문학석사 학위논문

Sentiment Analysis of Online Reviews based on Genre-specific Discourse Patterns

장르 특정적 담화 유형 기반의 온라인 리뷰의 감정분석

2015년 8월

서울대학교 대학원
언어학과 언어학전공
Otmakhova Yulia

Abstract

Sentiment Analysis of Online Reviews based on Genre-specific Discourse Patterns

Otmakhova Yulia

Department of Linguistics

The Graduate School

Seoul National University

Though in recent years sentiment analysis has evolved from simple lexicon-based and statistical models to methods involving discourse information, the major problem with the current approaches is that they use the same set of features for sentiment classification of texts of all genres and types (tweets, editorials, discussion board posts, online reviews etc.). Moreover, features that were used by previous researchers reflect only one aspect of discourse, namely, coherence, and they are limited to explicit ways of ensuring coherence, such as conjunctions. To be more specific, these are such features as *implicit coherence*, realized through

adjacency of two sentences, *continuity*, which shows that two sentences have the same sentiment and is commonly reflected through the use of such conjunctions as *and* or *moreover,* and *contrast*, which is indicated by such conjunctions as *but* and shows the shift of the opinion's polarity.

In this study we propose a new set of features which reflects the specific traits of a particular genre – online reviews: *implicit contrast*, realized through usage of such limiting expressions as *the only drawback*; *background* patterns, which are expressions that help to establish a review author's identity; and *involvement* features, which are used to interact with the reader.

To show the effectiveness of these features, we annotated a corpus of 120 product reviews and represented each review as a set of non-discourse, generic and genre-specific discourse features extracted from it (together with the target label from the annotation). Such feature sets were used in two series of experiments: fine-grained and coarse grained. At the sentence level we conducted the experiments with and without lexical features, while at the document level we performed 5-, 3- and 2-class classification. Our experiments showed that genre-specific features in general perform better than the generic ones, ensuring greater improvements in precision and recall. If generic features led to minor increases or even deteriorated the performance (as in case of *implicit coherence*), genre-specific features (especially *background)* were more stable and allowed us to achieve better recall and precision across all experiments. These tendencies were especially remarkable in the fine-grained classification with lexical features, where adding generic discourse features to the lexical ones deteriorated the results. Moreover, the performance of genre-specific features is not only statistically reliable but also reflects the theoretical properties of online reviews discourse outlined in our study.

Keywords: sentiment analysis, opinion mining, online reviews, product reviews, discourse analysis

Student number: 2012-23882

Table of Contents

List of Figures

List of Tables

1. Introduction

1.1 Subject Matter

Sentiment analysis deals with automatic prediction of semantic orientation, or polarity[1], of a text or its part – that is, determining if a particular sentence or text is positive, negative or lacking any opinions and sentiments (objective). Driven by social and economic needs, such as the necessity to automatically learn public opinions on a particular incident, issue or product, in recent years sentiment analysis has become a major field of study for natural language processing (NLP) and computational linguistics and has made a significant progress from simple lexicon-based and statistical methods to complex models in which context plays an increasingly important role.

In addition to such tasks as disambiguation of polarity cues at the lexical context level (Wilson et al., 2009; Wu and Jin, 2013) and resolving the scope of negation that switches the polarity at the syntactic context level (Polanyi and Zaenen, 2006; Councill et al, 2010; Lapponi et al., 2012), more and more tasks

[1] In early studies *semantic orientation* and *polarity* were defined as properties of a word: these terms were used to refer to the "evaluative character of a word" (Turney, 2002) or the "direction the word deviates from the norm for its semantic group or lexical field" (Hatzivassiloglou and McKeown, 1997). However, as evaluation can be done at a higher level than a word, and opinions can be implicit, that is, lacking any sentiment words (Liu, 2012), in this study we use these terms at any level where an opinion can be expressed (word, phrase, clause, sentence or a document).

nowadays involve higher levels of pragmatics and discourse. However, the main drawback here is that the majority of sentiment analysis studies employing discourse information use the same set of discourse relations whenever they deal with online reviews, discussion boards, tweets etc. Though researchers have been long aware of the need of genre-specific discourse features and genre constraints (Polanyi and Zaenen, 2006), in practice almost all attempts at using discourse information boil down to a standard approach based on coherence markers and corresponding relations, such as contrast or continuation. To be more specific, these are conjunctions or connectives which indicate continuity (*and, moreover*) or a shift (*but, although*) of sentiment.

While such generic relations are important to any discourse and universally applicable, we argue that discourse relations which are unique to a particular genre[2] can be as – or even more – valuable for sentiment analysis. To prove it, in this study we determine and describe several discourse patterns which frequently appear in online reviews, and show how using them improves the results of sentiment analysis at the sentence and document level. While the task of determining genre-specific relations and extracting the corresponding patterns is hindered by the fact that most of them are implicit, that is, not expressed by discourse markers that form a closed set[3], in this study we follow the idea of

[2] In this study we use the term "genre" to refer to a linguistic genre - a discourse form which is motivated by some social needs and has some predictable structural conventions, such as an invitation, a recipe or a lecture (Strauss and Feiz, 2014), not a literary form (such as tragedy or a poem).

[3] Though Taboada (2009) claims that all discourse relations are signaled (explicit) by different means in addition to discourse markers (such as semantic, morphological,

Mackiewicz (2010), who maintained that "close analysis of reviews provides benefits that text mining cannot", and show that linguistically informed features, based on theoretical findings and corpus study of a particular genre, can be more meaningful and useful for sentiment analysis than the generic ones.

1.2 Purposes of the Study

The purposes of this study can be briefly summarized as follows:

1) **Describing and formalizing generic discourse features** which are commonly used in sentiment analysis of all genres and types of texts. Though such features have been widely studied and applied in opinion detection, we formalize them in terms of Rhetorical Structure Theory (Mann and Thompson, 1988) and simplify them for practical purposes.

2) **Describing and formalizing genre-specific features** of online reviews applicable to sentiment analysis. Though reviews have been studied as a genre in sociolinguistics (Vasquez, 2014), there are no NLP studies which employ discourse features or unique characteristics of online reviews. In our study, we try to discover some patterns which are likely to appear in online reviews and signify some sentiment, change of opinion or absence of it. Where applicable, we formalize such

---

syntactic, or pragmatic mechanisms), in this study we use the terms "explicit" and "implicit" discourse relations in a conventional way (Renkema, 2004): the relations which can be identified by well-known discourse markers (conjunctions, adverbs, adverbial phrases), forming a closed set, are referred to as *explicit* in this thesis, while relations which lack such predictable indicators will be called *implicit*.

patterns based on Rhetorical Structure Theory (RST).

3) Comparing the effect of generic and genre-specific discourse features on sentiment analysis of online reviews. We perform two series of experiments that show how genre-specific features can improve the results of sentiment analysis which relies only on generic discourse features. We discuss how such features improve the results of opinion detection both at the sentence and document level.

1.3 Contributions of the Study

The main contributions of this study are outlined below:

1) We annotated a corpus of 120 user reviews about products belonging to different domains (movies, electronic and household appliances, books, clothes, food etc.). The annotation was done at the clause or sentence level, and a reliability study showed its high accuracy.

2) Through the corpus study based on another corpus[4] we determined several patterns that are specific to online reviews, generalized them and used them to extract features from our annotated corpus in order to perform the experiments.

3) We were first to conduct experiments that included genre-specific discourse features, and we showed that such features outperform the

[4] Darmstadt Service Review Corpus, available from https://www.ukp.tu-darmstadt.de/data/sentiment-analysis/darmstadt-service-review-corpus (Toprak et al., 2010).

generic ones in precision and recall. Thus we proved their usefulness for sentiment analysis both at the sentence and the document level.

4) In addition to experiments conducted to prove the validity of genre-specific features, we were the first to perform the following two experiments:

4a) An experiment which showed that for the sentence-level analysis removing all lexical information can improve the results, provided there is a set of reliable discourse features and information on the global polarity of the text (such as a product's rating indicated by the number of stars assigned to the review).

4b) An experiment which showed that the importance of discourse information for document-level sentiment analysis increases with the number of classes, that is, while it is insignificant for two-class (positive vs. negative) classification, it helps to substantially improve the accuracy of detailed classification (3 or 5 class).

1.4 Structure of the Study

In this thesis, we firstly describe previous studies related to sentiment analysis of online reviews and using discourse features for sentiment analysis. Then we give an overview of generic and genre-specific discourse features and describe the theoretical background of this study. Next, we describe the corpus we used for our experiments, its annotation and features extracted from it for the purposes of this study. In the next two chapters we discuss our experiments conducted at various

levels of granularity and in different settings: Chapter 5 deals with the sentence-level sentiment analysis with and without lexical features, while Chapter 6 is devoted to predicting review ratings and discusses the degree in which discourse information influences the results of two-class or multi-class classification. Finally, in Chapter 7 we summarize the results of our study.

2. Previous Studies

2.1 Previous Studies on Sentiment Analysis of Online Reviews

Online reviews was the first genre to which the methods of sentiment analysis were applied. Before the seminal work of Turney (2002), which dealt with classifying reviews as positive or negative, sentiment analysis was done at the word and phrase level. To give just a few examples, Hatzivassiloglou and McKeown (1997) determined the semantic orientation of adjectives based on supervised learning; Wiebe in her early works (2000, 2001) annotated a corpus of 1001 sentences as subjective or objective and used it to learn adjectives expressing subjectivity; Riloff et al. in a later work (2003) used such patterns as "expressed <direct_object>" and a bootstrapping algorithm to extract nouns representing private states[5].

While early schemes for sentiment analysis were very detailed (they aimed at detecting not only the opinion itself, but also its *holder* and *target*), this level of granularity was not required for more coarse tasks, such as determining what users think about the product. Here the target of the opinion (product) was predefined, the holder was not important, and the only thing to determine was the general opinion. Such low requirements can explain why online reviews became the first genre to which sentiment analysis methods were applied at the document level. A simple two-way classification ("thumbs up/thumbs down", "positive/negative"),

[5] The term *private state* was introduced by Quirk et al. (1985) and then used in Wiebe (1990) and later works of other researchers to refer to internal states which cannot be observed and verified, such as opinions, emotions and beliefs.

introduced in Turney's work (2002), for some time remained the most popular task for document-level analysis (Pang and Lee, 2004; Dang, 2010). However, there were also attempts at 3-, 4- or 5-class classification, that is, predicting the actual number of stars the review has (Pang and Lee, 2005; Ganu et al., 2009).

Over the years the focus again switched to a more fine-grained classification. Though *holders* of opinions were still considered unimportant (reviews are always first-person discourses, and though their authors can be determined by IDs, their identity is largely irrelevant to the task), *targets* were generalized and grouped into *aspects* or *features* of a product (such as camera's lens or weight), and researchers recognized the need to extract opinions regarding them (Popescu and Etzioni, 2007; Baccianella et al., 2009; Jo and Oh, 2010). Moreover, as the review's author can express different opinions about various aspects of a product or service, which leads to conflicting evaluations in the review, determining sentiment at the sentence or clause level became an important task (Täckström and McDonald, 2011; Zirn et al., 2011). Further, though at first a large number of opinions could not be detected because they did not have any overt lexical cues, recently there were some attempts at identifying implicit opinions, or polar facts, in online reviews (Toprak et al., 2010) and at determining the reasons (which can also bear some sentiment, but have no lexical cues) behind negative or positive opinions of product features (Zhang et al., 2013).

In this study, we conduct experiments both at the fine-grained level (determining the sentiment of a particular clause or sentence in a review) and at the coarse-grained level (predicting the product rating on a two-point and a multipoint scale).

2.2 Previous Studies on Discourse in Sentiment Analysis

As the drawbacks of lexicon-based bag-of-words classification became obvious, more attention was paid to the structure of the analyzed texts and the discourse features used in them. Pang and Lee (2004) were probably the first to consider sentences not in isolation and to use Wiebe's (1994) idea that the sentiment should be consistent between two adjacent sentences. They used pairwise interaction information in a graph-based method relying on minimum cuts. In an attempt to detect domain-dependent polar clauses (opinions which cannot be learned automatically based on a training corpus, as the ways of expressing them depend on a domain) Kanayama and Nasukawa (2006) collected candidate items adjacent to domain-independent (lexicon-based) polar items. McDonald et al. (2007) relied on a similar approach, using the label of the previous sentence as a feature. Moreover, McDonald et al. (2007) and Täckström and McDonald (2011) claimed that the local sentence-level sentiment is influenced by and should be consistent with the global polarity of the document, and used coarse-level predictions (review ratings) to correct misclassified opinions at the sentence level. Global polarity also appeared as one of the features in a study of Qu et al. (2012), where they used potentially noisy indicators (review ratings) and a small set of base predictors (phrase-level lexical predictions, language heuristics and co-occurrence counts) in a Gaussian model with a multi-expert prior.

Apart from methods based on adjacency or consistency of sentiments, there were some more linguistically-informed approaches. The first attempts at incorporating discourse structure and relations into sentiment analysis were purely theoretical: Polanyi and Zaenen (2006) explained how one can calculate the overall

sentiment of a text taking context into account, while Asher et al. (2008) showed how sentiments combine in a text, using SDRT discourse theory (Lascarides and Asher, 2007). Somasundaran (2010), whose work was mainly based on product discussion meetings and discussion boards, was the first to go further than annotation and theoretical schemes and develop a practically applicable discourse framework. She enforced discourse constraints (in forms of reinforcing and non-reinforcing relations) on the polarity of segments with the same or alternative targets using opinion frames based on explicit coherence cues such as conjunctions. Lazaridou et al. (2013) discard opinion frames due to their complexity, but use three similar discourse relations which encode aspect/polarity change (*AltSame*, *AltAlt*, *SameAlt*) to jointly induce discourse, sentiment and topic information in a weakly supervised way. Even more simplified approach was proposed by Trivedi and Eisenstein (2013), who use only two kinds of discourse connectors (*shift* and *continuation*) in combination with proximity features and document-level annotations to disambiguate and correct sentence-level predictions made by OpinionFinder[6].

The majority of other sentiment analysis methods relying on discourse information use simplified versions of Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) relations and explicit cues (conjunctions) denoting them. Zhou et al. (2011) grouped RST relations and their cues into more general classes (*Contrast*, *Condition*, *Continuation*, *Cause* and *Purpose*) and used nucleus-satellite relations to determine which clause has the primary polarity in order to disambiguate sentiments. In a similar approach, Yang and Cardie (2014) discover

[6] Available at http://mpqa.cs.pitt.edu/opinionfinder/

opinionated sentences which do not have strong sentiment signals (implicit opinions) using RST-based discourse cues (*Expansion, Contingency, Comparison, Temporal, List*), coreference relations which are similar to *SameTarget* relations in Somasundra's work (2010) and global polarity information (review ratings).

While it was as early as 2006 that the researchers noticed the necessity of incorporating genre information in opinion mining tasks (Polanyi and Zaenen, 2006), until now there were no attempts at using genre-specific discourse features. More particularly, Polanyi and Zaenen describe some genre constraints specific for movie reviews and give an example of how genre conventions can influence valence (polarity) assignment. However, they do not propose any practical algorithm for using such genre-specific information and assert that it is infeasible unless we learn to determine the genre structure automatically. Thus this study (and previously published related study by Otmakhova and Shin (2015) which introduced several genre-specific discourse patterns) is the first attempt at describing, extracting genre-specific discourse features and applying them for sentiment analysis of online reviews at the coarse- and fine-grained level.

3. Generic and Genre-specific Discourse Features for Sentiment Analysis

3.1 Theoretical Background

Before giving an account of discourse theories that lay the foundation of our study, we briefly explain what we mean by *generic* and *genre-specific* features.

In almost all studies that attempt to improve the lexical baseline by employing discourse features (Somasundaran, 2010; Zhou et al., 2010, Yang and Cardie, 2014), one can expect to find at least one of the following discourse cues:

1. **Implicit coherence**, which is realized through adjacency (that is, two consecutive sentences are supposed to have the same semantic orientation unless there is strong evidence to the contrary) (Wiebe, 1994).
2. **Cause and continuity markers**, which indicate continuation of discourse flow.
3. **Contrast and concession markers**, which show changes, or breaks, in the sentiment flow, and
4. (for fine-grain classification) **Global polarity features** (such as the review's score or overall sentiment of a text), which determine the general semantic orientation of the document and help disambiguate the sentiment of a particular sentence or phrase.

All of these features have been extensively used by researchers under various names and in different combinations, and they were shown to improve results of sentiment analysis whether one tries to classify ideological debates, discussions or product reviews. Taking into account their universal character, we refer to them as *generic discourse features*. However, one might note that all such features are related to only one of discourse properties – its cohesion[7], – and though universally applicable, they are limited in their effect.

If we truly want to employ discourse information, we need to overcome the limits of generic cohesion-based approach and turn to a higher level of discourse relations – discourse patterns which ensure connectedness of a cognitive representation of a text even without explicit cohesion cues such as conjunctions (Sanders and Maat, 2006). Though some discourse patterns, such as *cause-consequence, problem-solution*, are commonly described in literature, theoretically their number is not limited (Kirkpatrick, 1991), and their function, unlike that of connectives[8], can depend on a text's genre. We refer to such features as "*genre-specific*": some of them can appear only in texts of a particular genre, while some

[7] To be even more exact, they represent only one of aspects of cohesion, which, alongside with *conjunction*, includes, according to Halliday and Hasan's (1976) classification, *reference, ellipsis, substitution* and *lexical cohesion*.

[8] It should be noted that connectives can also have additional functions depending on the mode and genre of discourse: as Sanders and Maat (2006) explain, such connectors as *but*, which express a contrastive relation, can have an additional function of signaling turn-taking in conversations. However, their primary function and discourse relations they represent are the same in all genres, so we refer to them as *generic* features.

can be more widespread, but their purpose and meaning may depend on the kind of text they are used in.

Consider, for instance, the following sentence:

(1) Let's start off by saying that I am a big supporter of Apple and their products as well as a proud owner of an iMac, Macbook, iPods, and now the latest and greatest iPhone 4.

One might expect to find it at the beginning of an online review or a debate, but not in a tweet or an editorial. This sentence gives us some information about the author of the review and is related to the following sentence by an RST (Mann and Thompson, 1988) relation of *background*. In terms of sentiment analysis, this sentence is objective: though it has lexical expressions which may cause it to be misclassified as positive (*big supporter, proud, greatest*), the genre-specific *background feature* [I am a] takes precedence and shows that this sentence is not an evaluation of the product but a claim made by the author about him/herself and thus should not be assigned any polarity. On the other hand, because the author has taken such an effort to convince us that he or she is an expert on Apple products, we might expect that the next sentence – and the whole review of another Apple product – is negative. As Vasquez (2014) showed, at the beginning of a review authors often construct their identity to gain trust of their audience and prove their expertise to consequently support their negative opinion of product.

On the other hand, such discourse features as *involvement* (Vasquez, 2014), which indicates interaction between the participants of a discourse and is often realized through the use of questions (Chafe, 1982; Gumperz, 1982), have different

functions depending on the genre of the text. More specifically, in discussions, which are dialogical in nature, questions can be used to challenge an opponent and thus express a negative opinion, while in online reviews, which are monological, questions are used to build rapport with the reader and usually carry no sentiment.

The examples discussed above represent two high-level approaches to discourse studies – Rhetorical Structure Theory approach (Mann and Thompson, 1987), which focuses on discourse (rhetoric) patterns in a text, and a sociolinguistic approach, which views discourse as a social interaction and analyzes text patterns in term of their social function. They are discussed in more detail in the following sections.

3.2 Discourse in Rhetorical Structure Theory

Rhetorical Structure Theory (RST) is a descriptive framework for text representation, which was originally developed to aid description and analysis of texts and computer-based text generation (Mann and Thompson, 1988). While the Rhetorical Structure Theory received criticism as theoretically ungrounded and inconsistent[9], it remains a useful and computationally viable tool for a wide variety of NLP tasks. In particular, in addition to natural language generation systems, RST was used for such applications as discourse parsing, text summarization, indexing, information extraction, machine translation, and sentiment analysis (Taboada and

[9] The most common criticism is probably the one related to the number of RST relations – the authors of RST admit that the list of relations is open and can be extended when needed (which was done by the authors themselves), and the critics claim that this can lead to creation of an unnecessary large number of arbitrary relations (Kehler, 2002).

Mann, 2006).

According to Rhetorical Structure Theory, a text has a hierarchical structure, which can be represented in the form of a dependency tree (Figure 1):

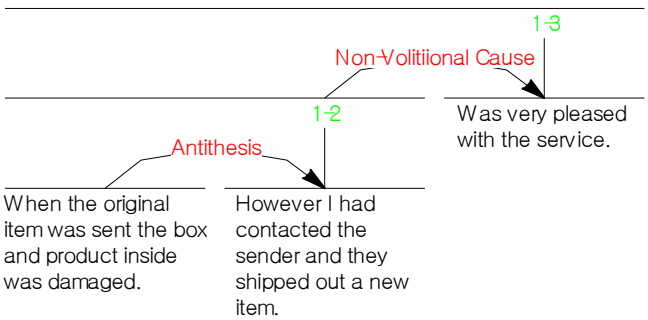

Figure 1. RST representation of an online review[10]

Figure 1 represents a review of a guitar, which contains both positive and negative opinions about the product. The arrows in the diagrams point from the satellite (the span expressing a subordinate idea, or, in our case, opinion) to the nucleus (the span which is semantically more important). If both spans are equally important, as in case of multinuclear relations, they are connected by an arc without arrows.

The arrows here are labeled with the names of relations taken from (Mann and Thompson, 1988). Each RST relation is defined by four aspects (some of which can be optional): constraints on the nucleus, constraints on the satellite, constraints on the combination of the nucleus and satellite, and the effect. For example, for the relation of *Antithesis* the writer should have a positive attitude towards the nucleus

[10] All RST trees in this thesis were drawn by the author using RSTTool v.3 (available from http://www.sfu.ca/rst/06tools/index.html)

(constraint on the nucleus), the situations described in the nucleus and the satellite must be in contrast (constraint on the combination of the nucleus and satellite), and thanks to this conflict the positive regard towards the situation expressed in the nucleus should increase (constraint on the effect) (Mann and Thompson, 1988). Thus the relations between the nucleus and the satellite and the constraints imposed on them can help us to determine which of opinions is more important and deduce the general polarity of a review. Moreover, in case of imperfect lexical predictions such relations can ensure consistency and coherence of sentiments and help to disambiguate them.

In this study, we simplify RST relations and combine them into several relation classes. In doing so, we relax some of the constraints and generalize the effects while maintaining the concepts of the nucleus and the satellite. For example, *volitional cause* and *non-volitional cause* are merged into a single relation of *Cause*, and thus the constraint on the nucleus (that the action expressed by it must be volitional or non-volitional) disappears. Though such generalization makes our framework less accurate theoretically, in practice it helps to reduce noise in datasets used for machine learning (Marcu and Echihabi, 2002). Moreover, as we are interested only in one aspect of discourse relations between the spans – we need to know whether they indicate continuation of opinion, its change or a switch from/to no opinion at all – we can discard or merge some of the relations based on this principle. In fact, in a similar attempt to generalize RST relations and make them more applicable for practical use, Zhou et al. (2011) showed that only 13 relations accounted for over 70% of all relations in their annotated corpus, and these 13 relations could be further grouped into 5 classes for disambiguating polarity.

3.3 Discourse in Sociolinguistics

If Rhetorical Structure Theory views discourse relations as a way to achieve coherence of cognitive representation of a text, sociolinguistic theories are more concerned with roles and functions such relations have in communication. It would be incorrect though to view these two approaches as incompatible: all RST relations, apart from ensuring coherence of a text, are used by writers with some communicative purpose, such as convincing the reader. For instance, the RST relation of *background*, briefly described above, not only helps to connect an objective sentence with an opinionated one, but provides a way for the author to show his or her expertise and construct their identity (Mackiewicz, 2010a). One major exception is the relation of *involvement* (Vasquez, 2014) which does not have a corresponding RST relation. This can be explained by the fact that it reflects the interaction between the writer and the reader and thus has a dialogical nature, while RST was not designed to be used for dialogs and multilogs (Mann and Thompson, 1988).

Sociolinguistics, and in particular interactional sociolinguistics (Gumperz, 1982), studies the way speech events (spoken or written) are organized: how speakers and writers build their identities (Bucholtz and Hall, 2005), how they interact with each other (Gumperz, 1982), how they introduce or switch topics, how they tell stories (Stubbs, 1983) or construct narratives (Bamberg, 1997), etc. All of these questions require discourse analysis to one degree or another; in fact, in some aspects discourse analysis and sociolinguistics are so close that it is hard to make distinction between them (Boutet and Maingueneau, 2005). In our study, we will keep them separate and use the term *discourse analysis* to refer to the study of

semantic or pragmatic connectedness of a text (such as RST analysis), and the term *sociolinguistic approach* to denote the functional side of speech events.

As this study focuses on identifying discourse features and patterns of online reviews that can be applicable to sentiment analysis, we will again describe only those patterns of communication, narration or interaction which can help us to identify and disambiguate opinions in this particular genre. These patterns and features, together with relevant RST relations, are outlined in the next chapter.

4. Data and Features

In this chapter we describe the corpus we used and the features extracted from it for our experiments.

4.1 Data and Annotation

4.1.1 Corpus

For the experiments conducted during this study we used Filatova's Amazon online reviews corpus (Sarcasm Corpus)[11], which consists of 1254 reviews (437 sarcastic and 817 regular) (Filatova, 2012). We chose this corpus because the presence of ironic reviews makes it difficult to classify using standard methods. Moreover, the corpus contains reviews of a wide variety of products (books, music albums, DVDs, electronics, household goods, food, clothes etc.), which ensures that the experiment results are domain-independent. The distribution of reviews in the corpus according to their rating (number of stars) is shown in Table 1.

Number of stars	**1**	**2**	**3**	**4**	**5**
Number of reviews	326	44	55	110	719
% of reviews	26%	3.5%	4.4%	8.8%	57.3%

Table 1. Distribution of reviews in Sarcasm corpus

As can be seen from Table 1, the corpus is not well balanced (the number of

[11] Available at http://storm.cis.fordham.edu/~filatova/SarcasmCorpus.html

reviews with a particular rating differs greatly). To prevent a skew towards positive labels we used equal-size random samples of reviews with all possible scores. This resulted in a subset of 120 reviews (24 reviews for each rating). During 6-fold cross-validation they were further subdivided into 100 reviews for training and 20 reviews for test data while maintaining the proportion of ratings.

4.1.2 Annotation Guidelines and Results

Each review in the corpus was annotated both at the document and sentence level. The label for document-level annotation was assigned automatically based on the review's rating (from the set of {1-5}). The labels for sentences or clauses (*positive, negative* or *objective*) were assigned manually according to guidelines outlined in this section. The reviews were annotated at the sentence level if there were no conflicting opinions in a sentence, and at the clause level in case of conflict.

Before we describe the guidelines, though, we feel the need to justify our choice of the three-way classification (*positive/negative/objective*) for the sentence level. While the studies in sentiment analysis usually make distinction between subjective and objective sentences on one hand and between negative, positive and neutral sentences on the other (Liu, 2012), we make a twofold distinction, first classifying a segment as objective or subjective, and then, for subjective (polar) sentences, further subdividing them into positive and negative. To our mind the classification into positive, negative and neutral sentences, commonly adopted for online reviews, is incorrect, as neutral sentiments rarely, if ever, appear in reviews. What is often referred to as neutral sentences should be classified as objective segments, as they do not carry any sentiment related to the subject matter.

Because the corpus contained a large number of ironic sentences and implicit opinions (Liu, 2012), when annotating it we considered the intended semantic orientation of a segment, not its literal meaning and the presence and polarity of lexical cues. This lead to establishment of the following guidelines:

A. Segments without any lexical cues can be annotated both as subjective and objective:

(2) I bought this mobo from Amazon, afte[12] buying the same month the DG31PR Classic for my wife. (*objective*)

(3) After I install my new PC, the 2do. day of use, the LAN failed. (*subjective*, *negative*)

Both of these sentences do not have any lexical cues indicating the presence of sentiment, however the second one expresses a negative opinion and thus should be labeled as NEGATIVE.

B. Segments with a lexical cue of a certain polarity can be annotated both as positive and negative:

(4) The ring is **nice** and heavy. (*positive,* from a review of a ring)

(5) It's going to be a **nice** paperweight. (*negative,* from a review of a camera)

[12] We keep the original spelling and grammar in all examples from the corpus here.

In the second review the positive word *nice* is used ironically, so the sentence should be labeled as NEGATIVE.

C. Segments where an alternative product was praised or preferred were understood to be a criticism towards the reviewed product:

(6) I will never buy another Panasonic product. There are plenty of other brands that are loyal to their customers. (both segments are *negative)*

In the second sentence the author praises competitors' brands, which makes the evaluation of the reviewed product NEGATIVE.

D. The annotators were asked to consider the sentiment of each segment in context, taking the polarity of the previous and the next segment into account.

The subset used for experiments was annotated by the author based on the guidelines above. A sample of it (including 54 sentences from 5 randomly selected reviews – 1 per each rating) was also annotated by a native speaker according to the same guidelines. To estimate the reliability of annotation we used the following two measures:

Fleiss' kappa κ[13] (0.912 for our annotation)**.** Fleiss's kappa (Fleiss, 1971) is a

[13] We use Fleiss' kappa instead of a more commonly used Cohen's kappa (Cohen, 1960) as, according to Hayes and Krippendorff (2007), it is more reliable: unlike Cohen's kappa, it satisfies the condition of reliable interpretability of its numerical scale (0 represents a situation where there is no statistical relation between annotations). While the equations for Cohen's and Fleiss' kappas appear to be similar, the difference lies in how P_e is calculated.

statistical measure of inter-rater agreement which supports multiple raters and categories and is based on Scott's pi (Scott, 1955). It is defined as $\kappa = \frac{\overline{P} - \overline{P}_e}{1 - \overline{P}_e}$, where the numerator expresses the degree of agreement achieved, and the denominator stands for the degree of agreement that can be possibly achieved above chance. The agreement scale is from 0 (no agreement) to 1 (perfect agreement), and according to Landis and Koch (1977) $0.81 < \kappa < 1$ means almost perfect agreement between annotators.

Krippendorff's alpha α (0.913 for our annotation). As the sample we used for the reliability study was small, we also validated it using Krippendorff's alpha, which is not sensitive to the sample size and thus well-suited for small data sets (Krippendorff, 2004). In its general form, $\alpha = 1 - \frac{D_o}{D_e}$, where D_o is the observed disagreement between annotation labels and D_e is the disagreement which can be expected by chance[14]. A value of $\alpha \geq 0.08$ is required for annotation to have an acceptable level of agreement (Krippendorff, 2004).

As can be seen from the values above, the reliability study showed a high inter-annotator agreement, which lets us conclude that the annotation is reliable. All cases of disagreement between annotators involve objective sentences, for example:

(7) The long wooden handle has a curve to it if you look down its length. (in a review of a brush)

This sentence, annotated as objective by one of the annotators and as negative

[14] The details of calculation can be found in (Krippendorff, 2007).

by another, is a typical example showing the inherent difficulty of distinguishing between real facts (objective sentences) and polar facts (implicit opinions).

Overall, the annotated set consisted of 988 segments (sentences or clauses). The distribution of negative, positive and objective segments in the data is shown in Table 2. There is a major skew towards negative labels, so we can expect that for the fine-grained task the majority class voting, which classifies all segments as *negative*, will achieve the accuracy of 52.2%. On the other hand, objective sentences are underrepresented and thus can be expected to be difficult for machine learning.

Negative	**Positive**	**Objective**
516	309	163
52.2%	31.3%	16.5%

Table 2. Distribution of negative, positive and objective segments in annotated data

4.2 Features Used for Experiments

In this section, we describe three sets of features (non-discourse, generic discourse and genre-specific discourse features) extracted from the corpus for our experiments. For each feature we give a brief definition and the reason for its usage, and then provide the details of its extraction.

4.2.1 Non-discourse Features

We use two types of features not related to discourse: lexical and global polarity features. Though global polarity features were treated by some researchers as discourse ones, we believe they are closer to metadata than to linguistic features and thus regard them as basic non-discourse features used to determine polarity.

4.2.1.1 Lexical Features

Lexical features have always been considered to be an indispensable basis for sentiment analysis: the majority of opinion mining studies use some kind of lexical resources with polarity annotation, whether compiled by hand (Wilson et al., 2005) or automatically extracted from a large corpus (Hu and Lui, 2004). Some tools that use lexical resources to determine the semantic orientation of a sentence or a clause have also been developed for research and commercial purposes, the first one being OpinionFinder[15] (Wilson et al., 2005).

In this paper, to determine the lexical polarity of each segment we use a state-of-art lexical classifier – Stanford Sentiment Analysis Classifier from Stanford CoreNLP toolkit[16]. This classifier, based on a Recursive Deep Learning Model, takes into account semantic compositionality and scope of negation, and thus achieves 80.7% accuracy of fine-grained prediction on the original dataset (Socher et al., 2013). It considers only lexical features available in a particular segment without looking at neighboring sentences or discourse cues, so we use it as one of non-discourse baselines.

Stanford Sentiment Analysis Classifier assigns each segment one of five labels

[15] Available from http://mpqa.cs.pitt.edu/opinionfinder/

[16] Available from http://nlp.stanford.edu/sentiment/code.html

(*very negative*, *negative*, *neutral*, *positive*, *very positive*). In our study, however, we use only three classes, so we merge very negative/negative and very positive/positive classes (into *negative* and *positive* classes accordingly). Neutral segments are the ones the classifier failed to classify as either positive or negative, which makes them *objective* in our framework.

4.2.1.2 Global Polarity Features[17]

We use review scores (star ratings) to predict their primary semantic orientation (polarity). Review scores have been used to "even out" incorrect lexical-level predictions by a number of researchers (Yang and Cardie 2014, McDonald et al., 2007). The intuition behind this is that the reviews with a higher rating will contain more positive sentences than reviews with a lower score, and thus global polarity information might help us to amend incorrect predictions of a lexical classifier. This is supported by the statistics of our corpus: the polarity of sentences in a review in general correlates with its score. As can be seen from Table 3, highly positive (5-star) and highly negative (1-star) reviews contain few segments of the opposite polarity, and even reviews with a less extreme score demonstrate a clear preference of one of the polarities. Thus it can be predicted that the classifier using this feature will tend to assign the primary polarity (positive for 4- and 5-star reviews, negative for 1-, 2-, and 3-star reviews) unless there is some strong evidence against it.

[17] Global polarity features are used only for sentence-level classification, as for document-level task they are dependent variables, not predictors.

Review score	Positive	Negative	Objective
1	0.01	0.85	0.13
2	0.10	0.77	0.12
3	0.22	0.65	0.13
4	0.62	0.23	0.15
5	0.68	0.04	0.27

Table 3. Percentage of positive, negative and objective segments in reviews with different ratings

4.2.2 Generic Discourse Features

Generic discourse features have been widely studied and used previously, so we will only summarize them in brief.

4.2.2.1 Implicit Coherence

Implicit coherence does not have any lexical cues and is realized through adjacency or proximity (Kanayama and Nasukawa, 2006; Pang and Lee, 2004): two consecutive sentences are supposed to bear the same sentiment unless there are some adversative expressions (such as contrast markers). To capture implicit coherence, we also determine the lexical polarity of the previous and the next sentence (if they exist) and use the sequence of {previous_polarity, current_polarity, next_polarity} as a feature (a similar approach is taken by Somasundaran (2010)). This is done to disambiguate and, if necessary, to correct

the polarity of misclassified instances that are sandwiched between the correctly classified ones. For example, if the lexical classifier fails to detect an implicit opinion in a sentence that appears between two explicit opinions, it might correct it as follows:

POSITIVE OBJECTIVE POSITIVE ->

POSITIVE POSITIVE POSITIVE

4.2.2.2 Continuity

Continuity features also indicate that the polarity remains the same. It is an umbrella term that covers such RST relations as volitional/non-volitional cause, volitional/non-volitional result, conjunction, joint, list and other relations which show that there is no shift in polarity. They are detected by such common cues as *and*, *so*, *moreover*, *because* etc. Segments connected by this relation in general have the same sentiment. Below we describe two main relations included into this group.

Cause

This is an umbrella term for such RST relations as *Volitional cause*, *Non-volitional cause*, *Volitional result*, *Non-Volitional result* and *Evidence* (Zhou et al. (2011)). As *Cause* includes semantically opposite relations (such as *cause* and *result*), we remove the constraints on the nucleus and the satellite and generalize the constraint on their relation as "There must be a cause-effect relation between the nucleus and the satellite".

Though not all causal relations are expressed explicitly, in most studies, including this one, they are determined by the presence of such explicit cues as *so*, *because*, *thus* etc. These cues can appear in the nucleus (*so*) or in the satellite (*because*):

(8) [After much work we could not get this item to work.]$_{satellite}$ [**So** we returned it.]$_{nucleus}$

(9) [**Because** after much work we could not get this item to work,]$_{satellite}$ [we returned it.]$_{nucleus}$

Table 4 lists *Cause* cues used in our experiments and shows if they appear in the nucleus or satellite.

In terms of sentiment, both satellite and nucleus have the same polarity, and the satellite includes supporting evidence for the opinion expressed in the nucleus. An example of *cause* relation is shown in Figure 2.

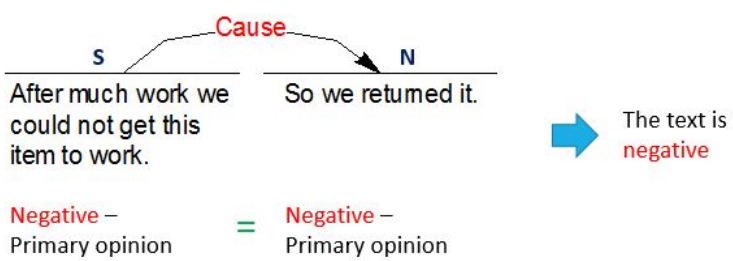

Figure 2. An RST scheme for the generalized *Cause* relation

In this example, Stanford classifier fails to label the second sentence as negative, because it lacks any sentiment cues. However, the presence of *so* shows

that these two sentences are connected by *Cause* relation and thus must have the same polarity. Also, disambiguating the polarity of the nucleus allows to classify the whole review as negative.

Cue	Nucleus/Satellite
accordingly	nucleus
as	satellite
as a result	nucleus
because	satellite
consequently	nucleus
hence	nucleus
since	satellite
so	nucleus
that's why	nucleus
therefore	nucleus
thus	nucleus

Table 4. Discourse markers for *Cause* relation

Conjunction

This relation includes not only *Conjunction* itself, but also other RST relations such as *Disjunction*, *Joint, List* and *Sequence*.

It can be determined by the presence of such cues as *and, also, moreover* etc. (more examples can be found in Table 5).

again	first	nor
also	firstly	on top of that
and	further	or
another	furthermore	second
as well	in addition	secondly
besides	likewise	too
finally	moreover	what is more

Table 5. Discourse markers for *Conjunction* relation

This is a multinuclear relation, and the cue appears in the second nucleus. From the point of view of sentiment analysis, both nuclei express the same opinion and thus have the same polarity:

(10) Amazing paper saving options. **Also**, a very helpful automatic on-screen error guide.

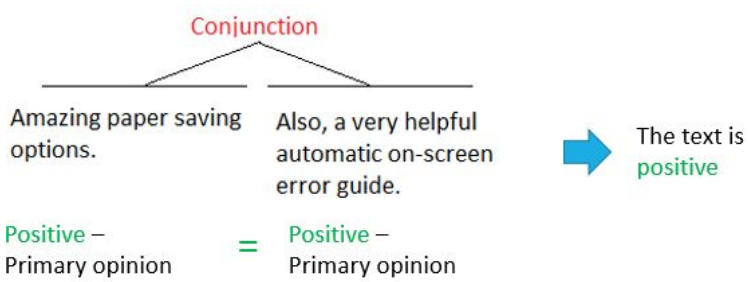

Figure 3. An RST scheme for the generalized *Conjunction* relation

In the example in Figure 3, Stanford classifier incorrectly assigns negative polarity to the second sentence (probably because of the negative word *error*), but

the *Continuity* cue also shows that it should have the same sentiment as the previous sentence, and thus its polarity is corrected.

Extraction of *Continuity* Relations

As explained above, we add the *Continuity* feature only to the relations that have an explicit cue, in hope that relations lacking any cues will be captured by *Implicit Coherence* feature. To detect *Continuity* relations, we use regular expression matching for the corresponding cues, add *Continuity* feature to the matched clause, and then, depending on whether the cue appears in the satellite or the nucleus, add it to the previous or next segment.

4.2.2.3 Explicit Contrast

Explicit contrast features, unlike *Continuity* features, indicate the change of polarity. Again, it is a simplified relation including such RST relations as *Concession, Antithesis, Otherwise* (Zhou et al., 2011)[18]. The common discourse markers here are *but, although, however* etc., and, depending on their type, they can appear either in the nucleus (which expresses the opinion matching the primary polarity of the text) or in the satellite (which includes the opposite opinion). We

[18] Unlike Zhou (2011), we do not include the RST *Contrast* relation here, as it is a multinuclear relation in which both opinions are equally important (the relation is also called *Neutral Contrast* for this reason):

I like IPhone but my brother hates it.

As the sentence has multiple opinion holders, we cannot say that the opinion expressed in it can be unambiguously summarized as positive or negative.

add *Explicit contrast* feature to the satellite segment expressing an opposite opinion:

A. The discourse marker is in the nucleus (*but*) type:

(11) The Phillips screwdriver on the end of one of the tines is helpful for things like tightening eyeglasses, POSITIVE CONTRAST
but it is slightly offset from the opposing blade and I've nicked or jabbed myself with it more than once while it's in my pocket. NEGATIVE NCONTRAST

B. The discourse marker is in the satellite (*although*) type:

(12) **Although** it has 10 workable buttons which come in handy for some games, POSITIVE CONTRAST
it has some major flaws. NEGATIVE NCONTRAST

Table 6 lists *Explicit contrast* cues we used and shows if they appear in the nucleus or satellite.

Figure 4 demonstrates how this relation can be used to correct a misclassified opinion in the following review:

(13) This tape shows you some really great exercises **but** I found that it took some time to see real results.

Cue	Nucleus/ Satellite	Cue	Nucleus/ Satellite
all the same	nucleus	nevertheless	nucleus
although	satellite	nonetheless	nucleus
anyway	nucleus	notwithstanding	nucleus
at any rate	nucleus	on the contrary	nucleus
at the same time,	nucleus	regardless	nucleus
despite that	nucleus	still	nucleus
even so	nucleus	that said,	nucleus
even though	satellite	though	satellite
for all that,	nucleus	while	satellite
however	nucleus	yet	nucleus

Table 6. Discourse markers for *Explicit Contrast* relation

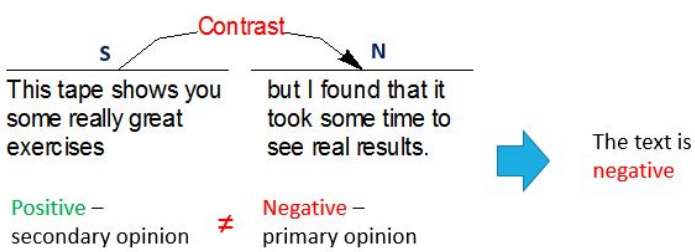

Figure 4. An RST scheme for the generalized *Explicit Contrast* relation

Stanford classifier misinterprets both clauses as positive because of similar lexical cues (*really, real*), and the negative sentiment in the second one cannot be detected because of its complicated structure. However, the contrast marker *but*

shows that one of the clauses is positive. Moreover, *but* appears in the nucleus, which should reflect the primary polarity of the text, and as this review is negative (indicated by the *Global polarity* feature), the second clause can be disambiguated as negative.

From a sociolinguistic perspective, such concessive conjunctions as *but* are often used to 'hide' a complain, preceding it with a positive statement to position oneself as a reasonable, objective person (Vasquez, 2011). As was first noted by Sacks (Edwards, 2005), complaints are often structured according to the following pattern: positive opinion – *but* – negative opinion, and thus we can expect *Contrast* relations to be widely used in negative reviews.

4.2.3 Discourse Features Specific to Online Reviews

Though there is a large number of discourse features that are specific to online reviews, such as, to name just a few, addressing the reader in the second person to build rapport, using professional words to prove one's expertise, using present tense to make categorical claims and using past tense sequences to construct a narrative (Vasquez, 2014), in this paper we introduce only those which, according to our experiments, are reliably helpful for mining opinions in online reviews.

4.2.3.1 Implicit Contrast

Implicit contrast features in online reviews are often realized through the use of hedges – words or phrases that mitigate the impact of an utterance. These are mostly such limiting expressions as *one good point, only drawback, would only*

recommend etc, and thus we detect these features by the presence of *only* and *one*. While hedges are widely used in texts of all genres for different purposes such as expressing politeness or avoiding responsibility (Markkanen R. and Schröder, 1997b), in online reviews these patterns indicate that the author wants to mention some negative side of a product they like (or a positive aspect of a product they hate) without an unnecessary emphasis. Though it is problematic to define such expressions as hedges from a traditional point of view (as modifiers of the writer's responsibility for the truth value of propositions (Markkanen R. and Schröder, 1997a)), in terms of sentiment analysis they behave as weakeners of the writer's opinion. Compare the following two reviews (the second one is taken from the corpus, the first one is modified to illustrate our point):

(14) A rather sappy love story is the center of the movie and there's very little suspense. **But** the redeeming feature is Alyssa Milano's performance as Lily.

(15) A rather sappy love story is the center of the movie and there's very little suspense. The **only** redeeming feature is Alyssa Milano's performance as Lily.

In the first review – the one with an explicit contrast – the emphasis is on the second opinion (the opinion in nucleus), which makes the overall review positive. The first, hedged, review is negative even though there is a positive assessment of one of the aspects of a movie. In fact, *only* weakens the positive assessment in such a way that it makes the negative opinion of the movie even stronger. Thus, in terms of RST relations the segment with the hedge is a satellite and does not affect the

overall opinion. Figure 5 shows how this principle helps to disambiguate sentiments at the document level:

(16) I think I finally feel that it's worth the spending to buy my first mac. My **only complain** is that it's still a lot more expensive than PC laptops with similar specs.

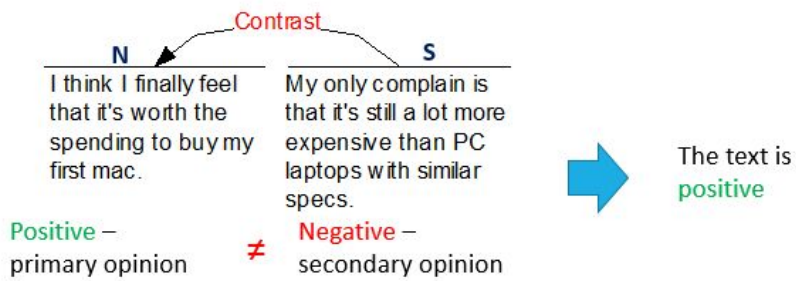

Figure 5. An RST scheme for the generalized *Implicit Contrast* relation

Though the second sentence is negative (which is correctly detected by the lexical classifier), thanks to hedging it does not affect the overall polarity of the text, and thus the review can be correctly classified as positive.

From the sociolinguistic viewpoint, such patterns are considered to be extreme case formulations (ECFs) (Pomerantz, 1986), and they were shown to be used while complaining (Edwards 2000, 2005) to legitimize one's claims and present one's opinion as well-grounded and objective. Thus we can predict that they will appear in negative reviews or positive reviews where the writer has a minor complaint about some aspect of a product.

4.2.3.2 Background Features

Background features represent the RST relation of *Background* and satisfy all constraints of this relation[19]:

constraints on the nucleus: the reader cannot comprehend the nucleus sufficiently before reading the satellite;

constraints on the nucleus and the satellite: the satellite should increase the reader's comprehension of the nucleus.

According to these constraints, the writer uses the relation of *Background* to make the reader better understand the following part of a review. More specifically, the satellite of this relation contains information about the author which shows that she/he is qualified to write the review and provides some basis for the opinionated claims that follow. The satellite, to which the *Background* feature is added, is an objective sentence and thus does not affect the overall polarity of the document. Consider the following example (Figure 6):

(17) When I saw the DC25 Animal, I **decided to spend** the money hoping that this vacuum would do the job. It has lived up to my wildest dreams.

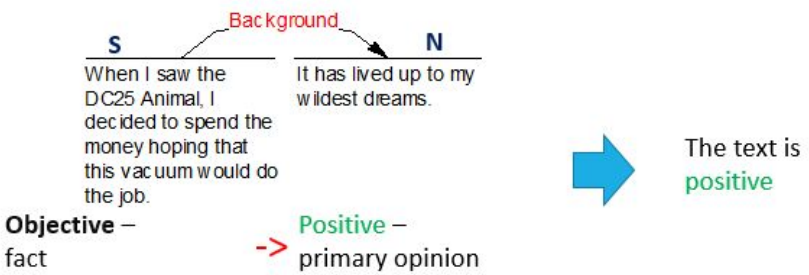

Figure 6. An RST scheme for the *Background* relation

[19] As listed on RST site http://www.sfu.ca/rst/01intro/definitions.html

Though the Stanford classifier incorrectly labels the first sentence as negative because of the lexical cue *spend money*, the presence of a *Background* feature (*decided to spend – Acquirement* pattern, see below*)* helps to correctly classify it as objective. Thus the first segment does not influence the overall opinion in the text, and the review can be correctly classified as positive.

From the sociolinguistic perspective, such patterns are used by writers to construct their identity, and, more specifically, to build their credibility and validate their opinion on the reviewed product (Mackiewicz, 2007, Vasquez, 2014). According to Mackiewicz (2010b), credibility can be built based on several assertions, such as assertion of a product-specific experience, assertion of familiarity with related products, or assertions of a relevant role. Vasquez (2014) also considers identities which help the writer gain trust of the readers, such as references to one's age, lifestyle, tastes etc. We generalize these categories and use three patterns to capture the *Background* feature. As, according to Mackiewicz (2010a), background information, which shows that the reviewer is capable of accurately evaluating the product, usually appears at the beginning of the review, we add this feature only to matching sentences in the first 25% of the text.

We use the following three patterns to extract this feature:

Acquirement Patterns

These patterns provide an explanation of the way the writer acquired the product:

(18) **I bought** this camera for my deployment to Iraq. (*objective*)

It was in my cargo pants pocket one day I took it out and the lens was cracked and the silver trim ring had fallen off. (*negative*)

We formalize this feature as follows:

[I | we] [verb synonymous to "acquire"|verb of decision + verb synonymous to "acquire"],

or, more specifically:

[I | we] [ordered | bought | got .* as a gift | purchased | decided to buy…]

We used such words as "buy", "get", "order" as seeds and added their synsets from WordNet 3.1.[20] Table 7 lists the verbs we used to match this pattern:

acquire	get	order
be given	get hold of	purchase
buy	obtain	receive

Table 7. Cues for *Acquirement* patterns

All verbs are in past simple tense, as in this tense they are unlikely to bear any sentiment. Compare, for instance, sentences with the same verbs in present perfect tense:

(19) However, I am glad that **I have bought** a mac. (*positive*)

[20] Princeton University (2010). About WordNet. <http://wordnet.princeton.edu>

(20) This is probably the worst book **I've bought**. (*negative*)

Personal Background Patterns

In these patterns, the authors disclose their personal information that is relevant to the subject matter of the review and can support their opinion. For instance, in the following review the author refers to his pets as the major reason for buying a particular vacuum cleaner:

(21) **I have a** cat and a dog, and there is lots of shedding hair, all the time. (*objective*, *Personal background*)

When I saw the DC25 Animal, I **decided to spend** the money hoping that this vacuum would do the job. (*objective*, *Acquirement*)

It has lived up to my wildest dreams, it is wonderfully easy to handle, so easy to maneuver, the 16 lbs make such a difference compared to those very heavy machines I had before, I had no problem carrying it upstairs. (*positive*)

We formalize this feature as follows:

[I|we] [am (a|an)|have (a|an)|'m (a|an)|am not (a|an)]

The indefinite article is used to prevent matching polar expressions containing positive or negative adjectives:

(22) I'm very pleased with the quality of this product. (*positive*)

Personal Experience Patterns

These patterns also serve to provide some background information about the user's experiences, achievements or expertise to back up his or her opinion on a product:

(23) Usually **I am a** huge fan of hats that look like food. (*objective*, *Personal background*).

My meatloaf hat **has been** a hit for years. (*objective*, *Personal experience*)

When I received my turkey hat I carefully unwrapped the bubble wrap and gazed upon its tan beauty. (*positive*).

To capture this pattern we search for verbs in perfect forms (except for the verbs of possession and acquirement, see Table 7). We exclude verbs in perfect continuous forms, as they are more often used to describe positive or negative results of using a product. Compare, for example:

(24) I **have been using** it for almost a month and my lashes are so long, they touch my eyebrows... (*positive*)

We also exclude phrases that have *should* or *would* before *have*, as they often express negative sentiments (Liu et al., 2014):

(25) **Would have been** nice if the stilts could accommodate multiple/varying heights. (*negative*)

4.2.3.3 Involvement Features

Involvement features do not have a corresponding RST relation, as they are of spoken (dialogical) nature, and Rhetorical Structure Theory was designed for analysis of written texts (Mann and Thompson, 1988). In sociolinguistic terms, it is one of the aspects of interactivity, which manifests itself not only in speech, but also in online discourse. By involvement we understand the ways of engaging one's audience and building rapport with it, such as using second-person pronouns (*you*), imperatives, conventional speech-act formulas (*I'm sorry*), etc. (Vasquez, 2014). Though we tried using many of these features to aid sentiment classification, most of them turned out to be too noisy (as in case of second-person pronouns whose function depends on the context) or difficult to extract reliably (as in case of imperatives). In this study we use questions to detect this feature, as they represent the interactional strategy in discourse (Chafe, 1982; Gumperz, 1982) and reflect quasi-interaction between the reader and the writer in reviews, which was first noted by Polanyi and Zaenen (2006). Questions are usually objective sentences, as they engage the reader without carrying any sentiment:

(26) Guess what? (*objective*)

They should be distinguished from rhetorical questions that have a strong sentiment and are often used as indirect accusations (Neurauter-Kessels, 2011):

(27) It's a waffle maker...not the Space Shuttle...how hard could it be to make it last more than 20 cycles?! (*negative*)

As distinguishing real and rhetorical questions is a difficult task in itself, in this study we have to rely on punctuation marks, excluding all sentences that end in ?!.

4.3 Feature Validation

Before using the features in the experiments we checked if they are not only justified linguistically, but are also statistically valid.

We used the following tests[21] to assess if the proposed features correlated with the sentence labels from the annotation[22]:

A) **Pearson's chi-squared test** (Plackett, 1983) was used to check if features and labels were independent. The value of p below the threshold $\alpha = 0.05$ shows that the features and labels are not independent.

B) **Contingency coefficient** and **Cramer's V** (Cramér, 1946) were used to measure the degree of association between the features and sentence labels (the values range from 0 to 1 with 1 indicating perfect associations).

[21] All tests were performed in R using VCD package.

[22] For three of the features – *Explicit Contrast*, *Implicit Contrast* and *Continuity* – we checked for correlation with a supplementary feature *same* or *different polarity* (*same* feature was assigned if the current and the previous segments had the same annotation labels, *different* feature was assigned otherwise), because they represent change or continuation of polarity rather than polarity itself.

Features	Chi-square	Contingency coef.	Cramer's V
Lexical	p < 2.2e-16	0.42	0.328
Global polarity	p < 2.2e-16	0.562	0.481
Continuity	p = 0.1099	0.072	0.072
Explicit contrast	p <3.892e-12	0.242	0.25
Implicit contrast	p < 2.2e-16	0.35	0.374
Acquirement	p < 2.2e-16	0.324	0.324
Personal backgr.	p = 1.411e-10	0.226	0.232
Personal exper.	p = 6.609e-11	0.23	0.236
Involvement	p = 7.354e-148	0.259	0.268

Table 8. Degree of correlation between features and labels

The results of the correlation tests are shown in Table 8 above. As can be seen from the table, all features, with the exception of *Continuity*, are not independent from the labels and correlate with them. The strongest correlation is between the review's score (*Global polarity*) and the sentence labels, which lets us predict that this feature will be particularly useful. In fact, *Global polarity* correlates with the sentence labels better than the *Lexical* labels from Stanford's Sentiment Classifier. Another surprising result is that *Implicit contrast* has a stronger correlation with the sentence-level annotations than *Explicit contrast,* and thus none of the generic discourse features correlate strongly with the dependent variable. This allows us to hypothesize that genre-specific features will have a more noticeable effect on the accuracy of sentiment prediction than the generic ones.

In the following chapters we use these features in two series of experiments to verify whether this conclusion is valid.

5. Predicting Sentence Polarity Using Discourse Features

5.1 Experiment Setup

In this set of experiments, we aim to classify sentiments at the fine-grained level, that is, to predict whether a particular clause or sentence is negative, positive or objective. The experiments are divided into two parts: in the first experiment we performed "classic" sentiment analysis relying on *Lexical* features, while in the second one we removed *Lexical* features and use only *Global polarity* and discourse features (both generic and genre-specific). We also conducted a preliminary study, adding features one at a time to the lexical baseline to check if each of them has an individual effect on the accuracy of predictions.

While such machine learning algorithms as Naïve Bayes or Support Vector Machines (SVMs) are still the primary tools for sentiment analysis, lately such texts as online reviews have been recognized as having an internal structure and inter-sentential relations, and thus structural conditional frameworks are nowadays used for their classification. One popular tool is Conditional Random Fields (CRF), which was used, among others, by Zhao (2008) to classify sentiments on a sentence level, by Breck (2007) to identify subjective expressions, and by Li et al. (2010) to summarize product reviews taking their structure into account. Following this trend, for the fine-grained classification we treat the sentiment analysis problem as a sequence labelling task, modelling each review as a sequence of opinions and thus discarding the bag-of-opinions model of non-structural frameworks.

In Conditional Random Fields (CRF) method (Lafferty et al., 2001) the probability of a sequence is defined as follows:

$$p_\lambda(Y \mid X) = \frac{\exp \lambda \cdot F(Y, X)}{Z_\lambda(X)}$$

where

$$Z_\lambda(x) = \sum_y \exp \lambda \cdot F(y, x)$$

where X is a set of input random variables, Y is a set of labels, and λ is a weight for the feature function $F(Y,X)$ (Sha and Pereira, 2003).

For this task we used a C++ implementation of a linear Conditional Random Fields classifier (CRF++)[23]. Though more complex or constrained types of CRF classifiers proved to be more efficient and suitable for sentiment analysis (Mao and Lebanon, 2006; Yang and Cardie, 2014), in this study we use the simplest model as a proof of concept.

Each review in the training and test set is converted into a sequence of polarity labels assigned to it. For example, the following short review:

(28) The ring is nice and heavy. Have been wearing if for almost a month and still not a scratch!

is represented as a sequence of target tokens POSITIVE POSITIVE, based on the sentiment labels from the annotation. Each target token is assigned features, as was described in the previous chapter, which are then fed into the classifier.

[23] Available from http://taku910.github.io/crfpp/

5.2 Evaluation of Experiments

5.2.1 Measures

To evaluate the results of classification, we use the standard set of measures borrowed from information retrieval – precision, recall, F1 and accuracy (Manning et al., 2008):

Precision is defined as a number of correct guesses for a given class (true positives) divided by the total number of guesses for this class (including true and false positives):

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

On the other hand, **recall** shows the proportion of the correct guesses to the total number of correct labels (including true positives – correct guesses – and false negatives, which were assigned an incorrect label):

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

F-measure is a weighted harmonic mean of precision and recall, and in this study we use its standard variant – the balanced F-measure, which assigns the same weights to precision and recall (F1):

$$F1 = \frac{2PR}{P + R}$$

Lastly, **accuracy** denotes the fraction of predictions which are correct:

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

All results reported for the experiments are averaged across all folds of the 6-fold validation.

5.2.2 Results

5.2.2.1 Preliminary Classification

Before the main experiments we conducted a preliminary study to check if each of the features had a significant effect on the accuracy of classification. We added the features one at a time to the lexical baseline to check if it will result in some improvement. The results of the preliminary experiment are shown in Table 9:

Features	Accuracy
Lexical	0.6138
Lexical + Global polarity	**0.7172**
Lexical + Implicit coherence	0.6758
Lexical + Continuity	0.6413
Lexical + Explicit Contrast	0.6137
Lexical + Implicit Contrast	0.6413
Lexical + Background	**0.6896**
Lexical + Involvement	0.6482

Table 9. The gain in accuracy achieved by adding individual features to the lexical baseline

As can be seen from the table, all features except for *Explicit Contrast* show an improvement over the lexical baseline. The strongest feature is *Global polarity*, followed by *Background* patterns. We can expect these two features to have a

major influence on the results of the classification, while the other features might have an inferior performance.

5.2.2.2 Classification with Lexical Features

For this experiment we used all features defined in the previous chapter, adding them one by one to the lexical baseline and all previously used features. The results of the experiment with lexical features are summarized in Table 10. As no features were removed when adding the new ones, the last row (+*Involvement*) represents the results for the full set of features. The numbers in bold indicate the best results for the corresponding measure.

	Subjective				**Objective**		**Total**			
	Negative		**Positive**							
	Prec	**Rec**	**Prec**	**Rec**	**Prec**	**Rec**	**Prec**	**Rec**	**F1**	**Acc**
Lexical	0.71	0.77	0.61	0.54	0.29	**0.27**	0.60	0.61	0.61	0.6138
+ Global	0.77	0.88	0.64	**0.85**	0	0	0.59	0.72	0.65	0.7172
+ Coherence	0.73	0.88	0.61	0.73	0	0	0.56	0.68	0.62	0.6827
+ Continuity	0.74	**0.90**	0.61	0.73	0.50	0.04	0.66	0.70	0.64	0.6965
+ Ex. Contr.	0.74	**0.90**	0.64	0.73	0.50	0.08	0.67	0.70	0.65	0.7034
+ Im. Contr.	0.77	0.87	0.62	**0.85**	**1**	0.04	0.77	0.72	0.66	0.7172
+ Backgr.	**0.79**	0.87	0.66	**0.85**	**1**	0.23	**0.79**	0.75	0.72	0.7517
+ Involvem.	**0.79**	0.87	**0.67**	**0.85**	**1**	**0.27**	**0.79**	**0.76**	**0.73**	**0.7586**

Table 10. Precision, recall, F1 and accuracy scores for fine-grained classification with lexical labels

As can be seen from Table 10, the full feature set ensures the best results for all measures except for the recall of negative segments. Using all discourse features leads to improvement of total F1 score by 12% and of accuracy by over 14%. However, if we compare the results by groups (non-discourse, generic discourse and genre-specific discourse features), it becomes clear that generic discourse features (in white) actually have a negative effect on the results – adding them makes the results worse compared to the results achieved by non-discourse features (in light gray): F1 score remains the same, but the accuracy drops by over 1%. This is due to the addition of *Implicit Coherence* feature, which degrades both F1 and accuracy by 3%. On the other hand, genre-specific discourse features lead to the relative improvements of 8% and 5% for F1 score and accuracy respectively, the best performing feature being *Background* (relative gain of 6% and 4% respectively).

In general, it can be noticed that the performance for negative segments is higher than for positive ones, and classification of objective segments is performed much worse than that of other classes. This is in line with our predictions based on the distribution of labels in the annotated data (see section 4.1.2). General results from Table 10 are more clearly presented in Figures 7 to 9 below, which show F1 scores for negative, positive and objective segments separately.

For negative sentences, generic discourse features have a worse performance than the one achieved by using only lexical features and *Global polarity* (see Figure 7). It can be explained by the fact that the *Implicit coherence* feature tends to overapply the polarity of neighboring sentences. However, adding other generic and genre-specific discourse features helps to steadily improve the performance over the non-discourse baseline.

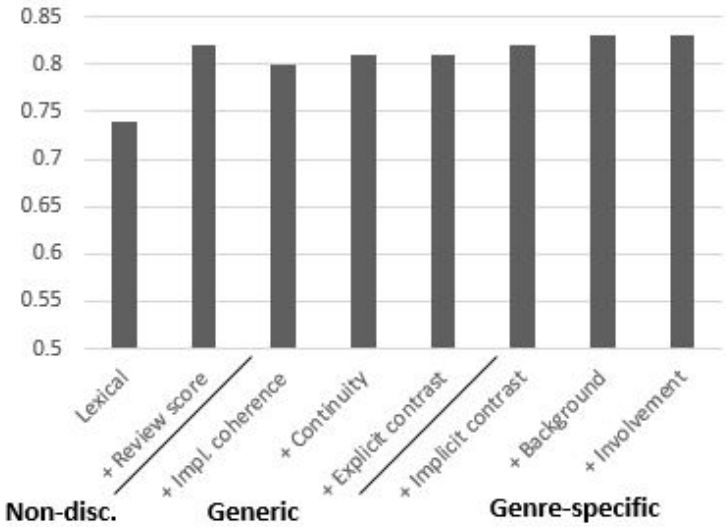

Figure 7. F1 scores for negative sentences

The same trend appears in Figure 8 for positive sentences:

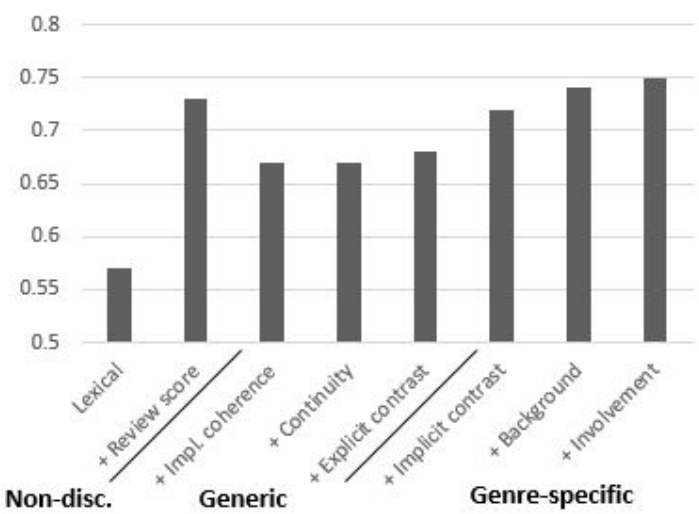

Figure 8. F1 scores for positive sentences

The performance drops when the *Implicit Coherence* is introduced, and the improvement over the *Global polarity* is achieved only when the *Background* feature is added. Moreover, *Continuity* does not have any effect on classification of positive sentences. On the other hand, genre-specific discourse features again show a steady improvement.

However, the influence of genre-specific features can be most clearly seen in Figure 9, which represents objective sentences. As the objective sentences do not have a strong correlation with the review's score (though they tend to be used more in negative reviews), adding the review score feature completely removes them, and generic discourse features do little to correct this. Moreover, the gain in recall achieved by using *Explicit Contrast* is annulated by adding the *Implicit Contrast* feature. However, the other genre-specific features perform much better. They are strong predictors of objectivity that help us to classify objective sentences with almost perfect precision, though a somewhat low recall.

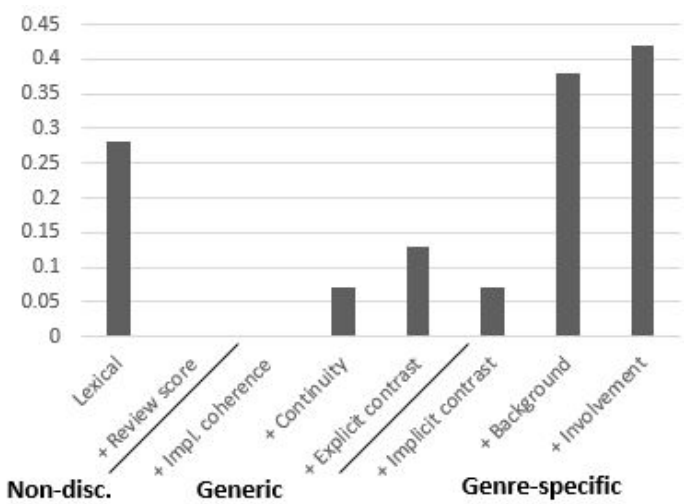

Figure 9. F1 scores for objective sentences

The overall effect of generic and genre-specific discourse features is shown in Figure 10. It confirms the overall trends we outlined above:

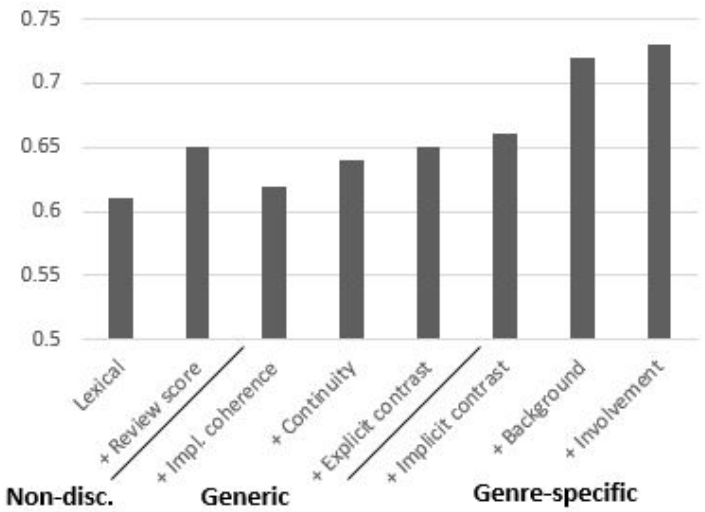

Figure 10. Average F1 scores

5.2.2.3 Classification without Lexical Features

The trends outlined in the previous section can be seen even more clearly if we remove all lexical features and use only review ratings and discourse features for classification. In this experiment we use *Global Polarity* as a baseline and add generic and genre-specific discourse features to it. As *Implicit coherence* feature is a sequence of the previous, current and next lexical labels, we cannot use it for this experiment and thus do not report any results for it. Also, there is no sense in replacing it with the sequence of {previous_Global_polarity, current_Global_polarity, next_Global_polarity}, as *Global polarity* for all segments in a review is the same. We add the features to the baseline in the same

way as in the previous experiment: the final row in the results table represents the performance of the whole feature set.

As can be seen from Table 11, *Global polarity* feature alone helps achieve better average F1 and accuracy than the combination of Lexical and Global polarity features (see section 5.2.2.2) Thus removing lexical features helps us to reduce noise and leads to better results overall:

	Subjective				**Objective**		**Total**			
	Negative		**Positive**							
	Prec	**Rec**	**Prec**	**Rec**	**Prec**	**Rec**	**Prec**	**Rec**	**F1**	**Acc**
Global	0.75	**0.94**	0.71	0.83	0	0	0.61	0.74	0.66	0.7379
+ Continuity	0.75	**0.94**	0.71	0.83	0	0	0.61	0.74	0.66	0.7379
+ Ex. Contr.	0.77	0.92	0.69	0.85	0	0	0.61	0.74	0.67	0.7379
+ Im. Contr.	0.80	0.91	0.68	**0.93**	0	0	0.62	0.75	0.68	0.7517
+ Backgr.	**0.82**	0.91	**0.75**	**0.93**	**1**	**0.27**	**0.83**	**0.80**	**0.77**	**0.8**
+ Involvem.	**0.82**	0.91	**0.75**	**0.93**	**1**	**0.27**	**0.83**	**0.80**	**0.77**	**0.8**

Table 11. Precision, recall, F1 and accuracy scores for fine-grained classification without lexical labels

However, the major drawback of using *Global polarity* as a baseline is that it does not assign any objective labels, so the recall and precision for the objective class are zero. This is corrected only when *Background* features are introduced: the *Background* patterns help to retrieve objective segments with a perfect precision, though their coverage (and thus recall) is much lower than of the other features. Thus in this experiment *Background* proves to be a strong feature. On the other

hand, the other objectivity feature – *Involvement* – does not ensure any additional gain.

The generic discourse features used here also do not lead to any (in case of *Continuity*) or almost any (in case of *Explicit Contrast*) improvement in performance: the total increase in F1 after adding all generic discourse features is 1%, while accuracy stays the same. On the other hand, genre-specific discourse features improve the F1 score by 10% and accuracy by over 6%.

Figure 11 shows F1 scores for all classes and helps to better visualize the trends we described above:

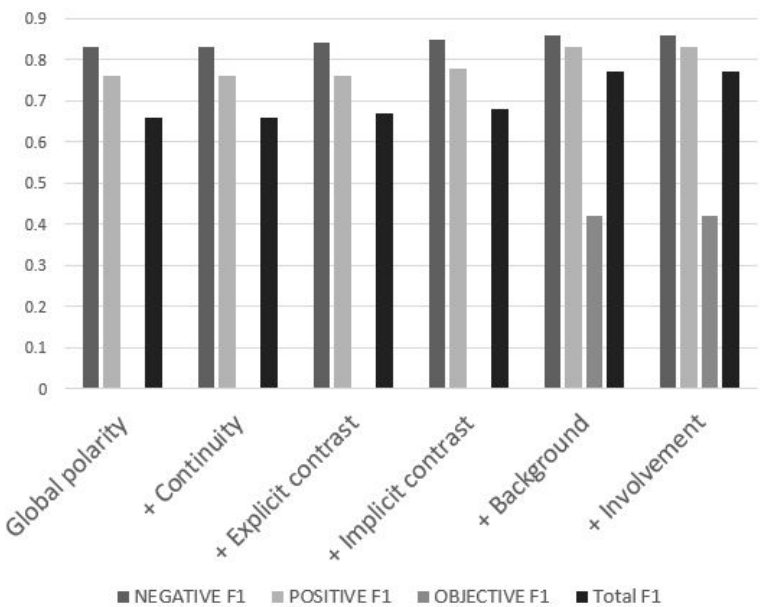

Figure 11. F1 scores for sentence-level classification without lexical features

5.3 Discussion

As was shown above, generic discourse features tend to be more noisy and less effective when it comes to sentiment analysis. In particular, there are two

features that proved to have a lower performance than we expected: *Implicit Coherence* deteriorates the results of classification with lexical features, while *Explicit Contrast* performs worse than *Implicit Contrast:*

The *Implicit Coherence* caused a major drop of precision and recall for both positive and negative labels. It seems that this feature tends to overapply the polarity of neighboring sentences and thus ignore the shifts in sentiment flow.

Though it intuitively appears that *Explicit Contrast* should perform better than *Implicit Contrast*, this feature is inherently noisy, as overt contrast markers such as *but* can indicate other relations except for *Contrast*. In fact, according to Mann and Thomson (1988), there are no reliable and unambiguous signals for any of RST relations.

Even if *but* and similar cues do indicate *Contrast*, in some cases it is not the contrast between positive and negative opinions, but the contrast between expectations and reality. For example, a satellite connected to a negative nucleus by the relation of *Contrast* is not always positive: it can be an objective sentence expressing the writer's expectations before using the product:

(29) After having devoured The Girl With The Dragon Tattoo (Vintage) I was very much looking forward to The Girl Who Played With Fire. (objective)
But somehow the story did not captivate me as much as the first Lisbeth Salander / Mikael Blomkvist book. (negative)

The first sentence does not express an evaluation of The Girl Who Played With Fire: it simply shows the author's expectations, and thus is objective. However, our classifier misclassifies it as positive due to the contrast relation with

the following negative sentence.

Another result that deserves some attention is that *Lexical* features perform worse than more coarse-grained sentiment information, such as product ratings. Not only is the lexical baseline lower than the *Global Polarity* one, the discourse features added to it show less stable improvements. It appears that they are not able to generalize well because of noise introduced by *Lexical* features.

6. Predicting Review Ratings Using Discourse Features

6.1 Experiment Setup

In the second set of experiments, we aim to predict document ratings using lexical and discourse features. Again, we conduct two experiments: in the first one we perform 5-class (1 to 5 stars) classification and demonstrate how generic and genre-specific discourse features influence its results. In the second one we vary the number of classes, comparing the results of 2-, 3 and 5-class prediction to show that discourse features are especially important for detailed (multi-label) classification.

As document-level classification cannot be cast as a sequence labelling task (each document has only one rating, so we need to use a bag-of-features approach to predict one label per document), we cannot rely on a CRF classifier here. We use a multinomial Naïve Bayes model, which, unlike Support Vector Machines and other commonly used methods, natively supports multi-class classification. Though predicting review ratings has been often treated as a regression or a ranking task (Ganu et al., 2009), for simplicity and clarity of interpretation we treat review scores as class labels and not as ordinal or continuous variables.

In a Naïve Bayes model a review *r* is assigned a rating (class) *c* satisfying the following formula (Manning et al., 2008):

$$c = \arg\max_{c \in C} P(c \mid r) = \arg\max_{c \in C} P(c) \prod_{1 \le k \le n} P(f_k \mid c),$$

where C is the set of classes (star ratings) a review can have, *P(c)* is the prior probability of a review having a particular rating[24], f_k is the k-th feature in the review and *n* is the total number of features. The prior and posterior probabilities are estimations based on the training set. For both experiments we used a multinomial Naïve Bayes classifier from *skikit-learn* package for Python.

Each review is represented as a bag of features which were described in Chapter 4, together with their counts. *Global Polarity* feature is omitted, as it is a dependent variable which has to be predicted in this series of experiments. Unlike the experiments presented in Chapter 5, here we do not have reliable non-discourse features that we can use instead of lexical ones, so we have to rely on Stanford labels. The manually annotated labels were not used in this set of experiments.

In the next section we outline the results of experiments averaged across 6 folds during cross-validation. For estimation we use the same information retrieval measures (precision, recall, F1 score and accuracy) as explained in Chapter 5.

[24] The prior probability can be ignored in our experiments as the data set is well balanced and each rating is equally probable for any review.

6.2 Experiment Results

6.2.1 5-class Classification

In the first part of document-level experiments, we tried to predict review ratings on the scale of 1 to 5 stars. Because it is a five-class classification, the random baseline for such an experiment would have an accuracy of 20%, and a bag-of-words classifier using Naïve Bayes model does not perform much better, achieving accuracy of 24%.

This baseline is improved by 7% if instead of a bag of words we use lexical labels from a more linguistically-informed Stanford classifier (Table 12):

	Precision	**Recall**	**F1**	**Accuracy**
1 star	0.287	0.292	0.289	
2 stars	0.208	0.25	0.227	
3 stars	0.37	0.417	0.392	
4 stars	0.122	0.125	0.123	
5 stars	0.49	0.5	0.495	
Average	0.295	0.317	0.305	**0.317**

Table 12. Predicting review ratings with lexical features

Adding generic discourse features improves the accuracy by 3%. Their influence is especially remarkable for 1-star reviews, for which the F1 score

increases by 6.3%, and for 4-star reviews, which were poorly classified by the Stanford classifier and where the increase equals 6.4% (Table 13):

	Precision	**Recall**	**F1**	**Accuracy**
1 star	**0.332**	**0.375**	**0.352**	
2 stars	0.222	0.25	0.235	
3 stars	0.403	0.375	0.389	
4 stars	**0.213**	**0.167**	**0.187**	
5 stars	0.478	0.542	0.508	
Average	**0.33**	**0.342**	**0.334**	**0.342**

Table 13. Predicting review ratings with lexical and generic discourse features

However, genre-specific features seem to have even more evident effect on review's score, further improving the overall accuracy by 6 percent (Table 14). The most remarkable improvement is for 1-star reviews, where F1 increases by almost 17% thanks to the doubled recall and increased precision. The classifier also performs better for the "vague" classes – 2 and 4-star reviews, which are harder to classify than the extreme 1-star and 5-star classes and thus have very low lexical baseline scores. For 2 star reviews, F1 score improves by almost 18%, and it increases by 3.6% for 4 star reviews. Overall, the precision of prediction for all classes improves by 12.3% compared to the lexical baseline, while the gain in recall reaches 9.1%.

	Precision	Recall	F1	Accuracy
1 star	**0.412**	**0.708**	**0.521**	
2 stars	**0.545**	**0.333**	**0.414**	
3 stars	0.445	0.292	0.352	
4 stars	**0.24**	**0.208**	**0.223**	
5 stars	0.45	0.5	0.474	
Average	**0.418**	**0.408**	**0.397**	**0.408**

Table 14. Predicting review ratings with lexical, generic and genre-specific discourse features

6.2.2 Comparison of Results of 2, 3 and 5-class Classification

As discourse features have been primarily used for sentence level classification and, to the best of our knowledge, have not been employed for document-level sentiment analysis, one may wonder if they are necessary for such task. As can be seen from Figure 12, they are indeed not needed for the two-class ('thumbs up/thumbs down", "negative/positive") classification, where high accuracy (0.83) can be achieved using only lexical features and the discourse information does not improve it. However, for classification with 3 labels, which includes the middle class (3 star reviews) and two extreme classes (1 and 5-star reviews), discourse features have a statistical effect on the performance. Genre-specific features are especially important here: though generic discourse features improve the accuracy of the lexical classifier (0.75) only by 1%, the genre-specific ones ensure an accuracy increase of 7%. The same tendency can be seen in 5-class

classification we described in more detail in the previous section: the generic discourse features improve the lexical baseline (32%) by 2%, while the genre-specific patterns allow to further improve the accuracy by 6%.

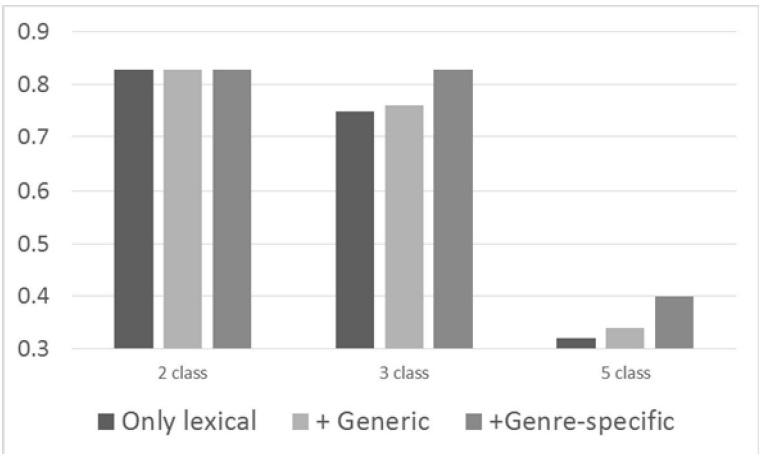

Figure 12. The role of discourse features in 2, 3 and 5-class classification

Thus it can be concluded that the lexical classifier's performance degrades with the number of classes used for classification, which makes the discourse features (especially genre-specific ones) more important.

6.3 Discussion

The results of the 5-class classification task highlight some important properties of genre-specific and generic discourse features which deserve a separate discussion. These are features that ensured the improvement of performance compared to the baseline, in particular, *Explicit Contrast*, *Implicit Contrast* and *Background*.

The influence of *Explicit Contrast* is especially remarkable for 1-star reviews, as, according to Vasquez (2014), contrast markers are often used in negative reviews where they help the reviewer appear reasonable and objective while complaining (see Section 4.2.2.3).

Implicit Contrast, on the other hand, is used to realize another discourse strategy (see Section 4.2.3.1) – it allows the reviewer to make mild statements about those aspects of a product which he or she evaluates differently than its other aspects. Thus opinion hedges, showing weak preference or dispreference of some aspect of a product ("the only good point" etc.), help to single out 2 and 4-star reviews which are not so categorical in evaluation as 1- or 5-star reviews.

Finally, *Background* relations are a strong feature of negative reviews (see Section 4.2.3.2), because they help the reviewer assert his or her identity as a person capable of objectively assessing the product and thus justify their negative evaluation (Vasquez, 2014).

7. Conclusion and Future Prospects

In this study, we compared generic discourse features, applicable to texts of any genre, and genre-specific discourse features, that is, features, which, in scope of our study, are relevant only to online reviews. Firstly, we defined genre-specific features and discovered the ways to extract them by analyzing a corpus of online reviews. Based on Rhetorical Structure Theory and studies of consumer reviews discourse, we explained how genre-specific features could improve the results of sentiment classification. Our experiments proved this claim both at the sentence- and document level: while some of generic features, such as *Implicit Coherence*, turned out to introduce noise and worsen classification performance, genre-specific features were overall more precise and helped to improve the accuracy of prediction. In particular, in terms of fine-grained classification, such features as *Background* and *Involvement* helped to detect objective sentences, while *hedges* (*Implicit Contrast*) served as an indicator of polarity change. On the level of a document, *Background* features improved the results for negative reviews, while *Implicit Contrast* helped to distinguish "middle" classes, such as 2- and 4-star reviews. We also showed that the role of discourse features – both generic and genre-specific – increases with the number of classes in the rating prediction task, which lets us conclude that they are important not only for sentence-level analysis, but also for detailed document-level classification.

Though genre-specific features proved to be highly precise predictors of subjectivity, one cannot but notice that the recall achieved by using these newly discovered patterns is rather limited. Therefore, discovering and formalizing new

genre-specific features which could aid sentiment analysis would be a promising line of research. Also, the features we already defined, such as *Background* relations, should be further generalized to be better suited for new datasets and domains. Ideally, we should find a method to extract such features automatically from any new text. This would not only improve the performance of opinion mining for online reviews, but also serve as a basis for studying and extracting genre-specific patterns in other genres.

In this thesis we showed that such commonly used explicit discourse features as discourse markers have a very limited effect on the accuracy of sentiment analysis tasks, while less overt and obvious features can be more reliably used for this purpose. As many of such implicit features are difficult to discover manually based on corpus study or theoretical research, extracting them by applying Deep Learning and Unsupervised Feature Learning methods appears to be a promising line of research.

We began this study in a hope that using carefully selected and theoretically grounded discourse features would help us achieve better results than using more obvious and general features, often collected automatically by statistical methods. As our experiments proved the viability of this idea, we hope that it will be extended and improved on in future research.

References

Asher, N., Benamara, F., & Mathieu, Y. Y. (2008). Distilling Opinion in Discourse: A Preliminary Study. *COLING (Posters),* 7-10.

Baccianella, S., Esuli, A., & Sebastiani, F. (2009). Multi-facet rating of product reviews. *Advances in Information Retrieval,* 461-472.

Bamberg, M. (1997), Oral versions of personal experience. *Journal of Narrative and Life History*, *7* (*1-4*).

Boutet J., & Maingueneau D. (2005), Sociolinguistique et analyse de discours: facons de dire, facons de faire, *Langage et societe 4/2005* (*No 114*), 15-47.

Breck E., Choi Y., & Cardie C. (2007). Identifying Expressions of Opinion in Context. *IJCAI*, *7*, 2683-2688.

Bucholtz, M., & Hall, K. (2005). Identity and interaction: A sociocultural linguistic approach. *Discourse studies*, *7*(4-5), 585-614.

Chafe, W. (1982). Integration and involvement in speaking, writing, and oral literature. *Spoken and Written Language: Exploring Orality and Literacy*. 35-53.

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

Councill, I. G., McDonald, R., & Velikovich, L. (2010). What's great and what's not: learning to classify the scope of negation for improved sentiment analysis. *Proceedings of ACL*, 51-59.

Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton: Princeton University Press

Dang, Y., Zhang, Y., & Chen, H. (2010). A lexicon-enhanced method for sentiment classification: An experiment on online product reviews. *Intelligent Systems, IEEE, 25(4)*, 46-53.

Edwards, D. (2000). Extreme case formulations: Softeners, investment, and doing nonliteral. *Research on language and social interaction*, *33*(4), 347-373.

Edwards, D. (2005). Moaning, whining and laughing: The subjective side of complaints. *Discourse studies*, *7*(1), 5-29.

Filatova, E. (2012). Irony and Sarcasm: Corpus Generation and Analysis Using Crowdsourcing. *LREC*, 392-398.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, *76*(5), 378-382.

Ganu, G., Elhadad, N., & Marian, A. (2009). Beyond the Stars: Improving Rating Predictions using Review Text Content. *WebDB, 9*, 1-6.

Gumperz, J. (1982), *Discourse Strategies*. Cambridge: Cambridge University Press.

Halliday, M. A., & Hasan, R. (1976). Cohesion in spoken and written English. *Longman's, London*.

Hatzivassiloglou, V., & McKeown, K. R. (1997). Predicting the semantic orientation of adjectives. In *Proceedings of the 35th annual meeting of the Association for Computational Linguistics and eighth conference of the European chapter of the Association for Computational Linguistics*, 174-181

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, *1*(1), 77-89.

Hu, M., & Liu, B. (2004). Mining opinion features in customer reviews. *AAAI, 4*(4), 755-760.

Jo, Y., & Oh, A. H. (2011). Aspect and sentiment unification model for online review analysis. *Proceedings of ACM*, 815-824.

Kanayama, H., & Nasukawa, T. (2006). Fully automatic lexicon expansion for domain-oriented sentiment analysis. *Proceedings of ACL*, 355-363.

Kehler, A. (2002). *Coherence, reference, and the theory of grammar*. Stanford, CA: CSLI publications.

Kirkpatrick, R. (1999). Text Analysis at Different levels: Schema theory, genre analysis, discourse analysis. *Language Issues, vol. 5-1, 49-66*

Krippendorff, K. (1970). Estimating the reliability, systematic error, and random error of interval data. *Educational and Psychological Measurement, 30 (1)*, 61-70.

Krippendorff, K. (2004). Reliability in content analysis. *Human Communication Research*, *30*(3), 411-433.

Krippendorff, K. (2007). Computing Krippendorff's alpha reliability. *Departmental Papers (ASC)*, 43-53.

Lafferty, J., McCallum, A., & Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proceedings of the Eighteenth International Conference on Machine Learning*, 282-289.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, 159-174.

Lapponi, E., Read, J., & Ovrelid, L. (2012). Representing and resolving negation for sentiment analysis. *Proceedings of the 2012 ICDM Workshop on Sentiment Elicitation from Natural Text for Information Retrieval and Extraction*, 687-692.

Lascarides, A., & Asher, N. (2007). Segmented discourse representation theory: Dynamic semantics with discourse structure. *Computing meaning,* 87-124.

Lazaridou, A., Titov, I., & Sporleder, C. (2013). A Bayesian Model for Joint Unsupervised Induction of Sentiment, Aspect and Discourse Representations. *Proceedings of ACL*, 1630-1639.

Li F., Han C., Huang M., Zhu X., Xia Y. J., Zhang S., & Yu H. (2010). Structure-aware review mining and summarization. *Proceedings of the 23rd International Conference on Computational Linguistics*, 653-661.

Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers.

Liu, Y., Yu, X., Liu, B., & Chen, Z. (2014). Sentence-Level Sentiment Analysis in the Presence of Modalities. Computational Linguistics and Intelligent Text Processing, 1-16.

Mackiewicz, J. (2007). Reviewer bias and credibility in online reviews. *Association for Business Communication Annual Convention*.

Mackiewicz, J. (2010a). Assertions of expertise in online product reviews. *Journal of Business and Technical Communication*, *24(1)*, 3-28.

Mackiewicz, J. (2010b). The co-construction of credibility in online product reviews. *Technical Communication Quarterly, 19(4)*, 403-426.

Mann, W. & Thompson, S. (1987). Rhetorical structure theory: A theory of text organization*, The Structure of Discourse,* 87-190.

Mann, W. & Thompson, S. (1988). Rhetorical structure theory: Towards a functional theory of text organization. *Text*, *8*(3), 243-281.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge: Cambridge University Press.

Mao, Y., & Lebanon, G. (2006). Isotonic conditional random fields and local sentiment flow. *Advances in neural information processing systems*, 961-968.

Marcu, D., & Echihabi, A. (2002). An unsupervised approach to recognizing discourse relations. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics,* 368-375.

Markkanen R. and Schröder, H. (1997a). *Hedging and Discourse*, 280 p.

Markkanen, R., & Schröder, H. (1997b). Hedging: A challenge for pragmatics and discourse analysis. *Research in Text theory*, 3-20.

McDonald, R., Hannan, K., Neylon, T., Wells, M., & Reynar, J. (2007). Structured models for fine-to-coarse sentiment analysis. *Annual Meeting-Association For Computational Linguistics, vol. 45, 1*, 432-439.

Neurauter-Kessels, M. (2011), Im/polite reader responses on British online news sites. *Journal of Politeness Research*, *7*, 187-214.

Otmakhova, Y. & Shin, H. P. (2015). Do We Really Need Lexical Information? Towards a Top-down Approach to Sentiment Analysis of Product Reviews. *Proceedings of NACCL*, 1559-1568.

Pang, B., & Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. *Proceedings of ACL*, 271-279.

Pang, B., & Lee, L. (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. *Proceedings of ACL,* 115-124.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. *Proceedings of EMNLP*, 79-86.

Plackett, R. L. (1983). Karl Pearson and the chi-squared test. *International Statistical Review/Revue Internationale de Statistique*, 59-72.

Polanyi, L., & Zaenen, A. (2006). Contextual valence shifters. *Computing attitude and affect in text: Theory and applications,* 1-10.

Pomerantz, Anita. (1986). Extreme case formulations: A way of legitimizing claims. *Human studies, 9.2-3*, 219-229.

Popescu, A. M., & Etzioni, O. (2007). Extracting product features and opinions from reviews. *Natural language processing and text mining*, 9-28. Springer London.

Qu, L., Gemulla, R., & Weikum, G. (2012). A weakly supervised model for sentence-level semantic orientation analysis with multiple experts. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning,* 149-159.

Quirk, R., Greenbaum, S. L. G. & Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. *Harlow: Longman*.

Renkema, J. (2004). *Introduction to Discourse Studies*. John Benjamins Publishing.

Riloff, E., Wiebe, J., & Wilson, T. (2003). Learning subjective nouns using extraction pattern bootstrapping. *Proceedings of the seventh conference on Natural language learning at HLT-NAACL,* 25-32.

Sanders, T., & Maat, H. P. (2006). Cohesion and coherence: Linguistic approaches. *Encyclopedia of Language and Linguistics, 2nd edition, volume 2,* 591-595.

Scott, W. A. (1955). Reliability of content analysis: The case of nominal scale coding. *Public opinion quarterly*.

Sha, F., & Pereira, F. (2003). Shallow parsing with conditional random fields. *Proceedings of NAC CL*, *1*, 134-141.

Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. *Proceedings of EMNLP*, 1642-1654.

Somasundaran S. (2010). *Discourse-Level Relations for Opinion Analysis* (Doctoral dissertation). University of Pittsburgh.

Strauss, S., & Feiz, P. (2014). *Discourse analysis: Putting our worlds into words*. Abingdon: Routledge.

Stubbs, M (1983). *Discourse Analysis: The Sociolinguistic Analysis of Natural Language.*

Taboada, M., & Mann, W. C. (2006). Applications of rhetorical structure theory. *Discourse studies*, *8*(4), 567-588.

Taboada, M. (2009). Implicit and Explicit Coherence Relations. *Discourse, of Course*, 125-138.

Täckström, O., & McDonald, R. (2011). Discovering fine-grained sentiment with latent variable structured prediction models. *Advances in Information Retrieval,* 368-374.

Toprak, C., Jakob, N., & Gurevych, I. (2010). Sentence and expression level annotation of opinions in user-generated discourse. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics,* 575-584.

Trivedi, R. S., & Eisenstein, J. (2013). Discourse Connectors for Latent Subjectivity in Sentiment Analysis. *HLT-NAACL,* 808-813.

Turney, P. D. (2002). Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. *Proceedings of ACL*, 417-424.

Vasquez, C. (2011), Complaints online: The case of TripAdvisor. *Journal of Pragmatics*, *43*, 1707-17

Vasquez, C. (2014). *The Discourse of Online Consumer Reviews*. Bloomsbury Publishing.

Wiebe, J. (1990). Identifying subjective characters in narrative. In *Proceedings of the 13th conference on Computational linguistics,* 401-406.

Wiebe, J. (1994). Tracking point of view in narrative. *Computational Linguistics*, *20(2)*, 233-287.

Wiebe, J. (2000). Learning subjective adjectives from corpora. *AAAI/IAAI*, 735-740.

Wiebe, J., Wilson, T., & Bell, M. (2001). Identifying collocations for recognizing opinions. *Proceedings of the ACL-01 Workshop on Collocation: Computational Extraction, Analysis, and Exploitation,* 24-31.

Wilson, T., Hoffmann, P., Somasundaran, S., Kessler, J., Wiebe, J., Choi, Y.. & Patwardhan, S. (2005a). OpinionFinder: A system for subjectivity analysis. *Proceedings of HLT/EMNLP,* 34-35.

Wilson, T., Wiebe, J., & Hoffmann, P. (2005b). Recognizing contextual polarity in phrase-level sentiment analysis. *Proceedings of the conference on human language technology and empirical methods in natural language processing*, 347-354.

Wilson, T., Wiebe, J., & Hoffmann, P. (2009). Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational linguistics, 35(3)*, 399-433.

Wu, Y., & Jin, P. (2013). Semeval-2010 task 18: Disambiguating sentiment ambiguous adjectives. *Language resources and evaluation*, *47*(3), 743-755.

Yang, B., & Cardie, C. (2014). Context-aware learning for sentence-level sentiment analysis with posterior regularization. *Proceedings of ACL, 325-335.*

Zhang, Q., Qian, J., Chen, H., Kang, J., & Huang, X. (2013). Discourse Level Explanatory Relation Extraction from Product Reviews Using First-Order Logic. *Proceedings of EMNLP*, 946-957.

Zhao J., Liu K. & Wang G. (2008). Adding Redundant Features for CRFs-based Sentence Sentiment Classification. *Proceedings of EMNLP,* 117-126.

Zhou L., Li B., Gao W., Wei Z. & Wong K. F. (2011). Unsupervised Discovery of Discourse Relations for Eliminating Intra-sentence Polarity Ambiguities. *Proceedings of EMNLP*, 162-171.

Zirn, C., Niepert, M., Stuckenschmidt, H., & Strube, M. (2011). Fine-Grained Sentiment Analysis with Structural Features. *IJCNLP*, 336-344.

초록

장르 특정적 담화 유형 기반의 온라인 리뷰의 감정분석

최근 감정분석 연구는 단순한 어휘기반 모형과 통계모형에서 더 나아가 담화 정보(discourse information)를 적극 활용하는 데까지 이르고 있으나, 텍스트의 다양한 장르와 유형(사회관계망, 언론 사설, 토론 게시판, 온라인 리뷰 등)을 고려하지 않고 동일한 자질 집합을 사용한다는 문제는 여전히 남아 있다. 더구나 기존 연구에서 사용된 자질이 표상하는 담화 측면은 응집성(coherence) 한 가지뿐이며, 응집성을 확보하는 방식도 접속사 등 명시적인 것으로 한정되어 있다. 구체적인 자질로는 두 문장이 인접한 데서 드러나는 암묵적 응집성(implicit coherence)과, 두 문장이 인접할 때 같은 감정극성을 가짐을 보여주는 주로 ‘and’(그리고)나 ‘moreover’(게다가)와 같은 접속사로 반영되는 연속성(continuity), 그리고‘but’(그러나)과 같은 접속사로 표시되고 의견의 극성의 전이를 보여주는 대조(contrast) 등이 있다. 본고에서는 온라인 리뷰라는 특정 장르의 구체적 특성을 반영하는 새로운 자질로 ‘the only drawback’(유일한 단점) 등과 같이 한계 짓는 표현으로 실현되는 암묵적 대조(implicit contrast)와, 상품평 작성자의 신원을 밝히는 데 도움을 주는 표현인 배경(background) 유형, 그리고 독자와 소통하는 데 쓰이는 개입(involvement)을 새로 도입한다.

이러한 자질들의 효과를 보이기 위하여 본 연구에서는 상품평 120개로 이루어진 말뭉치에 주석을 달고 각 상품평에서 추출된 비담화 자질,

포괄적 자질, 장르 특정적 담화 자질의 집합으로 (주석에서 붙인 대상 표지와 결합하여) 상품평을 구성하였다. 자질 집합은 문장 층위 및 문서 층위 두 단계의 실험에서 사용되었다. 문장 층위에서는 어휘 자질을 포함하는 실험과 제외하는 실험을, 문서 층위에서는 5개, 3개, 2개 분류를 수행하였다.

실험 결과 장르 특정적 자질이 일반적으로 포괄적 자질보다 좋은 성능을 보여 정밀도(precision)와 재현도(recall)가 모두 더 높았다. 암묵적 응집성의 경우에서처럼 포괄적 자질이 성능을 덜 향상시키거나 오히려 저하시켰다면, 배경을 비롯한 장르 특정적 자질은 더 안정적이었고 모든 실험에서 더 나은 재현도와 정밀도를 보여주었다. 이러한 경향은 특히 어휘 자질을 포함시킨 문장 층위 분류에서 더 특징적이었고, 같은 실험에서 포괄적 담화 자질을 추가했을 때는 오히려 성능이 떨어졌다. 따라서 장르 특정적 자질의 성능은 통계적으로 신뢰할 수 있는 동시에 온라인 리뷰 담화에 관하여 본고에서 서술한 이론적 속성을 반영한다고 볼 수 있다.

주제어: 감정분석, 오피니언마이닝, 온라인 리뷰, 상품평, 담화분석
학번: 2012-23882