Universitat d'Alacant
Universidad de Alicante

WASSA? 2010

Proceedings of the 1st Workshop on Computational Approaches to Subjectivity and Sentiment Analysis

Held in conjunction to

Lisbon
ECAI
2010
PORTUGAL

FOREWORD

A central point in the quality of communication using natural language is effectiveness, i.e. the extent to which the meaning conveyed through the message reaches the audience and is interpreted in the same manner as it was intended. Rhetorics, as the art of using language to communicate effectively, is based on three pillars: *logos* (the use of words), *ethos* (appealing to common concepts) and *pathos* (appealing to the emotions of the audience). It is thus of no wonder that the fields where communication effectiveness is vital, such as politics, advertisement or marketing, employ in their text the persuasive power of emotions.

Affect expression is not only important because of its argumentative power. It also helps to create a more natural, human way of interaction. This is why affect modelling has also become an important research theme in Artificial Intelligence, within the framework of Affective Computing. Research in this area is currently applied to Human-Computer Interaction, intelligent agents modelling and Robotics.

Research in automatic Subjectivity and Sentiment Analysis, as subtasks in Affective Computing within Natural Language Processing, has flourished in the past years, as the Social Web made it possible for people all over the world to express, comment or consult opinions on any given topic. The fact that so many people express themselves on these topics makes opinions less biased and more credible; their subjective nature makes them easily understandable by all people and leads to their growing influence on communities worldwide. Due to all these reasons, opinions expressed on the Web are more and more considered as basis for decision-making processes, for recommendation systems, business intelligence processes, image monitoring, marketing or for obtaining unbiased, massive feedback. And the list of applications for such data can go on, each with their impact on social, economical or psychological aspects.

Bearing in mind the abovementioned reflections, the main aim of this multidisciplinary workshop was to bring together researchers in Computational Linguistics who are working on Subjectivity and Sentiment Analysis, but also from other disciplines related to this area, such as psychologists, sociologists, economists etc. The objective was to facilitate an interdisciplinary dialogue on the analysis, requirements, issues and applications of the study of subjectivity and sentiment in the context of traditional and emerging text types.

The first edition of the Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2010) is a satellite workshop to the 19th European Conference on Artificial Intelligence - ECAI 2010 -, and it will be celebrated on the 17th of August 2010, in the Faculty of Science, University of Lisbon, in Lisbon, Portugal.

For this first edition of WASSA we received a total of 18 submissions, from Austria, Belgium, Bulgaria, France, Germany, Greece, India, Italy, Portugal, Spain, Switzerland, Tunisia, United Kingdom and United States, of which 10 were accepted, thus resulting in an acceptance rate of 0,56. Each paper has been reviewed by an average of 3 specialists. The accepted papers were all highly assessed by the reviewing committee; the best paper received a punctuation of 3/3 from two of the reviewers and 2/3 from the third reviewer. The predominant topics of the accepted papers are real-life applications, comparisons between approaches on different text types, in a monolingual or multilingual setting, joint topic and sentiment analysis, creation, annotation and evaluation of resources.

We would like to thank the ECAI 2010 Organizers for the help and support at the different stages of the workshop organization process. We are also especially grateful to the Program Committee members and the external reviewers for the time and effort spent assessing the papers and our invited speaker, for accepting to deliver the keynote talk.
Last, but not least, we would like to express our gratitude for the partial funding given by The Spanish Ministry of Science and Education of the Spanish Government (Ministerio de Ciencia e Innovación - Gobierno de España) through the grant no. TIN2009-13391-C04-01, and to the Education Council of the Valencia Community (Conselleria d’Educación - Generalitat Valenciana) through the grant no. PROMETEO/2009/119 and ACOMP/2010/286.

Organisers (Co-chairs):
Andrés Montoyo - University of Alicante, Spain
Patricio Martínez-Barco - University of Alicante, Spain
Alexandra Balahur - University of Alicante, Spain
Ester Boldrini - University of Alicante, Spain

Invited speaker:

Prof. Michael Thelwall - University of Wolverhampton, U.K.

Program Committee:

Eneko Agirre - University of the Basque Country, Spain

Khurshid Ahmad - Trinity College Dublin, Ireland

Nicoletta Calzolari - CNR Pisa, Italy

Jesús M. Hermida - University of Alicante, Spain

Veronique Hoste - University of Ghent, Belgium

Mijail Kabadjov - EC-Joint Research Centre, Italy

Mosche Koppel - Bar-Ilan University, Israel

Zornitsa Kozareva - Information Sciences Institute California, U.S.A.

Rafael Muñoz - University of Alicante, Spain

Günter Neumann - DFKI, Germany

Constantin Orasan - University of Wolverhampton, U.K.

Manuel Palomar - University of Alicante, Spain

Viktor Pekar - University of Wolverhampton, U.K.

Massimo Poesio - University of Trento, Italy

Paolo Rosso - Polytechnic University of Valencia, Spain

Ralf Steinberger - EC- Joint Research Centre, Italy

Veselyn Stoyanov - Cornell University, U.S.A

Carlo Strapparava - FBK, Italy

Hristo Tanev - EC- Joint Research Centre, Italy

Mike Thelwall - University of Wolverhampton, U.K.

José Antonio Troyano - University of Seville, Spain

Dan Tufis - RACAI, Romania

External reviewers

Gozde Ozbal, FBK, Italy

Michael Wiegand, University of Saarland, Germany

TABLE OF CONTENTS

Sentiment strength detection in short informal text

Mike Thelwall[1]

Abstract. An increasing proportion of human communication takes place via short informal text messages. These can be part of exchanges, as is typically the case for instant messaging, social network site (SNS) comments, internet chatrooms, online forums and mobile phone text messages. In contrast, they may be one offs, as in web site customer feedback comments, or of a more hybrid nature, such as twitter posts or comments on blog posts, both of which may attract responses or may be attempts to broadcast thoughts. Analysing sentiment in short informal text is therefore important not only for traditional commercial applications but also for the social science purposes of understanding interpersonal communication and new internet communication styles. This paper argues that it is useful to measure sentiment strength rather than just polarity and subjectivity in this context, argues that even short messages can be usefully analysed for positive and negative sentiment simultaneously, and describes a sentiment strength classification algorithm for short informal text.

1 SHORT INFORMAL TEXT

Short informal text is loosely defined here as up to one paragraph of text intended as an individual message or part of an exchange and in a context where the norms of standard grammar and spelling do not apply or are widely flouted. The increase in this form of communication primarily reflects widespread adoption of new electronic interpersonal communications technologies, such as those listed above, and the ability to post comments on anything from YouTube videos and Flickr photographs to blog posts and news stories, in addition to the more hidden online customer feedback facilities. Common, but not universal, features of such texts include a lack of concern for grammar and spelling (by definition), standard and improvised abbreviations (e.g., lol, xox (kiss hug kiss)), local dialect and slang, and emoticons. Another feature is the partial transference of shorthand between technologies. Such text provides an opportunity for sentiment analysis because of the ease of access to huge amounts in some cases (e.g., twitter) and because sentiment is important in interpersonal communication. Hence the results may be relevant to social science goals such as understanding successful interpersonal communication and identifying typical communication patterns and strategies. Nevertheless, sentiment analysis poses challenges due to frequent disobeying of linguistic rules and the non-standard methods of encoding sentiment discussed below.

2 SENTIMENT STRENGTH

Sentiment strength has been ignored in most previous sentiment analysis studies on the basis that the most important tasks are to differentiate between subjective and objective text and between polarities in subjective text. Moreover, to some extent it is possible to measure the collective strength of feeling via the proportion of positive and negative texts. Nevertheless this is clearly insufficient for individualised analyses involving sentiment strength. As an example of this, in order to assess the impact of the strength of feeling expressed in interpersonal communication on the longevity of online friendship or debates, sentiment strength measures would be needed.

3 SIMULTANEOUS POSITIVE AND NEGATIVE SENTIMENT STRENGTH

Simultaneous positive and negative sentiment strength measures are those that give separate positive and negative sentiment measures to individual texts. In emotion psychology, although there is a tradition of separating positive and negative sentiment (i.e., valence), it is recognised that this is an oversimplification and that it is common to experience a combination of positive and negative sentiment. This may be due to fluctuations over time or to events triggering opposing signals. In short informal texts a common example of this is I miss you which has positive connotations expressing warmth towards the recipient and a negative surface meaning that the sender is feeling bad. Another type of example is My leg was also in agony but rapidly got better which conveys a positive message and expresses empathy by discussing a past strongly negative episode. As an example of a natural social science hypothesis requiring simultaneous positive and negative measures, is mixing sentiment polarity a more engaging communication strategy than being universally positive?

4 THE SENTISTRENGTH ALGORITHM

An algorithm designed for simultaneous positive and negative sentiment strength detection in short informal text, SentiStrength, classifies texts as positive on an integer scale of 1 (no positive sentiment) to 5 (strong positive sentiment) and negative on a similar scale of 1 to 5. It applies simple linguistic rules without relying upon correct grammar. For instance, negating words are recognised and used to invert the polarity of subsequent emotion-bearing terms but there is no attempt to use tagging to semantically disambiguate terms such as like. The hear of the system is a dictionary of common emotion-bearing terms pre-coded by human annotators with their expected sentiment strengths as commonly used. It is complemented by shallow linguistic features, such as negating and boosting terms, as well as a range of text emotion-related expressive features. The latter includes emoticons, the use of repeated letters for emphasis (e.g., haaaaapy) and the use of excess punctuation (e.g., happy birthday!!!!!). Experiments with this approach indicate that it outperforms a range of machine learning approaches for positive but not for negative sentiment strength on a corpus of 1041 MySpace comments.

[1] University of Wolverhampton, UK, email: m.thelwall@wlv.ac.uk

Automatic Identification of Subjectivity in Morphologically Rich Languages: The Case of Arabic

Muhammad Abdul-Mageed[1] and **Mohammed Korayem** [2]

Abstract. As more user-generated content becomes available online, the need for mining that content becomes increasingly critical. One related area that has been witnessing a flurry of research is that of subjectivity and sentiment analysis. We report our efforts to annotate a corpus of 200 documents from the Penn Arabic Treebank, which is composed of news texts, for subjectivity, along with attempts to automatically classify that data at the sentence level. We investigate the performance of three different machine learning methods on the task with various features and vector settings. We achieve a very high accuracy (i.e., 99.48%) using a support vector machines classifier. We finally briefly discuss issues related to performing text classification on Arabic, a morphologically rich language, and suggest future directions.

1 INTRODUCTION

The web is no longer a static platform where users are passive consumers of content provided by institutions, organizations, or governmental agencies. What we are witnessing is a read-write Web where user-generated content is becoming an integral part of what we now perceive as the Web. Social network sites, online news sites, video- (and other content [e.g., photo]) sharing sites, online forums, blogs, online review sites, among others, are examples of interactive Web platforms where users are more of prosumers (i.e., simultaneously consumers and producers) than mere consumers. This revolutionary change poses both unprecedented challenges as well as great opportunities to researchers in various fields (including natural language processing [NLP], information retrieval [IR], and machine learning [ML]), but equally to business and governmental bodies interested in mining this data for various purposes. One related area where research has been witnessing a flurry of research is that of subjectivity and sentiment analysis where attempts are made to sort out objective (i.e., factual) from subjective (non-factual) information. Non-factual information can be positive, negative, mixed, or neutral. Subjectivity and sentiment analysis is, in general, done in two flavors: (1) *targeted*, where the holders of opinion/sentiment and its targets are identified, and (2) *non-targeted*, where the identification of the opinion/sentiment holders and targets is disregarded. The current research falls in the second category.

Although research on subjectivity and sentiment analysis on some Indo-European languages (e.g., English) is flourishing and has achieved considerable successes, it is not yet clear whether and how comparable success could be achieved on morphologically rich languages (MRLs). The term MRLs refers to languages in which significant information concerning syntactic units and relations is expressed at word-level [16]. MRLs are known to pose many challenges for NLP and IR, and Arabic is a significant case in point (see e.g., [4]). Given that Arabic was judged as the language with fastest growth rate in terms of Internet users in 2009 by Internetworldstats (http://www.internetworldstats.com/stats7.htm), there is a great potential in conducting subjectivity and sentiment analysis on Arabic. In addition, as far as we know, there is neither publicly available Arabic data annotated for subjectivity analysis nor opinion lexicons. Our overarching goal is thus to take one step further toward filling this gap by attracting attention to the problem, hoping to trigger interest in subjectivity and sentiment analysis of Arabic in particular, and MRLs in general.

We have a number of specific goals. First, we wanted to test our simple hypothesis that the challenges that MRLs pose to text classification in general, and subjectivity detection in particular, could be solved if only certain tokens are kept. With a tokenized and part-of-speech (POS) tagged text, this approach is highly successful (as we will show). Second, we wanted to explore the feasibility of employing a number of machine learning techniques (i.e., support vector machines [SVM], Nave Bayes [NB], and Instance-based learning [IB1]) for the task of Arabic subjectivity classification. Although it is by now known that SVMs generally perform well on text classification [8], it is not yet very clear whether or not this will still be the case on the task of Arabic-based subjectivity classification. Neither is it clear how the performance of SVMs compares to other ML algorithms on the same task, and to some extent on Arabic text classification in general. To the best of our knowledge, not many such comparisons have been conducted before on Arabic IR, let alone subjectivity classification. In this way, our work is an attempt to accumulate knowledge on these two fronts. Finally, we wanted to test how each of three ML methods would perform when two different settings (i.e., *frequency* vs. *presence* vectors) are employed. In section 2, we review the related literature. In sections 3 and 4, we present our methods and results, respectively. In our conclusion, section 5, we reflect on the contribution of our research and list a number of future directions.

2 RELATED WORK

Wiebe et al. [18] manually annotated a corpus of 1,001 sentences of the Wall Street Journal Treebank Corpus [11] with subjectivity clas-

[1] Department of Linguistics and School of Library & Information Science, Indiana University, Bloomington, IN-47405 USA, email: mabdulma@indiana.edu

[2] School of Informatics and Computing, Indiana University, Bloomington, IN-47405, USA, email: mkorayem@indiana.edu

sifications by instructing three humans to assign a subjective or objective label to each sentence. Annotators were instructed to consider a sentence to be subjective if they perceived any significant expression of subjectivity (of any source), and to consider the sentence to be objective, otherwise. Wiebe et. al [18] then trained a probabilistic classifier using five POS features, two lexical features, and a paragraph feature and performed 10-fold cross validation. They obtained an average accuracy on subjectivity tagging of 72.17%, more than 20 percentage points higher than a baseline accuracy obtained by always choosing the more frequent class.

Bruce & Wiebe [2] performed a statistical analysis of the assigned classifications in the corpus reported in [18]. The analysis showed that adjectives are statistically significantly and positively correlated with subjective sentences in the corpus on the basis of the log-likelihood ratio test statistic G2. Authors found that probability that a sentence is subjective, simply given that there is at least one adjective in the sentence, was 55.8%, even though there were more objective than subjective sentences in the corpus.

Wiebe [17] added more annotations to the same corpus used in [18]. Specifically, annotators were asked to identify the subjective elements in each subjective sentence and also to rate the strength of the elements on a scale of 1 to 3, with 3 being the strongest. Wiebe [17] then used the subjective elements identified by one judge to seed the distributional similarity [9] of such elements. Lin [9] used distributional similarity on a 64-million corpus consisting of news articles, to create a thesaurus entry for each word consisting of the 200 words of the same part-of-speech that are most similar to it.

Performing 10 fold cross validation, for each adjective of all adjectives with strength 3 extracted from subjective sentences Wiebe [17] identified the top 20 entries in Lins thesaurus entry and used these as the seed sets for each fold. Then she used a simple prediction method for subjectivity: a sentence is classified as subjective if at least one member of a seed set S of adjectives occurs in the sentence, and objective otherwise. Precision is measured by the conditional probability that a sentence is subjective, given that one or more instances of members of S appear. Thus, this metric assesses feature quality: if instances of S appear, how likely is the sentence to be subjective? Wiebe [17] acquired a precision of 61.2% using this method. She also repeated this process, but with the seeds synonyms in WordNet [12] in place of words from Lins thesaurus entry. She found that the performance is slightly better with WordNet (62.0%), but the coverage is lower.

The same method was then repeated using a sample of adjectives extracted by a technique described in [6]. Such a technique depends on identifying the semantic orientation/polarity of adjectives as originally described in [5], but is also enhanced by use of a morphological analyzer for identifying gradable adjectives. Wiebe [17] calls the sets derived in this later method the gradability/polarity sets and the features associated to them the lexical semantic features. She reports 9% improvement in precision when such lexical semantic features are triangulated with similarity clusters described above.

Wiebe & Riloff [19] developed subjectivity classifiers using only unannotated texts for training. More specifically, they use 298,809 unannotated sentences from the world press. The test set used consists of 535 texts (5104 sentences) also from the world press that were annotated for subjectivity and are part of the Multi-perspective Question answering (MPQA) corpus. First, they implement a rule-based classifier that learns a number of previously established subjectivity clues. This classifier achieves 34.2% subjective recall and 90.4% subjective precision, and the rule-based objective classifier achieves 30.7% objective recall and 82.4% objective precision. In an attempt to learn extraction patterns associated with objectivity, the authors use the AutoSlog-TS [14] algorithm since it does not need annotated texts for training, but rather needs relevant and irrelevant texts. The patterns extracted were fed to a classifier that classifies a sentence into subjective if it contains any of the learned subjective patterns, etc. The sentences that were extracted by such a classifier were incorporated into the rule-based classifier. The sentences that were labeled by the rule-based classifier were in turn fed into a nave Bayes classifier. Several set-valued features (e.g., POS features, pronouns, modals [except will], adjectives, cardinal numbers, adverbs [except not], etc.) were used with the nave Bayes classifier. The nave Bayes achieved 70.6% for subjective recall, 79.4% for subjective precision, 74.7% for subjective F-measure. As to objective classification, 77.6% recall, 68.4% precision, and 73.8% F-measure. Finally, authors use self-training methods in that they use the nave Bayes classifier to generate training data. This strategy results in substantial improve in recall and a slight drop in precision.

Riloff et al. [15] explored the idea of using subjectivity analysis to improve the precision of information extraction (IE) systems. They ran their experiments on the MUC-4 information extraction data set (MUC-4 Proceedings 1992). The MUC-4 IE task was to extract information about terrorist events from 1700 stories, mainly news articles about Latin American Terrorism. The MUC-4 set was accompanied by answer key templates containing the information that can be extracted from each story. Riloff et al. [15] used the system developed by [19] for classifying sentences based on subjectivity to extract information from individual sentences and then map that information into the answer key templates. When discarding all extractions that were found in subjective sentences, the authors achieved increased precision of the IE system by 2%, but recall dropped 8%. The authors then concluded that indiscriminately discarding all extractions is very aggressive and that subjective language can co-exist with factual information. Two strategies were followed to improve recall: (1) when a sentence is not strongly subjective and has a source of attribution (e.g., The Associated Press reported . . .) it was extracted and (2) they allowed certain indicator extraction patterns (e.g., *murder of* + *NP* and *NP* + *was assassinated*) to be extracted. These measures raised recall from 44% to 51%.

Pang, et al. [13] conducted a sentiment analysis of a corpus of 752 negative and 1301 positive movie reviews that were taken from the Internet Movie Database (IMDb) archive of the rec.arts.movies.reviews newsgroup (accessible at http://reviews.imdb.com/Reviews/). These authors used three machine learning techniques: Nave Bayes classification, maximum entropy classification, and support vector machines. With each classifier, various feature settings were used. These features included unigrams, unigrams + bigrams, bigrams, unigrams + POS, adjectives, and unigrams + position. For many of these features, the presence or absence of a feature as well as the frequency of a feature were tried. The best accuracy (i.e., 82.9%) was acquired using SVM with only presence or absence of unigram features. Interestingly, the performance of these machine learning techniques turned out to be better than the human-selected-unigram baselines. Also, NB was found to perform worst. Authors concluded with a useful comparison between topic-based classification and sentiment classification, using SVM.

Abbasi et al. [1] performed a sentiment analysis of English and Arabic Web forums, making use of not only syntactic but also stylistic features. The stylistic features they used included (1) letter N-grams (e.g., a, b, c), (2) character N-grams (e.g., ab, abo, etc. in about), (3) word-level lexical features (e.g., total number of words, % of char. per word), (4) character-level lexical features (e.g., total number of characters, % of char. per message) , (5) vocab. richness (e.g., hapax legomena), (6) special characters (i.e., occurrence of special characters e.g., #, $, %, etc.), total number of words, % of char. per word), (7) structural features (e.g., contains greeting, contains url, contains requoted content), etc. These authors used an entropy weighted genetic algorithm (EWGA) as a feature selection technique on (1) an English benchmark movie review database (e.g., Pang et. al. [2002]) taken from the IMDb movie review archive and (2) a testbed of messages from two major extremist forums (one U.S. [in English] and one Middle Eastern [in Arabic]). Their EWGA used the information gain (IG) heuristic to weight the various sentiment attributes and these weights are then incorporated into the GAs initial population as well as crossover and mutation operators. Abbasi et al. [1] found that stylistic features on their own were outperformed by syntactic features, but when triangulated with syntactic features a higher classification accuracy (about 5%) was achieved. A number of stylistic features were found to be specifically helpful, including the total number of characters, use of digits and emphasizing symbols, and vocabulary richness. Abbasi et al. [1], however, did not give much information about the pre-processing of the Arabic data. Given the rich morphology of arabic, pre-processing is a crucial step that we believe should be explained in detail.

3 METHOD

3.1 Dataset and annotation

We use the Penn Arabic Treebank [10] as our dataset. More specifically, we extracted and annotated 1552 sentences from the first 200 documents (i.e., more than one fourth [namely, 0.27%]) of Part 1 V 3.0 of the Penn Arabic Treebank. This version was released in 2004, and is encoded in Buckwalter transliteration [3]. This improved version uses a level of annotation that is more accurately described as morphological analysis than as part-of-speech (POS) tagging. To annotate our dataset, we followed the following procedure: For each sentence, each of the two annotators was required to provide a tag (i.e., either subjective or objective). We followed [18] in operationalizing these categories. In other words, if it was felt that the primary goal of a sentence is the objective reporting of information, it was labeled as objective. Otherwise, a sentence would be tagged as subjective. The two authors, who are both native speakers of Arabic, annotated the data. After extensive discussions and consideration of the literature on annotating data for subjectivity, each annotator independently labeled 10% of the data (i.e., 156 sentences). We then met and compared the annotations of each of us. Inter-annotator agreement was found to be 87.1% (with differences in 20 sentences). After discussing the differences, we reached 97.4% agreement (with inability to resolve 4 sentences). Once we reached this high level of agreement, we decided that each of us annotates the remaining 90% (i.e. 1396 sentences) of the data independently. After each of us finished annotating the rest of the corpus, we calculated inter-annotator agreement on that remaining part and it was found to be 91.6%. For all remaining cases of disagreement, the annotations of the first author, a (computational) linguist with prior background with similar tasks, were adopted.

Table 1 below provides the numbers of instances of each category labeled in our corpus:

Table 1. Number of instances per each category.

Category	# of instances
Objective	815
Subjective	737

3.2 Feature extraction

Since we wanted to identify which representation of a word and which words would boost the performance of a classifier, we extracted the words and corresponding POSs in the treebank in various ways, including the raw form of each word and its corresponding tokens. To reduce the number of features, we decided to practice some informed tuning by using only tokens, rather than words, with the POSs in table 2 below with the three classifiers.

Table 2. POS for selected tokens.

POS tag	Meaning
ADJ	Adjective
ADV	Adverb
EMPHATIC_PARTICLE	Emphatic particle
NON_ALPHABETIC	Non-alphabetic token
NON_ARABIC	Non-Arabic token
NOUN_PROP	Proper noun
NUM	Number
VERB_IMPERFECT	Imperfect verb
VERB_PASSIVE	Passive verb
VERB_PERFECT	Perfect verb

We had the intuition that tokens with these POSs would be the most important ones for our task, which turned out to be true.

3.3 Experiments

After extracting the 1552 sentences from the 200 documents from the treebank, we created a vector for each sentence and used the standard ten-fold cross-validation to determine the overall accuracy of each of our classifiers. Since our dataset is relatively small, we also report results in terms of five-fold cross validation. Below, we describe the three classifiers we have used along with our two vector settings of frequency and presence.

3.3.1 Classifiers

The SVM^{light} package [7] was used in all of the support vector machine calculations with the default settings i.e., the linear kernel. Since Nave Bayes and IB1 classifiers have also been previously used for text-classification, we decided to use it for comparison purposes. We used Wekas implementation with default parameters for these two last classifiers. Weka is a library of Machine Learning Algorithms in Java [20].

3.3.2 Building a lexicon of polarized adjectives

Since previous research on subjectivity and sentiment analysis has proved that polarized adjectives are generally barriers of subjective content, we have extracted all adjectives from all the first four parts of the Penn Arabic Treebank and manually selected those adjectives that we believed are either positive or negative. We have used two features related to the absence or existence of items of this list of polarized adjectives in each sentence vector, in order to give more weight to these adjectives as to the decision of the classifier. In the future, we plan to automatically expand our lexicon of polarized adjectives.

3.3.3 Vector settings

To test which vector settings will work better for the task, we ran our experiments with two different vector settings: frequency and presence. A sentence vector is a frequency sentence vector if the value of each coordinate is based on the frequency of a feature in the class to which the sentence belongs. The more a feature appears in a class the larger its coordinate value will be in sentence vectors of that class. A sentence vector is a presence sentence vector if the value of each coordinate is based on whether a feature is present in the sentence or not. Any feature that occurs at least once receives the same value (i.e., value of 1).

4 Results

As mentioned earlier, we tested the performance of three different classifiers, each with two sentence vectors (i.e., frequency vs. presence vectors). Further, we ran experiments with 10- as well as 5-fold cross validation. Table 3 below show the results, in terms of accuracy, of our various experiments:

Table 3. Results.

Type of vectors	# of folds	Type of classifier		
		SVM	IB1	NB
Frequency vectors	5 folds	99.48%	93.71%	97.27%
	10 folds	99.35%	93.91%	97.24%
Frequency vectors	5 folds	98.97%	81.59%	82.23%
	10 folds	99.23%	81.89%	82.90%

As shown in table 3 above, our best result, 99.48% accuracy, was obtained with the SVM and frequency vectors, 5-fold cross validation. A consideration of the results shows that the SVM classifier performs better than the other two methods, given the current feature setting. It has already been indicated by [8] that SVMs perform better on text classification and do not need parameter tuning. What feature settings other than those we have implemented could boost the performance of the IB1 and NB classifiers on the task of subjectivity classification and whether or not their performance could be comparable to those of SVM remain an open question. As table 3 shows, results acquired from 5-fold and 10-fold cross validation are not strikingly different (i.e., the difference is always below 0.5%). What turned out to be different for two of the classifiers are the vector settings (i.e., frequency vs. presence vectors). More specifically, whereas the type of vector setting used did not result in strikingly different results with respect to the SVM classifier, frequency vectors were found to very much benefit the IB1 and NB classifiers (with improvement of 12.07% in the case of the IB1 classifier and 14.69% for the NB classifier). Another interesting finding is that the classifier performed significantly better than our overall human inter-annotator agreement. Although this would seem surprising, it may be an artifact of the fuzziness of classifying the sentiment of texts for a human annotator.

5 CONCLUSION

Although there is one work (i.e., [1]) on Arabic sentiment classification, our work, to the best of our knowledge, is the first on Arabic-based subjectivity classification. In addition, our work is different from [1] in various ways. First, we use data from a different domain (i.e., news data) that is known to have less subjective content than extremist Web forums. Second, we use different classifiers and employ two different settings (i.e., frequency vs. presence vectors). Third, our work is at the sentence level, as opposed to their document-level classification. Fourth, we keep only tokens with certain POS, rather than extracting roots of words in the text using a dictionary-based approach. We believe that a dictionary-based approach is limited by the scope of the dictionary itself. One possible advantage of [1] over our approach, however, is that they worked on a potentially relatively noisy domain (although it is not clear how noisy their data was). This contrasts with our usage of gold-standard data from the Penn Arabic Treebank. However, our work is original in that it introduces what we believe to be a crucial solution (as the very high accuracy we achieve testifies) to the problem on modern standard Arabic (MSA). Given the current existence of high-performance tokenizers and POS taggers for MSA, we see no problem with using our approach with real world MSA data. Finally, we achieve significantly better results.

It was shown by [8] that being too aggressive in tuning leads to loss of information and less accurate results. This is why we decided to adopt what we call informed tuning, which turned out to be useful for Arabic. Our decision to adopt this approach was based on our observation that the tokens we have kept to use as features are those that are relevant for the current task. We believe that the sort of informed tuning (i.e., using only tokens with certain POSs rather than excluding a commonly-used standard stop list) we employ here would also generalize over other languages. In addition, while we believe that this approach could boost the performance of classifiers on a language like English, we also would like to claim that it, theoretically, should solve some text classification issues (e.g. subjectivity classification) for morphologically rich languages. If this turns out to be empirically true, the approach will boost the state-of-the-art in subjectivity (and, generally, text) classification in these languages.

Many interesting future extensions to our work seem possible. First, we plan to extend our work to automatic sentiment classification on POS-tagged Arabic texts to identify whether or not our approach will achieve the same very high accuracy as the one reported here. Second, we plan to perform sentiment and subjectivity classification on Arabic real world data, using a tokenizer and a POS tagger. Finally, we believe that the approach we describe here could be straightforwardly applied on other MRLs (e.g., Hebrew) and do hope that this claim will be tested.

ACKNOWLEDGEMENTS

We would like to thank Kiduk Yang and the anonymous reviewers for their helpful feedback.

REFERENCES

[1] A. Abbasi, H. Chen, and A. Salem, ‘Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums’, *ACM Trans. Inf. Syst.*, **26**, 1–34, (2008).

[2] R. Bruce and J. Wiebe, ‘Recognizing subjectivity. a case study of manual tagging’, *Natural Language Engineering*, **5(2)**, (1999).

[3] T. Buckwalter. Arabic morphological analyzer version 1.0. Linguistic Data Consortium, 2002.

[4] N. Habash, O. Rambow, and R. Roth, ‘Mada+tokan: A toolkit for arabic tokenization, diacritization, morphological disambiguation, pos tagging, stemming and lemmatization’, in *Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR)*, Cairo, Egypt, (2009).

[5] V. Hatzivassiloglou and K. McKeown, ‘Predicting the semantic orientation of adjectives’, in *ACL-EACL*, pp. 174–181, (1997).

[6] V. Hatzivassiloglou and J. Wiebe, ‘Effects of adjective orientation and gradability on sentence subjectivity’, in *International Conference on Computational Linguistics, (COLING-2000)*, (2000).

[7] T. Joachims. Svmlight: Support vector machine. http://svmlight.joachims.org/, Cornell University, 2008.

[8] T. Joachims, ‘Text categorization with support vector machines: Learning with many relevant features’, in *Proc. of the European Conference on Machine Learning (ECML)*, pp. 137–142, (1998).

[9] D. Lin, ‘Automatic retrieval and clustering of similar words’, in *Proc. COLING-ACL 98*, pp. 768–773, (1998).

[10] M. Maamouri, A. Bies, T. Buckwalter, and W. Mekki, ‘The penn arabic treebank: Building a large-scale annotated arabic corpus’, in *NEMLAR Conference on Arabic Language Resources and Tools*, pp. 102–109, (2004).

[11] M. Marcus, B. Santorini, and M. Marcinkiewicz, ‘Building a large annotated corpus of english: The penntreebank’, *Computational Linguistics*, **19(2)**, 313–330, (1993).

[12] G. Miller, ‘Wordnet: An on-line lexical database’, *International Journal of Lexicography*, **3(4)**, 235–244, (1990).

[13] B. Pang, L. Lee, and S. Vaithyanathan, ‘Thumbs up? sentiment classification using machine learning techniques’, in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 79–86, (2002).

[14] E. Riloff, ‘Automatically generating extraction patterns from untagged text’, in *Proceedings of the Thirteenth National Conference on Artificial Intelligence (AAAI-96)*, pp. 1044–1049, (1996).

[15] E. Riloff, J. Wiebe, and W. Phillips, ‘Exploiting subjectivity classification to improve information extraction’, in *Proc. 20th National Conference on Artificial Intelligence (AAAI-05)*, pp. 1106–1111, Pittsburgh, PA, (2005).

[16] R. Tsarfaty, D. Seddah, Y. Goldberg, S. Kuebler, Y. Versley, M. Candito, J. Foster, I. Rehbein, and L. Tounsi, ‘Statistical parsing of morphologically rich languages (spmrl) what, how and whither’, in *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages*, Los Angeles, CA, (2010).

[17] J. Wiebe, ‘Learning subjective adjectives from corpora’, in *Proc.17th National Conference on Artificial Intelligence (AAAI-2000)*, pp. 735–741, Austin, Texas, (2000).

[18] J. Wiebe, R. Bruce, and T. O’Hara, ‘Development and use of a gold standard data set for subjectivity classifications’, in *Proc. 37th Annual Meeting of the Assoc. for Computational Linguistics (ACL-99)*, pp. 246–253, University of Maryland: ACL, (1999).

[19] J. Wiebe and E. Riloff, ‘Creating subjective and objective sentence classifiers from unannotated texts’, in *Computational Linguistics and Intelligent Text Processing*, pp. 486–497, (2005).

[20] I. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann Publishers, San Mateo, CA, 2000.

Evaluation and Extension of a Polarity Lexicon for German

Simon Clematide and **Manfred Klenner** [1]

Abstract. We have manually curated a polarity lexicon for German, comprising word polarities and polarity strength values of about 8,000 words: nouns, verbs and adjectives. The decisions were primarily carried out using the synsets from *GermaNet*, a *WordNet*-like lexical database. In an evaluation on German novels, it turned out that the stock of adjectives was too small. We carried out experiments to automatically learn new subjective adjectives together with their polarity orientation and polarity strength. For this purpose, we applied a corpus-based approach that works with pairs of coordinated adjectives extracted from a large German newspaper corpus. In the context of this work, we evaluated two subtasks in detail. First, how good are we at reproducing the polarity classification – including our three-level strength measure – contained in our initial lexicon by machine learning methods. Second, because adding of training material did not improve the results at the expected rate, we evaluated the human inter-coder agreement on polarity classifications in an experiment. The results show that judgements about the strength of polarity do vary considerably between different persons. Given these problems related to the design and automatic augmentation of polarity lexicons, we have successfully experimented with a semi-automatically approach where a list of reliable candidate words (here: adjectives) is generated to ease the manual annotation process.

1 Introduction

A polarity lexicon of prior word or word sense polarities forms the basis of (almost) any sentiment detection approach. In order to avoid costly manual lexicon design, most approaches prefer to automatically or semi-automatically derive such lexicons from small seed lexicons with clear cut polarity by utilizing corpus-based cooccurrence effects [19, 13]. Or, they define conversion methods turning semantic databases as *WordNet* into a polarity tagged resource [2]. The second approach has the advantage that polarity is attributed to word senses (*synsets*), and not just lemmas (see [1] for the topic of subjectivity word sense disambiguation). As [9] showed, using both of this approaches, i.e. a lexical database as WordNet and the textual corpus contained in its glosses, improves the results. Yet, a third, translation-based approach is viable as in [20] where English sentiment resources are transferred into German by low-cost bilingual dictionary online services.

The work of [22] has shown that a carefully designed polarity lexicon of about 8,000 entries can lead to well performing sentiment analysis systems. We use the subjectivity lexicon of [22] as part of our multilingual (English, French, German) polarity detection system *PolArt*[2] [16]. For the German version, we have annotated synsets from *GermaNet*, a *WordNet*-like lexical database, with prior polarities (positive, negative) and polarity strength (three values: low, medium, high).[3] For instance, “Zärtlichkeit” (*tenderness*) is highly positive, while “Zaghaftigkeit” (*timidity*) is – with a low intensity – negative. We are aware of the problems with prior polarities, c.f. [10], i.e. that they might get contextually overridden (neutralized or even inverted, e.g. in ironical usages). Nevertheless they are useful and indispensable as a basis to more sophisticated sentiment processing. Especially in systems that carry out sentiment composition [16], where two polar words are combined to form a phrase-level (and eventually a sentence-level) polarity. So, ’disappointed$^-$ hope$^+$’ is negative as well as ’a perfect misery$^-$’. We deliberately did not tag ’perfect’ here with a polarity since it acts as an intensifier (with a strength value) rather than a positive adjective. As such it does not invert the polarity of the noun, but it pushes the negative strength value of it even higher.

In our evaluation of the German lexicon on the basis of literary texts we found that especially the stock of polarity tagged adjectives was far too low [15]. There are quite a number of misclassified phrase-level polarities due to missing adjectives in the lexicon. The crucial importance of this word class for sentiment detection performance is also described in [5]. In order to overcome these gaps, we carried out experiments to automatically induce a larger lexicon starting from our stock of adjective seed entries. The design of our experiments and the empirical evaluation of our new lexical resource are the main topics of our paper.

The rest of this paper is organized as follows: In Section 2 we explain our method for polarity orientation classification using cooccurrence data from coordinated adjectives and report the results. In Section 3 we describe our experiment on human polarity classification and discuss some of the implications regarding inter-coder agreement and the reliability of fine-grained strength classification. In Section 4 we present and discuss the adjective acquisition method which proved to be useful for semi-automatic extension of our lexicon.

2 Orientation classification experiments

In this section, we describe the different steps to assess and evaluate polarity orientation assignments using coordinated adjectives extracted from large text corpora.

Our initial seed lexicon consists of 2899 adjectives with polarity

[1] University of Zurich, Switzerland, email: {klenner,siclemat}@cl.uzh.ch

[2] See `http://kitt.cl.uzh.ch/kitt/polart` for a demonstration of the *PolArt* system and for access to our freely available resources.

[3] The lack of freely available German sentiment resources has also been remedied by [20]. This lexicon (called *GermanPolarityClues*) provides in total 3,220 negative and 5,848 negative word readings (so-called features) distributed over verbs, nouns, adjectives, and adverbs.

classification[4]. The polarity classes are not equally distributed as Table 1 shows. Entries with low polarity strength are rare. There are more negative items (55.4%) than positive ones (44.6%). More than a quarter of the adjectives are highly negative. For the classification experiments we sampled 2850 entries (henceforth evaluation corpus).

Table 1. Distribution of the polarity strength classes in our lexicon including a random example for each class. *h* is high strength, *m* is medium strength, *l* is low strength.

%	Freq	Pol	Examples
27.1	785	–h	sadistisch (*sadistic*)
19.5	566	–m	ablehnend (*refusing*)
19.5	565	+h	fachkundig (*expert*)
18.4	533	+m	kühn (*bold*)
8.8	255	–l	stiefmütterlich (*stepmotherly*)
6.7	195	+l	wuchtig (*bulky*)

2.1 Extraction of coordinated adjective pairs

For each lemma of our lexicon, we use the word form generation service of "Wortschatz Leipzig"[5] (henceforth WS) [17] to create inflected word forms. This service returns only word forms represented in the corpus. For adjectives derived from past or present participles, the verb forms are delivered as well. For instance, for a German adjective like "missraten" (*wayward*) we get "missrät, missriet, **missratener**, missrieten, **missratene**, **missratenes**" (true adjectives in bold). For our evaluation corpus, a set of 23,761 word forms arises. For each word form, we request a set of sample sentences from WS. However, the service delivers at most 256 of them for one word form. In total, we got 2,039,175 sentences for all word forms.

Each sample sentence is tagged and syntactically analyzed by the chunker "Chunkie" [18], which processes extremely fast, yet sometimes with imprecise results. After that, we extract all pairs of adjacent heads[6] of coordinated adjective phrases. For example, the sentence "Es ist ein veritables Labyrinth mit idyllischen, romantischen und gruseligen Zutaten" (*It is a veritable maze of idyllic, romantic, and scary ingredients*) gets the following analysis represented in Penn-Treebank format[7]:

(PPER Es) (VAFIN ist) (NP (ART ein) (ADJA veritables) (NN Labyrinth)) (PP (APPR mit) **(CAP (ADJA idyllischen) ($, ,) (ADJA romantischen) (KON und) (ADJA gruseligen))** (NN Zutaten)) ($. .)

From the tripartite coordination structure contained therein (marked in bold face), we extract the following two pairs :

1. "idyllisch/romantisch" (*idyllic/romantic*)
2. "gruselig/romantisch" (*scary/romantic*)

To achieve a canonical pair representation we always order the lemmas in alphabetical order. For reasons of precision, we dispense with transitive pairs as "gruselig/idyllisch". The heads are lemmatized by the morphological analyzer GERTWOL [12].

Table 2 shows the mean numbers of sentences and adjective pairs (tokens and types) computed on increasing partitions of our evaluation lexicon with the same binning sizes as used for the classification experiments in Section 2.3. Since the same sentence may contain more than one adjective, the average number of sentences per adjective lemma drops as the number of adjectives grows. For the whole evaluation lexicon, we find 29.2 pair types on average for every adjective (type-token ratio: 0.69). Only 12.3 of them combine 2 entries from our polarity lexicon. But note that this value grows linearly with the size of the lexicon starting from 2.4, although the number of pairs containing one polarity item doesn't change. About 77% of the pair types show identical binary polarity orientation, which already demonstrates a strong cooccurrence bias. In contrast, by taking the more fine-grained polarity strength distinction, only 36% of the pairs show identical polarity.

Table 2. Frequency rates of different sentences and coordinated pairs computed per seed adjective from our evaluation lexicon. The number of adjectives reported in each column Sn is $2.850 \times n/5$. The same binning size is used in the polarity classification experiment. Legend for columns: Small letters a report counts on occurrences, uppercase letters A report type counts. All pairs xy and yx are treated as one alphabetically ordered type. *Sent* is the number of different sentences containing at least one adjective from the lexicon; aa or AA is the mean number of coordinated adjective pairs; $\overline{aa}$ or $\overline{AA}$ is the mean number of coordinated pairs containing at least one seed adjective; $\bar{a}\bar{a}$ or $\bar{A}\bar{A}$ is the mean of coordinated pairs consisting of two seed adjectives; $_{\pm}\vec{a}\vec{a}$ or $_{\pm}\vec{A}\vec{A}$ is the mean of coordinated pairs where both adjectives have equal polarity direction (only $+$ or $-$); $_{\pm 3}\vec{a}\vec{a}$ or $_{\pm 3}\vec{A}\vec{A}$ is the mean of coordinated pairs where both adjectives have equal polarity values $\pm\{h, m, l\}$;

	S1	S2	S3	S4	S5
Lemmas	570	1140	1710	2280	2850
Sent	852.8	796.6	753.6	736.8	715.5
aa	66.0	62.7	58.1	54.8	51.9
$\overline{aa}$	40.0	42.9	42.8	42.8	42.2
$\bar{a}\bar{a}$	3.9	8.1	11.7	16.2	20.2
$_{\pm}\vec{a}\vec{a}$	2.7	6.2	9.2	12.7	15.9
$_{\pm 3}\vec{a}\vec{a}$	1.3	3.0	4.5	6.2	7.6
AA	50.3	45.6	41.4	38.2	35.6
$\overline{AA}$	29.4	30.6	30.3	29.8	29.2
$\bar{A}\bar{A}$	2.4	4.9	7.4	9.8	12.3
$_{\pm}\vec{A}\vec{A}$	1.8	3.7	5.7	7.5	9.5
$_{\pm 3}\vec{A}\vec{A}$	0.8	1.7	2.6	3.4	4.4

2.2 Coordination hypothesis

Our orientation classification approach relies on the idea from [13] that "conjunctions between adjectives provide indirect information about orientation". [13] proved for English that coordinated adjectives show same orientation to a degree which is way beyond chance.

In our case, we count 35,156 occurrences of adjective pairs ($\bar{A}\bar{A}$, c.f. Table 2) using the whole seed lexicon. The distribution of negative and positive lemmas in these pairs is as follows: + (54%), – (46%). Although adjectives with negative orientation are more frequent in our lexicon as shown in Table 1, we find more or less inverse proportions in the extracted corpus data.

[4] Our lexical sentiment resource contains additionally words which functions as polarity shifters or intensifier as mentioned above. Additionally we have a list of all adjectives considered as neutral when the seed lexicon was prepared from *GermaNet* data.

[5] See `http://wortschatz.uni-leipzig.de` and the documentation on the available SOAP services. We performed the lookup using the handy command line tool `wsws.pl`. It is part of the Perl package Lingua::DE::Wortschatz implemented by Daniel Schröer.

[6] Conjunctions are the only word class we allow between adjective heads. This restriction helps to remove noise in our data which otherwise would occur because of faulty chunking analyses. A detailed quantitative evaluation on the recognition quality of *Chunkie* for coordinated adjectives has been done in [6].

[7] The German version of Chunkie assigns part-of-speech tags and phrase tags from NEGRA corpus: `http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus/negra-corpus.html`

Table 3 shows the expected relative frequencies assuming equally distributed positive and negative adjectives and the corresponding empirical values. The chi-square goodness-of-fit test reveals an extreme low probability for an equal distribution of same- and different-orientation adjectives in coordinated pairs (X-squared = 10326.55, df = 2, p-value $< 2.2e\text{-}16$). This again supports the hypothesis that coordinated adjectives tend to share their polarity orientation.

Table 3. Distributions of orientation pairs of coordinated adjectives. The column labelled $+/-$ comprises all pairs with a positive and a negative item.

Pair orientations	$+/+$	$+/-$	$-/-$
Expected relative frequency	0.30	0.50	0.20
Empirical relative frequency	0.43	0.23	0.34

A closer look at our data reveals which of the pairs (expressed in terms of their polarity strength values) significantly exceed the expected value (winners) and which ones are below it (losers). Table 4 shows the winners of the same orientation bias predicted by the coordination hypothesis. High strength – especially negative direction, but also in the case of positive direction – boosts the same orientation bias most. Combinations with lower strength get quite a small benefit.

Table 4. Winners of the same orientation bias ordered by difference

Pair	Expected	Empirical	Difference
-h-h	5.2	11.1	+5.9
+h+m	11.5	16.6	+5.1
+h+h	6.9	11.0	+4.1
-h-m	7.3	10.3	+3.0
+m+m	4.8	7.1	+2.3
-m-m	2.5	4.6	+2.1
-m-l	2.1	3.5	+1.4
+m+l	2.9	3.8	+1.0
+h+l	3.4	4.1	+0.7
-h-l	3.0	3.7	+0.7
-l-l	0.4	0.7	+0.3
+l+l	0.4	0.7	+0.3

Table 5 shows the losers of the same orientation bias. As can be seen quickly from this table, the bigger the orientation difference, the bigger the empirical loss. The picture drawn by the descriptive statistics already exhibits the fact that the coordination hypothesis works best for high and medium orientation strengths. Throughout, low orientation strength means low bias, and in the case of +l/-l almost no bias. Therefore, we can draw the conclusion that the same orientation bias correlates positively with the orientation strength.

2.3 Automatic orientation classification

In the next section, we discuss and evaluate methods to classify automatically the direction and the strength of non-neutral adjectives.

2.3.1 Binary classification

Our coordinated pair data (i.e. pairs of type $\bar{a}\bar{a}$) that was gained through natural language processing as described in Section 2.1 can be exploited for automatic orientation classification. For the task of learning the positive or negative orientation of a subjective adjective x we used the following baseline decision procedure:

Table 5. Losers of the same orientation bias ordered by difference

Pair	Expected	Empirical	Difference
+h-h	12.1	4.4	-7.7
+m-h	10.0	3.7	-6.3
+h-m	8.3	3.6	-4.7
+m-m	6.9	3.6	-3.3
+h-l	3.4	1.8	-1.6
+l-h	3.0	1.5	-1.5
+m-l	2.9	1.8	-1.1
+l-m	2.1	1.4	-0.6
+l-l	0.9	0.8	-0.1

1. Count all occurrences of all known subjective adjectives which appear combined with x in a coordinated pair.
2. Set the orientation of x to the orientation of adjective z which co-occurs most often with x.

If there is no coordination pair in our data for a given adjective no decision is made. This is the case for 249 adjectives of our seed lexicon.

Table 6 gives the results of our baseline algorithm computed with ten-fold cross-validation. Because performing cross-validation is data-intensive we decided to compute the learning curves for binning sizes of 570 lemmas. The standard deviation numbers in Table 6 reveal that the variability of the results is considerably high. As evaluation measures we use precision (P), recall (R), and their harmonic mean, F1-measure (F). Recall grows quickly with larger training sets. The difference of 4% respectively 2% in recall between data set S4 and S5 suggests that more data would increase recall even more. To a less degree, this is also true for precision. There is a certain tendency towards positive orientation classification that reflects the general prevalence of positive items in our pair data.

Table 6. Learning rates and performance of our baseline algorithm for orientation classification. Column *E* specifies the evaluation measures: *P* is precision, *R* is recall, *F* is F1-measure. The corpus size reported in each column Sn is $2.850 \times n/5$. We performed a ten-fold cross-validation using 1/10 of the corpus as test material. Standard deviation is noted as a subscript; these values show that the results vary quite strongly for the smaller training sets.

Pol	E	S1	S2	S3	S4	S5
Lem		570	1140	1710	2280	2850
+	P	$75_{\pm 11}$	$81_{\pm 5}$	$84_{\pm 3}$	$83_{\pm 3}$	$84_{\pm 4}$
+	R	$63_{\pm 11}$	$72_{\pm 6}$	$74_{\pm 3}$	$77_{\pm 5}$	$81_{\pm 4}$
+	F	$67_{\pm 7}$	$76_{\pm 4}$	$79_{\pm 3}$	$80_{\pm 4}$	$82_{\pm 3}$
–	P	$82_{\pm 9}$	$87_{\pm 5}$	$90_{\pm 5}$	$91_{\pm 4}$	$93_{\pm 3}$
–	R	$42_{\pm 5}$	$63_{\pm 6}$	$72_{\pm 5}$	$74_{\pm 5}$	$76_{\pm 3}$
–	F	$55_{\pm 4}$	$73_{\pm 5}$	$80_{\pm 5}$	$81_{\pm 4}$	$83_{\pm 2}$

To improve beyond the baseline we experimented with methods assessing the orientation of an adjective by measuring the similarity between the set of lemmas it is coordinated with. Although a similar approach based on point-wise mutual information was used in [3] we could not beat the baseline algorithm in general. That means, although for smaller training sizes better results were possible, the same was not true for larger training sizes.

Therefore we tried a new approach based on conditional maximum entropy models [4], which allow easy integration of diverse and partially dependent features. The maximum entropy classifier *MegaM* [8] was used for our experiments. The best results for the orientation classification task used the following feature extraction procedure:

1. For each subjective adjective, compute the set of all other subjective adjectives that co-occur in an extracted coordination pair (so-called *coordination fellows*).
2. For each positive adjective each positive coordination fellow acts as a feature. In the same way for each negative adjective each negative coordination fellow acts as a feature.
3. To account for pure frequency effects which proved to be powerful in the baseline algorithm, several features based on raw counts were defined: For example, whether at least 60, 70, or 80 percent of all occurrences of coordination fellows of an adjective are positive or negative. Similar features were created for the positive or negative types of coordination fellows.

The resulting maximum entropy model computes the optimal weight for each feature on our training data. With regard to the coordination fellows of an adjective, we estimate the same-orientation impetus of a subjective word in coordinated contexts. With regard to the frequency features, we estimate their general importance for the classification.

Table 7 shows the results for the maximum entropy method computed with ten-fold cross-validation using the same metrics as in Table 6. Overall performance expressed by F-measure improves by almost 3%. The differences between the precision/recall values of negative and positive adjectives which were very strong for the baseline algorithm calibrate much better with the machine learning method. The variability of the results expressed by the standard deviation, however, is still high.

Table 7. Learning rates and performance of our maximum entropy method for polarity orientation classification.

Pol	E	S1	S2	S3	S4	S5
Lem		570	1140	1710	2280	2850
+	P	$77_{\pm 9}$	$84_{\pm 5}$	$87_{\pm 4}$	$87_{\pm 3}$	$87_{\pm 4}$
+	R	$61_{\pm 10}$	$71_{\pm 6}$	$75_{\pm 3}$	$78_{\pm 4}$	$80_{\pm 4}$
+	F	$68_{\pm 6}$	$77_{\pm 5}$	$81_{\pm 3}$	$82_{\pm 3}$	$84_{\pm 2}$
-	P	$78_{\pm 10}$	$84_{\pm 6}$	$89_{\pm 4}$	$90_{\pm 4}$	$90_{\pm 4}$
-	R	$51_{\pm 6}$	$69_{\pm 6}$	$78_{\pm 4}$	$80_{\pm 3}$	$82_{\pm 2}$
-	F	$61_{\pm 5}$	$76_{\pm 5}$	$83_{\pm 4}$	$85_{\pm 3}$	$86_{\pm 2}$

2.3.2 *Polarity strength classification*

We applied the same approaches for polarity orientation classification to the more difficult problem of polarity strength classification. Table 8 contains the results for our baseline algorithm. High standard deviation values indicate that we run into severe sparse data problems with this fine-grained classification. However, as Table 5 showed, there is considerably less same-orientation bias for low and medium strength orientation. Therefore, we expect our method to perform rather poorly for these classes anyway. Table 8 shows that additional training material does not improve the F-measure performance for all classes. The confusion matrix revealed that most of the errors originate from different strength classification (within the correct polarity orientation).

Using our maximum entropy model for polarity strength classification we were unable to beat the baseline algorithm this time. Assuming that our method is more or less sound this begs the question whether manual classification in our seed lexicon contains too much noise, i.e. whether these fine-grained classifications are enough consistent and reliable. The learning curves we receive for our seed lexicon suggest that the distinction between middle and low orientation is not worthwhile. We will turn back to this question in Section 3.

Table 8. Learning rates and performance of our baseline method for polarity strength classification. See Table 6 for the legend.

Pol	E	S1	S2	S3	S4	S5
Lem		570	1140	1710	2280	2850
+h	P	$48_{\pm 24}$	$47_{\pm 14}$	$45_{\pm 10}$	$46_{\pm 4}$	$49_{\pm 6}$
+h	R	$25_{\pm 13}$	$35_{\pm 7}$	$45_{\pm 9}$	$56_{\pm 6}$	$56_{\pm 4}$
+h	F	$32_{\pm 16}$	$39_{\pm 8}$	$45_{\pm 8}$	$50_{\pm 4}$	$52_{\pm 4}$
+m	P	$27_{\pm 12}$	$35_{\pm 10}$	$33_{\pm 11}$	$38_{\pm 8}$	$38_{\pm 12}$
+m	R	$21_{\pm 5}$	$28_{\pm 8}$	$21_{\pm 8}$	$22_{\pm 4}$	$25_{\pm 7}$
+m	F	$23_{\pm 8}$	$30_{\pm 8}$	$25_{\pm 9}$	$27_{\pm 5}$	$30_{\pm 8}$
+l	P	$0_{\pm 0}$	$33_{\pm 41}$	$38_{\pm 29}$	$19_{\pm 25}$	$26_{\pm 20}$
+l	R	$0_{\pm 0}$	$7_{\pm 8}$	$8_{\pm 5}$	$4_{\pm 6}$	$5_{\pm 4}$
+l	F	$0_{\pm 0}$	$12_{\pm 13}$	$13_{\pm 8}$	$6_{\pm 10}$	$9_{\pm 6}$
-l	P	$10_{\pm 32}$	$0_{\pm 0}$	$35_{\pm 41}$	$30_{\pm 20}$	$15_{\pm 16}$
-l	R	$2_{\pm 6}$	$0_{\pm 0}$	$5_{\pm 5}$	$5_{\pm 3}$	$3_{\pm 3}$
-l	F	$3_{\pm 11}$	$0_{\pm 0}$	$8_{\pm 7}$	$8_{\pm 4}$	$5_{\pm 4}$
-m	P	$36_{\pm 33}$	$49_{\pm 14}$	$47_{\pm 7}$	$49_{\pm 6}$	$48_{\pm 9}$
-m	R	$11_{\pm 11}$	$22_{\pm 8}$	$26_{\pm 7}$	$25_{\pm 4}$	$25_{\pm 4}$
-m	F	$17_{\pm 16}$	$30_{\pm 10}$	$33_{\pm 6}$	$33_{\pm 4}$	$33_{\pm 5}$
-h	P	$62_{\pm 21}$	$61_{\pm 7}$	$62_{\pm 6}$	$63_{\pm 9}$	$63_{\pm 7}$
-h	R	$25_{\pm 7}$	$42_{\pm 11}$	$48_{\pm 10}$	$53_{\pm 8}$	$58_{\pm 4}$
-h	F	$35_{\pm 10}$	$50_{\pm 10}$	$54_{\pm 6}$	$57_{\pm 7}$	$60_{\pm 4}$

3 Inter-coder agreement and reliability experiment

The work of [9] emphasizes the fact that sentiment detection contains two subtasks. First, we have to assess whether a term is subjective or objective (neutral). Second, if a term is subjective we have to decide whether it is positive or negative. And additionally, if polarity strength is an issue as in our setting we have to classify them accordingly.

For our user experiment we dropped the difference between medium and low strength which was present in our evaluation corpus. Partly because of the result of our classification learning experiments, partly because we could not expect to make this distinction obvious for our untrained test persons.

To assess the consistency of this three-level decision in our extended lexicon we conducted a user classification experiment. 20 test persons reclassified 60 adjectives randomly selected from our lexicon (including known neutral ones). The lemmas of the adjectives were presented in isolation without any textual context. For half of the lemmas, we presented a list of at most 10 coordination fellows.[8] The test persons had to decide between the following classes:

- Neutral: Objective adjectives
- Strongly positive: Adjectives with strong or purely positive connotations
- Medium positive: Adjectives with weaker or mixed positive connotations
- Strongly negative: Adjectives with strong or purely negative connotations
- Medium negative: Adjectives with weaker or mixed negative connotations
- Undecidable: Adjectives for which the test person could not make a decision

Each decision[9] had to be taken in a time frame of at most 12 seconds. When the classification of an adjective had been dealt with, the par-

[8] The reason for doing so were research questions not treated in this paper.

[9] Other experimental settings are feasible. For example a two-stage procedure where in a first step a decision has to be made between subjective, neutral or undecidable. And then for the subjective cases only, a second classification

ticipants could wait as long as they wanted before the next item was presented. This means, all test person had the same amount of time to classify an adjective but there was no time pressure in the whole experiment because of the breaks.

Table 9. Distributions of polarity orientation classifications of all test persons. The category *undecidable* and timeouts both appear labelled with N/A.

Orientation	+	–	neut	N/A
Relative frequencies (test persons)	0.43	0.26	0.21	0.10
Relative frequencies (lexicon)	0.53	0.37	0.10	

Table 9 contains the relative frequencies for the polarity classification task. Although we sampled randomly, there is quite a large bias towards positive polarity. It is striking that our test persons classified a lot more adjectives as neutral than our creators of the lexicon. One reason for the bias towards subjectivity in our lexicon is the comprehensible tendency to boost recall for a prior sentiment lexicons. Another reason may be given by the fact that the primary source for the classifications were *GermaNet* synsets, and therefore the adjectives appeared in groups of related word senses to the lexicographer. Thus, word senses with polarity orientation were in focus, and neutral senses were pushed into the background.

3.1 Polarity orientation voting

3.1.1 Inter-rater agreement on polarity orientation

The individual decisions of our test persons differ in several ways. In order to build a common sense classification, we select the decision which is in the majority (henceforth called *majority decision*). Ties are resolved towards the most frequent category in the data set.

Using Cohen's Kappa [7] we can assess the inter-rater agreement as well as the agreement between our lexicon and the majority decision. Mean Kappa between all test persons is 0.61[10]. This is a rather strong agreement and comparable to [14]. Mean Kappa between our lexicon and all test persons is only 0.52. Whereas mean Kappa between our lexicon and the majority decision is 0.72. The last value is interesting and shows that to a certain degree the majority decision of weak classifiers (untrained test persons) converges on our more expert classification present in the lexicon.

Table 10 presents accuracy and Cohen's Kappa of the test persons with respect to our lexicon and to the majority decision ordered by the Kappa values. These numbers allow us to calibrate human raters either on the majority opinion or on our lexicon.

In Table 11 we give the precision, recall, and F-measure on positive and negative orientation for our ten best human raters with respect to our lexicon. These numbers allow a more direct comparison to the automatic classification results in Table 7. Taking the mean of F-measure the human raters perform slightly better for positive orientation detection than our conditional entropy model. For negative polarity detection, the automatic methods shows better results. Note however that the human task included neutral orientation detection, and is harder therefore.

How can we quantify reliability of a majority decision? [21] use a metric scale from -1 to 1 for polarity strength which makes standard deviation a measure of variability. Since our classification is

Table 10. Agreement between test persons, our lexicon, and the majority decision. *Acc* is accuracy (percentage of correct classifications), which is the same as micro-average F-Measure.

Agreement with lexicon			Agreement with majority		
Person	Kappa	Acc	Person	Kappa	Acc
5	0.67	80.00	4	0.82	88.33
10	0.67	80.00	3	0.79	86.67
9	0.66	78.33	5	0.79	86.67
14	0.63	76.67	14	0.74	83.33
3	0.62	76.67	9	0.72	81.67
13	0.62	76.67	11	0.71	81.67
11	0.60	75.00	13	0.69	80.00
19	0.60	75.00	19	0.69	80.00
4	0.59	75.00	1	0.64	75.00
12	0.52	68.33	12	0.64	76.67
1	0.48	65.00	10	0.63	76.67
6	0.45	63.33	2	0.52	66.67
8	0.45	63.33	15	0.52	66.67
17	0.45	63.33	6	0.49	65.00
2	0.44	60.00	8	0.49	65.00
15	0.40	60.00	16	0.49	63.33
20	0.40	60.00	20	0.49	65.00
16	0.39	56.67	18	0.48	66.67
18	0.39	61.67	7	0.47	61.67
7	0.35	51.67	17	0.43	61.67

Table 11. Evaluation of the ten best human raters with respect to our lexicon. *P* is precision, *R* is recall, *F* is F1-measure. Standard deviation for mean values are noted as subscripts.

Person	Pol	P	R	F
3	+	90.00	84.38	87.10
3	–	93.75	68.18	78.95
4	+	87.10	84.38	85.71
4	–	86.67	59.09	70.27
5	+	90.32	87.50	88.89
5	–	89.47	77.27	82.93
9	+	89.29	78.12	83.33
9	–	100.00	77.27	87.18
10	+	89.66	81.25	85.25
10	–	100.00	81.82	90.00
11	+	89.29	78.12	83.33
11	–	88.89	72.73	80.00
12	+	92.31	75.00	82.76
12	–	92.31	54.55	68.57
13	+	86.67	81.25	83.87
13	–	94.12	72.73	82.05
14	+	92.86	81.25	86.67
14	–	100.00	72.73	84.21
19	+	96.30	81.25	88.14
19	–	83.33	68.18	75.00
Mean	+	$90.4_{\pm 3}$	$81.3_{\pm 4}$	$85.5_{\pm 2}$
	–	$92.9_{\pm 6}$	$70.5_{\pm 8}$	$79.9_{\pm 7}$

concerning the strength level is needed. With this setting one can pinpoint more exactly which part of the decision is harder, i.e. takes more time.

[10] Regarding the categories positive, negative, neutral, and N/A.

categorial another measure must be used. Therefore, we compute the relative entropy H_{rel} for our majority decisions as follows:

$$H_{rel} = -\frac{\sum_{i=1}^{n}(p_i \times log(p_i))}{log(n)} \quad (1)$$

$H_{rel} = 1$ means equally distributed classification, $H_{rel} = 0$ means uniform classification.

Table 12 shows for each polarity orientation all adjectives ordered from the highest to the lowest relative entropy. Low entropy correlates with high reliability, high entropy pinpoints towards more difficult cases. For instance, the German adjective “kritisch” (*critical*) exhibits the highest entropy value.[11] Applying subjectivity word sense disambiguation as done by [1] seems promising for such cases.

Our lexicon and the majority decision differ for 10 cases. 3 positive adjectives and 3 negative adjectives from our lexicon were rated as neutral by the majority. Most of them show polarity orientation ambiguity. The only adjective where our lexicon and the majority voting propose opposite orientation is “sorgenlos” (*carefree*), which our lexicon wrongly treats the same as “sorglos” (*carefree, careless*). The old-fashioned and rare word “auskömmlich” (*sufficient*) was classified as undecidable by the majority voting because many test persons reached the time out of 12 seconds.

3.1.2 Inter-coder agreement on polarity strength

The task of polarity strength classification is harder than polarity orientation classification. Therefore, a drop on inter-coder agreement is expected. Mean Kappa between our lexicon and all test persons is only 0.31, this is still a fair agreement. Mean Kappa between our lexicon and the majority decision reaches 0.47, this is a moderate agreement.

4 Lexicon extension

In this section we turn back to the question which is a feasible way to extend an existing lexicon. The set of coordinated adjective pairs contains a lot of adjectives that are not part of the seed lexicon. Among them there are many adjective compounds and deverbal adjectives which are often missing in *GermaNet*.

As fine-grained polarity strength classification is not reliable enough for our purposes, we decided to automatically prepare adjective candidates for human classification. Two criteria are crucial for this purpose: First, the candidates should be frequent. Second, false positives should be avoided to minimize human work.

In the first place, we tried an algorithm which favoured frequent and unknown candidates that share the most coordination fellows with a known adjectives. This proved to give unsatisfactory results as well as quite long computation times.

Therefore we tried a simpler algorithm that performed a lot better:

1. Select all unknown adjective lemmas beyond a certain frequency threshold.
2. Request sample sentences for all word forms of these adjectives and extract all coordinated adjective pairs as it was done for the seed lexicon.
3. Sort these adjectives along the criterion which prefers adjectives with the highest proportion of subjective coordinations fellows.

Table 12. Sample adjectives ordered by polarity and the number of votes. *Freq* is the number of votes for the majority decision. If the majority decision and the original PolArt (PA) classification differ it is reported in brackets. We tried to translate our German adjectives into English words with a similar semantic spectrum.

Adjective	Pol (PA)	Freq	H_{rel}
ehrlich (*honest*)	+	20	0.00
herzhaft (*hearty*)	+	20	0.00
praechtig (*magnificent*)	+	20	0.00
blitzschnell (*lightning*)	+	19	0.14
elegant (*elegant*)	+	19	0.14
flink (*agile*)	+	19	0.14
hoeherwertig (*higher quality*)	+	19	0.14
gedankenreich (*rich in ideas*)	+	18	0.23
wirksam (*effective*)	+	18	0.23
niveauvoll (*sophisticated*)	+	18	0.28
solidarisch (*showing solidarity*)	+	18	0.28
verzaubernd (*bewitching*)	+	18	0.28
gradlinig (*straight*)	+	17	0.30
sorgenlos (*carefree*)	+ (–)	17	0.30
namhaft (*substancial*)	+	16	0.44
schluessig (*conclusive*)	+	16	0.44
energisch (*energetic*)	+ (–)	16	0.46
spektakulaer (*spectacular*)	+	16	0.51
aufopfernd (*devoted*)	+	15	0.53
eintraechtig (*peaceful*)	+	15	0.58
vertraeglich (*compliant*)	+	15	0.60
leistungsfoerdernd (*efficiency increasing*)	+	14	0.59
folgerichtig (*consequential*)	+	13	0.47
konzis (*concise*)	+	13	0.62
anruehrend (*touching*)	+	12	0.49
genehmigt (*approved*)	+	12	0.49
schuldlos (*innocent*)	+	11	0.81
meistgespielt (*most often played*)	+	10	0.62
bereit (*willing*, *ready*)	+	10	0.68
atmosphaerisch (*atmospheric*)	+	10	0.79
antriebsarm (*lacking in drive*)	–	20	0.00
unausstehlich (*insufferable*)	–	20	0.00
widerrechtlich (*illegal*)	–	20	0.00
leichtfertig (*frivolous*)	–	19	0.14
populistisch (*populist*)	–	19	0.14
unoekologisch (*anti-ecological*)	–	19	0.14
desorientiert (*disoriented*)	–	18	0.23
veraltet (*outdated*)	–	18	0.23
unedel (*ignoble*)	–	17	0.37
unsolid (*unreliable*)	–	17	0.37
unnoetig (*unnecessary*)	–	15	0.50
unchristlich (*unchristian*)	–	14	0.59
uneinheitlich (*uneven*)	–	14	0.59
unangepasst (*unadapted*)	–	13	0.74
melodramatisch (*melodramatic*)	–	12	0.77
sprachbehindert (*speech impaired*)	–	11	0.70
betaeubt (*stunned*, *dazed*)	–	10	0.74
monarchisch (*monarchic*)	– (0)	9	0.68
zeichnerisch (*graphic*)	0	17	0.42
surreal (*surreal*)	0	14	0.66
schicksalhaft (*fateful*)	0 (–)	11	0.61
taubstumm (*deaf-mute*)	0 (–)	11	0.67
riesenhaft (*gigantic*)	0	11	0.81
dezentral (*decentralized*)	0	11	0.84
angenommen (*assumed*)	0	10	0.72
laeuferisch (*running*)	0 (+)	10	0.82
nichtbehindert (*non-handicapped*)	0 (+)	10	0.87
saturiert (*satisfied*)	0 (+)	8	0.94
kritisch (*critical*)	0 (–)	7	0.96
auskoemmlich (*sufficient*)	na (+)	8	0.78

[11] Our lexicon classifies this word with a medium negative strength whereas the majority votes for neutral.

4. Select the topmost adjectives for human classification.

So far we extended our lexicon iteratively in two rounds: The first round produced a list of 668 candidates of which only 43 were rated as fully neutral. The second round produced a list of 250 candidates of which 30 were rated as neutral. Further acquisition rounds seem feasible. The same method may be used to perform domain adaptation by adding sublanguage specific vocabulary from corresponding text corpora. Automatic extension of a prior polarity lexicon seems practical to us as long as only binary orientation classification is needed.

5 Conclusions

We presented a detailled evaluation on supervised learning of polarity orientations of German adjectives. As already shown by [13] for English, same-orientation of coordinated adjectives allow reliable classifications of binary polarity. We have shown that this also holds for German. However, fine-grained strength differentiations are a lot more difficult to learn. Starting from a set of adjective seed entries we enhanced a polarity lexicon exploiting coordinating adjectives. For the task of binary sentiment orientation detection automatic classification of polarity orientation is viable.

It has been shown that human raters also differ largely regarding polarity strength classification. This empirical finding points out the principled problems related to the automatic acquisition of strength values.

Another fundamental problem of prior polarity lexicons which needs a principled solution are word senses with different polarity orientation. We believe that a contextualization of polarity assignments is necessary. For instance, the German adjective "sorglos" means *carefree* in a positive sense or *thoughtless* in a negative sense depending on the noun it modifies. Quite often lower orientation strength is used to express word sense ambiguity in cases where neutral and subjective readings are common.

Another problem is the unavailability of huge German text corpora, or an easy and programmable way to access them. Missing or false classifications are often due to sparse data problems.

There are several ways to improve our results. The use of a better chunker would optimize the quantity and quality of our extracted pairs. In particular, we could analyze adversative coordinations in a similar vein as [13], or we could relax our adjacency criterion for pair extraction. Furthermore, other promising approaches for term orientation classification have been suggested and tested, and some of them (cf. [11]) have substantially lower requirements in terms of textual data and linguistic knowledge.

ACKNOWLEDGEMENTS

This work is partly funded by the Swiss National Science Foundation (grant 105211-118108). We would like to thank Michael Wiegand and Ronny Peter for their manual lexicon curation, and also to our students and group members who took part in the categorization experiment.

REFERENCES

[1] Cem Akkaya, Janyce Wiebe, and Rada Mihalcea, 'Subjectivity word sense disambiguation', in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pp. 190–199, Singapore, (August 2009). Association for Computational Linguistics.

[2] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani, 'SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining', in *Proceedings of the 7th Conference on Language Resources and Evaluation (LREC'10), Valletta, MT*, pp. 2200–2204, (2010).

[3] Marco Baroni and Stefano Vegnaduzzo, 'Identifying subjective adjectives through web-based mutual information', in *In Proceedings of the 7th Konferenz zur Verarbeitung Natürlicher Sprache (German Conference on Natural Language Processing – KONVENS'04*, pp. 613–619, (2004).

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

[5] Rebecca F. Bruce and Janyce M. Wiebe, 'Recognizing subjectivity: a case study in manual tagging', *Natural Language Engineering*, **5**(02), 187–205, (1999).

[6] Simon R. Clematide, *Koordination im Deutschen und ihre syntaktische Desambiguierung*, Diss., Universität Zürich, 2009.

[7] Jacob Cohen, 'A coefficient of agreement for nominal scales', *Educational And Psychological Measurement*, **20**, 37–46, (1960).

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

[9] Andrea Esuli and Fabrizio Sebastiani, 'Determining term subjectivity and term orientation for opinion mining', in *Proceedings of EACL-06, 11th Conference of the European Chapter of the Association for Computational Linguistics, Trento, IT.*, (2006).

[10] Andrea Fahrni and Manfred Klenner, 'Old wine or warm beer: Target-specific sentiment analysis of adjectives', in *Symposion on Affective Language in Human and Machine, AISB Convention*, pp. 60 – 63, (2008).

[11] Rafael Guzmán-Cabrera, Manuel Montes-y Gómez, Paolo Rosso, and Luis Villaseñor Pineda, 'Using the Web as corpus for self-training text categorization', *Information Retrieval*, **12**(3), 400–415, (December 2008).

[12] Mariikka Haapalainen and Ari Majorin, *GERTWOL: Ein System zur automatischen Wortformenerkennung deutscher Wörter*, Lingsoft Oy, Helsinki, 1994.

[13] Vasileios Hatzivassiloglou and Kathleen R. McKeown, 'Predicting the semantic orientation of adjectives', in *Proc. of ACL-97*, pp. 174–181, Madrid, ES, (1997). Association for Computational Linguistics.

[14] Soo-Min Kim and Eduard Hovy, 'Determining the sentiment of opinions', in *Proceedings of Coling 2004*, pp. 1367–1373, Geneva, Switzerland, (Aug 23–Aug 27 2004). COLING.

[15] Manfred Klenner, 'Süße Beklommenheit und schmerzvolle Ekstase. Automatische Sentimentanalyse in den Werken von Eduard von Keyserling.', in *Tagungsband der GSCL- Tagung, Gesellschaft für Sprachtechnologie und Computerlinguistik (GSCL)*, Potsdam, (30.9. - 2.10 2009).

[16] Manfred Klenner, Stefanos Petrakis, and Angela Fahrni, 'Robust Compositional Polarity Classification', in *Recent Advances in Natural Language Processing (RANLP)*, Borovets, Bulgaria, (2009).

[17] Uwe Quasthoff, Matthias Richter, and Christian Biemann, 'Corpus portal for search in monolingual corpora', in *Proceedings of the fifth international conference on Language Resources and Evaluation, LREC 2006, Genova*, pp. 1799–1802, (2006).

[18] Wojciech Skut, *Partial parsing for corpus annotation and text processing*, Saarbrücken Dissertations in Computational Linguistics and Language Technology 10, Saarland University, Saarbrücken, 1999.

[19] Peter Turney, 'Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews', in *Proc. of ACL*, pp. 417–424, (2002).

[20] Ulli Waltinger, 'GermanPolarityClues: A lexical resource for German sentiment analysis', in *Proceedings of the 7th Conference on Language Resources and Evaluation (LREC'10), Valletta, MT*, pp. 1638–1642. European Language Resources Association (ELRA), (2010).

[21] Gbolahan Williams and Sarabjot Anand, 'Predicting the polarity strength of adjectives using wordnet', in *AAAI International Conference on Weblogs and Social Media (ICWSM 2009)*, (2009).

[22] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann, 'Recognizing contextual polarity in phrase-level sentiment analysis', in *Proc. of HLT/EMNLP 2005*, Vancouver, CA, (2005).

Subjectivity Detection using Genetic Algorithm

Amitava Das[1] and **Sivaji Bandyopadhyay**[2]

Abstract. An opinion classification system on the notion of opinion subjectivity has been reported. The subjectivity classification system uses Genetic-Based Machine Learning (GBML) technique that considers subjectivity as a semantic problem using syntactic simple string co-occurrence rules that involves grammatical construction and linguistic features. Application of machine learning algorithms in NLP generally experiments with combination of various syntactic and semantic linguistic features to identify the most effective feature set. This is viewed as a multi-objective or multi-criteria optimization search problem. The experiments in the present task start with a large set of possible extractable syntactic, semantic and discourse level feature set. The fitness function calculates the accuracy of the subjectivity classifier based on the feature set identified by natural selection through the process of crossover and mutation after each generation. The proposed technique is tested for English and Bengali and for the news, movie review and blog domains. The system evaluation results show precision of 90.22%, and 93.00% respectively for English NEWS and Movie Review corpus and 87.65% and 90.6% for Bengali NEWS and Blog corpus.

1 Introduction

As a growing number of people use the Web as a medium for expressing their opinions, the Web is becoming a rich source of various opinions in the form of product reviews, travel advice, social issue discussions, consumer complaints, movie review, stock market predictions, real estate market predictions, etc. Present computational systems need to extend the power of understanding the sentiment/opinion expressed in an electronic text. The topic-document model of information retrieval has been studied for a long time and several systems are available publicly since last decade. On the contrary Opinion Mining/Sentiment Analysis is still an unsolved research problem. Although a few system like Bing[3] , Twitter Sentiment Analysis[4] Tool are available in World Wide Web since last few years still more research efforts are needed to match the user satisfaction level and social need.

The General Inquirer System by[1] IBM in the year of 1966 was probably the first milestone to identify textual sentiment. They called it a content analysis research problem in the behavioral science. The aim was to gain understanding of the psychological forces and perceived demands of the situation that were in effect when the document was written and counting positive or negative emotion instances. Later on an opinion was defined as a private state that is not open to objective observation or verification [2]. During 1970-1995 various research activities ([3], [4]) proves the necessity of an automated system that can identify sentiment in electronic text.

In the year of 1999 Jaynce Wiebe [5] defined the term Subjectivity in Information Retrieval perspective. Sentences are categorized in two genres as Subjective and Objective. Objective Sentences are used to objectively present factual information and subjective sentences are used to present opinions and evaluations.

Researchers have experimented with several methods to solve the problem of subjectivity detection using SentiWordNet, Subjectivity Word List etc. as prior knowledge database. But Subjectivity Detection is a domain dependent and context dependent problem [6]. Hence building a prior knowledgebase for Subjectivity Detection will never end up with an adequate list. Moreover Sentiment/opinion changes its polarity orientation during time. For example, during 90's mobile phone users generally report in various online reviews about their color phones but in recent time color phone is not just enough. People are excited about their touch screen or various software installation facilities. Hence Subjectivity detection needs a most sophisticated algorithm to capture and effectively use the sentiment pragmatic knowledge. The algorithm should be customizable for any new domain and language.

Previous works in subjectivity identification have helped developing a large collection of subjectivity clues. These clues include words and phrases collected from manually developed annotated resources.

The clues from manually developed resources include entries from adjectives manually annotated for polarity [7], and subjectivity clues listed in [8]. Clues learned from annotated data include distributionally similar adjectives and verbs [9] and n-grams [10]. Low-frequency words are also used as clues. Such words are informative for subjectivity recognition.

The subjectivity detection task in Bengali has started only recently. Several syntactic and semantic feature ensembles with a rule base topic-base model is reported in [11], [12].

Genetic Algorithms (GAs) are probabilistic search methods ([13], [14]). GAs are applied for natural selection and natural genetics in artificial intelligence to find the globally optimal solution from the set of feasible solutions. Nowadays GAs have been applied to various domains that include timetable, scheduling, robot control, signature verification, image processing, packing, routing, pipeline control systems, machine learning, and information retrieval ([15], [16]).

Only a few attempt [17] in the literature uses Genetic Algorithm to solve the opinion mining problem. They developed the Entropy Weighted Genetic Algorithm (EWGA) for opinion feature selection. The features and techniques result in the creation of a sentiment analysis approach geared towards classification of web discourse sentiments in multiple languages. The EWGA has been applied for English and Arabic languages. The Entropy Weighted Ge-

[1] A. Das .Department of Computer Science and Engineering, Jadavpur University. Kolkata 700032, West Bengal, India. email: amitava.santu@gmail.com

[2] S. Bandyopadhyay.Department of Computer Science and Engineering, Jadavpur University. Kolkata 700032, West Bengal, India. email: sivaji_cse_ju@yahoo.com

[3] http://www.bing.com/

[4] http://twittersentiment.appspot.com/

netic Algorithm (EWGA) uses the information gain (IG) heuristic to weight the various opinion attributes.They compared their result with SVM based method and previous existing methods in literature. The EGWA method outperform compared to existing methods and achieved approximately 94.00% accuracy score on both the languages English and Arabic.

Application of machine learning algorithms in NLP generally experiments with combination of various syntactic and semantic linguistic features to identify the most effective feature set. Here we viewed this as a multi-objective or multi-criteria optimization search problem. The experiments in the present task start with a large set of possible extractable syntactic, semantic and discourse level feature set. The fitness function calculates the accuracy of the subjectivity classifier based on the feature set identified by natural selection through the process of crossover and mutation after each generation. In the present paper we use GBML to identify automatically best feature set based on the principle of *natural selection* and *survival of the fittest*. The identified fittest feature set is then optimized locally and global optimization is then obtained by multi-objective optimization technique. The local optimization identify the best range of feature values of a particular feature. the Global optimization technique identifies the best ranges of values of given multiple feature. The proposed technique is tested for English and Bengali and for the news, movie review and blog domains. The system evaluation results show precision of 90.22%, and 93.00% respectively for English NEWS and Movie Review corpus and 87.65% and 90.6% for Bengali NEWS and Blog corpus.

2 Resource Organization

Resource acquisition is one of the most challenging obstacles to work with resource constrained language Bengali, the fifth popular language in the World, second in India and the national language in Bangladesh. NLP research in Bengali has kicked off in recent times and resources like annotated corpus, various linguistic tools are still unavailable for this language. Hence in this section we mainly describe the corpus acquisition and development of tools used in feature extraction for Bengali. English is a resource rich language, the resources for English are collected from various publicly available resources, mentioned in detail in relevance sections.

2.1 Corpus

The subjectivity classification technique presented in this paper is based on Genetic-Based-Machine-Learning (GBML) methodology and hence annotated data preparation is necessary for system testing and evaluation. The technique has been applied on both English and Bengali language texts. In case of English, the MPQA[5] corpus is chosen which is well known for its high inter-annotator agreement score. In the MPQA corpus the phrase level private states are annotated that has been used in the sentence level opinion subjectivity annotation as described in [11]. Manually annotated Subjective data is available for English in the form of International Movie Database (IMDB)[6] among others.

For the present task we have used a Bengali NEWS corpus, developed from the archive of a leading Bengali NEWS paper available on the Web. A portion of the corpus from the editorial pages, i.e., Reader's opinion section or Letters to the Editor Section containing 28K wordforms have been manually annotated with sentence level subjectivity. Detailed reports about this news corpus development in Bengali can be found in [11] and a brief statistics is reported in Table 1.

Table 1. Bengali Corpus Statistics

	NEWS	BLOG
Total number of documents	100	-
Total number of sentences	2234	300
Avgerage number of sentences in a document	22	-
Total number of wordforms	28807	4675
Avgerage number of wordforms in a document	288	-
Total number of distinct wordforms	17176	1235

2.2 Feature Organization

The experimentation started with the complete collection of identified lexicon, syntactic, semantic and discourse level features. The best feature set selection has been carried out by the GBML technique. Various features and the linguistics tools used for features extraction are reported below. The GBML trained with all the features are summarized in Table 3.

2.2.1 Lexico-Semantic Features

- Part of Speech (POS)

Number of research activities like [18], [19] etc. have proved that opinion bearing words in sentences are mainly adjective, adverb, noun and verbs. Many opinion mining tasks, like the one presented in [20], are mostly based on adjective words. The Stanford Parser[7] has been used for identifying the POS tags in English. The POS tagger described in [11] has been used for Bengali.

- SentiWordNet

Words that are present in the SentiWordNet carry opinion information. The English SentiWordNet ([21]) has been used in the present task. The SentiWordNet (Bengali)[8] as described in [22] is used as an important feature during the learning process. These features are individual sentiment words or word n-grams (multiword entities) with strength measure as strong subjective or weak subjective. Strong and weak subjective measures are treated as a binary feature in the supervised classifier. Words which are collected directly from the SentiWordNet are tagged with positivity or negativity score. The subjectivity score of these words are calculated as:

$$E_s = |S_p| + |S_n|$$

where E_s is the resultant subjective measure and S_p, S_n are the positivity and negativity score respectively.

- Frequency

Frequency always plays a crucial role in identifying the importance of a word in the document. After removal of function words and POS annotation, the system generates four separate high frequent word lists for the four POS categories: Adjective, Adverb, Verb and Noun. Word frequency values are effectively used as a crucial feature in the Subjectivity classifier.

[5] http://www.cs.pitt.edu/mpqa/databaserelease/
[6] http://www.cs.cornell.edu/people/pabo/movie-review-data/
[7] http://nlp.Stanford.edu/software/lex-parser.shtml
[8] http://www.amitavadas.com/sentiwordnet.php

- Stemming

Several words in a sentence that carry opinion information may be present in inflected forms. Stemming is necessary for such inflected words before they can be searched in the appropriate lists. Due to non availability of a good Bengali stemmer, a stemming cluster technique based Bengali stemmer [23] has been developed. The stemmer analyzes prefixes and suffixes of all the word forms present in a particular document. Words that are identified to have the same root form are grouped in a finite number of clusters with the identified root word as cluster center. The Porter Stemmer[9] has been used for English.

2.2.2 Syntactic Features

- Chunk Label

Chunk level information is effectively used as a feature in the supervised classifier. Chunk labels are defined as B-X (Beginning), I-X (Intermediate) and E-X (End), where X is the chunk label. A detailed empirical study [11] reveals that Subjectivity clue may be defined in terms of chunk tags. The Stanford Parser has been used for identifying the chunk labels in English. The Bengali chunker used in the present task is described in [11].

- Dependency Parser

Dependency feature is very useful to identify intra-chunk polarity relationship. It is very often a language phenomenon that modifiers or negation words are generally placed at a distance with evaluative polarity phrases. The Stanford Dependency Parser has been for English. A statistical parser [24] has been used for Bengali.

2.2.3 Discourse Level Features

- Positional Aspect

Depending upon the position of subjectivity clue, every document is divided into a number of zones. Various factors of this feature are Title of the document, the first paragraph and the last two sentences. A detailed study was done on the MPQA and Bengali corpus to identify the roles of the positional aspect (first paragraph, last two sentences) in the sentence level subjectivity detection task and these results are shown in the Table 2. Zone wise statistics could not be done for the IMDB corpus because the corpus is not presented as a document.

- Document Title

It has been observed that the Title of a document always carries some meaningful subjective information. Thus a Thematic expression bearing title words (words that are present in the title of the document) always get higher score as well as the sentences that contain those words.

- First Paragraph

People usually give a brief idea of their beliefs and speculations in the first paragraph of the document and subsequently elaborate or support their ideas with relevant reasoning or factual information. This first paragraph information is useful in the detection of subjective sentences bearing Thematic Expressions.

- Last Two Sentences

Table 2. Statistics on Positional Aspect.

Positional Factors	Percentage	
	MPQA	Bengali
First Paragraph	48.00%	56.80%
Last Two Sentences	64.00%	78.00%

It is a general practice of writing style that every document concludes with a summary of the opinions expressed in the document.

- Term Distribution Model

An alternative to the classical TF-IDF weighting mechanism of standard IR has been proposed as a model for the distribution of a word. The model characterizes and captures the informativeness of a word by measuring how regularly the word is distributed in a document. As discussed in [25] introduced the opinion distribution function feature to capture the overall opinion distributed in the corpus. Thus the objective is to estimate that measures the distribution pattern of the *k* occurrences of the word wi in a document *d*. Zipf's law describes distribution patterns of words in an entire corpus. In contrast, term distribution models capture regularities of word occurrence in subunits of a corpus (e.g., documents, paragraphs or chapters of a book). A good understanding of the distribution patterns is useful to assess the likelihood of occurrences of a word in some specific positions (e.g., first paragraph or last two sentences) of a unit of text. Most term distribution models try to characterize the informativeness of a word identified by inverse document frequency (IDF). In the present work, the distribution pattern of a word within a document formalizes the notion of topic-sentiment informativeness. This is based on the Poisson distribution. Significant Theme words are identified using TF, Positional and Distribution factor. The distribution function for each theme word in a document is evaluated as follows: where *n*=number of sentences in a document with a particular theme word S_i=sentence id of the current sentence containing the theme word and S_{i-1}=sentence id of the previous sentence containing the query term, is the positional id of current Theme word and is the positional id of the previous Theme word.

$$f_d(w_i) = \sum_{i=1}^{n} \frac{(S_i - S_{i-1})}{n} + \sum_{i=1}^{n} \frac{(TW_i - TW_{i-1})}{n}$$

where *n*=number of sentences in a document with a particular theme word S_i=sentence id of the current sentence containing the theme word and S_{i-1}=sentence id of the previous sentence containing the query term, TW_i is the positional id of current Theme word and TW_{i-1} is the positional id of the previous Theme word.

Distribution function for thematic words plays a crucial role during the Thematic Expression identification stage. The distance between any two occurrences of a thematic word measures its distribution value. Thematic words that are well distributed throughout the document are important thematic words. In the learning phase experiments are carried out using the MPQA Subjectivity word list distribution in the corpus and encouraging results are observed to identify the theme of a document. These distribution rules are identified after analyzing the English corpora and the same rules are applied to Bengali.

- Theme Words

[9] http://tartarus.org/~martin/PorterStemmer/

In the general practice of Information Retrieval term frequency plays a crucial role to identify document relevance. In many documents relevant words may not occur frequently and on the other hand irrelevant words may occur frequently. A rulebased Theme detection technique has been proposed in [9]. The theme of a document is described as a bag-of-words that describe the topic of the document.

In the general practice of Information Retrieval term frequency plays a crucial role to identify document relevance. In many documents relevant words may not occur frequently and on the other hand irrelevant words may occur frequently. A rule-based Theme detection technique has been proposed in [9]. The theme of a document is described as a bag-of-words that describe the topic of the document.

Table 3. Features.

Lexico-Syntactic Features
POS
SentiWordNet
Frequency
Stemming
Syntactic Features
Chunk Label
Dependency Parsing
Document Title
Discourse Level Features
First Paragraph
Term Distribution Model
Theme Word

3 Basic Principles of Genetic Algorithm

GAs are characterized by the five basic components as follows. Figure 1 displays a diagrammatic representation of the whole process.

1. Chromosome representation for the feasible solutions to the optimization problem.
2. Initial population of the feasible solutions.
3. A fitness function that evaluates each solution.
4. Genetic operators that generate a new population from the existing population.
5. Control parameters such as population size, probability of genetic operators, number of generation etc.

4 Proposed Technique

The experimentation starts with a large set of possible extractable set of syntactic, semantic and discourse level features. The fitness function calculates the accuracy of the subjectivity classifier based on the fittest feature set identified by natural selection through the process of crossover and mutation after each generation. The subjectivity classification problem can be viewed as a summation of the subjectivity probability of the set of possible features.

$$f_s = \sum_{i=0}^{N} f_i$$

Where is the resultant subjectivity function, to be calculated and is the ith feature function. If the present model is represented in a vector space model then the above function could be rewritten as:

$$f_s = \vec{f_i}.\vec{f_{i+1}}.\vec{f_{i+2}}.........\vec{f_n}$$

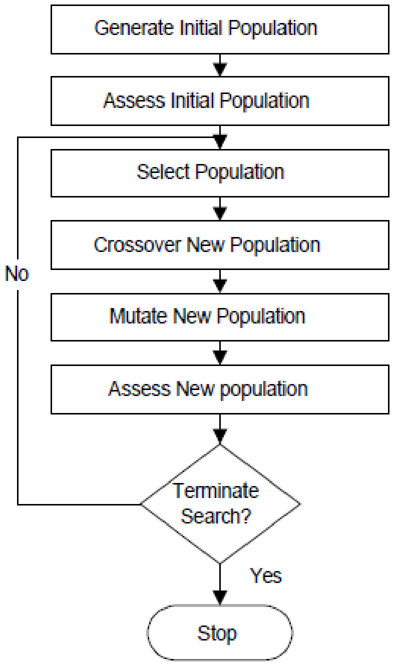


Figure 1. The Process of Genetic Algoriyhm

This equation specifies what is known as the dot product between vectors. Now, in general, the dot product between two vectors is not particularly useful as a classification metric, since it is too sensitive to the absolute magnitudes of the various dimensions.

From the previous research it is already proven that particular features have their own range of tentative values (Instead of features identified by us; Syntactic Chunk Label and Discourse Level feature). As example some special types of POS category reflects sentiment very well, hence it is simpler to infer that frequent occurrence of those special types of POS category into a sentence can increase the subjectivity value of any sentence. Another example: occurrence low frequency word is a well established clue of subjectivity but a sentence with only low-frequent word may not subjective always. In a multiple feature or multiple vector spaced model desired optimal solution may found by finding out the optimal range (highest or lowest) of value of every feature vector. Hence it is obvious that in single-criterion optimization, the notion of optimality scarcely needs any explanation in this particular category of problem. We simply seek the best value of assumedly well-defined multi-objective (utility or cost) optimization function.

4.1 Problem Formulation

To maximize the subjectivity probability, the occurrence of low-frequency words (LFW), title words (TW), average distributed words (ADW) and theme words (TD) and their position in each sentence are calculated. The matrix representation for each sentence looks like: [x, y]= [frequency in the entire corpus, position in the sentence]

LFW= [5, 2]
TW= [34, 11]
ADW= [21, 10]
TD= [25, 5]

The above data are plotted as position versus frequency in the Figure 2.

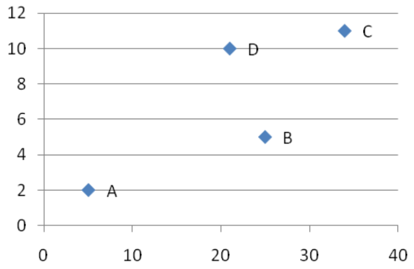


Figure 2. Position vs. Frequency plot of categorical words

Scanning the graph reveals that the best points are lower and to the right of the plot. In particular, scenarios A, B and C seem like good possible choices: even though none of the three points is best along both dimensions, we can see that there are tradeoffs from one of these three scenarios to another; there is gain along one dimension and loss along the other. In optimization terminology we say these three points are *nondominated* because there are no points better than these on all criteria.

The GBML provides the facility to search in the Pareto-optimal set of possible features. This Pareto-optimal set is being generated from crossover and mutation. To make the Pareto optimality mathematically more rigorous, we state that a feature vector x is partially less than feature vector y, symbolically x<p y, when the following condition holds:

$$(x < p\, y) \Leftrightarrow (\forall_i)\,(x_i \leq y_i) \bigwedge (\exists_i)\,(x_i < y_i)$$

This may be mapped to Pareto plane as shown in Figure 3, where Pareto front of *nondominated* points are highlighted in red color.

In the notion of Pareto optimality by multi-objective optimization we used GA in parallel fashion. The methodology used is as follows:

1. Generate chromosome for each feature.
2. Initialize population for each feature.
3. For i=1 to population size For j=1 to feature vector size Compute fitness value.
4. If termination condition satisfied go to Step 10.
5. Crossover.
6. Mutation.
7. Natural Selection.
8. Go to Step 3.
9. Output
10. End

The termination condition as mentioned in Step 4 is a pre-mature termination condition. If the fitness function wasn't improved for *n* consecutive generations then the iteration has been discarded. But there is no case of pre-mature termination case found during experimentation. Experimantally the threshold value of *n is 5.*

The parallelism is obtained here by generating n number of GA based subjectivity classifiers. Based on the principle of survival of the fittest, a few of the feature strings are selected. This parallelism provides the granularity for every feature. The GA based subjectivity classifiers are synchronous in nature. The n numbers of GA based subjectivity classifiers generate their population simultaneously. The fitness value is calculated after every iteration. The optimal solution is selected based on the theory of Pareto optimality. Pareto optimality helps to reach the fittest global solution from local best solution for each feature. The effectiveness of the present technique is observed in the experimental results.

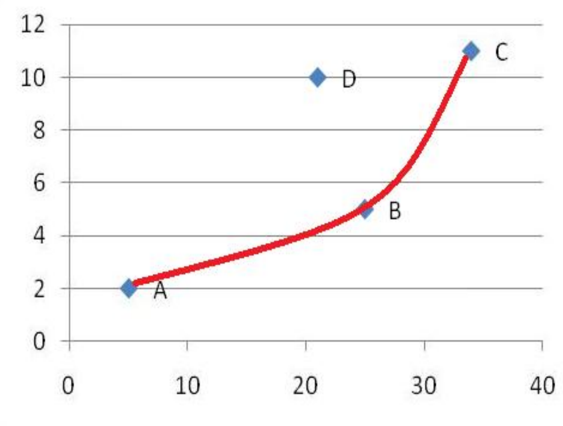


Figure 3. Pareto optimal plane

4.2 Chromosome Representation

The size of the chromosome for every feature varies according to the possible solution vector size. Tentative solutions are made of sequences of genes. Each gene corresponds to word sequence in the sentence to be tagged.

The chromosomes forming the initial population are created by randomly selecting from a dictionary one of the valid tags for each word. For the present task we have used real encoding. A sentence wise feature vector can be represented as.

Example. Imperialism/NNP is/VBZ the/DT source/NN of/IN war/NN and/CC the/DT disturber/NN of/IN peace/NN.

The encoded chromosome is represented in Table 5. The real values are the serial number of the corresponding tag from the POS Tag labeled dictionary. Table 4 reports how real values vary for every feature.

For POS feature values vary for languages as the tag set are different. There are 21 tags and 45 tags in the POS tagset for Bengali and English respectively. For sentiment words from SentiWordNet values are -1 for negative, 0 for neutral and +1 for positive words. For low frequency words features are considered as binary i.e. either a word is low-frequent or not. Any word occurring less than 5 times in the corpus has been considered as a low frequency word. This feature is encoded as a binary feature. Stems from the corpus are listed and the serial number of any stem within the list is used to encode the chro-

Table 4. Dimension of Chromosome Encoding.

Features	Real Values
POS	1-21 (Bengali) / 1-45 (English)
SentiWordNet	-1 to +1
Frequency	-1 to +1
Stemming	0 or 1
Chunk Label	1 to 17176/ 1 to 1235
Dependency Parsing	1-11 (Bengali) / 1-21 (English)
Title of the Document	1-30 (Bengali) / 1-55 (English)
First Paragraph	Varies document wise
Average Distribution	Varies document wise
Theme Word	Varies document wise

mosome. It is basically the range of unique wordforms in any corpus. Chunk label and Dependency parsing is encoded as the POS feature.

Discourse level features varies at each document level. For the three listed discourse level features three different dictionary of first paragraph word, eventually distributed words and theme words have been generated at each document level. Then index numbers from the dictionaries are used to generate the encoded chromosomes.

Table 5. Chromosome Representation.

NNP	VBZ	DT	NN	IN	NN	CC	DT	NN	IN	NN
1	12	6	2	18	2	4	6	2	18	2

4.3 Fitness Evaluation

Fitness function is a performance measure or reward function which evaluates how good each solution is. The following cost-to-fitness transformation is commonly used with GAs.

$$f\left(x\right)=C_{max}-g\left(x\right) when\, g\left(x\right)<C_{max} \quad or\; 0\; Otherwise$$

There are variety of ways to choose the coefficient C_{max}. C_{max} may be taken as an input coefficient, as the largest g value observed thus far, as the largest g value in the current population, or the largest of the last *k* generation.

When the natural objective function formulation is a utility function we have no difficulty with the direction of the function: maximized desired profit or utility leads to desired performance. But still there is some problems with negative utility function as in the particular case it occurs during the fitness calculation of *n* number of features fitness evaluation. To overcome this, we simply transform fitness according to the equation:

$$f\left(x\right)=u\left(x\right)+C_{min} \quad When\, u\left(x\right)+C_{min}>0 \quad or 0 Otherwise$$

For the present problem there is a single fitness function to select the best Pareto optimal plane.

4.4 Crossover

Crossover is the genetic operator that mixes two chromosomes together to form new offspring. Crossover occurs only with some probability (crossover probability). Chromosomes that are not subjected to crossover remain unmodified. The intuition behind crossover is the exploration of new solutions and exploitation of old solutions. GAs construct a better solution by mixing the good characteristic of chromosomes together. From the *n* solution strings in the population (simply n/2 pairs), certain adjacent string pairs are randomly selected for present crossover technique. In the standard GA, we use single-point crossover by selecting a pair of strings and swapping substrings at a randomly. No adaptive or probabilistic crossover technique has been used for current experimentation.

4.5 Mutation

Each chromosome undergoes mutation with a probability μ_m. The mutation probability is also selected adaptively for each chromosome as in [26]. The expression for mutation probability, μ_m, is given below:

μ_m= $k_2 \times \frac{(f_{max}-f)}{(f_{max}-\bar{f})}$ if $f > \bar{f}$,

μ_m= k_4 if $f > \bar{f}$, Here, values of k_2 and k_4 are kept equal to 0.5. This adaptive mutation helps GA to come out of local optimum. When GA converges to a local optimum, i.e., when $f_{max} - \bar{f}$ decreases, μ_c and μ_m both will be increased. As a result GA will come out of local optimum. It will also happen for the global optimum and may result in disruption of the near-optimal solutions. As a result GA will never converge to the global optimum. The μ_c and μ_m will get lower values for high fitness solutions and get higher values for low fitness solutions. While the high fitness solutions aid in the convergence of GA, the low fitness solutions prevent the GA from getting stuck at a local optimum. The use of elitism will also keep the best solution intact. For a solution with the maximum fitness value, μ_c and μ_m are both zero. The best solution in a population is transferred undisrupted into the next generation. Together with the selection mechanism, this may lead to an exponential growth of the solution in the population and may cause premature convergence.

Here, each position in a chromosome is mutated with probability μ_m in the following way. The value is replaced with a random variable drawn from a Laplacian distribution, $p\left(\epsilon\right)\alpha\; e^{-\frac{|\epsilon-\mu|}{\delta}}$, where the scaling factor δ sets the magnitude of perturbation. Here, μ is the value at the position which is to be perturbed. The scaling factor δ is chosen equal to 0.1. The old value at the position is replaced with the newly generated value. By generating a random variable using Laplacian distribution, there is a non-zero probability of generating any valid position from any other valid position while probability of generating a value near the old value is more.

4.6 Natural Selection

After we evaluate population's fitness, the next step is chromosome selection. Selection embodies the principle of 'survival of the fittest'. The mutant fittest chromosomes are selected for reproduction. A few poor chromosomes or lower fitness chromosomes may be selected. Each solution having a probability equal to its fitness score divided by the sum of the total solutions scores in the generation. The top *n* solutions at each generation automatically retained and carried over to the next generation. Roulette wheel selection is used to implement the proportional selection strategy.

5 Experimental Results

We have used Java API for Genetic Algorithm[10] application. Approximately 70% of every corpus has been used for training purpose and the rest 30% has been used for testing purpose. The following parameter values are used for the genetic algorithm: population size=50, number of generation=50.

[10] http://www.jaga.org/

Table 6. Results of final GA based classifier.

Languages	Domain	Precision	Recall
English	MPQA	90.22%	96.01%
	IMDB	93.00%	98.55%
Bengali	NEWS	87.65%	89.06%
	BLOG	90.6%	92.40%

The overall precision and recall values of the GBML based subjectivity classifier are shown in Table 6 for all the corpora selected for English and Bengali. It is observed that subjectivity detection is trivial for review corpus and blog corpus rather than for news corpus. In news corpus there is more factual information than review or blog corpus that generally contain people's opinion. Thus subjectivity classification task is domain dependent. But the proposed technique is domain adaptable through the use of natural selection. The difference of GA-based classifier with others statistical system is that a whole sentence could be encoded in GA and could be used as a feature. In other classifier system n-gram method has been followed. The fixed size of n in the n-gram does not fit into the variable string length of an input string.

5.0.1 Comparison

Present GBML sytem outperform than existing Subjectivity systems in literature. The CRF based subjectivity classification system as we reported previously in [12] perform experiment on same set of Bengali and English corpus and reported accuracy of the system was 72.16% and 74.6% for the news and blog domains respectively. In the previous Subjectivity Detection study the subjectivity problem was modeled as a text classification problem that classifies texts as either subjective or objective depending upon various experimentally choosen features. This paper illustrates a Conditional Random Field (CRF) based Subjectivity Detection approach tested on English and Bengali multiple domain corpus. Standard machine learning (ML) techniques needs rigorous permutation and combination wise experimentation to find out the best set of features for any particular problem definition. The GBML based methodology as we proposed here provide a best solution as natural selection method to overcome the classical feature engineering. The CRF based system was tested on the same dataset as reported in Table 7. Besides the novelty over feature engineering GBML technique is a better solution as it need no human interruption to find out best fetures and it choose the best fetures through *natural selection.*

Table 7. Results of final CRF-based subjectivity classifier.

Languages	Domain	Precision	Recall
English	MPQA	76.08%	83.33%
	IMDB	79.90%	86.55%
Bengali	NEWS	72.16%	76.00%
	BLOG	74.6%	80.4%

In compare to the previous subjectivity classification systems on MPQA corpus the present GBML system has an increment of near about 4.0%. The reported highest accuracy on MPQA using Naive Bayse was 86.3% as reported in [27]. The authors used Naive Bayes sentence classifier and the reported accuracy was as reported in Table 8.

Table 8. Results of previous subjectivity classifier on MPQA.

Languages	Domain	Precision	Recall
English	MPQA	86.3%	71.3%

The accuracy of previous subjectivity detection on the same movie review corpus is 86.4% reported in [28]. The authors proposed a interesting machine-learning method that applies text-categorization techniques to just the subjective portions of the document. Extracting these portions are then categorized using efcient techniques for finding minimum cuts in graphs to incorporate the cross-sentence contextual constraints. To capture cross-sentence contextuality we prefer the theme word features in present GBML based technique. Two standard machine learning (ML) techniques used in [28] as Naive Bayes (NB) and Support Vector Machine (SVM). The reported accuracy of the subjectivity system was as reported in Table 9.

Table 9. Results of previous subjectivity classifier on IMDB.

Classifier	Reported Accuracy
NB	86.40%
SVM	86.15%

6 Conclusion

Application of machine learning algorithms in NLP generally experiments with combination of various syntactic and semantic linguistic features to identify the most effective feature set. Here we viewed this as a multi-objective or multi-criteria optimization search problem. The experiments in the present task start with a large set of possible extractable syntactic, semantic and discourse level feature set. The fitness function calculates the accuracy of the subjectivity classifier based on the feature set identified by natural selection through the process of crossover and mutation after each generation. In the present paper we use GBML to identify automatically best feature set based on the principle of natural selection and survival of the fittest. The identified fittest feature set is then optimized locally and global optimization is then obtained by multi-objective optimization technique. The local optimization identify the best range of feature values of a particular feature. the Global optimization technique identifies the best ranges of values of given multiple feature.

In the present experimental setup it harder to identify feature wise performance value. The GBML identifies the best feature set and their optimal range value by *natural selection.* The present experiment by us is to determine contribution of each feature to the overall subjectivity problem.

The performance of the present multiple objective optimization tecnique based GBML strategy easily estublised that it is worthy than available ML techniques so far used in NLP. The novelty of the present task is not only towards finding the better way to detect subjectivity moreover it depicts a generation change in ML techiques so far used in NLP.

7 Reference

1. Philip J. Stone. The General Inquirer: A Computer Approach to Content Analysis. The MIT Press, 1966.
2. Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. A comprehensive grammar of the English language. Longman, 1985.

3. Wilks Yorick and Bein Janusz. Beliefs, Points of View,and Multiple Environments. In Cognitive Science 7. pp. 95-119 . 1983.
4. Janyce M. Wiebe and William J. Rapaport. A computational theory of perspective and reference in narrative. In Proceedings of the Association for Computational Linguistics (ACL), pages 131–138, 1988.
5. Janyce M. Wiebe, Rebecca F. Bruce, and Thomas P. O'Hara. Development and use of a gold standard data set for subjectivity classifications. In Proceedings of the Association for Computational Linguistics (ACL), pages 246–253, 1999.
6. A. Aue and M. Gamon, "Customizing sentiment classifiers to new domains: A case study," In the Proceedings of Recent Advances in Natural Language Processing (RANLP), 2005.
7. Vasileios Hatzivassiloglou and Kathy McKeown. 1997. Predicting the semantic orientation of adjectives. In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL-97), pages 174–181.
8. Janyce Wiebe. 1990. Recognizing Subjective Sentences: A Computational Investigation of Narrative Text. Ph.D. thesis, State University of New York at Buffalo.
9. Janyce Wiebe. 2000. Learning subjective adjectives from corpora. In Proceedings of the Seventeenth National Conference on Artificial Intelligence (AAAI-2000), pages 735–740.
10. Kushal Dave, Steve Lawrence, and David M. Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In Proceedings of WWW, pages 519–528, 2003.
11. A. Das and S. Bandyopadhyay. Theme Detection an Exploration of Opinion Subjectivity. In Proceeding of Affective Computing & Intelligent Interaction (ACII 2009b).
12. A. Das and S. Bandyopadhyay (2009a). Subjectivity Detection in English and Bengali: A CRF-based Approach., In Proceeding of ICON 2009, December 14th-17th, 2009, Hyderabad.
13. J. H. Holland. 1975. Adaptation in Natural and Artificial Systems. The University of Michigan Press, AnnArbor.
14. D. E. Goldberg. 1989. Genetic Algorithms in Search, Optimization and Machine Learning. Addison Wesley, New York.
15. Kraft, D.H. et. al. "The Use of Genetic Programming to Build Queries for Information Retrieval." In Proceedings of the First IEEE Conference on Evolutional Computation. New York: IEEE Press. 1994. PP. 468-473.
16. Martin Bautista and M.J. "An Approach to An Adaptive Information Retrieval Agent using Genetic Algorithms with Fuzzy Set Genes." In Proceeding of the Sixth International Conference on Fuzzy Systems. New York: IEEE Press. 1997. PP.1227-1232.
17. Abbasi, A., Chen, H., and Salem, A. "Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums," ACM Transactions on Information Systems, 26(3), 2008, no. 12.
18. Vasileios Hatzivassiloglou and Janyce Wiebe. Effects of adjective orientation and gradability on sentence subjectivity. In Proceedings of the International Conference on Computational Linguistics (COLING), pages 299-305, 2000.
19. Paula Chesley, Bruce Vincent, Li Xu, and Rohini Srihari. Using verbs and adjectives to automatically classify blog sentiment. In AAAI Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW), pages 27–29, 2006.
20. Tetsuya Nasukawa and Jeonghee Yi. Sentiment analysis: Capturing favorability using natural language processing. In Proceedings of the Conference on Knowledge Capture (K-CAP), pages 70-77, 2003.
21. Andrea Esuli and Fabrizio Sebastiani. SentiWordNet: A publicly available lexical resource for opinion mining. In Proceedings of Language Resources and Evaluation (LREC), 2006.
22. A. Das and S. Bandyopadhyay (2010a). SentiWordNet for Bangla., In Knowledge Sharing Event-4: Task 2: Building Electronic Dictionary , February, 2010, Mysore.
23. A. Das and S. Bandyopadhyay (2010b). Morphological Stemming Cluster Identification for Bangla., In Knowledge Sharing Event-1: Task 3: Morphological Analyzers and Generators, January, 2010, Mysore.
24. A. Ghosh, A. Das, P. Bhaskar, S. Bandyopadhyay (2009). Dependency Parser for Bengali : the JU System at ICON 2009., In NLP Tool Contest ICON 2009, December 14th-17th, 2009a, Hyderabad.
25. Giuseppe Carenini, Raymond Ng, and Adam Pauls. Multi-document summarization of evaluative text. In Proceedings of the European Chapter of the As-sociation for Computational Linguistics (EACL), pages 305–312, 2006.
26. M. Srinivas and L. M. Patnaik. 1994. Adaptive probabilities of crossover and mutation in genetic algorithms. IEEE Transactions on Systems, Man and Cybernatics, 24(4):656–667.
27. Janyce Wiebe and Ellen Riloff. Creating Subjective and Objective Sentence Classifiers from Unannotated Texts. In Proceeding of International Conference on Intelligent Text Processing and Computational Linguistics, Mexico City, Pages 475–486, 2006.
28. Bo Pang and Lillian Lee. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. In Proceedings of the Association for Computational Linguistics (ACL), pp. 271–278, 2004.

Private State in Public Media
Subjectivity in French Traditional and On-line News

Anne Küppers and **Lydia-Mai Ho-Dac**[1]

Abstract. This paper reports on ongoing work dealing with the linguistic impact of putting the news on-line. In this framework, we investigate differences in one traditional newspaper and two forms of alternative on-line media with respect to the expression of authorial stance. Our research is based on a comparable large-scale corpus of articles published on the websites of the three respective media and aims at answering the question to what extent the presence of the author varies in the different media.

1. Is it a matter of amount and mode of the author's presence?
2. Is it a matter of lexical choice and diversity?
3. If this were the case, what expressions are used in the respective media?

Our endeavour will be a methodological one. We firstly present our data, and thus describe the different news media included in our study, and the diverse computer aided and manual production steps we performed in order to build up the corpus. Secondly, we outline our working hypotheses that are linked to the chosen types of media and describe the theoretical framework within which they are situated. Thirdly, we present our research method as well as some first results and insights gained throughout the pilot study of our data.

1 Corpus

The main objective of our research is to contrast traditional newspaper language with the language used in alternative forms of journalism in order to determine whether we have to do with distinct genres, or merely different text types.

We therefore created a large-scale corpus consisting of articles published in one traditional newspaper and in two alternative written on-line mass media. Texts included in the corpus have been published between 2005 and 2009 and were collected directly from the respective website's archives. The media are briefly presented in the following and table 1 gives an overview of the sections included in each data set. Sections are chosen on the basis of comparison of the topics dealt with in order to ensure a higher degree of comparability of the different sub corpora.

The first data set consists of articles published in one of the principal Belgian, French-speaking traditional reference newspapers, namely *Le Soir*. This liberal, most read supra-regional newspaper was first published in 1887. Texts included in our corpus have been published in the printed or the on-line version of the newspaper.

The second data set is composed of articles published on the website of the French independent journalism project *Rue89*. This project started in 2007 and aims at unifying professional journalism and Internet culture. It works with a committee of professional journalists and young reporters ensuring a good portion of articles and the reviewing of texts submitted by external domain specialists.

Table 1. Newspaper Sections included in the Data Sets

Le Soir	Rue89	AgoraVox
News	World News	International News
		Europe
	Politics	Politics
	Society	Society
Culture	Culture	Culture
	Media	Media
		Religion
		Bizarre
		Life and Style
		People

The third data set is made up of articles published on the French alternative on-line information platform *AgoraVox*. This citizen press website was created in 2005 and follows the principle of editorial democracy, that is to say that any Internet user can subscribe and contribute articles. Reviews are done by members of the committee, composed of some anchormen and the sub-editors, who are potentially all members having published at least 4 articles on the AgoraVox website.

2 Data Processing

The creation of the corpus was realised through several steps. Processing procedures are different for the three media because the constitution of their on-line archives is not uniform. And, as there are no ready-made programs for automatically extracting articles from different web pages ([44]), the data collection was a first challenge. We will briefly describe the basic legs that were realised for the entire corpus.

In a first step, all articles are collected directly from the websites' archives by means of Perl scripts. The script extracts the list of articles published under a certain URL and saves them in a separate folder for each medium. A second script extracts the articles mentioned in these folders and converts the original files of different source codes into XML format. During this step, the script keeps track of information such as the medium, the section, the author, the title and the date of publication, whenever these are available. Besides, we already partially remove wire copies from our data sets, as they would falsify our analyses on the expression of authorial stance. For the same reason, interviews, poems and songs are removed man-

[1] Université catholique de Louvain (UCL), Institut Langage et Communicatiton (IL&C), Belgium, email: anne.kueppers@uclouvain.be, lydia.hodac@uclouvain.be

ually in a subsequent step.

All data are encoded following the TEIP5 model and are cleaned by means of computer aided and manual control sequences. The resulting files include information about the text structure and give titles, subtitles, formatted lists and paragraphs. They also indicate citations[2], bold or italic printing, a set of meta-information concerning the corpus, and each unique article. All this supplementary information is displayed by XML tags.

3 Hypotheses

On the basis of this multi-layered corpus of written mass media, we aim at bringing to light similarities and divergences in terms of linguistic and structural dimensions. Concerning the language use in the three different media, we have the following hypotheses:

1. Subjectivity is expressed differently in the three media, namely with respect to the amount and lexical choice of subjective expressions.
2. Articles in the alternative media are more subjective than those published in the traditional newspaper, as the author expresses more overtly her/his opinion and thoughts in the former ones.
3. On average, the writing style in articles published in AgoraVox is even more subjective than the style in those published in Rue89.

We formulate these hypotheses on the basis of the working processes prevailing in the editorial departments of the different media and their respective 'philosophy'. The traditional newspaper's data set is composed of two different types of articles, those published in the paper version and those published on the website. While most of the articles recorded for the printed version are based on investigation and are entirely written by one journalist, most of the articles published on the website are slightly modified wire copies that are simply adapted to the editorial line ([20]). This is a consequence of the prevailing working guideline for the web journalists: 'put the news on-line as fast as possible, be the first to publish, and get the scoop'.[3] This principle influences the writing style, which we assume to be less individual and thus less subjective. As a consequence, the presence of the on-line articles in our data set will probably have an impact on the amount of subjective expressions, their mode and lexical choice.

The two alternative media included in our research differ from the traditional newspaper in some points: First, these media are solely published on the Internet. Second, they work with professional journalists, domain specialists and non-professional Internet users. Third, they tend to cover current topics, but not to get the scoop at all costs. Fourth, the alternative media do not intend to cover all actual topics, but just those that seem relevant to the authors – either for themselves, or with respect to their audience.

Concerning their working mode and philosophy, Rue89 defines itself as being more comparable to a radio station than to a traditional newspaper with regard to reactivity, the absence of deadlines, the exchange of participants, and the informal writing style ([6]).

Comparing Rue89 and AgoraVox, the parameter of professionalism of the authors might influence the degree of subjectivity and the way it is expressed. Members of Rue89's editorial board are professional journalists and authors submitting articles to the website are professional journalists as well, or at least so called domain specialists ([6]). Topics to be dealt with in Rue89 may be suggested by Internet users, but the latter do not participate in contributing content themselves. Articles published in AgoraVox are written by professional journalists, domain specialists and non professional web users, and the review committee is composed of professional and non-professional writers.

We assume that the language style differs between the two alternative media, with Rue89 occupying an intermediate position between Le Soir and AgoraVox. More concretely, we suppose that the writing style in articles published in AgoraVox is more individual, and more subjective due to the articles written by individual web users that are not professional journalists or domain specialists. We hypothesize the intermediate position of Rue89 due to the professionalism of authors working on the project on the one hand (which is not the case for AgoraVox) and to the independent topic choice and the time and investigation for the recording of articles on the other (which is not the case for Le Soir, at least for the part of the on-line articles).[4]

Furthermore, we are interested in outlining whether we could speak of distinct genres or text types when comparing different types of news media. Accounting for the a priori difference between the three data sets, we aim to outline whether the supposed differences effectively exist. We do not intend to point out potential differences between on-line and printed journalistic genres like bulletin, reportage, editorial or comment ([1], [37]), but to detect evidence on a more general level, namely between the three types of media under consideration.[5]

As all data sets belong to the domain of journalism, we cannot presume that the three media belong to different genres. But we expect discrepancies between the data sets that are due to (1) the professionalism and education of authors writing for Le Soir, (2) the aim to diffuse objective information in the sense of reflecting news without judging in Le Soir, (3) the aim of Rue89 and AgoraVox to report differently from the traditional press, i.e. not following neither a particular editorial line, nor a given deadline for article publishing, (4) the aim to make non-professional writers participate in news coverage as is the case for AgoraVox. If the distinction of genres turns out to be too general for our purpose of distinguishing three types of journalese, we will still try to outline representative text types for each of them.[6]

Whether we can speak of different genres or text types when comparing the three media under consideration is tested by the phenomenon of speaker stance.[7] In order to outline the expression of subjectivity in our newspaper corpus, we use a twofold method including deductive and inductive quantitative approaches which are presented in section 5, as well as a qualitative analysis.

[2] As compared to interviews, single citations included in articles are kept and tagged in order to easily identify and exclude when wanted, as the choice of a citation reflects a personal state, not only of the source, but also of the author, and citations thus may serve our subsequent analyses of subjectivity.

[3] As the traditional newspaper aims to treat all actual topics also in the paper version, journalists may have time pressure due to deadlines when recording for the printed support as well, depending on the topic of the article.

[4] Texts for which this might be the case when regarding the paper version of Le Soir, such as wire copies or newsflashes, have been excluded from our corpus as explained in section 2.

[5] As studies on subjectivity are often based on corpora build up of texts dealing with the same topic, we plan to compare different sub corpora in subsequent studies.

[6] In the domain of corpus linguistics, the term *text type* was first introduced by Biber ([9]: 68) who defines it by means of inner-textual linguistic characteristics, as opposed to *register* (previously *genre*), which is defined in terms of external and cultural criteria linked to the author's purpose.

[7] To determine in a more general way whether we effectively have to do with different genres or text types will need further investigation concerning other linguistic dimensions, but our pilot study already allows for the detection of tendencies.

4 Theoretical Framework and Research Method

Our quantitative analysis is based on two axes, namely (1) discourse organisation through initial position (i.e. the first preverbal zone of a given sentence) and (2) subjectivity through PSEs. While the first focuses on the evaluation of the typological differences between the media, the second is devoted to subjectivity in order to observe the variation of the author's presence in the texts. Nevertheless, the two axes interact: Not only subjective language, but also the order of information reflects speaker stance, namely by choosing the information included, by mentioning certain aspects before others, or by linking different texts parts (phrases, sentences, paragraphs).

Before we introduce our research methods, we briefly sketch the theoretical principles our research is based on. The work in progress presented is situated in a corpus linguistic framework of discourse analysis. The present paper aims to outline first tendencies in our data. We describe and evaluate the typological differences between the three media by applying a corpus-based methodology providing a description of the global discourse organisation of our data sets and the expression of subjectivity by certain predefined cues (section 4.3).

4.1 Discourse Organisation through Initial Position

Because of the apparent incompatibility between the qualitative nature of discourse analysis and the quantitative requirements of corpus linguistics, discourse organisation is usually difficult to study by means of corpus linguistic methods ([8]). Ho-Dac [27] proposes a method providing a solution to this incompatibility, allowing for a data-driven approach to discourse organisation based on automatic tagging and quantitative analysis of the discourse roles of sentence-initial elements in different text positions given by the layout. The theoretically-based hypothesis is that the initial position – defined as the starting point of the message and composed of the first elements that the reader receives – has an important function in discourse organisation. The analysis of the distribution of these elements according to their text position gives an overview of the textual organisation of different text types. Therefore, two text positions are distinguished: *P1* corresponding to sentences introducing a paragraph, and *P2* corresponding to intraparagraphic sentences. Elements in P1 are by definition associated with a paragraph break, i.e. a visual cue of discontinuity. As a consequence, they have a greater capacity of signalling high-level discontinuities and orienting high-level segments. Because discourse organisation is complex and texts are organised according to different structuring principles, we have to consider different types of discourse segments. In this study we focus on cues that potentially signal topical continuity, rhetorical articulation, setting discontinuity, and textual discontinuity.

Topical continuity is outlined by means of co-referential grammatical subjects covering pronouns, possessive noun phrases, reiterations, and detached appositions in initial position. Several studies in cognitive linguistics showed that linguistic means available to refer to a given entity already mentioned in the text are associated with different degrees of accessibility (e.g. [41], [2], [23]). On their basis we assume that (1) co-referential expressions, especially when occurring in grammatical subject position, have an instructional meaning indicating topical continuity, and that (2) the type of this expression indicates different levels of topical continuity. For example, a first person personal pronoun in grammatical subject position indicates a strong topical continuity while reiteration may be used to reintroduce a topic or to reinforce a topical continuity when there is a discourse shift e.g. a paragraph break, a setting or a textual discontinuity cue ([46]). Another topical continuity cue is apposition, which is an attributive construction communicating supplementary information on a given sentence constituent from which it is syntactically detached. Concerning discourse organisation, and especially when occurring in initial position just before the first grammatical subject, appositions may indicate topical continuity just like to referential links ([18]), and it has been shown that the more narrative a text, the more appositions and pronouns occur in P1 ([27]).

For **rhetorical articulations** we only consider connectives occurring in absolute first position. When introducing a sentence or a paragraph, they may acquire a high-level discourse function in order to signal a rhetorical articulation taking place inside in the course of a given continuity (concerning topic or setting). Ho-Dac ([27]) shows that the more argumentative a text, the more connectives occur in initial position.[8]

Setting discontinuity is outlined by means of detached setting adverbials. When occurring in sentence-initial position, setting adverbials may orient the reader by indicating the domain of applicability within which the following proposition holds (e.g. [13], [21] and [22]). In the present study, we focus on time, space, and notional adverbials, i.e. elements which set a notion that may be a domain of knowledge (*in linguistics*), a defined object (*concerning the case of adverbials*), a specific point of view (*in line with Halliday*), etc. The text part introduced by these adverbials is labelled discourse frame and characterized by temporal, spatial, or notional homogeneity ([17]). Ho-Dac ([27]) shows that the more descriptive a text, the more setting adverbials occur in initial position.

Concerning **textual discontinuity**, we focus on sequencers (linking adverbials and grammatical subjects introducing items) that serve to indicate discourse organisation attributing limits of different text parts and information sources by explicitly indicating the position of a given segment in discourse (e.g. *Firstly,... Secondly,... Finally,... Moreover,... Besides,... etc*).

4.2 Private State

Subjectivity generally refers to the expression of personal state, covering devices of opinion, evaluation, attitude and emotion or sentiment when generally speaking. Depending on the underlying theory and the linguistic means at focus, the phenomenon is amongst others designated as *stance* ([9], [11]), *appraisal* ([38], [51]), *hedging* ([35], [29]), *commitment* ([47]), *private state* ([42]) or *evaluation* ([5]).[9] Diverse means can serve to express subjectivity in texts. Usually any subjective element is linked to its *emitter* who can either be the writer or some other person referred to or cited in the text. In the same way, subjective elements are generally linked to a *goal* that the personal state relates to. In line with Thompson and Hunston ([48]), we define private state as

> the broad cover term for the expression of the speaker's or writer's attitude or stance towards, viewpoint on, or feelings about the entities or propositions that he or she is talking about. That attitude may relate to certainty or obligation or desirability or any of a number of other sets of values.[10]

[8] Connectives are more often used to link several continuous clauses inside a given sentence.

[9] For further description see Bednarek ([5]).

[10] While the defined phenomenon is labelled *evaluation* by Thompson and Hunston ([48]), we use the terms *subjectivity*, *stance* and *private state* as equivalents to it in the course of this article.

Besides the lexical choice (including single words, collocations and complex phrases), also morphology and syntax can communicate a personal state in written texts.[11] It is very important to note that the discrete occurrence of a subjective expression is not by force used in its subjective meaning, which is true for objective devices as well. Depending on the context, an a priori subjective expression can be used objectively and vice versa. The distinction of subjective and objective elements thus demands more detailed and qualitative analyses. Following Wiebe et al. ([52]: 281f), we therefore speak of Potential Subjective Elements (PSE) to refer to those "linguistic element[s] that **may** be used to express subjectivity" by means of their primary meaning (our emphasis). Whether a PSE is effectively used subjectively is dependent on the context of a given utterance.

4.3 Potential Subjective Elements

By the second axis, we explore the use of Potential Subjective Elements (PSE) in the three data sets, accounting for occurrences of first person personal and possessive pronouns[12], stance adverbials, *it*-extrapositions, cleft sentences, and hapax, i.e. words that occur just once in a given data set.

(i) Ces fameuses années 68–70, qui **nous** submergent aujourd'hui, ça commence à **m'**énerver. `Rue89_2850`
(ii) **Il faut donc** pour les africains francophones abandonnés le F CFA, fabriqué en France - près de Clermont-Ferrand... `AgoraVox_2709`
(iii) Pour Alain Menand, **il est de toute façon hasardeux** de prétendre "classer" les différentes licences : [...] `Rue89_3876`
(iv) **Je** ne reviendrai pas sur les questions rhétoriques toujours aussi efficaces. `AgoraVox_3667`
(v) Alors que l'anthropologie et la sociologie ont souvent pensé les cultures selon des modèles de groupe **nous** verrons ici ce que le concept de culture doit à la prise en compte des besoins de l'individu au plan personnel [...] `AgoraVox_2913`
(vi) Vendredi dernier dans **nos** colonnes, les recteurs de l'Université libre de Bruxelles (ULB), Pierre de Maret, et de la Vrije Universiteit Brussel (VUB), Ben Van Camp, signaient une Carte blanche. `LeSoir_5359`
(vii) **On** assiste ainsi à une soirée organisée en l'honneur des Amis américains de Versailles dans la Galerie des Glaces du fameux château. `LeSoir_4341`
(viii) **C'est l'amer constat que** l'**on** peut faire soit qu'**on** y habite ou qu'**on** y arrive pour la première fois dans cette ville qui jadis, présentait fière allure. `AgoraVox_3584`
(ix) Car **il serait évidemment bien dangereux de** se replier frileusement sur les égoïsmes nationaux d'antan. `Rue89_987`
(x) **C'est d'abord parce qu**'ils gardent au début l'espoir insensé d'un miracle, et qu'ensuite il est trop tard. `AgoraVox_1451`

The use of a first person pronoun is one of the most conspicuous means used to express subjectivity as it "refers to the act of individual discourse in which it is pronounced, and by this it designates the speaker" ([7]: 226) (examples (i), (iv)-(vi)). Several linguistic investigations on speaker stance – most of them including English as (at least one of) the language(s) under investigation and focusing on academic and scientific discourse – outline the importance of the choice of personal pronouns in order to express the degree of involvement in relation to the propositional content (e.g. [24], [25], [32], [33], [34], [49]). The third person singular pronoun can fulfill the communication of subjectivity when occurring as grammatical subject, too. In French, this is the case for the third person singular pronoun *on* used as an alternative to the first person plural pronoun *nous*, and taking over the speaker-inclusive meaning (examples (vii)-(viii)).

In addition to pronouns we include stance adverbials and two special constructions that are potentially related to subjectivity: *it*-extrapositions and cleft sentences. We are especially interested in these cues of subjective language as they offer the author the possibility to express stance in an indirect way.

Stance adverbials (like *évidemment* in example (ix)) can be used for reasons linked to content (e.g. when information about a topic is not sufficiently accessible), or for interpersonal reasons (e.g. when the author does not want to impose a personal point of view to the readers) ([30], [31]). Originally, stance adverbials have been defined as a means of hedging by rendering the affiliation of an object to a certain category fuzzier ([35]), while they are accredited now a more general function, including the expression of attitude, emotion and opinion. The particularity of *it*-extrapositions is that they express a subjective meaning while at the same time communicating a certain degree of distance between the author and the propositional content (example (ix)). As Charaudeau ([14]) points out, the use of *it*-extraposition is very frequent in journalese, as these formulations seem to be less subjective, so that we speak of *constructed objectivity* ([16]: 504). Cleft sentences can also express an indirect judgement on the propositional content of a message ([36]). Their main function is to focus on an extracted element that is detached from the other sentence components in order to be emphasised (examples (viii), (x)). By choosing a cleft sentence structure, the author can communicate the accentuation of a given propositional content.

Lastly, we focus on single word occurrences, so called hapax legomena.[13] Following Wiebe et al. ([52]), we assume that one word occurrences are especially interesting when exploring subjectivity in discourse.[14] Existing studies investigating stance by a corpus-based approach are all based on English language data ([28], [15], [10], [19], [40] amongst others). Until now – to our knowledge – there are no such analyses based on French large-scale corpora.

5 Quantitative Analysis

Our quantitative analyses of the different cues presented in section 4 are based on automatic tagging which has already been used and evaluated twice ([27] and [39]).

5.1 Cues Marking

The quantitative analyses presented are based on an automatic labelling of features concerning discourse organisation on the one

[11] In spoken or direct discourse, still other indicators like intonation, gesture, and mimicry can perform this task.

[12] The occurrences include those of the third person singular pronoun *on* in the use of *nous* (*we*).

[13] Instead of focusing just on the most frequent and statistically significant word occurrences in a given data set, what has been custom in corpus linguistics for several years, more recent studies also take into account the least frequent phenomena, namely hapax ([4], [50], [3], [52]).

[14] The inclusion of hapax in large-scale corpus investigation could also be a first step to work against the criticism mentioned in qualitative research (e.g. [5]) that bottom-up data mining is not an adequate method to outline expressions of subjectivity, as it could never detect all occurrences of subjective language due to the undefined and unlimited diversity of possible formulations.

hand, and subjectivity on the other hand. This labelling is based on the results of a POS tagging (*TreeTagger*, [45]), a syntactic parser (*Syntex*, [12]), and on layout information directly extracted from the TEI encoding of the corpus (section 2).

Concerning discourse organisation, the automatic marking extracts a selection of potential organisational cues occurring in initial position, distinguishing connectives occurring in absolute first position, detached elements, and grammatical subjects. These elements are automatically characterized by their POS, their function (setting vs. textual adverbials, sequencers, appositions, etc.), the semantic meaning (e.g. temporal, spatial, and notional setting adverbials), and the properties of reiteration (when an NP's head restates a noun already mentioned in a given section). Moreover, these elements are associated with their textual position, i.e. P1 (if their host sentence introduces a paragraph) or P2 (if the host sentence is intraparagraphic).[15] The automatic characterisation is based on a Perl script (1) delimiting the first preverbal zone for all sentences, (2) identifying all syntactic blocs composing the preverbal zone (based on *Syntex* results), 3) categorising each bloc by applying a set of regular expressions associated with lexical lists concerning functional and semantic features.

Detecting the PSEs, we use the POS database in order to outline hapax legomena. Concerning cues marking, we have adapted the Perl script used for discourse organisation to extract first person pronouns (*j*e, *n*ous, and *o*n in subject position and *m*e/*m*', *m*oi, *n*ous, *s*e/*s*' in other positions), and possessive NPs (*m*on/*ma*/*mes X*, *n*os/*notre* X). Moreover, three other cues, namely stance adverbials and *it*-extrapositions as well as cleft sentences, are automatically extracted, based on the POS tagging, a lexical list for the former, and syntactic patterns for the latter.

Table 2. Extracted Cues

DISCOURSE ORGANISATION	
SETTING	Setting adverbials
SEQ	sequencers
CONNECT	connectives (coordinations, adverbs)
APPOS	appositions
COREF_r	proper nouns, definites, demonstratives, possessive and undetermined NPs with a syntactic head reiterating a noun already mentioned
COREF_p	pronoun and possessive NPs
POTENTIAL SUBJECTIVE ELEMENTS (PSE)	
STANCE	stance adverbials
LOCpro	1st person personal pronouns (including *on*)
LOCposs	NPs with 1st person possessive determiner
IT-ex	*it*-extrapositions
CLEFT	cleft sentences

5.2 Frequency Analysis

We firstly describe the main differences between the three data sets in terms of layout, discourse organisation, and occurrences of PSEs. Secondly, we expose occurrence frequencies of the diverse cues for each data set by using contingency tables (comparing the data sets two by two) and the log-likelihood ratio (henceforth LL[16]) in order to measure the significant relative frequency differences between them.

The higher the LL value, the more significant the difference is between two frequency scores. In this study principally aiming at describing main tendencies, we only focus on LL corresponding to $p < 0.0001$ (i.e. higher than 15.13). The resulting tendencies will or will not support our hypotheses and will constitute the starting point for further detailed, quantitative and qualitative analyses.[17] Before presenting tendencies concerning subjectivity in our corpus, the next section describe the linguistic characteristics of the three media in terms of layout and discourse organisation.

6 Interim Results and Tendencies

This section presents the first results and insights gained throughout the quantitative analysis. Its main concern is to expose observed general tendencies on the basis of frequency analyses of the cues for each data set and LL statistics for their comparison.

6.1 Linguistic Characterisation of the Three Media

To give an overview of the **general characteristics** of the three data sets, we describe in the following their layout and lexical diversity. While the first has to do with discourse organisation, the second may be linked to subjectivity ([52]). Table 3 suggests different units of measurement, more or less related to layout, in order to describe their size and textual segmentation.

Table 3. Layout Segmentation

	Rue89	AgoraVox	Le Soir	total
Words	2,187,333	**3,281,208**	2,744,270	8,212,811
Headings	687	**896**	715	2,298
Articles	3,879	4,368	**5,873**	14,120
Words/Article	564	**751**	467	582
Sentences/Article	50	**68**	42	53
Paragraphs/Article	112	**14**	8	11
Sentences/Paragraph	2.7	**3.5**	2.5	2.9

AgoraVox is the largest data set concerning the overall number of words, paragraphs, and headings. It also contains the longest articles (on average 751 words/text) and longer paragraphs as compared to the two other media. In contrast, Le Soir shows shorter articles (on average 467 words/text) and paragraphs. As a consequence, Le Soir is the larger data set with respect to the total number of articles (5,873). Rue89, the smallest sub corpus, occupies an intermediate position.[18] Paragraph size may play an important role in discourse organisation, allowing for simple structures in short paragraphs as compared to longer ones.[19]

Lexical diversity is evaluated in the present by using the type/token ratio based on the idea that the more types as compared to the number of tokens, the more varied is the vocabulary. And the closer to 1 the ratio, the more lexically diverse is the data set. Lexical diversity might be linked to authorial presence and the expression of private state in the text as outlined in sections 3 and 4.3.

[15] For more methodological details see Ho-Dac ([27]).

[16] See [43] for details on the use of this ratio for large-scale corpus comparison.

[17] Our research prospects include the investigation of more quantitative analyses on modal expressions, stance adverbials, and adjectives expressing subjectivity, as well as *verba dicendi et sentiendi*. We also intend to carry out qualitative analyses on occurrences of all mentioned subjectivity cues.

[18] Rue89 is the youngest media founded in 2007, explaining the comparatively smaller size.

[19] Sentence length will be analysed in subsequent investigations, being another discourse organisation factor.

Table 4. Type/Token Ratio

Rue89	AgoraVox	Le Soir
.019447	.015928	**.018259**

The type/token ratio shows that lexical diversity is more elevated in Rue89 and Le Soir, with a higher degree of diversity in Rue89.

To characterise the three media in terms of their **discourse organisation**, we first compare them concerning the frequencies of discourse organistion cues, and second with respect to the content of P1 and P2, applying the methodology described in section 4.

Table 5 gives the LL statistics for the sentences beginning with at least one organisational cue.

Table 5. Organisational Cues' Distribution – LL Statistics

Sentences with	R vs. A	A vs. S	R vs. S
All. org. cues	**116.57 (R)**	**880.27 (S)**	**262.73 (S)**
SETTING	316.12 (R)	427.01 (S)	[P2: 39.45 (S)]
SEQ	21.97 (A)[P2]	33.00 (A)[P1]	
CONNECT	16.32 (A)[P1]	38.38 (A)[P1]	
Topical cues		**647.2 (S)**	**498.37 (S)**
APPOS	72.41 (R)[P2]	1,716.75 (S)	832.41 (S)
COREF_r	91.35 (A)[P2]	[P1: 34.55 (S)]	93.73 (S)
COREF_p		[P1: 151.29 (A)]	15.64 (R)[P1]

R = Rue89, A = AgoraVox, S = Le Soir
(R,A,S) indicates corpus with overuse
[P1,P2] indicates position with overuse

LL statistics indicate diverse differences between the three media,and support our hypothesis concerning the typological difference between them. If we look at the first rows, Le Soir appears to be the media with the highest number of cues signalling discourse organisation. But this overuse is only effective for topical continuity cues and especially via appositions and reiterations as shown in the last three rows. Nevertheless, this overuse is not effective for all topical cues. Indeed, each topical cue is significantly associated with different media: appositions with Le Soir (and in a weaker proportion with Rue89), reiterations with AgoraVox and Le Soir, and pronouns and possessive NPs with Rue89. AgoraVox is the media with the lowest amount of organisational cues. Nevertheless, this weaker proportion of organisational cues must be qualified by looking at the detailed LL indicating that there are significantly more sequencers and connectives in AgoraVox. Concerning setting cues, Rue89 and Le Soir seem to be alike, being significantly more present in the first than in AgoraVox. If we now look at the columns, Le Soir emerges as the most specific data set in contrast to Rue89 and AgoraVox that are closer in terms of discourse organisation. Nevertheless, Rue89 and AgoraVox are not similar. They weakly differ for all different cues: (1) while AgoraVox prefers reiteration, Rue89 shows a higher amount of strong topical continuity devices (appositions and pronouns), (2) while AgoraVox shows more sequencers and connectives, Rue89 shows more setting adverbials, comparable to Le Soir.

Taking into account variations according to textual position (indicated by brackets), we find significantly more setting adverbials, sequencers, reiterations, and appositions in P1 and significantly more connectives, pronouns, and possessive NPs in P2 (Ho-Dac's ([27]) results are in line with our insights). It is only if we focus on variations between media in each textual position that new insights appear. Setting adverbials in Le Soir are overused only in P2 when comparing Le Soir to Rue89, i.e. setting adverbials are overused in Le Soir when they are not associated with an effective structuring power ([26]). When connectives are overused in a media, it is generally in P1. In AgoraVox vs. Rue89 and Le Soir, but also in Le Soir vs. Rue89, the difference between the three data sets concerning the use of these argumentative elements is conspicuous. Pronouns and possessive NPs are overused in Rue89 when occurring in P1, underlying global topical continuity. In contrast, it is intraparagraphic reiterations that are overused in AgoraVox.

All these observations allow for assuming that the three data sets under investigation are different. Although further analyses are needed in order to better understand the differences, we may state here that the media show more characteristics of the argumentative text type (as compared to descriptive or expository texts), considering the use of connectives in P1 (associated with argumentative text types ([27])) as compared to the use of setting adverbials in P2 (associated with descriptive text types([27])).

6.2 PSEs as lexico-syntactic elements

The present subsection describes results concerning the research axis on subjectivity by the use of potential subjective elements, outlined in section 4.3.

Table 6. PSE Distribution – Number of Sentences

Sentences with		Rue89	AgoraVox	Le Soir
STANCE	Nb	3,229	6,059	3,201
	%	*1.65*	*2.03*	*1.28*
LOCpro	Nb	29,468	42,952	24,413
	%	*15.08*	*14.38*	*9.77*
LOCposs	Nb	4,715	7,467	4,531
	%	*2.41*	*2.50*	*1.81*
IT-ex	Nb	1,219	2,814	1,103
	%	*0.62*	*0.94*	*0.44*
CLEFT	Nb	4,567	6,256	4,762
	%	*2.34*	*2.09*	*1.91*
total		195,395	298,636	249,830

Table 6 displays the overall occurrences of PSEs included in the present investigation for each data set. As can be seen, LOCpro constitutes the most prolific cue, with about 15% of sentences containing a personal pronoun refering to the first person in the alternative media and 9% in the traditional newspaper. All other cues occur much less frequently and divergences between traditional and alternative media are not that striking. *It*-extrapositions are the least frequent means for expressing subjectivity in all data sets (R = 0.62%, A = 0.94%, S = 0.44%), while first person posssessive pronouns, stance adverbials and cleft sentences occupy an intermediate position with alike frequencies in the differnt sub corpora. It is striking that the number of sentences any of the given PSE is never hiher for Le Soir than for Rue89 or AgoraVox when considering percentages.

Table 7 represents the LL realised for the selected PSEs occurring in our corpus, comparing the data sets two by two. The first striking result is that the Le Soir data set never corresponds to the one with overuse for any of the subjectivity cues under investigation. This is in line with the comparison of percentages in table 6. Second, the divergence between alternative media on the one hand and the traditional newspaper on the other is eye-catching, especially when comparing the frequency of sentences containing a first person personal

Table 7. PSE Distribution – LL Statistics

total	R vs. A	A vs. S	R vs. S
PSE		3320.99 (A)	2829.75 (R)
Sentences with	R vs. A	A vs. S	R vs. S
STANCE	90.4 (A)	460.28 (A)	103.85 (R)
LOCpro	39.19 (R)	2,395.78 (A)	2,528.77 (R)
LOCposs		297.28 (A)	188.14 (R)
IT-ex	151.86 (A)	499 (A)	69.28 (R)
CLEFT	31.47 (R)	24.2 (A)	96.66 (R)

R = Rue89, A = AgoraVox, S = Le Soir
(R,A,S) indicates the corpus with overuse

pronoun (A vs. S LL = 2,395.78 and R vs. S LL = 2,528.77). Third, the differences comparing Rue89 and AgoraVox are much less conspicuous. Rue89 displays significantly more personal pronouns (LL LOCpro = 39.19) and cleft sentences (LL = 31.47), while AgoraVox overuses stance adverbials (LL = 90.4) and *it*-extrapositions (LL = 151.86). The frequency of possessive pronouns does not differ significantly between the two alternative media, in which they are respectively overused as compared to the traditional newspaper (A vs. S LL = 297.28 and R vs. S LL = 188.14). Rue89 seems to overuse cleft sentences (R vs.A LL = 31.47 and R vs. S LL = 96.66) that generally serve to point out an element by detachment, and which may as well be an indication for a more informal language style in Rue89. In contrast, the high frequency of *it*-extrapositions in the AgoraVox data set (R vs.A LL = 151.86 and A vs. S LL = 499) reflects an overuse of these constructions commonly associated with an impersonal expression of private state. These findings may be associated with the smaller amount of first person personal pronouns in AgoraVox as compared to Rue89. The assumption of a more informal language use in Rue89 and a more impersonal expression of subjectivity linked to it ask for further investigation.[20]

6.3 PSE as Hapax

Table 8 concerning subjectivity cues analyses occurrence patterns of hapax legomena.

Table 8. Distribution and LL Statistics Concerning Hapax Legomena

Distribution		Rue89	AgoraVox	Le Soir
Token		3,131,675	5,092,708	3,836,117
Hapax	Nb	26,849	38,006	29,777
	%	*0.86*	*0.75*	*0.78*
LL statistics		R vs. A	A vs. S	R vs. S
Hapax		300.17 (R)	25.80 (S)	139.16 (R)

R = Rue89, A = AgoraVox, S = Le Soir
(R,A,S) indicates the corpus with overuse

As can be seen, their frequency is significantly higher in Rue89 than in the two other media (R vs. S LL = 139.16 and R vs. A LL = 300.17). AgoraVox shows the lowest type/token ratio (table 4: .015928) and also the lowest amount of one word occurrences as compared to the two other media (R vs. A LL = 300.17 and A vs. S LL = 25.80). But as articles published in AgoraVox have a longer mean length (table 3: A = 751.19 words/text and S = 467.27 words/text), this first tendency has to be put into perspective and controlled by further research. Because even if ”people are creative when they are being opinionated” ([52]: 286), the corpus of journalese texts may show a high amount of hapax due to technical terms and specific language linked to a given subject. This might be an explanation for the low amounts of hapax and type/token ratio in AgoraVox – linked to the participation of non professional journalists publishing in this media, the highest amount of type/token ratio in Rue89, which is due to the professionalism of authors on the one hand side and the aim to report 'differently' from the traditional newspapers on the other, and the intermediate position of Le Soir, associated with a professional and thus probably more technical but less individual language use.

7 Conclusion

The present paper outlines the occurrences patterns of potential subjective elements in three different types of written mass media. In order to outline the expression of subjectivity, we carried out quantitative analyses by which we draw first tendencies to respond to our research questions and tested our hypotheses. The results show that the use of the different PSEs varies in the three data sets, and percentages (table 6) and raw frequencies (table 7) show that their use is less frequent in Le Soir than in the two alternative media, which is consistent with our first hypothesis. While articles in the alternative media seem to be alike, they clearly differ from the traditional newspaper. First tendencies support our second hypothesis as well: The strikingly higher use of first person personal pronouns in AgoraVox and Rue89 reflects an overt presence of the author in the these two media, as compared to Le Soir. The amount of the other PSEs under consideration is also slightly lower in the traditional newspaper. By contrast, our third hypothesis was not confirmed. Our data betoke that the two alternative media seem to prefer different PSEs, but we cannot declare an intermediate position for Rue89. Concerning the presence of the author in the text, it even seems to be more overt in Rue89, given the higher amount of personal pronouns and hapax. The high frequency of *it*-extrapositions in AgoraVox may indicate a subjectivity that is expressed via constructed objectivity as compared to Rue89, where overuses of cleft sentences and first person personal pronouns may be an indication for a more informal or direct expression of subjectivity. These cues will have to be investigated in subsequent research steps, including further quantitative analyses on supplementary PSE such as adjectives, *verba dicendi et sentiendi*, or modal expressions, as well as more detailed qualitative analyses with regard to the presented PSEs and the creation of different sub corpora. Furthermore, our results support our hypothesis concerning a typological difference between the three media. The results effectively indicate differences between the three data sets, being less well-defined when comparing the two alternative media, but being conspicuous when opposing the former to the traditional newspaper.

ACKNOWLEDGEMENTS

This research was supported by grant number ARC 08-12-001 from the Belgian French Speaking Community. We thank Liesbeth Degand and Anne Catherine Simon for proofreading and useful re-

[20] As the occurrences of the different PSEs do not vary conspicuously, neither concerning their amount within the three data sets, nor when comparing the sub corpora concerning a given cue, we intend to carry out qualitative analyses for all of them. We expect from this detailed investigation insights concerning the mode of subjectivity expression (formal vs. informal) and the judgement's value (positive, negative, neutral) in order outline distinguishing means for the three media.

marks as well as the three anonymous reviewers for their critical but notwithstanding constructive comments.

REFERENCES

[1] J.-M. Adam, ‘Genres de la presse écrite et analyse de discours’, *Semen*, **13**, 9–15, (2001).

[2] Mira Ariel, *Accessing Noun Phrase Antecedents*, London: Routledge, 1990.

[3] Harald Baayen, *Word Frequency Distributions*, Dordrecht: Kluwer Academic, 2001.

[4] Harald Baayen and Richard Sproat, ‘Estimating lexical priors for low-frequency morphologically ambiguous forms’, *Computational Linguistics*, **22**(2), 155–166, (1996).

[5] Monika Bednarek, *Evaluation in Media Discourse: Analysis of a Newspaper Corpus*, Continuum, 2006.

[6] Francoise Benhamou, Julie Lambert, and Marc-Olivier Padis, ‘Le journalisme en ligne: Transposition ou réinvention? entretien avec laurent mauriac et pascal riché’, *Esprit*, **3-4**, (2009).

[7] Emile Benveniste, *Problems in General Linguistics*, Coral Gables: University of Miami Press, 1971.

[8] D. Biber, U. Connor, and T.A. Upton, *Discourse on the move, using corpus analysis to describe discourse structure*, volume 28 of *Studies in corpus Linguistics*, John Benjamins Publishing Company: Amsterdam/Philadelphia, 2007.

[9] Douglas Biber, *Variation Across Speech and Writing*, Cambridge: Cambridge University Press, 1988.

[10] Douglas Biber and Edward Finegan, ‘Styles of stance in english: Lexical and grammatical marking of evidentiality and affect’, *Text*, **9**, 93–124, (1989).

[11] Douglas Biber, Stig Johansson, Geoffry Leech, Susan Conrad, and Edward Finegan, *Longman Grammar of Spoken and Written English*, London: Longman, 1999.

[12] Didier Bourigault, *Un Analyseur Syntaxique Opérationnel : SYNTEX*, Ph.D. dissertation, Mémoire d’HDR en Sciences du Langage, CLLE-ERSS, Toulouse, France, 2007.

[13] Wallace Chafe, *Subject and Topic*, chapter Givenness, Contrastiveness, Definiteness, Subjects, Topics, and Point of View, 25–55, New York/San Francisco/London: Academic Press, 1976.

[14] Patrick Charaudeau, ‘Discours journalistique et positionnements énonciatifs. frontières et dérives’, *Semen*, **22**, (2006).

[15] Maggie Charles, ‘”this mystery...”: A corpus-based study of the use of nouns to construct stance in theses from two contrasting disciplines’, *Journal of English for Academic Purposes*, **2**, 313–326, (2003).

[16] Maggie Charles, ‘Construction of stance in reporting clauses: A cross-disciplinary study of theses’, *Applied Linguistics*, **27**(3), 492–518, (2006).

[17] Michel Charolles, ‘L’encadrement du discours; univers champs domaines et espaces’, *Cahier de Recherche Linguistique, LanDisCo université Nancy2*, (6), (1997).

[18] Bernard Combettes, ‘Les constructions détachées comme cadres de discours’, *Langue Francaise*, **148**, 31–44, (2005).

[19] Susan Conrad and Douglas Biber, *Evaluation in Text. Authorial Stance and the Construction of Discourse*, chapter Adverbial Marking of Stance in Speech and Writing, 56–73, Oxford: Oxford University Press, 2000.

[20] Amandine Degand, ‘Le multimédia face à l’immédiat’, *Communication*, (submitted).

[21] Simon C. Dik, *Theory of Funtional Grammar Complex and Derived Constructions*, Berlin/New York: Mouton de Gruyter, 1997.

[22] Peter Fries, *On Subject and Theme: A Discourse Functional Perspective*, chapter Themes Method of Development and texts, 317–359, John Benjamins: Amsterdam/Philadelphia, 1995.

[23] Jeanette K. Gundel, Nancy Hedberg, and Ron Zacharski, ‘Cognitive status and the form of referring expressions in discourse’, *Language*, **69**, 274–307, (1993).

[24] Nigel Harwood, ‘’nowhere has anyone attempted...in this article i aim to do just that’: A corpus-based study of self-promotional i and we in academic writing across four disciplines’, *Journal of Pragmatics*, **37**, 1207–1231, (2005a).

[25] Nigel Harwood, ‘’we do not seem to have a theory...the theory i present here attempts to fill this gap’: Inclusive and exclusive pronouns in academic writing’, *Applied Linguistics*, **26**(3), 343–375, (2005b).

[26] L.-M. Ho-Dac and M.-P. Pry-Woodley, ‘Mthodologie exploratoire outille pour l’tude de l’organisation du discours’, in *Actes du Congrs Mondial de Linguistique Franaise (CMLF-08)*, Paris, (2008).

[27] Lydia-Mai Ho-Dac, *A Mosaic of Corpus Linguistics. Selected Approaches.*, chapter An exploratory data-driven analysis for describing discourse organization, 79–100, Frankfurt/Berlin: Peter Lang, 2010.

[28] Susan Hunston and Geoff Thompson, *Evaluation in Text. Authorial Stance and the Construction of Discourse*, Oxford: Oxford University Press, 2000.

[29] Ken Hyland, ‘Hedging in academic writing and eap textbooks’, *English for Specific Purposes*, **13**, 239–256, (1994).

[30] Ken Hyland, ‘Writing without conviction? hedging in science research articles’, *Applied Linguistics*, **17**, 433–454, (1996).

[31] Ken Hyland, ‘Persuasion and context: The pragmatics of academic metadiscourse’, *Journal of Pragmatics*, **30**, 437–455, (1998).

[32] Ken Hyland, ‘Stance and engagement: A modal of interaction in academic discourse’, *Discourse Studies*, **7**(2), 173–192, (2005).

[33] Ken Hyland and Polly Tse, ‘Metadiscourse in academic writing: A reappraisal’, *Applied Linguistics*, **25**(2), 156–177, (2004b).

[34] Chih-Hua Kuo, ‘The use of personal pronouns: Role relationships in scientific journal articles’, *English for Specific Purposes*, **18**(2), 121–138, (1999).

[35] George Lakoff, ‘Hedges: A study in meaning criteria and the logic of fuzzy concepts’, in *Papers from the Eighth Regional Meeting*, eds., Paul Peranteau, Judith Levi, and Gloria Phares, pp. 183–228. Chicago Linguistics Society (CLS 8), (1972).

[36] Knud Lambrecht, *Language Typology and Language Universals*, chapter Dislocation, 1050–1079, Berlin/New York: Mouton de Gruyter, 2001.

[37] Gilles Lugrin, ‘Le mélange des genres das l’hyperstructure’, *Semen*, **13**, (2001).

[38] James R. Martin, ‘Reading positions/positioning readers: Judgement in english’, *Prospect: a Journal of Australian TESOL*, **10**, 27–37, (1995).

[39] M.-P. Péry-Woodley, N. Asher, P. Enjalbert, F. Benamara, M. Bras, C. Fabre, S. Ferrari, L.-M. Ho-Dac, A. Le Draoulec, Y. Mathet, P. Muller, L. Prévot, J. Rebeyrolle, L. Tanguy, M. Vergez-Couret, L. Vieu, and A. Widlöcher, ‘Annodis : une approche outillée de l’annotation de structures discursives’, in *TALN 2009*, Senlis, (June 2009). ATALA, LIPN.

[40] K Precht, ‘Stance moods in spoken english: Evidentiality and affect in british and american conversation’, *Text*, **23**, 239–257, (2003).

[41] Ellen Prince, *Radical Pragmatics*, chapter Toward a Taxonomy of Given-New Information, 223–255, New York: New York Academic Press, 1981.

[42] Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik, *A Comprehensive Grammar of the English Language*, London: Longman, 1985.

[43] Paul Rayson and Roger Garside, *Comparing corpora using frequency profiling*, 1–6, October 2000.

[44] Marina Santini, ‘Web pages, text types, and linguistic features: Some issues’, *ICAME Journal*, **4**, 67–86, (2006).

[45] Helmut Schmid, *TreeTagger*, IMS, Universität Stuttgart, Germany.

[46] Catherine Schnedecker, *Noms Propres et Chaînes de référence*, Université de Metz : Metz, 1997.

[47] Michael Stubbs, ‘A matter of prolonged fieldwork: Notes towards a modal grammar of english’, *Applied Linguistics*, **7**, 1–25, (1986).

[48] Geoff Thompson and Susan Hunston, *Evaluation in Text. Authorial Stance and Construction of Discourse*, chapter Evaluation: An Introduction, 1–27, Oxford: Oxford University Press, 2000.

[49] Irena Vassileva, ‘Who am i/who are we in academic writing?’, *International Journal of Applied Linguistics*, **8**(2), 163–190, (1998).

[50] Marc Weeber, Rein Vos, and R. Harald Baayen, ‘Extracting the lowest-frequency words: Pitfalls and possibilities’, *Computational Linguistics*, **26**(3), 301–318, (2000).

[51] Peter R. R. White, *Appraisal Outline*, www.grammatics.com/appraisal, 2001.

[52] Janyce Wiebe et al., ‘Learning subjective language’, *Computational Linguistics*, **30**(3), 277–308, (2004).

Multi-view Learning for Text Subjectivity Classification

Dinko Lambov[1] and **Gaël Dias**[2] and **João V. Graça** [3]

Abstract. In this paper we consider the problem of building models that have high sentiment classification accuracy across domains. For that purpose, we present and evaluate a method based on co-training using both high-level and low-level features. In particular, we show that multi-view learning combining high-level and low-level features with adapted classifiers can lead to improved results over text subjectivity classification. Our experimental results present accuracy levels across domains of 86.4% combining LDA learning models over high-level features and SVM over bigrams.

1 Introduction

Over the past few years, there have been an increasing number of publications focused on the detection and classification of sentiment and subjectivity in texts. However, as stated in ([1], [2],[4], [6]), most research have focused on the construction of models within particular domains and have shown difficulties in crossing domains. In this paper, we propose to use multi-view learning to maximize classification accuracy across topics. For that purpose, we combine high-level features (e.g. level of affective words, level of abstraction of nouns) and low-level features (e.g. unigrams, bigrams) as different views to learn models of subjectivity which may apply to different domains such as movie reviews or newspaper articles. As stated in [15], SVM classifiers have usually been adopted for sentiment classification based on unigrams and bigrams. However, improvements over high-level features have been reached using LDA (Linear Discriminant Analysis) classifiers. So, our approach combines both SVM and LDA classifiers in the co-training algorithm [3] to obtain maximum performance over two views (high-level and low-level features). Experimental results show that the proposed approach outperforms over 9.2% the methodology proposed by [7] i.e. the SAR (Stochastic Agreement Regularization) algorithm and reaches 86.4% accuracy on average over four different data sets embodying different domains.

2 Related Work

The subjectivity and polarity[4] of language has been investigated at some length. Many features have been used to characterize opinionated texts at different levels: words [8], sentences [10] and texts ([8], [20], [25], [5]). In this section, we will only enumerate research works which focus on cross-domain classification.
One possible approach is to train a classifier on a domain-mixed set of data instead of training it on one specific domain as it is proposed in ([1], [6], [4]). Another possibility is to propose high-level features which do not depend so much on topics such as part-of-speech statistics or other semantic resources as in ([6], [15]). In this case, higher level representations do not reflect the topic of the document, but rather the type of text used. Just by looking at high-level features statistics, improved results can be obtained comparatively to unigram or bigram models (low-level models) when trying to cross domains.
Another approach is to find anchor terms which cross domains and evaluate the correlation between those words and words which are specific to the domain [2]. In this case, pivot features are discovered based on domain mutual information to relate training and target domains. The overall approach extends to sentiment classification the SCL (Structural Correspondence Learning) algorithm. Then, they identify a measure of domain similarity that correlates well with the potential for adaptation of a classifier from one domain to another. Best results across domains reach 82.1% accuracy. Finally, over the past few years, semi-supervised and multi-view learning proposals have emerged. [7] propose a co-regularization framework for learning across multiple related tasks with different output spaces. They present a new algorithm for probabilistic multi-view learning which uses the idea of stochastic agreement between views as regularization. Their algorithm called SAR (Stochastic Agreement Regularization) works on structured and unstructured problems and generalizes to partial agreement scenarios. For the full agreement case, their algorithm minimizes the Bhattacharyya distance between the models of each of the two views. [24] proposes a co-training approach to improve the classification accuracy of polarity identification of Chinese product reviews. First, machine translation services are used to translate English training reviews into Chinese reviews and also translate Chinese test reviews and additional unlabeled reviews into English reviews. Then, the classification problem can be viewed as two independent views: Chinese view with only Chinese features and English view with only English features. They then use the co-training approach to make full use of the two redundant views of features. An SVM classifier is adopted as a basic classifier in the proposed approach. Experimental results show that the proposed approach can outperform the baseline inductive classifiers and more advanced transductive classifiers.
Unlike all proposed methods so far, our approach aims at taking advantage of different view levels. We propose to combine high-level features (e.g. level of affective words, level of abstraction of nouns) and low-level features (e.g. unigrams, bigrams) to learn models of subjectivity which may apply to different domains. For that purpose, we propose a new scheme based on the classical co-training algorithm over two views [3] and join two different classifiers LDA and SVM to maximize the optimality of the approach.

[1] University of Beira Interior, Portugal, email: d.lambov@gmail.com
[2] University of Beira Interior, Portugal, email: ddg@di.ubi.pt
[3] L2F INESC-ID INESC-ID,Portugal, email: joao.graca@l2f.inesc-id.pt

[4] Most papers deal with polarity as the essence of subjectivity. However, subjectivity can be expressed in different ways. In this paper, we will focus on subjectivity classification and not just polarity.

3 Characterizing Subjectivity

Many works to date have been concerned with the less ambitious goal of identifying the polarity of sentiment in texts. However, subjectivity can be expressed in different ways as summarized in [4] who identify the following dimensions: evaluation (positive or negative), potency (powerful or unpowerful), proximity (near or far), specificity (clear or vague), certainty (confident or doubtful) and identifiers (more or less), direct expressions, elements of actions and remarks. Based on these assumptions, our methodology aims at classifying texts at the subjectivity level (i.e. subjective vs. objective and not, (positive, negative) vs. objective) taking into account both high-level features which cross domains easily [15] as well as low-level features (unigrams or bigrams) which evidence high precision results within domains [20].

3.1 High-Level Features

Intensity of Affective Words: sentiment expressions mainly depend on some words which can express subjective sentiment orientation. [22] use words from the WordNet Affect lexicon [23] to annotate the emotions. For example horror and hysteria express negative fear, enthusiastic expresses positive emotion, glad expresses joy, and so on and so forth. So, we propose to evaluate the level of affective words in texts as shown in Equation 1.

$$K_1 = \frac{total\ affective\ words\ in\ text}{total\ words\ in\ text} \tag{1}$$

Dynamic and Semantically Oriented Adjectives: [9] consider two features for the identification of opinionated sentences: (1) semantic orientation, which represents an evaluative characterization of word deviation from its semantic group and (2) dynamic adjectives which characterize word ability to express a property in varying degrees. For the present study, we use the set of all adjectives automatically identified in a reference corpus i.e. the set of dynamic adjectives manually identified by [9] and the set of semantic orientation labels assigned as in [8]. So, we propose to evaluate the level of these adjectives in texts as shown in Equation 2.

$$K_2 = \frac{total\ specific\ adjectives\ in\ text}{total\ adjectives\ in\ text} \tag{2}$$

Classes of Verbs: [5] present a method using verb class information. The verb classes they use express objectivity and polarity. To obtain relevant verb classes, they use InfoXtract [21], an automatic text analyzer which groups verbs according to classes that often correspond to their polarity. As InfoXtract is not freely available, we reproduce their methodology by using the classification of verbs available in Levins English Verb Classes and Alternations [17]. So, we propose to evaluate the level of each class of verbs (i.e. conjecture, marvel, see and positive) in texts as in Equation 3.

$$K_3 = \frac{total\ specific\ verbs\ in\ text}{total\ verbs\ in\ text} \tag{3}$$

Level of Abstraction of Nouns: There is linguistic evidence that level of generality is a characteristic of opinionated texts, i.e. subjectivity is usually expressed in more abstract terms than objectivity [15]. Indeed, descriptive texts tend to be more precise and more objective and as a consequence more specific. In other words, a word is abstract when it has few distinctive features and few attributes that can be pictured in the mind. One way of measuring the abstractness of a word is by the hypernym relation in WordNet [19]. In particular, a hypernym metric can be the number of levels in a conceptual taxonomic hierarchy above a word (i.e. superordinate to). For example, chair (as a seat) has 7 hypernym levels: $chair \Rightarrow furniture \Rightarrow furnishings \Rightarrow instrumentality \Rightarrow artifact \Rightarrow object \Rightarrow entity$. So, a word having more hypernym levels is more concrete than one with fewer levels. So, we propose to evaluate the hypernym levels of all the nouns in texts as shown in Equation 4.

$$K_4 = \frac{total\ hypernym\ levels\ for\ nouns\ in\ text}{total\ nouns\ in\ text} \tag{4}$$

Calculating the level of abstraction of nouns should be preceded by word sense disambiguation. Indeed, it is important that the correct sense is taken as a seed for the calculation of the hypernym level in WordNet. However, in practice, taking the most common sense of each word gives similar results as taking all the senses on average [15].

3.2 Low-Level Features

The most common set of features used for text classification is information regarding the occurrences of words or word ngrams in texts. Most of text classification systems treat documents as simple bags-of-words and use the word counts as features. Here, we consider texts as bags-of-words of lemmatized unigrams or lemmatized bigrams for which we compute their TF.IDF weights as in Equation 5 where w_{ij} is the weight of term j in document i, tf_{ij} is the normalized frequency of term j in document i, N is the total number of documents in the collection, and n is number of documents where the term j occurs at least once.

$$w_{ij} = tf_{ij} * \log_2 \frac{N}{n} \tag{5}$$

4 The Multi-View Approach

4.1 Co-Training

The co-training algorithm [3] is a typical bootstrapping method, which starts with a set of labeled data, and increases the amount of annotated data using some amounts of unlabeled data in an incremental way. One important aspect of co-training is that two conditional independent views are required for co-training to work, but the independence assumption can be relaxed. The co-training algorithm is illustrated in Figure 1. In the algorithm, the class distribution in the labeled data is maintained by balancing the parameter values of p and n at each iteration (e.g. positive (resp. negative) examples will be subjective (resp. objective) texts). The intuition of the co-training algorithm is that if one classifier can confidently predict the class of an example, which is very similar to some labeled ones, it can provide one more training example for the other classifier. But, of course, if this example happens to be easy to be classified by the first classifier, it does not mean that this example will be easy to be classified by the second classifier, so the second classifier will get useful information to improve itself and vice versa [13].

In the context of cross-domain sentiment classification, each labeled or unlabeled text has two views of features: high-level features (V1) and low-level features (V2). A basic classification algorithm is also required to construct both models H1 and H2. Typical sentiment classifiers include Support Vector Machines and Maximum Entropy. In this study, we adopt the widely used SVM classifier [11] as well as the LDA classifier which has proved to provide better results than

```
Given a set L of labeled examples
Given a set U of unlabeled examples
Loop for k iterations
    • Train a classifier H1 on view V1 of L
    • Train a classifier H2 on view V2 of L
    • Allow H1 and H2 to label U
    • Add the p positive and n negative most confidently predicted
      textsto L
    • Retrain H1 and H2 on L
```

Figure 1. The co-training algorithm.

SVM for high-level features [15]. So, we will present results both with SVM or LDA classifiers for the view V1 while only SVM will be applied to the view V2 due to its huge number of features. Moreover, it is important to notice that the unlabeled set of examples U will be from a different domain than the labeled set of examples U. Indeed, the overall idea is that each classifier gets useful information from the other view to improve itself to cross domains.

4.2 SAR Algorithm

[7] propose the SAR (Stochastic Agreement Regularization) algorithm. It models a probabilistic agreement framework based on minimizing the Bhattacharyya distance [12] between models trained using two different views. They regularize the models from each view by constraining the amount by which they permit them to disagree on unlabeled instances from a theoretical model. Their co-regularized objective which has to be minimized is defined in Equation 6 where Li for i = 1,2 are the standard regularized loglikelihood losses of the models p1 and p2, Eu[B(p1,p2)] is the expected Bhattacharyya distance between the predictions of the two models on the unlabeled data, and c is a constant defining the relative weight of the unlabeled data.

$$MinL_1(\theta_1) + L_2(\theta_2) + cE_u[B(p_1(\theta_1), p_2(\theta_1))] \qquad (6)$$

In the context of sentiment classification and multi-view learning, [7] is certainly the best reference up-to-date, reaching accuracy levels of 82.8% for polarity detection upon reviews from the kitchen and the dvd domains using random views of unigrams. In this work, we will test SAR on our dataset both on random views of unigrams and random views of bigrams and take its results as baselines[5].

5 Multi-Domain Corpora

To perform our experiments, we used three manually annotated standard corpora and built one corpus based on Web resources which could be automatically annotated as objective or subjective.

The Multi-Perspective Question Answering (Mpqa) Opinion Corpus[6] contains 10.657 sentences in 535 documents from the world press on a variety of topics. All documents in the collection are marked with expression-level opinion annotations. The documents are from 187 different news sources in a variety of countries and date from June 2001 to May 2002. The corpus corpus has been collected and manually annotated with respect to subjectivity as part of the summer 2002 NRRC Workshop on Multi-Perspective Question Answering. Based on the work done by [20] who propose to classify texts based only on their subjective/objective parts, we built a corpus of 100 objective texts and 100 subjective texts by randomly selecting sentences containing only subjective or objective phrases. This case represents the ideal case where all the sentences in texts are either subjective or objective.

The second corpus (Rotten/Imdb) is the subjectivity dataset v1.0[7] which contains 5000 subjective and 5000 objective sentences collected from movie reviews data [20]. To gather subjective sentences, [20] collected 5000 movie review snippets from http://www.rottentomatoes.com. To obtain (mostly) objective data, they took 5000 sentences from plot summaries available from the Internet Movie Database http://www.imdb.com. Similarly to what we did for the Mpqa corpus, we built a corpus of 100 objective texts and 100 subjective texts by randomly selecting only subjective or objective sentences.

The third corpus (Chesley) has been developed by [5] who manually annotated a dataset of objective and subjective documents[8] . It contains 496 subjective and 580 objective documents. Objective feeds are from sites providing content such as world and national news (e.g. CNN, NPR), local news (e.g. Atlanta Journal and Constitution, Seattle Post-Intelligencer), and various sites focused on topics such as health, science, business, and technology. Subjective feeds include content from newspaper columns (e.g. Charles Krauthammer, E.J. Dionne), letters to the editor (e.g. Washington Post, Boston Globe), reviews (e.g. dvdver-dict.com, rottentomatoes.com), and political blogs (e.g. Powerline, Huffington Post). For our purpose, we randomly selected 100 objective texts and 100 subjective texts.

The fourth corpus is based on the idea that Wikipedia conveys objective contents whereas Web Blogs provide subjective contents to its audience [16]. As a consequence, [15] built the automatically annotated Wiki/Blog[9] corpus. They downloaded part of the static Wikipedia dump archive[10] and automatically spidered Web Blogs from different domains. The final corpus contains 200 Mb of downloaded articles from Wikipedia and 100 Mb of downloaded texts from different Web Blogs. These texts are in English and cover many different topics. Due to their characteristics, Wikipedia texts were automatically labeled as objective and Web Blogs automatically labeled as subjective. From this data set, we finally randomly selected 100 objective texts and 100 subjective texts.

6 Experiments

In order to evaluate the difference between high-level features with low-level features, [15] performed a comparative study on the four data sets presented in the previous section. For the high-level features, they took into account 7 features: affective words, semantically oriented adjectives, dynamic adjectives, conjecture verbs, marvel verbs, see verbs and level of abstraction of nouns. For the unigram and bigram models, they used all the lemmas inside the corpora withdrawing their stop words. In Table 1, we summarize the results obtained for the single view classification task using high-level or low-level features.

[5] The SAR package has been implemented for unigrams and bigrams only [7]. Future work will aim at adapting the SAR to other views.

[6] http://www.cs.pitt.edu/mpqa/

[7] http://www.cs.cornell.edu/People/pabo/movie-review-data/

[8] http://www.tc.umn.edu/ ches0045/data/

[9] The corpus is available on the web (url omitted for anonymity)

[10] http://download.wikimedia.org/enwiki/

Table 1. Accuracy for high-level features (HL) and low-level features (LL) across domains in %.

		MPQA	Rotten	Chesley	Wiki
HL	SVM	52.6	69.5	73.9	71.0
	LDA	67.6	70.9	73.6	74.5
LL	Unigram	53.8	63.9	59.9	61.1
	Bigram	54.4	67.1	55.0	57.5

All experiments were performed on a leave-one-out 5 cross validation basis combined with both SVM and LDA classifiers for high-level features and only SVM for low-level features due to the high level of features which does not suit to LDA classifiers. In particular, they used Joachims SVMlight package[11] [11] for training and testing with SVM and the implementation of LDA in the R[12] software for statistical computing. As part-of-speech tagger, they used the MontyTagger module of the free, common sense-enriched Natural Language Understander for English MontyLingua[13] [18]. In order to test models across domains, they proposed to train different models based on one domain only at each time and test the classifiers over all domains together. So, each percentage can be expressed as the average results over all data sets. Best results overall are obtained for high-level features with the Wiki/Blog data set as training set and the LDA classifier with an average accuracy of 74.5%. This result will represent our baseline for single view classification as we aim at showing that multi-view learning can lead to improved results to cross domains. In all our experiments, we will use the same process as in [15] to evaluate accuracy so that values are comparable.

6.1 Results for the SAR algorithm

We first propose to show the results obtained with SAR [7] which represents the state-of-the-art in multi-view learning to cross domains in the field of sentiment analysis. To perform SAR experiments, we used two views generated from a random split of low-level features together with the maximum entropy classifiers with a unit variance Gaussian prior. Indeed, the actual implementation of SAR does not allow to testing it with different views but only with random subsets of views (e.g. unigrams are divided into two subsets: unigrams1 and unigrams2), nor with different classifiers. The results are illustrated in Table 2 exactly in the same way they have been processed in [15].

Table 2. SAR accuracy for low-level features across domains in %.

	MPQA	Rotten	Chesley	Wiki
Unigram	65.3	73.5	72.2	59.2
Bigram	71.6	75.2	77.2	65.1

The results show indeed interesting properties. Models built upon bigrams constantly outperform models based on unigrams. Higher accuracy compared to [15] is reached with less knowledge. Indeed, the baseline with single view classification is 74.5% while 77.2% can be obtained with the SAR algorithm upon a random split of bigrams. One great advantage of only using low-level features is the ability to reproduce such experiments on different languages without further resources than just texts. However, a good training data set will have to be produced as the best results are obtained from the manually annotated corpus Chesley.

6.2 Results for Co-Training

In this subsection, we propose to use the co-training algorithm to combine a first view which contains 7 high-level features (7F) and a second view which contains low-level features (unigrams or bigrams). As a consequence, we expect that the low-level classifier will gain from the decisions of the high-level classifier and will self-adapt to different domains based on the high results of high-level features for crossing domains. In Table 3, we show the results obtained using two SVM classifiers i.e. one for each view. In Table 4, we show the results obtained using an SVM classifier for the low-level view and an LDA classifier for the high-level classifier as we know that LDA outperforms SVM for high-level features.

Table 3. Co-training accuracy with two SVM classifiers across domains in %.

		MPQA	Rotten	Chesley	Wiki
7F	Unigram	61.0	72.3	78.8	62.75
7F	Bigram	66.4	78.1	75.3	85.6

Table 4. Co-training accuracy with one SVM and one LDA classifiers across domains in %.

		MPQA	Rotten	Chesley	Wiki
7F	Unigram	63.3	74.9	79.0	63.5
7F	Bigram	67.4	78.1	68.5	86.4

The benefit from the high-level features is clear based on the results of Tables 3 and 4. The best result is obtained by the combination of high-level features with the LDA classifier and bigram low-level features with the SVM classifier trained over the automatically annotated corpus Wiki/Blogs. In this case, the average accuracy across domains is 86.4% outperforming SAR best performance 77.2%. It is interesting to notice that in almost all cases, bigram low-level features provide better results than only unigrams. The only exception is the Chesley training set. But, it is especially evident for the Wiki/Blog training data set that bigrams drastically improve the performance of the co-training as the difference between unigrams or bigrams as second views is huge. Accuracy results were obtained from the second view classifier, i.e. the low-level classifier. Indeed, while the high-level classifier accuracy remains steady iteration after iteration, the low-level classifier steadily improves its accuracy based on the correct guesses of the high-level classifier[14] . We illustrate the behavior of each classifier in Figure 2.

In order to better understand this situation, we propose a visual analysis of the distribution of the data sets in the space of high-level and low-level features. The goal of this study is to give a visual interpretation of the data distribution to assess how well co-training may perform using high-level and low-level features. If objective and subjective texts can be represented in a distinct way in a reduced space of features, one may expect good classification results. To perform this study, we use a MDS (Multidimensional Scaling) process which is a traditional data analysis technique. MDS [14] allows to displaying

[11] http://svmlight.joachims.org/

[12] http://www.r-project.org/

[13] http://web.media.mit.edu/ hugo/montylingua/

[14] After each iteration, 2 positive (subjective) examples and 2 negative (objective) examples from each classifier are added to the set of labeled data L.

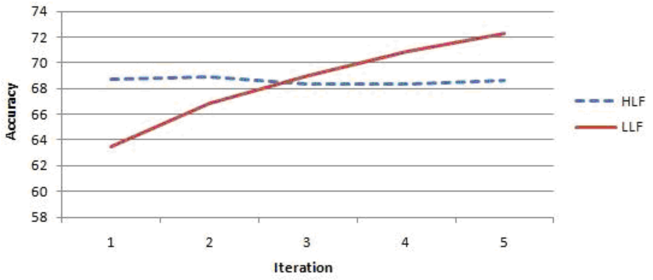


Figure 2. Low-level and high-level accuracies iteration after iteration for the Rotten/Imdb data set with LDA over 7F and SVM over bigrams.

the structure of distance-like data into an Euclidean space. In practice, the projection space we build with the MDS from such a distance is sufficient to have an idea about whether data are organized into classes or not. For our purpose, we performed the MDS process over pairs of corpora represented by low-level features and high-level features to try to visualize how texts evolve in the multidimensional space before and after co-training.

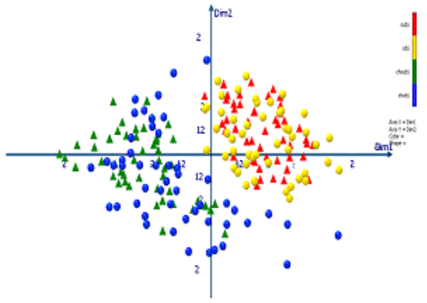


Figure 3. Low-level feature representation of subjective (red and green triangles) and objective (blue and yellow dots) texts before co-training.

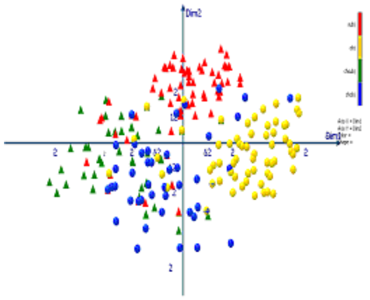


Figure 4. Low-level feature representation of subjective (red and green triangles) and objective (blue and yellow dots) texts after co-training.

In Figures 3 and 4, we graphically represent texts of Rotten/Imdb and Chesley in a reduced space of the low-level features space. Red and green triangles represent subjective texts from Rotten/Imdb and Chesley respectively. Yellow and Blue dots represent objective texts from Rotten/Imdb and Chesley respectively. This visualization clearly shows that after co-training subjective and objective texts from different domains tend to approximate. Comparatively, in Figures 5 and 6, we graphically represent the same texts in a reduced space of the high-level features space. In this experiment, we clearly see that texts do not tend to approximate and remain difficult to separate, as such comforting us in the choice of using low-level classifiers for our classification task using the co-training approach.

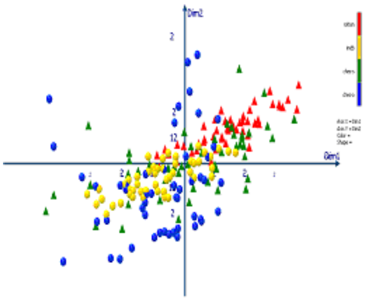


Figure 5. High-level feature representation of subjective (red and green triangles) and objective (blue and yellow dots) texts before co-training.

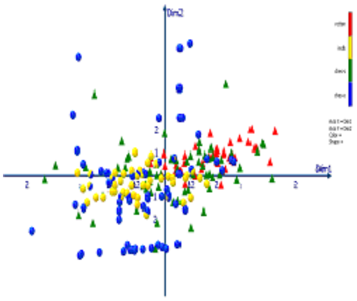


Figure 6. High-level feature representation of subjective (red and green triangles) and objective (blue and yellow dots) texts after co-training.

7 CONCLUSION

Sentiment classification is a domain specific problem i.e. classifiers trained in one domain do not perform well in others. At the same time, sentiment classifiers need to be customizable to new domains in order to be useful in practice. In this paper, we proposed to use the co-training approach to address the problem of cross-domain sentiment classification. For that purpose, we presented different experiments based on multi-view learning algorithms using high-level and low-level features to learn subjective language across domains. The experimental results showed the effectiveness of the proposed approach. Best results showed accuracy of 86.4% across domains compared to 77.2% for the SAR algorithm proposed by [7] and 74.5% for single view classification with LDA proposed by [15]. In future work, we plan to improve the subjectivity classification accuracy by using more than two views as well as customizing the SAR algorithm to receive different types and numbers of views.

REFERENCES

[1] A. Aue and M. Gamon, ‘Customizing sentiment classifiers to new domains: a case study’, in *Proceedings of International Conference on Recent Advances in Natural Language Processing (RANLP)*, (2005).

[2] J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. Wortman, 'Biographies, bollywood, boom-boxes, and blenders: Domain adaptation for sentiment classification', in *Proceedings of Association for Computational Linguistics*, (2007).

[3] A. Blum and T. Mitchell, 'Combining labeled and unlabeled data with co-training', in *Proceedings of Conference on Computational Learning Theory*, (1998).

[4] E. Boiy, P. Hens, K. Deschacht, and M-F. Moens, 'Automatic sentiment analysis of on-line text', in *Proceedings of the 11th International Conference on Electronic Publishing, Openness in Digital Publishing: Awareness, Discovery & Access*, (2007).

[5] P. Chesley, B. Vincent, L. Xu, and R. Srihari, 'Using verbs and adjectives to automatically classify blog sentiment', in *Proceedings of AAAI Spring Symposium*, (2006).

[6] A. Finn and N. Kushmerick, 'Learning to classify documents according to genre', *Journal of American Society for Information Science and Technology, Special issue on Computational Analysis of Style*, **57**, (2006).

[7] K. Ganchev, J. Graca, J. Blitzer, and B. Taskar, 'Multi-view learning over structured and non-identical outputs', in *In Uncertainty in Artifitial Intelligence*, (2008).

[8] V. Hatzivassiloglou and K.R. McKeown, 'Predicting the semantic orientation of adjectives', in *In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th Conference of the European Chapter of the Association for Computational Linguistics*, (1997).

[9] V. Hatzivassiloglou and J. Wiebe, 'Effects of adjective orientation and gradability on sentence subjectivity', in *In Proceedings of International Conference on Computational Linguistics*, (2000).

[10] Y. Hong and V. Hatzivassiloglou, 'Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences', in *In proceedings of Conference on Empirical Methods on Natural Language Processing*, (2003).

[11] T. Joachims, *Learning to Classify Text Using Support Vector Machines*, Springer, 2002.

[12] T. Kailath, 'The divergence and bhattacharyya distance measures in signal selection', *IEEE Transactions on Communications*, **15**, 5260, (1967).

[13] S. Kiritchenko and S. Matwin, 'Email classification with co-training', in *In Proceedings of the 2001 Conference of the Centre for Advanced Studies on Collaborative Research*, (2001).

[14] J. B. Kruskal and M. Wish, *Multidimensional Scaling*, Sage Publications, Beverly Hills, 1977.

[15] D. Lambov, G. Dias, and V. Noncheva, 'High-level features for learning subjective language across domains', in *Proceedings of International AAAI Conference on Weblogs and Social Media ICWSM*, (2009).

[16] D. Lambov, G. Dias, and V. Noncheva, 'Sentiment classification across domains', in *14th Portuguese Conference on Artificial Intelligence EPIA*, (2009).

[17] B. Levin, *English Verb Classes and Alternations*, University of Chicago Press, 1993.

[18] Hugo Liu. Montylingua: An end-to-end natural language processor with common sense. available at: http://web.media.mit.edu/˜hugo/montylingua, 2004.

[19] G.A. Miller, *Wordnet: A Lexical Database*, In Communications of the ACM 38, 1995.

[20] P. Pang and L. Lee, 'A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts', in *In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, (2004).

[21] R. Srihari, W. Li, C. Niu, and T. Cornell, 'Infoxtract: A customizable intermediate level information extraction engine', in *In Proceedings of the HLT-NAACL 2003 Workshop on Software Engineering and Architecture of Language Technology Systems*, (2003).

[22] C. Strapparava and R. Mihalcea, 'Learning to identify emotions in text', in *In Proceedings of the Symposium on Applied Computing*, (2008).

[23] C. Strapparava and A. Valitutti, 'Wordnet-affect: An affective extension of wordnet', in *In Proceedings of the Language Resources and Evaluation International Conference*, (2004).

[24] X. Wan, 'Co-training for cross-lingual sentiment classification', in *In Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics*, (2009).

[25] J. Wiebe, T. Wilson, R. Bruce, M. Bell, and M. Martin, 'Learning subjective language', *Computational Linguistics*, **30**, 277308, (2004).

Packed Feelings and Ordered Sentiments: Sentiment Parsing with Quasi-compositional Polarity Sequencing and Compression

Karo Moilanen[1] and **Stephen Pulman**[2] and **Yue Zhang**[3]

Abstract. Recent solutions proposed for sentence- and phrase-level sentiment analysis have reflected a variety of analytical and computational paradigms that include anything from naïve keyword spotting via machine learning to full-blown logical treatments, either in pure or hybrid forms. As all appear to succeed and fail in different aspects, it is far from evident which paradigm is the optimal one for the task. In this paper, we describe a quasi-compositional sentiment learning and parsing framework that is well-suited for exhaustive, uniform, and principled sentiment classification across words, phrases, and sentences. Using a hybrid approach, we model one fundamental logically defensible compositional sentiment process directly and use supervised learning to account for more complex forms of compositionality learnt from mere flat phrase- and sentence-level sentiment annotations. The proposed framework operates on quasi-compositional sentiment polarity sequences which succinctly capture the sentiment in syntactic constituents across different structural levels without any conventional *n*-gram features. The results obtained with the initial implementation are highly encouraging and highlight a few surprising observations pertaining to role of syntactic information and sense-level sentiment ambiguity.

1 INTRODUCTION

Language affords a wonderfully rich array of devices for expressing subjectivity, sentiments, affect, emotions, stances, opinions, arguments, points of view, perspectives, slurs, and the many other forms of non-factuality. From the viewpoint of a computational algorithm, non-factual content is bound to appear noticeably fuzzier than what is usually the case in traditional, more factual NLP tasks such as sentence breaking, part-of-speech tagging, or topic categorisation, to name a few. On the other hand, recent advances in computational Sentiment Analysis, Opinion Mining, and Affect/Emotion Analysis (and other related areas) have produced applications which, while still leaving much to be desired, are already highly useful in practice and can in some cases mimic human sentiment interpretation relatively well.

All proposals made in the above areas ultimately face the same fundamental challenge which is to determine what happens when individual expressions with rich (non-)sentiment properties interact with each other. A wide range of different solutions can be found encompassing mere frequency-based keyword spotting with no or naïve analytical additions, various machine learning approaches that have incorporated shallow-structural or -semantic features, and explicit direct fully- or shallow-compositional sentiment logics (§5).

If the goal is to be able to fully understand and account for the very behaviour of sentiment in language, then the task of explaining a given expression is to be approached using some form of principled logical reasoning that tries to *systematically* analyse all different parts of the expression in order to arrive at a logically defensible, coherent, and interpretable explanation in each case. Logical reasoning gives rise to a number of fundamental **compositional sentiment processes** many of which are simple enough to be modelled directly (e.g. [18], [13]). The most basic process involves **sentiment charge** which effectively involves inserting sentiment into an otherwise neutral expression. For example, when the neutral sentence “*[This report will make you* $\_\_\__i$ *for hours]*$^{(N)}$” is modulated by a positive sentiment carrier (e.g. “*laugh*$_i^{(+)}$”), the **non-neutral propagation** process causes the latter to propagate its non-neutral sentiment across the entire sentence (vice versa for “*weep*$_i^{(-)}$”). Another obvious process is **null composition** which simply involves combining expressions displaying the same polarity (e.g. “*[evil]*$^{(-)}$*[wars]*$^{(-)}$” = “*[evil wars]*$^{(-)}$”). Somewhat less frequent is the **direct reversal** process in which reversive expressions reverse other expressions’ polarities (e.g. “*[avoid]*$^{[\neg]}$*[trouble]*$^{(-)}$” = “*[avoid trouble]*$^{(+)}$”). More challenging are the numerous cases where clashing non-neutral polarities interact: in these cases, some form of **conflict resolution** is necessary whereby the ensuing conflicts are resolved using either syntactic or semantic means (e.g. “*[benefit]*$^{(+)}$*[fraud]*$^{(-)}$” = “*[benefit fraud]*$^{(-)}$”).

Although it is still unclear which computational paradigm is optimal for practical purposes, explicit sentiment logics that implement the above kinds of fundamental sentiment processes have generally been observed to be very precise. They however commonly require manual rules which specify how individual expressions are to interact in the analysis, and are not unlikely to suffer from limited recall levels. Less focused machine learning approaches typically offer greater coverage but risk becoming too domain-dependent. They have yet to explain how the many contextual factors that ultimately govern the interaction of individual expressions across different structural levels are to be captured in a uniform and exhaustive manner.

In this paper, we seek to bridge these two paradigms and propose a hybrid sentiment learning and parsing framework for (sub)sentential sentiment analysis that implements only one of the above logical sentiment processes directly while leaving the rest to be learnt proba-

[1] Oxford University Computing Laboratory, UK. email: Karo.Moilanen@oucl.ox.ac.uk

[2] Oxford University Computing Laboratory, UK. email: Stephen.Pulman@oucl.ox.ac.uk

[3] University of Cambridge Computer Laboratory, UK. email: Yue.Zhang@cl.cam.ac.uk

bilistically from annotated sentiment data. The justification for a hybrid strategy is that the most basic compositional processes are simple enough to be modelled directly while the more complex ones may necessitate a more data-driven approach (cf. [4]). The framework is conceptually simple yet surprisingly powerful, and lends itself naturally to uniform sentiment parsing across words, phrases, and sentences. It is abstract enough to reduce domain and structural dependency effects but specific enough to capture one of the most important behavioural properties of sentiment. The proposed framework can be implemented easily without any complex linguistic processing as it requires only flat phrase- and/or sentence-level sentiment annotations, a sentiment lexicon, and, optionally, a part-of-speech tagger and a syntactic parser.

2 POLARITY SEQUENCE MODEL

Consider the following sample sentence (Ex. 1) (non-neutral and reversive words underlined).

(1) *["Our* `lives`$^{(+)}$*will* `never`$^{[\neg]}$*be the same again, having* `lost`$^{(-)}$*our* `loved`$^{(+)}$*ones and everything [we had] having been* `destroyed`$^{(-)}$*," Moussa told IRIN.]*$^{(-)}$

In a baseline approach, a classifier could first consider the polarity frequencies in the sentence (cf. [6]). It would however fail in this case since a `POS`/`NEG` tie count of 2 would ensue. In a more popular approach, the classifier could instead consider the distribution of various *n*-grams in the sentence: the unigram *"lost"* and the bigram *"and everything"* might, for example, help the classifier to make a decision if it has seen them amongst its negative polarity training examples (cf. [17]). Although they can reach moderate accuracy levels, the major drawback in *n*-gram models is that they seldom generalise well beyond the training data used, ignore all positional and temporal aspects of the sentiment carriers, and also perform markedly worse at lower (sub)sentential levels where evidence is always scarcer.

More complex features can then be harnessed to account for limited contexts around individual sentiment carriers by using crude, fixed windows (e.g. ±5 words) or by considering the sentiment around and above the sentence. For example, the fact can be exploited that the reversive adverb *"never$^{[\neg]}$"* affects the positive noun *"lives$^{(+)}$"* (via the head verb *"be"*). However, structural features still struggle to cope with the fact that sentiment carriers and other expressions modifying them can occur in any syntactic position, with many nested long-distance dependencies involved (cf. [25]). When more global structural features (e.g. checking if the sentence is surrounded by negative sentences or if it is in a negative document) are used (cf. [12]), evidence may erroneously be amassed from structures whose sentiment properties have nothing at all to do with each other. If positional information (e.g. the verb *"destroyed$^{(-)}$"* is the last sentiment carrier in the sentence) is included (cf. [15]), the problem remains that the most salient carrier can occur anywhere in the sentence. A further complication arises from the fact that many sentiment carriers suffer from context-dependent polarity ambiguity which confounds the problem even further.

2.1 Quasi-Compositional Sentiment Sequencing

On the basis of these kinds of complications that have hampered learning-based approaches, we instead investigate the possibility of an alternative, much simpler, route around the problem by dropping all but one of the conventional assumptions: specifically, we focus *solely* on the linear order in which atomic sentiment polarities occur in a sentence. If we represent the above sentence (Ex. 1) based on the prior out-of-context sentiment polarities of the words in it, the following **raw polarity sequence** representation emerges (Ex. 2):

(2) `1:NTR` **`2:POS`** `3:NTR` **`4:REV`** `5:NTR 6:NTR 7:NTR 8:NTR 9:NTR` **`10:NEG`** `11:NTR` **`12:POS`** `13:NTR 14:NTR 15:NTR 16:NTR 17:NTR 18:NTR 19:NTR` **`20:NEG`** `21:NTR 22:NTR 23:NTR`

Note that in addition to the three sentiment polarities proper (`POS`, `NTR`, `NEG`), the sentiment reversal potential (`REV`) of a word is used here as a fourth 'polarity' (cf. [4]). Raw polarity sequences such as this can then be turned into learning features by treating each step (i.e. **slice**) in the polarity sequence as a separate feature. However, because sentences and phrases vary a great deal in terms of their length (i.e. the number of raw feature slices that they yield), raw polarity sequences risk generating too sparse feature vectors and do as such necessitate very large amounts of training data to cover the probability of each of the four polarities occurring in each of the slices. Hence, we seek to employ some form of feature reduction instead.

The fundamental compositional sentiment process of **null composition** described in §1 offers a simple, yet logically defensible, means to shrink the feature space. If it is the case that two expressions which display the same polarity (e.g. *"[evil]$^{(-)}$[wars]$^{(-)}$"*) cannot but result in a compositional expression with the very same polarity (e.g. *"[evil wars]$^{(-)}$"*), then the same holds for three, four, and, by extension, *n* expressions. Hence, all subsequences in a raw polarity sequence that display the same consecutive polarity can axiomatically be collapsed into a single feature slice. We accordingly observe that the present sentence reduces into the following **compressed quasi-compositional polarity sequence** (Ex. 3) (with old and new slice IDs):

`1:NTR` **`2:POS`** `3:NTR` **`4:REV`** `5:NTR` **`6:NEG`** `7:NTR` **`8:POS`** `9:NTR` **`10:NEG`** `11:NTR`

(3)

Polarity (sub)sequence		
Raw		Compressed
1:NTR	▷	1:NTR
2:POS	▷	**2:POS**
3:NTR	▷	3:NTR
4:REV	▷	**4:REV**
5:NTR 6:NTR 7:NTR 8:NTR 9:NTR	▷	5:NTR
10:NEG	▷	**6:NEG**
11:NTR	▷	7:NTR
12:POS	▷	**8:POS**
13:NTR 14:NTR 15:NTR 16:NTR 17:NTR 18:NTR 19:NTR	▷	9:NTR
20:NEG	▷	**10:NEG**
21:NTR 22:NTR 23:NTR	▷	11:NTR

From the raw polarity sequence that originally had 23 feature slices, a compressed quasi-compositional polarity sequence (i.e. *"Our$_1$ lives$_2$ will$_3$ never$_4$ be$_5$ ... lost$_6$ our$_7$ loved$_8$ ones$_9$ ... destroyed$_{10}$ Moussa$_{11}$"*) with only 11 feature slices (compression rate 52.17%) can therefore be derived. By 'quasi-compositional' we mean that the framework is aware of the fact that each compressed slice is composed of *n* sub-slices but does not attempt to analyse the composition: in other words, we jump directly from atomic prior sentiment (stemming from individual words) to more global sentiment without explaining the mapping(s) in between. The main assumption behind the quasi-compositional model is that, because of the null composition process, the compressed slices can still be expected to represent the *very same* sentiment information as their raw source slices: in the present example, although nearly half of the words were discarded, the sentiment information in the compressed 11 slices can be equated with that in the raw 23 slices.

Note that compressed polarity sequences can match a potentially very large number of unseen expressions regardless of which or how many words they contain because what is considered is the positions of the individual relevant sentiment polarities - *not the surface words* - in them. For example, a classifier trained on the present training example ought to be able to reason that an unseen chunk of text - be it a phrase with 11 words, a sentence with 25 words, or a document with 58 words - that contains compressed polarity slices ordered as NTR_POS_NTR_REV_NTR_NEG_NTR_POS_NTR_NEG_NTR can be negative. More importantly, if an unseen chunk of text fails to match any known sequence fully (e.g. when it is longer or shorter than any of the training examples), it is still likely to match many of the individual slice positions in the training data which means that the framework fails gracefully as the most optimal submatch can be expected in each case.

The advantage of the proposed polarity sequence model over simple *n*-gram modelling is that more information can in fact be captured because all key evidence can be accessed pertaining to the temporal (and hence positional) development of sentiment involving the smooth mixing, blending, figure/ground, and fading in/out behaviour amongst the three polarities ([10], [11]). Moreover, its advantage over more complex structural features is that polarity sequences may get rid of some unnecessary and untrue structural dependencies amongst words and syntactic constituents.

2.2 Feature Representation

In order for the compressed sequences to be used in supervised learning, we generate from each slice four separate features reflecting the polarity of the slice (i.e. POS, NTR, NEG, REV) represented with binary true/false values. The base polarity features can further be augmented with other information pertaining to various other properties of the words to which the feature slices point such as their word classes or grammatical roles. We consider further **non-sentiment-related features** from non-neutral and reversive words encompassing (i) word class tags (as output by a part-of-speech tagger) (§3.2), (ii) grammatical role tags (as output by a dependency parser) (§3.2), (iii) polarity word sense (WSD) ambiguity tags (as specified in a sentiment lexicon) (§3.1), and (iv) various combinations thereof. These additional non-sentiment-related features can be incorporated in two distinct ways. If **composite tags** are used, then additional non-sentiment-related evidence can be represented with more specific features. For example, the features from the above compressed polarity sequence can be enriched to include information such as the following (Ex. 4) (two sample slices shown):

	Word class	2:POS:N|8:POS:ADJ| ..
(4)	Syntax	4:REV:ADV|6:NEG:MAIN-V| ...
	WSD	2:POS:NTRPOS|10:NEG:NONE| ...

The classifier could then consider whether the eighth slice points to a positive adjective, whether the sixth one is a negative main verb, or whether the second slice points to a positive word that can also be neutral, for example. Another logical choice involves **parallel tags** amongst which additional non-sentiment-related evidence is scattered around multiple features. For example, parallel features such as the following can be had from the above compressed polarity sequence (Ex. 5) (two sample slices shown):

	Word class	2:POS|2:N|8:POS|8:ADJ| ...
(5)	Syntax	2:POS|2:SUBJ|6:NEG|6:MAIN-V| ...
	WSD	2:POS|2:NTRPOS|10:NEG|10:NONE| ...

In this case, the classifier could consider whether the second slice (i) is positive, (ii) points to a noun, (iii) functions as the subject in the sentence, and (iv) can be neutral or positive, respectively.

2.3 Training Data and Classifier

The learning models that we explore in this study were trained on two public domain data sets. The first ternary POS/NTR/NEG source, the *MPQA Opinion Corpus Version 2.0*[4] ([24]) (henceforth MPQA), yields 20822 (3993 (19.18%) POS, 7493 (35.99%) NEG, 9336 (44.84%) NTR) hand-labelled flat phrase- and sentence-level annotations from general news articles (inter-annotator agreement .72∼.82). Of the many different annotation types offered by the database, only *expressive subjectivity* and *direct subjectivity* annotations (*intensity* ∈ {*low*, *medium*, *high*, *extreme*}; *polarity* ∈ {*positive*, *negative*, *neutral*}) were included. Most of the training examples are short, with an average token count of ca. 2.69 (min. 1, max. 34, stdev. 2.29).

The second binary POS/NEG source, the *Sentence Polarity Data Set v1.0*[5] ([16]) (henceforth P&L), offers 10662 (5331 (50%) POS, 5331 (50%) NEG) flat sentence- and snippet-level annotations from (unverified) movie review star ratings mapped automatically onto binary sentiment polarities (inter-annotator agreement unknown). The P&L training examples are much longer, with an average token count of ca. 21.02 (min. 1, max. 59, stdev. 9.41).

In total, 18 models were trained from the two sources, in the conditions given in Table 1. The feature group label *pol* refers to base sentiment polarity features (§3.1), *wsd* to lexical polarity ambiguity features (§3.1), *pos* to word class features (§3.2), and *syn* to grammatical role features (§3.2). It can be seen that both training data sets could be captured with only a handful of slices (min. 20...27) which in turn translated into a small number of features (min. 58...99). Note that these figures are by a magnitude smaller than what would be the case if typical *n*-gram features were used as default unigrams would alone generate ca. 7800 (MPQA) vs. 18000 (P&L) features.

As a classifier of choice for the study, we used the Support Vector Machine implementation in the SVM.NET package with a linear kernel and all default parameters[6].

3 SENTIMENT PARSING

The previous sections illustrated the proposed framework that represents sentiment as compressed polarity sequences. The framework enables uniform sentiment parsing across words, phrases, and sentences without having to develop separate classifiers for different structural levels (e.g. running a sentence-level classifier to classify very short phrases). We combine the framework with a syntactic dependency parser to classify each individual syntactic constituent

[4] http://www.cs.pitt.edu/mpqa/

[5] http://www.cs.cornell.edu/People/pabo/movie-review-data/rt-polaritydata.tar.gz

[6] Johnson, M. (2008). SVM.NET 1.4. (www.matthewajohnson.org/software/svm.html). Based on Chang, C. & Lin, C. (2001). LIBSVM. (www.csie.ntu.edu.tw/˜cjlin/libsvm/).

Table 1. Summary of learning models

Feature Groups	Feature Type	Features		Slices	
		MPQA	P&L	MPQA	P&L
pol	composite	58	99	20	27
pol.wsd	composite	183	394	22	33
pol.pos	composite	157	347	21	28
pol.syn	composite	380	977	26	32
pol.wsd	parallel	270	975	24	34
pol.pos	parallel	303	1048	33	47
pol.syn	parallel	501	1842	38	50
pol.wsd.pos.syn	composite	1031	3449	28	36
pol.wsd.pos.syn	parallel	1331	5507	55	78

in a piecemeal fashion, one sentence at a time. Fully compositional sentiment parsing can be achieved by allowing the sentiment polarity sequence model to base its decisions on its own previous decisions amongst constituents and their subconstituents in an incremental and recursive manner. We however focus in this initial study on the general properties of sentiment polarity sequencing at various non-interacting structural levels and leave the investigation of full composition for future work.

3.1 Sentiment Lexicon

The underlying sentiment knowledge that our framework draws on comes in the form of an extensive sentiment lexicon which contains 57103 manually classified entries tagged with various properties relevant to compositional sentiment interpretation across adjectives (22402, 39.2%), adverbs (6487, 11.4%), nouns (19004, 33.3%), and verbs (9210, 16.1%). Included are positive (21341, 37.4%), neutral (7036, 12.3%), and negative (28726, 50.3%) entries as well as reversive operators (1700, 3.0%) which are words and phrases that can directly reverse the polarity of a non-neutral expression (e.g. "*reduce*$^{[\neg]}$", "*no*$^{[\neg]}$", "*prevention*$^{[\neg]}$"). The lexicon also contains for each entry sentiment word sense ambiguity (WSD) tags that specify whether a given entry (i) unambiguously displays only one polarity across its senses (NONE) (e.g. "*woefully*$^{(\,-)}$"); is binary-ambiguous within the binary choice space (ii) positive or neutral (POSNTR) (e.g. "*brilliant*$^{(+)(N)}$"), (iii) negative or neutral (NEGNTR) (e.g. "*rat*$^{(N)(\,-)}$"), (iv) positive or negative (POSNEG) (e.g. "*proud*$^{(+)(\,-)}$"); or (v) is fully ternary-ambiguous (ANY) (e.g. "*high*$^{(N)(+)(\,-)}$"). The proposed framework is not tied to our current lexicon as any sentiment lexica can be used instead.

3.2 Grammatical Analysis

Each sentence is input into an initial grammatical analysis which involves part-of-speech tagging and syntactic dependency parsing. The chosen dependency parser[7] (i) tokenises the sentence into individual tokens, (ii) lemmatises them, (iii) assigns word class and other morphological features to them, (iv) creates syntactic links between them, and (v) labels the links according to their syntactic and dependency functions and types. The resultant raw dependency links between individual words in the sentence are converted into a flat, non-binary constituent tree in which each word in the sentence is treated as a head of a syntactic constituent for which sets of optional immediate (non-recursive) pre-head and post-head dependents are constructed. The proposed framework is not dependent in any way on this parser as any component that offers part-of-speech tags and marks syntactic constituent boundaries can be plugged in.

3.3 Recursive Sentiment Analysis

1st Pass. For each parsed sentence, we then assign prior sentiment polarities and polarity reversal values to all tokens based on the sentiment lexica (§3.1). All unknown words are asserted as neutral by default. Sentiment parsing involves first identifying plausible entry points into the dependency tree of the sentence which typically encompass (i) the main lexical head verb of the root clause, (ii) the head noun of a main clausal verbless NP, or (iii) a stranded word not linked to any other word in the sentence. The parser first descends recursively down to the lowermost atomic child leaf constituent under an entry constituent, and then climbs the tree upwards recursively to calculate a sentiment polarity for each intermediate constituent until all constituents - and hence the whole sentence - have been analysed.

When parsing a constituent, the parser follows a fixed head-dependents combination schema in combining the constituent head (H_i) with *k* pre- ($L_{i-k\,:\,i-1}$) and *j* post-head ($R_{i+1\,:\,i+j}$) dependents in a specific sequence, namely 1) first combining post-heads ([*R*]) with the head in a rightward direction (starting with the post-head nearest to the head), and 2) then combining the pre-heads ([*L*]) with the head-post-heads set ([*HR*]) in a leftward direction (starting with the pre-head nearest to the head). Each time a head is combined with a dependent, a chunk of text which reflects the surface words subsumed by the head-dependent pair is input into the sentiment sequence classifier. The resultant predicted polarity class label is then considered as the current global polarity in the analysis so far.

We accept the probabilistic predictions in all but one situation: in cases where a constituent head lacks any dependents (i.e. is made of just a singular word), we bypass the classifier and instead resort to the polarity assigned to the word in the lexicon. The reason for this simple exception is that there is no guarantee that the probabilistic classifier does not (i) override the prior polarity assigned to a word in the lexicon or (ii) render a neutral word non-neutral (e.g. inputting a NTR word into a binary POS/NEG model) in which case the framework would cease to be grounded on lexical knowledge. Note that our goal is to classify *combinations* of words, not individual words.

2nd Pass. The above 1st pass in the sentiment parsing process assigns sentiment to all syntactic constituents in a given sentence which ultimately results in all individual surface words displaying the final top-level compositional sentiment polarity/ies. In real-world use scenarios, the success (or the failure) of a sentiment algorithm will be judged based on whether or not the sentential polarities that individual surface words display make sense and 'read well'. It is unfortunately possible that some surface words end up displaying a polarity that appears incongruous with respect to the rest of the sentence. Such anomalies can stem from fragmentary grammatical analyses or arise when the classifier suggests a neutral polarity for a sentence even though it contains words which bear a *known* non-neutral polarity in the lexicon.

A further 2nd pass is therefore required to hide any traces of fragmentary or inconsistent analyses at the top sentence level. On the basis of the general tendency towards a coherent polarity flow within/across sentences (cf. [10], [11]), we accordingly account for 1) **neutral polarity gaps** (i.e. stranded neutral words amidst non-neutral words), and for 2) **non-neutral islands** (i.e. stranded non-neutral words that clearly disagree with the global majority sentiment of the sentence). For both gaps and islands, we simply execute a bidirectional lookup method around each incongruous surface word, and use the polarity evidence from their neutral/non-neutral neighbours as a heuristic masking polarity.

[7] Connexor Machinese Syntax 3.8.1. http://www.connexor.com/

4 EXPERIMENTS

Evaluating the performance of the proposed framework is not as straightforward as it seems. Firstly, because the sentiment sequence model is applied across all structural levels as part of exhaustive sentiment parsing, the targeted classification task is ultimately a ternary POS/NTR/NEG one for not all constituents are non-neutral: however, most public-domain gold standards come with binary POS/NEG annotations only. Accordingly, if a ternary classifier's output is evaluated against a binary gold standard (or vice versa), any conclusions that may be drawn are partial in the strictest sense. Secondly, since our framework assigns sentiment labels to all constituents in sentences, it is by no means clear which constituents ought to be evaluated. For example, if a gold standard contains expressions with arbitrarily chosen boundaries, there is no guarantee that the classifier's syntactic constituents map fully onto them (in fact they rarely do). As we are not aware of any manually-annotated and verified multi-level sentiment treebanks for English at the time of writing, we instead resort to three different gold standards which collectively shed light on the strengths and weaknesses of the framework at different structural levels. Due to these complications, we focus mainly on strictly binary evaluation conditions (whereby neither NTR predictions by the classifier nor NTR cases in the gold standard (if present) are considered) as they are much more indicative of core sentiment judgements.

4.1 Gold Standard Data Sets

Headlines [SEMEVAL]. The first data set comprises 1000 news headlines from the SemEval-2007 Task #14 annotated for polarity along the scale [-100...-1|0|1...100] (46.80% POS), 0.60% NTR, 52.60% NEG) (six annotators, inter-annotator agreement *r* .78) ([23])[8]. We included only the POS ([+1...+100]) and NEG ([-100...-1]) entries in the evaluation, and compare the classifier's sentential polarity against each headline. Ex. 6 illustrates sample headlines from the data set.

(6) [+32] *Test to predict breast cancer relapse is approved*
[-48] *Two Hussein allies are hanged, Iraqi official says*

Phrases [MPQA]. Evaluation targeting phrase-level expressions is based on the MPQA data set (§2.3) which we utilise for both ternary POS/NTR/NEG and binary POS/NEG evaluation. Ex. 7 illustrates a sample expression annotation in a sentence (annotation underlined).

(7) [LOW][POS] *Private organizations are also being encouraged to help fight sandstorms, according to the administration's vice-director Li Yucai.*

The MPQA expressions are considered in isolation without any contextual evidence from their hosting sentences in the MPQA database in order to avoid any subjective mappings or overlapping measures between the MPQA expression boundaries and our parser's constituents. In this condition, we compare the top-level polarity output by the classifier against each expression.

Snippets [P&L]. Further sentence- and snippet-level evaluation data come from the P&L data set (§2.3). Because a given snippet may consist of multiple sentences, we evaluate the majority 'document-level' polarity output by the classifier against each snippet in this condition. Ex. 8 illustrates a sample sentence from the data set.

(8) [NEG] *it wouldn't be my preferred way of spending 100 minutes or $7.00.*

4.2 Evaluation Measures and Baselines

A large number of different evaluation measures can be used to characterise the performance of the models, each of which highlights a different evaluative aspect. We hence evaluate the models using multiple complementary measures. The first measure family targets the conventional notion of 'accuracy' used in traditional factual classification tasks encompassing **Accuracy**, **Precision**, and **Recall** measures. For these, individual pairwise polarity decisions (POS vs. NOT-POS, NTR vs. NOT-NTR, NEG vs. NOT-NEG) were used. The second measure family focuses on different levels of **agreement** and **correlation** between human sentiment judgements and our models by calculating chance-corrected rates based on the standard **Kappa** *k*, **Pearson**'s *r* product moment correlation coefficient, and **Krippendorff**'s α reliability coefficient measures. In ternary POS/NTR/NEG classification, not all classification errors are equal because classifying a POS case as NTR is more tolerable than classifying it as NEG, for example. We lastly characterise three distinct **error types** between human *H* and algorithm *A*, namely 1) FATAL errors ($H^{(\alpha)}A^{(\neg\alpha)}$ $\alpha\in\{+ -\}$), 2) GREEDY errors ($H^{(N)}A^{(\alpha)}$ $\alpha\in\{+ -\}$), and 3) LAZY errors ($H^{(\alpha)}A^{(N)}$ $\alpha\in\{+ -\}$).

The models are further compared against three baselines, namely **positive** (POS_BASE), **negative** (NEG_BASE), and **majority** sentiment using raw polarity frequency counting (FREQ_BASE).

4.3 Results

[SEMEVAL]. Starting with the short headlines, Table 2 highlights the performance of the models in the 2-way POS/NEG condition. In overall, the results are highly encouraging on both training data sets and are comparable with sample levels reported in other studies ([23])[9]. MPQA training data yielded clearly better scores than P&L data because (i) the former contains much more training data, and (ii) the MPQA expressions and the SEMEVAL headlines are of similar lengths. Both training data sets surpassed the POS_BASE (47.08) and NEG_BASE (52.92) baselines while the P&L models struggled to outperform the very high FREQ_BASE level at 71.53. Binary accuracy levels range from 71.03 to 77.94 while precision varies interestingly between the two polarities in that positive sentiment (72.47∼84.14) is more precise than negative sentiment (71.45∼76.94). Recall in turn displays a reverse pattern as positive sentiment has a considerably lower recall (62.80∼66.88) than negative sentiment (84.41∼92.41). Agreement levels point towards moderate levels at around 52.47∼54.49.

[MPQA]. Models trained on the P&L training data reached even more promising rates on the MPQA data set which is shown in Table 3 (2-way POS/NEG condition). All models surpassed the accuracy baselines (POS_BASE (34.76), NEG_BASE (65.24), FREQ_BASE (70.32)). The scores are especially significant because the slices from the MPQA and P&L training data differ considerably in length. Again, the models perform well against reported levels reached in other studies[10]. While binary accuracy rose to 84.73, agreement

[8] http://www.cs.unt.edu/~rada/affectivetext

[9] Cf. the highest reported 3-way figures in [23] are ∼55.10 (accuracy), ∼61.42 (precision), and ∼66.38 (recall). Note, however, that their evaluation conditions are not strictly identical with ours.

[10] Cf. the highest reported 2-way figures in [25]: 421 are ∼74.5 (POS precision), ∼87.8 (NEG precision), ∼77.8 (POS recall), and ∼98.3 (NEG recall). [4] report ∼88.5...90.7 2-way accuracies. Note, however, that their evaluation conditions are not strictly identical with ours.

Table 2. Experimental results on the SEMEVAL data set, 2-way POS/NEG condition (↑= boost over pol features)

	Trained on 3-way MPQA [20882], tested on 3-way Semeval headlines [784...841]								
Feature Groups	pol	pol wsd	pol pos	pol syn	pol wsd	pol pos	pol syn	pol wsd pos syn	pol wsd pos syn
Feature Types	composite				parallel			composite	parallel
Acc POS/NEG	**77.94**	74.88	76.74	77.30	76.68	76.96	77.38	73.20	74.79
Prec POS	**84.14**	82.09	80.50	79.46	76.17	82.13	81.74	76.29	76.75
Prec NEG	75.52	72.47	75.17	76.43↑	**76.94**↑	74.87	75.53↑	71.99	73.86
Rec POS	57.36	49.85	57.57↑	57.42↑	**62.80**↑	56.93	58.63↑	51.75	58.26↑
Rec NEG	**92.41**	92.39	90.21	90.30	86.34	91.19	90.70	88.52	86.98
Kappa	**52.24**	44.88	49.89	50.12	50.48	50.36	51.43	42.21	46.72
Pearson	**54.49**	48.01	51.57	51.64	51.08	52.38	53.15	44.09	47.85
Krippendorff	51.43	47.57	51.14	50.35	**52.76**↑	51.12	51.87↑	48.11	51.28

	Trained on 2-way P&L [10662], tested on 3-way Semeval headlines [994]								
Feature Groups	pol	pol wsd	pol pos	pol syn	pol wsd	pol pos	pol syn	pol wsd pos syn	pol wsd pos syn
Feature Types	composite				parallel			composite	parallel
Acc POS/NEG	70.52	**71.03**↑	65.90	69.11	67.40	70.62↑	70.82↑	62.47	64.29
Prec POS	69.40	72.06↑	72.01↑	70.18↑	68.95	70.66↑	**72.47**↑	68.06	63.75
Prec NEG	**71.45**	70.31	63.34	68.40	66.45	70.60	69.73	60.47	64.67
Rec POS	**66.88**	62.82	45.09	59.83	55.98	64.32	61.32	38.25	55.98
Rec NEG	73.76	78.33↑	**84.41**↑	77.38↑	77.57↑	76.24↑	79.28↑	84.03↑	71.67
Kappa	40.73	**41.44**↑	30.12	37.51	33.90	40.75↑	40.95↑	22.83	27.84
Pearson	40.75	**41.75**↑	32.29	37.89	34.46	40.90↑	41.40↑	25.21	28.03
Krippendorff	**52.47**	52.21	45.30	50.61	48.94	52.21	51.84	41.61	47.21

with human sentiment annotations is closer to substantial levels (57.84∼67.20). A clear asymmetry towards negative sentiment can be attested as both negative precision (90.60) and negative recall (89.20) are higher than positive precision (78.10) and recall (83.06) (cf. similar observations in [25]: 421).

[P&L]. The P&L data set interestingly appears more challenging for models trained on the MPQA training data as can be seen in the markedly lower levels shown in Table 4 (2-way POS/NEG condition). Binary accuracy decreased to 61.65 while agreement rates dropped to the level of only fair agreement (23.20∼43.50). Although they surpassed the POS_BASE and NEG_BASE baselines (50), the models are just below the FREQ_BASE baseline (61.57). A polarity asymmetry can once again be observed between higher positive precision (72.76 vs. 58.23 (NEG)) vs. higher negative recall (87.40 vs. 42.79 (POS)). The unexpectedly lower performance stems from the disparity in the number of slices in the (3-way) MPQA and (2-way) P&L data sets. An alternative conclusion drawable from the cross-training and -testing between the MPQA and P&L data sets is that the polarity sequence model may work better when the training data (P&L) contains more slices than the test data (MPQA). Note however that the P&L data set is replete with sarcasm, irony, and unknown words not found in our lexica.

Neutral Polarity. Since our goal was to maximise the amount of training data for the models, we employed the MPQA data set in its entirety. We moreover aimed at emulating real-world conditions by using strictly separate data sets for training and testing instead of cross-validation conditions of any kind (e.g. [4], [25]). Unfortunately, no unseen testing data with neutral polarity instances were then available for our experiments as only the MPQA data set contains ternary annotations. In order to estimate the neutral polarity performance of the models, we examined the relative performance of neutral polarity against non-neutral polarities using the base polarity *pol* model on the MPQA data set itself. Note that because we train and test on the same data set, the figures are understandably higher that what can be expected from unseen neutral annotations in the future. Nevertheless, many useful observations can be made based on the figures in Table 5. The inclusion of neutral polarity is likely to have an adverse effect on overall performance - an observation which concurs with the general trend in the area (e.g. [13], [25]). In our experiments, neutral recall was somewhat low (62.11) but its accuracy (72.03) and precision (71.72) were still high relative to the non-neutral levels. If we consider the error types in the ternary condition, only 14.14% of the errors were FATAL: the high level of GREEDY errors (52.15) indicates that the models may display oversensitivity towards non-neutral sentiment. For reference, we also report the corresponding ternary rates offered by the same model trained on binary P&L data. Note however that all neutral predictions in this condition come from singular words that bypassed the classifier altogether (see §3.3). The general pattern is the same, albeit somewhat more pronounced.

Features. We lastly consider the relative merits of individual feature groups across all data sets. The first clearly evident pattern is that mere polarity features (*pol*) are generally highly effective - especially considering that *no n-gram evidence was used in any form.* It is in fact surprising that so few features (58 (MPQA), 99 (P&L)) can even reach such high rates with highest accuracies touching on 84.73, precision levels up to 90.60, and recall levels up to 92.41 in some cases. More intriguing is the evidence pertaining to the expected utility of the extra non-sentiment-related feature groups. On the one hand, sentiment WSD, word class, and syntactic information do facilitate the analysis in many cases. On the other hand, they also hurt the performance of the base features in a number of cases. Although all of the extra features help in some condition, none of them can be said to help categorically. The single most useful supporting role is played by word-level sentiment WSD features which gave a boost most often in 24 conditions (13 composite, 11 parallel), indicating that the WSD tags can crudely mask the sentiment ambiguity amongst the slices. The support given by word class and syntactic information was not as high as expected since both boosted the base features in 19 conditions (word class: 10 composite, 9 parallel; syntax: 6 composite, 13 parallel). This in turn seems to suggest that either more training data are required or that morphosyntactic information is subservient to mere linear polarity sequences. Against the conventional

Table 3. Experimental results on the MPQA data set, 2-way POS/NEG condition (↑= boost over pol features)

	Trained on 2-way P&L [10662], tested on 3-way MPQA [10709]								
Feature Groups	pol	pol wsd	pol pos	pol syn	pol wsd	pol pos	pol syn	pol wsd pos syn	pol wsd pos syn
Feature Types	composite				parallel			composite	parallel
Acc POS/NEG	84.21	**84.73**↑	83.89	83.31	83.79	83.95	80.53	77.82	81.04
Prec POS	74.28	75.17↑	76.64↑	73.13	**78.10**↑	74.12	70.19	68.93	72.93
Prec NEG	90.38	**90.60**↑	87.68	89.63	86.51	89.98	86.59	82.06	85.16
Rec POS	82.76	**83.06**↑	76.49	81.41	73.48	81.95	75.43	64.69	71.39
Rec NEG	84.97	85.61↑	87.77↑	84.31	**89.20**↑	85.00↑	83.20	84.71	86.11↑
Kappa	65.94	**67.00**↑	64.29	64.00	63.57	65.31	57.61	50.13	57.79
Pearson	66.18	**67.20**↑	64.29	64.22	63.64	65.51	57.70	50.19	57.80
Krippendorff	57.58	**57.84**↑	55.59	56.89	54.56	57.26	54.25	50.04	53.14

Table 4. Experimental results on the P&L data set, 2-way POS/NEG condition (↑= boost over pol features)

	Trained on 3-way MPQA [20882], tested on 2-way P&L [9743...10313]								
Feature Groups	pol	pol wsd	pol pos	pol syn	pol wsd	pol pos	pol syn	pol wsd pos syn	pol wsd pos syn
Feature Types	composite				parallel			composite	parallel
Acc POS/NEG	60.37	60.13	61.01↑	59.86	**61.55**↑	60.12	60.42↑	59.00	60.67↑
Prec POS	**72.76**	72.48	69.08	69.91	68.93	68.46	68.22	67.76	67.00
Prec NEG	56.65	56.41	57.70↑	56.71↑	**58.23**↑	56.88↑	57.20↑	56.03	57.68↑
Rec POS	33.49	33.41	40.13↑	33.58↑	42.57↑	38.14↑	39.63↑	34.35↑	**42.79**↑
Rec NEG	**87.40**	87.17	81.97	85.76	80.67	82.28	81.38	83.65	78.72
Kappa	20.85	20.51	22.08↑	19.41	**23.20**↑	20.38	20.97↑	18.01	21.48↑
Pearson	24.78	24.38	24.33	22.69	**25.12**↑	22.75	23.11	20.70	23.04
Krippendorff	39.89	39.82	42.51↑	39.62	**43.50**↑	41.52↑	42.12↑	39.70	43.16↑

principle of crude force, wielding all of the features did not help as only 9 conditions benefited from them (2 composite, 7 parallel): in many cases, they too proved counterproductive. We hypothesise that this is due to the sparser feature spaces involved. Regarding which feature representation option - composite vs. parallel - is optimal, no firm conclusions can be drawn.

In overall, the boost given by the extra non-sentiment-related features over the base polarity features can range between only 1.02 (Pearson) and as much as 10.65 (negative recall) (cf. +1.18 (agreement), +1.58 (negative precision), +2.35 (Kappa), +3.61 (Krippendorff), +3.83 (positive precision), +9.31 (positive recall)). However, their adverse effects are much more pronounced, potentially ranging from as much as -5.76 (positive precision) to -28.63 (positive recall) (cf. -8.05 (agreement), -8.68 (negative recall), -10.86 (Krippendorff), -10.99 (negative precision), -15.99 (Pearson), -17.90 (Kappa)).

5 RELATED WORK

Sentence and Phrase-level Sentiment Analysis. A wide range of different approaches have been attempted. At the base level, mere frequency counting ([6]) with naïve analytical or learning additions ([3], [6], [10]) can offer moderate accuracies in some tasks. Various more complex machine learning approaches have incorporated shallow structural features ([1], [3], [25]), or joint classification models that target the structural co-dependency between individual sentences and documents using constrained inference ([12]). At the other end of the spectrum, a number of explicit direct fully- or shallow-compositional sentiment logics have been developed most of which rely on hand-written combinatory rules and lexical sentiment seeds in conjunction with semantic scope-driven valence shifters ([18]); fully compositional syntax-driven parsing ([13], [21]); structured inference-based learning with lexical, negator, and voting features ([4]); cascaded pattern matching with shallow phrasal chunking ([8]); learning-based topic classifiers with shallow phrasal chunking ([14]); verb-centric event frames with scored knowledge bases ([20]); or other heuristic linking and ranking patterns ([15]).

Positional Features. Even though they appear intuitively useful, positional features have so far been somewhat underrepresented in the area. Past attempts have focused on simple positional information within sentences ([9]), documents ([17]), or discourse ([22]). The solution closest to our sequence model is the sequential approach in [11] who model global document-level sentiment using a temporal trajectory function from local sentential polarities calculated by an Isotonic Conditional Random Field-based classifier. None of the above are driven by any compositional sentiment processes.

Feature Reduction and Compression. Various feature reduction techniques have been used in conjunction with sentiment learning. Typically, they operate on *n*-gram features and remove redundant or weak features through subsumption ([19]), abstraction ([7]), log likelihood ratio filters ([5]), or more sophisticated search criteria ([2]) amongst others. The guiding force behind our proposed feature reduction mechanism is in contrast the fundamental, linguistically justified, null composition principle. A conceptually analogous approach to sentiment compression is mentioned in [26] who, in measuring controversy in social media, construct polarity 'micro-state vectors' from words' polarity intensities and then similarly try to compress them. However, they leave all feature reduction decisions to standard compression algorithms agnostic of any compositional sentiment processes.

6 CONCLUSION

We have described a simple, yet effective, hybrid sentiment learning and parsing framework which is grounded on one basic logically defensible compositional sentiment process and which uses additional supervised learning to deal with more complex sentiment processes. The proposed framework, which offers a natural, yet principled basis for sentiment reasoning, operates on quasi-compositional sentiment polarity sequences which succinctly capture the sentiment in syntactic constituents across different structural levels without any conven-

Table 5. Experimental results on the MPQA data set, 3-way POS/NTR/NEG condition

	Tested on 3-way MPQA [20882]											
	Accuracy			Precision			Recall			Error Severity		
Trained on	POS	NTR	NEG	POS	NTR	NEG	POS	NTR	NEG	FATAL	GREEDY	LAZY
3-way MPQA [20882]	79.97	72.03	82.85	48.48	71.72	78.55	71.25	62.11	72.00	14.14	52.15	33.72
2-way P&L [10662]	77.37	65.51	71.89	44.73	79.05	57.96	76.36	31.39	79.66	19.06	72.19	8.76

tional *n*-gram features. It can be used for uniform sentiment classification across words, phrases, and sentences, and requires only simple flat phrase- or sentence-level sentiment annotations, a sentiment lexicon, and, optionally, a part-of-speech tagger and a syntactic parser. The results obtained with the initial implementation are highly encouraging and suggest that simple linear polarity sequence features alone operate effectively.

REFERENCES

[1] Apoorv Agarwal, Fadi Biadsy, and Kathleen Mckeown, 'Contextual phrase-level polarity analysis using lexical affect scoring and syntactic n-grams', in Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2009), pp. 24–32, Athens, Greece, (March 30 - April 3 2009).

[2] Edoardo Airoldi, Xue Bai, and Rema Padman, 'Markov blankets and meta-heuristics search: Sentiment extraction from unstructured texts', in Advances in Web Mining and Web Usage Analysis: Revised Selected Papers from the 6th International Workshop on Knowledge Discovery on the Web (WebKDD 2004), 167–187, Seattle, WA, USA, (August 22-25 2006).

[3] Alina Andreevskaia and Sabine Bergler, 'CLaC and CLaC-NB: Knowledge-based corpus-based approaches to sentiment tagging', in Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), pp. 117–120, Prague, Czech Republic, (June 23-24 2007).

[4] Yejin Choi and Claire Cardie, 'Learning with compositional semantics as structural inference for subsentential sentiment analysis', in Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2008), pp. 793–801, Honolulu, Hawaii, (October 25-27 2008).

[5] 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 (COLING 2004), pp. 841–847, Geneva, Switzerland, (August 23-27 2004).

[6] Minqing Hu and Bing Liu, 'Mining and summarizing customer reviews', in Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2004), pp. 168–177, Seattle, Washington, USA, (August 22-25 2004).

[7] Mahesh Joshi and Carolyn Penstein-Rosé, 'Generalizing dependency features for opinion mining', in Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP (ACL-AFNLP 2009), pp. 313–316, Singapore, (August 4 2009).

[8] Manfred Klenner, Angela Fahrni, and Stefanos Petrakis, 'Polart: A robust tool for sentiment analysis', in Proceedings of the 17th Nordic Conference on Computational Linguistics (NODALIDA 2009), pp. 235—238, Odense, Denmark, (May 14-16 2009).

[9] Lun-Wei Ku, I-Chien Liu, Chia-Ying Lee, Kuan hua Chen, and Hsin-Hsi Chen, 'Sentence-level opinion analysis by CopeOpi in NTCIR-7', in Proceedings of the 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Cross-Lingual Information Access (NTCIR 7), pp. 260–267, Tokyo, Japan, (December 16-19 2008).

[10] Hugo Liu, Henry Lieberman, and Ted Selker, 'A model of textual affect sensing using real-world knowledge', in Proceedings of the ACM International Conference on Intelligent User Interfaces (IUI 2003), pp. 125–132, Miami, Florida, USA, (January 12-15 2003).

[11] Yi Mao and Guy Lebanon, 'Isotonic conditional random fields and local sentiment flow', in Advances in Neural Information Processing Systems: Proceedings of the 2006 Conference (NIPS 2006), number 19, 961–968, Vancouver, British Columbia, Canada, (December 4-7 2007).

[12] Ryan McDonald, Kerry Hannan, Tyler Neylon, Mike Wells, and Jeff Reynar, 'Structured models for fine-to-coarse sentiment analysis', in Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL 2007), pp. 432–439, Prague, Czech Republic, (June 23–30 2007).

[13] Karo Moilanen and Stephen Pulman, 'Sentiment composition', in Proceedings of the Recent Advances in Natural Language Processing International Conference (RANLP 2007), pp. 378–382, Borovets, Bulgaria, (September 27-29 2007).

[14] Kamal Nigam and Matthew Hurst, 'Towards a robust metric of opinion', in Computing Attitude and Affect in Text: Theory and Applications; Papers from the 2004 AAAI Spring Symposium (AAAI-EAAT 2004), Stanford, USA, (March 22-24 2004).

[15] Alexander Osherenko, 'Towards semantic affect sensing in sentences', in Proceedings of the Symposium on Affective Language in Human and Machine at the Communication, Interaction and Social Intelligence Convention (AISB 2008), pp. 41–44, Aberdeen, UK, (April 1-4 2008).

[16] 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 (ACL 2005), pp. 115–124, Ann Arbor, USA, (June 25-30 2005).

[17] 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 (EMNLP 2002), pp. 79–86, Philadelphia, PA, USA, (July 6-7 2002).

[18] Livia Polanyi and Annie Zaenen, 'Contextual valence shifters', in Computing Attitude and Affect in Text: Theory and Applications; Papers from the 2004 AAAI Spring Symposium (AAAI-EAAT 2004), 106–111, Stanford, USA, (March 22-24 2004).

[19] Ellen Riloff, Siddharth Patwardhan, and Janyce Wiebe, 'Feature subsumption for opinion analysis', in Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-06), pp. 440–448, Sydney, Australia, (July 22-23 2006).

[20] Mostafa Al Masum Shaikh, An analytical approach for affect sensing from text, Ph.D. dissertation, The Graduate School of Information Science and Technology, University of Tokyo, 2008.

[21] František Simančík and Mark Lee, 'A CCG-based system for valence shifting for sentiment analysis', in Advances in Computational Linguistics: Proceedings of 10th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2009), volume 41 of Research in Computing Science, 93–102, Mexico City, Mexico, (March 1-7 2009).

[22] Swapna Somasundaran, Janyce Wiebe, and Josef Ruppenhofer, 'Discourse level opinion interpretation', in Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008), pp. 801–808, Manchester, UK, (August 18-22 2008).

[23] Carlo Strapparava and Rada Mihalcea, 'Semeval-2007 task 14: Affective text', in Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), pp. 70–74, Prague, Czech Republic, (June 2007).

[24] Janyce Wiebe, Theresa Wilson, and Claire Cardie, 'Annotating expressions of opinions and emotions in language', Language Resources and Evaluation, **39**(2-3), 165–210, (May 2005).

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

[26] Zhu Zhang and Xin Li, 'Controversy is marketing: Mining sentiments in social media', in Proceedings of the 43rd Hawaii International Conference on System Sciences (HICSS-43 2010), pp. 1–10, Hawaii, USA, (January 5-8 2010).

Using Text Summaries for Predicting Rating Scales

Horacio Saggion[1] and **Elena Lloret**[2] and **Manuel Palomar**[3]

Abstract. This paper presents a detailed analysis of a wide range of text summarization approaches within the rating-inference task. This task consists of associating a fine-grained numerical rating to an opinionated document. We collect a small dataset of bank reviews that have been rated from 1 to 5 by real users. Then, we use a Support Vector Machines learning paradigm to predict the correct rating, employing both the full review and review's automatic summaries. We suggest three types of summaries – generic, query-focused and sentiment-based – of five compression rates (10% to 50%) to further investigate whether they are useful or not for associating the correct star-rating to a review in comparison to the use of the whole review. We compute the *Mean Squared Error* in an attempt to find a specific type of summary and compression rate that performs better over all the rest. The results obtained are very encouraging, and although they are very preliminary to claim a strong tendency for a particular summarization approach, they show that query-focused and sentiment-based summaries may be the most appropriate kinds of summaries for tackling the rating-inference problem.

1 INTRODUCTION

The Web 2.0 (also known as social web) has led to the emergence of new types of websites, such as blogs, forums, or social networks, where everybody can express his/her emotions and likes with respect to a specific product or service. This way of stating opinions has become very popular on the Internet, providing a good source of recommendations that people always take into account for a decision-making process. For example, between 73% and 87% of Internet users have reported that reviews had a significant influence on their purchase [38]. What people think of a product, service, etc. is of great value for companies, and therefore business analysts are turning their eyes to the Web in order to monitor public perception on products, services, policies, and managers. However, the exponential and fast growth of the information available on the Internet makes this task a real challenge difficult to cope with.

Natural Language Processing (NLP) groups a series of tasks (e.g. information retrieval, text summarization, sentiment analysis) capable of providing methods and tools to efficiently deal with all this information. Recently, sentiment analysis has emerged as an interesting area of NLP, which is present in a number of relevant conferences (*Text REtrieval Conference*[4], *Text Analysis Conference*[5], *DEfi Fouille de Textes program* [18]). Its main aim is to provide methods and tools for identifying and classifying opinions or subjective text. Normally, this is done by considering the task as an opinion-classification problem (*subjective vs. objective,* or *positive vs. neutral vs. negative*) ([48], [10]). However, it is frequent to find texts where users give a score, depending on how much they liked or not a product, movie, restaurant, hotel, service, etc. which is normally associated with a scale rating (1=worst,...5=best). This particular task within sentiment analysis is known as rating-inference, and can be defined as the task of identifying the author's evaluation of an entity with respect to an ordinal-scale based on the author's textual evaluation of the entity [37]. Moreover, the use of text summaries has been proven to be useful for several NLP tasks, such as information retrieval [49], question answering [6], sentiment analysis [24] or text classification [22], obtaining successful results when summaries are used instead of full documents to accomplish the task. This also allows us to evaluate a summary extrinsically, whose main objective is to assess the summarization approach based on how it affects the completion of some other task [30].

Text summarization and sentiment analysis have been combined together in order to produce opinion-oriented summaries. However, very little research has been carried out to exploit summarization as a text processing step for the rating-inference problem. The objective of this paper is to present an initial analysis of the usefulness of different types of text summaries for the rating-inference problem. Furthermore, this work can also be considered a case of extrinsic evaluation methodology in text summarization where the objective is to verify if, although imperfect in its current state of the art, text summarization can help solving this particular problem. We therefore put forward the following research question "what type of summary and compression rates could be used instead of the full document for the rating-inference task?". Although the framework we are presenting here can accommodate any type of summarization (i.e., extractive, non-extractive) we are only reporting experiments with a limited set of summary types which we consider have widespread use in the NLP community. The results will show that query-focused and sentiment-based summaries are the most appropriate kinds of summaries for this particular task over other kinds of summarization strategies.

This paper is organized as follows: In Section 2 some existing work regarding the rating-inference task, as a specific problem of sentiment analysis, is discussed. Next, in Section 3 the different summarization approaches are described. The whole process concerning the rating-inference is explained in Section 4. Afterwards, the extensive experimental framework is provided in Section 5, together with the corresponding evaluation and discussion. Finally, the most important conclusions and some research lines for further work are presented in Section 6.

[1] Department of Infomation and Communication Technologies, Grupo TALN, Universitat Pompeu Fabra, Barcelona, Spain, e-mail: horacio.saggion@upf.edu

[2] Department of Software and Computing Systems, University of Alicante, Spain, e-mail: elloret@dlsi.ua.es

[3] Department of Software and Computing Systems, University of Alicante, Spain, e-mail: mpalomar@dlsi.ua.es

[4] http:trec.nist.gov/

[5] http://www.nist.gov/tac/

2 BACKGROUND

Recently, the surge of interest in new systems that deal directly with opinions has resulted in a major attention to the problem of identifying and classifying sentiment in text by the NLP research community [38]. Whereas most of the literature addresses the problem of detecting and classifying opinions at a sentence level [48], [1], [10], there is less research which aims to detect the overall sentiment of a document [39], [33], [19]. Furthermore, the vast amount of on-line information available these days has brought great interest in text summarization. The goal of this NLP field is to provide methods and algorithms that condense information providing users with shorter texts and therefore, helping them to manage all the information in a more efficient way.

Although work in text summarization started in the late fifties [28], the development of the Internet, the availability of massive textual databases together with international evaluation efforts have fuelled research in this field. Most summarization today is based on a sentence-extraction paradigm where a list of features believed to indicate the relevance of a sentence in a document or set of documents [29] is used as interpretation mechanism. Basic statistical approaches to the selection of sentences have relied on the use of frequency computation (e.g., tf*idf) to identify relevant document keywords [28], position of sentences in text [11], [23], presence of title words in sentences [11], or presence of cue-words [11]. These features are usually combined to produce sentences' scores [11], or they are tuned by a machine learning approach [20]. Knowledge-rich approaches to text summarization incorporate knowledge from lexical resources such as WordNet [14] or apply discourse organization theories. Lexical cohesion has been used to measure sentence relevance in summarization [5] and Rhetorical Structure Theory (RST) [31] has been used to select key sentence components to create summaries [32]. In more formal domains (e.g. legal texts, scientific literature) modelling information types which are likely to be relevant for summaries are usually employed [45], [43]. Information Extraction [17] has also been used for document analysis and mapping to stereotypical template structures which are in turn used as the basis for summary generation [36]. Current summarization methods also use graph-based algorithms [34], [12], where the degree of relatedness of the sentences is measured in order to identify relevant information. Similarly, the idea of using centroids to account for central pieces of information serves for determining the important topic of the document and then how close a sentence is from the centroid can be measured, thus establishing its degree of importance within the document [40].

In recent years, the subjectivity appearing in documents has led to a new emerging type of summaries: sentiment-based summaries, which have to take into consideration the sentiment a person has towards a topic. Examples of approaches that generate this type of summaries can be found in [7], [4] or [3]. In this way, the combination of sentiment analysis and summarization techniques can result in great benefits for stand-alone applications of sentiment analysis, as well as for the potential uses of sentiment analysis as part of other NLP applications [44]. One of these sentiment analysis applications is that of rating-inference problem. For example, given a review and star-rating classification scale (e.g. star values ranging from 1 to 5), this task should correctly predict the review's rating, based on the language and sentiment expressed in its content. This differs from the traditional approach of binary sentiment classification, where documents are classified into *positive vs. negative*, or into *subjective vs. objective*, according to their polarity and subjectivity degree, respectively. Specific work dealing with this problem is addressed for example in [37]. Here, the rating-inference problem is analyzed for the movies domain and different configurations of Support Vector Machines (SVM) (one vs. all, regression and metric labelling) are employed. The ratings ranged from 1 up to 4 stars, depending on the degree the author of the review liked or not the film. Focusing also on movie reviews, the approach described in [21] suggests the use of collaborative filtering algorithms together with sentiment analysis techniques to obtain user preferences expressed in textual reviews. Once opinion words from user reviews have been identified, the polarity of those opinion words together with their strength need to be computed and mapped to the rating scales to be further inputted to the collaborative filtering algorithm. In [42], the rating-inference problem is addressed for product reviews, in which the text associated to the review is rather short. They use Support Vector Machines, reporting a classification accuracy of 80% for binary classification (thumbs up or thumbs down) and a 74% for a 1 to 5-star rating scale, when dealing with very-short texts (2 or 3 sentences).

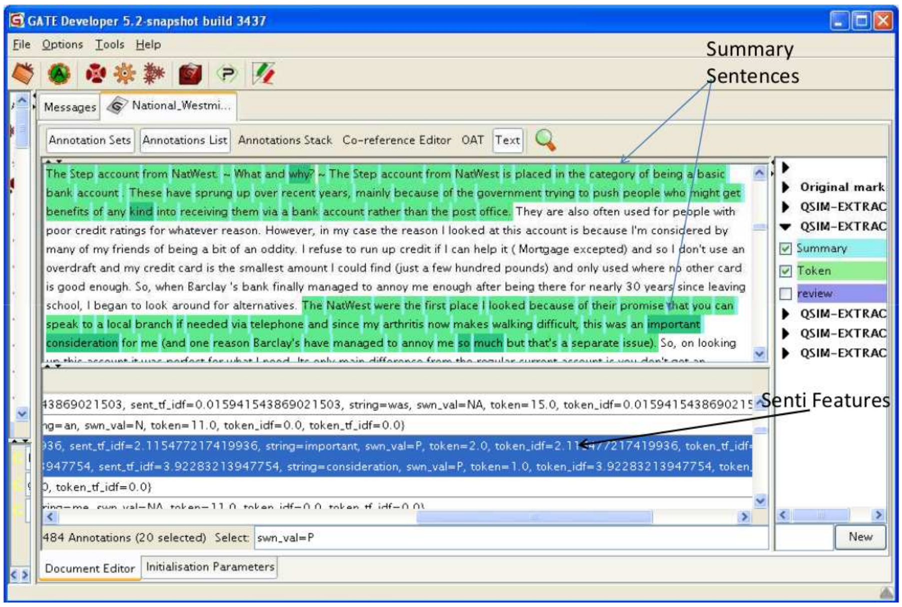
Differently from the previous approaches, in [2] the rating of different features regarding hotel reviews (cleanliness, location, staff, etc.) is addressed by analyzing three aspects involved in the generation of product review's representations. First, a part-of-speech tagger is used to extract complex features based on patterns. Then, an opinion lexicon is employed in order to make the extracted features more robust; and finally, discriminative features are selected using ordinal regression techniques. Other approaches [9], [46] face this problem by grouping documents with closer stars under the same category, i.e. positive or negative, simplifying the task into a binary classification problem.

However, in all above mentioned approaches the whole review is used to determine its rating, and the complexity of the rating-inference task is reduced in most of the cases by dealing only with three classes at most. Moreover, the existing works attempting to carry out a finer-granularity classification tackle the problem employing very short documents. This makes us think of the appropriateness of using summaries instead of full documents for the rating-inference task, when dealing with long documents and more than three classes. Text summaries are very useful for selecting and extracting the most important information of a document, thus avoiding noisy and unimportant information of the full document. In addition, they can vary in size, allowing us to analyze whether particular lengths can be more suitable than others.

3 TEXT SUMMARIZATION

Three types of summaries and four baselines are taken into consideration for the analysis carried out for the rating-inference task. The types of summaries we deal with are: i) generic; ii) query-focused; and iii) sentiment-based. Generic summaries are those which contain the main ideas of one or various documents; query-focused summarization generates summaries which contain the most important facts associated to a specific query, entity or topic; finally, sentiment-based summarization takes into account the subjective information within a document and produces a summary based on this information. This could be either positive information, negative, or a combination of both. For each of these summarization types, we propose the same gradual combination of several techniques which have been proven to be appropriate for extractive summarization[6] [26]. These techniques

[6] Extractive summarization is a particular kind of summarization, where the most relevant sentences of one or more documents are grouped forming the final summary, after they have been identified and extracted.



The Step account from NatWest ~ What and why? ~ The Step account from NatWest is placed in the category of being a basic bank account . These have sprung up over recent years, mainly because of the government trying to push people who might get benefits of any kind into receiving them via a bank account rather than the post office. They are also often used for people with poor credit ratings for whatever reason. However, in my case the reason I looked at this account is because I'm considered by many of my friends of being a bit of an oddity. I refuse to run up credit if I can help it (Mortgage excepted) and so I don't use an overdraft and my credit card is the smallest amount I could find (just a few hundred pounds) and only used where no other card is good enough. So, when Barclay 's bank finally managed to annoy me enough after being there for nearly 30 years since leaving school, I began to look around for alternatives. The NatWest were the first place I looked because of their promise that you can speak to a local branch if needed via telephone and since my arthritis now makes walking difficult, this was an important consideration for me (and one reason Barclay's have managed to annoy me so much but that's a separate issue). So, on looking

```
43869021503, sent_tf_idf=0.015941543869021503, string=was, swn_val=NA, token=15.0, token_idf=0.0159415438690215
ng=an, swn_val=N, token=11.0, token_idf=0.0, token_tf_idf=0.0}
936, sent_tf_idf=2.115477217419936, string=important, swn_val=P, token=2.0, token_idf=2.115477217419936, token_tf_idf=
947754, sent_tf_idf=3.92283213947754, string=consideration, swn_val=P, token=1.0, token_idf=3.92283213947754, token
0, token_tf_idf=0.0}
```

484 Annotations (20 selected) Select: swn_val=P

Figure 1. Linguistic analysis of a summary using GATE

rely on statistical and linguistic knowledge, as well as on different levels of language analysis (lexical, syntactic and semantic). In particular, the suggested techniques are word frequency, the code quantity principle, and textual entailment. Next, we are going to describe each technique briefly.

- **Term frequency (TF)**: following Luhn's idea [28], it is assumed that the most frequent words in a document are indicative of its topic and may contain relevant information. Consequently, sentences with frequent words will have more chance to belong to the final summary.
- **Code quantity principle (CQP)**: this is a linguistic principle which proves the existence of a proportional relation between how important the information is, and the number of coding elements it has [16]. In particular, it states that "more important information will be given more coding material" (i.e. syllables, words, or phrases). Since noun-phrases within sentences can vary in size depending on the level of detail a piece of information wants to be stressed, we assume that they can be good candidates to represent the "coding material" in Givon's principle. Consequently, sentences containing longer noun-phrases will be more important and will be selected for the final summary.
- **Textual entailment (TE)**: the TE module presented in [15] is used to detect redundant information. This TE system performs well according to the state-of-the-art (around a 63% in the last RTE[7] track), and as a consequence it is appropriate for its integration within a text summarization approach. The identification of these entailment relations helps to avoid incorporating redundant information in summaries as it has been previously proven in [25]. The way redundant information is detected is by computing the entailment between two consecutive sentences and discarding the entailed ones.

These techniques are combined together in the following way: the word-frequency (*tf*) is always taken into account, and then the remaining techniques are coupled with it, having two more approaches (textual entailment+word frequency –*te+tf*– and the code quantity principle+word frequency –*cqp+tf*–). Finally, the combination of all techniques is analyzed as well (*te+cqp+tf*). These initial approaches represent the generic summarization approach. For the query-focused summarization approach, we select those sentences in the review about the topic (name of a bank) by means of different patterns manually identified before applying the techniques previously mentioned. The main limitation of the manual generated patterns is that we do not deal with coreference, and some sentences which include pronouns (e.g."it", "this", etc) are not taken into account, affecting the performance of the summarizer. We plan to overcome this limitation in the future. Finally, sentiment-based summarization first identifies the polarity of each sentence (positive, negative or neutral) using a similar approach to [4], and then from the set of positive and negative ones, the most relevant sentences according to the proposed techniques are extracted.

In addition, the baselines suggested comprise a *lead* and a *final* baseline, in which a summary is produced by extracting the first or the last sentences of a document, respectively; and besides them, a query-focused (*qf*) and sentiment-based baseline (*sent*) are also provided. The former relies on the SUMMA toolkit [41] to compute a query[8] similarity value (cosine similarity) for each sentence in the review, which serves to rank the sentences in a document, whereas the latter extracts the strongest opinionated sentences of a document as a summary.

4 THE RATING INFERENCE PROCESS

In order to predict the correct numerical rating of a review, we use SVMs. In particular, we rely on the implementation included in the General Architecture for Text Engineering (GATE) framework [8], since the feature extraction and conversion from documents to the

[7] http://www.nist.gov/tac/2009/RTE/

[8] The query is the name of the entity being reviewed (e.g. National Westminster).

machine learning implementation is completely transparent to the researcher. We also use GATE for carrying out a linguistic analysis of the textual input documents. A basic pre-processing comprising tokenization, part of speech tagging, and morphological analysis is performed to produce document annotations, which are later used to extract features for document representation. Moreover, sentiment features are also computed using SentiWordNet [13], a lexical resource in which each synset of WordNet [14] is associated with three numerical scores: *obj* (how objective the word is), *pos* (how positive the word is), and *neg* (how negative the word is). Taking this resource as a basis, we calculate an aggregated score of the general sentiment of a word with respect to the number of times a word appears in SentiWordNet as more positive than negative or vice-versa. It is worth remarking that it is not trivial to associate sentiments to words, due to the ambiguity that may appear in some contexts. For instance, the word “poor” could indicate either objectivity or subjectivity. It can be used to state a factual situation like “a poor neighbourhood” or in contrast, it can be also used to express the bad quality of a specific service, for example “this company offers a poor service”. In this context we give an opinion, and consequently a subjectivity nature is associated to the word. The way we compute the sentiment score associated to a word avoid us having to carry out any disambiguation process, which could also introduce some errors in the language analysis stage. Figure 1 depicts an example of the GATE framework where the information annotated for a summary is shown.

Once all the input documents have been annotated, different features are used to train the SVM classifier with 10-fold cross validation. Previous preliminary experiments showed that the best performing features for full documents are on the one hand, the root of each word, and on the other hand the combination of the root of each word together with its part of speech category and the SentiWordNet score [27]. For this reason, we also use these features for predicting a review’s rating using summaries as training and test set.

5 EXPERIMENTAL SETUP

This section describes all the experimental framework carried out. First of all, the dataset and some related properties will be described. Then, the set of experiments, the evaluation metric used, and the results obtained will be shown. Finally, a discussion of the results will be provided in order to allow us to draw some conclusions.

5.1 Dataset

The corpus employed consisted of 89 bank reviews gathered directly from the Internet. In particular, we used Ciao[9] website to collect reviews from several English banks, such as Abbey, Barclays or Halifax. Some statistical properties concerning the dataset are shown in Table 1. It is worth mentioning that the documents we deal with are long documents, since the length of the reviews is 2,603 words on average, and the longest review has 5,730 words in total.

Table 1. Corpus properties.

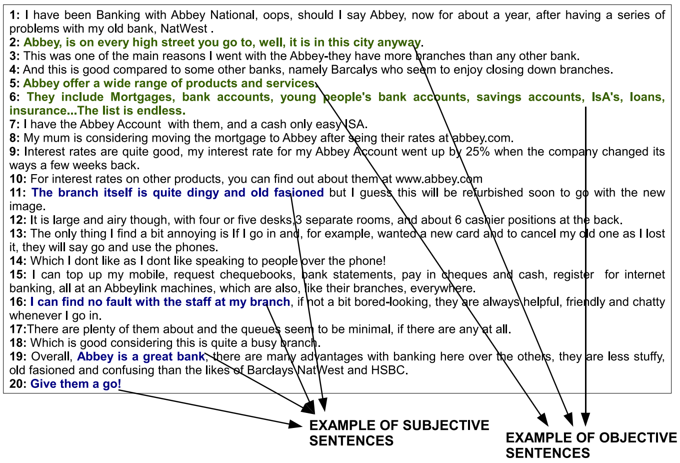
# Reviews	89
Avg length	2,603
Max length	5,730
Min Length	1,491

Moreover, Table 2 shows the star-class distribution of the reviews. As can be seen, the corpus is unbalanced, being the 4-star class the most predominant one in the dataset with a 32% of the total, whereas only the 10% of the reviews are rated with 3-stars.

Table 2. Class Distribution.

Star-rating	# reviews	%
1-star	17	19
2-star	11	12
3-star	9	10
4-star	28	32
5-star	24	27

In order to provide an example of the kind of documents we deal with, Figure 2 shows a fragment of a 5-star rating review (the maximum value in the scale). In particular, this fragment belongs to a review concerning the Abbey National bank. As can be seen, there are some objective sentences (e.g. sentence 2, 5 and 6), which express factual information about the bank (e.g. its location, its main products), but the review also contains opinionated sentences, for instance “overall, Abbey is a great bank” (sentence 19). Apart from this, sentences 11, 16, and 20, among others, also state subjective information. In this case, the user is giving his/her own opinion based on the experience he/she had with the bank, thus being directly related to the rating he/she will assign later. Therefore, it can be deduced that the analysis of this type of sentences is very important to be able to correctly predict the rating for any review.



1: I have been Banking with Abbey National, oops, should I say Abbey, now for about a year, after having a series of problems with my old bank, NatWest .
2: **Abbey, is on every high street you go to, well, it is in this city anyway**.
3: This was one of the main reasons I went with the Abbey-they have more branches than any other bank.
4: And this is good compared to some other banks, namely Barcalys who seem to enjoy closing down branches.
5: **Abbey offer a wide range of products and services.**
6: **They include Mortgages, bank accounts, young people's bank accounts, savings accounts, IsA's, loans, insurance...The list is endless.**
7: I have the Abbey Account with them, and a cash only easy ISA.
8: My mum is considering moving the mortgage to Abbey after seing their rates at abbey.com.
9: Interest rates are quite good, my interest rate for my Abbey Account went up by 25% when the company changed its ways a few weeks back.
10: For interest rates on other products, you can find out about them at www.abbey.com
11: **The branch itself is quite dingy and old fasioned** but I guess this will be refurbished soon to go with the new image.
12: It is large and airy though, with four or five desks,3 separate rooms, and about 6 cashier positions at the back.
13: The only thing I find a bit annoying is If I go in and, for example, wanted a new card and to cancel my old one as I lost it, they will say go and use the phones.
14: Which I dont like as I dont like speaking to people over the phone!
15: I can top up my mobile, request chequebooks, bank statements, pay in cheques and cash, register for internet banking, all at an Abbeylink machines, which are also, like their branches, everywhere.
16: **I can find no fault with the staff at my branch**, if not a bit bored-looking, they are always helpful, friendly and chatty whenever I go in.
17: There are plenty of them about and the queues seem to be minimal, if there are any at all.
18: Which is good considering this is quite a busy branch.
19: Overall, **Abbey is a great bank**, there are many advantages with banking here over the others, they are less stuffy, old fasioned and confusing than the likes of Barclays NatWest and HSBC.
20: **Give them a go!**

EXAMPLE OF SUBJECTIVE SENTENCES
EXAMPLE OF OBJECTIVE SENTENCES

Figure 2. Fragment of a 5-star review

5.2 Experiments

The main purpose of the paper is to analyze the usefulness of different types of summaries for the rating-inference problem. Therefore, each experiment consists of using each type of summarization approach described in Section 3 for training and testing the SVM classifier, and predict the star associated to a review. We deal with three types of summaries (generic, query-focused and sentiment-based), and for each type we use the following techniques (word frequency,

[9] http://www.ciao.co.uk

textual entailment, and the code quantity principle) resulting in four different combinations (*tf*; *te+tf*; *cqp+tf*; *te+cqp+tf*). In addition, we propose four baselines (*lead*, *final*, *qf* and *sent*), and summaries are generated according to different compression rates, which range from 10% to 50%. As a consequence, for each review we generate 80 different summaries, resulting in 7,120 summaries in total. Finally, each summary is trained and tested using 10-fold cross validation through two sets of features, as it was previously explained in Section 4. On the one hand, only the root of the words is used, while on the other hand, the root of the word is combined together with the part of speech category and the SentiWordNet score.

It is worth stressing upon the fact that due to the class distribution shown in Table 2, a totally uninformed baseline (4-star rating baseline) is also taken into account for predicting the rating of a review. This baseline consists of considering all documents as if they belonged to the most predominant class, i.e. 4-star rating.

Regarding the evaluation, we take into consideration the *Mean Squared Error* (MSE) [47], which is capable of capturing the deviation of the prediction from the true class label. The MSE can be defined as

$$MSE = \frac{\sum_{i=1}^{n} (Y_i - \widehat{Y_i})^2}{n}$$

where n is the total number of samples; Y_i is the true class label and $\widehat{Y_i}$ is the predicted class.

This measure has been previously used for text classification purposes [35] and this is also appropriate for our experiments because we deal with classification at a fine-grained granularity, and we are interested in analyzing which summarization approaches minimize the error in the star-rating prediction. As a consequence, if the original review's rating belonged to the 5-star class, and after using summaries for the prediction, it was rated as 4 stars, the MSE would be lower than if it was rated as 1 star. Other evaluation metrics such as F-measure or accuracy would not account for this fact, and they would have only focused on the number of reviews correctly rated (i.e. in the aforementioned example both results would have obtained the same performance). Furthermore, after all experiments were conducted, a paired t-test [47] was used to assess significance of the results at a 95% confidence interval. In the next subsection, the results obtained will be shown and discussed.

5.3 Results and Discussion

Table 3 and Table 4 show the results obtained for all experiments, using only the root feature for classification purposes and the combination of root, part of speech and SentiWordNet value, respectively. In both tables, the MSE obtained when using the whole reviews as text input, the baseline which assigns all reviews a 4-star rating, as well as the different types of summaries can be seen. Moreover, the results improving both the full document and the baseline are emphasized in bold. Concerning the statistical t-test we performed of the most relevant results obtained, we could not report any significance value with a 95% confidence interval in most of the cases. Only the approach *qf+cqp+tf* for a compression rate of 40% was statistically significant with respect to *qf* baseline of 30%, 40% and 50%, when employing the root as a single feature for the classification process. Nonetheless, this does not mean that we cannot analyze the results in depth in order to find possible tendencies about which summaries could be more suitable for the rating-inference task. From Table 3 it is worth noting that the word frequency is performing well in most of the cases. Moreover, it seems that query-focused summaries are also beneficial for the rating classification. However, the generation of sentiment-based summaries do not seem to be of great help when using only the root of a word as a feature for the classification process. This may happen due to the fact that the semantic information associated to the words is not being employed, relying only on a lexical level when using the root of the words. Similarly, the results shown in Table 4 also indicate a clear tendency that query-focused summarization is more appropriate than other strategies for the rating-inference task. However, it is important to stress upon the fact that when using also sentiment-based features (SentiWordNet) in the classification process, sentiment-based summarization is suitable as well.

Concerning the comparison of the proposed features for training and testing the SVM algorithm, the MSE results obtained for the *lead*, *qf* and *sent* baselines, as well as the sentiment-based summarization approaches, when using the root of the word, the part of speech and the SentiWordNet value, are better with respect to the same approaches when using only the root of words. However, this does not happen when query-focused summaries are combined with the proposed text summarization techniques, where the results are better in the latter case. For generic summarization, there is not an obvious tendency, and therefore for some techniques (*tf*, and *te+cqp+tf*) the results when taking also into account the part of speech and SentiWordNet are better, but for *cqp+tf* or *te+tf* they are not.

Another aspect that it is worth analyzing from the tables is whether it exists or not an optimal compression rate for summaries. If we have a look at both tables, and count the number of summarization approaches which perform better than the full review and the 4-star rating baseline for each compression rate, we can observe that 30% compression rate contains the highest number of better MSE results. This might mean that this compression rate could be an appropriate one, when deciding the length the summaries should have. However, it is important to remark that compression rates of 40% and 50% are also quite predominant, because the longer the summary is, the more relevant information it may contain. Therefore, an in-depth analysis needs to be performed out of the scope of this paper in order to draw definitive conclusions regarding the optimal summary length for the rating-inference task.

From the results, it can be seen that the summarization approach using the combination of techniques *te+cqp+tf* for a 10% compression rate obtains the best results with respect to the remaining combination techniques of the same compression rate in the three summarization strategies in turn (generic, query-focused, and sentiment-based), except for the query-focused strategy with word frequency alone, when predicting the rating with the root, the part-of-speech and the SentiWordNet value. This may mean that despite not having obtained the best results for the rating-inference task, the combination of these three features could be very appropriate for generating summaries of the 10% of the source documents independently or in conjunction with other NLP tasks.

Although we cannot claim that there exists a particular feature or strategy that contributes the most to the rating-inference problem, we can observe a general trend with respect to the summarization approaches that leads to the best results, obtaining lower MSE than when employing the full review. In this case, the most appropriate types of summarization strategies are query-focused and sentiment-based. Also, some specific types of generic summarization, such as *tf* or *te+tf* can also be adequate for this task.

Table 3. MSE results for summaries using only *root* as feature for rating classification (*lead* = first sentences; *final* = last sentences; *tf* = term frequency; *te* = textual entailment; *cqp* = code quantity principle with noun-phrases; *qf* = query-focused summaries; and *sent* = sentiment-based summaries).

Approach		**Compression Rate**				
Full document		**10%**	**20%**	**30%**	**40%**	**50%**
Full review	MSE	2.59	2.59	2.59	2.59	2.59
4-star rating baseline	MSE	2.58	2.58	2.58	2.58	2.58
Summarization method						
lead	MSE	3.10	3.00	3.10	3.30	3.10
final	MSE	2.74	3.00	**2.13**	2.64	**2.48**
qf	MSE	**2.49**	2.70	3.58	3.78	3.88
sent	MSE	3.89	3.16	3.03	2.90	2.66
generic-tf	MSE	3.21	**2.33**	**2.39**	**2.37**	**2.44**
generic-te+tf	MSE	3.39	3.23	**2.52**	**2.38**	**2.29**
generic-cqp+tf	MSE	3.01	3.34	2.61	3.17	3.03
generic-te+cqp+tf	MSE	2.70	2.93	3.00	3.10	2.71
qf-tf	MSE	**2.11**	**2.19**	**2.18**	**2.46**	**2.37**
qf-te+tf	MSE	**2.44**	**2.08**	**2.30**	**2.27**	**2.42**
qf-cqp+tf	MSE	2.83	3.00	2.70	**1.80**	**2.00**
qf-te+cqp+tf	MSE	**2.11**	**2.51**	**2.28**	**2.40**	**2.10**
sent-tf	MSE	2.83	**2.16**	**2.47**	**2.43**	**2.29**
sent-te+tf	MSE	3.20	2.80	**2.40**	2.69	2.71
sent-cqp+tf	MSE	3.01	3.27	2.62	3.21	3.10
sent-te+cqp+tf	MSE	2.69	3.21	3.46	2.90	2.93

Table 4. MSE results for summaries using *root*, *category* and *SentiWordNet* as features for rating classification (*lead* = first sentences; *final* = last sentences; *tf* = term frequency; *te* = textual entailment; *cqp* = code quantity principle with noun-phrases; *qf* = query-focused summaries; and *sent* = opinion-oriented summaries).

Approach		**Compression Rate**				
Full document		**10%**	**20%**	**30%**	**40%**	**50%**
Full review	MSE	2.59	2.59	2.59	2.59	2.59
4-star rating baseline	MSE	2.58	2.58	2.58	2.58	2.58
Summarization method						
lead	MSE	2.63	2.62	2.72	2.86	2.60
final	MSE	3.20	2.66	**2.16**	**2.13**	**2.50**
qf	MSE	**2.45**	**2.53**	2.90	2.91	2.81
sent	MSE	3.69	2.87	**2.44**	**2.34**	**2.27**
generic-tf	MSE	3.21	2.68	**2.19**	**2.29**	**2.33**
generic-te+tf	MSE	3.98	2.90	2.79	3.06	2.64
generic-cqp+tf	MSE	3.37	3.45	3.31	3.19	2.84
generic-te+cqp+tf	MSE	3.02	2.87	2.83	3.18	**2.53**
qf-tf	MSE	**2.44**	**2.56**	**2.37**	2.77	**2.50**
qf-te+tf	MSE	3.21	**2.52**	**2.26**	2.62	2.72
qf-cqp+tf	MSE	2.73	3.48	**2.51**	**2.48**	**2.27**
qf-te+cqp+tf	MSE	2.60	2.83	**2.52**	**2.44**	**2.53**
sent-tf	MSE	3.02	**2.47**	**2.16**	**2.12**	**2.31**
sent-te+tf	MSE	2.99	3.07	**2.21**	2.93	2.64
sent-cqp+tf	MSE	3.54	2.81	**2.51**	**2.43**	2.70
sent-te+cqp+tf	MSE	2.77	**2.31**	**2.14**	2.82	**2.47**

6 CONCLUSION AND FUTURE WORK

In this paper we proposed an extensive analysis of a wide range of summarization strategies of different compression rates for the rating-inference task, which can be considered as a particular problem within the sentiment analysis research area. Therefore, we generated summaries of different types (generic, query-focused and sentiment-based) and we analyzed several techniques (word frequency, textual entailment and the code quantity principle) within each type. Moreover, four summarization baselines were also produced in order to provide a more detailed analysis. These baselines were: *lead* baseline, which selects the first sentence of a document up to a desired length; *final* baseline, where the last sentences of a document are extracted; a query-focused baseline (*qf*), selecting the most similar sentences to a given query, and a sentiment-based baseline (*sent*), accounting only for the most opinionated sentences. For all these summarization approaches five compression rates (10%-50%) were also tested. As far as the rating-inference process is concerned, we approached the task employing a SVM paradigm within the GATE framework. We also proposed two types of features for document representation. On the one hand, the root of each word in the text was used, whereas on the other hand, the root of each word was also combined with the part of speech category and a sentiment scored, that we computed using the SentiWordNet lexical resource. The corpus we dealt with consisted of 89 bank reviews that had been rated by users from 1 to 5 stars, according to their experiences with these banks. After carrying out the classification process, the results were evaluated using *Mean Squared Error*, which is very appropriate for such fine-grained class granularity. Moreover, in order to assess the summaries' performance, we compared them to the full review, and a naive baseline considering all the reviews as a 4-star rating, which was the most frequent rating assigned by users. The results obtained showed that, although there is a variability in the summaries' performance, there seem to be a clear tendency for query-focused and sentiment-based summaries, indicating that these types are more appropriate for the rating-inference task than other types. However, we could not report strong evidence on this, and consequently, several things have to be further explored.

On the one hand, in the short-term we would like to extend the size of the dataset, in order to investigate whether the rating performance obtained using summaries could be affected by such small dataset. We also are interested in analyzing the difficulty of the fine-grained classification in depth. Consequently, we would like to study if a simplification of the problem, for instance with a scale of 3 rating values, could result in an increase of performance by minimizing the error obtained. On the other hand, in the long-term we would like to replicate the experiments but within a different domain, for example movie reviews. This would allow us to compare our results with already existing ones, and assess the performance of summarization in the rating-inference task.

ACKNOWLEDGEMENTS

This research has been supported by the project PROMETEO "Desarrollo de Técnicas Inteligentes e Interactivas de Minería de Textos" (2009/119) and project ACOMP/2010/286 from the Valencian Government, as well as by the Spanish Government (grant no. TIN2009-13391-C04-01). Moreover, Horacio Saggion is supported by a Ramón y Cajal Fellowship and Elena Lloret is funded by the FPI program (BES-2007-16268) under the project TEXT-MESS (TIN2006-15265-C06-01), both from the Spanish Ministry of Science and Innovation. The authors would also like to thank Alexandra Balahur for helping to process the dataset with her Opinion Mining approach.

REFERENCES

[1] Apoorv Agarwal, Fadi Biadsy, and Kathleen R. Mckeown, 'Contextual Phrase-level Polarity Analysis Using Lexical Affect Scoring and Syntactic N-grams', in *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 24–32, (2009).

[2] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani, 'Multi-facet Rating of Product Reviews', in *ECIR '09: Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval*, pp. 461–472, (2009).

[3] Alexandra Balahur, Elena Lloret, Ester Boldrini, Andres Montoyo, Manuel Palomar, and Patricio Martinez-Barco, 'Summarizing threads in blogs using opinion polarity', in *Proceedings of the International Workshop on Events in Emerging Text Types (eETTs)*, pp. 5 – 13, (2009).

[4] Alexandra Balahur-Dobrescu, Mijail Kabadjov, Josef Steinberger, Ralf Steinberger, and Andrés Montoyo, 'Summarizing Opinions in Blog Threads', in *Proceedings of the 23rd Pacific Asia Conference on Language, INformation and Computation (PACLIC)*, pp. 606–613, (2009).

[5] Regina Barzilay and Michael Elhadad, 'Using Lexical Chains for Text Summarization', in *Inderjeet Mani and Mark Maybury, editors*, Advances in Automatic Text Summarization, pp. 111–122. MIT Press, (1999).

[6] M. Biryukov, R. Angheluta, and M.-F. Moens, 'Multidocument Question Answering Text Summarization Using Topic Signatures', *Journal on Digital Information Management*, **3**(1), 27–33, (2005).

[7] Giuseppe Carenini and Jackie C. K. Cheung, 'Extractive vs. nlg-based abstractive summarization of evaluative text: The effect of corpus controversiality', in *Proceedings of the Fifth International Natural Language Generation Conference, ACL 2008*, pp. 33–40, Ohio, USA, (2008).

[8] Hamish Cunningham, D. Maynard, K. Bontcheva, and V. Tablan, 'GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications', in *Proceedings of the* 40^{th} *Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*, Philadelphia, USA, (Jul 2002).

[9] Ann Devitt and Khurshid Ahmad, 'Sentiment polarity identification in financial news: A cohesion-based approach', in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 984–991, (2007).

[10] Weifu Du and Songbo Tan, 'An Iterative Reinforcement Approach for Fine-Grained Opinion Mining', in *Proceedings of the North American Chapter of the ACL*, pp. 486–493, (2009).

[11] H. P. Edmundson, 'New Methods in Automatic Extracting', in *Inderjeet Mani and Mark Maybury, editors*, Advances in Automatic Text Summarization, pp. 23–42. MIT Press, (1969).

[12] Güneş Erkan and Dragomir R. Radev, 'LexRank: Graph-based Lexical Centrality as Salience in Text Summarization', *Journal of Artificial Intelligence Research*, **22**, 457–479, (2004).

[13] Andrea Esuli and Fabrizio Sebastiani, 'SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining', in *Proceedings of LREC-06, 5th Conference on Language Resources and Evaluation*, pp. 417–422, Genova, IT, (2006).

[14] C. Fellbaum, *WordNet: An Electronical Lexical Database*, The MIT Press, Cambridge, MA, 1998.

[15] Óscar Ferrández, Daniel Micol, Rafael Muñoz, and Manuel Palomar, 'A Perspective-Based Approach for Solving Textual Entailment Recognition', in *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pp. 66–71, (June 2007).

[16] Talmy Givón, *Syntax: A functional-typological Introduction, II*, John Benjamins, 1990.

[17] R. Grishman, 'Information Extraction: Techniques and Challenges', in *Information Extraction. A multidisciplinary approach to an Emerging Information Technology*, ed., Maria Teresa Pazienza, number 1299 in Lecture Notes in Artificial Intelligence, Springer, (1997).

[18] Cyril Grouin, Martine Hurault-Plantet, Patrick Paroubek, and Jean-Baptiste Berthelin, 'Deft'07 : une campagne d'valuation en fouille d'opinion', in *Fouille de donnes d'opinion*, volume E-17, 1–24, Cpadus ditions, (2009).

[19] Yi Hu, Wenjie Li, and Qin Lu, 'Developing Evaluation Model of Topical Term for Document-Level Sentiment Classification', in *Proceedings of the 10th Pacific Rim International Conference on Artificial Intelligence (PRICAI)*, pp. 175–186, (2008).

[20] Julian Kupiec, Jan Pedersen, and Francine Chen, 'A Trainable Document Summarizer', in *SIGIR '95: Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 68–73, (1995).

[21] Cane W. K. Leung, Stephen C. F. Chan, and Fu-Lai Chung, 'Integrating Collaborative Filtering and Sentiment Analysis: A Rating Inference Approach', in *Proceedings of The ECAI 2006 Workshop on Recommender Systems*, pp. 62–66, (2006).

[22] Cong Li, Ji-Rong Wen, and Hang Li, 'Text Classification Using

Stochastic Keyword Generation', in *Proceedings of the Twentieth International Machine Learning Conference (ICML 2003)*, pp. 464–471, (2003).

[23] C. Lin and E. Hovy, 'Identifying Topics by Position', in *Fifth Conference on Applied Natural Language Processing*, pp. 283–290. Association for Computational Linguistics, (31 March-3 April 1997).

[24] Elena Lloret, Alexandra Balahur, Manuel Palomar, and Andrés Montoyo, 'Towards Building a Competitive Opinion Summarization System: Challenges and Keys', in *Proceedings of the North American Chapter of the Association for Computational Linguistics. Student Research Workshop and Doctoral Consortium*, pp. 72–77, (2009).

[25] Elena Lloret, Óscar Ferrández, Rafael Muñoz, and Manuel Palomar, 'A Text Summarization Approach Under the Influence of Textual Entailment', in *Proceedings of the 5th Natural Language Processing and Cognitive Science Workshop*, pp. 22–31, (2008).

[26] Elena Lloret and Manuel Palomar, 'A Gradual Combination of Features for Building Automatic Summarisation Systems', in *Proceedings of the 12th International Conference on Text, Speech and Dialogue*, pp. 16–23, (2009).

[27] Elena Lloret, Horacio Saggion, and Manuel Palomar, 'Experiments on summary-based opinion classification', in *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pp. 107–115, Los Angeles, CA, (June 2010). Association for Computational Linguistics.

[28] H. P. Luhn, 'The Automatic Creation of Literature Abstracts', in *Advances in Automatic Text Summarization*, pp. 15–22. MIT Press, (1958).

[29] Inderjeet Mani, *Automatic Text Summarization*, John Benjamins Publishing Company, 2001.

[30] Inderjeet Mani, 'Summarization Evaluation: an Overview', in *Proceedings of the North American chapter of the Association for Computational Linguistics (NAACL). Workshop on Automatic Summarization*, (2001).

[31] W.C Mann and S.A. Thompson, 'Rhetorical Structure Theory: Toward a functional theory of text organization', *Text*, **8**(3), 243–281, (1988).

[32] Daniel C. Marcu, *The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts*, Ph.D. dissertation, Toronto, Ont., Canada, Canada, 1998. Adviser-Hirst, Graeme.

[33] Ryan McDonald, Kerry Hannan, Tyler Neylon, Mike Wells, and Jeff Reynar, 'Structured Models for Fine-to-Coarse Sentiment Analysis', in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 432–439, (2007).

[34] Rada Mihalcea, 'Graph-based Ranking Algorithms for Sentence Extraction, Applied to Text Summarization', in *Proceedings of the Association for Computational Lingusitics*, pp. 170–173, (2004).

[35] Rahman Mukras, Nirmalie Wiratunga, Robert Lothian, Sutanu Chakraborti, and David Harper, 'Information Gain Feature Selection for Ordinal Text Classification using Probability Re-distribution', in *Proceedings of the Textlink workshop at IJCAI-07*, (2007).

[36] Michael P. Oakes and Chris D. Paice, 'The Automatic Generation of Templates for Automatic Abstracting', in *21st BCS IRSG Colloquium on IR*, Glasgow, (1999).

[37] Bo Pang and Lillian Lee, 'Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales', in *Proceedings of the ACL*, pp. 115–124, (2005).

[38] Bo Pang and Lillian Lee, 'Opinion Mining and Sentiment Analysis', *Foundations and Trends in Information Retrieval*, **2**(1-2), 1–135, (2008).

[39] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan, 'Thumbs up? Sentiment Classification using Machine Learning Techniques', in *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 79–86, (2002).

[40] Dragomir R. Radev, Sasha Blair-Goldensohn, and Zhu Zhang, 'Experiments in Single and Multi-Document Summarization using MEAD', in *First Document Understanding Conference, New Orleans, LA*, pp. 1–7, (2001).

[41] H. Saggion, 'SUMMA: A Robust and Adaptable Summarization Tool', *Traitement Automatique des Languages*, **49**, 103–125, (2008).

[42] H. Saggion and A. Funk, 'Extracting Opinions and Facts for Business Intelligence', *RNTI*, **E-17**, 119–146, (2009).

[43] Horacio Saggion and Guy Lapalme, 'Generating Indicative-Informative Summaries with SumUM', *Computational Linguistics*, **28**(4), 497–526, (2002).

[44] Veselin Stoyanov and Claire Cardie, 'Toward Opinion Summarization: Linking the Sources', in *Proceedings of the Workshop on Sentiment and Subjectivity in Text*, pp. 9–14, (2006).

[45] Simone Teufel and Marc Moens, 'Summarizing Scientific Articles: Experiments with Relevance and Rhetorical Status', *Computational Linguistis*, **28**(4), 409–445, (2002).

[46] Peter D. Turney, 'Thumbs up or Thumbs down?: Semantic Orientation Applied to Unsupervised Classification of Reviews', in *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pp. 417–424, (2002).

[47] Timothy C. Urdan, *Statistics in Plain English*, Psychology Press, 2005.

[48] Theresa Wilson, Janyce 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*, pp. 347–354, (2005).

[49] Xiao-Peng Yang and Xiao-Rong Liu, 'Personalized Multi-document Summarization in Information Retrieval', *Machine Learning and Cybernetics, 2008 International Conference on*, **7**, 4108–4112, (July 2008).

The smallest, cheapest, and best: Superlatives in Opinion Mining

Silke Scheible[1]

Abstract. This paper introduces superlatives as special indicators for product features in customer reviews. The investigation shows that one type of superlative (called 'ISA') is of particular relevance, as instances in this class tend to contain both a feature string and its associated opinion word. An identification of the components of such superlative comparisons can therefore help to solve two Opinion Mining tasks at once: Feature and Opinion Word Identification. The study further introduces and evaluates a novel tool that can reliably identify such superlatives, and extract from them potential product feature strings and opinion words.

1 INTRODUCTION

In recent years, the domain of product reviews has attracted much attention in the area of Sentiment Analysis and Opinion Mining. While the main goal of the former is classification of documents, sentences, phrases or words as positive or negative, the interest in Opinion Mining lies in extracting information about which entities or features of entities are considered as positive or negative, and to summarise this information ([8]; [11]; [4]). This is of great benefit not only for companies who want information about customer's opinions on their products, but also for recommendation systems whose purpose is to assist customers in deciding which product to buy. In general, Opinion Mining systems are required to solve the following main tasks (e.g. [8]):

1. Feature Identification
2. Opinion Word Identification
3. Sentiment Classification
4. Opinion Summarisation

The first step is to identify features of the products that customers are interested in, usually by using data mining and natural language processing techniques. [8] define the term "product feature" as representing both components of an object (e.g. zoom) and their respective attributes (e.g. size).[2] The next step is to identify sentences in the reviews that express opinions about these features. This involves distinguishing opinion words from factual words (*subjectivity recognition*). To address (3), the system has to determine whether a statement of opinion is positive or negative. Finally, the system also requires techniques for summarising this information ([3]; [2]).

So far, none of the studies in Sentiment Analysis or Opinion Mining have specifically looked at the role of superlatives in these areas. While it has been generally acknowledged that there is a positive correlation between subjectivity and the use of *adjectives* (e.g. [7]), there has not yet been a thorough investigation of *superlative* adjectives and adverbs in this context. This paper aims to show that some types of superlative represent a special linguistic means of expressing opinions about products. Consider for example:

(1) The Panasonic TC-P54G10 is the *best* <u>plasma TV</u> on the market.
(2) It has the *clearest* <u>picture</u> I have ever seen.

I claim that superlative constructions like (1) and (2) act as special indicators of product features, which contain both the opinion word (the superlative, italicised) and the feature string (underlined). This means that the identification of the components of such superlative comparisons addresses two Opinion Mining tasks at once: Feature and Opinion Word Identification. This paper provides evidence for this claim, and introduces a novel tool which can be used to reliably identify superlatives of interest and extract potential product feature strings from them.

2 PREVIOUS APPROACHES

Existing work on identifying product features (Task 1) often relies on the simple heuristic that explicit features are expressed as noun phrases. While this narrows down the set of product feature candidates, it is clear that not all noun phrases represent product features. Various approaches to further limiting this set have been proposed. The two most notable ones are [8] and [11].

[8] suggest that nouns or noun phrases that occur frequently in reviews for a particular product are likely to be features. To identify frequent features they use association mining, and then apply heuristic-guided pruning to further refine their results. They further assume that adjectives appearing in the same sentence as frequent features are opinion words, thereby solving Task 2 (however, at the cost of precision). In addition, retrieving nouns and noun phrases that co-occur with these opinion words in other sentences helps their system to identify so-called *infrequent* features, which are also of great interest [10].

[11], on the other hand, consider product features to be concepts that stand in particular semantic relationships with the product (for example, a camera may have "properties" size, weight, etc., while the lens, flash, etc. stand in a "part" relationship with the camera). Their strategy for identifying such features is to search for corresponding meronymy discriminators. This approach achieves better performance than the one employed by [8], but no sentiment analysis is carried out, and opinion words have to be identified in a second step.

Although a previous study by [9] investigated graded adjectives in the context of customer reviews, their study is not suitable for identifying product features. They investigate the topic of comparative sen-

[1] University of Manchester, UK, email: Silke.Scheible@manchester.ac.uk
[2] In this definition, the object itself is also a feature.

tence mining, whose goal is to identify sentences in evaluative texts on the web that express “an ordering relation between two sets of entities with respect to some common features”, and to extract comparative relations from the identified sentences. A follow-up study by [6] builds on these findings and aims to determine which of the extracted entities in a comparison are preferred by its author. However, as [9] apply their vector approach to *every* graded adjective in the corpus, this involves a large amount of cases which do not modify “product features” (as identified and annotated by [8] in the same corpus). As a consequence, their system is not suitable for the task of identifying product features. Furthermore, even though Jindal and Liu’s system aims to identify the components of superlative comparisons, a closer study showed that their approach does not distinguish between different types of superlatives, leading to incorrect analyses of superlative constructions [12]. The current study takes different superlative surface constructions into account, and suggests that a particular subclass of superlatives (namely, ’ISA superlatives’) is especially useful in identifying product features.

3 SUPERLATIVES IN OPINION MINING

In text books, superlatives are usually introduced alongside comparatives as special forms of adjectives or adverbs which are used to compare two or more things, as for example in:

(3) Bill is *taller* than Sue. [comparative]
(4) {Joe} is the *tallest* [boy at school]. [superlative]

Superlative constructions like (4) express a comparison between a target entity T (Joe; curly brackets) and its comparison set CS (the other boys at school; square brackets). An investigation of superlative forms showed that two types of relation hold between a superlative target and its comparison set [12]:

Relation 1: Superlative relation
Relation 2: IS-A relation (hypernymy)

The superlative relation specifies a property which all members of the set share, but which the target has the highest (or lowest) degree or value of. The IS-A relation expresses the membership of the target in the comparison class (e.g. its parent class in a generalisation hierarchy). For example, in (4), the superlative relation implicitly specifies the property *height*, which applies to all members of the comparison set *boys at school*. Of this set, the target *Joe* has the greatest *height* value. The IS-A relation states that *Joe* is a member of the set *boys at school*. Both relations are of great interest for relation extraction, and [14] discusses their use in applications such as Question Answering (QA) and Ontology Learning.

Superlatives occur in a variety of syntactic structures which usually represent different types of comparisons. [14] developed a classification of superlatives based on surface forms (illustrated in Table 1). Superlatives belonging to the ISA class are incorporated in a definite NP and contain a clear-cut comparison between a target item and its comparison set. In example (a) in Table 1, the Panasonic TC-P54G10 is compared to other plasma TVs on the market with respect to its overall quality. The difference between the ISA-1 and ISA-2 subclasses lies in the way in which the relation between target and comparison set is expressed. In the case of ISA-1 superlatives, the verb “to be” or appositive form is used, while ISA-2 superlatives involve other forms (e.g. other copula verbs). While superlatives classified as DEF are also incorporated in a definite NP, they differ from members of the ISA class in that the target of comparison is not independently specified in the context. In example (c) the comparison remains implicit as the target is not specified in the sentence, except as that which satisfies the superlative NP. When superlative forms are incorporated in an indefinite NP they are classified as INDEF (d). Members of this class are often used as intensifiers. In the FREE class, on the other hand, superlative forms are not incorporated in a noun phrase but occur freely in the sentence. This often makes the comparison less easy to pinpoint: (e) does not compare the 37” size with other screen sizes, but rather the quality of the 37” size viewed from different locations in the room. Superlatives that are derived from adverbs form their own class, ADV (f). Finally, the IDIOM, PP, and PROP classes contain superlatives which do not express proper comparisons: IDIOM contains superlatives that occur as part of an idiom (g), PP contains so-called PP superlative constructions (h), and PROP includes uses of *most* as a proportional quantifier (i).

This study argues that superlatives of the type ISA are of particular importance in Opinion Mining as they make explicit the IS-A relation that holds between target and comparison set (cf. Relation 2 above). This means that both their target and comparison set are explicitly realised in the text, where the target string often expresses the product, the CS string expresses a feature while the superlative itself expresses the opinion word (as in (a) and (b)). The present study rests on the following claims:

1. ISA superlatives are special indicators for sentences containing product features.
2. The product feature usually appears within their T or CS string, while the superlative expresses its respective opinion word.

The next section briefly describes the data used to support these claims.

4 DATA

The investigation described in this paper uses Hu and Liu’s corpus of customer reviews, which was not only the basis of their own study of opinion feature mining [8], but has been used as test set by other studies as well (e.g. [11]). The corpus contains reviews of five products: two digital cameras (Canon G3 and Nikon Coolpix 4300), one mobile phone (Nokia 6610), an mp3 player (Creative Labs Nomad Jukebox Zen Xtra 40GB), and a dvd player (Apex AD2600 Progressive-scan)[3]. Sentences in this corpus have been manually annotated with information about product features. Each feature is taken to express an opinion, and labelled as *positive* or *negative* in terms of values on a six-point scale, where [+3] and [+1] stand for the strongest positive and weakest positive opinions, respectively, and [-3] and [-1] stand for the strongest and weakest negative opinions.

Hu and Liu’s corpus contains 4259 sentences altogether, of which 1728 include at least one product feature (40.6%). The remaining sentences in the corpus either contain no product feature (2217 altogether, 52.1%), or describe a review title, in which case they have been excluded from consideration (314 instances, 7.4%). The corpus contains a total of 230 superlatives in 4259 sentences, which means that there is around one superlative in every 18 sentences. All 230 superlatives found in the corpus were annotated with class labels as shown in Table 1.

[3] http://www.cs.uic.edu/ liub/FBS/CustomerReviewData.zip

Table 1. Superlative classes

Example	Class	Example
(a)	ISA	ISA-1: The Panasonic TC-P54G10 is the *best* plasma TV on the market.
(b)		ISA-2: The Samsung is considered the *most stylish* plasma TV.
(c)	DEF	I bought the *cheapest* plasma TV.
(d)	INDEF	Plasma TVs represent a *most compelling* option for home entertainment.
(e)	FREE	The 37” size is *best* when you are 8-10 feet away from the screen.
(f)	ADV	HD TVs *most commonly* use progressive scan for 1280x720.
(g)	IDIOM	The 42PC1RR won the *Best* Plasma TV Award this year.
(h)	PP	The TV weighs about 57 pounds at *most*.
(i)	PROP	*Most* cheap TVs have poor quality scalers.

5 ISA-SUPERLATIVES AS PRODUCT FEATURE INDICATORS

This section aims to provide support for the claim that superlatives are special indicators of product features in customer reviews. In particular, I will show that this especially applies to a subgroup of superlatives (ISA) by analysing the distribution of feature labels across the eight superlative classes in Hu and Liu’s corpus of customer reviews.

Table 2 shows the overall distribution of superlative classes in the corpus (columns 1 and 2). The ISA class is the most frequent with 71 instances (30.9%) (of which ISA-1 accounts for 63 instances, and ISA-2 for 8). The table further shows the proportion of title sentences (T), feature-containing sentences (F), and non-feature containing sentences (N) among the 230 superlative-containing sentences (S) in the corpus. The last row (TOTAL) indicates that the proportion of feature-containing sentences among them is higher (at 51.7%) than the average for all sentences (which is 40.6%, cf. Section 4). What is especially striking is that features are particularly highly represented among sentences containing ISA superlatives: Of 71 ISA superlatives in the data set, 53 occur in a sentence involving a feature (74.6%). This suggests that membership in the ISA class is a good indicator of the sentence containing a product feature.

Table 2. Distribution of features

Class	#S	#T	#F	#N
ISA	71	2	53	16
DEF	45	9	16	20
INDEF	15	10	3	2
FREE	3	1	2	0
ADV	10	0	4	6
IDIOM	12	0	5	7
PP	27	1	13	13
PROP	47	0	23	24
TOTAL	230	23	119	88
	100%	*10%*	*51.7%*	*38.3%*

A closer investigation of the data reveals further interesting results. Among the 119 superlative sentences that contain a feature (column “F”), not all superlatives directly contribute to the evaluation of the feature. For example, the superlative “most” in (5), which belongs to the PROP class, is not directly involved in the evaluation of the feature “firewire” as [-1]. In contrast, the ISA superlative “best” in (6) is directly responsible for the positive [+3] rating of the feature “dvd player”.

(5) it does n’t have firewire , not a real complaint since *most* windows users do n’t generally have firewire cards themselves . [Creative]

(6) i think , {apex} is the *best* [dvd player you can get for the price] . [Apex]

An assessment of all feature-containing sentences with respect to the involvement of the superlative in the feature-rating shows that the IDIOM, PP, and PROP classes are of little relevance, while ISA-1 and ISA-2 clearly are, with the superlative form acting as opinion word evaluating the feature, or acting as intensifier of an opinion word, as for example “complaint” in (7).

(7) [my] *biggest* [complaint] is {the battery life or lack there of} . [Creative]

Furthermore, in 34 out of the 46 feature-containing ISA-1 instances (73.9%) and in 6 out of 7 ISA-2 instances (85.7%), the feature is a substring of either the target (as shown in (7)) or the comparison set spans (6).

The importance of the ISA class is further supported by an investigation which showed that Hu and Liu’s annotation is not always consistent. Several of the 16 ISA-1 instances that did *not* receive a feature label in Hu and Lu’s annotation (column “N” in Table 2) do in fact modify a feature. For example, (8) and (9) make a similar positive statement about a camera, however only (8) was annotated with a feature (player[+3]). To be consistent, (9) should receive the same feature label. Example (10), on the other hand, is similar to (7) in that the superlative intensifies a negative evaluation (*drawback*, vs. *complaint* in (7)) of a feature (*software*, vs. *battery life*), however only (7) received a feature label (battery life[-3]). Given the structural and semantic similarities of the examples, one could clearly argue for adding a feature label “software[-3]” to (10).

(8) compared to everything else in this category , {this} is most definately [the] *best* [bang for the buck] . [Creative]

(9) i did a good month ’s worth of research before buying this over other similar priced digital cameras , and {this} is [the] *best* [buy for the buck] . [Canon]

(10) [the] *biggest* [drawback that people have about the zen xtra] is {the software} . [Creative]

The findings of this section corroborate the claim that ISA superlatives are special indicators of product features. Their identification could simultaneously help to solve Opinion Mining tasks 1 and 2 (see above) as they frequently contain a product feature within their T or CS string, and at the same time express its associated opinion word. As this strategy for finding product features does not depend on frequency (unlike Hu&Liu’s approach), ISA superlative identification also represents an efficient way of locating so-called *infrequent* features, which are also of great interest in Opinion Mining.

Table 3. List of abbreviations

Abbreviation	Description
CS	Comparison set of a superlative comparison
T	Target of a superlative comparison
SRE	Superlative Relation Extractor
SUP-Finder	Component of SRE used to identify superlatives in text
SUP-Classifier	Component of SRE used to classify superlative instances according to the surface forms described in Table 1
ISA1-Identifier	SUP-Classifier module used to identify ISA1-superlatives
T/CS-Identifier	Component of SRE used to identify the spans of the target and comparison sets of superlatives classified as ISA-1
CS-Identifier	Sub-component of T/CS-Identifier used to identify comparison set spans of ISA-1 superlatives
T-Identifier	Sub-component of T/CS-Identifier used to identify target spans of ISA-1 superlatives
CSDet	Determinative phrase of the superlative NP, e.g. *the* in *the best TV on the market*
CSHead	Head of the superlative NP, e.g. *TV* in *the best TV on the market*

6 AUTOMATIC IDENTIFICATION OF POTENTIAL PRODUCT FEATURES USING SUPERLATIVES

Having established a positive correlation between ISA superlatives and product features, the following sections describe how instances of this superlative type can be automatically identified and how potential product feature strings can be extracted from them, using Hu and Liu's corpus of customer reviews as data set. The tool used to achieve this is SRE ('Superlative Relation Extractor'), a novel system implemented in Python[4] which can be used to:

1) Identify superlatives in text;
2) Classify superlative instances according to the surface forms described in Table 1;
3) For superlatives classified as ISA-1, identify the spans of the target and comparison sets.

Initially, component 1) (called 'SUP-Finder') is used to find superlative instances in Hu and Liu's corpus of customer reviews. Next, the Classifier in 2) ('SUP-Classifier') is used to identify ISA-1 types among the retrieved superlatives, which are then input into component 3) ('T/CS-Identifier') to extract potential product feature strings (which have been shown to occur as substrings of the target or comparison set spans). Table 3 shows an overview of common abbreviations used in the following sections.

The SRE tool was originally developed on a corpus of Wikipedia texts (`TextWiki` corpus, [13]]). It employs a rule-based approach based mainly on tag sequences and dependency relations (using the output of the C&C tools, cf. [5]). SRE employs rules rather than machine learning due to the relatively small size of the gold-standard data set and the low frequency of some superlative types, which would represent a problem for a learner. An additional difficulty concerns the fact that the tools used to obtain the tags and dependency relations will have been optimised to correctly tag frequently occurring phenomena in its target text type, in order to achieve the highest possible performance score. As superlatives are relatively low frequency phenomena, with most types occurring far down the end of low frequency patterns (part of "the long tail"), even a relatively high-performance tagger like C&C may perform poorly at tagging them, because it will make little difference to the tagger's overall performance score. SRE's approach involves highly flexible and fine-tuned rules which can take these factors into account wherever necessary.

The following sections describe the three components of SRE and assess their suitability for the purpose of identifying potential product features in customer reviews. As SRE was originally developed on Wikipedia texts, its performance is expected to be affected by the non-standard nature of the data and the tagging/parsing errors that are likely to result from this.

6.1 Superlative detection

6.1.1 Method

As a first step, superlatives in the corpus are automatically identified using the SUP-Finder component of SRE. In general, superlatives are derived from their base adjective/adverb in two different ways: inflectionally or analytically. In the first case, the inflectional suffix *-est* is appended to the base form of the adjective or adverb (e.g. *largest*), while in the second case they are preceded by the analytical markers *most*/*least* (e.g. *most beautiful*). In addition, there is a (limited) number of irregular forms, such as *best*, *worst*, or *furthest*.

Previous automatic approaches to identifying superlatives have mainly focussed on techniques involving a search for the POS tags JJS and RBS (e.g. [1]), usually without carrying out a detailed error analysis due to the large amount of manual intervention that is required for a gold standard. The SUP-Finder tool aims to improve on the POS-based approach by using a pattern matcher based on regular expressions and a list of "superlative distractors" (i.e. a list of clear cases of non-superlatives, such as *nest*, *protest*, or *honest*), which are excluded from consideration. As superlatives form a well-defined class with a limited number of irregular forms, this pattern-based search works very well, and has been shown to outperform a POS-based approach by 2-3% with 99.0% precision and 99.8% recall[5] on Wikipedia texts [14].

6.1.2 Results and discussion

Unlike the POS-based approach, which has been optimised to work well on a particular text type, SUP-Finder is independent from text type and can be assumed to work equally well on customer review data. With its recall value nearing 100%, SUP-Finder was only assessed for precision in this study. The list of 231 superlatives returned by the tool was manually checked. Only one false positive was found, which had been missing from the list of "superlative distractors" (*hobbiest*, a mistyped version of *hobbyist*). The precision value is therefore 99.6% (230/231).

[4] SRE is freely available upon request (email the author of this paper at Silke.Scheible@manchester.ac.uk.)

[5] The only error affecting recall was due to incorrect tokenisation of quotes.

6.2 Identifying ISA superlatives

6.2.1 Method

The task of the second component of SRE, SUP-Classifier, is to classify superlatives as ISA-1, DEF, INDEF, etc. SUP-Classifier consists of a cascade of modules, each of which applies a set diagnostic tests to determine which class a given superlative instance belongs to. Here the focus is on the module that identifies an instance as belonging to ISA-1, called ISA1-Identifier.[6] This module requires substantial syntactic information, for example on whether the superlative form is bound in a definite NP, and if so, what the indices of the NP head and the determiner are. Furthermore, as the target of comparison needs to be explicitly mentioned in the sentence (cf. Section 3), the ISA1-Identifier component makes extensive use of the Grammatical Relations output of the C&C parser. Two main cases are distinguished: Instances where the IS-A relation between target and comparison set is expressed via the verb "to be", or via apposition. The strategy for the former case is as follows:

- Step 1: Locate the position of the comparison set head (`CSHead`) within the sentence
- Step 2: Test whether the relation word between the `CSHead` and its dependant is a form of "to be"
- Step 3: Find the corresponding target entity

If all three steps succeed, the instance is classified as ISA-1. The first step is addressed by testing whether the head of the superlative NP (`CSHead`) occurs in subject (`ncsubj`) or complement (`xcomp`) position, as for example in (11).

(11) The Panasonic is the best [TV]$_{\{CSHead\}}$.

The output of the C&C parser for this sentence is shown in Table 4. To fulfil Step 1, the Identifier first searches for a GR tuple where `CSHead` (here: *TV*) stands in an `xcomp` position (Row 4 in Table 4). Step 2 is then met by checking if the item in the second slot of this tuple is a form of "to be". If it is, Step 3 is addressed by searching the GR list for another tuple where the identified verb stands in an `ncsubj` relation with another word (the suspected target, cf. Row 5).

Table 4. GR output for "*The Panasonic is the best TV.*"

Row	GR output
1	`(det Panasonic_1 The_0)`
2	`(ncmod _ TV_5 best_4)`
3	`(det TV_5 the_3)`
4	`(xcomp _ is_2 TV_5)`
5	`(ncsubj is_2 Panasonic_1 _)`

6.2.2 Results

SUP-Classifier is tested on the output of SUP-Identifier, i.e. all superlative-containing sentences in Hu and Liu's corpus (230 altogether).[7] The results are displayed in Table 5.

The results show that SUP-Classifier clearly outperforms a random baseline system. With 94.6% precision and 85.5% recall, it can be reliably used to identify ISA-1 superlatives in customer reviews.

Table 5. Results of SUP-Classifier

Class	Precision	Recall	F-Measure
ISA-1	(53/56) **94.6%**	(53/62) **85.5%**	**89.8%**
Baseline	(33/115) **28.7%**	(33/62) **53.2%**	**37.3%**

6.2.3 Discussion

The non-standard nature of the data in customer reviews does not seem to have had the anticipated negative effect on the performance of the Classifier. Surprisingly, its performance is better on this text type than on the corpus of Wikipedia texts used in [14], where ISA-1 achieved 82.4% precision and 84.3% recall. A closer investigation of the gold-standard ISA-1 superlatives shows that this improvement is likely to be due to a simpler syntactic structure of ISA-1 cases in customer reviews, leading to better parser performance.

The C&C tool's inability to handle non-standard language mainly affected recall. For example, (12) was classified as INDEF because the system failed to identify "it 's" as erroneous variant of the possessive pronoun "its" (incorrectly tagged as personal pronoun, PRP, and 3rd person singular present tense verb, VBZ). Example (13) was not recognised by the parser because "about" is interpreted as preposition (IN) rather than as a preceding adverb (RB).

(12) i think this is it$_{\{PRP\}}$'s$_{\{VBZ\}}$ *biggest* flaw .

(13) if you do any research into digital cameras , you 'll quickly find that this camera is just$_{\{RB\}}$ about$_{\{IN\}}$ the *best* value out there.

6.3 Identification of potential product feature strings

6.3.1 Method

The third component of SRE, T/CS-Identifier, identifies potential feature-containing strings by extracting the target and comparison set strings of ISA-1 superlatives. The tool consists of two parts: a comparison set span identifier (CS-Identifier), and a target span identifier (T-Identifier). Their goal is to identify all relevant constituents of the T and CS phrases, which is a major challenge because both can have pre- and postmodifiers, the latter of which may be restrictive or non-restrictive [13]. To achieve maximum accuracy, T/CS-Identifier uses a fine-grained set of rules based on the lexical annotation output of the C&C tools. This approach was chosen as the GR output by the C&C parser proved to be unreliable due to the non-standard nature of the data. Similar problems are described by [1].

The present task assumes that both target (T) and comparison set (CS) comprise a single span. The CS span is defined as consisting of a determinative phrase (`CSDet`) and the main CS phrase (`CSMain`). To identify the determinative phrase, the tool uses a purely pattern-based approach (based on POS tags). The main CS span is determined by rules which aim to identify all pre- and postmodifiers of `CSHead` (cf. Section 6.2). Generally, tokens occurring between the superlative form and `CSHead` are included as premodifiers. Postmodifiers are identified using a set of patterns which were devised to match common types of superlative postmodifiers. Target identification involves locating the target in the sentence, and identifying all restrictive pre- and postmodifiers. The following sentences are examples of superlatives for which T/CS-Identifer is able to correctly identify the target (curly brackets) and comparison set spans (square brackets), with the product feature underlined.

[6] Due to the low frequency of ISA-2 types, I will restrict this investigation to ISA-1 types only.

[7] However, five of the 230 instances were excluded from evaluation as the C&C parser failed to parse them.

(14) i think , {apex} is [the] best [dvd player$_{\{+3\}}$ you can get for the price] .

(15) in my opinion [the] *worst* [issue on this phone] is {the side-mounted volume control$_{\{-3\}}$} .

6.3.2 *Results*

Table 6 shows the results of running T/CS-Identifier on the ISA-1 superlatives in Hu and Liu's data set. The baseline system assumes "the" as `CSDet`, and the first word following the superlative as the beginning of the `CSMain`, and the first word tagged as NN.* in that sequence as the end. The CS span is marked as correct only if both components `CSDet` and `CSMain` are exact matches with the gold standard. The baseline target identifier chooses the sequence of NP chunks closest to the superlative as target span.

Table 6. Performance of T/CS-Identifier (Accuracy)

Component	SRE	Baseline
CS-Identifier	62.9%	17.7%
- CSDet	98.4%	88.7%
- CSMain	64.5%	22.6%
T-Identifier	66.1%	37.1%

Both components clearly outperform their respective baselines.

6.3.3 *Discussion*

The majority of errors in the `CSMain` span were caused by the tagger/parser, in cases where a restrictive "bare" relative clause starting with the pronoun "i" follows the `CSHead`. In (16), the parser falsely interprets "i" as the NP head because of its non-standard spelling (which caused it to be tagged as plural noun NNS instead of personal pronoun PRP). A quick test confirmed this: Running the same sentence through the tagger with "I" capitalised resulted in the correct analysis.

(16) {this} is [the] *best* [dvd player i] 've purchased .

(17) {this} is [one of the] *nicest* [phones nokia] has made .

Similarly, in (17), the token "nokia" was tagged as common noun (NN) and not recognised as a new NP chunk ('B-NP') indicating the start of a relative or subordinate clause. In both cases, the CS span breaks off incorrectly (square brackets).

While CS-Identifier performs worse on customer reviews compared to its original domain (Wikipedia texts, where it achieved 88.8%), the situation is the reverse for T-Identifier, despite the non-standard nature of the data (66.1% vs. 58.4% in Wikipedia). This is largely due to shorter sentences and fewer appositions, which positively affect the target location methods. Furthermore, the target heads are often pronouns ("this", "it") or simple NPs such as "Apex" with no pre- nor postmodifiers (30 out of 62 instances), which do not represent a problem to T-Identifier.

The fact that a large proportion of targets are represented by pronouns immediately raises the question of pronoun resolution. However, a first investigation of the data suggests that the great majority of the pronouns "this" and 'it" refer to the entity under review.[8] With respect to the goal of the current investigation (i.e. identifying product features), pronouns in the target string do not represent a problem, as most product features occur in the comparison set string.

7 CONCLUSION AND FUTURE WORK

This paper established ISA-1 superlatives as special indicators of product features in Opinion Mining, which not only contain the feature strings (in most cases as part of the CS), but also the opinion word (usually the superlative itself), addressing two Opinion Mining tasks at once. Although superlatives are of relatively low frequency, the study supports previous findings that superlatives are perceived as interesting and important by people [12], and Section 5 highlights their importance in customer reviews. The study further introduced SRE as a tool to reliably identify ISA-1 superlatives automatically, and to extract from them potential product feature strings. As this strategy for finding product features does not depend on frequency, it represents an efficient way of locating *infrequent* features, which are also of great interest in Opinion Mining. SRE can be used as a stand-alone system for finding product features involving ISA comparisons, or it could be incorporated as an additional component in an existing Opinion Mining system.

Having automated the detection of ISA-1 superlatives and their components, the important final question is how these results can be used to arrive at the product features they are assumed to contain. As previously mentioned, the feature is a substring of either the target or the comparison set in 34 out of the 46 instances (73.9%). As the majority of them (27) occur as part of the comparison set, one strategy would be to assume that the product feature substring is the NP-chunk containing the `CSHead`. This simple approach would work for 25 of the 27 cases. Crucially, as most of the errors in automatically detecting the CS span were in recognising postmodification, product features can still be correctly identified as they only require identification of the `CSHead` chunk.

Finally, while this paper has focused on the role of ISA-1 superlatives in Opinion Mining, another interesting and potentially useful class is represented by DEF, illustrated by (18) and (19), which express positive statements about the features "image quality" and "lens adapter", respectively.

(18) overall , the g3 delivers what must be considered the *best* image quality of any current $>$ 4 megapixel digicams , from a detail , tonal balance and color response point of view .

(19) they got the *best* lens adapter for the g3-better than canon 's .

While the distribution of product features across the DEF class does not hint at their importance (cf. Table 2), one needs to consider that the DEF class is based on surface forms and contains a variety of different semantic types, of which only the so-called "relative set comparisons" type may be of interest. Future work will therefore involve finding techniques to distinguish this type from the other semantic types found in the DEF class.

ACKNOWLEDGEMENTS

I would like to thank Bonnie Webber for her helpful feedback on this paper.

REFERENCES

[1] Johan Bos and Malvina Nissim, 'An Empirical Approach to the Interpretation of Superlatives', in *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pp. 9–17, (2006).

[2] Giuseppe Carenini and Jackie Chi Kit Cheung, 'Extractive vs. NLG-based Abstractive Summarization of Evaluative Text: The Effect of Corpus Controversiality', in *International Conference on Natural Language Generation (INLG-08)*, pp. 33–41, (2008).

[8] This claim would however have to be verified by a thorough investigation of the context.

[3] Giuseppe Carenini, Raymond Ng, and Adam Pauls, ‘Multi-document Summarization of Evaluative Text’, in *Proceedings of 11th Conference of the European Chapter of the Association for Computational Linguistics, (EACL-06)*, pp. 305–312, (2006).

[4] Giuseppe Carenini, Raymond T. Ng, and Ed Zwart, ‘Extracting Knowledge from Evaluative Text’, in *K-CAP ’05: Proceedings of the 3rd International Conference on Knowledge Capture*, pp. 11–18, (2005).

[5] Stephen Clark and James R. Curran, ‘Parsing the WSJ using CCG and log-linear Models’, in *ACL ’04: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, (2004).

[6] Murthy Ganapathibhotla and Bing Liu, ‘Mining Opinions in Comparative Sentences’, in *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pp. 241–248, (2008).

[7] Vasileios Hatzivassiloglou and Janyce M. Wiebe, ‘Effects of Adjective Orientation and Gradability on Sentence Subjectivity’, in *Proceedings of the 18th Conference on Computational Linguistics*, pp. 299–305, (2000).

[8] Minqing Hu and Bing Liu, ‘Mining Opinion Features in Customer Reviews’, in *Proceedings of AAAI 2004*, pp. 755–760, (2004).

[9] Nitin Jindal and Bing Liu, ‘Mining Comparative Sentences and Relations’, in *Proceedings of AAAI 2006*, (2006).

[10] Bo Pang and Lillian Lee, ‘Opinion Mining and Sentiment Analysis’, *Foundations and Trends in Information Retrieval*, **2**(1-2), 1–135, (2008).

[11] Ana-Maria Popescu and Oren Etzioni, ‘Extracting Product Features and Opinions from Reviews’, in *HLT ’05: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pp. 339–346, (2005).

[12] Silke Scheible, ‘Towards a Computational Treatment of Superlatives’, in *Proceedings of the ACL 2007 Student Research Workshop, Prague, Czech Republic*, pp. 67–72, (2007).

[13] Silke Scheible, ‘Annotating Superlatives’, in *Proceedings of the Sixth International Language Resources and Evaluation Conference (LREC’08), Marrakech, Morocco*, pp. 923–928, (2008).

[14] Silke Scheible, *A Computational Treatment of Superlatives*, Ph.D. dissertation, University of Edinburgh, 2009.

Bootstrapping Supervised Machine-learning Polarity Classifiers with Rule-based Classification

Michael Wiegand and **Dietrich Klakow**[1]

Abstract. In this paper, we explore the effectiveness of bootstrapping supervised machine-learning polarity classifiers using the output of domain-independent rule-based classifiers. The benefit of this method is that no labeled training data are required. Still, this method allows to capture in-domain knowledge by training the supervised classifier on in-domain features, such as bag of words.

We investigate how important the quality of the rule-based classifier is and what features are useful for the supervised classifier. The former addresses the issue in how far relevant constructions for polarity classification, such as word sense disambiguation, negation modeling, or intensification, are important for this self-training approach. We not only compare how this method relates to conventional semi-supervised learning but also examine how it performs under more difficult settings in which classes are not balanced and mixed reviews are included in the dataset.

1 Introduction

Recent years have seen a growing interest in the automatic text analysis of opinionated content. One of the most popular subtasks in this area is polarity classification which is the task of distinguishing between positive utterances (Sentence 1) and negative utterances (Sentence 2).

1. The new iPhone looks *great* and is *easy* to handle.
2. London is *awful*; it's *crime-ridden*, *dirty* and full of *rude* people.

Various supervised classification approaches, in particular classifiers using bag of words, are heavily domain-dependent [2], i.e., they usually generalize fairly badly across different domains. Yet the costs to label data for any possible domain are prohibitively expensive.

Semi-supervised learning tries to solve this issue by reducing the size of the labeled dataset. The lack of labeled training data is compensated by a large unlabeled dataset of the target domain. The latter is much cheaper to obtain.

Rule-based classification does not require any labeled training data. In polarity classification, the rule-based classifier relies on domain-independent polar expressions. Polar expressions are words containing a prior polarity, such as *great* and *awful*. One typically counts the number of positive and negative polar expressions in a test instance and assigns it the polarity type with the majority of polar expressions. Since the classifier is restricted to domain-independent polar expressions, it lacks the knowledge to recognize domain-specific polar expressions, such as *crunchy*$^{+}$ in the food domain or *buggy*$^{-}$ in the computer domain.

In this paper, we explore the effectiveness of an alternative, which like most semi-supervised learning algorithms is based on *self-training*, i.e., the process of labeling the unlabeled data with a preliminary classifier and then training another (more robust) classifier by using the expanded annotated dataset. Unlike traditional semi-supervised learning, we do not use an initial classifier trained on a small labeled dataset but the output of a domain-independent rule-based classifier. (For reasons of simplicity, we will often refer to this specific version as plain *self-training* in the following sections.) While the rule-based classifier is restricted to the knowledge of (domain-independent) polar expressions, the supervised classifier trained on in-domain data labeled by the rule-based classifier can make use of domain-specific features, such as bag of words. Hopefully, the supervised classifier can effectively use this domain-specific knowledge and thus outperform the rule-based classifier.

Though this kind of self-training has already been applied to tasks in opinion mining, including polarity classification, there are certain aspects of this method which have not yet been fully examined:

Firstly, what are good features for the (pseudo-)supervised polarity classifier which is trained on the data labeled by the rule-based classifier? Do the insights hold from common supervised learning or semi-supervised learning?

Secondly, what is the impact of the robustness of the rule-based classifier on the final classifiers, i.e., does the supervised classifier improve when the rule-based classifier improves? This addresses the issue of in how far relevant constructions for polarity classification that can be incorporated into a rule-based classifier, such as word disambiguation, negation modeling, or intensification, are important for this kind of self-training approach.

Thirdly, how does this type of self-training compare to state-of-the-art semi-supervised learning algorithms?

Finally, does this method work in realistic settings in which – in addition to definite polar reviews – also mixed polar reviews are part of the dataset and the distribution of the classes is imbalanced?

2 Related Work

There has been much work on document-level polarity classification using supervised machine learning methods. Various classifiers and feature sets have been explored [10, 11]. Support Vector Machines (SVMs) [5] usually provide best results [11]. Unigram and bigram features outperform complex linguistic features [10].

Rule-based polarity classification usually requires an open-domain polarity lexicon with polar expressions. One typically counts the number of positive and negative polar expressions occurring in a test document and assigns it the polarity type with most polar expressions. This method can be enhanced by disambiguating polar expressions in their respective contexts. A framework in which scores are

[1] Spoken Language Systems, Saarland University, Germany, email: {michael.wiegand, dietrich.klakow}@lsv.uni.saarland.de

heuristically assigned to polar expressions depending on their individual contexts is proposed in [12]. The contextual modeling mainly focuses on *negation* and *intensification*. Implementations inspired by that formalism have been empirically proven effective [7, 8, 9].

Semi-supervised learning for polarity classification has been shown to be effective on inducing polarity lexicons from lexical resources [3, 14] but on text classification, the effectiveness is heavily dependent on the parameter settings. Significant improvement over supervised classification can usually only be achieved in presence of few labeled training data and a predictive feature set, such as in-domain adjectives or polar expressions from a polarity lexicon [17]. Another effective semi-supervised approach suggests to apply unsupervised learning (i.e., clustering) to classify unambiguous data instances and restrict manual annotation to hard data instances [4].

Bootstrapping supervised machine-learning classifiers with the help of rule-based classification has already been effectively applied to subjectivity detection of sentences [16]. The method has also been applied to polarity classification, but so far only on Chinese data [13, 15]. While the performance with out-of-domain supervised classifiers is compared in [15], this method is embedded into a complex bootstrapping system which also extends the vocabulary (or feature set) of the rule-based classifier in [13]. Neither of these works examine the impact of the rule-based classifier on the final result, the relation towards semi-supervised learning, nor discusses various settings of the self-training algorithm, in particular, different feature sets for the supervised classifier.

3 Data

In this paper, we use both the dataset of *IMDb* movie reviews [11] and reviews extracted from *Rate-It-All*[2]. We evaluate on the former because it is considered a benchmark dataset for polarity classification. The additional data are used to show that our findings are valid throughout different domains. Moreover, they have also been used in prior work on semi-supervised learning [17] which we also make use of in our experiments. Table 1 lists the properties of the corpora from the different domains. Note that on the *Rate-It-All* datasets we labeled 1 and 2 star reviews as *negative* and 4 and 5 star reviews as *positive*. 3 star reviews are labeled *mixed*. The actual class of these reviews is unknown. Usually a 3 star review should be neutral in the sense that it equally enumerates both positive and negative aspects about a certain topic, so that a definite verdict in favor or against it is not possible. That is also why we cannot assign these instances to either of the other two groups previously mentioned, i.e., *positive* and *negative*. During a manual inspection of some randomly chosen instances, however, we also found definite positive and negative reviews among 3 star reviews. For this work, we leave these instances in the category of mixed reviews.

4 Method

4.1 Rule-based Classifier

In the following, we describe how a polarity lexicon is converted to a rule-based polarity classifier. The polarity lexicon, the list of other important word classes being intensifiers, negation expressions (including the rules to disambiguate them) and polarity shifters are taken from the *MPQA* project [18].

4.1.1 Feature Extraction

Any word in a review that is not included in a polarity lexicon is discarded. Positive words (e.g., *excellent*) are assigned the value $+1$, negative words (e.g., *awful*) -1, respectively.

4.1.2 Basic Word Sense Disambiguation with Part-of-speech Tags

The polarity lexicon we use has part-of-speech tags attached to polar expressions in order to disambiguate them, e.g., the word *like* is either a polar verb or a preposition (in which case it is meaningless for polarity classification). We identify words as polar expressions only if their part-of-speech tags also match the specification in the lexicon. This can be considered as some basic form of word sense disambiguation. For part-of-speech tagging we use the *C&C* tagger[3].

4.1.3 Negation Modeling

If a polar expression occurs within the scope of a negation, its polarity is reversed (e.g., $[$*not* nice$^{+}]^{-}$). By scope, we define the five words immediately preceding the polar expression in the same sentence. Since some negation words are ambiguous and do not express negations when used in certain constructions, such as *not* in *not only . . . but also*, we also apply some rules disambiguating negation words.

In addition to common negation expressions, such as *not*, we also consider *polarity shifters*. Polarity shifters are weaker than ordinary negation expressions in the sense that they only reverse a particular polarity type. For example, the shifter *abate* only modifies negative polar expressions as in $[$*abate the damage*$^{-}]^{+}$.

4.1.4 Heuristic Weighting

So far, all polar expressions contained in the polarity lexicon are assigned the same absolute weight, i.e., $(\pm)1$. This does not reflect reality. Polar expressions differ in their individual polar intensity or, in case of ambiguous words, in their likelihood to convey polarity. Therefore, they should not obtain a uniform weight.

The polarity lexicon we use [18] includes a binary feature expressing the prior intensity of a polar expression. It distinguishes between *weak* polar expressions, such as *disordered*, and *strong* polar expressions, such as *chaotic*. Intuitively, strong polar expressions should obtain a higher weight than weak polar expressions.

When a polar expression is modifed by a so-called *intensifier*, such as *very* or *extremely*, its polar intensity is also increased. An ordinary weak polar expression has a similar polar intensity when it is modified by an intensifier as a strong polar expression, e.g., *extremely disordered* and *chaotic*.

The part of speech of a polar expression usually sheds light on the level of ambiguity of the word. If a polar expression is an *adjective*, its prior probability of being polar is much higher than the one of polar expressions with other parts of speech, such as verbs or nouns [11, 17]. Therefore, polar adjectives should obtain a larger weight than polar expressions with other parts of speech.

Since there are no development data in order to adjust the weights for the previously mentioned properties, we propose to simply *double* the value of a polar expression if either of these properties apply. If n of these properties apply for a polar expression, then its value is

[2] `http://www.rateitall.com`

[3] `http://svn.ask.it.usyd.edu.au/trac/candc`

Table 1. Properties of the different domain corpora (†only relates to the *Rate-It-All* data).

Domain	Source	Positive (4 & 5 Stars†)	Mixed (3 Stars†)	Negative (1 & 2 Stars†)	Vocabulary Size
computer	*Rate-It-All*	952	428	1253	15083
products	*Rate-It-All*	2292	554	1342	21975
sports	*Rate-It-All*	4975	725	1348	24811
travel	*Rate-It-All*	9397	1772	3289	38819
movies	*IMDb*	1000	0	1000	50920

doubled n times. For instance, an intensified adjective is assigned the value of 4, i.e., $2 \cdot 2$.

The properties considered for heuristic weighting have already been motivated and proven effective in previous work [7, 11].

4.1.5 Classification

For each data instance the *contextual* scores assigned to the individual polar expressions are summed. If the sum is positive, then the instance is classified as positive. It is classified as negative, if the sum is negative. We assign to all cases in which the sum is 0 the polarity type which gives best performance on that individual dataset (which is usually negative polarity). Thus, we have a stronger baseline that is to be beaten by self-training.

Note that the prediction score of a data instance, i.e., the sum of contextual scores of the polar expressions, can also be interpreted as a confidence score. This property is vital for effectively using this rule-based classifier in self-training. Thus, previously mentioned instances with a score of 0, for example, are unlikely to occur in the labeled training set since it only includes instances labeled with a high confidence score. The sum of contextual scores is normalized by the overall number of tokens in a test instance. This normalization additionally encodes the density of polar expressions within the instance. The greater the density of polar expressions of a particular type is in a text, the more likely the text conveys that polarity.

Figure 1 summarizes all steps of the rule-based classifier.

1. Lexicon loading, i.e., polar expressions, negation words, and intensifiers
2. Preprocessing:
   (i) Stem test instance.
   (ii) Apply part-of-speech tagging to test instance.
3. Polar expression marking:
   (i) Check whether part-of-speech tag of potential polar expression matches lexical entry (*basic word sense disambiguation*).
   (ii) Mark strong polar expressions.
4. Negation modeling:
   (i) Identify potential negation words (including polarity shifters).
   (ii) Disambiguate negation words.
   (iii) Reverse polarity of polar expression in scope of (genuine) negation.
5. Intensifier marking
6. Heuristic weighting: double weight in case polar expression is:
   (i) a strong polar expression
   (ii) an intensified polar expression
   (iii) a polar adjective.
7. Classification: assign test instance the polarity type with the largest (normalized) sum of scores.

Figure 1. Rule-based classifier.

4.1.6 Different Versions of Classifiers

We define four different types of rule-based classifiers. They differ in complexity. The simplest classifier, i.e., RB$_{Plain}$, does not contain word sense disambiguation, negation modeling or heuristic weighting. RB$_{bWSD}$ is like RB$_{Plain}$ but also contains basic word sense disambiguation. RB$_{Neg}$ is like RB$_{bWSD}$ but also contains negation modeling. The most complex classifier, i.e., RB$_{Weight}$, is precisely the algorithm presented in the previous sections. Table 2 summarizes the different classifiers with their respective properties.

4.2 Semi-Supervised Learning

Semi-supervised learning is a class of machine learning methods that makes use of both labeled and unlabeled data for training, usually a small set of labeled data and large set of unlabeled data. A classifier using unlabeled and labeled training data can produce better performance than a classifier trained on labeled data alone. This is usually achieved by harnessing correlations between features in labeled and unlabeled data instances and thus making inferences about the label of these unlabeled instances. Since labeled data are expensive to produce, semi-supervised learning is an inexpensive alternative to supervised learning.

In this paper, we exclusively use Spectral Graph Transduction (SGT) [6] as a semi-supervised algorithm since it produced consistently better results than other algorithms on polarity classification in previous work [17]. In SGT, all instances of a collection (i.e., labeled and unlabeled) are represented as a k nearest-neighbor graph. The graph is transformed to a lower-dimensional feature space, i.e., its spectrum, and then divided into two clusters by minimizing the graph cut. The two clusters that are chosen should preserve the highest possible connectivity of edges within the graph.

4.3 Self-Training a Polarity Classifier using the Output of a Rule-based Classifier

The idea of this bootstrapping method is that a domain-independent rule-based classifier is used to label an unlabeled dataset. Unlike in semi-supervised learning (Section 4.2), no labeled training data are used. The only available knowledge is encoded in the rule-based classifier. The data instances labeled by the rule-based classifier with a high confidence serve as labeled training data for a supervised machine-learning classifier. Ideally, the resulting supervised classifier is more robust on the domain on which it was trained than the rule-based classifier. The improvement can be explained by the fact that the rule-based classifier only comprises domain-independent knowledge. The supervised classifier, however, makes use of domain-specific features, i.e., words such as *crunchy*$^+$ (food domain) or *buggy*$^-$ (computer domain), which are not part of the rule-based classifier. It may also learn to correct polar expressions that are specified in the polarity lexicon but have a wrong polarity

Table 2. Properties of the different rule-based classifiers.

Properties	**RB**$_{Plain}$	**RB**$_{bWSD}$	**RB**$_{Neg}$	**RB**$_{Weight}$
basic word sense disambiguation		✓	✓	✓
negation modeling			✓	✓
heuristic weighting				✓

type on the target domain. A reason for a type mismatch may be that a polar expression is ambiguous and contains different polarity types throughout the different domains (and common polarity lexicons usually only specify one polarity type per entry). For instance, in the movie domain the polar expression *cheap* is predominantly negative, as it can be found in expressions, such as *cheap films*, *cheap special-effects* etc. In the computer domain, however, it is predominantly positive as it appears in expressions such as *cheap price*. If such a polar expression occurs in sufficient documents which the rule-based classifier has labeled correctly, then the supervised learner may learn the correct polarity type for this ambiguous expression on that domain despite the fact that the opposed type is specified in the polarity lexicon.

We argue that using a rule-based classifier is more worthwhile than using few labeled (in-domain) data instances – as it is the case in semi-supervised learning – since we thus exploit two different types of features in self-training being domain-independent polar expressions and domain-specific bag of words which are known to be complementary [1]. The traditional semi-supervised approach usually just comprises one homogeneous feature set.

Figure 2 illustrates both semi-supervised learning and self-training using a rule-based classifier for bootstrapping.

4.4 Feature Sets

Table 3 lists the different feature sets we examine for the supervised classifier (within self-training) and the semi-supervised classifiers. We list the feature sets along their abbreviation with which they will henceforth be addressed. The first three features (i.e., Top2000, Adj600, and MPQA) have been used in previous work on semi-supervised learning [17]. They all remove noise contained in the overall vocabulary of a domain corpus. The last two features (i.e., Uni and Uni+Bi) are known to be effective for supervised polarity classification [10]. Bigrams can be helpful in addition to unigrams since they take into account some context of polar expressions. Thus, crucial constructions, such as negation (*[not nice]*$^{-}$) or intensification (*[extremely nice]*$^{++}$), can be captured. Moreover, multiword polar expressions, such as *[low tax]*$^{+}$ or *[low grades]*$^{-}$, can be represented as individual features. Unfortunately, bigram features are also fairly sparse and contain a considerable amount of noise.

Table 3. Description of the different feature sets.

Feature Set	**Abbrev.**
the 2000 most frequent non-stopwords in the domain corpus	Top2000
the 600 most frequent adjectives and adverbs in the domain corpus	Adj600
all polar expressions within the polarity lexicon	MPQA
all unigrams in the domain corpus	Uni
all unigrams and bigrams in the domain corpus	Uni+Bi

5 Experiments

For the following experiments – with the exception of those presented in Section 5.4 – we mainly adhere to the settings of previous work [17]. We deliberately chose these settings in favor of semi-supervised learning in order to have a strong baseline for the proposed self-training method. We use a balanced subset (randomly generated) for each domain. The *Rate-It-All* dataset consists of 1800 data instances per domain, whereas the *IMDb* dataset consists of 2000 data instances. We just consider (definite) positive and (definite) negative reviews. The rule-based classifiers and the self-trained classifiers (bootstrapped with the help of rule-based classification) are evaluated on the entire domain dataset. The 1000 most highly-ranked data instances (i.e., 500 positive and 500 negative instances) are chosen as training data for the supervised classifier. This setting, which is similar to the one used for semi-supervised learning [17], provided good performance in our initial experiments. For the supervised classifier, we chose SVMs. As a toolkit, we use *SVMLight*[4]. Feature vectors were always normalized to unit length and additionally weighted with *tf-idf* scores. All words are stemmed. We report statistical significance on the basis of a paired t-test using 0.05 as the significance level.

5.1 Comparison of Different Rule-based Classifiers

Table 4 shows the results of the different rule-based classifiers across the different domains. On average, the more complex the rule-based classifier gets, the better it performs. The only notable exceptions are the *products* domain (from RB$_{Neg}$ to RB$_{Weight}$) and the *sports* domain (from RB$_{Plain}$ to RB$_{bWSD}$). On average (i.e., considering all domains), however, the improvements are statistically significant.

5.2 Self-Training with Different Rule-based Classifiers and Different Feature Sets

Table 5 compares self-training (SelfTr) using different rule-based classifiers and different feature sets for the embedded supervised classifier. In addition to accuracy, we also listed the F(1)-scores of the two different classes. The results are averaged over all domains. With the exception of RB$_{Neg}$ in combination with Top2000 and MPQA, there is always a significant improvement from a rule-based classifier to the corresponding self-trained version. If Top2000 or MPQA is used, there is a drop in performance from RB$_{Neg}$ to SelfTr in the *sports* domain. Improving a rule-based classifier also results in an improvement of the self-trained classifier. With exception of SelfTr(RB$_{Plain}$) to SelfTr(RB$_{bWSD}$) this is even significant.

The feature set producing the best results is Uni+Bi. Uni+Bi is statistically significantly better than Uni. This means that, as far as feature design is concerned, the supervised classifier within self-training behaves similar to ordinary supervised classification [10]. Unlike in semi-supervised learning [17], a noiseless feature set is not necessary. Best performance of SelfTr using a large set of polar

[4] `http://svmlight.joachims.org`

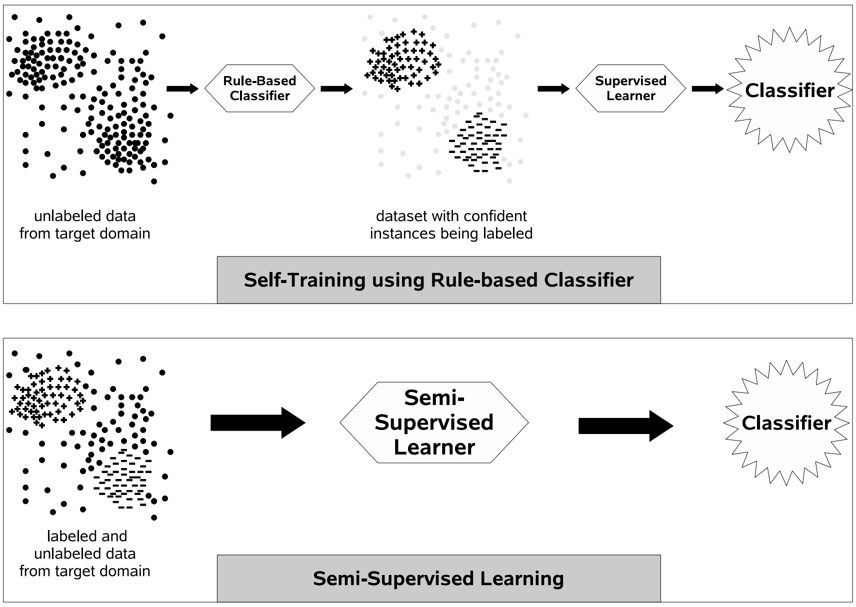

Rule-Based Classifier
Supervised Learner
Classifier
unlabeled data from target domain
dataset with confident instances being labeled
Self-Training using Rule-based Classifier
Semi-Supervised Learner
Classifier
labeled and unlabeled data from target domain
Semi-Supervised Learning

Figure 2. Comparison of semi-supervised learning and self-training using a rule-based classifier for bootstrapping.

Table 4. Comparison of accuracy between different rule-based classifiers (RB) and self-trained classifiers (SelfTr) trained on best feature set (Uni+Bi) on different domains (for each domain, performance is evaluated on a balanced corpus).

	RB$_{Plain}$		**RB**$_{bWSD}$		**RB**$_{Neg}$		**RB**$_{Weight}$	
Domain	**RB**	**SelfTr**	**RB**	**SelfTr**	**RB**	**SelfTr**	**RB**	**SelfTr**
computer	64.11	80.22	70.61	81.72	73.56	83.67	74.28	83.50
products	60.78	70.78	66.06	73.89	71.06	77.00	70.94	77.00
sports	64.33	66.44	64.39	64.94	67.50	68.89	68.89	72.78
travel	64.61	69.56	67.39	69.83	70.72	73.33	72.61	76.89
movies	61.75	72.70	64.80	72.45	67.85	73.55	71.30	77.75
average	63.12	71.94	66.65	72.57	70.14	75.29	71.60	77.58

expressions is reported in [13]. The feature set comprises an open-domain polarity lexicon and is automatically extended by domain-specific expressions. Our results suggest a less complex alternative. Using SelfTr with unigrams and bigrams (i.e., SelfTr$_{Uni+Bi}$) already provides better classifiers than SelfTr with a polarity lexicon (i.e., SelfTr$_{MPQA}$). The increase is approx. 3%.

It is also worth pointing out that the gain in performance that is achieved by improving a basic rule-based classifier (i.e., RB$_{Plain}$) by modeling constructions (i.e., RB$_{Weight}$) is the same as is gained by just self-training it with the best feature set (i.e., SelfTr$_{Uni+Bi}$).

The relation between the F-scores of the two different classes differs between RB and SelfTr. In RB, the score of the positive class is always significantly better than the score of the negative class. This is consistent with previous findings [1]. The gap between the two classes, however, varies depending on the complexity of the classifier. In RB$_{Plain}$, the gap is 17.45%, whereas it is less than 6% in RB$_{Neg}$ and RB$_{Weight}$. In SelfTr, the F-score of the negative class is usually better than the score of the positive class[5]. This relation between the two classes is typical of learning-based polarity classifiers [1]. However, it should also be pointed out that the gap is much smaller (usually not greater than 2%). Moreover, the size of the gap does not bear any relation to the gap in the original RB, i.e., though there is a considerable difference in size between the gaps of RB$_{Plain}$ and RB$_{Neg}$, the size of the gaps in the self-trained versions is fairly similar.

We also experimented with a combination of bag of words and the knowledge encoded in the rule-based classifier, i.e., the two features: the number of positive and negative polar expressions within a data instance. The performance of this combination is worse than a classifier trained on bag of words. The correlation between the two class labels and the two polarity features is disproportionately high since the polarity features essentially encode the prediction of the rule-based classifier. Consequently, the supervised classifiers develop a strong bias towards these two features and inappropriately downweight the bag-of-words features.

Table 4 compares rule-based classification and self-training on individual domains. In some domains self-training does not work. This is most evident in the *sports* domain using self-training on RB$_{bWSD}$.

[5] The only exception where the reverse is always true is SelfTr$_{MPQA}$. This does not come as a surprise since this feature set resembles RB most.

Apparently, the better the rule-based classifier is, the more likely a notable improvement by self-training can be obtained. Note that in the *sports* domain the self-trained classifier using the most complex rule-based classifier, i.e., SelfTr(RB$_{Weight}$), achieves the largest improvement compared to the rule-based classifier. These observations are also representative for the remaining feature sets examined but not displayed in Table 4.

5.3 Self-Training using Rule-based Classifiers Compared to Semi-Supervised Learning

In the following experiments, we use Spectral Graph Transduction (SGT) [6] as a semi-supervised classifier, since it provided best performance in previous work [17]. As a toolkit, we use *SGTLight*[6]. For each configuration (i.e., training and test partition) we randomly sample 20 partitions from the corpus. Labeled training and test data are always mutually exclusive but the test data (500 positive and 500 negative instances) can be identical to the unlabeled training data.

Figure 3 compares self-training bootstrapped on the output of rule-based classification (SelfTr) to supervised learning (SL) and semi-supervised learning (SSL). We compare two variations of SelfTr. SelfTr-A, as SSL, uses the same 1000 randomly sampled data instances for both unlabeled training and testing[7]. (Again, we report the averaged result over 20 samples.) SelfTr-B (like in previous sections) selects 1000 training instances by confidence from the entire dataset. The test data are, however, the same as in SelfTr-A. Unlike in previous work in which Top2000 is used for SL [17], we chose Uni+Bi as a feature set. It produces better results than Top2000 on classifiers trained on larger training sets (i.e., ≥ 400)[8]. For SSL, we consider Uni+Bi and Adj600, which is the feature set with the overall best performance using that learning method. For SelfTr, we consider the best classifier, i.e., SelfTr$_{Uni+Bi}$.

Though SSL gives a notable improvement on small labeled training sets (i.e., ≤ 100), it produces much worse performance than SL on large training sets (i.e., ≥ 200). Adjectives and adverbs are a very reliable predictor. However, the size of the feature set is fairly small. Too little structure can be learned on large labeled training sets using such a small feature set. Using larger (but also noisier) feature sets for SSL, such as Uni+Bi, improves performance on larger labeled training sets. However, even with Uni+Bi SSL does not reach a performance comparable to SL on large training sets and it is significantly worse than Adj600 on small training sets.

Whenever SSL outperforms SL, every variation of SelfTr also outperforms SSL. SelfTr-B is significantly better than SelfTr-A which means that the quality of labeled instances matters and SelfTr is able to select more meaningful data instances than are provided by random sampling. Unfortunately, SSL-methods, such as SGT, do not incorporate such a selection procedure for the unlabeled data. Further exploratory experiments using the *entire* dataset as unlabeled data for SSL produced, on average, results similar to those using 1000 instances. This proves that SSL cannot internally identify as meaningful data as SelfTr-B does. Whereas SSL significantly outperforms SL on training sets using less than 200 training instances, the best variation of SelfTr, i.e., SelfTr-B, significantly outperforms SL on training sets using less than 400 instances. This difference is, in particular, remarkable since SelfTr does not use any labeled training data at all whereas SSL does.

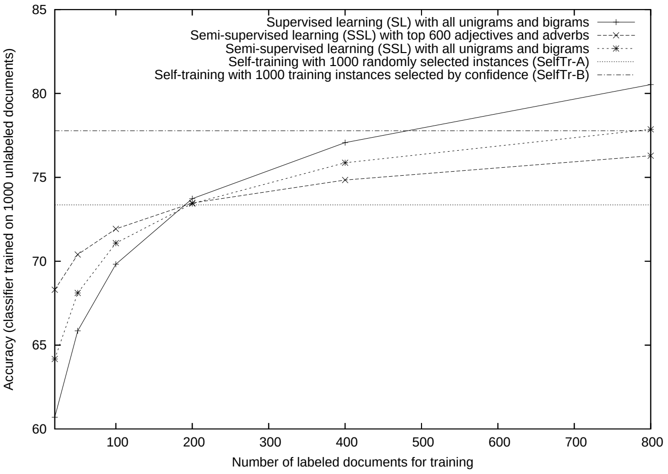


Figure 3. Comparison of self-training and semi-supervised learning (performance is evaluated on balanced corpus and results are averaged over all domains).

5.4 Natural Class Imbalance and Mixed Reviews

In this section, we want to investigate what impact natural class imbalance has on bootstrapping polarity classifiers with a rule-based classifier since this aspect has only been marginally covered in previous work [13, 15]. In those works, different class ratios on the test set are evaluated. However, the same amount of positive and negative reviews is always selected for training. We assume that the optimal performance of self-training can be achieved when the class distribution of training and test set is identical and we will provide evidence for that. Moreover, we want to explore what impact different distributions between the two sets have on the accuracy of the classifier and how different class-ratio estimation methods perform.

Previous work dealing with bootstrapping polarity classifiers using unlabeled data also focuses on datasets exclusively consisting of definite positive and negative reviews [4, 13, 15, 17]. In this section, the unlabeled dataset will also include mixed reviews, i.e., 3 star reviews (see Section 3). Due to the availability of such data the experiments are only carried out on the *Rate-It-All* data. We also add the constraint that the test data must be disjoint from the unlabeled training data[9].

Test data are exclusively (definite) positive reviews (i.e., 4 & 5 star reviews) and (definite) negative reviews (i.e., 1 & 2 star reviews). From each domain, we randomly sample 200 data instances 10 times. We state the results averaged over these different test sets. The class ratio on each test set corresponds to the distribution of definite polar reviews, i.e., 3 star reviews are ignored.

The unlabeled training dataset is the dataset of a domain excluding the test data. As labeled training data for the embedded supervised classifier within self-training, we use 70% of data instances labeled by the rule-based classifier ranked by confidence of prediction (across all domains/configurations, this size provided best results). Hopefully, most mixed reviews are among the remaining 30%.

[6] `http://sgt.joachims.org`

[7] We use this configuration since it is required by *SGTLight*.

[8] Note that previous work in particular focused on small training sets [17].

[9] We can include this restriction in this section since we will not consider the semi-supervised learning algorithm SGT in this section.

Table 5. Performance of self-trained classifiers with different feature sets (experiments are carried out on a balanced corpus and results are averaged over all domains).

	RB$_{Plain}$			**RB**$_{bWSD}$			**RB**$_{Neg}$			**RB**$_{Weight}$		
Type	**F1+**	**F1−**	**Acc**	**F1+**	**F1−**	**Acc**	**F1+**	**F1−**	**Acc**	**F1+**	**F1−**	**Acc**
RB (*Baseline*)	69.81	52.36	63.12	70.39	61.79	66.65	72.42	67.40	70.14	74.26	68.30	71.60
SelfTr$_{Top2000}$	70.15	70.88	70.53	70.26	71.55	70.92	72.78	73.88	73.40	74.79	74.18	75.73
SelfTr$_{Adj600}$	68.94	69.92	69.44	70.08	71.41	70.76	72.46	73.90	73.20	74.34	75.82	75.10
SelfTr$_{MPQA}$	69.18	67.85	68.55	70.03	69.46	69.75	72.50	72.19	72.15	74.57	75.47	75.04
SelfTr$_{Uni}$	69.82	71.16	70.51	70.53	72.41	71.50	73.17	74.87	74.05	75.73	77.67	76.74
SelfTr$_{Uni+Bi}$	71.14	74.69	71.94	71.41	73.64	72.57	74.39	76.12	75.29	76.43	78.62	77.58

5.4.1 Class Imbalance and Rule-based Classification

In the first experiment, we just focus on class imbalance (i.e., 3 star reviews are excluded). We examine a self-trained classifier using the class-ratio estimate of a rule-based classifier as it is the most obvious estimate since the rule-based classifier is also used for generating the labeled training data. In particular, we want to explore whether there is a systematic relationship between the class distribution, the class-ratio estimate of the rule-based classifier and the resulting self-trained classifier. Table 6 lists the actual distribution of classes on the test set, the deviation between the distribution as it is predicted by the rule-based classifier and the actual distribution along the information towards which class the rule-based classifier is biased. Finally, we also list the absolute improvement/deterioration of the self-trained classifier in comparison to the rule-based classifier. We will only consider the best rule-based classifier, i.e., RB$_{Weight}$, and for self-training, we will exclusively consider the best configuration from the previous experiments, i.e., SelfTr$_{Uni+Bi}$. The table shows that the quality of class-ratio estimates of rule-based classifiers varies among the different domains. The deviation is greatest on the *computer* domain. This is also the only domain in which the majority class are the negative reviews. With exception of the *sports* domain, the rule-based classifier always overestimates the amount of positive reviews. This overestimation is surprising considering that the polarity lexicon we use contains almost twice as many negative polar expressions as positive polar expressions. This finding, however, is consistent with our observation from Section 5.2 that rule-based classifiers have a bias towards positive reviews, i.e., they achieve a better F-score for positive reviews than for negative reviews[10]. Table 6 also clearly shows that the deviation negatively correlates with the improvement of the self-trained classifier towards the rule-based classifier. The improvement is greatest on the *sports* domain where the deviation is smallest and the greatest deterioration is obtained on the *computer* domain where the deviation is largest.

In summary, the class distribution of the data has a significant impact on the final self-trained classifier. In case there is a heavy mismatch between actual and predicted class ratio, the self-training approach will not improve the rule-based classifier.

5.4.2 Class Imbalance, Class Ratio Estimates and 3 Star Reviews

In the following experiment we will compare how alternative class-ratio estimates relate to each other when applied to self-training. We compare the actual distribution (Ratio-Oracle) with the balanced class ratio (Ratio-Balanced), the class ratio as predicted by the rule-based classifier over the entire dataset (Ratio-RB) and estimates gained from a small amount of randomly sampled data instances from the dataset. We randomly sample 20 (Ratio-20), 50 (Ratio-50) and 100 (Ratio-100) instances. For each configuration (i.e., 20, 50, and 100), we sample 10 times, run SelfTr for each sample and report the averaged result. We compare the self-trained classifier with a classifier always assigning a test instance to the majority class (Majority-Cl) and the rule-based classifier (RB$_{Weight}$). This time, we also include the 3 star reviews in the unlabeled dataset.

Table 7 displays the results. We also display results of the datasets without using 3 star reviews in brackets. SelfTr using Ratio-Balanced produces the worst results among the self-training classifiers. This was the only method used in previous work (in Chinese) [13, 15]. Apparently, English data are more difficult than Chinese and, in English, SelfTr is more susceptible to deviating class-ratio estimates since in [13, 15] SelfTr with Ratio-Balanced scores rather well. Ratio-Oracle produces best results which comes to no surprise since the class distribution in training and test set is the same. On average, Ratio-100 produces the second best result as it also gives fairly reliable class-ratio estimates (the deviation is 3.3% on average, whereas the deviation of Ratio-Balanced is 18.16%). Both Ratio-50 and Ratio-100 produce results which are significantly better than Majority-Cl and RB$_{Weight}$.

As Ratio-Oracle, Ratio-Balanced, Ratio-20, Ratio-50, and Ratio-100 suggest, the presence of mixed polar reviews does not produce significantly different results. It is very striking, however, that the results of Ratio-RB are better using the 3 star reviews which seems counter-intuitive. We found that this is a corpus artifact. As already stated in Section 3, 3 star reviews do not only contain indefinite polar reviews but also positive and negative reviews. We also noted that Ratio-RB has a bias towards predicting too many positive instances. The bias is stronger if 3 star reviews are not included in the ratio-prediction (deviation of 8.5% instead of 6%). We, therefore, assume that among the 3 star reviews the proportion of negative-like reviews is greater than among the remaining part of the dataset and RB within SelfTr detects them as such. Thus, the bias towards positive polarity is slightly neutralized.

In summary, using small samples of labeled data instances is the most effective way for class ratio estimation enabling SelfTr to consistently outperform Majority-CL and RB$_{Weight}$. Mixed reviews only have a marginal impact on the final overall result of SelfTr.

6 Conclusion

In this paper, we examined the effectiveness of bootstrapping a supervised polarity classifier with the output of an open-domain rule-based classifier. The resulting self-trained classifier is usually significantly better than the open-domain classifier since the supervised classifier

[10] We also observed that this bias is significantly larger on simple classifiers, such as RB$_{Plain}$, which is plausible since on this classifier the gap between F-scores of positive and negative reviews is also largest (see Table 5).

Table 6. Class imbalance and its impact on self-training.

Domain	Class distribution ($+ : -$)	Deviation of predicted distribution from actual distribution	Class towards which predicted distribution is biased	Difference in Accuracy between RB and SelfTr(RB)
computer	$43.17 : 56.83$	16.30	$+$	-3.60
products	$63.07 : 36.93$	6.65	$+$	-0.25
sports	$78.68 : 21.32$	2.10	$-$	$+3.15$
travel	$74.07 : 25.93$	3.71	$+$	$+1.30$

Table 7. Accuracy of different classifiers tested on naturally imbalanced data: for self-trained classifiers the unlabeled data also contain 3 star reviews; numbers in brackets state the results on a dataset which excludes 3 star reviews.

			SelfTr					
Domain	**Majority-Cl**	**RB**$_{Weight}$	**Ratio-Oracle**	**Ratio-Balanced**	**Ratio-RB**	**Ratio-20**	**Ratio-50**	**Ratio-100**
computer	56.83	73.80	82.80 (83.35)	**83.25** (**82.95**)	75.95 (70.20)	77.36 (77.95)	80.43 (80.91)	80.96 (81.47)
products	63.07	76.00	80.90 (81.70)	75.40 (76.05)	77.50 (75.75)	77.61 (78.10)	80.45 (80.86)	**80.69** (**81.27**)
sports	78.68	77.35	81.25 (81.10)	62.55 (60.30)	**80.75** (**80.50**)	79.10 (79.01)	79.94 (79.94)	80.62 (**80.50**)
travel	74.07	79.50	81.70 (81.60)	66.95 (66.10)	**81.15** (**80.80**)	77.96 (76.59)	80.64 (80.52)	80.76 (80.58)
average	68.16	76.66	81.66 (81.94)	72.04 (71.35)	78.84 (76.81)	78.01 (77.91)	80.37 (80.56)	**80.76 (80.96)**

exploits in-domain features. As far as the choice of the feature set is concerned, the supervised classifier within self-training behaves very much like an ordinary supervised classifier. The set of all unigrams and bigrams performs best.

The type of rule-based classifier has an impact on the performance of the final classifier. Usually, the more accurate the rule-based classifier is, the better the resulting self-trained classifier is. Therefore, modeling open-domain constructions relevant for polarity classification is important for this type of self-training. It also suggests that further improvement of rule-based polarity classifiers by more advanced linguistic modeling is likely to improve self-training as well.

In cases in which semi-supervised learning outperforms supervised learning, self-training at least also performs as well as the best semi-supervised classifier. A great advantage of self-training is that it chooses instances to be added to the labeled training set by using confidence scores whereas in semi-supervised learning one has to resort to random sampling. The resulting data from self-training are usually much better.

Self-training also outperforms a rule-based classifier and a majority-class classifier in more difficult settings in which mixed reviews are part of the dataset and the class distribution is imbalanced, provided that the class-ratio estimate does not deviate too much from the actual ratio on the test set. A class-ratio estimate can be obtained by the output of the rule-based classifier but, on average, using small samples from the data collection produces more reliable results.

Since this self-training method works under realistic settings, it is more robust than semi-supervised learning, and its embedded supervised classifier only requires simple features in order to produce reasonable results, it can be considered an effective method to overcome the need for many labeled in-domain training data.

Acknowledgements

Michael Wiegand was funded by the German research council DFG through the International Research Training Group "IRTG" between Saarland University and University of Edinburgh and the BMBF project NL-Search under contract number 01IS08020B. The authors would like to thank Alexandra Balahur for insightful comments.

REFERENCES

[1] A. Andreevskaia and S. Bergler, 'When Specialists and Generalists Work Together: Overcoming Domain Dependence in Sentiment Tagging', in *Proc. of ACL/HLT*, (2008).

[2] A. Aue and M. Gamon, 'Customizing Sentiment Classifiers to New Domains: a Case Study', in *Proc. of RANLP*, (2005).

[3] S. Baccianella, A. Esuli, and F. Sebastiani, 'SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining', in *Proc. of LREC*, (2010).

[4] S. Dasgupta and V. Ng, 'Mine the Easy, Classify the Hard: A Semi-Supervised Approach to Automatic Sentiment Classification', in *Proc. of ACL/IJCNLP*, (2009).

[5] T. Joachims, 'Making Large-Scale SVM Learning Practical', in *Advances in Kernel Methods - Support Vector Learning*, (1999).

[6] T. Joachims, 'Transductive Learning via Spectral Graph Partitioning', in *Proc. of ICML*, (2003).

[7] A. Kennedy and D. Inkpen, 'Sentiment Classification of Movie Reviews Using Contextual Valence Shifters', in *Computational Intelligence (Special Issue)*, volume 22, (2006).

[8] M. Klenner, S. Petrakis, and A. Fahrni, 'Robust Compositional Polarity Classification', in *Proc. of RANLP*, (2009).

[9] K. Moilanen and S. Pulman, 'Sentiment Construction', in *Proc. of RANLP*, (2007).

[10] V. Ng, S. Dasgupta, and S. M. N. Arifin, 'Examining the Role of Linguistic Knowledge Sources in the Automatic Identification and Classification of Reviews', in *Proc. of COLING/ACL*, (2006).

[11] B. Pang, L. Lee, and S. Vaithyanathan, 'Thumbs up? Sentiment Classification Using Machine Learning Techniques', in *Proc. of EMNLP*, (2002).

[12] L. Polanyi and A. Zaenen, 'Context Valence Shifters', in *Proc. of the AAAI Spring Symposium on Exploring Attitude and Affect in Text*, (2004).

[13] L. Qiu, W. Zhang, C. Hu, and K. Zhao, 'SELC: A Self-Supervised Model for Sentiment Classification', in *Proc. of CIKM*, (2009).

[14] D. Rao and D. Ravichandran, 'Semi-Supervised Polarity Lexicon Induction', in *Proc. of EACL*, (2009).

[15] S. Tan, Y. Wang, and X. Cheng, 'Combining Learn-based and Lexicon-based Techniques for Sentiment Detection with Using Labeled Examples', in *Proc. of SIGIR*, (2008).

[16] J. Wiebe and E. Riloff, 'Creating Subjective and Objective Sentence Classifiers from Unannotated Texts', in *Proc. of CICLing*, (2005).

[17] M. Wiegand and D. Klakow, 'Predictive Features in Semi-Supervised Learning for Polarity Classification and the Role of Adjectives', in *Proc. of NoDaLiDa*, (2009).

[18] T. Wilson, J. Wiebe, and P. Hoffmann, 'Recognizing Contextual Polarity in Phrase-level Sentiment Analysis', in *Proc. of HLT/EMNLP*, (2005).

Comparable English-Russian Book Review Corpora for Sentiment Analysis

Taras Zagibalov [1] and **Katerina Belyatskaya**[2] and **John Carroll** [3]

Abstract. We present newly-produced comparable corpora of book reviews in English and Russian. The corpora are comparable in terms of domain, style and size. We are using them for cross-lingual experiments in document-level sentiment classification. Quantitative analyses of the corpora and the language differences they exhibit highlight a number of issues that must be considered when developing systems for automatic sentiment classification. We also present experiments with a sentiment classification system applied to the corpora. The results suggest that differences in the way that sentiment is expressed in the two languages lead to large difference in sentiment classification accuracy.

1 INTRODUCTION

Automatic classification of document sentiment (and more generally extraction of opinion from text) has recently attracted a lot of 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 analyse market or media reaction towards their brands, products or services, or search engines that help potential purchasers make an informed choice about a product they are considering buying.

Most extant sentiment classification systems use approaches based on supervised machine learning, which require substantial manually-produced or -curated resources including texts annotated at the document level and below, sentiment term dictionaries and thesauri, and some level of language analysis.

There are a number of publicly available sentiment-annotated corpora, such as MPQA [15], and Pang and Lee's Movie Review corpus [8]. However, most of these corpora consist just of English text. As for other languages, we are aware of only one publicly available corpus, of Chinese product reviews [20]. There are other 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 [12]).

As part of an on-going research effort in sentiment analysis, we have designed and built comparable corpora of book reviews in English and Russian, which we are making publicly available, in the expectation that they will contribute to research in cross-lingual sentiment processing[4]. The Russian corpus is probably the first sentiment-annotated resource in that language.

In this paper, as well as describing the corpora and quantifying various relevant aspects of them, we analyse some important language-specific and domain-specific issues that would be likely to impact on automatic sentiment processing. We also describe experiments with standard machine learning sentiment classification technique applied to the corpora.

The paper is structured as follows. Section 2 surveys related work in sentiment classification. Section 3 describes the corpora. Section 4 presents experiments with the corpora and Section 5 concludes.

2 RELATED WORK

Most work on sentiment classification has used approaches based on supervised machine learning. For example, Pang et al. [9] collected movie reviews that had been annotated with respect to sentiment by their authors, and used this data to train supervised classifiers. A number of studies have investigated the impact on classification accuracy of different factors, including choice of feature set, machine learning algorithm, and pre-selection of the segments of text to be classified. For example, Dave et al. [2] experiment with the use of linguistic, statistical and n-gram features and measures for feature selection and weighting. Pang and Lee [8] use a graph-based technique to identify and analyse only subjective parts of texts. Yu and Hatzivassiloglou [18] use semantically oriented words for identification of polarity at the sentence level. Most of this work assumes binary classification (positive and negative), sometimes with the addition of a neutral class (in terms of polarity, representing lack of sentiment).

Turney [13] carried out an early and influential study into unsupervised sentiment classification. The approach starts from two 'seed' words and builds lists of positive and sentiment vocabulary from large amounts of text using a technique based on pointwise mutual information. For sentiment classification of movie reviews the approach achieves a relatively modest 65% accuracy (although reviews of automobiles are classified with 84% accuracy). Turney attributes this discrepancy in accuracy between domains to the much more complex structure of movie reviews. Popescu and Etzioni [10] extend the approach, applying hand-made rules, linguistic information and WordNet resources. Kobayashi et al. [4] employ

[1] University of Sussex, email: T.Zagibalovl@sussex.ac.uk

[2] Siberian Federal University, email: e.o.belyatskaya@gmail.com

[3] University of Sussex, email: J.A.Carroll@sussex.ac.uk

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

an iterative semi-automatic approach to extracting opinion-bearing expressions, although this requires human input at each iteration. Unsupervised and semi-supervised techniques may offer the promise of overcoming domain dependence since they do not require training data in order to be applied to a new domain. Wiebe and Riloff [14] present an unsupervised sentence-level subjectivity classifier that uses an extensive set (about 8000) of rules (subjectivity clues). Li, Zhang and Sindhwani [5] used labelled documents to adjust a hand-built sentiment lexicon to a domain. The extensive use of knowledge (rule or lexicons) make these approaches language-dependent.

An alternative approach to overcoming domain dependence is presented by Aue and Gamon [3], who attempt to solve the problem of the absence of large amounts of labelled data by customizing sentiment classifiers to new domains using training data from other domains. Blitzer et al. [1] investigate domain adaptation for sentiment classifiers using structural correspondence learning.

There has been little previous work on applying sentiment analysis to languages with scarce relevant language resources. A notable exception is the work towards producing cross-lingual subjectivity analysis resources from English data by Mihalcea et al. [7]. They use a parallel corpus to adjust a subjectivity lexicon translated from English to Romanian. Other multilingual opinion mining work (in English, Japanese and Chinese) was carried out by Zagibalov and Carroll ([19] and [21]), using techniques requiring limited manual input to classify newswire documents with respect to subjectivity and to extract opinion holders and targets.

A number of studies include development of linguistic resources for sentiment analysis. The text corpora are quite often annotated by a several annotators to produce different kinds of annotation. For example, Read [11] developed an annotation scheme with about 30 different tags that closely follows the Appraisal Theory [6]. Wilson and Wiebe [16] developed a detailed annotating scheme for expressions of opinions, beliefs, emotions, sentiment and speculation. To ensure annotation robustness, the authors calculate inter-annotator agreement. Another approach uses tags produced by authors ('self tagged') of the documents included to the corpus [2].

3 THE CORPORA

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, ensuring that the language used is 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 some others. The reviews were published on the website www.amazon.co.uk.

We annotated each review as 'POS' if positive sentiment prevails or 'NEG' if the review is mostly negative based on the tags assigned by reviewers, but moderated where the tag was obviously incorrect. 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]
```

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 1). It is not possible to compare these figures directly across the languages as they have different grammar structures which makes English more 'wordy' as it has function words (articles, auxiliary verbs) which are almost completely absent in Russian.

Russian, being a synthetic language, has a lot of forms of the same lemma. This results in a large number of distinct word forms: the corpus contains a total of 13472 word forms, with 6589 (42%) in positive reviews and 8993 (58%) in negative. The total number of words in the corpus is 50745, which means that every word form was used a little more than 3 times on average. The English corpus has only 7489 word forms in the whole corpus, 4561 (47%) in positive reviews, and 5098 (53%) in negative. The re-use of word forms in English is much higher: every word form was used 9 times on average (the total number of word in the corpus is 87539). These figures suggest that the Russian reviewers used a richer vocabulary for expressing *negative* opinion (compared to the number of unique words used in Russian positive reviews) than English reviewers.

Further evidence of the different ways in which people distinguish sentiment polarity in Russian compared with English is the distribution of lengths of positive and negative reviews. The Russian corpus has a large number of short reviews (less than 50 words) with a median of 15 words for positive reviews and 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 from English reviews in terms of length. The English reviews feature a more or less equal number of documents of different lengths (mostly in the range 15 to 75). The prevalence of short reviews in the Russian corpus, together with the rich morphological variation, may lead to data sparseness which would be a problem for many current sentiment classification techniques.

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

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

Table 1. Overall quantitative measures of the English and Russian corpora.

and hand-annotated. There were a lot of re-occurring short reviews like: Хорошо (*Good*); Интересная книга (*Interesting book*); Супер! (*Superb!*); Нудятина!! (*Boring!!*); Ниже среднего (*Below 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.

3.1 Ways of Expressing Sentiments

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

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 2 and 3). As the Tables show, the two languages express sentiment in slightly different ways. English makes heavy use of adjectives to express sentiment (this class of words is used to express sentiment in a third of all documents). In contrast, Russian uses verbs as often as adjectives to express sentiment (both of these classes are used in about quarter of all reviews) and makes 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 4083 compared to 4716 in Russian, which means that given equal number of reviews for each language, Russian reviews tend to have more different ways of expressing sentiment per document.

	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 2. 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 3. Ways of expressing sentiment in the Russian Book Review Corpus (numbers of documents).

Lexical Level

Adjectives Adjectives are the most frequent means of expressing opinions in both languages, closely followed by verbs in the Russian corpus. 1215 Russian reviews use adjectives to express sentiment and 1070 reviews use verbs. In the English corpus there are 1360 reviews that use adjectives, but only 701 use verbs to express opinion.

Apart from adjectives, which are recognised as the main tool for 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 As observed by some researchers, opinions delivered by 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. Appraisal verbs 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 a whole 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.

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*).

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

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:

(4) A BIG FAT ZEEROOOOOOOOOOOOOO for M.A

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

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 negative attitude to an author about his book:

(6) 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:

(7) 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

(8) 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 like the following are also used to express sentiment:

(9) 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.

(10) Run away! Run away!

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

(12) Читать однозначно.
Definitely should read.

(13) Читать !!!!!!!!!!! ВСЕМ
Read!!!!!!!! EVERYONE

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

(14) It certainly leaves you hungering for more!

(15) Buy at your peril. Mines in the bin!

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

(16) И это автор вычислителя и леммингов? ...
so this author calculator and lemmings? ...
НЕ ВЕРЮ! Садись, Громов, два.
(DO)NOT BELIVE! sit 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 used to express negative sentiment. This review also lacks any explicit sentiment markers. The negative appraisal is expressed by the verbs 'stab' and 'burn' that only in this context show negative attitude.

(17) Stab the book and burn it!

Discussion 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 the noun, in 39% of reviews. Sentence-level and discourse-level sentiment phenomena are found in 56% of reviews. 3% of reviews contain phonetic phenomena.

[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'.

3.2 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 number 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:

(18) 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 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:

(19) 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 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:

(20) 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 (lit.) let it (GG) burn with blue flames (≈ I do not at all care about GG)

(21) 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.

4 EXPERIMENTS

We used Naïve Bayes multinomial (NBm) and a Support Vector Machine classifiers[7] to investigate performance of standard supervised classifiers on the two corpora . The feature sets were the lexical units extracted from the relevant corpora. We extracted all words from the corpora but did not process them in any way (no stemming or lemmatisation). 15582 words were extracted from the Russian corpus and 9659 words were found in the English book reviews. The evaluation technique is 10-fold cross-validation.

	NBm			SVM		
Corpus	P	R	F	P	R	F
English book reviews	0.88	0.88	0.88	0.84	0.84	0.84
Russian book reviews	0.81	0.81	0.81	0.78	0.78	0.78

Table 4. Supervised classification results (Precision, Recall and F_1, 10-fold cross-validation)

Table 4 show the results of supervised classification, Russian review classification being 6-7 percentage points worse the results obtained from the English corpus.

5 CONCLUSION

In this paper we presented comparable corpora of English and Russian book reviews, providing the research community with a resource that can be used for cross-lingual sentiment classification experiments. We examined language-specific features

[7] We used WEKA 3.4.11 [17] (http://www.cs.waikato.ac.nz/~ml/weka)

of the reviews that are relevant to sentiment classification and showed that sentiment in different languages is expressed in slightly different ways, covering all levels of the language: from phonetic to discourse. The experiments suggest that these differences have an impact on the accuracy of a standard, supervised sentiment classification technique.

In future work, we intend to investigate in more depth which specific characteristics of different languages lead to differences in sentiment classification accuracy, using sentiment-annotated corpora of English, Russian, Chinese and Japanese.

References

[1] John Blitzer, Mark Dredze, and Fernando Pereira, 'Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification', in *Proceedings of the 45th Annual Meet- ing of the Association of Computational Linguistics.*, pp. 440–447, Prague, Czech Republic, (June 2007). Association for Computational Linguistics.

[2] 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*, pp. 519 – 528, Budapest, Hungary, (2003). ACM Press.

[3] 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*, pp. 57–64. Association for Computational Linguistics, (2005).

[4] Nozomi Kobayashi, Kentaro Inui, Yuji Matsumoto, Kenji Tateishi, and Toshika Fukushima, 'Collecting evaluative expressions for opinion extraction', *Natural Language Processing–IJCNLP 2004*, **13**(12), 596–605, (December 2004).

[5] Tao Li, Yi Zhang, and Vikas Sindhwani, 'A Non-negative Matrix Tri-factorization Approach to Sentiment Classification with Lexical Prior Knowledge', in *Proceeding of Association for Computational Linguistics*, number August, pp. 244—-252, Morristown, NJ, USA, (2009). Association for Computational Linguistics.

[6] J.R. Martin and Peter Robert Rupert White, *The language of evaluation: Appraisal in English*, Palgrave Macmillan, 2005.

[7] Rada Mihalcea, Carmen Banea, and Janyce M Wiebe, 'Learning multilingual subjective language via cross-lingual projections', in *976 Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, volume 45, pp. 976—-983, Prague, Czech Republic, (2007).

[8] Bo Pang and Lillian Lee, 'A sentimental education: Sentiment analysis using subjectivity summarization based on Minimum Cuts', in *the 42nd Annual Meeting on Association of Computational Linguistics*, Barcelona, Spain, (2004).

[9] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan, 'Thumbs up?: sentiment classification using machine learning techniques', in *Conference on Empirical Methods in Natural Language Processing*, pp. 79—-86, (2002).

[10] Ana-Maria Popescu and Oren Etzioni, 'Extracting product features and opinions from reviews', in *Natural Language Processing and Text Mining*, pp. 9–28, Vancouver, Canada, (October 2005). Springer.

[11] Jonathon Read, David Hope, and John Carroll, 'Annotating expressions of appraisal in English', *ACL 2007*, 93, (2007).

[12] 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', *Proceedings NTCIR-7, NII, Tokyo*, 185–203, (2008).

[13] Peter D. Turney, 'Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews', in *Annual Meeting of Assosiation of Computational Linguistics*, pp. 417–424, Philadelphia, Pennsylvania, (2002).

[14] Janyce M Wiebe and Ellen Riloff, 'Creating subjective and objective sentence classifiers from unannotated texts', *Computational Linguistics and Intelligent Text Processing*, 486–497, (2005).

[15] Janyce M Wiebe, Theresa Ann Wilson, and Claire Cardie, 'Annotating expressions of opinions and emotions in language', *Language Resources and Evaluation*, **39**(2), 165–210, (2005).

[16] Theresa Ann Wilson and Janyce M Wiebe, 'Annotating opinions in the world press', in *4th SIGdial Workshop on Discourse and Dialogue (SIGdial-03)*, pp. 13–22, (2003).

[17] I.H. Witten and Eibe Frank, *Data Mining: Practical machine learning tools and techniques*, Morgan Kaufmann Pub, San Francisco, 2nd edn., 2005.

[18] Hong Yu and Vasileios Hatzivassiloglou, 'Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences', in *Proceedings of EMNLP*, volume 3, pp. 129–136. Association for Computational Linguistics, (2003).

[19] Taras Zagibalov and John Carroll, 'Almost Unsupervised Cross Language Opinion Analysis at NTCIR 7', in *NTCIR-7*, Tokyo, (2008).

[20] 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*, pp. 1073—-1080, Manchester, United Kingdom, (2008).

[21] Taras Zagibalov and John Carroll, 'Multilingual Opinion Holder and Target Extraction using Knowledge-Poor Techniques', in *Language and Technology Conference*, Poznañ, Poland, (2009).