004; Chen et al., 2005; Feiguina and Lapalme, 2007) or given as a query (Eguchi and Lavrenko, 2006). The latter approach is close to opinionated information retrieval as it ranks documents or sentences according to both topic and sentiment relevance.

Some approaches use a variety of tools for opinion summarisation. In the domain of film review summarisation, Zhuang et al. (2006) describe a multi-knowledge based approach that uses WordNet, movie casts and labelled training data (1100 reviews), as well as grammatical rules linking feature words and opinion words.

Ku et al. (2006b) present a comprehensive system that summarises web blogs on a given topic (e.g. animal cloning). The summarisation is then presented by representative sentences augmented by an opinionated curve showing supportive and non-supportive degree along the time-line. The authors use a multi-level (word - sentence - document) sentiment classification system for detecting opinion direction.

Opinion summarisation can be combined with other techniques to produce an all-round practical application. Liu et al. (2005) describes a system called Opinion Observer which is capable of semi-automatic sentiment extraction, sentiment summarizing and visualisation. The system is able to compare sentiments about different products. The system is based on supervised rule discovery from a hand-labelled training corpus.

Opinion Extraction and Mining

Opinion extraction and opinion mining (the two terms are commonly used interchangeably) are concerned with extraction of certain aspects of opinion. One such aspect is the opinion holder (a person or a group that expresses an opinion) and another is the opinion target (something which is being discussed or evaluated). Feature-based opinion mining finds to find opinions about particular features of a product or service (as opposed to an overall opinion about something).

Opinion Holder Extraction There are two main types of approach to opinion holder extraction: one based on machine learning and the other using knowledge-based techniques. An example of the first type is presented by Kim and Hovy (2006) who used a machine learning technique for opinion holder extraction. As features for their Maximum Entropy classifier they used selected structural features from a deep parse, based on a frame representation of opinionated expressions. The frame was built around an opinion word, with semantic relations between it and opinion holder and target derived from semantic role labelling within the frames. Choi et al. (2005) consider opinion holder extraction to be an information extraction task and use a combination of two techniques: named entity recognition (by training Conditional Random Fields) and information extraction (AutoSlog, a supervised extraction pattern learner). The former models source identification as a sequence-tagging task; the latter learns extraction patterns.

Knowledge-based approaches utilise hand-build lexicons, parsing, heuristics and ontologies. For example, Bloom et al. (2007) describe an opinion holder extraction approach based on a hand-built lexicon, a combination of heuristic shallow parsing and dependency parsing, and expectation-maximization word sense disambiguation; they match phrases in the text with domain-dependent holder type taxonomies.

Kim et al. (2008) exploited a set of communication and appraisal verbs, SentiWordNet, a named entity recognizer, and a syntactic parser for opinion holder extraction. In each sentence they looked for the most opinionated word and then ascended the tree to its first ancestor node with verbal part of speech, and looked for its subject (a noun phrase) which was assumed to contain opinion holder candidates. If a subject was not found, then 'author' was set as the opinion holder of the sentence. If a subject was found, then from the NP chunk, any named entities or opinion holder candidates were extracted as the opinion holder. If no named entity or opinion holder candidate was found, then the holder again defaulted to the 'author' of the document. Regardless of the previous step, if a sentence included quotation marks, then the speaker of the quote was extracted as the opinion holder.

Kim and Hovy (2004) present a system that combines sentiment summarisation and opinion mining: it finds people who expressed opinion on a given topic as well as the orientation of the opinion. The system operates in four steps. First it selects sentences that contain both the topic phrase and holder candidates, found by means of BBN's named entity tagger. Next, it delimits the holder-phrase region. Then the sentence sentiment classifier calculates the polarity of all sentiment-bearing words individually. Finally, the

system combines word-based sentiments to find the holder's sentiment for the whole sentence. Ku et al. (2007b) use opinion operators as clues to find the locations of opinion holders. Opinion operators are words that are often associated with expressing opinions: *say*, *think*, *believe* etc.

The two main approaches can be used within a single system. For example, Seki (2008) used a combination of different techniques, including machine learning, parsing, rules and some in-test clues for detecting opinion holder and orientation.

Opinion Target Extraction Stoyanov and Cardie (2008) define an opinion topic (target) as "the real-world object, event or abstract entity that is the subject of the opinion as intended by the opinion holder". Thus opinion target detection facilitates detecting positive or negative sentiments for specific entities referred to in a document, instead of classifying the whole document into positive or negative. For example, the sentence *Product A is good but expensive.* contains two sentiment related statements: *Product A is good* and *Product A is expensive*, each describing different features (quality and price) of an object (the *Product*). In general, for this task, researchers use techniques similar to those used for opinion holder extraction.

For opinion target extraction, Kim and Hovy (2006) used the same approach as they used for the opinion holder extraction: semantic role labelling. Bloom et al. (2007) also used a similar technique for both tasks: their manually created taxonomies also included opinion targets. Reasoning that opinion targets share similar features with opinion holders (each being a noun phrase, but acting as object rather than subject), Kim et al. (2008) used a technique similar to that of Kim and Hovy (2006) for opinion holder extraction, adopting a a statistical machine learning technique based on syntactic features (syntactic path and dependency) and other heuristic features, such as topic words and named entities. Nasukawa and Yi (2003) utilised a sentiment analysis dictionary consisting of more than 3,000 items and a set of rules, as well as shallow parsing.

Product Feature Extraction A more fine-grained version of opinion target extraction extracts evaluations of product features. Unlike opinion targets, a product may have many different features that could be evaluated, all of them may have different importance for reviewers. Gamon et al. (2005) used a clustering algorithm to find a product feature taxonomy. The algorithm used a stop-word list, which should not be used for building clusters, and 'go-words' known to be salient in the domain. Sentences were then clustered according to the product feature taxonomy, and processed by a sentiment classifier trained

on a corpus bootstrapped from a small manually-created corpus. Popescu and Etzioni (2005) present a system and claim to be the first to report precision and recall on the tasks of opinion phrase extraction and opinion phrase polarity determination in the context of known product features and sentences. This system intensively uses the knowledge mining tool, KnowItAll, a Web information-extraction system (Etzioni et al., 2005), to extract product features and opinions regarding them.

Zhang and Varadarajan (2006) identify a new task in opinion extraction: predicting the utility (or, reliability, usefulness, informativeness) of product reviews. Utility is defined as a multi-aspect feature of customer reviews that combines subjectivity with deep technical analysis of a product's features. The authors build regression models by incorporating a diverse set of features including lexical similarity, part of speech tags and lexical subjectivity clues.

Titov and McDonald (2008) present a novel framework for extracting the features of objects from online user reviews. They build statistical models to induce multi-grain topics. The models not only extract features, but also cluster them into coherent topics, e.g., *waitress* and *bartender* are part of the same topic, *staff*, for restaurants. This differentiates it from much of the previous work which extracts aspects through term frequency analysis with minimal clustering.

Question Answering

Question answering (QA) is well-established research topic in NLP. A new facet of it is presented by opinion QA. Yu and Hatzivassiloglou (2003) study separating opinions from fact, at both the document and sentence level, in the context of QA. Ku et al. (2007a) define six opinion question types and use an information retrieval system to detect question focus. The retrieved information is then processed to match the sentiment of the query.

2.2.2 Techniques

Research in sentiment analysis uses a number of techniques, such as supervised machine learning, rule- and knowledge-based and some others described beneath.

Supervised Machine Learning

Supervised machine learning is the most frequently used technique in sentiment classification. To date, the majority of studies have used support vector machines (SVM) and Naïve Bayes (NB). A study of the effectiveness of machine learning techniques was carried out

by Pang et al. (2002), who explored three different supervised machine learning techniques (NB, maximum entropy and SVM). All these were applied to a movie-review corpus in the task of text-based sentiment classification. A baseline was produced by means of a list of manually (with some help from statistics) selected words (mostly adjectives). The authors also tested different feature sets for each classifier: unigrams, bigrams, unigrams with POS-tags and adjectives. Document feature vectors either encoded the frequency of a feature, or just its presence (a binary value). The best result was obtained by a support vector machine (SVM) with unigrams as features and with presence encoded in the feature vector (although Naïve Bayes was not far behind). Pang et al. also noted a slight increase in performance by using a simple negation check. POS-tags increased accuracy in Naïve Bayes and maximum entropy, but decreased in it SVM. Many authors have also demonstrated a higher accuracy for SVM compared to other machine learning techniques. For instance, Gamon (2004) showed that large feature vectors in combination with feature reduction help train linear support vector machines which achieve high classification accuracy on data that present classification challenges even for a human annotator. However Boiy et al. (2007) suggest that it can still be advantageous to use the Naïve Bayes multinomial technique, as it is considerably faster in practice.

Some researchers combine different machine learning techniques using classifier voting (Das and Chen, 2007) or combine machine learning with other techniques. For example, Watanabe et al. (2004) used an existing transfer-based machine translation engine (Watanabe, 1992) to translate from Japanese documents to a set of sentiment units (there are 3,752 Principal patterns, the size of the lexicon is not reported). To do so they replaced the translation patterns and bilingual lexicons with sentiment patterns and a sentiment polarity lexicon.

Comparing machine learning to symbolic techniques for sentiment analysis, Boiy et al. (2007) conclude that machine learning approaches are more promising.

Weakly Supervised and Unsupervised Techniques

A disadvantage of the supervised techniques is that they need a sufficient amount of human annotated training data to obtain acceptable results. Developing such data is a difficult and costly process and this has motivated researchers to look for methods that do not require training data or need only a relatively small amount of it.

Bootstrapping One of the most widely used weakly supervised methods is bootstrapping. Abney (2002) defines it as "a problem setting in which one is given a small set of

labelled data and a large set of unlabelled data, and the task is to induce a classifier".

Hatzivassiloglou and McKeown (1997) describe a semi-supervised method based on the idea that similarly oriented adjectives might be conjoined: "The conjoined adjectives and conjunctions usually have similar orientation, though 'but' is used with opposite orientation". This approach was also used by Brody and Elhadad (2010) who used automatically extracted seeds to build a conjunction graph. To find seeds they used morphological information (such as the prefixes 'un', 'in', 'dis', 'non') and explicit negation to find pairs of opposite polarity.

Another approach to building a sentiment lexicon is based on point-wise mutual information between lexical items, assuming that items that tend to be used together might share the same sentiment orientation. Turney (2002) proposes a technique for finding the semantic orientation (recommended or not recommended) of a phrase (containing adjectives and adverbs) from unlabelled text by comparing its association with a positive word (*excellent*) and a negative word (*poor*). The author uses point-wise mutual information to calculate each association using the World Wide Web as a corpus.

Turney's approach inspired Baroni and Stefano (2004) to design a similar technique for ranking a large list of adjectives according to a subjectivity score without resorting to any knowledge-intensive external resources (such as lexical databases, parsers or manual annotation). Baroni and Stefano describe a simple way of finding subjective adjectives by means of the Web used as a corpus and a small list of seed words (35 adjectives).

Gamon and Aue (2005) describe a bootstrapping technique similar to Turney's, which they use for finding the sentiment vocabulary in a domain. This method rests on three special properties of the sentiment task: (1) the presence of certain words can serve as a proxy for the class label; (2) sentiment terms of similar orientation tend to co-occur and, (3) sentiment terms of opposite orientation tend not to co-occur at the sentence level. They used the latter property to mine the sentiment vocabulary which was to be submitted to the Turney-style technique to find their semantic orientation. Another substantial difference is that the authors do not use a huge corpus (like the Web) for bootstrapping but rely solely on in-domain data. This work is also notable for applying 'a second layer' of classification by using machine learning techniques to the found data.

A weakly supervised sentence-level sentiment classifier is described by Gamon et al. (2005). The system classifies sentence sentiment using a small training corpus (2,500 sentences, enlarged by means of bootstrapping) and produces three classes: positive, negative and other. Banea et al. (2008a) use a list of sixty seeds to create a subjectivity lexicon for

languages with scarce resources using on-line dictionaries.

Riloff and Wiebe (2003) describe a a semi-supervised technique that learns extraction patterns from a training corpus produced by high-precision classifiers and then applies the newly found patterns to find more subjective sentences. The classifiers use a manually created set of features (words and n-grams) to produce two sets of sentences: objective and subjective. The two sets are then used by a pattern learner to find patterns that are mostly used in subjective sentences. The process of learning is based on application of a large set of syntactic templates to the corpus and extracting all possible patterns that match the templates. The frequencies of the patterns obtained for each of the classes of the sentences (objective and subjective) are compared and the most subjectivity-associated patterns are used to enlarge the feature set of the classifiers. In a later study, Wiebe and Riloff (2005) extend the system by applying machine learning techniques to the extracted sentences to increase recall.

Reference Data A different approach to unsupervised sentiment classification is described by Ghose et al. (2007). The authors use an economic context to find out what makes a review positive or negative. The approach is based on the observation that on-line merchants on eBay with positive feedback can sell products for higher prices than competitors with negative evaluations. This makes it possible to use techniques from econometrics to identify the 'economic value of text' and assign a 'dollar value' to each text snippet, measuring sentiment strength and polarity effectively and without the need for any annotated resources.

An alternative approach was explored by Read (2009). To find a document's sentiment orientation Read compared the document with some prototypes (positive and negative texts) using their constituents (words and phrases).

Linguistic Resources Subasic and Huettner (2001) present an approach based on a fusion of natural-language processing and fuzzy logic techniques for analysing affect content in free text. The linguistic resource for the approach is a hand-crafted fuzzy affect lexicon, from which other resources are generated: a fuzzy thesaurus and affect category groups. A text is tagged with affect categories from the lexicon, and the affect categories' centralities and intensities are combined using techniques from fuzzy logic to produce affect sets – fuzzy sets that represent the affect quality of a document.

Zhuang et al. (2006) use WordNet, statistical analysis and movie knowledge for movie review mining and summarisation.

Smrz (2006) uses linguistic resources, especially WordNet extensions, to collect and identify different opinions on a given topic and to report a diversity of opinions across languages and countries from various information sources available on the Web such as newspapers, Internet blogs and forums.

Sentiment Scores The use of sentiment lexicons often relies on score-based techniques in which classification is based on the total sum of positive or negative sentiment features present in a text. This technique is used in many of the studies mentioned above. Manually created phrase pattern matching (e.g. Nasukawa and Yi, 2003; Fei et al., 2004) requires checking text for manually created polarized phrase tags (positive and negative). Similarly, but with automatically found phrases, Turney (2002) and Hatzivassiloglou and McKeown (1997) classify documents with more positive items as positive and vice versa. Subasic and Huettner (2001) use a more fine-grained approach for affect analysis: documents are scored according to the degree of intensity of an emotion class. Ku et al. (2005) describe a technique based on finding opinion words from a semi-automatically created list and concept words also taken from a predefined list. The underlying idea of the approach is that the opinion of the whole is a function of the sentiments of the parts, so all individual scores are summed to produce an overall sentiment score for a document.

Negation Several studies have experimented with negation detection as part of sentiment classification. Ku et al. (2007b) found that negation is important for opinion polarity classification in Chinese. Boiy and Moens (2008) studied the influence of negation tagging in English, French and Dutch opinion mining and conclude that negation detection although helpful, depends on the specificities of the language. Wilson et al. (2009) explore features for phrase-level sentiment analysis and find that negation features give the best performance improvements. The most widely used techniques of negation detection involve n-grams (Dave et al., 2003) (e.g., "not worth") or reversing the sentiment of every word that follows a negation until the next punctuation token (Pang et al., 2002).

Link Analysis An approach that can also be applied to sentiment analysis in certain domains is based on analysis of links between documents. Efron (2004) used co-citation analysis for classification of website opinions on different topics, and Agrawal et al. (2003) used reply links between messages to classify USENET newsgroups as supporting or opposing some idea. Thomas et al. (2006) classify the transcripts of U.S. Congressional floor debates into speeches that represent support of, or opposition to, proposed legis-

lation. The authors exploit the fact that these speeches occur as part of a discussion and use sources of information regarding relationships between discourse segments, such as whether a given utterance indicates agreement with the opinion expressed by another. Obviously the approach is limited to domains that feature explicit links between messages.

In general, weakly supervised and unsupervised methods are less accurate than well-trained supervised machine learning classifiers, but have the potentially very important advantage of requiring little or no manually annotated training data. Another way of improving the performance of sentiment classification is combining different classifiers into one system. Prabowo and Thelwall (2009) describe a combined classifier consisting of a rule-based classifier, supervised machine-learning classifier and unsupervised learning. The classifiers may also contribute to each other to improve classification results. This approach makes the whole system less data-dependent, as each of the classifiers pre-process data for the others.

2.2.3 Features

A diverse set of features have been used in sentiment classification. This section describes them in more detail, starting with the most frequently used type: semantic features intended too capture the meaning of lexical items as relevant to sentiment classification; continuing with sequences of lexical items and relations between them constituting syntactic features; and, finally, lexical items that contribute to stylistic features.

Semantic

Any kind of annotation of the sentiment polarity of words or phrases (e.g. sentiment / polarity scores) is in fact a representation of a part of their meaning. Hatzivassiloglou and McKeown (1997) proposed a semantic orientation method for sentiment classification. This was extended by Turney (2002) who used a web-based mutual information method to find the semantic orientation of phrases. Often researchers use seeds to semi-automatically build a list of lexical items with marked polarity (Wiebe, 2000; Kobayashi et al., 2004; Kim and Hovy, 2006). Esuli and Sebastiani (2005, 2006a) use semi-supervised learning from human-labelled texts to tag words positive or negative. Ku et al. (2006b) use thesauri to extend the list of sentiment terms found in multi-lingual lexicographical resources. Others use external resources (WordNet, lexicons and dictionaries) to infer the sentiment polarity of lexical items. Smrz (2006) uses wordnets in different languages to create sentiment

lexicons in these languages. Kim and Hovy (2004) assume that WordNet synonyms share sentiment. Esuli and Sebastiani (2005) and Esuli and Sebastiani (2006a) also compare WordNet glosses of words, assuming that words with similar orientation have "similar" glosses.

There are also a number of lexical resources that can be used for sentiment classification: SentiWordNet by Esuli and Sebastiani (2006b), a WordNet-like resource developed for sentiment analysis; Ku et al. (2006b) developed a NTU Sentiment Dictionary for Chinese sentiment analysis; the General Inquirer (GI) lexicon (Stone et al., 1966) is often used for mining sentiment-bearing words (Esuli and Sebastiani, 2006a; Ku et al., 2005).

A more labour-intensive way of creating sentiment lexicons is based on manual tagging. Whitelaw et al. (2005a) manually tagged all phrases according to Appraisal Theory (Martin and White, 2005). Nasukawa and Yi (2003) manually built a sentiment lexicon incorporating information about each item's POS, canonical form and arguments (such as subject and object): for example, *gVB admire obj* indicates that the verb "admire" is a sentiment term that indicates favourability towards its noun phrase object. Abbasi et al. (2008) used manually constructed affect lexicons for analysis of hate and violence in extremist web forums. Subasic and Huettner (2001) developed a fuzzy affect lexicon which was used as a primary linguistic resource for fuzzy semantic typing.

Lexical resources are widely used in sentiment analysis; they might not, however, always be the most effective tool. Dave et al. (2003) found that using collocations as features, even after putting noun-adjective relationships into a canonical form, was ineffective. The authors observed that their corpus of reviews was highly sensitive to minor details of language: stemming performed below the baseline in some tests because, for example, negative reviews tend to occur more frequently in the past tense, since the reviewer might have returned the product. Airoldi et al. (2006) particularly found that the sentiment orientation of words is contextual and is "captured by conditional dependence relations among words, rather than by keywords or high-frequency words".

Syntactic

Syntactic features include word n-grams (Pang et al., 2002; Gamon, 2004), part of speech tags (Pang et al., 2002; Nasukawa and Yi, 2003; Choi et al., 2005; Gamon, 2004) and punctuation (Pang et al., 2002; McDonald et al., 2007; Choi et al., 2005; Abbasi et al., 2008). POS-tag n-grams were tested by Nasukawa and Yi (2003) and Fei et al. (2004). Fei et al. found, for example, that the combination *noun + adjective* is usually used to

convey negative sentiment, while *adjective + noun* is often used for expressing positive sentiment. Wiebe et al. (2004) used collocations to identify fixed n-grams, for example: worst-*adj* of-*prep* all-*det*. They also proposed a generalised version of collocations, where certain classes of words are represented by a POS-tagged variable. For example, U-*adj* as-*prep* represents a phrase that consists of a unique (occurring only once) adjective and the preposition 'as'. This generalised collocation matches phrases like 'drastic as', 'perverse as' and 'predatory as'.

Gamon (2004) analysed the effectiveness of linguistic features and found that part of speech trigrams and an NP consisting of a pronoun followed by a punctuation character were important for sentiment classification of customer reviews.

A broader context was used by Riloff et al. (2003). They created discourse features to capture the density of sentiment indicators in the text surrounding a sentence. Pang and Lee (2004) combined traditional bag-of-words features with inter-sentence level contextual information in a *minimum cut* formulation.

Stylistic

Some studies have used stylistic attributes for sentiment analysis tasks. Wiebe et al. (2004) used words that occurred only once (*hapax legomena*) to improve the accuracy of subjectivity classification. They observed a significantly higher presence of unique words in subjective texts compared to objective documents in a Wall Street Journal corpus and noted that "Apparently, people are creative when they are being opinionated". Gamon (2004) used the length of constituents (sentence, clauses, adverbial/adjectival phrases, and noun phrases) for sentiment classification of feedback surveys. Abbasi et al. (2008) used a wide array of English and Arabic stylistic attributes including lexical, structural, and function word style markers and reported high accuracy in blog sentiment analysis.

Feature Selection

Gamon (2004) describes a series of experiments for determining an optimal set of features for the supervised sentiment polarity classification task. He tested three kinds of features: linguistic features, surface features and word n-grams. The first kind was obtained by means of a tool that provided a phrase structure tree and a logical form for each string. The second kind consisted of word n-grams, function word frequencies and POS ngrams. Gamon observed that the presence of very abstract linguistic analysis features improves the performance of the classifiers and concluded that affect and style are linked in a more

significant way than was thought before. He also observed that some of the most effective features were absolutely unpredictable and very domain-dependant. Thus it is preferable to start without an artificially limited "hand-crafted" set of features: relevant patterns in the data that may not have been obvious to the human intuition can be identified by means of automatic data analysis.

2.2.4 Levels

Sentiment analysis can be carried out at a number of levels: words, phrases, sentences and (sets of) documents. Another classification can be done by separating out-of-context classification and context-based classification (*a priori* and *contextual* sentiment classification (Wilson, 2008)). These two classifications are not strictly orthogonal: for example, words can be classified into positive and negative for building a sentiment dictionary which is supposed to be used in different contexts (thus the dictionary should be context-free) as in Turney (2002). However words can also be classified bearing their context in mind (e.g. Hatzivassiloglou and McKeown, 1997), but in this case the resulting word list can be applied to the same or similar context and eventually is a part of document-level classification. Wilson et al. (2009) study how prior (context-free) polarities affect the performance of sentiment classifiers and find that certain words may change their polarity and become neutral, and this affects performance of a classifier. This sections overviews two major levels of sentiment classification: words and phrases (as a stand-alone, *a priori* classification) and sentences and document level (the level that utilise contextual information)

Words and Phrases: Context-free Sentiment Classification

The context-free sentiment classification is usually done on the lexical level and considers words and phrases. The aim of such a classification is the creation of linguistic resources that can be used without a relation to a certain context (domain, style or genre). Turney and Littman (2003) tested pointwise mutual information (PMI) and latent semantic analysis (LSA) techniques for sentiment classification of words and phrases. Baroni and Stefano (2004) used a technique similar to PMI for subjectivity classification of adjectives. An extended version of Turney's PMI method was proposed by Gamon and Aue (2005) who augmented the approach with an idea that sentiment terms of opposite orientation tend not to co-occur at the sentence level. Yuen et al. (2004) described a morpheme-based sentiment classification of Chinese words.

Esuli and Sebastiani (2005) analyse the glosses of on-line dictionaries to find the orientation of subjective terms. In another study Esuli and Sebastiani (2006a) test the technique for finding not only the orientation but also the subjectivity of words. Esuli and Sebastiani (2007) applies the PageRank technique to WordNet synsets to find sentiment orientation of words.

Sentences and Documents: Contextual Sentiment Classification

Contextual classification is possible at all levels of the language. But contextual sentiment classification of words and phrases is useless if it is not a part of sentence- or document-level classification. The latter two levels set the context for words and phrases. This suggests that sentence- and document-level classifications are indeed contextual for lexical units. Even generic sentiment lexicons when applied to document classification are often adjusted to the domain by using contextual features (surrounding words, POS, shallow parsing).

The study of contextual polarity was done by Wilson et al. (2005) who recognise contextual sentiment orientation in phrases. A weakly supervised sentence-level sentiment classifier is described in Gamon et al. (2005).

Often classification of documents is based on a chain of classification at all levels. Pang and Lee (2004) investigate sentiment classification of text at varying levels of granularity: an initial model classified each sentence as being subjective or objective and the top subjective sentences are then input into a standard document level polarity classifier. McDonald et al. (2007) do similar type of classification but using a joint structured model for all levels.

2.2.5 Text Types and Domains

Sentiment analysis studies are applied to a number of different types of text in a number of domains. The choice of domains is based on practical applicability of sentiment analysis. For example, customer reviews might be of interest to companies who would like to track customer opinions to improve their products or marketing. News stories and news provider forums for reader comments provide much information about public sentiment about current events. And social media (blogs, internet-forums, social networking websites and others) are a hot topic in many marketing and media studies as they are not only a valuable source of opinion-related information but also a medium where opinions are formed.

Customer Reviews

A particular (and a very specific) type of customer review is the film review which has become one of the most well-studied domains mostly due to availability of a movie review corpus created by Pang et al. (2002) and then enlarged and improved (Pang and Lee, 2004). But even before the corpus was created, Turney (2002) used film reviews for his studies, reporting this domain to be particularly difficult to process (as compared to reviews of automobiles, banks and travel destinations). Read and Carroll (2009) and Zhuang et al. (2006) studied this domain using the corpus. Boiy et al. (2007) used the movie review corpus and added blogs, discussion boards and other websites on a number of film titles and car brands.

A number of sets of product reviews have been used for sentiment classification experiments. Kobayashi et al. (2004) collected 15,000 reviews from several review sites on the Web about cars and 9.700 reviews of computer games. Car reviews were also processed by Gamon et al. (2005). Dave et al. (2003) mined reviews of different products from CNet and Amazon. McDonald et al. (2007) compiled a corpus of 600 on-line product reviews from three domains: car seats for children, fitness equipment, and MP3 players. Feiguina and Lapalme (2007) studied a corpus of electronic consumer goods (MP3 players, digital cameras, mobile phones, DVD players) partly based on the review corpus developed by Hu and Liu (2004). Zhou et al. (2008) mined Chinese customer reviews of different products. Brody and Elhadad (2010) used a corpus of over 50,000 restaurant reviews from Citysearch New York developed by Ganu et al. (2009). Reviews of Chinese public health system were studied by Zhang et al. (2008).

News

The domain of news features a more complex structure of sentiment expression than reviews. Apart from the objects(s) of discussion (opinion target) and the opinion itself, news items often report a subject who expresses the opinion (opinion holder), while in reviews, the opinion holder is usually the author. This paves the way to experiments in opinion mining, such as those presented in the series of Multi-Lingual Opinion Analysis Task Workshops (Seki et al., 2008).

Wilson (2008) investigated the manual and automatic identification of linguistic expressions of private states in a corpus of news documents from the world press.

An economic news domain was studied by Ku et al. (2006a) who detected event bursts from the tracking plots of opinions. Nasukawa and Yi (2003) processed general news

stories extracting sentiments for specific items. Read and Carroll (2009) studied domain and temporal dependency in news items. Ku et al. (2006b) carried out a number of opinion summarisation experiments on news and web blog articles related to the issue of animal cloning.

Social Media

This type can cover a lot of domains, for example Ku et al. (2006b) and Boiy et al. (2007) used blogs and other social media to study opinions in news and product reviews. However there still are some studies that do not belong to product review sentiment classification or news-based opinion mining. For example, Abbasi et al. (2008) applied sentiment analysis to web forum opinions in multiple languages studying propaganda dissemination. Agrawal et al. (2003) used three datasets from the archives of the Usenet postings: abortion, gun control and immigration. Mihalcea and Liu (2006) analysed dominating sentiments ("happiness") in blogs along the dimensions of time of day and day of the week.

2.3 Resource Development

Resources for sentiment analysis include datasets (corpora) and lists of lexical items. A comprehensive list of such resources is presented by Pang and Lee (2008). This section describes research efforts towards the development of these resources.

The approach described by Hatzivassiloglou and McKeown (1997) was the first attempt to automatically develop linguistic resources for opinion mining. The method finds two groups of adjectives by learning constraints from conjunctions on the positive or negative semantic orientation of the conjoined adjectives. Subsequently, many researchers have developed lists of words with different opinion orientation. One of the best known is SentiWordNet by Esuli and Sebastiani (2006b), a lexical resource in which each WordNet synset has three scores representing its positivity, negativity and neutrality. Ku et al. (2006b) developed the Chinese NTU Sentiment Dictionary using a number of external resources including Chinese thesauri, the General Inquirer lexicon and the Chinese Network Sentiment Dictionary.

Building a corpus suitable for a research can be a costly and time-consuming task. Several researchers have tried using user annotations on reviews ('stars', 'thumbs up and thumbs down' etc) for building corpora. Dave et al. (2003) noted that there is a number of specific problems that must be considered when collecting data for experiments in sentiment analysis. Rating inconsistency is often an issue when a researcher tries to build

a corpus from user-rated reviews. It has been observed that people often have their own quite different scales of appraisal which make their ratings very inconsistent especially in a multi-level rating system (e.g. one to five stars). People sometimes are not sure about their opinion or have rather 'mixed feelings' which may result in inconsistency between what they write and how score. This is one of the reasons why despite a huge amount of different reviews, editorials, customer feedbacks etc., there are not many tagged corpora for training and testing freely available. Most of research corpora have only text-level orientation tags, which makes it particularly difficult to carry out sentence-level experiments.

Another specific problem of sentiment analysis has been skewed distribution of sentiments. It has been observed by many researchers that positive texts quite often predominate in collections and this may affect experimental results since, for example, machine-learning techniques are often sensitive to data skew. A possible solution is manual tagging of research corpora. However, manual tagging requires human annotators who may also have different subjective scales of sentiment.

Wilson and Wiebe (2003) developed a detailed annotation scheme for expressions of opinion, belief, emotion, sentiment and speculation. The development of the annotation scheme had two goals: "to develop a representation for opinions and other private states that was built on work in linguistics and literary theory on subjectivity" and "to develop an annotation scheme that would be useful for corpus-based research on subjective language and for the development of applications such as multi-perspective question-answering systems". The scheme includes such features as subjectivity (affectiveness) represented by the tag *onlyfactive=yes/no*; *overall-strength* and *on-strength* describe the strength of a subjective clause and its particular constituents. The scheme also differentiates explicit and implicit sentiments and deals with nested constructions. Finally, the annotation ranks subjective clauses according to their type (*attitude-type*) and targets (*attitude-toward*). The study particularly found that removing sentences that are not clearly subjective ("borderline cases") helps increase inter-annotator agreement. The annotation scheme was further developed by Wiebe et al. (2005). Continuing the paradigm, Wilson (2008) developed an annotation scheme for fine-grained subjectivity analysis and created the Multiperspective Question Answering (MPQA) Opinion Corpus.

Read et al. (2007) developed an annotation scheme that closely follows the Appraisal Theory of Martin and White (2005). Read et al. applied a very detailed annotation scheme featuring more than 30 tags at different levels of abstraction to a corpus of book reviews. They observed a generally low level of inter-annotator agreement especially at

the most detailed level of annotation.

2.4 Challenges of Sentiment Analysis

The ways in which opinions are expressed vary between languages and also within a single language (so-called "domain-dependency"). For example, the word *horrible*, in a description of a plot of a horror film does not necessarily bear any sentiment-related meaning. However these word is a reliable indicator of negative sentiment in most other domains (e.g. *horrible performance*). Turney (2002) observes that "for example, the adjective "unpredictable", may have a negative orientation in an automotive review, in a phrase such as "unpredictable steering" but it could have a positive orientation in a movie review, in a phrase such as "unpredictable plot"". This problem is further complicated by ambiguity of word meaning in different contexts. This problem was studied by Wilson et al. (2005) who give an example of the word *trust*:

(1) Philip Clapp, president of the National Environment **Trust**...

The word *trust*, which has positive prior polarity, in this context has neutral meaning since it is part of named entity.

Domain-dependency decreases the performance of classifiers trained, or using data from a different domain (Engström, 2004). Read (2005) also noted a temporal dependency where even in the same domain people use different means of expressing sentiment over time. A major current challenge is how to automatically extract sentiment information from documents in different languages and in different domains. Most existing approaches are based on adapting systems designed for one language (or domain) to another. Obviously, there are differences between cultures, languages and even within a language (consider the difference between evaluations of company financial prospects in a business newspaper and reviews of a hard-rock festival in a participant's blog). Such differences make adaptation difficult.

2.4.1 Cross-Domain Approaches

Aue and Gamon (2005) try to overcome the problem of domain-dependency of sentiment analysis by means of using labelled data from other domains. They investigate and compare four approaches:

1. training on a mixture of labelled data from other domains where such data are

available;

2. training a classifier as above, but limiting the set of features to those observed in the target domain;

3. using ensembles of classifiers from domains where there is available labelled data;

4. combining small amounts of labelled data with large amounts of unlabelled data in the target domain. This approach does not use any out-of-domain data; instead, it uses a generative Naïve Bayes classifier using the Expectation Maximization algorithm.

The four approaches were tested on four different corpora: movie reviews, book reviews, product support services and knowledge base web survey data. It was found that the approaches that used some data from the target domain (approaches 3 and 4) performed better than ones that used only out-of-domain training data (1 and 2). The best accuracy was achieved by the last approach, which still requires (small) amounts of annotated in-domain data.

Blitzer et al. (2007) describe another way of overcoming domain-dependency by means of the adaptation of a classifier trained in one domain to another. The authors raise the problems of accuracy loss and domain similarity. The main idea underlying the approach is Structural Correspondence Learning (SCL) developed by the authors in previous papers. Since the authors use Mutual Information for finding new 'pivot features' in unlabelled domains, the full name of the approach is SCL-MI. The main intuition is that even when key opinion words are completely distinct for each domain, if they have high correlation with *excellent* and low correlation with *awful* in unlabelled data, then it is possible to align them. The approach consists of three steps:

1. Using a labelled corpus from one domain and unlabelled corpora from both a new domain and the old one, find pivot features which occur frequently in both domains.

2. SCL models the correlations between the pivot features and all other features by training linear pivot predictors to predict occurrences of each pivot in the unlabelled data from both domains (Ando and Zhang, 2005; Blitzer et al., 2006). This is based on the calculation of correlation (MI) of pivot features (such as *excellent*) and non-pivot features (like *fast*, *dual-core*).

3. For some domains the features found are not well-aligned (thus not good enough for sentiment classification). To correct misalignment the authors manually label 50 top

features of the target domain.

Domain-adaptation of a generic sentiment lexicon was tested by Li et al. (2009) who used labelled documents to adjust a hand-built sentiment lexicon to a domain.

Another way of improving the accuracy of domain adaptation is by selecting the most suitable source domain by means of A-distance (Ben-David et al., 2007). The key intuition behind the A-distance is that while two domains can differ in arbitrary ways, only a degree of difference of the relevant part affects the accuracy of classification.

An attempt to use extralinguistic data to overcome domain dependency is presented by Read (2005) who describes experiments with emoticons, as a way of learning sentiment-relevant linguistic expressions from large amounts of unlabelled text.

2.4.2 Cross-Language Approaches

Cross-language sentiment analysis has attracted attention in recent years. For example, there is a yearly evaluation workshop dedicated to multi-lingual opinion mining (NTCIR) at which research groups present their approaches to this problem (Seki et al., 2008).

One possible way of overcoming language dependency is the re-use of resources in one language for sentiment analysis in another. Mihalcea et al. (2007) describe a method for generating subjectivity analysis resources in a new language by using tools and resources available in English. As a medium the approach uses freely available cross-lingual resources, such as bilingual dictionaries or a parallel corpus. The authors used a subjectivity lexicon by Wiebe and Riloff (2005) as a source of subjective information and two English-Romanian dictionaries to translate the lexicon, dealing with such problems as inflections, multiple senses and multi-word expressions. The resulting Romanian lexicon was then tested on a corpus in this language. This method was further developed by Banea et al. (2008b), who suggested the use of machine translation for the generation of resources for subjectivity analysis in other languages (Spanish and Romanian in their study). The research explores two possible scenarios: 1) translating (bi-directional) an existing resource and 2) combining automatic subjectivity analysis with a machine translation system. Banea et al. (2008a) propose a bootstrapping method based on seed words and an on-line dictionary. The candidate words produced by this method are then ranked by LSA and top lexical items from the resulting list are regarded as a reliable subjectivity lexicon in a new language.

Abbasi et al. (2008) also used a translation-based approach for generating resources for Arabic sentiment analysis. A similar approach was used by Smrz (2006) who used

national versions of the WordNet lexicon to identify subjective expressions.

Boiy and Moens (2008) performed a number of machine learning experiments in sentiment analysis in Dutch, English and French. Although the experiments treated these languages separately (no specific multi-lingual adaptation techniques were used), they note language-specific particularities that affect sentiment analysis. The importance of such language-specific features for multilingual processing is discussed by Bender (2009), who argues that even approaches encoding little linguistic information can benefit from language-specific specialisation.

Chapter 3

Features for Chinese Sentiment Classification[1]

There are some distinctive characteristics of the Chinese language that are known to affect language processing. This chapter presents an investigation of these in connection with sentiment classification. Section 3.1 outlines problems with conceptualising Chinese text as comprising a sequence of 'words'. In particular, the problem of automatically segmenting text into words is discussed and tested in an experiment. The difficulty of splitting Chinese text into words raises the issue of what kind of basic unit of processing to use in sentiment analysis. Section 3.2 describes kinds of units to be experimented on and the data for the experiments as well as basic concepts, algorithms and evaluation metrics. Section 3.3 reports experiments in sentiment classification and discusses the results. Section 3.4 describes extensions to the techniques presented previously and discuses the results. All the experimental results are summarised in section 3.5.

3.1 The 'Word' in Chinese Language Processing

One of the central problems in Chinese NLP in general and in Chinese sentiment analysis in particular is what the basic unit of processing should be. The problem is caused by a distinctive feature of the Chinese language: the absence of orthographically marked word boundaries, while it is widely assumed that a word is of extreme importance for computational language processing. The absence of word delimiters cannot be solved by simply using dictionary lookup (or any other method) to segment a text into words,

[1]The experiments and part of the discussion in this chapter were presented in a condensed form at the Student Workshop at the 45th Meeting of the Association for Computational Linguistics and at the 2007 EUROLAN Doctoral Consortium (Zagibalov, 2007a,b)

because the language has a rather specific structure: a single vocabulary word (e.g. 吃饭 *to eat*) can include a part with no separate meaning as in examples (1-a) and (1-b), but the same 'meaningless' part may be a separate word in other cases (see examples (2-a) and (2-b))

(1) a. 他 吃 饭
he eat (food)
He is eating.

b. 他 吃 半个 小时 的 饭
he eat half hour DE food
He has been eating for half an hour.

(2) a. 他 吃 好 饭
he eat good food
He is eating good food.

b. 饭 他 应该 吃
food he must eat
He must eat food.

Example (1-a) demonstrates that the character sequence 吃饭 (*to eat*, lit. *eat food*) is one unit and is a vocabulary word which is not to be segmented into smaller units. The same word is split in (1-b), but the second part still does not have a separate meaning and is used as a way of introduction of an adverbial phrase. However in example (2-a) the second character is not only separated from the first one, but also becomes a word in its own right: a noun with a preceding adjective. In the last example (2-b) the word 饭 (*food*) is used as a topicalized object and is clearly used as a separate word.

The example above is not an exception, but representative of a very frequent morphological phenomenon in Chinese. One of the characteristics of the morphology of the Chinese language is that in many cases words are built in the same way as phrases, which results in words having the same structure as phrases. One of the most widely used patterns is *VERB + OBJECT* as in the example above which is also used for phrases consisting of separate words. Such patterns are very productive which results in a potentially endless number of phrase-like words.

This characteristic of the language makes it difficult even for human beings to segment texts into separate 'words'. Tsai (2001) and Hoosain (1991) show that segmentation is not a part of human understanding of written texts by native speakers of Chinese. They found that a segmented text was more difficult to read for native Chinese speakers as evidenced by a significant slowdown of reading. Tsai also described an experiment where the Chinese had to break a text into words. The results showed substantial disagreement

on where to divide the characters into words.

3.1.1 Preliminary Word Segmentation of Chinese Texts

Even in cases where words can be segmented quite easily by a human, these cases might be very difficult for a computer. A major problem is caused by segmentation ambiguity. There are two types of segmentation ambiguity (Liang, 1987; Guo, 1997): overlapping ambiguity: e.g. 大学 |生活 (*university life*) vs. 大学生 |活 (*(a) student lives*) as shown in examples (3-a) and (3-b); and hidden ambiguity: 个人 vs.个 |人 , as shown in examples (4-a) and (4-b)[2].

(3) a. **大学** 生活 很 有趣
university life very interesting
University life is very interesting.

b. **大学生** 活 不 下去 了
student life not continue LE (sentence-final particle LE)
University students can no longer make a living.

(4) a. 个人 的 力量
individual DE power
the power of an individual

b. 三 个 人 的 力量
three GE person DE power
the power of three persons

These examples show that automatic segmentation needs understanding of context even in such 'easy' cases, which makes complete segmentation a very difficult task. However, many researchers report good results for segmenters they have developed. This can be explained by the fact that in word segmentation experiments in many cases researchers have adopted their subjective understanding of what a word is in Chinese, such that training and test corpora are tagged not according to objective criteria but to ones that the research community have agreed. Xue (2003) comments: "In practice, noting the difficulty in defining wordhood, researchers in automatic word segmentation of Chinese text generally adapt their own working definitions of what a word is, or simply rely on native speakers' subjective judgements. The problem with native speakers' subjective judgements is that native speakers generally show great inconsistency in their judgements of wordhood, as should perhaps be expected given the difficulty of defining what a word is in Chinese".

[2]These examples are taken from Li (2000).

This problem is also crucial for sentiment analysis since some sort of basic unit needs to be defined in order for sentiment information to be associated with it. In many cases, NLP researchers working with Chinese use an initial segmentation module that is intended to break a text into 'words' before it is subjected to further processing. Although this can facilitate the use of subsequent computational techniques, there is no a clear definition of what a 'word' is in the Chinese language, so the use of such segmenters is of dubious theoretical status; indeed, good results have been reported from systems which do not carry out such pre-processing (Foo and Li, 2001; Xu et al., 2004).

Another drawback of using segmenters is that it makes an NLP system language-dependent, as segmenting relies on external language resources or extensive manual annotation. This does not accord with the research programme reported in this thesis which focuses on unsupervised and semi-supervised language processing. Nevertheless it is important to perform an initial investigation of the contribution of segmentation.

3.1.2 Preliminary Segmentation Experiment

To measure the impact that preliminary segmentation has on sentiment classification of Chinese documents, I compared the performance of two supervised classifiers: Naïve Bayes multinomial (NBm) and Support Vector Machine (SVM) [3]. I used the entries in a sentiment dictionary. In the first series of experiments the corpus was split into words (segmented), whereas in the second the features were extracted directly from the text without preliminary segmentation. All the experiments used 10-fold cross-validation.

Sentiment dictionary

For this and all subsequent experiments I used the NTU sentiment dictionary (NTUSD) (Ku et al., 2005)[4]. The dictionary has 2809 items in the 'positive' part and 8273 items in the 'negative'. For these experiments, the dictionary was converted from Traditional Chinese encoding (Big5) into Simplified Chinese encoding (UTF8) and all duplicate entries removed, which resulted in 2,598 items in the 'positive' part and 7,692 items in the 'negative' part.

[3] I used the Weka toolkit (Witten and Frank, 2005)

[4] Ku et al. (2005) automatically generated this dictionary by enlarging an initial manually created seed vocabulary by consulting two thesauri, including 同义词词林 (*The Dictionary of Synonyms*) and the Academia Sinica Bilingual Ontological Wordnet 3.

Test Corpus

All experiments were carried out on a corpus comprised of product reviews downloaded from the web-site IT168[5]. All the reviews were tagged by their authors as either positive or negative. Most reviews consist of two or three parts: positive opinion, negative opinion and comments ('other'), though some reviews have only one part. After all duplicate reviews were removed the final version of the corpus comprised 29,531 reviews of which 23,122 were positive (78%) and 6,409 were negative (22%). The total number of different products in the corpus totalled 10,631, the number of product categories was 255, and most of the reviewed products are items of either software or consumer electronics.

From manual inspection it seemed that some users misused the sentiment tagging facility on the web-site and quite a lot of reviews were tagged erroneously. However, the parts of the reviews were tagged much more accurately so I used only relevant (negative or positive) review parts as the documents in the corpus. The final version of the corpus included only the first 10,000 reviews, whose parts were extracted to make a balanced test corpus. As the corpus consisted of 10 thematic domains (mostly electrical appliances such as digital cameras, mobile phones and computers), I also balanced each of these domains. The resulting corpus contains 8,140 reviews, of which 4,073 are positive and 4,067 are negative[6].

Segmenter

To split the corpus into words I used a publicly available segmenter implemented by Peterson (1999)[7]. The segmenter uses a 138,000 word vocabulary and works with a version of the maximal matching algorithm. Thus when looking for words, it attempts to match the longest word possible. This simple algorithm is surprisingly effective, given a large and diverse lexicon: its segmentation accuracy can be expected to lie around 95% (Wong and Chan, 1996), although one should note the methodological and language-specific issues discussed above in Section 3.1.

The results presented in Table 3.1 show that segmenting the corpus into words affected the performance in a negative way. This suggests that using preliminary segmentation may negatively affect performance of a sentiment classifier.

[5]http://product.it168.com

[6]The corpus is available at http://www.informatics.sussex.ac.uk/users/tz21/.

[7]Available at http://www.mandarintools.com/segmenter.html

	Accuracy	Precision	Recall	F-Measure
NBm (Segmented)	83.59	0.84	0.84	0.84
NBm (Not segmented)	85.61	0.86	0.86	0.86
SVM (Segmented)	81.67	0.83	0.82	0.82
SVM (Not segmented)	85.50	0.86	0.86	0.86

Table 3.1: Results of sentiment classification of product reviews from the web-site IT168, with and without segmentation. The features are NTU sentiment dictionary items.

3.2 Words and Characters as Features for Sentiment Classification

In the absence of preliminary word segmentation, there are two possible types of feature that could be used in Chinese sentiment classification: (vocabulary) words[8] and characters. This section reports experiments into these two types The experiments evaluate various techniques that can facilitate classification including a simple negation check, as there is no a general agreement as to whether feature is useful for sentiment classification. This section also describes and tests an approach which divides the text into *zones*.

Processing based on words and characters are tested separately and in combination. The latter approach is inspired by results published by Nie et al. (2000) who found that for Chinese processing (IR in particular) the most effective kinds of features were a combination of dictionary look up (using the longest-match algorithm) together with single-character unigrams. Yuen et al. (2004) showed that Chinese characters constitute a distinct sub-lexical unit which, though having a smaller number of distinct types, has greater linguistic significance than words. Their experiments on sentiment classification of words by means of characters proved to be effective, achieving a precision of 80.23% and a recall of 85.03% with only 20 characters.

3.2.1 Basic Concepts

To introduce the approach I present some definitions of the concepts that are used in the experiments.

[8]The notion of used is that of *Vocabulary Word* as defined by Li (2000) being the set of of vocabulary items listed in a dictionary.

Basic Units

A basic unit is the smallest linguistic unit used for processing. In this Chapter I experiment with two kinds of basic units: words and characters.

- **Word** Noting the theoretical and practical difficulty of word segmentation in the Chinese language, I use the notion of 'vocabulary word', which is any sequence of characters that forms a vocabulary item in the NTU sentiment dictionary. To avoid confusion, I will also use term 'dictionary item' (DI) as a synonym of 'vocabulary word'.

- **Character** A character is any Chinese character (hieroglyph), excluding punctuation marks and other symbols (stars, bullet points etc.).

Classification Units

A classification unit is a contiguous segment of a document and can be either of the basic units or a larger unit, as indicated below.

- **Unigram** Unigram is a classification unit that consists of a single instance of a basic unit.

- **Zone** Zone is a classification unit that includes one or more basic units and usually is a sub-sentence unit. Zones are delimited by any non-character symbol (comma, full-stop, semicolon, quotation marks etc). If a sentence does not have any delimiters except for the final full-stop, the whole sentence is a zone. The idea of using zones for classification comes from the observations that sentiment classification benefits from consideration of word context, but that sentences may contain two or more opposite sentiments. Thus I decided to include a unit that is usually longer than a word but smaller than a sentence.

- **Sentence** Sentence is a sequence of basic units that ends with a full-stop, question mark, exclamation mark or similar symbol that usually marks the end of a sentence.

Frequency

The sentiment score (see below) is based on a basic unit's relative (normalised) frequency:

$$F_a = \frac{N_a}{N} \tag{3.1}$$

where N_a is the number of times a occurred in a collection of documents and N is the total number of basic units (lexical units or characters, as appropriate) in the collection of documents.

Sentiment Score

Each word (dictionary item) occurring in the positive side of the dictionary is assigned a positive sentiment score of 1 and negative sentiment score 0, and vice versa for words in the negative side.

- **Word Score** The unsupervised approach does not suppose obtaining any data from the test corpus. So initially all the words had a score 1 for the class (sentiment) they present and 0 for the class they are not present.

- **Character Scores** The characters for the experiments are extracted from the NTU sentiment dictionary. Most of the characters occur in both sides of the dictionary: positive and negative. The score for a character with respect to sentiment i (positive or negative) is:

$$Sa_i = \frac{F_i}{F_j} \tag{3.2}$$

  where F_i is the unit's frequency in a document collection of sentiment i, F_j is the character's relative frequency in the opposite side of the dictionary.

  The experiments also test modified sentiment scores: scores with a low or zero frequency 'penalty' and presence-based binary scores. Apart from the sentiment score as described above, the experiments test four **score modifications**[9]

  1. All characters were assigned the basic scores based on the relative frequency calculations, but if $Sa_i < 1$, then $Sa'_i = Sa_i - 1$. The intuition is that if a character is less frequent in one side of the dictionary than in the other, then it should be 'penalised' by being assigned a negative score.

  2. If $Sa_i > 0$, then $Sa'_i = 1$. This score is based on presence of a character in the relevant side of the dictionary, regardless of its frequency.

  3. If $Sa_i \geq 1$, then $Sa'_i = 1$, else $Sa'_i = 0$. This score is a binary version of the basic score.

[9] In the experiments the score modifications are represented by the numbers 1, 2, 3, 4.

4. The same as the first modification, but those characters that do not occur in any item in sentiment class i in the dictionary are assigned the lowest score: $Sa'_i = -1$. In this modification both parts of the character list have an equal number of items.

There are more characters in the negative part of the character list; this can be attributed to the larger size of the negative side of the dictionary. The equal number of characters in the fourth modification is because all characters in both the positive and negative parts of the dictionary receive a score: those characters which do not occur in a given class are assigned -1; the positive part ends up 1386 items with this score (out of 2385).

- **Classification Unit Score** The score of a classification unit is based on the sum of the sentiment scores of the basic units it contains. Thus the score of a unigram is equal to the score of that basic unit. But because Zones and Sentences are composite classification units containing one or more basic units, their scores are equal to the sum of sentiment scores of those basic units, i.e. $Sz_i = \sum_{a \in Z} Sa_i$.
- **Document Score** The score of a document is calculated as the sum of the scores of the classification units it contains.

3.2.2 Experimental Data and Classification Algorithm

The experiments in the remainder of this chapter use the same sentiment dictionary and test corpus as in the previous segmentation experiments (see 3.1.2).

Basic Classification Algorithm

Classification is done by summing up the sentiment scores of all the classification units found in a document. Since there are two classes (positive and negative) the algorithm does this twice to obtain positive and negative scores for a document, which are then compared to make a decision about its sentiment (see Algorithm 1).

3.2.3 Evaluation Metrics and Statistical Significance Test

Accuracy

Since the product review test corpus is balanced with respect to positive and negative documents, I chose accuracy as evaluation metric for all the experiments. I present accuracy

Algorithm 1 Basic Sentiment Classifier

Require: List of basic units a each with sentiment scores Sa_{pos} and Sa_{neg}

Require: Collection of documents D

for each d in D **do**

$$Sd_{pos} = \sum_{a \in d} Sa_{pos}$$

$$Sd_{neg} = \sum_{a \in d} Sa_{neg}$$

end for

for each d in D **do**

if $Sd_{pos} > Sd_{neg}$ **then**

tag d as POS

end if

if $Sd_{neg} > Sd_{pos}$ **then**

tag d as NEG

end if

if $Sd_{pos} == Sd_{neg}$ **then**

do not tag

end if

end for

return Sentiment tags for all classified documents in D

for the whole corpus as well as for each class. Accuracy is calculated as

$$\frac{number\ of\ documents\ classified\ correctly}{total\ number\ of\ documents}$$

Coverage

To measure what proportion of the test data was classified (regardless of correctness), I use coverage:

$$\frac{number\ of\ documents\ classified}{total\ number\ of\ documents}$$

Classification Skew

Sentiment classification in the experiments presented here can be split into two subtasks: finding positive documents and finding negative documents. Both of the subtasks can be evaluated by accuracy. It is very important to consider both positive and negative classification accuracy as the overall accuracy does not reflect the subtask performance: for example a classifier may have accuracies 0.50 and 1.00 for the two classes and overall accuracy of 0.75, while another classifier may have 0.76 and 0.74 with the same overall accuracy. Obviously, despite equal overall accuracy the second classifier is performing much better.

Precision

I also use precision for evaluation of classification performance:

$$\frac{number\ of\ documents\ classified\ correctly}{total\ number\ of\ documents\ classified}$$

Statistical Significance

I use the paired t-test to test if the results of any two experiments are significantly different at the 95% level.

3.3 Experiments with Classification Units

In the experiments presented in this section I test performance of the basic units applying them to classification units. As mentioned above, a classification unit is a unit which is used to define the overall sentiment direction of a document. In the experiments to follow I use three kinds of such units: unigrams, zones and sentences.

	Accuracy				
Basic Unit Kinds	Overall	Positive	Negative	Precision	Coverage
Chars	0.68	0.82	0.54	0.68	1.00
Chars 1	0.66	0.87	0.45	0.66	1.00
Chars 2	0.49	0.01	0.96	0.52	0.95
Chars 3	0.64	0.48	0.80	0.68	0.94
Chars 4	0.70	0.73	0.68	0.70	1.00
Words	0.68	0.71	0.66	0.87	0.79
Words and Chars	0.72	0.84	0.59	0.72	1.00
Words and Chars 1	0.69	0.88	0.50	0.69	1.00
Words and Chars 2	0.54	0.11	0.97	0.57	0.95
Words and Chars 3	0.71	0.58	0.83	0.74	0.95
Words and Chars 4	0.73	0.75	0.71	0.73	1.00

Table 3.2: Results of unigram-based sentiment classification using different types of features

3.3.1 Unigram-Based Classification

Unigram-based classification is based on computing the sum of all the sentiment scores of the basic unit instances found in a document. In the experiments presented here I test the performance of characters, words and combination of words and characters for sentiment classification.

Character-Based Classification Performance

Table 3.2 shows that all character-based classifiers performed reasonably well with only exception being score modification 2. The highest accuracy was achieved by modification 4: the difference between the top two results (modification 4 and the basic score) is significant according to the t-test at the 99% level. What is more important though is the classification skew: only modification 4 produced a balanced classification of both positive and negative documents. The results of the basic score and modification 1 are highly unbalanced and tend to be more accurate in classification of positive documents. In contrast, modifications 2 and 3 are more skewed towards the negative class. This prevalence of negative classification can be attributed to the highly skewed lists used in

	Overall	Positive	Negative	Coverage
One-class chars only	0.53	0.13	0.97	1.00

Table 3.3: Results of sentiment classification with the characters present only in a single class

the experiment, which resulted in very different numbers of characters in the positive and negative parts of the character list: the ratio of positive characters in the list to the negative ones is 1 : 2.18. The results of the basic score and modification 1 are skewed to the positive class because all characters have scores based on their normalised frequency in the appropriate side of the sentiment dictionary. Thus for the basic score and for the score modification 1 the sum of the scores of all characters in the positive side is 1,803.05 and 2,016.16 for the negative side, which makes 1 : 1.12 ratio. Bearing in mind that the number of negative characters is twice as many as the number if the positive ones, on average an item in the positive part of the list has a score almost twice as big as the score of an average item in the negative part. Modification 4 has equal numbers of items in both parts with a increased importance of the items that occur only in one side of the sentiment dictionary. To test if the characters that are present in only one class (the positive or negative side of the dictionary) can produce a good result on their own I ran such a test, but the results were poor (see Table 3.3). This result reflects the degree of skew of the characters that are present in one class only: only 206 such characters were found in the positive word list while 1386 characters were present in the negative side.

Word-Based Classification Performance

The word-based classifier performed at the same level as the second best character-based classifier (basic score): although the word-based classifier produced a more balanced classification, the t-test showed no significant difference between these two classifiers. In contrast, the performance of the best character-based classifier (modification 4) is significantly better than the word-based classifier. But it should be noted that the word-based classifier used only binary scores and in this respect it is closer to character-based classifiers modifications 2 and 3, which performed significantly worse. However, a particular disadvantage of the word-based classifier is its low coverage: 21% of all documents were omitted by the classifier. But in terms of precision the word-based classier performed much better than any other classifier.

Word and Character Combination Performance

The best result in this test was achieved by combining words and characters: the combination of words with the characters with the score modification 4 achieved an accuracy of 0.73, which is significantly better than the character-only classifier. All other combinations of words with characters (basic score and modifications 1 – 3), also performed much better than the character-only version of the classifier. On the other hand, these combinations with the word-based classifier still inherited the degree of skew of the character-based classifiers.

3.3.2 Zone-Based Classification

The zone-based approach to classification is different to classification by means of unigrams: in the zone-based approach the basic units are used to classify zones and the zone classifications then used for document classification. In contrast, unigram-based classification is de-facto a classification based on basic-units (words and characters). The following experiments test if the approach can increase performance and how the length of the classification units affects classification.

Classification of a zone is a simple process identical to classification of documents described above. The sentiment score of a zone is 1 for positive and -1 for negative. If both sentiment scores are equal in a zone then the zone has no sentiment and its score is 0. The sentiment of a document is calculated as the sum of the sentiments of all zones: if the sum is greater than zero then the overall document sentiment is positive, if the sum is less than zero then the sentiment is negative (see Algorithm 2 below).

As stated above (see Section 3.2.2) a zone in these experiments is a sub-sentence unit, consisting of a sequence of characters between punctuation marks. So, for example, the sentence 价格实在太高，这种鼠标普及起来好像不太可能 (*The price is really too high, this mouse will hardly become popular*) would be split into two zones: 价格实在太高 (*the price is too high*) and 这种鼠标普及起来好像不太可能 (*this (computer) mouse will hardly become popular*). Thus instead of immediate classification of documents, the classifier first classifies zones and then uses the zones to classify documents.

This approach did not perform well (see Table 3.4) compared to unigram-based classification (see Table 3.2) for almost all classifiers except the word-based one (this classifier performed very similarly). The character-based classifiers suffered the most significant drop in performance, although binary modifications of the scores (based on the presence of a character in a class rather than on its frequency) do not differ too much. In fact, the

Algorithm 2 Zone-based Sentiment Classifier

Require: List of basic units a each with sentiment scores Sa

Require: document d

split d into zones Z

for each z in Z **do**

$Sz_{pos} = \sum_{a \in z} Sa_{pos}$

$Sz_{neg} = \sum_{a \in z} Sa_{neg}$

if $Sz_{pos} > Sz_{neg}$ **then**

$Sz = 1$

end if

if $Sz_{neg} > Sz_{pos}$ **then**

$Sz = -1$

end if

end for

$Sd = \sum_{z \in Z} Sz$

if $Sd > 0$ **then**

tag d as POS

end if

if $Sd < 0$ **then**

tag d as NEG

end if

return Sentiment tag for d

	Accuracy				
Basic Unit Kinds	Overall	Positive	Negative	Precision	Coverage
Chars	0.61	0.71	0.50	0.69	0.88
Chars 1	0.61	0.79	0.42	0.68	0.88
Chars 2	0.49	0.02	0.96	0.52	0.94
Chars 3	0.62	0.51	0.73	0.72	0.86
Chars 4	0.62	0.65	0.60	0.71	0.97
Words	0.68	0.71	0.66	0.88	0.78
Words and Chars	0.64	0.72	0.55	0.72	0.88
Words and Chars 1	0.62	0.79	0.46	0.70	0.88
Words and Chars 2	0.53	0.11	0.95	0.58	0.92
Words and Chars 3	0.66	0.57	0.76	0.76	0.88
Words and Chars 4	0.64	0.66	0.62	0.73	0.88

Table 3.4: Results of zone-based sentiment classification

zone-based approach introduces a 'score-binarization' level to classification: all character scores are converted into a binary 'zone-score'. It also explains why the word-based classifier performed almost exactly as previously in unigram-based classification: the scores of the words are also binary. Another disadvantage of the approach is a decrease in coverage; again, non-binary classifiers were more affected.

3.3.3 Sentence-Based Classification

The sentence-based classifier uses almost the same algorithm as the zone-based classifier (see Algorithm 2): the zone is replaced by the sentence, but nothing else is changed. The results of the sentence-based classifier are presented in Table 3.5.

Similarly to zone-based classification, the performance of the non-binary classifiers when applied to sentence-based classification is significantly worse. It is also evident that the size of the classification units (zone or sentence) does not influence accuracy.

3.3.4 Discussion

The experiments described above tested two kinds of basic units for sentiment classification, characters and words, applying them separately and in combination under three different settings: unigram-based classification, zone-based classification and sentence-

	Accuracy				
Basic Unit Kinds	Overall	Positive	Negative	Precision	Coverage
Chars	0.63	0.76	0.50	0.69	0.88
Chars 1	0.62	0.82	0.41	0.67	0.92
Chars 2	0.49	0.01	0.96	0.52	0.94
Chars 3	0.61	0.47	0.75	0.70	0.88
Chars 4	0.65	0.67	0.62	0.71	0.92
Words	0.67	0.70	0.66	0.88	0.77
Words and Chars	0.67	0.78	0.55	0.72	0.92
Words and Chars 1	0.64	0.83	0.46	0.70	0.92
Words and Chars 2	0.53	0.11	0.96	0.58	0.92
Words and Chars 3	0.67	0.56	0.78	0.75	0.89
Words and Chars 4	0.67	0.69	0.65	0.73	0.92

Table 3.5: Results of sentence-based sentiment classification

based classification. The main purpose of the experiments was to find the best kind of basic units for sentiment classification and investigate how classification units affect the performance of the classifiers.

Basic Units

The highest accuracy (0.73) in the experiments was achieved by the combination of words with characters (score modification 4) in the unigram-based classification test (see Table 3.2). In terms of accuracy the best character-based classifier reached 0.70 in the same settings. The word-based classifier in all tests achieved approximately same accuracy of 0.68. Only in the zone-based experiments did the word-based classifier perform slightly better than the former two. But in terms of precision the word-based classifier performed best. The differences in performance among all these classifiers are significant at the 99% level.

Characters There are five variants of the sentiment score for characters: one basic and four modifications of it (as described in Section 3.2.1). The binary modifications (2 and 3) did not perform well in any of the tests, while the basic score and its modification 1 and especially modification 4 performed much better. The best performance achieved by a

character-based classifier was 0.70 in unigram-based classification, and the worst was 0.49 (modification 2). The performance of the character-based classifiers depends on the kind of score that is used for sentiment classification: the presence-based score (modification 2) performed very poorly. The main reason for this is that characters do not usually form semantically independent units (unlike words and phrases) and often have rather vague and ambiguous meanings. This was reflected in their distribution across the sentiment classes: the most frequent characters were present in both classes and so the presence-based score could not contribute to classification. Score modification 3, also being binary, to a certain degree reflected the predominant distribution of the characters and performed better, but was still much inferior to the words (also having binary scores). The best character-based classifiers used normalised frequency based scores, which represented the actual distribution of the characters between the two classes.

Words The performance of the word-based classifier was almost independent of the classification units. It was relatively high (about 0.68), but was significantly worse than the best scores achieved by the two other kinds of units. Still, taking the binary nature of the word score into consideration, the word-based classifier clearly outperformed characters with the same kind of score (modifications 2 and 3). This suggests that words might have even higher performance if scores based on normalised frequency were used. The drawback of the word-based classifier is its relatively low coverage: up to 23% of documents were not classified in the classification experiments. The low coverage might be a result of the more domain-dependent nature of words: although the list of sentiment words is quite large, it does not include all the words used in the corpus to express attitude since many of these words have sentiment-related meaning only in the context of a particular topic. However, the high precision (up to 0.88) indicates the importance of capturing a bigger context: words are longer than characters and cover bigger portions of text.Indeed, many of the 'words' are actually sentiment-bearing phrases which cover all relevant context.

Classification Precision Although the coverage of the word-based classifier was not high, it achieved a very high precision, compared to the other classifiers (see Table 3.6). This can be attributed to the more context-dependent nature of the word as compared to the character. Table 3.6 summarises the experiments with respect to precision: the word-based classifier performs significantly better in all the tests.

Basic Unit Kinds	Unigram	Zone	Sentence
Chars	0.68	0.69	0.69
Chars 1	0.66	0.68	0.67
Chars 2	0.52	0.52	0.52
Chars 3	0.68	0.72	0.70
Chars 4	0.70	0.71	0.71
Words	**0.87**	**0.88**	**0.88**
Words and Chars	0.72	0.72	0.72
Words and Chars 1	0.69	0.70	0.70
Words and Chars 2	0.57	0.58	0.58
Words and Chars 3	0.74	0.76	0.75
Words and Chars 4	0.73	0.73	0.73

Table 3.6: Precision of the unigram, zone-based and sentence-based sentiment classifiers

Words and Characters Words and characters when combined together performed relatively well, showing the best features of both: accuracy was never too bad, and coverage was fairly good. In unigram-based classification, three out of five combinations (with the basic score and modifications 3 and 4) performed significantly better (at 99% level) than the other kinds of basic units, with the highest accuracy of 0.73 (see Table 3.2). The combination of characters and words was able to classify many more documents than the word-based classifier (at least 86% against 77%). It is also worth noting that all character-based classifiers benefited from combination with words and performed better in all the tests.

Classification Units

Another task of the experiments was to explore the influence of the classification unit on classification performance. I compared the performance of the classifiers based on unigrams, zones and sentences.

Unigrams The highest accuracy achieved with unigram-based classification was 0.73 (characters combined with words), the average accuracy was 0.66 (0.67 if the lowest and the highest results are excluded).

Zones The introduction of zones decreased performance significantly: the highest accuracy was achieved by the word-based classifier (0.68) and average accuracy was 0.61.

Sentences The results of sentence-based classification are very close to zone-based: the average was 0.62 with the top result being 0.67.

The results obtained from the experiments indicate that the best classifier is one based on the combination of words and characters. It is also possible to conclude that scoring based on normalised frequency is better for Chinese sentiment classification than a binary score. The presence-based binary score is not suitable for character-based classification, but performs well with words. The results also suggest that for a sentiment classification a unigram-based approach is the best.

3.4 Sentiment Score Extensions

Although the preliminary experiments reported above produced some promising results, the characteristics of sentiment, and language more generally, suggest some possible extensions to the techniques which might lead to improved results. The extensions include score calculation adjustments for negation, input data degree of skew and basic unit length. This section presents the results of the experiments carried out using the same classifier as above (see Algorithm 1 and Algorithm 2) with the only difference being in the score calculation.

3.4.1 Negation Check

Negation plays an important role in language. It is also important in evaluative language, as *good* and *not good* express different sentiments in most contexts. Most researchers agree that including information about negation improves sentiment classification accuracy but detecting and integrating this information may be a difficult task (see Section 2.2.2). In this study the negation check is a very simple routine, based on regular expression patterns to find out if a word or a character is preceded by a negation up to 2 characters previously. If a negation is found the score is multiplied by -1:

$$Sa' = Sa * -1 \tag{3.3}$$

Basic Unit Kinds	Accuracy			Precision	Coverage
	Overall	Positive	Negative		
Chars	0.71	0.83	0.59	0.71	1.00
Chars 1	0.70	0.87	0.53	0.70	0.92
Chars 2	0.48	0.04	0.92	0.70	0.94
Chars 3	0.66	0.53	0.80	0.71	0.94
Chars 4	0.73	0.74	0.73	0.73	1.00
Words	0.70	0.69	0.70	0.87	0.81
Words and Chars	0.75	0.84	0.66	0.75	1.00
Words and Chars 1	0.73	0.88	0.58	0.73	1.00
Words and Chars 2	0.54	0.14	0.94	0.58	0.94
Words and Chars 3	0.73	0.62	0.84	0.76	0.95
Words and Chars 4	0.76	0.76	0.75	0.76	1.00

Table 3.7: Results of unigram-based sentiment classification with negation

I used only five of the most widely used negations in Chinese: 不 (*bu*), 没有 (*meiyou*), 不会 (*buhui*), 摆脱 (*baituo*), 免去 (*mianqu*), 避免 (*bimian*)[10].

The negation check was applied to all the classifiers in all the settings used in previous experiments: the character-based, word-based and combined classifiers were re-run in unigram-, zone- and sentence-based classification settings.

Unigram-Based Classification

Table 3.7 presents the results of the unigram-based experiments with negation. All of the classifiers performed significantly better compared to the same classification settings without negation (see Table 3.2). The only exception is the character-based classifier with the score modification 2 and its combination with the word list. The biggest improvement (+0.04) was achieved by the classifiers with the character score modification 1. The better performance is mostly due to improvement in classification of negative documents: from 0.45 to 0.53 for the character-based classifier and from 0.50 to 0.58 for the combined word and character classifier. It should be noted also that all of the classifiers produced a more balanced classification.

[10]The first two negation words cover most of the negation in the Chinese language (Tan, 2002), the other four negations are also common in general usage.

	Accuracy				
Basic Unit Kinds	Overall	Positive	Negative	Precision	Coverage
Chars	0.66	0.73	0.58	0.75	0.88
Chars 1	0.67	0.81	0.53	0.76	0.88
Chars 2	0.48	0.02	0.93	0.51	0.93
Chars 3	0.66	0.55	0.78	0.76	0.87
Chars 4	0.67	0.67	0.68	0.76	0.88
Words	0.72	0.71	0.72	0.90	0.79
Words and Chars	0.69	0.74	0.64	0.78	0.89
Words and Chars 1	0.69	0.81	0.57	0.78	0.89
Words and Chars 2	0.54	0.12	0.95	0.59	0.91
Words and Chars 3	0.71	0.60	0.81	0.80	0.88
Words and Chars 4	0.72	0.71	0.72	0.78	0.89

Table 3.8: Results of zone-based sentiment classification with negation

Zone-Based Classification

The zone-based classification results (see Table 3.8) show the same kind of improvement: all of the classifiers improved their classification on the class on which they performed worse in the previous experiments (see Table 3.4).

Sentence-Based Classification

Table 3.9 shows significant improvements in sentence-based classification compared to classification without the negation check.

Overall, the experiments show that negation significantly improved the performance of all the classifiers (except modification 2) by producing more balanced output. Another notable difference introduced by the negation check is a significant improvement of the word-based classifier using zones: in previous experiments this classifier did not show any significant variation in performance between the various classification settings (see Tables 3.2, 3.4 and 3.5).

	Accuracy				
Basic Unit Kinds	Overall	Positive	Negative	Precision	Coverage
Chars	0.67	0.77	0.57	0.73	0.92
Chars 1	0.67	0.83	0.51	0.73	0.92
Chars 2	0.47	0.03	0.92	0.51	0.93
Chars 3	0.65	0.52	0.77	0.73	0.88
Chars 4	0.69	0.69	0.68	0.75	0.92
Words	0.69	0.69	0.69	0.89	0.78
Words and Chars	0.71	0.78	0.63	0.77	0.92
Words and Chars 1	0.70	0.83	0.56	0.75	0.92
Words and Chars 2	0.53	0.13	0.94	0.58	0.91
Words and Chars 3	0.70	0.59	0.81	0.78	0.90
Words and Chars 4	0.72	0.71	0.71	0.77	0.92

Table 3.9: Results of sentence-based sentiment classification with negation

3.4.2 Length Ratio

Unlike characters, words (dictionary items) have different lengths and can capture various portions of context. For example, if a dictionary item covers most of a phrase a classifier can more reliably detect the phrase's sentiment. For example in the sentence 实在是不伦不类！(*It's really neither fish nor fowl!*) there are two matching dictionary items in the sentiment dictionary: 实在 (*really*) and 不伦不类 (*neither fish nor fowl*). The first item is in the positive side of the dictionary and the second is in the negative. If a classifier compares their scores (1 for positive and -1 for negative), then it will not be able to make any decision, but if it were to compare their lengths (2 and 4) and combine this with their scores ($2 * 1 = 2$ and $4 * -1 = -4$), the whole sentence would be tagged negative.

A length-sensitive sentiment score can be defined as:

$$Score = \frac{L_w^2}{L_{cu}} \tag{3.4}$$

where L_w is the length of a word and L_{cu} is the length of the relevant enclosing classification unit. The numerator L_w is squared to influence importance of longer units.

Since all characters have length 1, there is no point in testing character-only classifiers in conjunction with the length ratio.

	Accuracy				
Basic Unit Kinds	Overall	Positive	Negative	Precision	Coverage
Words	0.69	0.71	0.67	0.86	0.80
Words and Chars	0.78	0.85	0.70	0.78	1.00
Words and Chars 1	0.75	0.88	0.62	0.75	1.00
Words and Chars 2	0.72	0.52	0.93	0.75	0.97
Words and Chars 3	0.78	0.72	0.77	0.80	0.97
Words and Chars 4	0.78	0.78	0.77	0.78	1.00

Table 3.10: Results of unigram-based sentiment classification with length ratio

	Accuracy				
Basic Unit Kinds	Overall	Positive	Negative	Precision	Coverage
Words	0.69	0.71	0.66	0.88	0.78
Words and Chars	0.69	0.73	0.64	0.77	0.89
Words and Chars 1	0.67	0.78	0.56	0.76	0.89
Words and Chars 2	0.61	0.29	0.93	0.68	0.89
Words and Chars 3	0.71	0.62	0.79	0.80	0.88
Words and Chars 4	0.68	0.68	0.68	0.77	0.89

Table 3.11: Results of zone-based sentiment classification with length ratio

Unigram-Based Classification The results presented in Table 3.10 show some increases in performance of all the classifiers. But for the word-based classifier the improvement is not statistically significant, nor is it for the combined classifier with score modification 4. The word-only classifier's failure to increase performance can be explained by the fact that more than 70% of all the sentiment words used in the corpus have length 2, and these words are the most frequent ones. This means that on most occasions the length did not affect the score.

Zone-Based Classifier

In zone-based classification (see Table 3.11) only combined classifiers with score modifications 0, 1 and 4 showed improved performance. The word-based classifier did not show any significant improvement.

	Accuracy				
Basic Unit Kinds	Overall	Positive	Negative	Precision	Coverage
Words	0.67	0.70	0.65	0.87	0.78
Words and Chars	0.73	0.80	0.66	0.78	0.93
Words and Chars 1	0.70	0.83	0.57	0.76	0.93
Words and Chars 2	0.68	0.45	0.91	0.75	0.91
Words and Chars 3	0.72	0.73	0.71	0.81	0.91
Words and Chars 4	0.68	0.68	0.68	0.78	0.92

Table 3.12: Results of sentence-based sentiment classification with length ratio

Sentence-Based Classifier

In sentence-based classification, the word-based classifier and the combined classifier with score modification 4 also did not perform any better compared to non-length based classification (see Table 3.12).

Despite a very low impact on performance, the experiments with the length-based score extension revealed an interesting tendency: only the combined classifiers improved performance. Since the length ratio cannot affect characters (they have a constant length), the only explanation is that the length ratio increases the relative importance of words during classification, as they are longer than characters. But it is not possible to say that only words contribute to performance: the combined classifiers perform significantly better in terms of accuracy and coverage than the word-only classifier.

3.4.3 Discussion

It seems that the length-ratio and negation check extensions are useful for Chinese sentiment classification. Both of these significantly increased the performance of the combined classifiers and in some occasions the word-based classifier. The two techniques combined led to further improved performance (see Table 3.13).

The highest accuracy was achieved by the combined classifiers with the basic character score modification 3 and 4. Although the accuracy of the former is a little bit higher, its classification results differ for positive and negative classes. This is not the case for the combined classifier with modification 4. The word-based classifier also produced a fairly balanced classification. But this classifier showed no significant difference in performance compared with the negation check classification.

	Accuracy				
Basic Unit Kinds	Overall	Positive	Negative	Precision	Coverage
Words	0.70	0.69	0.70	0.86	0.81
Words and Chars	0.80	0.84	0.75	0.80	1.00
Words and Chars 1	0.78	0.87	0.69	0.78	1.00
Words and Chars 2	0.72	0.53	0.91	0.75	0.97
Words and Chars 3	0.79	0.73	0.85	0.81	0.97
Words and Chars 4	0.79	0.79	0.80	0.79	1.00

Table 3.13: Results of unigram-based sentiment classification with length ratio and negation check combined

3.5 Summary

The experiments in this chapter tested two different basic units for Chinese sentiment classification: words and characters, as well as the combination of the two. The main aim of the experiments was to find the best basic unit for classification. Judging from the results obtained, the best approach is the combination of the two basic units: the performance of this combination was the best in terms of accuracy and degree of classification skew.

Another aim was to measure the performance for different classification units. In most occasions unigrams performed much better than the zones and sentences.

The final experiments investigated whether extensions to the sentiment score computation can improve performance. The experiments showed that the most useful extension is the negation check, which improved the performance of almost all of the classifiers. Factoring in word length was useful for the combined classifiers, but did not improve the word-based classifier.

3.5.1 Accuracy

The best accuracy achieved in the tests was 0.80. The combined classifier with the basic score performed this well in unigram-based classification with the negation check and length ratio (see Table 3.13). The same settings also helped to achieve high accuracy (0.79) in the combined classifiers with score modifications 4 and 3.

3.5.2 Coverage

In terms of coverage, the character-based and combined classifiers usually performed much better at classifying all of the documents in the corpus. Word-based classifiers covered fewer documents in all experiments.

3.5.3 Skew

The overall performance measure combines classification performance over two classes. The most balanced classification was produced by the combined classifier with score modification 4: in almost all the tests classification results for the two classes matched. The word-based classifier also performed equally well on both classes of documents.

3.5.4 Precision

Although the main evaluation metric was accuracy, precision is also important and particularly so for the work reported in subsequent chapters. In terms of precision the word-based classifier outperforms all the others. It is also important to note that in most cases the highest precision classification was achieved in zone-based classification (see Table 3.6). Table 3.14 presents the top classifiers with the respect to precision. The table shows that the highest precision was achieved by the combination of zones and negation. The difference between the top two classifiers is statistically significant at the 95% level. The best non-word-based classifier is a combined classifier with the length-based extension: it achieved a precision of 0.81; however this is far behind the 0.90 of the best word-based classier. Another interesting observation is that complex classification units are more important for precision than the negation check: compare lines 3 and 4 with lines 5 and 6. The precision achieved is also higher than in the experiment reported at the beginning of this Chapter with supervised classifiers (0.85, see Table 3.1).

3.5.5 Conclusion

The best classifier for Chinese sentiment classification is the unigram-based combined classifier with score modification 4 with the length ratio and negation check: it achieved one of the highest accuracies while maintaining balanced classification and high coverage. However, for high precision the best choice is the word-based classifier using zones as classification units and the negation check.

Precision	Chars modifications	Negation	Length ratio	Classification Unit
0.90	-	yes	no	zone
0.89	-	yes	no	sentence
0.88	-	no	no	zone
0.88	-	no	no	sentence
0.87	-	yes	no	unigrams
0.87	-	yes	no	unigrams
0.81	3	yes	yes	zones

Table 3.14: Results of word-based sentiment with different features

Chapter 4

Classifier Improvements and Extensions[1]

The previous chapter presented and evaluated a number of sentiment classifiers based on different kinds of features and demonstrated that out of the techniques tested the best performance was achieved by a classifier that used words and characters combined with a check for negation and a length-based weighting of lexical units (Section 3.4.2). All the classifiers were based on a generic sentiment dictionary (Section 3.2.2), the biggest disadvantage of which is that it is not domain-specific: it contains no domain-specific sentiment-bearing lexical units. Although all of the classifiers used a generic sentiment dictionary that is supposed to have good coverage sentiment-bearing words, the dictionary cannot include all possible sentiment words for all possible contexts: some words have sentiment-relevant meaning only in a certain context or with respect to a particular topic. This means that even a linguistically flawless list of words cannot be equally effective for all possible domains. This chapter investigates if it is possible to improve the results by adapting a classifier to a domain. All of the techniques presented are based on an unsupervised approach as this makes it unnecessary to have annotated data in each domain of application and to facilitate application to different languages.

This chapter is structured as follows. The first set of experiments tests if the dictionary can improve performance if all its items are assigned corpus-relevant sentiment scores (see Section 4.1). This section also presents experiments on automatically building a corpus-relevant list of lexical units using manually chosen seed words (see 4.1.1). A technique for automatically finding such seed words is tested in Section 4.2.

[1]The experiments and part of the discussion in section 4.2 were presented in a condensed form at the 22nd International Conference on Computational Linguistics (Zagibalov and Carroll, 2008b)

Seeds on their own cannot produce a good classification due to their small number. Section 4.3 describes a way to overcome this problem by applying an iterative approach. This section also tests two techniques for increasing the precision of the iterative classifier: filtering scores of found lexical units, to reduce the number of non-discriminative lexical units and using difference between positive and negative zones to rank classification results by their reliability. Further classification accuracy improvements are based on extending the unsupervised classifier with supervised techniques: Naïve Bayes (multinomial) and Support Vector Machine. The machine-learning extension is based on using classification data produced by an unsupervised classifier to train supervised classifiers.

Section 4.5 summarises the experimental results described in this Chapter.

4.1 Dictionary Adjustment

A major disadvantage of a generic sentiment dictionary is that it does not take into account domain-specific ways of expressing sentiments. Quite often the same word might have opposite meanings in different contexts (e.g. '*unpredictable plot*' and '*unpredictable steering*'). One possible solution is to assign domain-dependent sentiment scores to every dictionary item. These scores would reflect how an item is connected with sentiment in a particular domain. This section presents experiments on dictionary adjustment by means of calculating domain-dependent sentiment scores. The scores can be obtained from a preliminary tagged corpus, but such an approach would no longer be unsupervised. To keep the system unsupervised I used a classifier described in the previous Chapter (Section 3.2.2) to extract a sentiment-classified subcorpus from a raw corpus. The most important feature of such a subcorpus is precision (providing the recall is high enough) rather than accuracy. As the experiments described in the previous chapter show, the highest precision was achieved by a word-based classifier with the negation check and using zones as the unit of classification. This classifier was used as the basis for the experiments described in this Chapter.

4.1.1 Adjustment to Corpus

I used the classifier to extract a subcorpus by labelling documents in the raw corpus according to the classification results. The extracted subcorpus, consisting of 6447 documents of which 3178 are classified as positive and 3269 are classified as negative, was used as a training corpus in subsequent experiments. The corpus built using this data did not have a very high accuracy (0.72), but it was balanced having similar number of positive and

Positive	Translation	Score	Negative	Translation	Score
独特	unique	66.95	不清楚	not clear	76.19
优秀	outstanding	55.38	根本就	absolutely (not)	58.04
良好	fine	52.07	突然	suddenly	51.99
与.*匹敌	matching	50.41	不爽	out of sorts	51.99
轻松	easy	46.69	郁闷	gloomy; depressed	49.57
迅速	fast	45.45	失去.*的	(having) lost smth	47.45
效率	efficiency	42.97	严重	severe(ly)	44.33
独特的	unique	34.70	不合理	not suitable	42.31
强大	powerful	34.70	严重的	severe	42.31

Table 4.1: List of top 10 words

negative documents. The extracted subcorpus had high precision (0.90) which, as noted above, is important for any subsequent training process. I split the corpus into two parts: one containing only documents which were tagged as positive, and the other containing only negative documents. I used the same approach to the sentiment score calculation as I did for the characters (Section 3.2.1)[2]. Those lexical units that were present only in one of the parts of the training subcorpus were assigned the minimal frequency ($N_a = 1$): similar to the score modification 4, which proved to be the best for characters (Section 3.2.1). The resulting positive and negative word lists contained 639 and 1524 items respectively, each word having a sentiment score (see Table 4.1). The sentiment word lists obtained were then used to re-run the two classifiers: a word-based classifier and a combined word- and character-based classifier. The results for both of them showed statistically significant improvements in performance compared to using the lexical units without any score adjustment (see Table 4.2 and Table 4.3).

To check if the proposed approach helps to adapt a classifier to a domain (rather then to a set of documents), I randomly split the corpus into two parts (with $4:1$ ratio). The larger part was used to calculate the scores for the lexical units in the sentiment dictionary. The smaller part was used for testing the effect the adjusted scores have on classification. The experiment was run five times, each time with a new random split.

[2]Since I did not use a word segmenter I assumed that the average length of a word in Chinese is 2.5 characters and divided the total number of characters by this figure to obtain the total number of 'words' in corpus.

	Accuracy	Precision	Recall	F-measure
Before adjustment	0.72	0.90	0.72	0.80
After adjustment	0.74	0.91	0.74	0.82

Table 4.2: Results of word-based sentiment classification before and after feature adjustment

	Accuracy	Precision	Recall	F-measure
Before adjustment	0.79	0.79	0.79	0.79
After adjustment	0.83	0.83	0.83	0.83

Table 4.3: Results of combined classifier sentiment classification before and after feature adjustment

	Accuracy	Precision	Recall	F-measure
Before adjustment	0.72	0.90	0.72	0.80
After adjustment	0.74	0.91	0.74	0.81

Table 4.4: Average of the results of five runs on a test corpus of the word classifier sentiment classification before and after feature adjustment

Corpus/product type	*Number of Reviews*
Mobile phones	2317
Digital cameras	1705
MP3 players	779
Monitors	683
Office equipment (copiers, multifunction devices, scanners)	611
Printers (laser, inkjet)	569
Computer peripherals (mice, keyboards, speakers)	457
Video cameras and lenses	361
Networking (routers, network cards)	350
Computer parts (CD-drives, motherboards)	308

Table 4.5: Product types and sizes of the test corpora.

Table 4.4 shows that words with adjusted scores perform slightly better (the improvement is statistically significant) than without.

4.1.2 Adjustment to Topic

The corpus used in the previous experiments consisted of customer reviews of consumer electronics of different kinds. This provides me an opportunity to split the corpus into different topic-based subcorpora (*topics* for short) and test the ability of the approach to find topic-dependent scores for the items in the sentiment dictionary. The experiments presented below used the same corpus as described in Section 3.1.2, but in order to to extract domain-specific scores, the corpus was split into 10 topics (see Table 4.5).

Five of the corpora combine smaller ones of 100–250 reviews each (as indicated in parentheses in Table 4.5) in order to have reasonable amounts of data in each. Each corpus has equal numbers of positive and negative reviews so that it is possible to derive strong comparator accuracy figures by applying supervised classifiers[3] (studying the effect of skewed class distributions is out of the scope of this study).

Table 4.6 compares the results of two classifications. The left side of the table presents the results of classification using the sentiment dictionary without any topic-specific adjustment. The right side contains results of classification using the same dictionary but with scores calculated on the basis of the extracted subset of documents. Although all

[3]This corpus is publicly available at http://www.informatics.sussex.ac.uk/users/tz21/

Corpus	No Scores			Scores		
	P	R	F	P	R	F
Mobile phones	0.87	0.71	0.78	0.87	0.72	0.79
Digital cameras	0.88	0.63	0.74	0.87	0.64	0.74
MP3 players	0.90	0.71	0.79	0.89	0.72	0.80
Monitors	0.87	0.71	0.78	0.87	0.74	0.80
Office equipment	0.90	0.72	0.80	0.87	0.74	0.80
Printers	0.90	0.71	0.79	0.88	0.71	0.79
Computer peripherals	0.93	0.79	0.85	0.91	0.81	0.86
Video	0.90	0.75	0.82	0.86	0.73	0.79
Networking	0.85	0.65	0.74	0.83	0.68	0.74
Computer parts	0.84	0.65	0.73	0.82	0.62	0.71
Macroaverage	0.88	0.70	0.78	0.87	0.71	0.78

Table 4.6: Classification results of different topics with the sentiment vocabulary with (*Scores*) and without topic-adjusted scores (*No Scores*). P is precision, R is recall, F is F-measure. Difference in the results for all corpora is statistically significant.

the results are significantly different (in terms of the paired t-test) there is only a slight increase in recall at the expense of precision.

4.1.3 Discussion

Calculating domain-specific scores for lexical items improved performance across the corpus but only marginally altered results of classification of the same corpus split into separate topics. This may be due to the generic nature of the dictionary: it contains only generic indicators of sentiment and is missing a lot of domain- and topic-specific ones. Thus a larger corpus has a better chance to improve performance with this generic sentiment dictionary as its items occur more frequently than in a small corpus. But if the same collection is split into topical corpora where the role of domain-relevant words is more important (the smaller collection is the more important every lexical unit becomes) then a generic dictionary fails to improve even after being adjusted with domain-related scores. Another important feature of a sentiment corpus is its topical coherence. The more closely related (in terms of the topic) documents are, the more important topic-related words may be and the smaller the improvement one can expect with a generic sentiment

dictionary. This explains why the generic dictionary performed better on a more generic corpus compared to the smaller more topic-oriented collections extracted from it.

4.2 Vocabulary Extraction

The experiments in the previous section suggest that a generic sentiment dictionary has limited potential to improve performance even with domain-specific scores used for adjustment of the dictionary item scores. If it is not possible to substantially increase performance by adjusting an existing generic dictionary then the next possibility to explore is creating domain-specific vocabularies.

4.2.1 Seed-Based Approach

Although the experiments described above suggest that classification results can potentially be improved by adjusting the vocabulary to the domain, the inflexibility of the precompiled vocabulary prevents it from full adjustment to a domain. Moreover, the vocabulary-based approach prevents a system from being multilingual as the very need for a comprehensive dictionary inevitably makes the system language-dependent. Another problem of the dictionary-based approach is that it is virtually impossible to include all important domain-related words. One way to solve the problem may be finding domain-related lexical units from a subcorpus which was extracted by an unsupervised classifier and calculating their sentiment scores for a given topic. This would pave the way to creating a domain-specific vocabulary to be used for classification. But this technique requires extraction of a subcorpus from a corpus to be classified so that words can be extracted from it and scores calculated for them. Such a subcorpus is a product of classification that needs some input data to start with. This input could be several lexical units (seeds) used for initial classification and extraction of the subcorpus.

Seeds

The experiments below test a number of seeds, which were selected intuitively without any special preliminary study of their potential effectiveness for the task of sentiment classification. This approach is justified by the unsupervised paradigm of the research, as any 'learned' data would contradict it. Two types of seed word lists were investigated: six one-word seed lists (see Table 4.7) and three multi-word seed lists consisting of the single seeds in various combinations (see Table 4.8). All the seeds had their sentiment scores set to 1 and the classifier was run with the seed lists taking the place of the sentiment

Seed list name	Seed	Translation	Sentiment
good	好	good	POS
very_good	很好	very good	POS
comfortable	方便	comfortable, convenient	POS
bad	坏	bad	NEG
too_bad	太差	too bad	NEG
poor	差	poor	NEG

Table 4.7: Single word seed lists

Seed list name	Seeds	Translation	Sentiment
allPOS	好	good	POS
	很好	very good	POS
	方便	comfortable, convenient	POS
allNEG	坏	bad	NEG
	太差	too bad	NEG
	差	poor	NEG
all	*all above*	*see above*	*mixed, see above*

Table 4.8: Multi-word seed lists

dictionary. In single seed classification, negative zones are found by means of the negation check (so 'not' + 'good' = negative item).

Seed-based Classification Results

Table 4.9 shows the results produced by the classifier using the seed lists on the entire corpus. As would be expected the multi-seed lists produced better classifications in terms of recall, but the single seeds achieved much higher precision. The only exception was the seed 好 *good* which performed similarly to the multi-seed lists: relatively high recall and low precision. This performance can be attributed to the high frequency of the word in the corpus and its ambiguity[4]. The biggest shortcoming of the classification results is

[4]The word 好 (*good*) is relatively ambiguous: in some contexts it means *to like* or acts as the adverbial *very*, and is often used as part of other words (although usually contributing a positive meaning). But since it is one of the most frequent units in the Chinese language, it is likely to occur in a relatively large number of reviews.

Seed list name	P	R	F	Acc	AccP	AccN
good	0.75	0.23	0.35	0.23	0.33	0.13
very_good	0.94	0.06	0.11	0.06	0.10	0.02
comfortable	0.96	0.09	0.17	0.09	0.15	0.04
bad	0.88	0.04	0.07	0.04	0.00	0.07
too_bad	0.99	0.02	0.04	0.02	0.00	0.04
poor	0.88	0.09	0.17	0.09	0.00	0.18
allPOS	0.80	0.29	0.42	0.29	0.42	0.15
allNEG	0.86	0.12	0.21	0.12	0.00	0.24
all	0.85	0.37	0.51	0.37	0.41	0.32

Table 4.9: Results of the seed list classifier sentiment classification. P is precision, R is recall, F is F-measure; Acc is accuracy, $AccP$ is accuracy of the positive class and $AccN$ is accuracy of the negative class.

low and highly skewed accuracy. The results also suggest that seed selection can affect classification: the F-measure varies from 0.04 to 0.35 in single seed classification; the negative seeds have a lower frequency than positive ones which is reflected in lower recall.

To test how the seeds perform on separate topics extracted form the corpus I tested only the three seed lists that performed the best: *good*, *allPOS* and *all*. Table 4.10 presents results obtained after classification of the topics using these three seed lists. The results resemble their performance on the whole corpus: the largest seed list *all* outperforms *allPOS* and *good*. But these results also resemble the results of the dictionary adjustment experiment: classification of the whole corpus is better than average performance on the topics extracted from the corpus (see the bottom line in Table 4.9), which can also be attributed to the fact that the seeds used in these tests are also generic ones and do not scale down to smaller collections which are more topically coherent.

4.2.2 Automatic Seed Word Selection

The previous experiments showed that not all seeds perform equally. This may be attributed to the generic nature of the seeds used. The next is therefore to test the possibility of automatically finding domain-dependent seeds that could potentially outperform generic ones.

Corpus	good			allPOS			all		
	P	R	F	P	R	F	P	R	F
Mobile phones	0.77	0.27	0.40	0.81	0.32	0.46	0.85	0.41	0.55
Digital cameras	0.76	0.19	0.30	0.80	0.24	0.37	0.86	0.35	0.50
MP3 players	0.77	0.21	0.33	0.83	0.28	0.42	0.88	0.35	0.50
Monitors	0.68	0.22	0.34	0.73	0.28	0.41	0.79	0.34	0.47
Office equipment	0.81	0.22	0.35	0.86	0.31	0.45	0.89	0.39	0.55
Printers	0.76	0.20	0.31	0.80	0.27	0.40	0.86	0.33	0.48
Computer peripherals	0.71	0.24	0.36	0.75	0.30	0.43	0.79	0.35	0.48
Video cameras and lenses	0.75	0.19	0.31	0.82	0.29	0.43	0.87	0.36	0.51
Networking	0.63	0.21	0.31	0.67	0.25	0.37	0.75	0.31	0.44
Computer parts	0.69	0.18	0.28	0.73	0.21	0.32	0.81	0.30	0.44
Macroaverage	0.73	0.21	0.33	0.78	0.28	0.41	0.84	0.35	0.49
Difference	-0.02	-0.02	-0.02	-0.02	-0.01	-0.01	-0.01	-0.02	-0.02

Table 4.10: Classification results with the seed *good*, and seed lists *allPOS* and *all*. *P* is precision, *R* is recall, *F* is F-measure. *Difference* shows the change in performance compared with the corpus-wise classification (see Table 4.9). The differences in the results for all seed lists are statistically significant.

Lexical Unit

As discussed in the previous chapter (Section 3.1.1), the concept of 'word' segmentation in Chinese NLP and so the term 'seed word' is not very accurate since it is not possible to guarantee that extracted units will always form words in the normally understood sense. Fortunately, the results of the experiments with different kinds of features (Section 3.5.1) showed that high accuracy can be achieved by a combination of both words and characters, which makes it possible not to use words as basic units. Instead, I use lexical units which could be any combination of characters constituting parts of words, words or even phrases. This approach avoids the need for word segmentation, and can also capture some grammatical and syntactic information, because lexical units can incorporate grammar words and parts of grammatical constructions. Example (1) shows a combination of two words that was extracted as one unit. This unit provides a context for each of its two members and potentially is a better indicator of sentiment than either of them on their own. The lexical unit in Example (2) consists of two function words, the first being a grammar word with quite a complex meaning (mostly related to the sentence level) and

a modal verb. Separately these two words have no relation to sentiment but combined together they are often used to show that something can be easily done or improved, which relates to sentiment. Example (3) comprises a combination of a negated modal verb with the first part of a number of words with meaning "setting up; switching to" (e.g. 设置 – *install, set up*; 设成 – *set to (some value)*; 设为自动 – *switch to an automatic mode*). Thus the unit is capable of representing a whole set of similar phrases that describe the inability of a device or a piece of software to perform a certain action, which most probably expresses negative sentiment. This unit has also advantage of being more frequent than any of the full forms. To avoid confusion in what follows I will use the term 'lexical unit' (LU) rather than 'word'. In the context of these experiments the term 'seed' means a LU used as a seed.

(1) 外观 好
appearance good
the appearance is good

(2) 就 可以
already can
OK; has become possible

(3) 不 能 设
not able set...
not able to set...

Lexical Unit Extraction To find lexical units that are candidates for being seeds, the process starts by looking for the longest character sequences that occur in any two zones across all documents in the corpus (using the Longest Common Substring algorithm). Although the process is computationally quite expensive it needs be run only once[5]. The application of this approach to the corpus produced more than 121 thousand lexical units. The list was filtered to exclude non-character symbols (digits, Latin chars, hyphens, but other in-word symbols were preserved). To reduce the list, all lexical units that occurred less than 10 times in the corpus were excluded. The final version of the lexical item list comprised 5492 items.

[5]If efficiency were to be an issue, the corpus could be represented as suffix tree to facilitate faster extraction of lexical units that reoccur.

Approach

To find a seed automatically, I make two assumptions:

1. Tan (2002) showed that in Chinese attitude is often expressed through the negation of vocabulary items with the opposite meaning; for example in Chinese it is more common to say 不好 *not good* than 坏 *bad*. The base system uses this observation to find negative lexical units, while nevertheless starting only from a positive seed. This suggests that it is possible to find candidate seeds themselves by looking for sequences of characters which are used with negation.

2. Preliminary investigations indicated that positive lexical units are more frequent and more widely used together with negation in negative contexts in comparison to negative items in positive. This behaviour can be used as an indicator of a positive lexical item.

These observations served as the basis for identifying seed lexical units: lexical units which occur with negation but more frequently occur without it.

As well as detecting negation[6] I also use adverbials[7] to avoid hypothesizing non-contentful seeds: the characters following the sequence of a negation and an adverbial are in general contentful units, as opposed to parts of words, function words, etc. Consider example (4) where the negation 不 followed by the adverbial 很 modifies the adjective 好. Examples (5) and (6) demonstrate that a negation can be followed by modal or auxiliary verbs which are not good seeds for sentiment analysis. The following sections refer to constructions such as in example (4) as negated adverbial constructions.

(4) 不 很 好
not very good
not very good

(5) 不 能 去
not can go
can't go

(6) 不 是 他
not be he

[6] I use the same list of negation words as before: 不 (bu), 没 有 (meiyou), 不 会 (buhui), 摆 脱 (baituo), 免去 (mianqu), 避免 (bimian)

[7] I use five frequently occurring adverbials: 很 (hen), 非常 (feichang), 太 (tai), 最 (zui), and 比较 (bijiao).

(it) is not him

Method

The first two steps are to identify seed candidates and find suitable positive seeds among them for the given corpus, as specified in Algorithms 3 and 4.

Algorithm 3 Finding Seed Candidates

Require: list of negations $N = n_1...n_n$

Require: list of adverbials $A = a_1...a_m$

Require: list of non-characters $P = p_1...p_q$

return list of candidate seeds W

$W = \epsilon$

for each string s **do**

list of substrings $Sub[0...r] = split(s\ at\ P)$

for each sub in $Sub[0...r]$ **do**

if sub matches $P.*NA.*P$ **then**

list of substrings $w[0...s] = split(sub\ at\ NA)$

for each w in $w[0...s]$ **do**

add w to W

end for

end if

end for

end for

Algorithm 4 Finding Positive Seeds

Require: list of negations $N = n_1...n_n$

Require: list of adverbials $A = a_1...a_m$

Require: list of candidate seeds $W = w_1...w_q$

return list of positive seeds Wp

list of positive seeds $Wp = \epsilon$

for each w **do**

$x = f(NAw)$

$y = f(w\ without\ preceeding\ NA)$

if $x < y$ **then**

add w to Wp

end if

end for

Automatically Found Seeds

Using the approach described above, I extracted seed words from each of the ten topic-based subcorpora of reviews. Table 4.11 shows that for most of the corpora the algorithm found different and (highly domain-salient) seeds.

To see if the automatically extracted seeds perform better than generic seeds, I ran the classifier with one of the generic seed lists and with the extracted seeds. For comparison I chose the *allPOS* seed list as providing the most appropriate comparison because extracted seeds are all positive ones (negative seeds were not extracted).

Table 4.12 presents results of classification using the *allPOS* seed list and seeds extracted automatically from the corpora. In seven out of 10 corpora, extracted seeds performed significantly better in terms of F-measure. The *Printers* corpus performed poorly because only one seed was found in this corpus and it is a rather generic one: 好 (*good*). This lexical unit is also a member of the *allPOS* seed list which contains two further generic positive lexical units. The corpus *Networking* also produced only one seed but this one is a much more domain-specific lexical unit 稳定 (*stable*) which resulted in better precision although recall is 6 percentage points lower. But lower recall is expected when the number of seeds is smaller (one extracted seed vs. three generic ones in the list). However higher precision indicates that the extracted seeds are better descriptors of sentiment in a specific domain.

Corpus	*Seed*		*Corpus*	*Seed*	
Monitors	好	(good)	Video cameras and lenses	清晰	(clear - of sound or image)
	便	(convenient; cheap)		方便	(comfortable)
	清晰	(clear)		实用	(practical)
	直	(straight)		理想	(perfect)
	方便	(comfortable)		爽	(cool)
	满	(fill, fulfill)			
	锐利	(sharp)			
	舒服	(comfortable)			
	爽	(cool)			
Mobile phones	好	(good)	Digital cameras	好	(good)
	支持	(support)		便	(convenient; cheap)
	便	(convenient; cheap)		方便	(comfortable)
	方便	(comfortable)		清晰	(clear - of sound or image)
	清晰	(clear - of sound or image)		专业	(special)
	足	(sufficient)		爽	(cool)
	好用	(easy to use)		满意	(satisfied)
	舒服	(comfortable)		耐用	(durable)
	人性化	(user friendly)		舒服	(comfortable)
	流畅	(smooth and easy)		理想	(perfect)
	清楚	(distinct)		真实	(straight)
	爽	(cool)		稳定	(stable)
	好了	(has become better)		方便了	(has become comfortable)
	耐用	(durable)		客气	(polite)
	方便的	(comfortable)		详细	(detailed)
	满意的	(satisfied)			
	适应	(fit, suit)			
	方便了	(has become comfortable)			
	适用	(applicable)			
	顺手	(handy)			
	科学	(science, scientific)			
Networking	稳定	(stable)	Printers	好	(good)
MP3 players	好	(good)	Computer peripherals	好	(good)
	便	(convenient; cheap)		便	(convenient;cheap)
	方便	(comfortable)		方便	(comfortable)
	实用	(practical)		准	(precise)
	灵敏	(sensitive)		舒服	(comfortable)
	舒服	(comfortable)		习惯	(habitual)
	爽	(cool)		流畅	(smooth and easy)
	方便了	(has become comfortable)		稳定	(stable)
Computer parts	好	(good)	Office equipment	好	(good)
	稳定	(stable)		方便	(comfortable)
				稳定	(stable)
				实用	(practical)

Table 4.11: Seeds automatically identified for each corpus.

Corpus	allPOS			Extracted Seeds			Seeds
	P	R	F	P	R	F	$\sum$
Mobile phones	0.81	0.32	0.46	0.86	0.51	0.64	21
Digital cameras	0.80	0.24	0.38	0.83	0.36	0.50	15
MP3 players	0.84	0.29	0.43	0.83	0.35	0.49	8
Monitors	0.73	0.29	0.41	0.75	0.44	0.55	9
Office equipment	0.86	0.31	0.46	0.87	0.35	0.50	4
Printers	0.80	0.27	0.41	0.76	0.20	0.32	1
Computer peripherals	0.75	0.31	0.44	0.79	0.41	0.54	8
Video cameras and lenses*	0.82	0.30	0.44	0.94	0.29	0.44	5
Networking	0.68	0.25	0.37	0.93	0.19	0.31	1
Computer parts	0.73	0.21	0.33	0.76	0.29	0.42	2
Macroaverage	0.78	0.28	0.41	0.83	0.34	0.47	

Table 4.12: Classification results with the *allPOS* seed list and extracted seeds. Difference between the two sets of results which are statistically NOT significant difference are marked with *.

Corpus	Only Positive			Positive & Negative		
	Acc	AccP	AccN	Acc	AccP	AccN
Mobile phones	0.51	0.66	0.35	0.57	0.65	0.50
Digital cameras	0.35	0.56	0.14	0.45	0.54	0.36
MP3 players	0.34	0.50	0.18	0.41	0.49	0.32
Monitors	0.43	0.68	0.18	0.48	0.67	0.30
Office equipment	0.34	0.50	0.18	0.43	0.49	0.36
Printers	0.20	0.26	0.13	0.26	0.26	0.27
Computer peripherals	0.41	0.58	0.24	0.45	0.56	0.33
Video cameras and lenses	0.28	0.49	0.07	0.37	0.49	0.25
Networking	0.18	0.32	0.04	0.27	0.33	0.22
Computer parts	0.28	0.48	0.09	0.37	0.47	0.27
Macroaverage	0.33	0.50	0.16	0.41	0.50	0.32

Table 4.13: Classification results with only positive extracted seeds vs the same seeds augmented with generic negative seeds. *Acc* is overall accuracy, *AccP* is accuracy per class of positive documents, *AccN* is accuracy per class of negative documents. For all topics the differences between the two sets of results are statistically significant.

Negative Seeds

The biggest disadvantage of the technique for automatically finding the seeds is that it does not find negative seeds. But as was shown in previous experiments, negative seeds significantly improve performance of the classifier. Negative seeds combined with positive ones not only improve precision and recall (Table 4.9) but also produce a much more balanced classification. Table 4.13 shows that adding generic negative seeds to extracted seeds produces less skewed results; Table 4.14 shows that overall classification accuracies also improve significantly. The performance of the combination of extracted seeds with generic negatives is better (in terms of F-measure) than the performance of classifier with the *all* seed list for seven out of ten corpora, with only two performing worse (*Printers* and *Networking*) and one performing equally (without a statistically significant difference).

Corpus	Only Positive			Pos & Neg			*all* Seed List		
	P	R	F	P	R	F	P	R	F
Mobile phones	0.86	0.51	0.64	0.89	0.57	0.70	0.85	0.41	0.55
Digital cameras	0.82	0.35	0.49	0.88	0.45	0.60	0.86	0.35	0.50
MP3 players	0.83	0.34	0.48	0.87	0.41	0.55	0.88	0.35	0.50
Monitors	0.74	0.43	0.55	0.80	0.48	0.60	0.79	0.34	0.47
Office equipment	0.86	0.34	0.49	0.90	0.43	0.58	0.89	0.39	0.55
Printers	0.76	0.20	0.31	0.84	0.26	0.40	0.86	0.33	0.48
Computer peripherals	0.79	0.41	0.54	0.83	0.45	0.58	0.79	0.35	0.48
Video cameras and lenses	0.93	0.28	0.43	0.94	0.37	0.53*	0.87	0.36	0.51*
Networking	0.92	0.18	0.30	0.93	0.27	0.42	0.75	0.31	0.44
Computer parts	0.76	0.28	0.41	0.82	0.37	0.51	0.81	0.30	0.44
Macroaverage	0.83	0.33	0.46	0.87	0.41	0.55	0.84	0.35	0.49

Table 4.14: Classification results with only positive extracted seeds (*Only Positive*), the same seeds augmented with generic negative seeds (*Pos & Neg*) and *all* seed list (*all Seed List*). P is precision, R is recall, F is F-measure. For all corpora the differences between the results for all corpora are statistically significant except for those marked with *.

4.2.3 Iterative Approach

In the context of real-world applications, most of the results presented in the previous experiments would probably be acceptable in terms of precision; however they are very low in recall, especially compared to the vocabulary-based classifier described earlier. This means that the seeds on their own are not sufficient and the classifier needs more lexical units with appropriately calculated scores to perform better.

One way of extracting more lexical units from the corpus is to run the classifier iteratively. Each new iteration uses a subset consisting of classified documents from the corpus for extracting new lexical units and calculating their scores. The newly found set of lexical units with scores assigned is then used for creating a new set of classified documents that form a new subset for the next iteration (see Algorithm 5).

Iteration Stopping Criterion

An iterative approach requires a way to control the number of iterations. I used a goal driven stopping criterion which means that iterations should stop once the goal is achieved.

Algorithm 5 Iterative sentiment classifier

Require: list of negations, sentiment seed lexicon W, corpus of documents

loop

Run the classifier

Extract a subcorpus

Find new lexical units and add them to W

For each w in W adjust the score of w

end loop

As well as accuracy of classification, the goal of sentiment classification is to classify as many documents as possible. But preliminary experiments showed that after a certain number of iterations the number of classified documents starts to change periodically, going up and down. So the idea the stopping criterion is based on is quite simple: stop the iterations when the number of classified documents stops increasing. This idea is supported by a strong correlation between the F-measure and the number of classified documents: for all the topics the correlation ranges between 0.81 and 0.99. The actual rule that stops the iterations adds some flexibility to be able to overcome local maxima: the system is allowed to make a few more iterations to find if there is another iteration with even better results. The number of the 'lookahead' iterations is set to the number of iterations the system used for finding the current maximum but not less than 3. If after at least three iterations the number of classified documents is smaller or remains unchanged, the system stops the iterations and uses the classification results of the best iteration (in which the number of classified documents was maximal).

Table 4.15 presents the results of classification of documents from two topics (*Mobile phones* and *Monitors*) for eight iterations[8]. The stopping criterion described above would have stopped at iteration 4 for *Mobile phones* and iteration 5 for *Monitors* at the point where the number of documents that were not classified by the classifier stopped going down. Although in both cases these points would not be the best in terms of F-measure, the performance is still rather high (the second best in both cases). Given that an unsupervised classifier does not have access to a gold standard and thus cannot evaluate each iteration in terms of precision or recall, the iteration control described above seems to perform well in being able to stop at one of the best iterations.

[8]The correlations between the number of classified documents and the F-measure for these two topics are 0.99 and 1.00 respectively.

	Mobile phones				Monitors			
Iter	P	R	F	C	P	R	F	C
1	0.86	0.41	0.56	1209	0.79	0.34	0.48	386
2	0.87	0.80	0.83	189	0.83	0.76	0.79	57
3	0.86	0.80	0.83	157	0.85	0.80	0.82	34
4	*0.85*	*0.80*	*0.82*	156	0.83	0.79	0.81	33
5	0.85	0.79	0.82	158	*0.83*	*0.80*	*0.81*	29
6	0.85	0.79	0.81	163	0.83	0.79	0.81	29
7	0.84	0.79	0.81	157	0.83	0.80	0.81	31
8	0.84	0.78	0.81	162	0.83	0.80	0.82	30

Table 4.15: Results of sentiment classification of 10 iterations with seed list *all* applied to two topics *Mobile phones* and *Monitors*. *Iter* is the number of iterations, P is precision, R is recall, F is F-measure; C is the number of documents that were NOT classified.

Classification Results: Over the whole Corpus

The next set of experiments tests the performance of the same set of seeds as presented in Section 4.2.1 on the whole corpus but using the iterative technique. After a number of iterations the classifier produced good results with positive seeds (see Table 4.16) compared to the non-iterative classifier (Table 4.9). The most significant progress was made in overall accuracy of classification, but the results are also less skewed. The best results were for group of seeds *all*. All the other positive seeds also performed quite well regardless of how many seeds there were in the list. In contrast, all negative seeds performed poorly, barely improving over the naïve baseline. The reason for this is a very unbalanced classification: almost all documents get tagged as positive, which results in near-baseline performance. The skew towards positive classification (which is not expected from the negative seeds) is the result of the skew towards negative classifications during the first iteration: the extracted subcorpus contains many more negative documents compared to positive ones, which affects extraction of lexical units and score calculation for them. The system extracts too many negative lexical units with very low scores (because there are too many documents classified as negative) and several high frequency supposedly positive lexical units (with high scores as the number of positive documents is low). This leads to a skew towards positive classification in subsequent iterations. This suggests that such classifications should be avoided when the iteration control chooses the best iteration and

Seed list name	P	R	F	Acc	AccP	AccN	Iterations
good	0.79	0.72	0.75	0.72	0.77	0.68	9
very_good	0.77	0.71	0.74	0.71	0.74	0.68	12
comfortable	0.78	0.72	0.75	0.72	0.73	0.71	5
bad	0.53	0.50	0.52	0.50	0.94	0.06	2
too_bad	0.51	0.49	0.50	0.49	0.98	0.01	2
poor	0.54	0.50	0.52	0.50	0.93	0.07	2
allPOS	0.79	0.72	0.75	0.72	0.77	0.68	10
allNEG	0.55	0.51	0.53	0.51	0.93	0.09	2
all	0.85	0.78	0.81	0.78	0.81	0.75	3

Table 4.16: Results of sentiment classification after iterations. P is precision, R is recall, F is F-measure; Acc is accuracy, $AccP$ is accuracy of the positive class and $AccN$ is accuracy of the negative class.

that the iteration control should be extended with a skew-control rule.

Skew Control The motivation behind skew control is to prevent a classifier from producing highly skewed classifications. To do so, the skew control needs some approximate 'idea' of what a balanced classification is. Such a 'gold standard' can be provided by the first (seed-only) iteration:

$$G = \frac{min(C_i, C_j)}{max(C_i, C_j)} \tag{4.1}$$

where G is the 'gold standard' for the balance, and C_i and C_j are the number of classified documents of each class (either positive or negative). During the iterative classification procedure, if the classification skew deviates from G then the iterations are stopped.

This means that the skew control uses the balance of the initial classification to compare with all subsequent classifications. However, if the system uses the exact value of the 'gold standard' (which is likely not to be perfect), then good classifications which are slightly different in balance will be regarded as skewed and thus ignored. For this reason the system in fact does not use a strict comparison but instead use a 'window' of ±50%. For example, if the initial iteration classified 100 positive documents and 100 negative documents, then the 'gold standard' would be 1; an acceptable balance should be at least 0.5 (a smaller class can be half of the size of the bigger one). So if the next classification finds 100

positive and 200 negative documents, then this classification is regarded as acceptable.

Of course, since this system relies on the performance of the initial seeds, unreliable seeds should be excluded. So because all the negative seeds in the first iteration produced highly skewed classifications, all of these seed lists have to be excluded from further experiments.

Extracted Lexical Units On completion of iterations the systems extracted different sets of lexical units for the various groups of seeds (see Table 4.17). Apart from the expected lexical units that describe qualities of products, the sets contain many noun-based items whose relation to sentiment is not obvious. For example, 5英寸 (*5 inch*) was mostly used in phrases like '2.5英寸屏幕' (*2.5 inch*). The phrase 在游戏 (*in the game*) was often used in positive reviews of a computer mouse by a gamer. Another group of positive lexical units denote product features which were regarded as a positive attribute by users: 倍光学变焦 (x *optical zoom*) was a good feature of a digital camera and 卫星箱 (*satellite speakers*) are a good addition to a sound system.

Of course these lexical units can also be used in a negative context, but in the corpus they were used mostly as indicators of positive sentiment, which was quite difficult to predict. This illustrates how difficult it is to predict all sentiment-related lexical units in any given domain, and suggests that it would be impossible to build an universal sentiment dictionary.

Table 4.18 shows examples of negative lexical units found. Apart from quality-related lexical units (e.g. 量太差 *(quality) is too poor*), as discussed above, there are a lot of items that are related to time: they were used to describe short-lived faulty devices. The latter ones are also difficult to predict. For example, 待机时间短 (*short standby time*) is used to describe mobile phones whose batteries do not last long in standby mode, and lexical units like 维修站 (*repair shop*), 保修期 (*warranty term*) 去维修 (*went to repair*) are often used in reviews of devices that developed a fault and had to go in for repair.

Classification Results: Per-Topic

The experiments presented below test the performance of the iterative approach over the topics taken separately. The experiments also test and compare the performance of different seeds: generic vs extracted, and with negative seeds vs without them.

Since the extracted seeds do not have negative lexical units in them, the only matching generic seed list is *allPos* which also does not include negative seeds. For eight out of the ten topics the classification results are significantly different. The extracted seeds per-

Seed list name	Top 10 words in positive list
good	操作简 (control is (easy)), 做工精 (carefully made), 倍光学 (x optics) 具有 ((it) has), 质不错 (quality is rather good) 倍光学变焦 (x optical zoom), 操作简单 (easy control), 5英寸 (5 inch) 效果出 (output), 功能丰富 (rich in features)
very_good	提供了 (supplied, provided), 操作简 (control is (easy)), 做工精 (carefully made) 倍光学 (optics), 倍光学变焦 (x optical zoom), 具有 ((it) has) 质不错 (quality is rather good), 5英寸(5 inch), 操作简单 (easy control), DVD+
comfortable	倍光学 (optics), 倍光学变焦 (x optical zoom), 效果出 (output) 提供了 (supplied, provided), 常出色 ([extrem]emly outstanding) 非常出 (extremely out[standing]), dpi, 感舒适 (feel comfortable) 效果出色 (outstanding output), 做工精细 (carefully made)
bad	CRT, 这款音箱 (these speakers), 游戏中 (during the game) 显示器的 ((of) monitor), 显像管 (CRT) 低音炮 (subwoofer), 何失真 ((some) distortion), 在游戏 (in the game) 几何失真(geometric distortion), 卫星箱 (satellite speakers)
too_bad	采用了 (used), 具有 ((it) has), 色彩还原 (colour reduction), 观设计 (visual design) 外观设计 (design), 提供了 (supplied, provided), 采用 ((it) uses) 光学变焦 (optical zoom), 功能强 (reach in features), 操作简 (control is (easy))
poor	采用了 (used), 具有 ((it) has), 观设计 (visual design), 外观设计 (design) 提供了 (supplied, provided), 光学变焦 (optical zoom), 功能强 (rich in features) 操作简 (control is (easy)), 采用((it) uses), 能强大 (rich in features)
allPOS	做工精 (carefully made), 倍光学 (x optics), 倍光学变焦 (x optical zoom) 质不错 (quality is rather good), 5英寸 (5 inch), 操作简单 (easy control) 效果出 (output), 音质不错 (good sound quality) 功能齐全 (full of features), 具有 ((it) has)
allNEG	倍光学 (x optics), 倍光学变焦 (x optical zoom) 功能齐全 (full of features), 感舒适(feel comfortable) 效果出 (output), 非常出 (extremely out[standing]), 的调节 (control (of)), dpi 常出色 ([extr]emly outstanding), 效果出色 (outstanding output)
all	做工精 (carefully made), 倍光学 (x optics), 倍光学变焦 (x optical zoom) 提供了 (supplied, provided), 质不错 (quality is rather good), 5英寸 (5 inch) 操作简单 (easy control), 功能齐全 (full of features), 具有 ((it) has), 效果出 (output)

Table 4.17: Top 10 positive lexical units found on completion of iterations.

Seed list name	Top 10 words in negative list
good	换了一 (changed one), 出问题 (problems appeared), 不能用 (not usable) 不好用 (faulty), 机时间短 (short time), 不耐用 (unusable) 就坏了(got broken soon), 个月就 (just (numeral) month(s)) 了不到 ((used) less than), 待机时间短 (short standby time)
very_good	以为是 ((wrongly) thought that), 经常出 (often happens), 个月就 ((a) month and) 不要买 (don't buy), 就不能 (it's become impossible to) 换了一 (changed one), 用了不到 (used for less than), 不能用 (not usable) 就坏了 (broke down soon), 我刚买 (I've just bought)
comfortable	不好用 (faulty), 出问题 (problems appeared), 不能用 (not usable) 机时间短 (short time), 待机时间短 (short standby time) 打电话 (to phone), 个月就 (just (numeral) month(s)), 死机 (device died) 换了一 (changed one), 时间太 (the time is too)
bad	以为是 ((wrongly) thought that), 个月就 ((one) month and) 就发现 (found out soon), 知道怎 (know how) 维修站 (repair shop), 保修期 (warranty term), 买了不 (bought not (long ago)) 换了一 (changed one), 就没电 (no power), 去维修 (went to repair)
too_bad	用了不到 (used for less than), 用了不(used for (less than)) 买了不 (bought not (long ago)), 了不到((used) less than) 就发现 (found out soon), 维修站 (repair shop), 我买了一 (I bought one) 就没电 (no power), 就没电了 (power's gone), 保修期 (warranty term)
poor	的电话 ((some) phone), 66, 电话簿 ((dial) a phone) N7, 时间太 (time is too), 短消息 (SMS) 00条, 的短信 ((some) SMS), 手机上 (on mobile phone)
allPOS	了不到 ((used) less than), 出问题 (problems appeared), 不能用 (not usable) 就坏了 (broke down soon), 不好用 (faulty), 个月就 ((one) month and) 换了一 (changed one), 待机时间短 (short standby time), 机时间短 (short time)
allNEG	打电话 (dial a phone), 个月就 ((one) month and), 量太差 ((quality) is too poor) 不能用 (not usable), 换了一 (changed one), 太差了 (too bad) 出问题 (problems appeared), 就不能 (became impossible soon) 用了不到 (used for less than), 质量太 (quality is too)
all	了不到 ((used) less than), 出问题 (problems appeared), 不能用 (not usable) 就坏了 (broke down soon), 不好用 (faulty), 个月就 ((one) month and) 时间太 (time is too), 机时间短 (short time) 换了一 (changed one), 待机时间短 (short standby time)

Table 4.18: Top 10 negative lexical units found on completion of iterations.

Corpus	allPOS			Extracted		
	P	R	F	P	R	F
Mobile phones	0.82	0.76	0.79	0.86	0.80	0.83
Digital cameras	0.74	0.66	0.70	0.74	0.67	0.70
MP3 players*	0.76	0.71	0.74	0.75	0.70	0.72
Monitors	0.81	0.77	0.79	0.81	0.78	0.79
Office equipment*	0.79	0.71	0.75	0.80	0.73	0.76
Printers	0.80	0.73	0.76	0.75	0.68	0.72
Computer peripherals	0.61	0.56	0.58	0.61	0.57	0.59
Video cameras and lenses	0.67	0.63	0.65	0.50	0.47	0.48
Networking	0.68	0.25	0.37	0.81	0.72	0.76
Computer parts	0.55	0.51	0.53	0.50	0.46	0.48
Macroaverage	0.72	0.63	0.67	0.71	0.66	0.68

Table 4.19: Classification results with *allPos* seed list and only positive extracted seeds *Extracted*. *P* is precision, *R* is recall, *F* is F-measure. Differences between the two sets of results are statistically significant except for the corpora marked with *.

formed better in terms of recall but precision was almost the same as that of the generic seeds (see Table 4.19). In two topics (*Computer parts* and *Video*) the extracted seeds failed to perform better than the naïve baseline, and the generic seeds failed to do so in topics *Networking* and *Computer parts*. The result of classification of the topic *Networking* illustrates the importance of a seed's domain-relevance: only one extracted seed outperformed three generic ones. However in the topics *Video* and *Computer parts* generic seeds performed better. The performance of the extracted seeds was most probably compromised by a small size of these two topic corpora (only 361 and 308 documents respectively, see Table 4.5) and that the collections combined reviews of related but nevertheless different items (video cameras and lenses; CD-drives and motherboards). But on a big topic such as *Mobile phones* the extracted seeds performed much better, mostly due to a large number of extracted seeds (21 lexical units, see Table 4.11).

Another comparable pair of seed lists are the *all* seed list and the extracted seeds combined with generic negative seeds (the same as the ones in *all*). Negative seeds helped both of the seed lists to increase performance, but the generic seeds gained more compared to the extracted ones (see Table 4.20). Although slightly better in recall, the generic seeds

Corpus	all			ExtractedNeg		
	P	R	F	P	R	F
Mobile phones	0.85	0.80	0.82	0.89	0.83	0.86
Digital cameras	0.82	0.74	0.77	0.81	0.73	0.77
MP3 players	0.81	0.75	0.78	0.79	0.73	0.76
Monitors	0.83	0.80	0.81	0.83	0.80	0.81
Office equipment	0.81	0.75	0.78	0.83	0.76	0.80
Printers	0.82	0.75	0.78	0.82	0.75	0.78
Computer peripherals	0.82	0.78	0.80	0.84	0.79	0.81
Video cameras and lenses	0.77	0.73	0.75	0.70	0.66	0.68
Networking	0.75	0.31	0.44	0.83	0.72	0.77
Computer parts	0.67	0.63	0.65	0.67	0.63	0.65
Macroaverage	0.80	0.70	0.74	0.80	0.74	0.77

Table 4.20: Classification results with generic seeds (*all*) and extracted seeds combined with generic negative seeds (*ExtractedNeg*). P is precision, R is recall, F is F-measure.

are similar in terms of precision. Again, similarly to the previous experiments, on a large document collection (*Mobile phones*) the extracted seeds performed much better than the generic ones. Both classifiers performed well (much higher than the naïve baseline) on all of the topics, which confirms the importance of negative seeds.

4.2.4 Discussion

The experiments presented above showed that although features (vocabulary) adjusted to the domain produce better sentiment classification, a vocabulary-based approach has limited ability to adapt to a domain: it is not possible to foresee all possible sentiment-bearing lexical units in all possible domains. An alternative approach, based on using seeds for classification proved to be effective when used with multiple iterations. All seeds consisting of both positive and negative lexical units managed to bootstrap a better vocabulary from the corpus than the extracted ones. The biggest disadvantage of the latter is absence of negative lexical units. However, augmented with generic negative seeds, the extracted seeds performed quite well in terms of recall, especially on large document collections. Generally, iterations allow the bootstrapping of a domain-related sentiment vocabulary which in some cases performs better than the generic sentiment vocabulary

Corpus	Seeds			Vocabulary		
	P	R	F	P	R	F
Mobile phones	0.89	0.83	0.86	0.86	0.82	0.84
Digital cameras	0.81	0.73	0.77	0.85	0.77	0.80
MP3 players	0.79	0.73	0.76	0.88	0.84	0.86
Monitors	0.83	0.80	0.81	0.86	0.82	0.84
Office equipment*	0.83	0.76	0.80	0.88	0.81	0.84
Printers	0.82	0.75	0.78	0.86	0.79	0.82
Computer peripherals*	0.84	0.79	0.81	0.89	0.86	0.87
Video cameras and lenses*	0.70	0.66	0.68	0.86	0.81	0.84
Networking	0.83	0.72	0.77	0.88	0.81	0.84
Computer parts*	0.67	0.63	0.65	0.79	0.72	0.76
Macroaverage	0.80	0.74	0.77	0.87	0.80	0.83

Table 4.21: Classification results with the seed list *all* (*Seeds*) and the vocabulary-based classifier (*Vocabulary*) after a number of iterations. P is precision, R is recall, F is F-measure. Differences between the two sets of results are statistically significant except for the corpora marked with *.

(Table 4.6): on the larger collections (upper half of that table) the seeds performed at a similar level or even better than the vocabulary-based classifier. But smaller collections (lower half of the table) make it difficult for the seeds to extract a good enough vocabulary to perform better than the predefined generic one. Although large number of seeds can produce better results, the NTU sentiment dictionary taken as the seed list performed only six percentage points better (F-measure) than the extracted seed list (including negative seeds): see Table 4.21. Note that on the largest topic *Mobile phones* the extracted seeds performed significantly better. This means that much smaller (and much easier to produce) resources might perform almost as well (or even better) as ones comprising thousands of items.

4.3 Performance Improvements

The previous section showed that starting from a large sentiment vocabulary is not the only way to obtain effective sentiment-bearing lexical units. Instead, seeds combined with

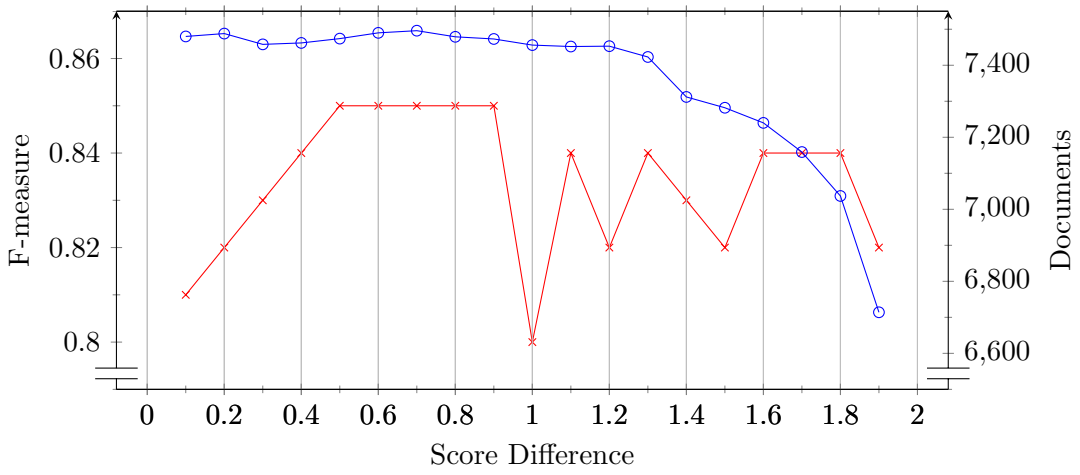

Figure 4.1: Classification results with the seed list *all* with the score difference technique. —×— is F-measure; —○— is the number of documents classified.

iteration produce results that are close to those derived from a predefined generic sentiment dictionary. This section investigates ways to improve performance of the seed-based classifiers. It describes a technique based on filtering scores of the lexical units to extract only highly discriminative ones (Section 4.3.1), a technique that helps to rank documents by reliability of their classification (Section 4.3.2), and experiments with extending the unsupervised classifier with supervised techniques (Section 4.3.3)

4.3.1 Score Difference

Ideally, a technique should not use lexical units that are not very discriminative, i.e. the difference between their positive and negative scores is low. To measure this difference I used the following formula:

$$D = \frac{|S_i - S_j|}{(S_i + S_j)/2} \tag{4.2}$$

If the difference $D > threshold$, then the difference between the two scores is taken to be significant and the lexical item associated with the scores is included in the final list. Different threshold values were tested: from 0.1 to 1.9 with steps of 0.1. The results of the corpus classification with the seed list *all* are presented in Figure 4.1.

The experimental results show that at different values of the score difference threshold, the classifier produces rather different results, with precision ranging from 0.85 at 0.1 to 0.91 at 1.9 and recall starting as low as 0.78 and reaching 0.82 at 0.6. The line —×— represents F-measure and shows a steady increase of performance between score filtering threshold values 0.1 and 0.5, after which it reaches a plateau stretching from 0.5 to 0.9.

At 1.0 the performance drops and becomes highly unstable ranging from 0.80 to 0.84. The increase in performance represented by the first half of the line (before 1.0) could be expected because the score difference approach is aimed at increasing precision with increasing values of the threshold as it filters out more lexical units whose positive and negative scores do not differ enough. But the higher the threshold, the more lexical units it filters out, as a consequence affecting recall. This is reflected in the second part of the graph which shows that lower recall and higher precision lead to an overall drop in F-measure. The latter, being a harmonic average of recall and precision, may be a good indicator of a classifier's performance that helps find the right balance between the two parameters. But how can an unsupervised system decide what threshold to choose and what result is the best if it cannot use a gold standard to calculate F-measure? To find the best result I used the same approach as used for the iteration control: the best result is taken to be the result with the highest number of classified documents. The highest number of classified documents (7496) is at threshold value 0.7 (Figure 4.1) which coincides with the plateau where F-measure is 0.85. This is the highest value and is significantly better than the results of the same seed list without the score difference threshold (Table 4.15) and the classification results of the vocabulary-based classifier even after adjusting scores of its items (Table 4.4). Another advantage of this threshold value is that it is situated within the more stable zone of the plateau far enough from the unstable zone of values > 0.9 thus ensuring more robust performance.

Score Difference on Topics

The next set of experiments tests the applicability of the score filtering approach to the classification of the reviews grouped in different topics. The classifiers used two seed lists that proved to be the most effective: *all* with generic seeds and the extracted seeds with generic negative lexical units. The classifiers used the same approach as described above for identifying of the best classification.

The macroaverage results in Table 4.22 show improvements in all aspects of performance for both seed lists as compared with the results of the same classifiers without score filtering (Table 4.20). However, only three topics performed significantly differently with list *all*, of which one topic (*MP3 players*) performed worse loosing two percentage points in precision. But the gains are much more substantial. The topic *Mobile phones* added 7 points in precision and 6 points in recall. A very large increase was also shown by *Networking* which increased performance by 26 percentage points (F-measure: from 0.44 to

	all				ExtractedNeg			
Corpus	ScDiff	P	R	F	ScDiff	P	R	F
Mobile phones	1.2	0.92	0.86	0.89	0.8	0.90	0.85	0.87
Digital cameras	0.3	0.80	0.72	0.76*	0.2	0.78	0.71	0.75*
MP3 players	0.4	0.79	0.75	0.77	0.6	0.82	0.77	0.79
Monitors	0.2	0.83	0.80	0.82*	0.1	0.83	0.80	0.81
Office equipment	0.0	0.81	0.75	0.78*	0.2	0.84	0.77	0.80*
Printers	0.1	0.82	0.76	0.79*	1.0	0.86	0.79	0.82*
Computer peripherals	0.1	0.82	0.79	0.81*	1.0	0.83	0.79	0.81
Video cameras and lenses	0.4	0.76	0.72	0.74*	0.0	0.70	0.66	0.68*
Networking	0.6	0.75	0.66	0.70	0.5	0.82	0.73	0.77*
Computer parts	0.1	0.67	0.63	0.65*	0.1	0.67	0.64	0.66
Macroaverage	0.3	0.80	0.74	0.77	0.5	0.81	0.75	0.78

Table 4.22: Classification results with seed list *all* and automatically extracted seeds with generic negative lexical units (*ExtractedNeg*). *ScDiff* is the score difference threshold value, P is precision, R is recall, F is F-measure. Differences between the results and the results in Table 4.20 are statistically significant except for those marked with *.

0.70), mostly because recall gained 36 points. The automatically extracted seeds together with generic negative lexical units performed better, with five topics showing significantly different results. Topics which performed better than those produced without the score difference technique were: *Mobile phones* increased by two percentage points, *Digital cameras* added three percentage points and *Computer parts* added one percentage point. It seems that the extracted seeds gained more with the score difference technique. Despite not all topics increasing their performance (and one topic even performing worse) the score difference technique appeared to be a useful way of improving the performance of the unsupervised sentiment classifier.

4.3.2 Zone Difference

The method described above utilized the difference of alternative scores of individual lexical units, but a similar approach can be applied to a whole document as its sentiment orientation is computed by comparing the number of zones of alternative orientation.

As described in Section 3.2.1 a document is assigned the sentiment of the majority

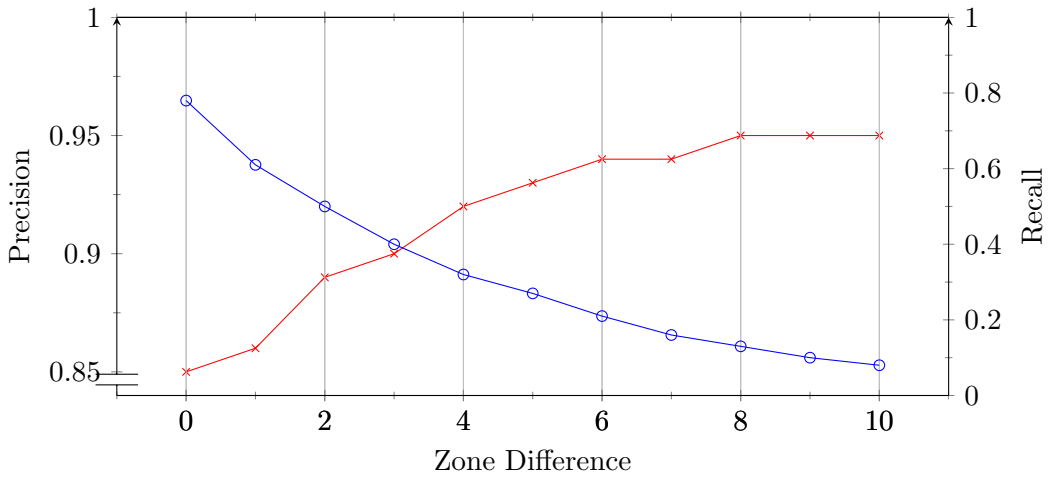

Figure 4.2: Classification results with the seed list *all* and with the zone difference technique. ×— is precision; ○— is recall.

of its zones. But a small prevalence of zones of one sentiment over zones of the other (for example 5 POS and 4 NEG) can be just a matter of chance or a result of poor zone classification. A larger difference (5 POS and 2 NEG) might indicate a more accurate document classification. Thus it might be useful to use a threshold value (T) for the zone difference as in equation 4.3,

$$S_d = \begin{cases} positive, if POS - NEG > T \\ negative, if NEG - POS > T \\ nil, otherwise \end{cases} \tag{4.3}$$

where POS is the number of positive zones and NEG is the number of negative zones.

However this method may adversely affect the performance in the initial iteration: since the number of initial seeds is low, the number of classified zones is also low and in most cases the difference between zones of alternative sentiment would be 1. In this circumstance the method described may dramatically reduce the size of the extracted subcorpus and thus adversely affect performance. To overcome the problem, the method is applied only to the final iteration of the classifier.

The graph in Figure 4.2 shows the classification results with the seed list *all* and zone difference threshold ranging from 1 to 10. There are two graphs: one represents precision and another one is recall. The precision at the final iteration of the classifier with threshold = 0 is 0.85 and steadily grows to 0.95 at threshold = 8. However recall drops from 0.78 to 0.13. Obviously even a high precision classification with such a small recall is of no use, but such control over precision might be very useful in practice in an opinionated

information retrieval system for ranking results according to their reliability. This means that the results with the highest precision might be treated as the most reliable ones and presented to the user before the others (e.g. on the first page(s)). The rest of the results could be presented according to their precision: results with higher precision would be placed before those with lower precision.

Zone Difference on Topics

Figures 4.3 and 4.4 show the classification performance on each of the topics with the zone difference technique applied to the classifiers based on the seed list *all* (the former) and on extracted seeds augmented with generic negative lexical units (the latter). In both cases on the majority of the topics the classifiers performed as expected: precision is increased by 10-15 percentage points as the zone difference threshold was increased. The only exception was *Office equipment* for the seed list *all* and *Video cameras and lenses* for the extracted seeds. The difference in distribution pattern is most probably connected with the average size of a document in a corpus: the longer a document is, the more zones it contains and the greater variability of the zone difference value it has.

Figures 4.3 and 4.4 show the classification results of the whole corpus with different sentiment zone values without ranking individual documents according to their zone difference value. To model the distribution of search results on different pages, simulating what might happen in an opinionated information retrieval system, I ranked all the classified documents by their zone difference value and split them into 'pages' each consisting of 100 documents. The results for the seed list *all* are presented in Figure 4.5, and for the extracted seeds the results are in Figure 4.6. Both Figures show that for all the topics the first 'page(s)' have much higher precision than the later ones.

4.3.3 Using Supervised Techniques to Extend Unsupervised Classifier

The previous experiments showed that the unsupervised classifier is capable of extracting collections of classified documents which can be used as a basis for subsequent iterations of the classifier. This suggests that the same approach may be used to extract classified corpora for training supervised classifiers. The feature sets of the latter could be lexical units extracted by the unsupervised classifier in the final iteration as well as the items of the NTU sentiment dictionary. Another option for the feature set is the whole set of the lexical units of the collection.

In the experiments below I chose two supervised techniques: Naïve Bayes multinomial

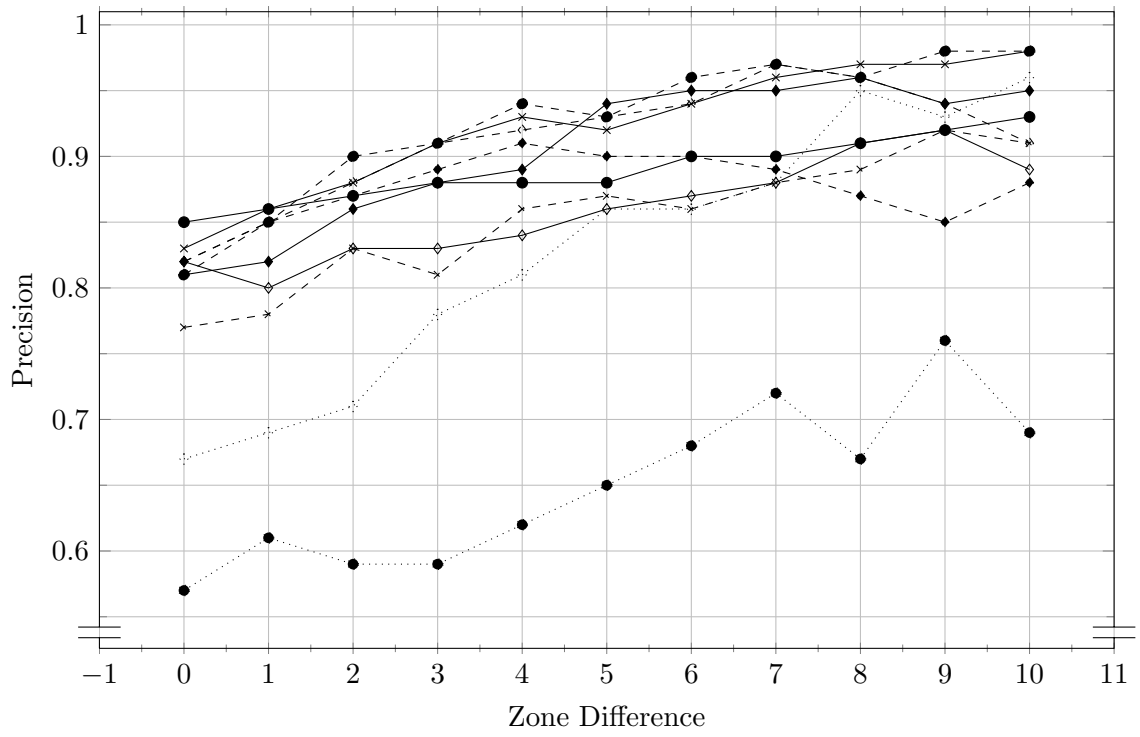

Figure 4.3: Classification results with seed list *all* with the zone distance technique. —●— is *Mobiles*; —◇— is *Digital Cameras*; —♦— is *MP3 Players*; —×— is *Monitors*; - -●- - is *Office Equipment*; - -◇- - is *Printers*; - -♦- - is *Computer peripherals*; - -×- - is *Video*; ···●··· is *Networking*; ···◇··· is *Computer parts* .

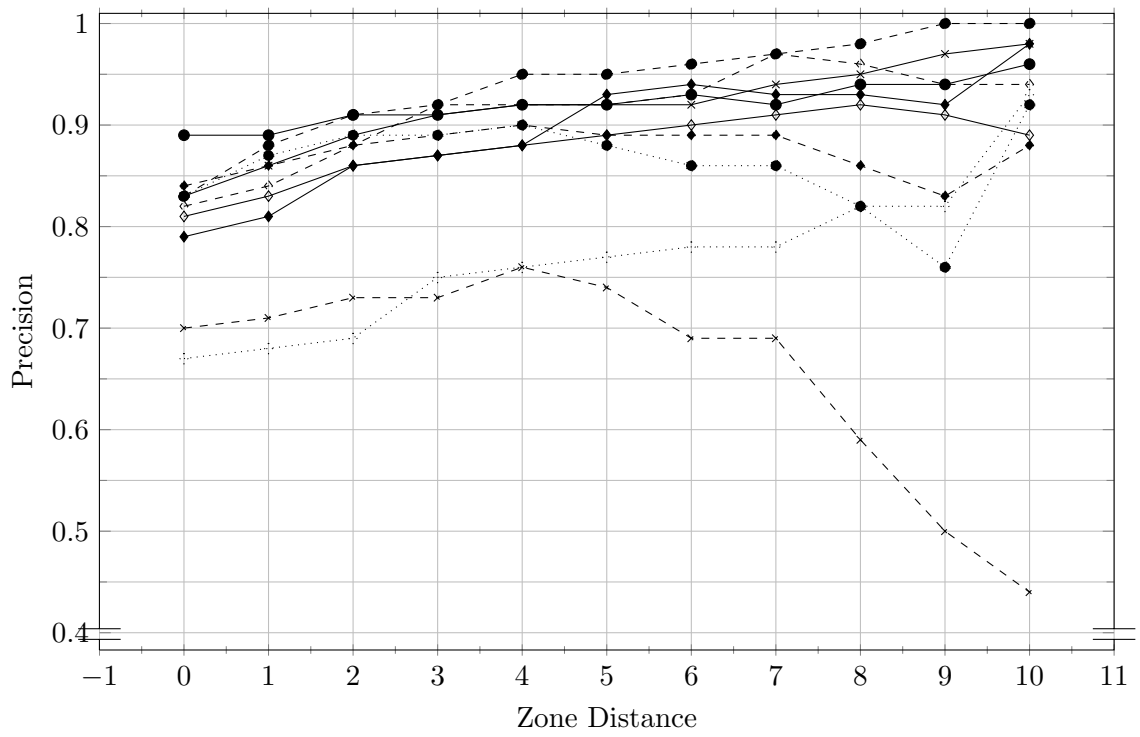

Figure 4.4: Classification results with extracted seeds with the zone distance technique. —●— is *Mobiles*; —◇— is *Digital Cameras*; —♦— is *MP3 Players*; —×— is *Monitors*; - -●- - is *Office Equipment*; - -◇- - is *Printers*; - -♦- - is *Computer peripherals*; - -×- - is *Video*; ···●··· is *Networking*; ···◇··· is *Computer parts* .

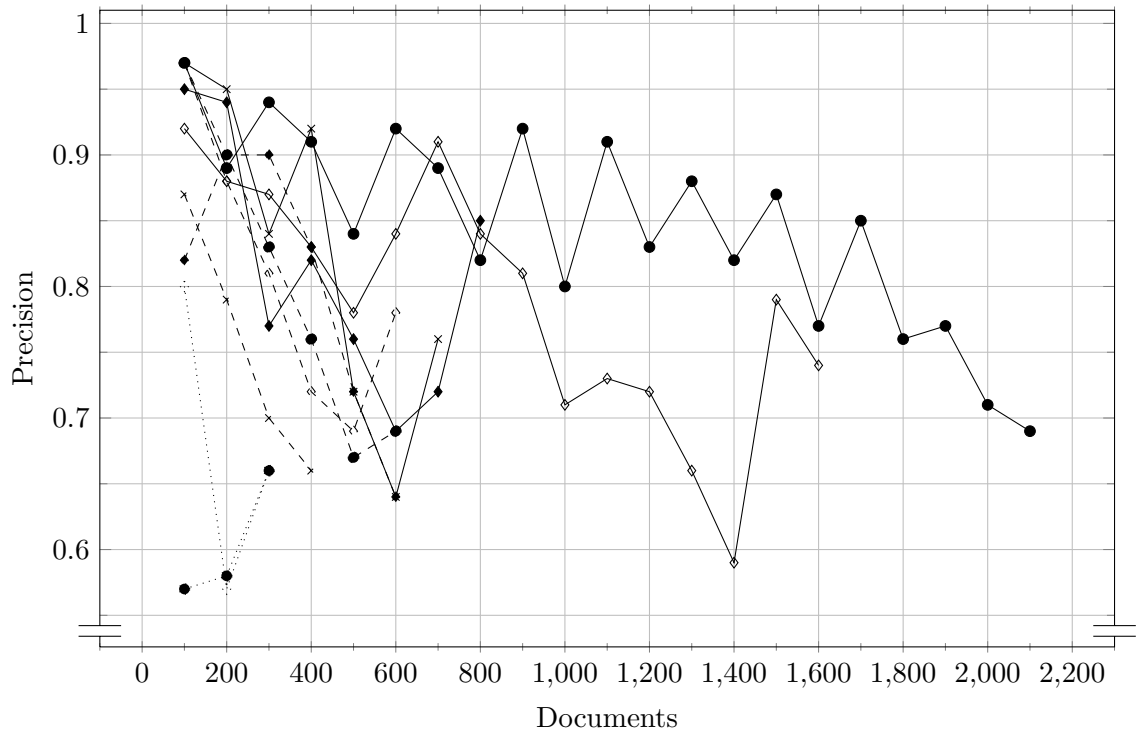

Figure 4.5: Information retrieval simulation results with seed list *all* with the zone distance technique. —●— is *Mobiles*; —◇— is *Digital Cameras*; —♦— is *MP3 Players*; —×— is *Monitors*; - -●- - is *Office Equipment*; - -▵- - is *Printers*; - -♦- - is *Computer peripherals*; - -×- - is *Video*; ···●··· is *Networking*; ····· is *Computer parts* .

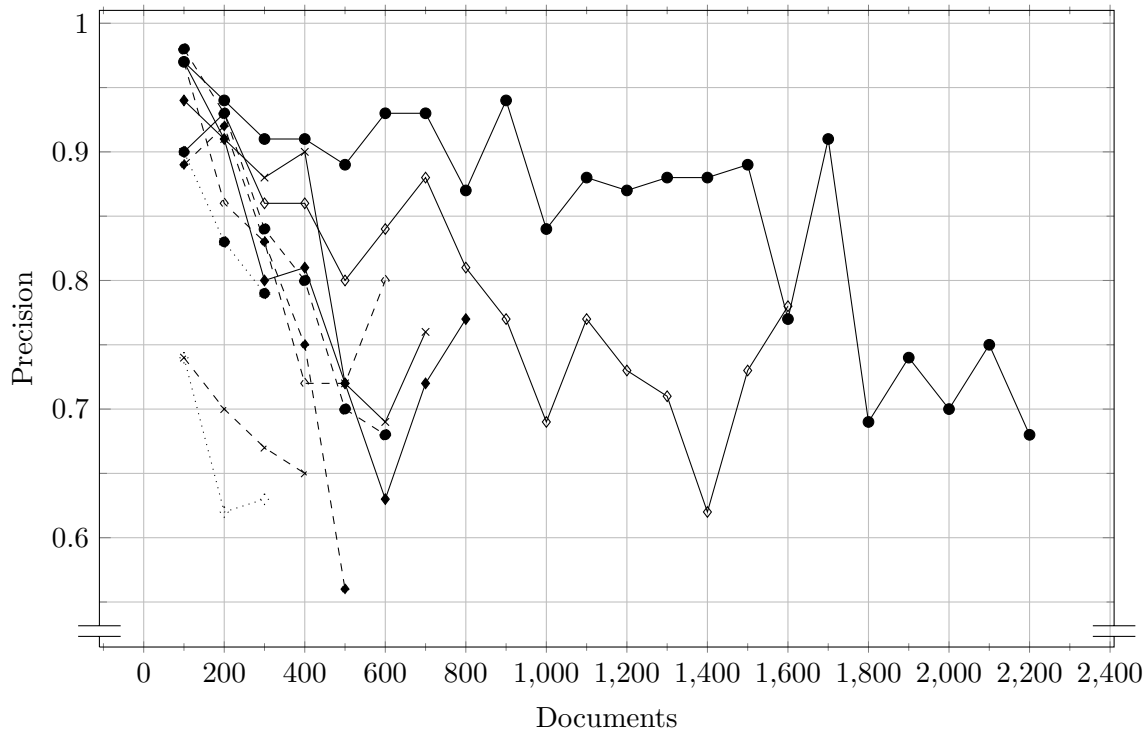

Figure 4.6: Information retrieval simulation results with extracted seed list with the zone distance technique. —●— is *Mobiles*; —◇— is *Digital Cameras*; —♦— is *MP3 Players*; —×— is *Monitors*; - -●- - is *Office Equipment*; - -▵- - is *Printers*; - -♦- - is *Computer peripherals*; - -×- - is *Video*; ···●··· is *Networking*; ···▵··· is *Computer parts* .

Corpus	SentiVoc		All LU		Extracted LU	
Classifiers	NBm	SVM	NBm	SVM	NBm	SVM
Mobile phones	0.87^{ae}	0.85^{ae}	0.90^{v}	0.90^{v}	0.90^{v}	0.90^{v}
Digital cameras	0.83^{ae}	0.82^{ae}	0.85^{v}	0.84^{v}	0.86^{v}	0.85^{v}
MP3 players	0.83^{ae}	0.82^{ae}	0.88^{ve}	0.87^{ve}	0.88^{va}	0.88^{va}
Monitors	0.86^{ae}	0.84^{e}	0.86^{v}	0.85	0.86^{v}	0.87^{v}
Office equipment	0.83	0.81	0.84^{e}	0.84^{e}	0.85^{a}	0.84^{a}
Printers	0.81	0.81	0.83	0.86	0.85	0.85
Computer peripherals	0.86	0.81	0.84	0.81	0.84	0.82
Video cameras and lenses	0.84	0.84	0.86^{e}	0.83	0.88^{a}	0.85
Networking	0.77	0.75	0.87	0.83	0.85	0.81
Computer parts	0.77	0.78	0.80^{e}	0.80^{e}	0.76^{a}	0.77^{a}
Macroaverage	0.83	0.81	0.85	0.84	0.85	0.84

Table 4.23: Supervised classifiers with the three feature sets (10-fold cross validation, weighted average accuracies for two classes); for each corpus, statistically significant differences indicated with respect to the NTU sentiment dictionary (v), all lexical units (a), and the extracted lexical units (e).

and Support vector machine[9]. Both are widely used in sentiment classification research and are therefore reasonable representative techniques.

Testing Features

Before extending the unsupervised classifier with the supervised machine learning techniques, it is necessary to identify which of the possible feature sets is the most effective. To test the performance of the feature sets I used a supervised technique with 10-fold cross-validation. There are three feature sets to be tested: the NTU sentiment dictionary, lexical items that were extracted by the unsupervised classifier during the final iteration and, finally, all lexical units of the corpus.

Table 4.23 presents the results of classification with the three different feature sets. The extracted lexical units perform very similarly to all lexical units of the corpus, but are much better than the NTU sentiment dictionary especially on a larger data sets (first four lines of the Table), where the differences are statistically significant.

[9]I used WEKA 3.4.11 (http://www.cs.waikato.ac.nz/~ml/weka)

Corpus	SentiVoc		All LU		Extracted LU	
Classifiers	10f	Extr	10f	Extr	10f	Extr
Mobile phones	0.87	0.84	0.90	0.86*	0.90	0.88
Digital cameras	0.83	0.83*	0.85	0.76	0.86	0.76
MP3 players	0.83	0.84	0.88	0.79*	0.88	0.80*
Monitors	0.86	0.84	0.86	0.84*	0.86	0.83*
Office equipment	0.83	0.82*	0.84	0.80*	0.85	0.80*
Printers	0.81	0.82*	0.83	0.83*	0.85	0.82*
Computer peripherals	0.86	0.83	0.84	0.82*	0.84	0.81*
Video cameras and lenses	0.84	0.77*	0.86	0.71	0.88	0.68
Networking	0.77	0.61	0.87	0.80	0.85	0.79
Computer parts	0.77	0.65	0.80	0.65	0.76	0.64
Macroaverage	0.83	0.79	0.85	0.79	0.85	0.78

Table 4.24: The NBm classifier with the three feature sets, 10-fold cross-validation (10f) vs trained on the extracted corpus (Extr), weighted average accuracies for two classes; differences between the three sets of results that are NOT statistically significant are marked with *.

Testing the Extracted Training Corpus

The impact of the extracted training corpus was measured by comparing supervised classifiers trained on the extracted training corpus with the performance of the same classifiers trained on 90% of the test corpus using 10-fold cross-validation.

Table 4.24 presents the results of the NBm classifier using the three different feature sets: the NTU sentiment dictionary, all lexical units of the corpora and lexical units extracted by the supervised classifier. The first two columns show the results of classification using the NTU sentiment dictionary. For 4 out of the 10 topics the classifier produced similar results, with statistically different results for six; in all but one case the results were inferior compared to the supervised technique. The classifiers that used all lexical units showed significantly decreased performance in half of the topics when trained on the extracted collection. Similar results were produced with extracted lexical units.

Table 4.25 summarises the results of the SVM classifier with the same three feature sets described above. The results follow the same pattern as with the NBm classifier: although in almost half of the topics the classification results did not differ significantly

the overall performance was worse compared to the fully supervised technique that used the same corpus for training.

Corpus	SentiVoc		All LU		Extracted LU	
Classifiers	10f	Extr	10f	Extr	10f	Extr
Mobile phones	0.85	0.88*	0.90	0.88*	0.90	0.88
Digital cameras	0.82	0.82*	0.84	0.82	0.85	0.76
MP3 players	0.82	0.81*	0.87	0.80	0.88	0.79
Monitors	0.84	0.84*	0.85	0.82*	0.87	0.82*
Office equipment	0.81	0.79*	0.84	0.82*	0.84	0.81*
Printers	0.81	0.81	0.86	0.82*	0.85	0.84
Computer peripherals	0.81	0.81*	0.81	0.80*	0.82	0.82*
Video cameras and lenses	0.84	0.72	0.83	0.74*	0.85	0.69
Networking	0.75	0.61	0.83	0.78	0.81	0.78
Computer parts	0.78	0.60	0.80	0.67*	0.77	0.66*
Macroaverage	0.81	0.77	0.84	0.80	0.84	0.79

Table 4.25: The SVM classifier with the three feature sets, 10-fold cross-validation (10f) vs trained on the extracted corpus (Extr), weighted average accuracies for two classes; differences between the three sets of results are NOT statistically significant are marked with *.

With both machine learning techniques the extracted collection used as training corpus for the machine learning classifiers decreased their performance. This outcome could be expected as none of the unsupervised classifiers produced a 100% accurate collection of classified documents.

4.3.4 Comparison of Supervised and Unsupervised Classifiers

The last comparison, presented in Table 4.26, is between two classifiers that use data obtained from the unsupervised classifier and one completely supervised classifier trained on the corpus and using the NTU sentiment dictionary as the feature set. The first two columns present accuracy of NBm and SVM classifiers evaluated using 10-fold cross validation. The first of the two other classifiers is a combination of the unsupervised classifier and the machine learning techniques that used classified documents as the training corpus and all lexical units as the feature set. The last classifier is similar to the previous one

but using a different feature set: the lexical units extracted by the unsupervised classifier. The results presented in the table suggest that the unsupervised classifiers perform better or equally on a large collection (*Mobile phones*), but they cannot match the supervised classifier on smaller collections, mostly because the unsupervised classifiers rely on bootstrapping their sentiment vocabulary which requires a larger amount of data. Another reason for poor performance on smaller topics is that some of them are not very topically homogeneous as they consist of reviews of different (albeit related) items (Table 4.5)

Corpus	Supervised		All LU		Extracted LU	
Classifiers	NBm	SVM	NBm	SVM	NBm	SVM
Mobile phones	0.87	0.85	0.86	0.85	0.88	0.88
Digital cameras	0.83	0.82	0.76*	0.76	0.76*	0.76
MP3 players	0.83	0.82	0.79*	0.78	0.80	0.79
Monitors	0.86	0.84	0.84*	0.85*	0.83*	0.82
Office equipment	0.83	0.81	0.80*	0.81	0.80*	0.81
Printers	0.81	0.81	0.83*	0.85	0.82*	0.84
Computer peripherals	0.86	0.81	0.82	0.82*	0.81*	0.82*
Video cameras and lenses	0.84	0.84	0.71	0.71*	0.68	0.69
Networking	0.77	0.75	0.80	0.81	0.79	0.78
Computer parts	0.77	0.78	0.65	0.65*	0.64	0.66*
Macroaverage	0.83	0.81	0.79	0.79	0.79	0.79

Table 4.26: Supervised classifiers compared with two classifiers trained on data extracted by the unsupervised classifier (weighted average accuracies for two classes); differences between the three sets of results that are NOT statistically significant are marked with *.

Although unable to outperform the supervised classifier, the combination of the unsupervised classifier with the machine learning techniques increased performance by 3 percentage points on average (in terms of recall, which equals accuracy in binary classification on a balanced corpus), with higher gains on larger collections (see Table 4.27)

4.4 Discussion

The techniques presented in this section raised performance of the unsupervised classifier by almost five percentage points (compare macroaverage recall of the *Seeds* in Table

	Unsupervised			Unsupervised + NBm		
Corpus	P	R	F	P	R	F
Mobile phones	0.92	0.86	0.89	0.89	0.89	0.89
Digital cameras	0.79	0.71	0.75	0.77	0.77	0.77
MP3 players	0.81	0.76	0.78	0.80	0.80	0.80
Monitors	0.84	0.80	0.82	0.82	0.82	0.82
Office equipment	0.83	0.76	0.80	0.81	0.81	0.81
Printers	0.85	0.76	0.79	0.82	0.81	0.81
Computer peripherals	0.83	0.78	0.80	0.81	0.81	0.81
Video cameras and lenses	0.70	0.65	0.69	0.69	0.68	0.68
Networking	0.82	0.73	0.77	0.81	0.80	0.80
Computer parts	0.67	0.64	0.66	0.65	0.64	0.64
Macroaverage	0.81	0.75	0.78	0.79	0.78	0.78

Table 4.27: Classification results with extracted seeds (with negative lexical units and score difference) and the same classifier with added NBm classifier. *ScDiff* is the score difference threshold value, P is precision, R is recall, F is F-measure.

4.21 and macroaverage accuracy[10] of the unsupervised classifiers in Table 4.26). It is only slightly (2-4 percentage points) inferior to a completely supervised classifier using a specialised sentiment vocabulary as feature set. However, the unsupervised approach is less effective on smaller document collections due to difficulty in bootstrapping vocabulary from limited or not very homogeneous data. But since automated means of data processing are usually aimed at processing large datasets, it is more important that the unsupervised classifier is able to classify larger collections with better or similar accuracy compared to supervised techniques.

4.5 Conclusion

The most important result of the experiments presented in this Chapter is that an unsupervised approach to sentiment classification can produce results very similar to a supervised approach. This opens up a possibility to avoid expensive development of training corpora

[10]For a corpus with all its elements belonging to a class that is to be classified, accuracy and recall values are the same.

and sentiment vocabularies for sentiment classification.

A number of other conclusions can also be drawn from the experiments in this Chapter.

- Current techniques for sentiment classification are sensitive to domain; this problem can be addressed automatically by ensuring that vocabulary items which are more discriminative in a given domain are assigned higher scores and contribute more to the overall performance of the classifier. But this approach has rather limited ability of adjustment in the cases where the amount of text available is limited.

- Even more improvement can be seen from extracting a vocabulary from a corpus using a small set of generic seeds. Automatic extraction of sentiment-related vocabulary from corpus helps find lexical units which have domain-dependent sentiment and would be difficult to predict, such as time-related expressions, product features which are regarded as good or bad by users, lexical units used to describe situations related to performance or quality (e.g. visits to a repair shop).

- Positive seeds have higher frequencies and can be used on their own with negation compensating for the absence of negative seeds. Negative seeds are quite sparse so do not produce good results on their own. The highest performance was achieved by a list comprising both positive and negative seeds.

- Positive seeds can be extracted automatically from corpus. They may improve the performance of a classifier, but their performance is compromised by the absence of negative seeds. The combination of automatically found positive seeds and generic negative seeds increases the performance of the classifier and outperforms generic seeds.

- An iterative technique can further improve performance of the classifier and eliminate the difference between generic and extracted seeds. Maintaining a count of classified documents is an effective way of determining when the iteration should finish.

- Score filtering, a technique that eliminates lexical units which are not discriminative enough, can further improve precision.

- The zone difference technique is an effective way of ranking results by their precision. This could be very useful for sentiment analysis in IR as means of presenting more reliable results (with higher precision of classification) before less reliable ones.

- A fully unsupervised technique based on automatic extraction of seeds performed well on large corpora, and much better than a naïve baseline (F-measure 0.20 – 0.30

over the baseline for 9 out of 10 test corpora). The technique performed better than supervised classifiers, except on smaller collections.

- Further improvements for large corpora can be achieved by applying of supervised techniques to the data extracted. The unsupervised classifiers managed to produce a good feature set for supervised classifiers. Although the extracted subcorpus used for training was not of the highest quality, the better feature set compensated for it: overall performance of the unsupervised classifier augmented with a supervised machine learning technique was only 1 - 3 percentage points behind in terms of macroaverage results and equal or better on bigger collections.

- All the unsupervised techniques that were applied depend heavily on the amount of data: the larger the corpus is the better the results are. This affects their success on small datasets, but means they can be useful for processing large amounts of data.

Chapter 5

Multilingual Sentiment Classification[1]

The previous Chapter described a unsupervised sentiment classifier, as well as a number of additional techniques that help improve performance of the classifier: including iterative expansion of initial seed vocabulary, score-based filtering of vocabulary items, thresholding of sentiment zone score, and integration with supervised machine learning. The results achieved are reasonably close to the performance of supervised classifiers. However, one of the main motivations for the research in this thesis is creation of sentiment analysis techniques which could be applicable not only to different domains of the same language, but also to different languages.

This chapter further investigates the iterative sentiment classification technique described in Chapter 4 and tested there on Chinese, by applying it to data in two other languages, English and Russian. For testing purposes I used three different corpora: two in English and one in Russian. Section 5.1 describes the data used for the experiments and discusses language-specific means of expressing sentiment that may influence automatic sentiment classification. The next Section (5.2) presents experiments with supervised classifiers that set an upper bound and expose specific aspects of the multi-lingual and multi-domain data used in this Chapter. Application of the unsupervised technique to different languages is presented in Section 5.3. Section 5.4 draws conclusions regarding the results obtained.

[1]The experiments and part of the discussion in this section were presented in a condensed form at the 1st Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (Zagibalov et al. (2010))

5.1 Data

There are a number of publicly available sentiment-annotated corpora, such as MPQA (Wiebe et al., 2005), and Pang and Lee's Movie Review corpus (Pang and Lee, 2004). However, most of these corpora consist only of English text. There are some corpora designed for cross-lingual evaluations, but these seem not to be publicly available (for example the NTCIR MOAT corpora of English, Japanese and Chinese (Seki et al., 2008).

For this study, I[2] have designed and built comparable corpora of book reviews in English and Russian, which are publicly available[3]. The corpora are comparable in terms of domain, style and size. The Russian corpus is probably the first sentiment-annotated resource in that language.

This section, as well as describing the corpora and quantifying their various relevant aspects, also analyses some important language-specific and domain-specific issues that would be likely to impact on automatic sentiment processing.

5.1.1 Language-Specific Issues

The data used in this chapter belongs to the English and Russian languages. These two languages are substantially different from the Chinese language used for the experiments in the previous Chapter. The most obvious (and visible) difference is the presence of formal word delimiters such as the space which is used to separate graphical[4] words in writing. However the languages have a number of features that need to be addressed carefully in their processing. Both of the languages have more complex morphology than Chinese.

Russian

Morphology In Chinese most forms are analytical, in English there are a small number of morphological processes, but there are many in the Russian language. The latter has a relatively complex morphology that comprises gender, case and number forms of adjectives and nouns as well as inclination and tenses, and aspect forms of verbs. For example, the adjective хороший (*good*) has the following forms:

1. хороший – masculine, singular

2. хорошая – feminine, singular

[2]These corpora were developed with the help of Katerina Belyatskaya.

[3]The corpora are available for download from http://www.informatics.sussex.ac.uk/users/tz21/.

[4]Other notions of word, such as semantic word or phonetic word are not affected, but since they are not directly connected to this research, they are not discussed here.

3. хорошее – neuter, singular

4. хорошие – plural (same for all genders).

Each of these forms may be used with different cases having various endings (see Table 5.1)

Cases	m. sing	f. sing	n. sing	plural
Nominative	хороший	хорошая	хорошее	хорошие
Genitive	хорошего	хорошую	хорошего	хороших
Dative	хорошему	хорошей	хорошему	хорошим
Accusative	хорошего / хороший	хорошую	хорошее	хороших / хорошие
Ablative	хорошим	хорошей	хорошим	хорошими
Prepositional	хорошем	хорошей	хорошем	хороших

Table 5.1: Case forms of Russian adjectives

Also there are comparative and superlative forms of the adjective: лучше and наилучший / самый лучший (the latter is an analytical superlative form). The word can also be used in a short form: хорош. The number of forms (16 unique forms) suggests that unsupervised Russian language processing could be difficult especially if the processing is to be language-independent and not relying on the language-specific tools (for example morphological analysers).

English

The English verb has morphological means of expressing grammatical tense and aspect, and noun morphology covers singular and plural. Probably the most important part of speech for sentiment analysis – adjectives – also have comparative and superlative forms which sometimes are formed irregularly (e.g. *good – better – best* and *bad – worse – worst*). Still, the variation of grammatical forms in English is not as complex as in Russian.

Discussion

Unlike the Chinese language, English and Russian feature graphical words separated by space. However, some words have a complex structure so may require lexical processing (morphological parsing, stemming or lemmatization) before a document can be further processed. Otherwise, keeping all the word forms intact, one may have the problem of

data sparseness as numerous word forms would 'hide' a single word even in a large corpus, making grammatical features (expressed by affixes) more significant compared with the meaning of the word. However, lexical processing of this type is necessarily language-dependent, making a system very much more resistant to multilingual use. The significant difference in the word structure of the languages used in the experiments complicates language processing if using unsupervised techniques. An even bigger challenge is multilingual processing that assumes using as few language-specific tools as possible. These issues constitute a strong test for the concept of the lexical unit as the basic unit for multilingual sentiment classification (Section 4.2.2).

5.1.2 Book Review Corpora

Corpora Content

The English and Russian book review corpora consist of reader reviews of science fiction and fantasy books by popular authors. The reviews were written in 2007, so the language used is fairly current.

The Russian corpus consists of reviews of Russian translations of books by popular science-fiction and fantasy authors, such as S. King, S. Lem, J.K. Rowling, T. Pratchett, R. Salvatore, J.R.R. Tolkien as well as by Russian authors of the genre such as S. Lukianenko, M. Semenova and others. The reviews were published on the website www.fenzin.org.

The English corpus comprises reviews of books by the same authors, if available. If some of the authors were not reviewed on the site or did not have enough reviews, they were substituted with other writers of the same genre. As a result, the English corpus contains reviews of books such as: S. Erickson (*Guardians of the Moon*, *Memories of Ice*), S. King (*Christine*, *Duma Key*, *Gerald's Game*, *Different Season* and others), S. Lem (*Solaris*, *Star Diaris of Iyon Tichy*, *The Cybriad*), A. Rise (*Interview with the Vampire*, *The Tale of the Body Thief* and others), J.K. Rowling (*Harry Potter*), J.R.R. Tolkien (*The Hobbit*, *The Lord of the Rings*, *The Silmarillion*), S. Lukyanenko (*The Night Watch*, *The Day Watch*, *The Twilight Watch*, *The Last Watch*), and a few others. The reviews were published on the website www.amazon.co.uk.

Although both of the sites from which the reviews were collected feature review-ranking systems (e.g. one to ten stars), many reviewers did not use the system or did not use it properly. For this reason all of the reviews were read through and hand-annotated. There were a lot of reoccurring short reviews like: Хорошо (*Good*); Интересная книга (*Interesting book*); Супер! (*Superb!*); Нудятина!! (*Boring!!*); Ниже среднего (*Below*

	Mean tokens POS	Mean tokens NEG	Total types POS	Total types NEG
English	58	58	7349	8014
Russian	30	38	9290	12309

Table 5.2: Overall quantitative measures of the English and Russian corpora.

average); *Awesome!*; *Amazing!*; *The best book I've ever read!*; *Boring*, and so on. These reviews were added to the corpus only once. Also both sites had a number of documents which did not have any direct relation to book reviewing, such as advertisements, announcements and off-topic postings. Such texts were excluded as irrelevant. The documents that were included in the corpora were not edited or altered in any other way.

Each review was manually annotated as 'POS' if positive sentiment prevails or 'NEG' if the review is mostly negative. Each corpus consists of 1500 reviews, half of which are positive and half negative. The annotation is simple and encodes only the overall sentiment of a review, for example:

```
[TEXT = POS]
Hope you love this book as much as I did. I thought
it was wonderful!
[/TEXT]
```

The English reviews contain a mean of 58 words (the mean length for positive and negative reviews being almost the same). Positive Russian reviews have a mean length of only 30 words; negative reviews are slightly longer, at 38 words (see Table 5.2). It is not possible to compare these figures directly between the languages as they have different grammar structures which makes English more 'wordy' as it has function words (articles, auxiliary verbs) which are almost absent in Russian.

As noted above, the Russian language, being a synthetic language, has many forms of the same word. This results in a large number of unique words (word-forms): the corpus contains 18,913 unique words, with 9,290 words (43%) in the positive part and 12,309 (57%) in the negative. The English corpus in the whole corpus, 7,349 (48%) in its positive part, and 8,014 (52%) in its negative part. These figures also suggest that Russian reviewers used a richer vocabulary for expressing negative sentiments than English readers. Further evidence of different attitudes to expressing alternative sentiments in

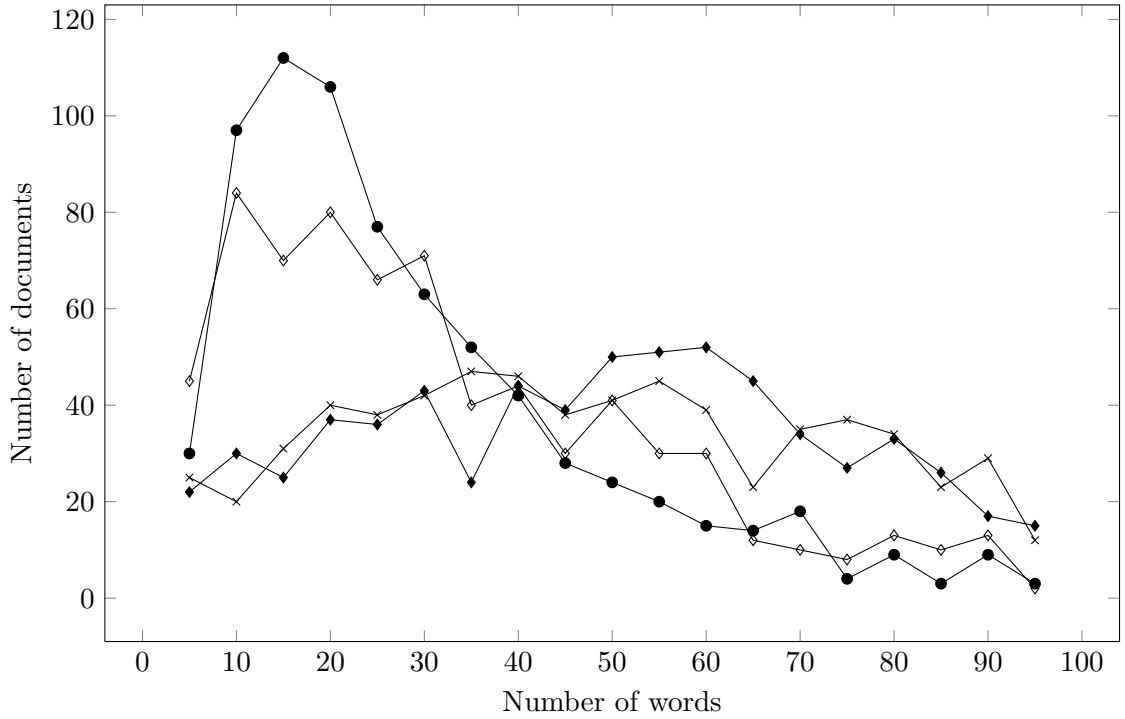

Figure 5.1: Distribution of documents according to the number of words in them. Russian book reviews: —●— is positive reviews, —◇— is negative reviews; English corpus: —♦— is positive reviews, —×— is negative reviews

Russian is the different length of positive and negative reviews. Figure 5.1 shows that in the Russian corpus, there are many short reviews ($<$ 50 words) with the mode at 15 words for positive reviews and at 10 words for negative reviews. Apart from the language-specific differences mentioned above that partly account for the smaller number of words in Russian documents, there is a clear difference with respect to English reviews in terms of the length distribution. The English reviews are more evenly spread featuring more or less an equal number of documents of different length (mostly in the range between 15 and 75). The prevalence of short reviews in the Russian corpus, compounded by the diversity of morphological variation, may lead to data sparseness that could adversely affect the performance of unsupervised classifiers.

Ways of Expressing Sentiments

To better understand the difference between the English and the Russian corpora, I have investigated the means used to express opinion and how this may impact on automatic sentiment classification[5].

[5]All the numerical data presented below comes from manual counting and is not represented in the corpus annotation.

	Syntactic	Lexical				Phonetic
		Verb	Adj	Noun	Other	
Positive	432	312	708	225	325	12
Negative	367	389	652	238	407	16
Total	799	701	1360	463	732	28

Table 5.3: Ways of expressing sentiment in the English Book Review Corpus (numbers of documents).

	Syntactic	Lexical				Phonetic
		Verb	Adj	Noun	Other	
Positive	417	492	648	374	367	27
Negative	475	578	567	334	394	43
Total	892	1070	1215	708	761	70

Table 5.4: Ways of expressing sentiment in the Russian Book Review Corpus (numbers of documents).

Sentiment can be expressed at different levels in a language: from lexical and phonetic levels up to the discourse level.

This range is reflected in the corpora (see Tables 5.3 and 5.4). As the Tables show, the authors of reviews in the two languages express sentiment in slightly different ways. In English they make heavy use of adjectives to express sentiment (this class of words is used to express sentiment in a third of all documents). In contrast, in Russian they use verbs as often as adjectives to express sentiment (both of these classes are used in about quarter of all reviews) and make more use of nouns (expressing sentiment in 15% of all documents compared to 11% in English). The Russian corpus also demonstrates a tendency to combine different ways of expressing sentiments in a document: the total number of uses of different ways in the English corpus is 4,083 compared to 4,716 in Russian, which means that given an equal number of reviews for each language, Russian reviews tend to have more different ways of expressing sentiment per document.

Lexical Level

Adjectives Adjectives are the most frequent way of expressing opinions in both corpora, closely followed by verbs in the Russian corpus. 1,215 Russian reviews use adjectives to express sentiment and 1,070 reviews use verbs. In the English corpus there are 1,360 reviews that use adjectives, but only 701 use verbs to express opinion.

Apart from adjectives, which are recognised as the main means of expressing evaluation, other parts of speech are also often used in this function, most notably verbs and nouns. The English reviews also feature adverbials, and both languages also use interjections.

Verbs Akimova and Maslennikova (1987) observed that opinions delivered by means of verbs are more expressive compared to opinions expressed in other ways. This is explained by the fact that a verb's denotation is a situation and the semantic structure of the verb reflects linguistically relevant elements of the situation described by the verb. Verbs of appraisal not only name an action, but also express a subject's attitude to an event or fact. Consider the following examples:

(1) I truly loved this book, and I KNOW you will, too!

(2) понравилось, научная фантастика в хорошем исполнении
I liked it, it's science fiction in a very good implementation

The English verbs *loved* and *liked* describe an entire situation which is completed by the time of reporting it. This means that a subsequent shift in sentiment polarity is all but impossible:

(3) *I truly loved this book, but it turned out to be boring.

However, adjectives usually describe only attributes of certain members of a situation leaving a significant amount of context aside:

(4) The story is pretty good but it stretches on and on.

In the example above a positive sentiment towards the story is shifted to negative. A verb is less usual in such a context:

(5) (?) I liked the story but it stretches on and on.

Nouns Nouns can both identify an object and provide some evaluation of it. But nouns are less frequently used for expressing opinion compared to verbs. Nonetheless in the Russian corpus, nouns were used more than in the English corpus. There are 708 Russian reviews that have opinions expressed by nouns, however, only 463 English reviews made use of a noun to describe opinion. The most frequent such nouns used in Russian reviews are чудо (*miracle*), классика (*classics*), шедевр (*masterpiece*), гений (*genius*), прелесть (*delight*), бред (*nonsense*), мура (*raspberry*), жвачка (*mind-numbing stuff*), ерунда (*bugger*).

Phonetic Level Although the corpora consist of written text and do not have any speech-related mark-up, some of the review authors used speech-related methods to express sentiment, for example:

(6) This was a sloooow, frail story

(7) A BIG FAT ZEEROOOOOOOOOOOO for M.A

(8) i have to say is a good booooooooooooooooooook!

(9) Ну что сказать. . . чепуха. . . ЧЕ-ПУ-ХА.
What should I say... boloney... BO-LO-NEY

(10) Ндаааааа.....такую муть давно не видел
Weeeeelll..... I haven't seen such a stinkaroo for long

(11) абалденная книшкаааааа!!!!!!!!!!!!!!!))) оч давно её люблю))
jaw-droppin' booooooook!!!!!!!!!!!!!!!))) been lovin' it for long

(12) Мозг ломиться от этого несоответствия... и получает ооочень большой кайф!!!
My brain is bursting because of this inconstancy... and it enjoys it veeery much!!!

(13) Читать ВСЕЕЕЕЕЕЕЕЕМ
Read, EVERYBOOOOODY

Another way to express opinion in Russian is based on the use of a sub-culture language, Padonky. This sociolect has distinctive phonetic and lexical features that are distant from 'standard' Russian (both official and colloquial). For example, a phrase usually used to express a negative attitude to an author about his book:

(14) Аффтор, выпей ЙАДУ
(lit) Autor, drink some POIZON

Padonky is close to some variants of slang (corresponding in English to expressions such as *u woz*, *c u soon* etc.), however it is more consistent and is used quite often on the Web.

Sentence Level Sentence-level means of expressing sentiment (mostly exclamatory clauses, imperatives or rhetorical questions) is slightly more frequent in the Russian corpus than in the English: 892 and 799 respectively. The distribution of positive and negative sentiments realised at the sentence level is opposite in the two corpora: syntactic means are used more frequently in negative reviews in Russian but they are more frequent in positive reviews in English.

One particularly common sentiment-relevant sentence-level phenomenon is the rhetorical question. This is a question only in form, since it usually expresses a statement. For example:

(15) И откуда столько восторженных отзывов? Коробит от крутости главных героев
Why are there so many appreciative reviews? The 'coolness' of the main characters makes me sick

(16) Что же такого пил/принимал/нюхал автор, чтобы написать такое?

What did the author drink / eat / sniff to write stuff like that?

Some 'borderline' cases such as the following are also used to express sentiment:

(17) Интересно, кто-нибудь дотянул хотя бы до середины? Лично я - нет.
I wonder if anyone managed to get to the middle? I failed.

Considering imperatives, the review author is telling their audience 'what to do', which is often to read a book or to avoid doing so.

(18) Run away! Run away!

(19) Pick up any Pratchett novel with Rincewind and re-read it rather than buying this one

(20) Читать однозначно.
Definitely should read.

(21) Читать !!!!!!!!!! ВСЕМ
Read!!!!!!!! EVERYONE

Another way of expressing sentiment through syntactic structure is by means of exclamatory clauses, which are, by their very nature, affective. This type of sentence is widely represented in both corpora.

(22) It certainly leaves you hungering for more!

(23) Buy at your peril. Mine's in the bin!

Discourse Level Some means of sentiment expression are quite complex and difficult to analyse automatically:

(24) И это автор вычислителя и леммингов? ... НЕ ВЕРЮ! Садись,
so this author calculator and lemmings? ... (DO)NOT BELIEVE! sit,
Громов, два.
gromov, two.
So is this the author of The Calculator and of The Lemmings? ... Can't believe it!
Sit down, Gromov, mark 'D'!

This short review of a new book by Gromov, the author of the popular novels *The Calculator* and *The Lemmings*, consists of a rhetorical question, an exclamatory phrase and an imperative. All of these means of expression are difficult to process. Even the explicit appraisal expressed by utilising a secondary school grade system is problematic as it requires specialised real-word knowledge. Otherwise the numeral 'two'[6] has nothing to do with appraisal per se.

The example below also features an imperative sentence is used to express negative sentiment. This review also lacks any explicit sentiment markers. The negative appraisal is expressed by the verbs 'stab' and 'burn' which only in this context show negative attitude.

(25) Stab the book and burn it!

5.1.3 Issues that may Affect Automatic Processing

One of the features of web content not mentioned above is a high level of **mistakes and typos**. Sometimes authors do not observe the standard rules on purpose (for example using sociolects, as outlined above). For example, in the corpora 52% of all documents contain spelling mistakes in words that have sentiment-related meaning. The English corpus is less affected as authors do not often change spelling on purpose and use contractions that have already become conventional (e.g. *wanna*, *gonna*, and *u*). However, the number of spelling mistakes is still high: 48% of reviews contain mistakes in sentiment-bearing words. The proportion of misspelled words in the Russian corpus is higher, at 58%.

Of course, a spelling error is not always fatal for automatic sentiment classification of a document, since reviews usually have more sentiment indicators than just one word. However, as many as 8% of the reviews in both corpora have all of their sentiment-bearing words misspelled. This would pose severe difficulties for automatic sentiment classification.

Another obstacle that makes sentiment analysis difficult is **topic shift**, in which the majority of a review describes a different object and compares it to the item under review. The negative review below is an example of this:

[6] Russian schools use a 5-grade marking system, with 5 as the highest mark. Thus 2 can be thought of as equivalent to 'D'.

(26) Дочитала с трудом. Ничего интересного с точки зрения информации. Образец интеллектуального детектива – романы У.Эко. И читать приятно, и глубина философии, и в историческом плане познавательно. А в эстетическом отношении вообще выше всяких похвал.

Hardly managed to read to the end. Nothing interesting from the point of view of information. An example of intellectual detective stories are novels by U.Eko. It's a pleasure to read them, and (they have) deep philosophy, and are quite informative from the point of view of history. And as for aesthetics it's just beyond praise.

The novel being reviewed is not the one being described, and all the praise goes to novels by another author. None of the positive vocabulary has anything to do with the overall sentiment of the review's author towards the book under review.

Other reviews that are difficult to classify are those that describe some positive or negative aspects of a reviewed item, but in the end give an overall **sentiment of the opposite direction**. Consider the following positive review:

(27) Сюжет довольно обычен, язык изложения прост до безобразия. Много грязи, много крови и смерти. Слишком реально для сказки коей является фэнтези. Но иногда такие книги читать полезно, ибо они описывают неприглядную реальность.

The plot is quite usual, the language is wickedly simple. A lot of filth, a lot of blood and death. Too true-to-life for a fairy-tale, which a fantasy genre actually is. But it is useful to read such books from time to time, as they depict ugly reality.

The large number of negative lexical units may mislead an automatic classifier to a conclusion that the review is negative.

The three issues described above are present in approximately one-third of all reviews in the corpora. This suggests that a sentiment classifier using words as features could only correctly classify around 55–60% of all reviews.

This performance may be even worse for the Russian corpus as many of its reviews feature very unexpected ways of expressing opinion. Unlike most of the English reviews, in which a reviewer simply gives a positive or negative appraisal of a book, backing it with some reasoning and probably providing some description and analysis of the plot, Russian reviews often contain **irony, jokes, and use non-standard words and phrases**, making use of a variety of language tools, as illustrated in the following examples:

(28) Скушнаа. дошёл до бегства ГГ в мир Януса, и внезапно понял (я), что гори он (ГГ) хоть синим пламенем
Booorin'. got to the (episode of) GG fleeing to the world of Janus, and suddenly (I) realised that let it (GG) burn with blue flames (≈ I do not at all care about GG)

(29) Я эту муть не покупал. Shift+del.
I didn't buy this garbage. Shift+del.

Since there are more reviews of this kind in the Russian corpus than in the English, it is very likely that a Russian sentiment classifier would have lower accuracy.

Summary The reviews in English and in Russian often use different means of expressing sentiment, many of which are difficult (if at all possible) to process automatically. Often opinions are described through adjectives (86% of reviews contain adjectives). The second most frequent way of expressing sentiment is through verbs (59% of reviews have sentiment-bearing verbs). Less frequent is expression through nouns, in 39% of reviews. Sentence-level and discourse-level sentiment phenomena are found in 56% of reviews. 3% of reviews contain sentiment-related phonetic phenomena. Other issues that may affect automatic processing include mistakes and typos, topic shift and expressing an overall sentiment that is opposite to the sentiment direction of most of the review.

5.1.4 Movie Review Corpus

The corpus of film reviews created by Bo Pang and Lilian Lee (Pang and Lee, 2004) contains 1,000 positive and 1,000 negative reviews all written before 2002, with a cap of 20 reviews per author (312 authors total) per category[7]. This corpus is widely used for sentiment classification experiments and researchers report different results, ranging from 70% of accuracy in weakly supervised experiments by Read and Carroll (2009) to more than 86% in supervised classification by Pang and Lee (2004).

The domain of film reviews has been argued to be difficult for automatic sentiment analysis (Turney, 2002). Indeed, the collection of film reviews consists of mostly long and very well-written reviews featuring rich vocabulary and a professional writing style. The average length of a positive review is 788 words, a negative review is on average shorter: 707 words. Positive and negative reviews have vocabularies which are very similar in size, consisting of 36,806 and 34,542 words respectively, with 50,920 unique words in the entire

[7]Available at www.cs.cornell.edu/people/pabo/movie-review-data/ (review corpus version 2.0)

corpus. The large size of the vocabulary can be attributed not only to professional writing but also to the many proper names (film titles, names of actors, characters, film directors, different locations where an action takes place and so on). The high variety of the words used in the reviews means that many occur with low frequency and this may adversely affect performance of a classifier due to sparsity of data.

The content of the reviews is also difficult to analyse automatically. The main reason for this is the often complex and ambiguous structure of reviews which usually touch upon different aspects of a film, including its plot, performance of actors, camera work, historical background etc. All of these aspects may receive different sentiments which can be at variance with the overall opinion. This phenomenon was also noticed in the book reviews (Section 5.1.3).

Words may be used that have some appraisal meaning but this is not necessarily connected with the evaluation of a film. Consider the following example of a positive review of a film:

(30) on a return trip from new york where he was trying to get a job , dunne is in a horrible train accident that he is the only survivor of .

The word *horrible* can bear negative sentiment but in this review it is used to describe a plot, not the film. In general, most horror films may have a lot of negative words in their descriptions regardless of their overall quality. The opposite is true of romantic love stories that may contain excessive amounts of positive vocabulary despite being rated very poorly.

(31) if there are any positive things to say about " message in a bottle , " it is that the performances by robin wright penn and paul newman , as garrett's stubborn , but loving father , are far above par to be in such a wasteful , " shaggy dog " love story , and that the cinematography by caleb deschanel takes great advantage of the beautiful eastern coast , and paints chicago as an equally alluring city .

5.2 Supervised Classification Experiments

Following the same procedure as in the previous experiments with Chinese customer reviews (Chapter 4), the first set of experiments was designed to determine a 'supervised upper bound' with which to compare unsupervised approaches.

5.2.1 Lexical Unit Extraction

As in the previous experiments with Chinese reviews, I used the same technique of extracting lexical units from the corpus by finding the longest common string in any two zones of the corpus (Section 4.2.2). Since the Movie review corpus is comparatively large, the resulting vocabulary is large consisting of more than 1,250,000 items. The large number of lexical units made processing very slow so I filtered the list of extracted lexical units to exclude ones with low frequency (less than 50 occurrences in the corpus) which resulted in a list of 38,116 items. The English book review corpus, being much smaller than the Movie review corpus, produced only 7,913 lexical units.

This approach appears to work well for English, as it permits the extraction, for example, of word sequences expressing features that are discussed by reviewers such as *the supporting cast* or *the special effects*, as well as phrases that could be used for appraisal such as *good performance*, *best performance* and *interesting and*.

The same approach was applied to the Russian book review corpus; despite the language's complex morphology one might hope the technique would be able to capture more unchangeable (stable) units as well as word forms that may also be frequent. This indeed turns out to be the case, since the approach extracts some 'semi-stemmed' forms that comprise the most important part of the word, leaving out affixes denoting minor grammatical features, for example, the lexical unit *бессмыленн* which is a common part of the word forms *бессмыленный, бессмыленная, бессмыленных, бессмыленного* and many others meaning *senseless*. The Russian corpus produced 8,372 lexical units.

In addition, for the English language corpus which features explicit word boundaries and does not have complex morphology, it is possible to use another technique of extracting lexical units. While finding the longest common string, I split all strings at space and filtered out those items that occurred less than 10 times in the corpus. This approach produced a list of 7,452 items from the Movie review corpus. Unlike the previous approach the items extracted by this simple method are, in fact, graphical words or their combinations.

5.2.2 Experimental results

I used two machine learning algorithms: Naïve Bayes multinomial (NBm) and Support Vector Machines (SVM). The feature sets were the lexical units extracted from the relevant corpora. The evaluation technique was 10-fold cross-validation.

Table 5.5 shows rather satisfactory results of supervised classifiers applied to English book reviews. Russian book reviews do not perform very well especially with the SVM

	NBm			SVM		
Corpus	P	R	F	P	R	F
English movie reviews	0.82	0.82	0.82	0.83	0.83	0.83
English book reviews	0.86	0.86	0.86	0.86	0.86	0.86
Russian book reviews	0.81	0.81	0.81	0.76	0.76	0.76

Table 5.5: Supervised classification results (10-fold cross-validation, lexical units).

classifier, which may be the result of the linguistic features described above. Film reviews also perform reasonably well but not as well as the book reviews.

Apart from the results of supervised classification presented in the Table, other researchers' sentiment classification results may also be used as strong upper bounds. For example, the authors of the English movie review corpus achieved an accuracy of 86% in their experiments using supervised classifiers and preliminary subjectivity classification Pang and Lee (2004). Li et al. (2009) achieved almost an accuracy of 0.80 with 50% of all documents labelled.

LU and Words

To test the impact lexical units have on sentiment classification in English and Russian, I also ran the same supervised classifiers using words extracted from the corpora as features. To make the resulting lexicons comparable (in terms of their elements' frequencies) I filtered out all words that occurred less than 10 times. I extracted all words from the corpora but did not process them in any way (no stemming or lemmatisation) as any of these techniques are language-dependent and would run counter to the unsupervised research paradigm. 1,075 words were extracted from the Russian corpus, 1,247 words from in the English book reviews, and 12,554 words from the movie reviews.

	NBm			SVM		
Corpus	P	R	F	P	R	F
English movie reviews	0.81	0.81	0.81	0.83	0.83	0.83
English book reviews	0.85	0.85	0.85	0.83	0.83	0.83
Russian book reviews	0.78	0.78	0.78	0.73	0.73	0.73

Table 5.6: Supervised classification results (10-fold cross-validation, words).

Table 5.6 shows that the results were worse for all the corpora performed worse compared with the LU-based classification. This could be expected for the Russian corpus as the abundance of word forms makes the data sparse. The small difference in performance on the English corpora could also be expected because of the smaller number of possible word forms in English.

5.3 Unsupervised Classification Experiments

The following experiments are based on the same techniques as those used for the Chinese data. Lexical units are the basic unit of processing, with all documents being split into lexical units using the same algorithm (the Longest common substring algorithm, as described in Section 4.2.2). The experiment use the zone-based classifier described in Chapter 4.

5.3.1 Seed-Based Classification

The multilingual experiments used three different sets of seeds: two manually selected sets of seeds, and a set of semi-automatically extracted seeds. For comparison purposes, I also used a pre-exising sentiment lexicon for English.

Manually Selected Seeds

For each of the languages under consideration, I manually selected two kinds of seed lists: 'short' and 'long'. The former consists of only two seeds (one for each sentiment direction) and the latter comprises six seeds (three for each sentiment). I did this intuitively without any preliminary study of their effectiveness for sentiment classification. The only requirement was that they should express positive or negative sentiment unambiguously.

The short list comprised the two lexical units: 'good' and 'bad' for the English corpus experiments. Choosing seeds for experiments in Russian was more difficult in the absence of a morphological parser, since the grammatical form of seeds may affect performance. To avoid this, I used the shortest possible forms: *плох* and *хорош* as most of the other forms include them as a part. The long list for the Russian language is shown in Table 5.7[8].

The seeds selected for English were: *good, wonderful, magnificent; bad, terrible, disgusting.*

[8]Note that all endings related to grammatical forms were deleted thus making the seeds ungrammatical (except for 'good' and 'bad' which are used in correctly formed short forms of masculine singular).

Seed	Gloss	Sentiment
хорош	good	POS
замечательн	outstanding	POS
великолепн	magnificent	POS
плох	bad	NEG
ужасн	horrible	NEG
отвратительн	disgusting	NEG

Table 5.7: The manually selected Russian seeds.

Automatically Extracted Seeds

For Russian and English, I used only a partially unsupervised version of the Chinese positive word extraction technique, manually selecting from a candidate seed list produced by extracting lexical units preceded by negations and adverbials—since these frequently indicate sentiment orientation. Tests with a fully unsupervised technique produced too many irrelevant candidates due to language and domain-specific issues. Specifically, in English and Russian, negative sentiment is more often expressed by separate words rather than by negated positives; and the domain of book and especially of movie reviews features a diverse vocabulary only part of which is relevant to sentiment.

This approach produced a list of 68 positive and 15 negative Russian terms, presented in Table 5.8. 65 positive and 46 negative terms were extracted for the English book reviews and 38 positive and 6 negative seeds were extracted from the film reviews corpus (see Table 6.3). The small number of seeds found in the film reviews corpus is the result of its rich vocabulary and extensive use of contextual means of expressing sentiment which cannot be capture in the out-of-context filtering process.

Sentiment Vocabulary

For the English experiments I also used a list of sentiment-related words which was compiled on the bases of the subjectivity clues created by Wilson et al. (2005). I used only those clues (words) that were marked as strongly subjective with their sentiment direction specified. The resulting sentiment word list has 2,718 positive items and 4,912 negative items.

Corpus	Seeds
Russian book reviews (Positive seeds)	нравится, хорошо, перспективный, красиво, понравилось, интересная, глубокая, увлекательная, познавательная, здорово, жизненны, хорошая, круто, неплохой, приятным, хорошоая, завораживает, лихо, интересный, увлекательно, задела, интересно, понравился, понравилась, умно, живые, старается, неплохо, ховошая, реалистично, удачная, своеобразно, понравились, хорошее, любил, интересные, нравиться, советую, детально, чётко, прилично, влюбился, хорош, милая, красивое, глубоким, доходчиво, яркая, понравилось, прекрасный, тщательно, сильное, приятное, неплохая, красочно, добротная, реалистичный, одарённому, долгие, цельные, необычный, яркий, удачные, хороший, правдоподобно, оригинальный, интересной, компактно,
Russian book reviews (Negative seeds)	плохой, зря, слабый, примитивно, занудным, средненько, картонны, мутно, бледно, предсказуемо, плохо, утомляет, слабо, плоско, слабенькая

Table 5.8: Semi-automatically extracted Russian seeds.

Corpus	Seeds
English book reviews (Positive seeds)	believable, seductive, likable, pretty, good, well, great, happy, impressed, humourous, funny, clever, familiar, enjoyable, glad, pleased, likeable, popular, worthwhile, exciting, beautiful, real, best, absorbing, strong, entangling, honest, explosive, grounded, realistic, extensive, cleverly, gripping, nice, readable, particular, fine, dynamic, easy, captivating, descriptive, interesting, challenging, greatly, erudite, imaginative, knowledgable, moving, emotional, human, inspiring, graphic, heartwarming, addictive, intresting, touching, generous, neatly, talented, interested, unique, detailed, important, intersting, entertaining
English book reviews (Negative seeds)	shallow, hard, predictive, confusing, odd, weak, dull, complicated, badly, difficult, mediocre, wooden, worst, offensive, silly, poor, onedimensional, awful, thin, uninvolved, boring, disappointing, lengthy, poorly, ordinary, cynically, disheartening, thinly, disapointing, dissappointed, wrong, tedious, predicable, untastefully, disturbing, selfcentered, predictable, harsh, complex, dissapointing, obvious, depressing, unrealistic, bad, loosely, sorry
English film reviews (Positive seeds)	good, great, funny, original, interesting, nice, deep, wise, strong, entertaining, surprising, important, successful, involving, happy, involved, smart, clever, convincing, believable, appropriate, memorable, bright, interested, charming, spectacular, satisfying, lucky, fond, impressed, faithful, carefully, coherent, keen, pleased, helpful, believeable, humerous
English film reviews (Negative seeds)	bad, hard, difficult, dumb, shabby, heavy

Table 5.9: Semi-automatically extracted English seeds.

5.3.2 Classification Results

The first set of experiments test the iterative sentiment classifier without the sentiment score feature (only with the negation check).

The experiments with the Russian corpus test the three set of seeds: 2 seeds, 6 seeds and extracted seeds. Table 5.10 presents results of the iterative classification of the Russian book review corpus. The ultimate performance correlated with the number of seeds used for the initial iteration.

	P	R	F
Russian books			
2 seeds	0.66	0.61	0.63
6 seeds	0.69	0.63	0.66
Extracted	0.73	0.67	0.70

Table 5.10: Russian book reviews: results of classification.

For the experiments with the English corpora, apart from the three sets of seeds I also used the sentiment vocabulary described above. To investigate whether different sets of lexical units extracted from a corpus affect performance, I also tested two different sets of lexical units: *bigMovie* is a set of lexical units extracted from the Movie Review corpus by the same technique used for Russian and Chinese; the same approach applied to the English book reviews produced the *books* set. *smallMovie* is the set of words (not lexical units) produced by splitting graphical words at space (or other word delimiters used in English) as described in Section 5.2.1.

Table 5.11 shows the results of iterative classification running on the English book reviews and the Movie reviews. The results for the former also (as in the case of the Russian corpus) improve in line with the number of seeds used for the initial iteration. However, the latter shows opposite tendency, performing much better with only two seeds and hardly better than the naïve baseline with the extracted seeds. This suggests that the more complex structure of film reviews makes it difficult for a human to predict which words are reliable indicators of sentiments (Section 5.3.1). The 2 seeds may perform better as they are less dependent on a human's choice, leaving it to the system (and the corpus) to 'decide' what lexical items are good for sentiment classification.

The sentiment vocabulary does not seem to be effective for either corpus, being only slightly better than the 2-seed setting for the book reviews. In film reviews, the vocabulary

is only better than the extracted seeds, which performed extremely poorly. However, with the *smallMovie* set, the vocabulary performed better than any of the seed words, most probably because the set included word- and phrase-like lexical units. For book reviews, the *smallMovie* set combined with the extracted seeds proved to be the best. But its performance is only one percentage point (F-measure) better than the result of the extracted seeds on the *bigMovie* set. Interestingly, the *books* set turned out to be the worst (although only a couple of percentage points) for the book review corpus. This suggests that the larger number of extracted lexical units may compensate for their out-of-domain origin, at least for such related domains as book and film reviews.

	bigMovie			smallMovie			books		
	P	R	F	P	R	F	P	R	F
English books									
2 seeds	0.70	0.66	0.68	0.74	0.68	0.71	0.65	0.63	0.64
6 seeds	0.78	0.75	0.76	0.78	0.73	0.75	0.77	0.72	0.74
Extracted	0.80	0.76	0.78	**0.82**	**0.77**	**0.79**	0.79	0.74	0.77
Vocabulary	0.75	0.72	0.73	0.75	0.72	0.73	0.71	0.68	0.70
English films									
2 seeds	**0.70**	**0.69**	**0.69**	0.63	0.63	0.63	-	-	-
6 seeds	0.67	0.66	0.67	0.64	0.64	0.64	-	-	-
Extracted	0.56	0.56	0.56	0.61	0.61	0.61	-	-	-
Vocabulary	0.64	0.64	0.64	0.67	0.67	0.67	-	-	-

Table 5.11: English corpora: results of classification.

5.3.3 Score Difference

The score difference technique decreased performance of the classifier on all data sets and with all seeds as well as with the vocabulary (see Tables 5.12 and 5.13). Inspection of the results showed that the main culprit was the iteration control that failed to stop the classifier at the best classification. For example, the classifier managed to achieve a reasonably good performance on movie reviews (Precision 0.73, Recall 0.72, F-measure 0.72) with score difference 0.1, but the number of classified documents was not the biggest, so the classifier did not chose this result as the best.

	P	R	F
Russian books			
2 seeds	0.64	0.60	0.62
6 seeds	0.67	0.63	0.65
Extracted	0.71	0.66	0.68

Table 5.12: Russian book reviews: results of classification.

	bigMovie			smallMovie			books		
	P	R	F	P	R	F	P	R	F
English books									
2 seeds	0.76	0.74	0.75	0.74	0.70	0.72	0.65	0.63	0.64
6 seeds	0.78	0.75	0.76	0.80	0.75	0.77	0.76	0.75	0.77
Extracted	0.80	0.77	0.79	**0.82**	**0.77**	**0.80**	0.79	0.75	0.77
Vocabulary	0.76	0.74	0.75	0.75	0.72	0.73	0.71	0.68	0.69
English films									
2 seeds	0.66	0.66	0.66	0.64	0.64	0.64	-	-	-
6 seeds	0.66	0.66	0.66	0.65	0.65	0.65	-	-	-
Extracted	0.54	0.54	0.54	0.61	0.61	0.61	-	-	-
Vocabulary	0.64	0.64	0.64	**0.67**	**0.67**	**0.67**	-	-	-

Table 5.13: English corpora: results of classification using score difference.

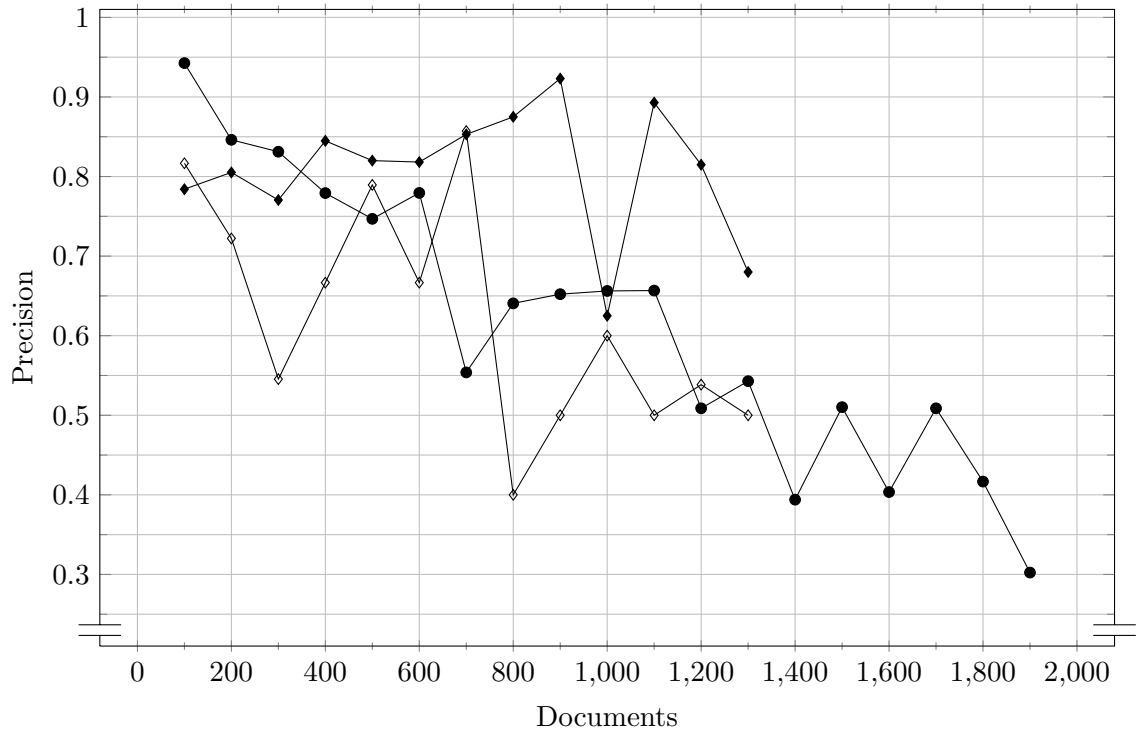

Figure 5.2: Information retrieval simulation results with the zone distance technique. —●— is *English movie review corpus*; —◇— is *English book review corpus*; —♦— is *Russian book review corpus*.

5.3.4 Zone Difference for Result Ranking

Section 4.3.2 described experiments with the zone difference technique. The main application of this technique was IR-like ranking of results according to their 'reliability', so that results with the most accurate classification were put on the first 'page' and those documents that probably were not classified very accurately were presented on the last 'page', each 'page' containing 100 documents. Figure 5.2 presents the results of this technique applied to the English and Russian corpora. Obviously, reviews from both of the book corpora are not distributed over the 'pages' properly: accurate results can be found in the middle of the graph, not only in the beginning. Movie reviews, however, show a very good distribution across the 'pages' with the most accurate (Precision = 0.94) on the first 'page' and the least accurate (Precision = 0.30) on the last 'page'.

5.3.5 Combining with Supervised Machine Learning Techniques

In an attempt to improve classification, I applied machine learning techniques to the results of unsupervised classification. Thus, the training corpus was the one extracted by the unsupervised classifier from the original corpus and the features were all the lexical

	P	R	F
Russian books			
2 seeds	0.69	0.68	0.67
6 seeds	0.72	0.70	0.69
Extracted	0.76	0.75	0.75

Table 5.14: Russian book reviews: results of classification.

units extracted from the corpus.

The results of classification of the Russian book reviews by means of the combined (unsupervised + machine learning) classifier (see Table 5.14) show improved performance over the initial classifier. The biggest improvement is in recall, which grew by 7-8 percentage points, with precision adding 3 percentage points. Compared with the supervised upper bound, these results are still far behind, although extracted seeds are only 6 percentage points worse. 6 seeds are 12 and 2 seeds are 14 percentage points worse.

Table 5.15 presents results for the English corpora. The English book review corpus performed better with the machine learning technique than without it (Table 5.11) gaining from 3 to 7 percentage points in recall and 3 to 5 in precision. Compared to the supervised upper bound, it is 6 to 11 percentage points worse (Table 5.5). The results for film reviews did not improve with the machine learning technique. This can be attributed to the poor performance of the initial classifier which produced bad training corpora for the NBm classifier. The latter produced skewed results (which is revealed by an unexpectedly low F-measure which is the weighted average of the classification results of the two classes). Only 2 seed-based classification performed on the same level as the initial classifier, but still being 14 percentage points behind the upper bound. However, the two seeds results are 3 percentage points better than the results reported by Turney (2002).

5.4 Discussion

This chapter presented two comparable corpora of book reviews in Russian and English. A study of the language-specific issues indicated a number of problems that may complicate sentiment classification. In particular, a complex morphology of Russian may affect the performance of a classifier that does not use any preprocessing techniques, such as stemming or lemmatisation. However, lexical units seem to be able to overcome this problem, proving their effectiveness as basic units for multilingual classification.

	bigMovie			smallMovie			books		
	P	R	F	P	R	F	P	R	F
English books									
2 seeds	0.70	0.69	0.69	0.74	0.73	0.73	0.70	0.67	0.65
6 seeds	0.78	0.78	0.78	0.79	0.79	0.78	0.80	0.79	0.79
Extracted	**0.81**	**0.81**	**0.81**	0.81	0.81	0.81	0.82	0.81	0.80
Vocabulary	0.76	0.75	0.74	0.78	0.75	0.74	0.75	0.72	0.71
English films									
2 seeds	0.69	0.69	0.68	0.71	0.64	0.61	-	-	-
6 seeds	**0.72**	**0.69**	**0.68**	0.70	0.65	0.62	-	-	-
Extracted	0.67	0.59	0.53	0.71	0.63	0.59	-	-	-
Vocabulary	0.68	0.65	0.63	0.70	0.68	0.67	-	-	-

Table 5.15: English corpora: results of classification using machine learning (NBm).

Unsupervised classification of the Russian and English book reviews and English film reviews performed well, achieving at least almost 0.70 F-measure for all the corpora. For the book reviews, the performance seems to depend on the size of the seed list. The best results were obtained by means of seeds extracted semi-automatically from the reviews. The English corpora were also classified using a pre-existing Sentiment Vocabulary comprising almost 8,000 items. Results with this were still inferior to the in-domain seeds: the movie reviews corpus performed better with only 2 and 6 generic seeds and the English book reviews performed better with the extracted seeds and 6 seeds. These results suggest that an in-domain vocabulary performs better than a generic one, even if the latter is bigger in size. This seems to be true despite the fact that the 2- and 6-seed lists comprised generic seeds too. Obviously their impact on performance was very small compared to the number of in-domain lexical units extracted with their help. Probably the large generic list was able to influence performance after the first iteration, but the difference in performance occurred after the first iteration; for example on the movie reviews corpus the first iteration with the Sentiment vocabulary resulted in Precision = 0.60, Recall = 0.58 and F1 = 0.59, while the 2-seed classifier achieved 0.63, 0.41, 0.50 respectively. This suggests that the vocabulary was not able to produce a classification of the same accuracy as only two seeds did (higher recall seems not to be of key importance). It is possible to

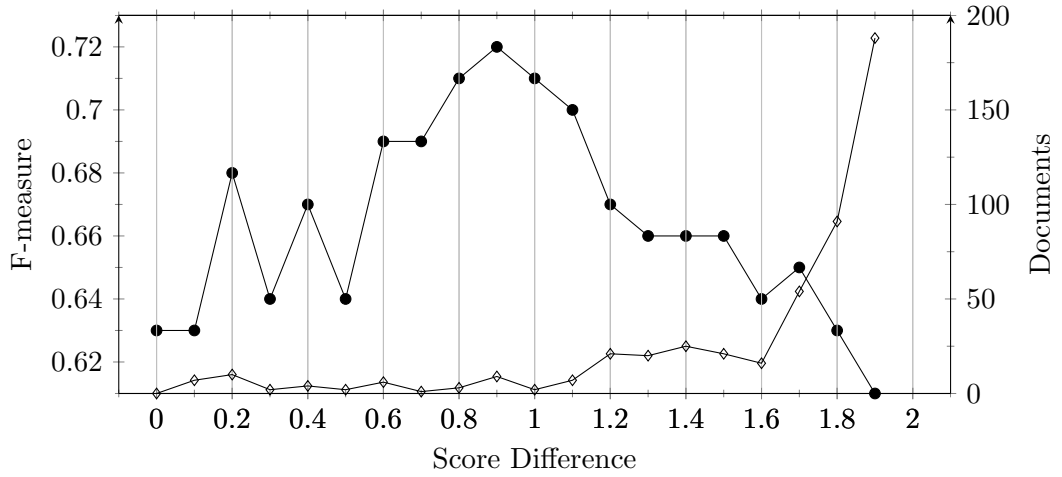

Figure 5.3: Score difference results for Movie review corpus: —●— is F-measure; —◇— is the number of documents that were NOT classified.

conclude that the quality of seeds might be more important than their number.

Another interesting finding is that the English book review corpus results were better using lexical units extracted from a larger corpus of movie reviews. The bigger list of LUs extracted from the reviews of films (*bigMovie*) included most of the book review LUs (6379 out of 7913). This suggests that 1) it is possible to use lexical units from a close domain and 2) the more lexical units a classifier can use, the better results it produces. Even better results on book reviews achieved using *smallMovie*, LUs extracted from the movie reviews split at space. This can be attributed to the larger average size of lexical units in this list: 8.50 against 6.89 of book review LUs.

The relatively poor performance of the score difference technique is a result of the iteration control subsystem failure to stop at the best iteration. Inspection of the score-difference results showed that the technique managed to increase performance of all the classifiers but the best performance did not coincide with the biggest number of classified documents. The reason for this is that the classifier was able to classify almost all of the documents beginning from the first iteration. The number of unclassified documents was very small (compared to the total number of documents), ranging between 0 and 10 (see Figure 5.3). This suggests that the iteration control cannot work effectively when a classifier is able to process almost all of the documents.

The zone difference technique in the multilingual experiments performed well only on the movie review corpus due to longer reviews, containing more zones. The book review corpora have rather short reviews containing few zones which makes the zone difference ranking ineffective.

The automatic seed extraction did not work for the languages used in the experiments. Apparently, the automatic seed extraction is language-specific because it benefits from certain features of the Chinese language.

In conclusion, despite a few problems revealed by the experiments, the unsupervised, knowledge-poor approach performed reasonably well in multilingual settings.

Chapter 6

Multi-Aspect Sentiment Analysis

The previous chapters dealt with unsupervised sentiment classification at the document level. As noted in the Introduction, the task of sentiment analysis is more complex and may require the extraction of more fine-grained information that is part of an opinion. Following the same unsupervised research paradigm, this Chapter investigates the possibility of unsupervised approaches to further aspects of sentiment analysis.

This Chapter describes investigations into three different aspects of sentiment analysis. Section 6.1 presents experiments on extending two-class sentiment classification by introducing a new, neutral class. This section also explores a possibility of simultaneous sentiment / subjectivity classification. The experiments use a novel approach to sentiment classification: scale-based classification (rather than binary classification). Section 6.2 further investigates unsupervised subjectivity classification and presents experiments on sentence-level subjectivity classification in English, Chinese and Japanese. This section also tests a new approach to seed list expansion. The same set of languages is used in Section 6.3, which describes experiments on opinion holder and opinion target extraction.

6.1 Three-Way Classification[1]

The previous Chapter assumed the existence of two classes of sentiment: positive and negative. However, there exist at least a third sentiment class – neutral. Neutral opinion does not express any support or criticism of a target but still expresses a subjective judgement, for example: *I think the table is big.* From this phrase it is not possible to conclude if this subjective utterance expresses a positive or negative opinion regarding the table so it should be classified as neutral.

[1]The experiments and part of the discussion in this section were presented in a condensed form at the Third International Joint Conference on Natural Language Processing (Zagibalov and Carroll, 2008a)

Considering sentiment classification in more general terms leads to the insight that positive and negative sentiments are extreme points in a continuum of sentiment, and that intermediate points on this continuum are of potential interest. For instance, in a real-world application context, someone might want to get an idea of the types of things people are saying about a particular product through reading a sample of reviews covering the spectrum from highly positive, through balanced[2], to highly negative. In another scenario, a would-be customer might only be interested in reading balanced reviews, since they often present more reasoned arguments with fewer unsupported claims. Such a person might therefore want to avoid reviews such as Example (1) – written by a Chinese purchaser of a mobile phone.

(1) 软件不行，发送短信时有时对方接收不 到；兼容性也不行，有的手机收到的短 信是乱码！还有死机现象！拍照效果次！ 不是循环或自定义式闹铃，每次都要调，太麻烦了！后盖不够严密！原装配件中 无座充！

The software is bad, some sent SMS are never received by the addressee; compatibility is also bad, while on some mobile phones messages received are in a scrambled encoding! And sometimes the phone 'dies'! Photos are horrible! It doesn't have a cyclic or programmable alarm-clock, you have to set it every time, how cumbersome! The back cover does not fit! The original software has many holes!

In a third scenario, someone might decide they would like to read only opinionated, weakly negative reviews such as Example (2), since these often contain good argumentation while still identifying the most salient bad aspects of a product.

(2) 这机子的反应速度超慢的哦，彩信必须 要30KB以下才能收，也不支持MP3铃声，自带铃声也不好听，时不时的还会死机， 本来买的时候挺喜欢的，样子挺独特，红色白色搭配的，挺有个性，也不贵， 但是用着实在是总出状况，让人头疼

The response time of this mobile is very long, MMS should be less than 30kb only to be downloaded, also it doesn't support MP3 ring tones, (while) the built-in tunes are not good, and from time to time it 'dies', but when I was buying it, I really liked it: very original, very nicely matching red and white colours, it has its individuality, also it's not expensive, but when used it always causes trouble, it makes one's head ache

This review contains both positive and negative sentiment covering different aspects of

[2] A review is balanced if it is an opinionated text with an undecided or weak sentiment direction.

the product, and the fact that it contains a balance of views means that it is likely to be useful for a would-be customer. Moving beyond review classification, more advanced tasks such as automatic summarisation of reviews (e.g. Feiguina and Lapalme, 2007) might also benefit from techniques which could distinguish more shades of sentiment than just a binary positive / negative distinction.

A second dimension is subjective / factual. When shopping for a product, one might be interested in the physical characteristics of the product or what features the product has, rather than opinions about how well these features work or about how well the product as a whole functions. Thus, if one is looking for a review that contains more factual information than opinion, one might be interested in reviews such as in Example (3).

(3) 总的感觉这台机器还不错，实用的有： 开（关）机闹钟5个，800条（500个人）电话本，阴阳历显示，时间与日期快速 转换，WAP上网，日程表，记事本等。
(My) overall feeling about this mobile is not bad, it features: 5 alarm-clocks that switch the phone on (off), a phone book for 800 items (500 people), lunar and solar calendars, fast switching between time and date modes, WAP networking, organizer, notebook and so on.

This review is mostly factual, but contains information that could be useful to a would-be customer which might not be in a product specification document, for example fast switching between different operating modes. Similarly, would-be customers might be interested in retrieving completely factual documents such as technical descriptions and user manuals. Again, as with sentiment classification, subjective and factual texts are not easily distinguishable separate sets, but form a continuum. In this continuum, intermediate points cam be of interest as well as the extremes.

6.1.1 Sentiment Classification

In this investigation, computation of sentiment is carried out in the same way as described previously in Chapter 4. For the experiments, I used the Chinese corpus of customer reviews of mobile phones, consisting of 2,317 documents (1,158 positive and 1,159 negative). The classifier starts out with a seed vocabulary consisting of the single word 好 (*good*), and bootstraps a domain-specific list of lexical units as described in Section 4.2.1. As discussed in Section 3.4.1, in order to determine the sentiment direction of the whole document, the classifier computes the difference between the number of positive and negative zones. If the result is greater than zero the document is classified as positive, and vice versa. If the

result is zero, the document is balanced or neutral for sentiment.

Given a sentiment classification for each zone in a document, a quantity called **sentiment density** is computed as the proportion of opinionated zones with respect to the total number of zones in the document:

$$SentimentDensity = \frac{\sum Z_{opinionated}}{\sum Z_{total}} \tag{6.1}$$

.

Sentiment density measures the proportion of opinionated text in a document, and thus the degree to which the document as a whole is opinionated. It should be noted that neither sentiment score nor sentiment density are absolute values, but are relative and only valid for comparing one document with other. Thus, a sentiment density of 0.5 does not mean that the review is half-opinionated, and half not. It means that the review is less opinionated than a review with a density of 0.9.

This section started by arguing that sentiment and subjectivity should both be considered as continua, not binary distinctions. The technique described above compares the number of positive and negative zones for a document and treats the difference as a measure of the 'positivity' or 'negativity' of a review. The document in Example (2), with 12 zones, is assigned a score of -1 (the least negative score possible): the review contains some positive sentiment but the overall sentiment direction of the review is negative. In contrast, Example (1) is identified as a highly negative review, as would be expected, with a score of -8, from 11 zones. Similarly, with regard to subjectivity, the sentiment density of the text in Example (3) is 0.53, which reflects its more factual character compared to Example (1), which has a score of 0.91. I represent sentiment and subjectivity on two scales: *positive – negative* and *factual – subjective*. The scales can be combined into a single coordinate system. Most product reviews could be expected to be placed towards the top of the coordinate system (i.e. opinionated), and stretch from left to right.

Figure 6.1 plots the results of sentiment and subjectivity classification of the test corpus in this two-dimensional coordinate system, where `X` represents sentiment (with scores scaled with respect to the number of zones so that -100 is the most negative possible and +100 the most positive), and `Y` represents sentiment density (0 being factual and 1 being highly subjective). Most of the reviews are located in the upper part of the coordinate system, indicating that they have been classified as subjective, with either positive or negative sentiment direction. Looking at the overall shape of the plot, more opinionated documents tend to have more explicit sentiment direction, while less opinionated texts

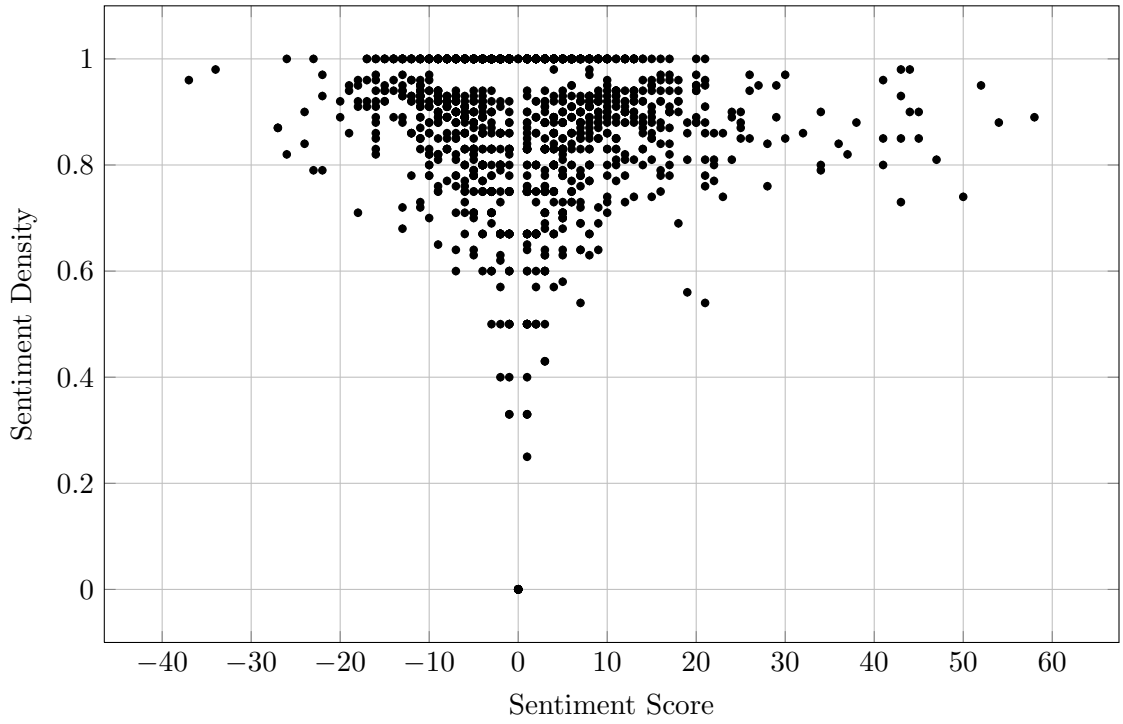

Figure 6.1: The distribution of Chinese customer reviews with respect to Sentiment Score and Sentiment Density.

stay closer to the balanced / neutral region (around $X = 0$).

6.1.2 Subjectivity Classification

As can be seen in Figure 6.1, the classifier managed to map the reviews onto the coordinate system with the predicted type of distribution. There are very few points in the neutral region, that is, on the same $X = 0$ line as balanced but with low sentiment density; this is expected, bearing in mind that the corpus is of reviews that express opinions towards certain products. To see if the system is capable of finding factual documents, I conducted a further experiment. I took Wikipedia[3] articles written in Chinese on mobile telephony and related issues, as well as several articles about the technology, the market and the history of mobile telecommunications, and split them into small parts (about a paragraph long, to make their size close to the size of the reviews) resulting in a corpus of 115 documents, which can be assumed to be mostly factual. I processed these documents with the classifier using lexical units and their scores extracted from the sentiment corpus and found that they were mapped almost exactly where balanced documents should be (see Figure 6.2).

Most of these documents have weak sentiment direction (X = -5 to +10), but are

[3]www.wikipedia.org

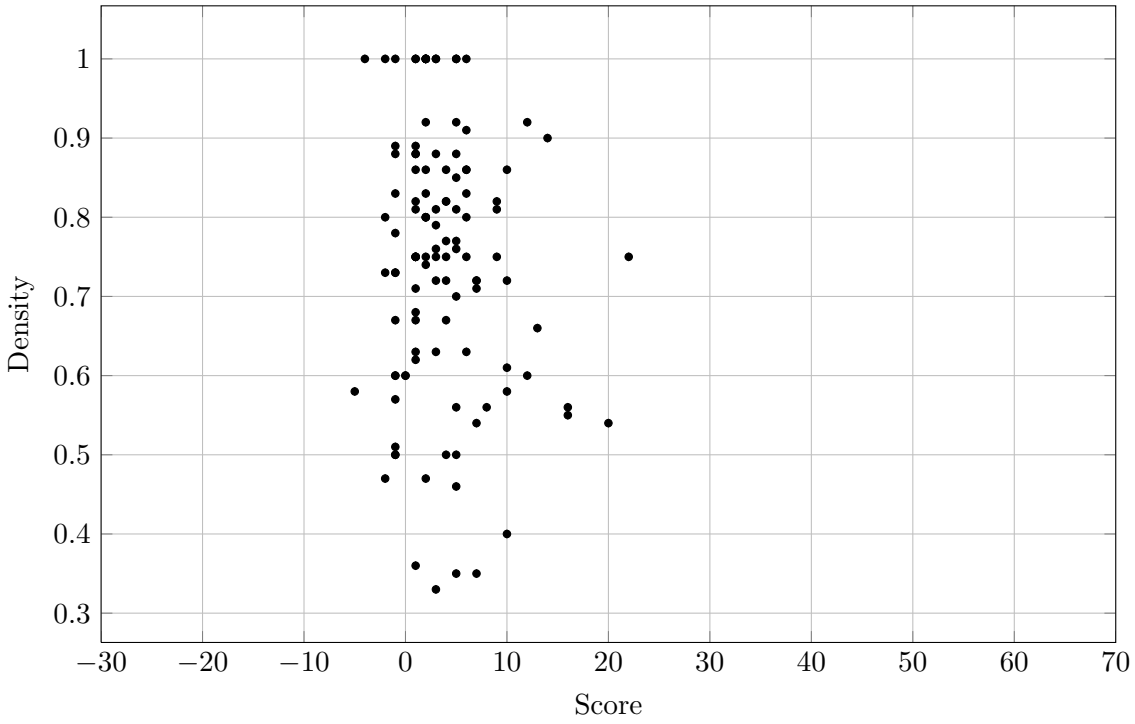

Figure 6.2: The distribution of factual documents with respect to Sentiment Score and Sentiment Density.

classified as relatively opinionated ($Y > 0.5$). The former is to be expected, whereas the latter is not. When investigating the possible reasons for this behaviour I noticed that the classifier found not only feature descriptions (such as 手感很好 *nice touch*) and expressions which describe attitude (喜欢 *(one) like(s)*), but also product features (for example, 彩信 *MMS* or 电视 *TV*) to be opinionated. This is because the presence of some advanced features such as MMS in mobile phones was often regarded as a positive by authors of reviews. In addition, the classifier found words that were used in reviews to describe situations connected with a product and its features: for example, 服务 (*service*) was often used in descriptions of quite unpleasant situations when a user had to turn to a manufacturer's post-sales service for repair or replacement of a malfunctioning phone, and 用户 (*user*) was often used to describe what one can do with some advanced features. Thus, the classifier was able to capture some product-specific as well as market-specific sentiment markers, however, it was not able to distinguish the context in which these generally objective words were used. This resulted in relatively high sentiment density of neutral texts which contained these words but used in other types of context.

To verify this hypothesis, I applied the same processing to the corpus derived from Wikipedia articles, but using as the vocabulary list the NTU Sentiment Dictionary. The results (Figure 6.3) show that most of the neutral texts are now mapped to the lower part

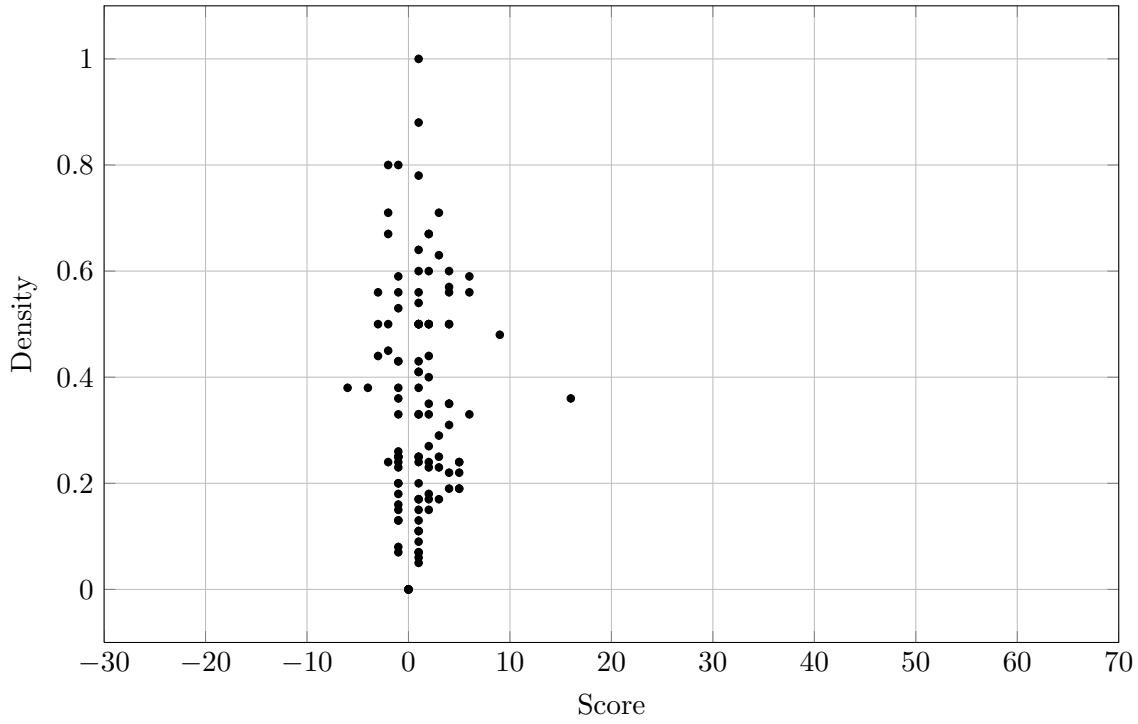

Figure 6.3: The distribution of factual documents with respect to Sentiment Score and Sentiment Density with the NTU Sentiment Dictionary.

of the subjectivity scale ($Y < 0.5$), as expected. Therefore, to successfully distinguish between balanced reviews and neutral documents, a classifier should be able to detect when product features are used as sentiment markers and when they are not.

The results also suggest that product attributes and descriptions of product-related situations play some role in expression of sentiment. However, these elements are very context-dependent in terms of their sentiment markedness.

6.2 Sentence-Level Subjectivity and Sentiment Classification[4]

The previous Section showed that the sentiment classifier is capable of subjectivity classification using the NTU Sentiment dictionary as the vocabulary list. In this section, I use a new data set of news items that contains both subjective and factual sentences. The sentences are also marked according their polarity. This data makes it possible to experiment with combined sentiment and subjectivity classification at the sentence level in the news domain.

[4]The experiments and part of the discussion in this section were presented in a condensed form at the NTCIR-7 MOAT Workshop Meeting (Zagibalov and Carroll, 2008c).

Applying sentiment classification to the data is difficult since the data includes both subjective and objective sentences which makes sentiment classification dependent on the accuracy of subjectivity classification. Simultaneous classification of sentiment direction and subjectivity is problematic as shown by Esuli and Sebastiani (2006a) and illustrated in the previous section. Another difficult point is a three-way classification which adds a class of neutral sentiment.

The subjectivity classification was done by marking as subjective all sentences whose sentiment score equals zero (either because no zones contain sentiment markers, or the number of positive zones equals the number of negative ones). Neutral sentences are those that may show a difference between positive and negative sentiment scores but where this difference is smaller than a threshold.

6.2.1 Data

For the experiments I used the NTCIR-7 MOAT (Multilingual Opinion Analysis Task) English, Chinese and Japanese test data collections. The English data runs from 1998 to 2001 with news items from the Mainichi Daily News, Korea Times, Xinhua News, Hong Kong Standard, and the Straits Times. It consists of 142 documents split into 14 topics (4312 sentences). The Traditional Chinese data contains documents from 1998 to 2001 from the China Times, Commercial Times, China Times Express, Central Daily News, China Daily News, United Daily News, Economic Daily News, Min Sheng Daily, United Evening News, and Star News, consisting of 188 documents in 14 topics (4655 sentences). The Simplified Chinese data contains documents from Xinhua News and Lianhe Zaobao from 1998 to 2001, consisting of 252 documents in 14 topics (4877 sentences). The Japanese data consists of 249 Japanese news items from 1998 to 2001 from the Mainichi newspapers split into 18 topics (5885 sentences)[5]. All documents in the test corpus in each language were annotated using a pool of six annotators (Seki et al., 2008).

6.2.2 Classification Using an Existing Classifier

To set a baseline, I applied the existing classifier at the level of individual sentences.

Traditional Chinese

The first experiment tests three different sets of seeds. The first set consisted of the six seeds used in previous experiments (Section 4.2.1). The second set was comprised of seeds

[5]The Simplified Chinese data was used only for experiments described in Section 6.2.4.

extracted semi-automatically from the corpus (Table 6.1), and the third set used all of the seeds.

Positive seed	Translation	Negative seed	Translation
成功	success	慘	tragic
穩定	stable	不幸	unlucky
樂觀	optimistic	困難	difficulty
完整	complete	難過	hardship
合理	reasonable	遺憾	regret
簡便	cheap and easy		

Table 6.1: Extracted seeds

All the seeds performed rather poorly, with only the 6-seed set performing slightly better than a naïve baseline (0.47 if all sentences are marked as opinionated) in the subjectivity classification task, and at the level of the worst performing supervised classifiers participating in NTCIR-7 in the sentiment classification task (see Table 6.2).

	Subjectivity			Sentiment		
Seeds	P	R	F	P	R	F
6 seeds	0.52	0.66	0.58	0.20	0.26	0.23
extracted	0.50	0.67	0.57	0.19	0.25	0.22
all	0.50	0.74	0.60	0.20	0.29	0.24

Table 6.2: Subjectivity and sentiment classification results

The classifier extensions failed to improve performance. The Score difference technique did not improve performance of either of the seed lists. The Zone difference approach is hardly applicable to sentence-based classification as most of the sentences consist of a very small number of zones.

An error analysis showed that the most important factor that influenced performance in the subjectivity classification task was the proportion of subjective sentences in a topic. For example, in topics 07, 13 and 16, in which more than 60% of sentences are subjective, the classifier performed well, achieving precision of about 0.70 and recall of about 0.40–0.60. However, on those topics that have a small proportion of subjective sentences (topics 08 and 11 have less than 30% of sentences that are subjective) performance was very poor.

This means that the classifier tends to produce too many false positives. This can be explained by the fact that was designed to process collections of subjective documents and tries to increase the number of documents classified.

The accuracy of sentiment classification is also affected by the performance of subjectivity classification. For the best-performing 6-seed classifier the correlation between the precision of subjectivity classification and sentiment classification is 0.64 which is usually considered to be strong. For example, for the best three topics the sentiment classification accuracy was 0.54–0.59, however for the worst two it was 0.21 and 0.22.

English

The English subjectivity classification used seeds presented in Table 6.3. The results feature low precision (however, not the lowest compared to some supervised systems in NTCIR-7) but rather high recall. Despite a high F-measure value, precision was about the level of the naïve baseline (P = 0.25, R = 0.68, F = 0.36).

In sentiment classification, the unsupervised classifier performed relatively well (P = 0.18, R = 0.32, F = 0.23), given that many of the supervised systems tested in the workshop performed poorly in terms of both precision and recall (with precision ranging from 0.03 to 0.50 and recall from 0.02 to 0.55).

Corpus	Seeds
Positive seeds	great, strong, important, popular, clean, easily, pleased, convincing, proud, profitable, attractive
Negative seeds	sad, difficult, weak, poor, critical, dangerous, tough, pessimistic, ashamed, afraid, expensive, disgraceful, traumatic, risky

Table 6.3: Semi-automatically extracted English seeds

Japanese

The experiments with the Japanese corpus required extraction of seeds which was not a trivial procedure, owing to the specific structure of the language. The seed word extraction technique used for the English and Chinese languages would not work for Japanese because negation is usually expressed only at the very end of a sentence, so does not mark the position of a possible seed. Thus negation is a bad indicator of opinion-bearing words. A

quick analysis of the use of Japanese adjectives suggested the use instead of two kinds of indicators: prepositional and post-positional. The first group consists of three items: より, 最も, 最. The first is an indicator of the comparative case and is often followed by an adjective; the other two are adverbs meaning "the most". To find a possible end of an adjective I used the particle い which is often used at the end of adjectives.

This approach produced a very small list of seed candidates of which I chose three positive seeds and four negative ones. The positive seeds were 良 *good*, 好 *fine*, 安定 *stability*, 美 *beautiful*, and the negative seeds were 難 *difficult*, 困難 *difficulty*, 悪 *evil*, 遅 *to retard*.

The subjectivity classification results were better than a naïve baseline (0.27) by only several percentage points, reaching precision 0.31 with 0.85 recall, which is the worst precision and the highest recall compared to the supervised systems at NTCIR-7. The precision of the supervised systems ranged from 0.31 to 0.81, and recall from 0.09 to 0.73.

Sentiment classification performance was also quite low with 0.10 precision and 0.09 recall. However, increasing zone difference threshold increased performance up to 0.75 precision and 0.68 recall which is much better than any other system. This improvement is due to the high proportion of neutral sentences in the Japanese corpus (about 86%), and since a higher zone difference produces more neutral classifications, it boosted performance.

6.2.3 Discussion

The classifier failed to produce acceptable results in sentence-based processing. This was due to a number of reasons including: small amounts of data preventing the extraction of useful seed vocabulary; and iteration control that is aimed at classification of as many items as possible, which results in a lot of false positive results as the corpora contain large proportions of factual data. These experiments not only confirm the difficulty of classifying sentiment and subjectivity in a combined process, but also show that the sentiment classifier is effective only if applied to a priori subjective texts (positive, negative or neutral).

6.2.4 Standalone Subjectivity Classification

The next set of experiments tests an unsupervised subjectivity classifier. The approach to subjectivity classification follows similar principles to the sentiment classifier described previously.

To determine whether a sentence is subjective, I used a semi-automatically generated

list of words which are considered to be indicators of subjectivity. Knowing that such indicators are domain- / topic-dependent, I first tried to derive lists of words specific to each topic. However, poor results in preliminary experiments suggested that none of the topic-specific sub-corpora in any of the four languages was large enough, so I merged all the topics together. The candidate list of subjectivity indicating words was created as follows. First, for each frequently occurring word, I found its immediate neighbours (words occurring either immediately before and after). Then for each word and neighbour, I calculated the χ^2 score; neighbours for which $\chi^2 > 3.84$ were retained and sorted in decreasing order of χ^2 score, and the others discarded. Words having similar sets of neighbours might be semantically close. However, I wanted to avoid words that are related syntactically and not semantically, which I filtered out by considering first-order co-occurrence. For example, assume words A, B and C, have neighbours as follows:

Word	Neighbours
A	X Y Z
B	A Y Z
C	B Y Z

The input corpus must have contained the string **AB** or **BA** (since **A** has been observed in the immediate context of **B**). Similarly, **BC** is also a first-order co-occurrence. On the other hand, **A** and **C** are probably related semantically rather than syntactically since there is no first-order co-occurrence and both appear in the context of **Y** and **Z**. So the pairs **AB** and **BC** are filtered out as syntactic, and **AC** remains as probably being semantic.

To estimate the degree of semantic association, I calculated a score S between every remaining pair of words, measuring the similarity of neighbours:

$$S = \sum \frac{1}{r} \tag{6.2}$$

where the sum is over the neighbours present in both neighbour lists, and r is the rank of a neighbour in the list of the first word. The word pairs were then filtered to leave only those with the highest associations. I used two filters. The first one filtered out all pairs with S less than $\bar{\chi} - 1.96\sigma$. The second filter deleted all words that occurred unusually often (threshold $\bar{\chi}+1.96\sigma$); such words are often function words without any task-relevant value. Finally I was left with a list of pairs of words that were highly semantically associated.

Chinese (Traditional)	Chinese (Simplified)	Japanese	English
難 (difficult)	太 (too)	難 (difficult)	important
功 (effort)	比 (compare)	激 (strike)	difficult
害 (damage)	最 (the most)	貧 (poor)	effective
感 (feeling)	强 (strong)	悲 (bad luck)	popular
好 (good)	欢 (welcome)	困難 (difficulty)	successful
才 (only)	好 (good)	良 (good)	easily
最 (the most)	良 (fine)	可能 (possibly)	troubled
太 (too)	可能 (possibly)	戦闘 (fighting)	striking
利 (luck)	善 (good)	深刻 (deeply)	best
效 (relatively)	害 (damage)	焦点 (disadvantage)	bad
利用 (make use of)	难 (hard)	犠牲 (sacrifice)	painful
認為 (suppose)	压力 (pressure)	強 (string)	strong
最 (the most)	紧 (tight)	最 (the most)	good
	强 (strong)	悪 (evil)	
	恐 (fear)	汚 (dirty)	

Table 6.4: Manually-selected opinionated words (all glosses are very approximate).

Subjective Word Selection

From the list of pairs of associated words I selected those words which are relevant to the task of subjectivity classification. Unfortunately, I was not able to devise an automatic technique of separating subjectivity markers from other words. Instead, I looked through the lists, manually selecting those words that looked most relevant to the task. In all, I spent less than one hour doing this for each language. Table 6.4 shows the lists of selected words, and Table 6.5 gives the numbers of words in the original and final lists. As I do not know any Japanese, I relied mostly on a dictionary when selecting Japanese words (although my knowledge of Chinese characters helped a lot). If I had known Japanese, I would undoubtedly have produced a better list. I also did not investigate which features are really relevant for subjectivity classification in any of the languages (for example, markers of modality, tense or aspect). Further work on these issues would be likely to lead to better results.

After the list of subjectivity markers was derived, it was applied to the corpus. If a sentence contained at least one of these words, it was classified as subjective. In the

Language	Automatically generated list	Number of selected words
Chinese (Traditional)	1154	13
Chinese (Simplified)	494	15
Japanese	491	15
English	1363	13

Table 6.5: Sizes of the lists of words.

overall results, this system is called NLCL-1. The NLCL-1 system in general achieves high precision but low recall. In order to improve recall I tried two ways of expanding the list of manually-selected subjectivity markers. The first way included all words that were associated with the manually selected subjectivity markers (system NLCL-3). An alternative method included only those words whose association score was higher than the arithmetic mean for this list (system NLCL-2). As an example, the list for the English NLCL-2 system was:

> *active, advanced, analysts, common, developed, developing, difficult, easily, economists, effective, frequent, grave, hotel, immediate, important, likely, long, nino, notably, obvious, optimistic, played, popular, possess, primary, recently, robust, scientists, striking, successful, supervision, surprising, they will be, threaten, troubled, urgent, vital, vulnerable*

6.2.5 Evaluation Results

Traditional and Simplified Chinese For the Chinese relevance and opinion sub-tasks (see Tables 6.6 and 6.7), the results are the lowest of all the systems presented at NTCIR-7, although not by much. More encouragingly, though, the results for each of the systems on the two sets of Chinese sub-tasks are numerically better than the results obtained for the other two languages.

Japanese I originally entered only a single system for the Japanese language sub-tasks, NLCL-1 (which uses just the manually selected list of 13 subjectivity markers). After the official submission, I also tested system NLCL-3 (which uses the manual list plus all associated words), to investigate whether the gains in recall would outweigh expected decreases (see Table 6.8).

	Sub-task	Precision (%)	Recall (%)	F-value
NLCL-1				
Lenient	Relevance	84.9	14.5	24.8
	Opinion	53.6	26.8	35.7
Strict	Relevance	92.4	18.0	30.1
	Opinion	62.6	29.3	39.9
NLCL-2				
Lenient	Relevance	86.4	28.6	43.0
	Opinion	49.4	50.6	50.0
Strict	Relevance	93.0	34.1	49.9
	Opinion	60.1	52.5	56.1
NLCL-3				
Lenient	Relevance	85.7	41.1	56.6
	Opinion	47.6	74.2	58.0
Strict	Relevance	92.8	48.5	63.7
	Opinion	58.3	74.1	65.3

Table 6.6: Relevance and opinion results for Chinese (Traditional).

	Sub-task	Precision (%)	Recall (%)	F-value
NLCL-1				
Lenient	Relevance	96.3	32.6	48.7
	Opinion	44.3	39.9	42.0
Strict	Relevance	97.4	33.3	49.6
	Opinion	38.6	40.2	39.2
NLCL-2				
Lenient	Relevance	97.5	28.0	43.5
	Opinion	48.2	36.9	41.8
Strict	Relevance	98.5	28.5	44.1
	Opinion	44.3	39.0	41.4
NLCL-3				
Lenient	Relevance	97.1	58.5	73.0
	Opinion	43.2	69.9	53.4
Strict	Relevance	98.3	59.0	73.7
	Opinion	36.7	70.6	48.3

Table 6.7: Relevance and opinion results for Chinese (Simplified).

	Sub-task	Precision (%)	Recall (%)	F-value
NLCL-1				
Lenient	Relevance	53.7	18.9	28.0
	Opinion	42.6	22.3	29.3
Strict	Relevance	30.1	21.1	24.8
	Opinion	31.4	22.6	26.3
NLCL-3				
Lenient	Relevance	47.7	63.8	54.6
	Opinion	30.2	91.0	45.3
Strict	Relevance	22.7	61.1	33.1
	Opinion	22.2	91.9	35.8

Table 6.8: Relevance and opinion results for Japanese.

	Sub-task	Precision (%)	Recall (%)	F-value
NLCL-1				
Lenient	Relevance	13.0	6.8	9.0
	Opinion	37.8	10.1	16.0
Strict	Relevance	5.3	8.5	16.0
	Opinion	11.7	10.5	11.1
NLCL-2				
Lenient	Relevance	17.5	14.4	15.8
	Opinion	33.8	18.6	24.0
Strict	Relevance	7.4	18.8	10.7
	Opinion	10.9	20.1	14.1
NLCL-3				
Lenient	Relevance	48.2	68.9	56.7
	Opinion	27.7	84.6	41.7
Strict	Relevance	16.4	72.7	26.8
	Opinion	8.4	86.1	15.3

Table 6.9: Relevance and opinion results for English.

English For the English language tasks, the NLCL-3 system performed well, delivering excellent results compared to other systems in the relevance subtask, under both lenient and strict scoring (see Table 6.9). In the opinion sub-task, NLCL-3 is in the third quartile.

6.2.6 Discussion

The results vary widely across the four languages: for the Japanese and English sub-tasks I obtained results which compare favourably with other systems, whereas in both Simplified and Traditional Chinese, the system performed poorly in comparison with the other systems – although my results were numerically superior to those obtained for the other languages. At this point, it is not clear why the system's performance varies so much across the languages, and in particular why the system performed comparatively less well on the Chinese data. One possible explanation is that the Chinese data is more homogeneous and so more tractable for competing approaches based on supervised machine learning. Another possibility is that the corpora were not very comparable (e.g. the number of subjective clauses differs significantly across the corpora), together with annotation

approaches which might be different across languages as a result of many factors, starting from different standards accepted in research groups that were doing the annotations and ending with a culture-specific (hence language-specific) understanding of subjectivity. Nevertheless, the system was not far behind other systems even in the Chinese language tests, so it achieved some success as a knowledge-poor portability-oriented system.

6.3 Opinion Holder and Target Extraction[6]

The final set of experiments tests an unsupervised approach to the task of opinion extraction, specifically the extraction of opinion holders and opinion targets. As discussed in Section 2.2.1, an opinion may have a holder (a person or a group that expresses the opinion) and a target (the object that is being discussed or evaluated). To explore if the research paradigm used in this study can be applied to opinion holder and target extraction, I use a knowledge-poor language-independent approach with some simple linguistic typology, as advocated by Bender (2009).

The opinion holder and opinion target extraction system described below consists of two major parts: a core system implementing a general approach to the extraction task, and a small set of language-specific extensions. The approach is based on the assumption that opinion holders and opinion targets are words or phrases which are topic-related and tend not to occur in other topics. A further assumption is that a language has markers of subjectivity and surface clues which can be used to find syntactic subjects. This set of assumptions together with a small amount of language-specific information constitutes the minimal task-related language description.

6.3.1 Overview of the Approach

The first assumption, that opinion holders and opinion targets are topic-related (with the exception of pronouns and generic phrases such as *our correspondent*)[7], requires that the system first finds topical words – words that are strongly related to the topic of a given text. In order to minimise language-specific input (such as word lists or automatic segmenters), I use the same basic unit as in previous studies – the lexical unit (see Section 4.2.2).

[6]The experiments and part of the discussion in this section were presented in a condensed form at the Language and Technology Conference (Zagibalov and Carroll, 2009)

[7]This is a purely empirical assumption. A better version could be to define a topic by 'holder – target' pairs, but this would be too restrictive for the relatively small corpus in these experiments.

Of course, the resulting list of extracted LUs contains a lot of noise. This problem is dealt with by filtering out those items that occur in too many different topics. Such items are filtered out on the basis of the number of different topics in which they occur. For the experiments described here, only one threshold was used: a LU is regarded as a topical LU if it is used in no more than 50% of the topics. This technique filters out most topic-irrelevant units. A preliminary investigation with lower thresholds showed that some potential holders may occur in many different topics (e.g. *President Bush*) so a higher threshold would significantly reduce coverage.

The next step is to find only those sentences that are subjective. The easiest way to do this is to use a lexical subjectivity marker (e.g. the word *said* in English). Attempts to automatically find such markers (usually they are words that introduce indirect speech), despite some success, turned out to be very complex and not particularly reliable, while making a list of such words (and extending it) is a very trivial task even for a person who does not know the language well.

Having a list of topic-relevant lexical units and a set of sentences that have been identified as subjective, the system then finds out which topic-relevant lexical items in these sentences are opinion holders and which are targets. To do this, the system uses a 'subject marker', a LU that denotes a subject in a sentence. This marker is language-dependent and for English and Chinese it is the same as a subjectivity marker, but for Japanese it is not. The relative position of a holder (subject) and a marker (predicate) is also a language-dependent feature which the system uses for finding holders. After opinion holders are identified, these LUs are removed from the list of topic-related LUs, and the remainder used to find opinion targets in the sentences. I make the assumption here that documents (news items) should be consistent on what a holder and a target are. Having found the lists of opinion holders and opinion targets, it is likely that there are other subjective sentences that were not found with the subjectivity marker, so I used the newly found holders and targets as a further set of subjectivity markers. Thus, all sentences that contain any of these words are assumed to be subjective, and opinion holders and targets are extracted from all of them. If a sentence contains a target, but a holder was not found, then the holder is tagged as 'AUTHOR'.

6.3.2 Language-specific Adjustment

The system described above cannot be used without any adjustment to the language being processed. First of all, to find noun phrases that could be holders or targets, it needs

to have well-formed lexical units, which implies finding word delimiters (such as space in English). This can be done automatically by counting the relative number of space symbols in the document collection: for English documents the number of space symbols will be very high, whereas it will be close to zero in Chinese and Japanese. Once it has such a delimiter it can form proper lexical items for English: meaningless sequences like *prose*, *rosec*, *cutor* and such like are eliminated, but the valid *prosecutor* is preserved as it occurs with delimiters (space or punctuation) on both sides. This task is more difficult for the Chinese and Japanese languages (it may require trimming out function words that 'stick' to the words within LUs). For further processing it is more important to find if there is such a delimiter as a space to avoid malformed phrases in English (or any other languages where words are separated by a space).

Another piece of language-specific information is the minimal LU length. This is not a particularly important parameter, but to save some time on filtering out 1-letter 'word-candidates' from a list of English lexical units, the minimum LU length was set to 4 letters. This variable was set to 2 for Chinese, and 3 for Japanese[8].

As outlined above, the system needs a list of subjectivity markers to find subjective sentences. The system uses the word *said* for English, the unit 说 (*say*, *says*, *said*) for Chinese, and for Japanese と言う, という, 言, 話, and 話し (which are equivalents of the English *said*). There is only one word for English and Chinese because in preliminary experiments I found that adding synonyms did not improve performance for either of these languages: the synonyms are too infrequent in the corpus used, as are modal verbs. However, since I do not know Japanese, I could not decide which of the words is the most important and left all of them in the list as they were found in an electronic dictionary.

Once subjective sentences are found, the system needs to find an opinion holder; this is assumed to be the subject of a sentence. Fortunately, the subjectivity markers for English and Chinese are verbs, and verbs in these languages are usually quite close to nouns denoting subjects. This allows for reuse of these words as subject markers. To find the opinion holder, the system finds the lexical item closest to the marker. It also considers the relative position of the holder: in English, the subject denoting the speaker can usually be found before the verb (as in *John said ...*), but the inverted construction (*..., said John*) can also be found in some genres of text. In Chinese, the corresponding verb-noun construction is almost impossible, so I had to adjust the extraction rule accordingly:

[8]These values are empirical trade-offs between the average length of words in a language and the number of candidate lexical items that could potentially be extracted.

(4) 布什说，政府可能还会采取更广泛的振兴经济措施。
((President) Bush said that ⋯)

The Japanese language is quite different from English and Chinese in its syntactic structure: it is a SOV (subject-object-verb) language. This means that the Japanese marker (the equivalent of *said*) cannot be near a holder (which is assumed to be a subject). However, there is a special function word in Japanese (は wa) that denotes the topic of a sentence which in conjunction with equivalents of *said* may often locate an opinion holder. So a simple rule finds a holder near and before this marker:

(5) 長崎大の谷川教授は支配層がコントロール能力を失えば「最悪の場合、スリランカのような内乱状態にならない 保証はない」と話す。
(Prof. Tanikawa from Nagasaki (University) said that ...)

6.3.3 System Summary

To summarise, the system performs the following steps:

1. Find lexical items.
2. Filter out noisy (not topic-relevant) lexical items.
3. Find all subjective sentences.
4. Find an opinion holder near a subject marker.
5. Find opinion targets.
6. Extract all found holders and targets from all sentences.

Language-specific information that is required is:

1. Word delimiter (can be found automatically)
2. Word-length (not critical, mostly for better performance)
3. Subjectivity marker (the word *said* and its equivalents, such words also can be found (semi-) automatically)
4. Subject marker (the same as in point 3 for English and Chinese, and the function word *wa* for Japanese)[9]

[9]This is a language-dependent information: for some languages (Slavic, Turkic) it could be morphological units, rather than lexical ones.

5. The relative position of a subject (usually before the marker in English, and always before in Chinese and Japanese).

As can be seen from this summary, the approach requires little language-specific information.

6.3.4 Experiments

The Gold Standard

The holder and target extraction experiments used the NTCIR-7 MOAT test data collections: English, Simplified Chinese and Japanese. The Simplified Chinese data as supplied by the task organisers had been annotated by twelve annotators, and all topics were annotated by three of them. The English data was annotated using a pool of six annotators. The same approach was taken for Japanese annotation. The gold standard authors provided two versions of the data: strict and lenient. The gold standard contains all variants of holders/targets that the annotators came up with (Seki et al., 2008).

Approximate Matches

Each test uses the standard NTCIR-7 MOAT evaluation metrics, consisting of precision, recall and F-measure (F1). Each test measures the number of correct matches, when a string (holder or target) extracted by the system exactly matches the one stored in the gold standard file. Since it is not always possible even for a human annotator to establish the exact boundaries of a string expressing target or holder, the evaluation script additionally counts all approximate matches. There are three kinds of such matches: *superstring*, *substring* and *overlap*.

A **superstring** is a string which is longer than the gold standard string and incorporates the latter entirely, for example:

(6) Gold standard: “don rodbell”

(7) System proposed: “mr don rodbell”

A **substring** is a shorter string that exactly matches part of a gold standard string:

(8) Gold standard: “former nuremberg prosecutor said’

(9) System proposed: “former nuremberg prosecutor”

An **overlap** of two strings is a substring that is present in both strings, but is not an exact match of either:

(10) Gold standard: "igor ivanov"

(11) System proposed: "mr ivanov"

The approximate matches described above may produce a lot of noise, matching, for example, short function words or phrases with a long string from the gold standard that also contains such words. To avoid this and to reduce the number of false positives, I set a limit of how different in length matching strings can be. For *superstring* and *substring* the shorter one should be at least half of the length of the longer one. For *overlapping* strings, the length of the shared part should at least one-third of the combined length of the two strings. For example: for the overlapping strings ABCD and BCDY, the overlapping part should be at least 2.6 characters long:

$(ABCD.length + BCDY.length)/3 = (4 + 4)/3 = 2.6$, so since $BCD.length = 4$, ABCD and BCDY is a valid approximate match. Manual inspection of the approximate matches indicated that the vast majority of approximate match strings are valid opinion targets or opinion holders.

Results

The results are summarised in Table 6.10, for holder and target identification in each of the three languages, English, Simplified Chinese and Japanese. Figures in brackets are results for approximate matches, which, as argued above, are reliable indicators of system performance. The low performance is rather typical for the task even for supervised monolingual systems presented at NTCIR-7. Nonetheless, the approach described here may form the basis for applications in web-based information retrieval where results can be aggregated and ranked.

Comparision

These results are numerically fairly low, but opinion holder and target extraction are very difficult tasks. The results compare reasonably well to those reported by the participants of the NTCIR-7 MOAT workshop, but in general are not the best. This can be expected since all of those systems were supervised, and moreover monolingual.

Specifically, there were 12 systems entered in the MOAT Chinese opinion holder extraction task. The system would have ranked 9th in terms of F1 (and 7th with respect

	Language	P	R	F1
holder	English	0.19 (0.28)	0.09 (0.13)	0.12 (0.18)
holder	Chinese	0.18 (0.24)	0.17 (0.22)	0.17 (0.23)
holder	Japanese	0.16 (0.16)	0.56 (0.56)	0.25 (0.25)
target	English	0.02 (0.16)	0.01(0.06)	0.01 (0.09)
target	Chinese	0.03(0.08)	0.03 (0.07)	0.03 (0.07)
target	Japanese	0.03 (0.08)	0.10 (0.25)	0.05 (0.13)

Table 6.10: Opinion holder and target performance on the NTCIR-7 MOAT test sets. Exact matches and approximate matches (in parentheses).

to approximate match): the best system's F-measure was 0.46, the worst was 0.02, and the macro-average for all systems was 0.19. In contrast, for opinion target extraction, the system would have been 2nd (1st) out of five submissions.

Only two systems extracted opinion holders in the English side of NTCIR-7, and the results obtained by this system would not have outperformed either of them. This can be attributed to the difference in evaluation approaches: at NTCIR-7 the English results were evaluated in a semi-automatic mode where if an automatic fuzzy match did not find any matching string, a human judge decided whether a string was an acceptable match. Obviously, the automatic evaluation cannot be as flexible and intelligent as a human judge, so a lot of potentially good output from the system was tagged as incorrect by the evaluation script.

Unfortunately, there were no submissions of opinion holder and target extraction systems for Japanese at NTCIR-7, which makes it impossible to compare the system with any others. But since the results are in line with those for the other languages, I assume that the results for Japanese are reasonable. It should be noted that most of the holders in the Japanese collection were tagged as 'AUTHOR', resulting in high recall, which might reflect the usual (impersonal) way of expressing opinion in the Japanese language.

6.3.5 Discussion

It is obvious that in principle it would be difficult for a knowledge-poor, unsupervised approach to outperform the best supervised (or knowledge-based) systems. But judging from the experiments presented in this section, it is possible to conclude that a system

which needs only very basic language-specific adjustments (minimal language description), may perform reasonably well. The previous section noted that a cross-lingual unsupervised system was being compared to monolingual supervised systems. A definitive study would involve comparison with supervised systems on a cross-lingual task.

Error Analysis

There are two possible types of errors: 1) a holder or a target is not present in a sentence in the gold standard, but the system "finds" them; and 2) a holder or a target is present in the gold standard, but the system proposes incorrect strings as holder or target. The majority of such errors are caused by the system finding too many candidate strings, many of which consist of function words, such as *but that cannot* (a system proposed holder). These errors could easily be eliminated by a list of stop-words applied to the candidate strings. Many mistakes were caused by lack of anaphora resolution, which led to too frequent use of pronouns as opinion holders (which was usually considered to be a mistake). One of the most widespread errors for target extraction was an inability to find correct boundaries of a target phrase. In preliminary experiments, I used the whole target subsentence (the remaining part of the sentence after an extracted holder) as the target. This approach produced much more appropriate and legible target strings, but such strings were too long compared with the correct targets.

From manual inspection of data, opinion holders seem to have a simpler structure than targets. This makes target extraction much more difficult. The complex structure of opinion targets also means that it is possible for different notions of 'target' to exist. Indeed, it is arguable which of the following variants of the same target is the most appropriate: *Russia and China* or *Non-status quo powers* or *Non-status quo powers, most notably Russia and China*? Should one incorporate all or any (which?) attributes into the target? Or should annotators tag only the shortest noun phrase without any attributes? This ambiguity might explain why results for target extraction are so low. The complex structure of opinion targets makes consistent tagging difficult: for example, the English gold standard seems to be less consistent, as in some cases annotators tagged only minimal noun sequences as targets but also frequently tagged long substrings as targets, for example:

(12) humanitarian intervention (along with cases of self-defense) has been made an exception from the general condemnation on the use of force when interfering in the domestic affairs of another state

Long strings such as this are difficult to extract using only topic words. The Chinese corpus annotators were more consistent, mostly tagging only the shortest noun phrases, which may explain the big difference between exact and approximate results for English.

6.4 Conclusion

This chapter showed that knowledge-poor approaches can be applied to a range of sentiment analysis tasks more complex than binary sentiment classification. These tasks include three-way sentiment classification, sentence-level subjectivity classification and opinion mining. The experiments also showed an advantage of a scale-based classification over binary (discrete) classification, in that it allows more flexible definition of classes and provides more information about classification instances. Sentence-level subjectivity classification experiments confirmed the ineffectiveness of a combined sentiment and subjectivity classification approach. A separate, almost unsupervised subjectivity classifier, however, performed well and this suggests that an unsupervised approach can also be applied to this task. Opinion mining is a difficult task even for supervised systems, but an unsupervised approach using only minimal task-relevant language descriptions proved to perform comparably to supervised systems.

Chapter 7

Conclusion

This Chapter summarises the research results presented in this thesis and proposes some possible directions for future work.

7.1 Unsupervised Sentiment Classification

The main contribution of the work presented in this thesis is the development of an unsupervised, knowledge-poor approach for sentiment analysis that is capable of domain-independent sentiment classification, as well as of sentiment classification in different languages. The approach does not require training data, large sets of rules or sentiment lexicons and is able to collect all the data required for classification from documents to be classified. The only input the classifier needs is a small number of seeds (up to six) labelled with their sentiment (either positive or negative). For some tasks (such as opinion mining), however, the approach may need some task-relevant language-specific information. This approach was implemented in a classifier described in Chapter 3. The classification is done by means of a classification score, which is based on relative frequencies of a lexical unit in positive and negative documents. The classifier used different units of classification: unigrams, zones and sentences. Zones, (subsentence unit consisting of a sequence of characters separated by punctuation) appeared to provide the best classification quality: capturing more context than a lexical unit, a zone is not as long as a sentence that may contain different sentiments.

This study also tests different kinds of features for the task of Chinese sentiment classification. The experiments show that the best performance can be achieved by combining information relating to dictionary items (words and phrases) and separate characters. This finding leads to the idea of a more universal notion of a basic unit. Rather than using

linguistic units (not always well defined) such as the character, the word or the phrase, I used a *lexical unit*, a sequence of characters that occurs at least twice in a corpus. Being a data driven unit, the lexical unit does not correspond to the traditional notion of part of speech and may be a part of word, a word or even a phrase. Lexical units are sub-sentence units, because they are extracted from zones as described in Chapter 4. This chapter also introduces a number of extensions to the sentiment classifier.

The first extension is iterative classification controlled by an iteration control system. Iterative classification allows for bootstrapping a list of domain-specific lexical units that performs better than a generic list of sentiment terms on large data sets. The iterative approach proved to be a highly effective means of increasing the performance of a sentiment classifier, significantly increasing recall without a large impact on precision. Iteration control stops iterations as soon as no more documents can be classified for three subsequent iterations. This technique proved to work well, stopping the classifier at an iteration with one of the best results, on big corpora, where a classifier cannot classify all documents. However, if a classifier manages to classify all or nearly all documents, the iteration control is obviously useless.

Another extension is sentiment score difference. This technique compares sentiment scores of opposite sentiments of each word. If the difference is smaller than a threshold, the word is considered less discriminatory and excluded from classification process. The technique helps increase performance of the classifier by eliminating lexical items that cannot contribute to classification accuracy. The performance of this technique was high on all of the test corpora, but its actual utility depends on the iteration control's ability to stop at the iteration that produced the best classification.

The final extension is zone difference. This technique ranks classified documents according to the difference between zones tagged as positive and ones tagged as negative. The larger the difference, the higher the precision (at the expense of recall), and the more accurate classification results are. This extension may be useful for opinionated information retrieval or similar applications, in which precision is more important than recall. However, it has a limitation: it does not work well if most documents are short and consist of a small number of zones. Test corpora with shorter documents failed to benefit from this technique, however a corpus of movie reviews, featuring very long reviews, showed very high sensitivity to the technique.

7.2 Other Tasks

The unsupervised approach was also applied to other sentiment analysis tasks: three-way sentiment classification, document- and sentence-level subjectivity classification, and opinion holder and opinion target extraction. The three-way classification, as described in Chapter 6.1 adds a neutral sentiment class. The approach is based on the unsupervised classifier and uses scale-based classification rather than a traditional binary (positive – negative) approach. The scale-based classification regards sentiment as a continuum stretching from positive to negative, and attempts to locate each document accurately on this continuum. The classifier relies on the zone difference to define a document position on the continuum: the more positive or negative the zone difference is, the more extreme position a document is placed at. Overall, scale-based classification appears to be a promising paradigm for sentiment classification. It is also possible to add a fourth class: objective documents (ones that do not express any sentiment, neutral included), by calculating sentiment density, the proportion of zones that express sentiment out of the total number of zones in a document. However, the classifier, being developed for sentiment classification, did not perform well in the task of subjectivity classification. Better performance on this task requires additional information about the class of objective documents (i.e. an appropriate list of terms). An attempt at standalone sentence-level combined opinion and subjectivity classification was not successful either. This suggests that sentiment classification and subjectivity classification are two distinct tasks.

An unsupervised approach was also applied to multilingual opinion holder and opinion target extraction. This task required a system, different from the one developed for sentiment classification. Although different the system was also developed within the unsupervised knowledge-poor paradigm and was based on a limited number of language-specific extraction rules. The system's performance was quite poor in absolute terms, however it compared well with a number of supervised techniques run by others on the same data set. This suggests that the unsupervised knowledge-poor approach may be a a viable alternative to supervised techniques in different aspects of sentiment analysis, especially for cross-domain real-time applications or for under-resourced domains / languages.

7.3 Cross-domain Sentiment Classification

The unsupervised sentiment classifier was tested on different domains: customer reviews (split into 10 different topics), film reviews, book reviews and news. The classifier achieved

very good results on larger data sets, sometimes even outperforming supervised classifiers. Smaller collections of documents performed poorly, because the classifier was unable to extract reliable markers of sentiment from them. Moreover, several customer review collections in Chinese were not topically homogeneous, being made up from the reviews of different (but related) products. This resulted in a diversity of terms used for product appraisal, which in combination with the small size of these collections, resulted in data sparseness.

Another type of problem affected the performance of the sentiment classifier applied to the movie review corpus. Despite being large, the results obtained on this corpus were not good. The domain of films reviews is known to be difficult for sentiment analysis owing to the complexity of language used in it. The professional style of writing adopted by film review authors features a wide variety of means of expressing sentiment. Another distinctive feature of this domain is the abundance of non-sentiment related text in reviews (e.g. descriptions of plots). All these result in data sparseness, preventing the classifier from finding sentiment bearing lexical items.

Results from the superficially similar domain of book reviews were better than for movie reviews. The most important features of the book reviews which helped the classifier gain about twenty percentage points over the naïve baseline, are short (compared to the Movie reviews) reviews and more simple language. Short reviews are more focused on evaluation and tend to use a simpler vocabulary for expressing sentiments.

7.4 Multilingual Sentiment Classification

The approach was tested on Chinese (Simplified and Traditional), English, Russian and Japanese and proved to be effective without any adjustment or modification (Chapter 5). The sentiment classifier was applied to Chinese (Simplified), Russian and English with only seeds changed, all other parameters remaining the same. In an opinion mining task, the relevant unsupervised classifier used a very limited task-related language description which included only a small set of markers and and information about the relative syntactic positions of objects and subjects. The performance on all the corpora achieved a satisfactory level being about 20 percentage points above a naïve baseline (50%). It is difficult to compare the performance of the classifier across the languages, because the corpora used for testing were not parallel.

Further experiments were carried out on comparable book review corpora in Russian and English. The performance of the classifier on these two corpora was different, with the

Russian corpus being 5 – 9 percentage points inferior. This difference in performance may be attributed to language-specific features, including grammar, pragmatics and lexicon. However, it is difficult to separate out the influence of each type of linguistic feature on the performance.

7.5 Hypotheses

The hypotheses stated in Chapter 1 were mostly supported by experimental data.

Hypothesis 1: *Unsupervised systems can be developed for performing sentiment analysis in different domains and in different languages that perform comparably with supervised systems.*

The research supports this hypothesis, but only for large datasets: it is possible to use unsupervised, knowledge-poor system for sentiment analysis in different domains and in different languages, but the performance of such a system to a great extent depends on data quality. Small, not very homogeneous datasets prevent the system from achieving performance comparable to the performance of supervised systems.

Hypothesis 2: *Unsupervised and knowledge-poor sentiment analysis may not require much domain- or language-specific input. Such a system might require only a basic indication of what positive and negative sentiments are, in the form of lexical 'seeds'.*

The experimental results support this hypothesis. The system does not require much domain- or language-specific input, being able to perform using as little as only two seeds.

Hypothesis 3: *A sentiment-related vocabulary automatically extracted from a corpus can produce similar or better results compared to a specialised hand-built sentiment vocabulary.*

The system with features that were automatically extracted from corpus performed significantly better compared to the manually created lists of sentiment indicators. So this hypothesis is supported.

Hypothesis 4: *An automatically acquired training corpus in conjunction with machine learning techniques can produce sentiment classification results similar or close to a standard supervised approach.*

This hypothesis is not fully supported by the experimental results. Machine learning-based system trained only on the automatically extracted data did not perform well.

Hypothesis 5: *A uniform notion of 'lexical unit' can be used across languages for*

sentiment analysis tasks

Lexical units proved to be useful for different languages, especially for those languages that either do not have explicit word boundaries (Chinese, Japanese) or have a very complex word structure (e.g. Russian). So this hypothesis is supported. It should be noted, however, that for other languages (e.g. English) the advantage of using lexical units may not be so evident.

7.6 Future Work

The unsupervised approach described in this thesis is based on seed lexical units. The experiments showed that the choice of seeds has a strong effect on the performance of the classifier. It is possible to find the seeds automatically (Section 4.2.2), but the seed finding technique seems to be very language-dependent. In an attempt to minimise the impact of generic (out-of-topic) seeds on classifier, I experimented with the minimum sets of seeds (2 seeds: 1 seed for each sentiment) which performed well, although worse than larger seed lists, especially ones extracted from the corpus. This makes the task of seed word extraction and/or selection one of the main directions of future work. I would like to experiment with semi-automatic techniques for seed selection and test the performance of bigger seed lists. One possibility for improving seed-selection could be extraction of adjectives associated with topical words (nouns, that are more frequent in a given corpus). This list of adjectives may be processed by a Turney-like technique (Turney, 2002), which measures the similarity of the adjectives with known sentiment words.

Another direction for further research is combining sentiment analysis with subjectivity analysis. It has been observed in several studies (e.g. Pang and Lee, 2004; Wiebe et al., 2004) that subjectivity classification may help improve the performance of sentiment analysis. However, the experiments in this thesis confirm a conclusion made by Esuli and Sebastiani (2006a) that sentiment classification and subjectivity classification are separate tasks and simultaneous subjectivity and sentiment analysis does not work well. Thus, for practical applications that have to deal with a mixture of objective and subjective documents, it might be beneficial to run a subjectivity classifier to exclude subjective documents. This could be done using the scale-based approach introduced in this thesis. Another possible improvement could be to exclude all factual (objective) zones from documents, before sentiment classification. Wiebe et al. (2004) showed that leaving only subjective portions of document helped increase accuracy of a sentiment classifier. This technique was partially implemented by the zone-difference technique at the

document-level, although this technique significantly reduced recall.

In all of the experiments in this study I used lexical units as the basic unit of processing because they are data-driven (extracted from the data set to be processed) and language-independent. Lexical units may consist of one or two parts of words, be an actual single word, consist of a word and part of other word, or consist of several words. Obviously, these types have different lengths and different frequencies. It would, therefore, be interesting to find out which kinds of lexical unit perform better and why. LUs may have different performance in different languages as compared with other units (characters, words and phrases) and it would be useful to know what (kinds of) languages benefit more from using these units. Answering all these questions would require a lot of research in different languages with different types of linguistic units. One of the first steps could be a more detailed study of the impact of the length of LUs on performance. A preliminary investigation done in this thesis did not show much influence on performance in Chinese sentiment classification, however, the influence of the LU's length on performance needs more experimental results to be able to make any firm conclusion. LUs might be especially useful for processing languages with complex morphology, and experiments with a Russian book review corpus showed the efficacy of LUs. However it is too early to conclude that this type of unit would be equally useful for other such languages (Turkish, Czech and others).

I believe that the scale-based sentiment and subjectivity classification may have a very big potential. Scale-based sentiment classification, as already mentioned, treats sentiment classification not as a binary classification problem, but as a problem of locating documents on a continuum stretching from extreme negativity to extreme positivity. This conception of the problem follows the dimensional paradigm introduced by Osgood et al. (1971). Experiments in these areas would require a special corpus that can be used to test the accuracy of placement of documents (or other units) on a sentiment scale. The corpus should follow the dimensional paradigm and use an appropriate annotation scheme. The development of such scheme, a prerequisite for development of the corpus, would require a significant research effort.

7.7 Practical Implementation

The research presented in this thesis was inspired by the idea of a designing an approach that can be free of most of the problems of domain-dependency and language-dependency. Such an approach could make possible a number of practical applications that are time-

and data-critical and cannot depend on a slow and expensive process of development of training data, rule-sets or lexicons (the approach, however, could also be a first step in development of such resources). Sentiment analysis is a task that is domain- and language-dependent and may benefit from the kind of an approach described above.

One of the possible applications based on the unsupervised, knowledge-poor approach may be opinionated information retrieval. This would be based on a search engine capable of real-time retrieval of information that contains some appraisal (negative or positive) of different products, events or personalities. One cannot predict all possible topics of queries and produce training datasets for supervised systems or rule-sets and lexicons for knowledge-based systems. The zone difference technique provides a suitable means of ranking the results.

Another application of the approach could be sentiment analysis in under-resourced languages. An example of such a language is Russian, which does not have any sentiment-related research corpora or any other relevant resources. The experiments with the Russian book review corpus presented in this thesis showed that application to Russian is possible and may produce useful results.

Real-time sentiment information monitoring may be useful for marketing departments in companies that are interested in how their customers perceive their products or services. A language-independent approach would make it possible to monitor different national markets and the absence of domain-dependency would allow a system to follow twists of language use that occurs in real-life human communication, (for example emerging new topics of discussion, different styles of language, and new colloquial words and phrases that are different to foresee).

Bibliography

Ahmed Abbasi, Hsinchun Chen, and Arab Salem. Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums. *ACM Transactions on Information Systems (TOIS)*, 26(3):12, 2008.

Steven Abney. Bootstrapping. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 360–367, Philadelphia, Pennsylvania, 2002.

Rakesh Agrawal, Sridhar Rajagopalan, Ramakrishnan Srikant, and Yirong Xu. Mining newsgroups using networks arising from social behavior. In *Proceedings of the Twelfth International Conference on World Wide Web - WWW '03*, pages 529–536, New York, NY, 2003.

Edoardo Airoldi, Xue Bai, and Rema Padman. Markov Blankets and Meta-heuristics search: sentiment extraction from unstructured texts. *Advances in Web Mining and Web Usage Analysis*, 3932:167–187, 2006.

Tatiana Akimova and A. Maslennikova. *Lingvisticheskie issledovanija*, chapter Semantika imperativa i ocenka (Semantics of imperatives and appraisal), pages 3–33. Moscow, 1987.

Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. Emotions from text: Machine learning for text-based emotion prediction. In *Proceedings of HLT/EMNLP*, pages 579–586, Vancouver, Canada, 2005.

Rie Kubota Ando and Tong Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853, 2005.

Anthony Aue and Michael Gamon. Customizing sentiment classifiers to new domains: a case study. In *Proceedings of RANLP*, 2005.

Carmen Banea, Rada Mihalcea, and Janyce M Wiebe. A Bootstrapping method for building subjectivity lexicons for languages with scarce resources. In *Proceedings of the Language Resources Evaluation Conference (LREC)*, pages 2764–2767, Marrakech, Morocco, 2008a.

Carmen Banea, Rada Mihalcea, Janyce M Wiebe, and Hassan Samer. Multilingual subjectivity analysis using machine translation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 127–135, Waikiki, Honolulu, Hawaii, 2008b.

Ann Banfield. *Unspeakable Sentences: Narration and Representation in the Language of Fiction*. Routledge & Kegan Paul Ltd., London, 1982.

Marco Baroni and Vegnaduzzo Stefano. Identifying subjective adjectives through web-based mutual information. In *Proceedings of the 7th Konferenz zur Verarbeitung Natürlicher Sprache (German Conference on Natural Language Processing – KONVENS'04)*, pages 613–619, Vienna, 2004.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In *Advances in Neural Information Processing Systems*, volume 20, page 137. MIT Press, Cambridge, MA, USA, 2007.

Emily M. Bender. Linguistically naive!= language independent: why NLP needs linguistic typology. In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*, pages 26–32, 2009.

John Blitzer, Ryan McDonald, and Fernando Pereira. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 120–128, 2006.

John Blitzer, Mark Dredze, and Fernando Pereira. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics.*, pages 440–447, Prague, Czech Republic, 2007.

Kenneth Bloom, Navendu Garg, and Shlomo Argamon. Extracting appraisal expressions. In *Proceedings of NAACL HLT*, pages 308–315, Rochester, NY, 2007.

Erik Boiy and Marie-Francine Moens. A machine learning approach to sentiment analysis in multilingual Web texts. *Information Retrieval*, 12(5):526–558, 2008.

Erik Boiy, Koen Deschacht, Marie-Francine Moens, and Pieter Hens. Automatic sentiment analysis in on-line text. In *Proceedings ELPUB2007 Conference on Electronic Publishing*, pages 349–350, Vienna, Austria, 2007.

Samuel Brody and Noemie Elhadad. An unsupervised aspect-sentiment model for online reviews. In *Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 804–812, Los Angeles, California, 2010.

Hsin-Hsi Chen, Li-Ying Lee, Lun-Wei Ku, and Tung-Ho Wu. Major topic detection and its application to opinion summarization. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 627–628, Salvador, Brazil, 2005.

Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. Identifying sources of opinions with conditional random fields and extraction patterns. In *Proceedings of HLT/EMNLP 2005*, pages 355–362, Vancouver, 2005.

Sanjiv R. Das and Mike Y. Chen. Yahoo! for Amazon: Sentiment extraction from small talk on the web. *Management Science*, 53(9):1375–1388, 2007.

Kushal Dave, Steve Lawrence, and David M. Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th International Conference on Information and Knowledge Management*, pages 519 – 528, Budapest, Hungary, 2003.

Stephen D. Durbin, J. Neal Richter, and Doug Warner. A system for affective rating of texts. In *Proceedings of the 3rd Workshop on Operational Text Classification at the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Washington, DC, 2003.

Miles Efron. Cultural orientation: Classifying subjective documents by cocitation analysis. In *Proceedings of the 2004 AAAI Fall Symposium on Style and Meaning in Language, Art, Music, and Design*, Washington, DC, 2004.

Koji Eguchi and Victor Lavrenko. Sentiment retrieval using generative models. In *Pro-*

ceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, pages 345–354, Sydney, 2006.

Paul Ekman and Wallace V. Friesen. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17(2):124–129, 1971.

Charlotte Engström. *Topic Dependence in Sentiment Classification.* MPhil dissertation, Computer Laboratory, University of Cambridge, 2004.

Brian Eriksson. Sentiment classification of movie reviews using linguistic parsing. In *Final Project Report*, volume 838 of *Final Project Report*. University of Wisconsin, 2006.

Andrea Esuli and Fabrizio Sebastiani. Determining the semantic orientation of terms through gloss classification. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, pages 617–624, New York, New York, USA, 2005.

Andrea Esuli and Fabrizio Sebastiani. Determining term subjectivity and term orientation for opinion mining. In *Proceedings of the 11th Meeting of the European Chapter of the Association for Computational Linguistics (EACL-2006)*, volume 2, pages 193–200, 2006a.

Andrea Esuli and Fabrizio Sebastiani. SentiWordNet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC*, 2006b.

Andrea Esuli and Fabrizio Sebastiani. PageRanking WordNet synsets: an application to opinion mining. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, number 1, pages 424–431, Prague, Czech Republic, 2007.

Oren Etzioni, Michael Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, and Alexander Yates. Unsupervised named-entity extraction from the web: An experimental study. *Artificial Intelligence*, 165(1):91–134, 2005.

Zhongchao Fei, Jian Liu, and Gengfeng Wu. Sentiment classification using phrase patterns. In *Proceedings of the 4th International Conference on Computer and Information Technology (CIT2004)*, pages 1147–1152, Wuhan, China, 2004.

Olga Feiguina and Guy Lapalme. Query-based summarization of customer reviews. In *Advances in Artificial Intelligence*, pages 452–463. Springer, Berlin / Heidelberg, 2007.

Schubert Foo and Hui Li. Chinese word segmentation accuracy and its effects on information retrieval. *TEXT Technology*, pages 1–11, 2001.

Michael 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*, pages 841–847, Geneva, Switzerland, 2004.

Michael Gamon and Anthony Aue. Automatic identification of sentiment vocabulary: exploiting low association with known sentiment terms. In *Proceedings of the ACL Workshop on Feature Engineering for Machine Learning in Natural Language Processing*, pages 57–64, Ann Arbor, Michigan, 2005.

Michael Gamon, Anthony Aue, Simon Corston-Oliver, and Eric Ringger. Pulse: Mining customer opinions from free text. *Lecture Notes in Computer Science*, 3646:121–132, 2005.

Gayatree Ganu, Noemie Elhadad, and Amelie Marian. Beyond the stars: improving rating predictions using review text content. In *Proceedings of the Twelfth International Workshop on the Web and Databases*, Providence, Rhode Island, USA, 2009.

Anindya Ghose, Panagiotis Ipeirotis, and Arun Sundararajan. Opinion mining using econometrics: A case study on reputation systems. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, volume 45, page 416, Prague, Czech Republic, 2007.

Stephan Charles Greene. *Spin: lexical semantics, transitivity, and the identification of implicit sentiment*. PhD thesis, University of Maryland, 2007.

Stephan Charles Greene and Philip Resnik. More than words. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 503–511, Morristown, NJ, USA, 2009.

Seth Grimes. The three secrets to successful sentiment analysis, 2010. URL `http://www.mycustomer.com/topic/customer-intelligence/seth-grimes-how-get-sentiment-analysis-right/103102`. Last accessed on 2010-02-16 22:11:34.

Jin Guo. *Chinese Language Modeling for Speech Recognition.* PhD thesis, National University of Singapore, 1997.

Vasileios Hatzivassiloglou and Kathleen R McKeown. Predicting the semantic orientation of adjectives. In *Proceedings of the Eighth Conference of European Chapter of the Association for Computational Linguistics*, volume pages, pages 174–181, Madrid, Spain, 1997.

Rumjahn Hoosain. *Psycholinguistic implications for linguistic relativity: A case study of Chinese.* Lawrence Erlbaum Associates Inc, Mahwah, NJ, 1991.

Mingqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 168–177, Seattle, WA, USA, 2004.

Jungi Kim, Hun-young Jung, Sang-hyob Nam, Yeha Lee, and Jong-Hyeok Lee. English opinion analysis for NTCIR7 at POSTECH. In *Proceedings of the NTCIR-7 Workshop Meeting*, pages 241–246, Tokyo, Japan, 2008.

Soo-min Kim and Eduard H. Hovy. Determining the sentiment of opinions. In *Proceedings of the 20th International Conference on Computational Linguistics*, volume 4, pages 1367–1373, Geneva, Switzerland, 2004.

Soo-min Kim and Eduard H. Hovy. Identifying and analyzing judgment opinions. In *Proceedings of the HLT/NAACL*, pages 200–207, New York, NY, 2006.

Nozomi Kobayashi, Kentaro Inui, Yuji Matsumoto, Kenji Tateishi, and Toshika Fukushima. Collecting evaluative expressions for opinion extraction. In *Proceedings of the IJCNLP*, pages 596–605, Heidelberg, 2004.

Moshe Koppel and Jonathan Schler. The importance of neutral examples for learning sentiment. *Computational Intelligence*, 22(2):100–109, 2006.

Lun-Wei Ku, Tung-Ho Wu, Li-Ying Lee, and Hsin-Hsi Chen. Construction of an evaluation corpus for opinion extraction. In *Proceedings of the Fifth NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access*, Tokyo, Japan, 2005.

Lun-Wei Ku, Hsiu-Wei Ho, and Hsin-Hsi Chen. Novel relationship discovery using opinions mined from the web. In *Proceedings of the 21st National Conference on Artificial intelligence*, pages 1357–1362, 2006a.

Lun-Wei Ku, Yu-Ting Liang, and Hsin-Hsi Chen. Opinion extraction, summarization and tracking in news and blog corpora. In *Proceedings of AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs*, pages 100–107, 2006b.

Lun-wei Ku, Yu-ting Liang, and Hsin-hsi Chen. Question analysis and answer passage retrieval for opinion question answering systems. In *Proceedings of the 19th Conference on Computational Linguistics and Speech Processing*, pages 177–190, Taipei, Taiwan, 2007a.

Lun-wei Ku, Yong-sheng Lo, and Hsin-hsi Chen. Using polarity scores of words for sentence-level opinion extraction an Chinese opinion system: CopeOpi Extraction. In *Proceedings of the 6th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access*, pages 316–322, Tokyo, Japan, 2007b.

Tao Li, Yi Zhang, and Vikas Sindhwani. A non-negative matrix tri-factorization approach to sentiment classification with lexical prior knowledge. In *Proceeding of the 47th Annual Meeting of the Association for Computational Linguistics*, pages 244–252, Morristown, NJ, USA, 2009.

Wei Li. On Chinese parsing without using a separate word segmenter. *Communications of COLIPS*, 10(1):17–67, 2000.

Nanyuan Liang. A written Chinese automatic word segmentation system. *Journal of Chinese Information Processing*, 1(2):44 – 52, 1987.

Bing Liu, Minqing Hu, and Junsheng Cheng. Opinion observer: Analyzing and comparing opinions on the web. In *Proceedings of the 14th International Conference on the World Wide Web*, pages 342–351, 2005.

James R. Martin and Peter R. R. White. *The language of evaluation: Appraisal in English.* Palgrave Macmillan, 2005.

Ryan McDonald, Kerry Hannan, Tyler Neylon, Mike Wells, and Jeff Reynar. Structured models for fine-to-coarse sentiment analysis. In *Annual Meeting of the 45th Assosiation of Computational Linguistics*, pages 432–440, Prague, Czech Republic, 2007.

Rada Mihalcea and Hugo Liu. A corpus-based approach to finding happiness. In *Proceedings of the AAAI Spring Symposium on Computational Approaches to Weblogs*, 2006.

Rada Mihalcea, Carmen Banea, and Janyce M Wiebe. Learning multilingual subjective language via cross-lingual projections. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 976–983, 2007.

Tetsuya Nasukawa and Jeonghee Yi. Sentiment analysis: Capturing favorability using natural language processing. In *Proceedings of the 2nd International Conference on Knowledge capture*, pages 70–77, Sanibel Island, Florida, USA, 2003.

Jian-Yun Nie, Jiangfeng Gao, Jian Zhang, and Ming Zhou. On the use of words and n-grams for Chinese information retrieval. In *Proceedings of the fifth international workshop on Information retrieval with Asian languages*, pages 148–156, 2000.

Andrew Ortony, Gerald L. Clore, and Allan Collins. *The cognitive structure of emotions*. Cambridge University Press, Cambridge, UK, 1988.

Charles Egerton Osgood. *Focus on Meaning: Explorations in semantic space*. Mouton Publishers, The Hague, The Netherlands, 2nd edition, 1976.

Charles Egerton Osgood, George J. Suci, and Percy H. Tannenbaum. *The measurement of meaning*. University of Illinois Press, Champaign, IL, 8th edition, 1971.

Charles Egerton Osgood, William H. May, and Murray S. Miron. *Cross-cultural universals of affective meaning*. University of Illinois Press, Urbana, 1975.

Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on Minimum Cuts. In *Proceedings of the 42nd Annual Meeting of the Association of Computational Linguistics*, Barcelona, Spain, 2004.

Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 115–124, Ann Arbor, Michigan, 2005.

Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2008.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 79–86, 2002.

Erik Peterson. A Chinese Named Entity Extraction System. In *Proceedings of the 8th Annual Conference of the International Association of Chinese Linguistics*, Melbourne, Australia, 1999.

Ana-Maria Popescu and Oren Etzioni. Extracting product features and opinions from reviews. In *Natural Language Processing and Text Mining*, pages 9–28, Vancouver, Canada, 2005.

Rudy Prabowo and Mike Thelwall. Sentiment analysis: A combined approach. *Journal of Informetrics*, 3:137–157, 2009.

Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. *A Comprehensive Grammar of the English language*. Pearson Longman, Harlow, 1985.

Jonathon Read. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In *Proceedings of the ACL Student Research Workshop*, pages 43–48, Ann Arbor, Michigan., 2005.

Jonathon Read. *Weakly-Supervised Techniques for the Analysis of Evaluation in Text*. DPhil thesis, University of Sussex, 2009.

Jonathon Read and John Carroll. Weakly supervised techniques for domain-independent sentiment classification. In *Proceeding of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion*, pages 45–52, 2009.

Jonathon Read, David Hope, and John Carroll. Annotating expressions of appraisal in English. In *Proceedings of the Linguistic Annotation Workshop (ACL)*, pages 93–100, Prague, Czech Republic, 2007.

Ellen Riloff and Janyce M Wiebe. Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pages 105–112, 2003.

Ellen Riloff, Janyce M Wiebe, and Theresa A. Wilson. Learning subjective nouns using extraction pattern bootstrapping. In *Proceedings of the seventh Conference on Natural language learning at HLT-NAACL*, pages 25–32, 2003.

Ellen Riloff, Janyce M Wiebe, and William Phillips. Exploiting subjectivity classification to improve information extraction. In *Proceedings of the 20th national Conference on Artificial intelligence*, volume 20, pages 1106–1111, 2005.

Klaus R. Scherer and Angela Schorr. *Appraisal Processes in Emotion: Theory, Methods, Research.* Oxford University Press, Canary, NC, 2001.

Yohei Seki. A Multilingual Polarity Classification Method using Mult-label Classification Technique Based on Corpus Analysis. In *Proceedings of the NTCIR-7 MOAT Workshop Meeting*, pages 284–291, Tokyo, Japan, 2008.

Yohei Seki, David K. Evans, Lun-Wei Ku, Le Sun, Hsin-Hsi Chen, and Noriko Kando. Overview of multilingual opinion analysis task at NTCIR-7. In *Proceedings of the NTCIR-7 MOAT Workshop Meeting*, pages 185–203, Tokyo, Japan, 2008.

Pavel Smrz. Using WordNet for opinion mining. In *Proceedings of the Third International WordNet Conference*, pages 333–335, Brno, Czeck Republic, 2006.

Philip J. Stone, Dexter C. Dunphy, and Marshal S. Smith. *The general inquirer: A computer approach to content analysis.* MIT Press Cambridge, MA, 1966.

Veselin Stoyanov and Claire Cardie. Topic identification for fine-grained opinion analysis. In *Proceedings of the 22nd International Conference on Computational Linguistics*, pages 817–824, 2008.

Pero Subasic and Alison Huettner. Affect analysis of text using fuzzy semantic typing. *IEEE Transactions on Fuzzy Systems*, 9(4):483–496, 2001.

Aoshuan Tan. *Problemy skrytoj grammatiki (Issues if Hidden Grammar).* Yazyki Slavjanskoy Kultury, Moscow, 2002.

Matt Thomas, Bo Pang, and Lillian Lee. Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. In *Proceedings of the 2006 Conference on empirical methods in natural language processing*, pages 327–335, Sydney, Australia, 2006.

Ivan Titov and Ryan McDonald. Modeling online reviews with multi-grain topic models. In *Proceedings of the 17th International Conference on the World Wide Web*, pages 111–120, Beijing, China, 2008.

Chih-Hao Tsai. *Word identification and eye movements in reading Chinese: A modeling approach.* PhD thesis, University of Illinois at Urbana-Champaign, 2001.

Peter D. Turney. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of Assosiation of Computational Linguistics*, pages 417–424, Philadelphia, Pennsylvania, 2002.

Peter D. Turney and Michael L. Littman. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems-TOIS*, 21:315–346, 2003.

Hideo Watanabe, Hiroshi Kanayama, and Tetsuya Nasukawa. Deeper sentiment analysis using machine translation technology. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 494–452, Geneva, Switzerland, 2004.

David Watson and Auke Tellegen. Towards a consensual structure of mood. *Psychological Bulletin*, 98, 1985.

Casey Whitelaw, Navendu Garg, and Shlomo Argamon. Using appraisal taxonomies for sentiment analysis. In *Proceedings of MCLC-05, 2nd Midwest Computational Linguistic Colloquium*, Columbus, USA, 2005a.

Casey Whitelaw, Navendu Garg, and Shlomo Argamon. Using appraisal groups for sentiment analysis. In *Proceedings of the 14th ACM International Conference on Information and knowledge management*, pages 625–631, 2005b.

Janyce M Wiebe. Tracking point of view in narrative. *Computational Linguistics*, 20(2): 233–287, 1994.

Janyce M Wiebe. Learning Subjective Adjectives from Corpora. In *Proceedings of the 17th National Conference on Artificial Intelligence*, 2000.

Janyce M Wiebe and Rada 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, Sydney, Australia, 2006.

Janyce M Wiebe and Ellen Riloff. Creating subjective and objective sentence classifiers from unannotated texts. In *Proceedings of the Sixth International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2005)*, pages 486–497, Mexico City, Mexico, 2005.

Janyce M Wiebe, Theresa A. Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. Learning subjective language. *Computational linguistics*, 30(3):277–308, 2004.

Janyce M Wiebe, Theresa A. Wilson, and Claire Cardie. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2):165–210, 2005.

Theresa A. Wilson. *Fine-grained Subjectivity and Sentiment Analysis: Recognizing the intensity, polarity, and attitudes of private states.* PhD thesis, University of Pittsburgh, 2008.

Theresa A. Wilson and Janyce M Wiebe. Annotating opinions in the world press. In *Proceedings of the 4th SIGdial Workshop on Discourse and Dialogue (SIGdial-03)*, pages 13–22, 2003.

Theresa A. Wilson, Janyce M Wiebe, and Rebecca Hwa. Just how mad are you? Finding strong and weak opinion clauses. In *Proceedings of the National Conference on Artificial Intelligence*, pages 761–769, 2004.

Theresa A. Wilson, Janyce M Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 354–362, 2005.

Theresa A. Wilson, Janyce M Wiebe, and Paul Hoffmann. Recognizing Contextual Polarity: an exploration of features for phrase-level sentiment analysis. *Computational Linguistics*, 35(3):399–433, 2009.

Ian H. Witten and Eibe Frank. *Data Mining: Practical machine learning tools and techniques.* Morgan Kaufmann, San Francisco, 2nd edition, 2005.

Pak-kwong Wong and Chorkin Chan. Chinese word segmentation based on maximum matching and word binding force. In *Proceedings of the 16th International Conference on Computational linguistics*, pages 200–203, Copenhagen, Denmark, 1996.

Jia Xu, Richard Zens, and Hermann Ney. Do we need Chinese word segmentation for statistical machine translation. In *Proceedings of the Third SIGHAN Workshop on Chinese Language Processing*, pages 257–264, Boston, MA, 2004.

Nianwen Xue. Chinese word segmentation as character tagging. *Computational Linguistics and Chinese Language Processing*, 8(1):29–48, 2003.

Hong Yu and Vasileios Hatzivassiloglou. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. *Proceedings of EMNLP*, pages 129–136, 2003.

Raymond W.M. Yuen, Terence Y.W. Chan, Tom B.Y. Lai, O.Y Kwong, and Benjamin K.Y. T'sou. Morpheme-based derivation of bipolar semantic orientation of

Chinese words. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 1008–1016, Geneva, Switzerland, 2004.

Taras Zagibalov. Kinds of features for Chinese opinionated information retrieval. In *Proceedings of the ACL Student Research Workshop*, pages 37–42, Prague, Czech Republic, 2007a.

Taras Zagibalov. Sentiment zones approach to extracting a training corpus in unsupervised sentiment classification. In *Proceedings of the Eurolan Doctoral Consortium*, pages 21–29, Iasi, Romania, 2007b.

Taras Zagibalov and John Carroll. Unsupervised classification of sentiment and objectivity in Chinese text. In *Proceedings of the Third International Joint Conference on Natural Language Processing*, pages pp. 304–311, Hyderabad, India, 2008a.

Taras Zagibalov and John Carroll. Automatic seed word selection for unsupervised sentiment classification of Chinese text. In *Proceedings of the 22nd International Conference on Computational Linguistics*, pages 1073–1080, Manchester, United Kingdom, 2008b.

Taras Zagibalov and John Carroll. Almost unsupervised cross language opinion analysis at NTCIR 7. In *Proceedings of the NTCIR-7 MOAT Workshop Meeting*, pages 204–210, Tokyo, 2008c.

Taras Zagibalov and John Carroll. Multilingual opinion holder and target extraction using knowledge-poor techniques. In *Proceedings of Language and Technology Conference*, Poznañ, Poland, 2009.

Taras Zagibalov, Katerina Belyatskaya, and John Carroll. Comparable English-Russian book review corpora for sentiment snalysis. In *Proceedings of the 1st Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, Lisbon, Portugal, 2010.

Changli Zhang, Daniel Zeng, Qingyang Xu, Xueling Xin, Wenji Mao, and F.Y. Wang. Polarity classification of public health opinions in Chinese. In *Intelligence and Security Informatics*, volume 5075 of *Lecture Notes in Computer Science*, pages 449–454, 2008.

Zhu Zhang and Balaji Varadarajan. Utility scoring of product reviews. In *Proceedings of the 15th ACM International Conference on Information and knowledge Management*, pages 51–57, 2006.

Chao Zhou, Guang Qiu, Kangmiao Liu, Jiajun Bu, Mingcheng Qu, and Chun Chen. SOPING: a Chinese customer review mining system. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and development in information retrieval*, pages 741–742, Singapore, 2008.

Li Zhuang, Feng Jing, and Xiao-Yan. Zhu. Movie review mining and summarization. In *Proceedings of the 15th ACM International Conference on Information and knowledge management*, pages 43–50, Arlington, Virginia, USA, 2006.