Mining Fine-grained Opinion Expressions with Shallow Parsing

Sucheta Ghosh
CLCS, Trinity College Dublin
ghoshs@tcd.ie

Sara Tonelli
FBK, Trento, Italy
satonelli@fbk.eu

Richard Johansson
University of Gothenberg, Sweden
richard.johansson@svenska.gu.se

Abstract

Opinion analysis deals with public opinions and trends, but subjective language is highly ambiguous. In this paper, we follow a simple data-driven technique to learn fine-grained opinions. We select an intersection set of Wall Street Journal documents that is included both in the Penn Discourse Tree Bank (PDTB) and in the Multi-Perspective Question Answering (MPQA) corpus. This is done in order to explore the usefulness of discourse-level structure to facilitate the extraction of fine-grained opinion expressions. Here we perform shallow parsing of MPQA expressions with connective based discourse structure, and then also with Named Entities (NE) and some syntax features using conditional random fields; the latter feature set is basically a collection of NEs and a bundle of features that is proved to be useful in a shallow discourse parsing task. We found that both of the feature-sets are useful to improve our baseline at different levels of this fine-grained opinion expression mining task.

1 Introduction

The explosion of data in all forms from blogs, online forums, Facebook, Twitter and other social media channels has given an opportunity of unprecedented reach to publicly sharing thoughts on events, products and services. However, there are some open issues related to this research area, commonly known as *Opinion Mining*, which can be summarized as follows: *(1)* Opinions are potentially ambiguous, and *(2)* Contextual interpretation of polarity is hard to achieve. Subsidiary important problem is the non-availability of large corpora with good annotation quality.

Fine-grained opinion analysis is a different task from the *coarse-grained* one (e.g. document level analysis), in that it classifies opinion phrases, chunks or expressions from a given text. In this work, we perform fine-grained analysis by focusing on higher-level linguistic structure like discourse, without rich linguistic or knowledge-intensive features, to classify subjective opinion expressions using the Multi-Perspective Question Answering corpus (MPQA) scheme Wiebe et al. (2005).

We perform two different experiments sets. We first exploit gold features based on shallow discourse structure[1] to classify fine-grained opinion expressions. In a second experiment, we use some syntax based features, those are found useful on a shallow discourse structure classification task, along with the named entities. Both of the experiments are found to be useful at different levels of fine-grained opinion expression mining. We use conditional random fields for this entire shallow parsing task. A set of documents from the Wall Street Journal (WSJ) corpus Marcus et al. (1993) annotated both in the Penn Discourse Treebank Prasad et al. (2008) and MPQA corpus is used. We also take advantage of the availability of several robust natural language processing tools pretrained on WSJ data.

2 Related Work

Fine-grained sentiment analysis methods have been developed by Hatzivassiloglou and McKeown (1997), Hu and Liu (2004) and Popescu and Etzioni (2007), among others. The first approach focuses on conjoined adjectives (i.e. the adjectives which are joined with discourse connectives) within the WSJ corpus. While the second one operates at the sentence level, the third one extracts

[1] By *shallow discourse structure* we mean the explicit discourse connective sense and its two argument spans Ghosh (2012).

Proceedings of Recent Advances in Natural Language Processing, pages 302–310,
Hissar, Bulgaria, 7-13 September 2013.

opinion phrases at the subsentence level for product features. Rich sets of linguistic features are used in the works of Choi et al. (2005), Wilson et al. (2005a), Breck et al. (2007). The first use conditional random models with information extraction patterns; the second is more focused on the classification of opinion phrases using contextual polarity; the third approach improved the performance of Wilson et al. (2005a), using conditional random fields and external knowledge sources.

Johansson and Moschitti (2013) developed a joint model-based sequence labeler for fine-grained opinion expression using relational features except discourse-level features, beside a set of classifier to determine opinion holder and also a multi-class classifier that assigns polarity to a given opinion expression. These classifiers were further used to generate the hypothesis sets for a re-ranking system that further improved the performance of the classification. Täckström and McDonald (2011) combine fully and partially supervised structured conditional models for a joint classification of the polarity of whole reviews and review sentences.

The impact of discourse relations for sentiment analysis is investigated in Asher et al. (2009). The authors conduct a manual study in which they represent opinions in text as shallow semantic feature structures. These are combined with overall opinion using hand-written rules based on manually annotated discourse relations. An interdependent classification scenario to determine polarity as well as discourse relations is presented in Somasundaran and Wiebe (2009). In their approach, text is modeled as opinion graphs including discourse information. In Somasundaran and Wiebe (2009) the authors try alternative machine learning approaches with combinations of supervised and unsupervised methods for the same task. However, they do not automatically identify discourse relations, but used task-specific manual annotations.

Polanyi and Zaenen (2006) investigate the usage of contextual valence shifters and discourse connectives inside a text. In the approach of Kim and Hovy (2006) the system makes use of conjunctions like “and” to infer polarities and applies a specific rule to sentences including the word “but”: if no polarity can be identified for the clause containing “but”, the polarity of the previous phrase is negated. In a more recent system, Zirn et al. (2011) incorporated this information using discourse relations. Zirn et al. (2011) studied a fully automatic framework for fine-grained sentiment analysis at sub-sentence level, combining multiple sentiment lexicons and neighbourhood as well as discourse relations. They used Markov logic to integrate polarity scores from different sentiment lexicons with information about relations between neighbouring segments, and evaluate the approach on product reviews. The authors used only contrast and no contrast discourse relations to achieve their results, conducting a survey on a small amount of data that showed that the contrast relation was the most frequent one. However, the survey presented in Hatzivassiloglou and McKeown (1997) on the WSJ corpus showed that contrast is actually the third most important relation in the corpus. Therefore the hypothesis made by Zirn et al. (2011) may be data specific.

The framework of Heerschop et al. (2011) achieved even better results than Zirn et al. (2011). The system uses deep discourse structure as well as SentiWordNet and WordNet in order to disambiguate words.

Kim and Hovy (2004) define opinion as a quadruple composed by topic, holder, claim and sentiment. The authors use a Named Entity tagger to identify the potential holder of the opinion. Later Stoyanov and Cardie (2008) argue that in fine grained subjectivity analysis, topic identification is very relevant, and treat the task from the perspective of topic coreference resolution. The authors use named entities beside other topic based features to represent the topical structure of text.

3 Data Resources

In order to test our hypothesis we used 80 Wall Street Journal documents Marcus et al. (1993) that are part both of the Penn Discourse TreeBank (PDTB) and of the Multi-Perspective Question-Answer (MPQA) bank.

3.1 Penn Discourse TreeBank (PDTB) 2.0

The Penn Discourse Treebank (PDTB) is a resource containing one million words from the Wall Street Journal corpus Marcus et al. (1993) annotated with discourse relations.

Connectives in the PTDB are treated as discourse predicates taking two text spans as arguments (Arg), i.e. parts of the text that describe events, propositions, facts, situations. Such two arguments in the PDTB are called Arg1 and Arg2, with the numbering not necessarily corresponding to their order in text. Indeed, Arg2 is the argument syntactically bound to the connective, while Arg1 is the other one.

In the PDTB, discourse relations can be either overtly or implicitly expressed. However, we focus here exclusively on explicit connectives and the identification of their arguments, including the exact spans. This kind of classification is very complex, since Arg1 and Arg2 can occur in many different configurations (see Table).

In PDTB the senses are assigned according to a three-layered hierarchy: the top-level classes are the most generic ones and include TEMPORAL, CONTINGENCY, COMPARISON and EXPANSION labels. We used these four surface senses only in our task.

We define our discourse structure as shallow since it includes only the discourse connective senses and its two argument spans, excluding other types of hierarchical annotation.

3.2 Multi-Perspective Question Answering (MPQA)

We use the version 2.0 of the MPQA corpus, whose central building block is opinion expression. Opinion expressions belong to two categories: Direct subjective expressions (DSEs) are explicit mentions of opinion, whereas expressive subjective elements (ESEs) signal the attitude of the speaker by the choice of words, other than these there are Objective Speech Events (OSEs). Opinions have two features: polarity and intensity, and most expressions are also associated with a holder, also called source. In this work, we only consider polarities, not intensities or holders. Polarity can be POSITIVE, NEUTRAL, NEGATIVE, and BOTH; for compatibility with Choi and Cardie (2010), we mapped BOTH to NEUTRAL.

4 Our Approach

The goal of our first experiment is to observe the effect of a limited number of gold label features from PDTB. Since no previous work documented the effect of PDTB senses on the task of opinion expression mining using MPQA, we use four PDTB surface senses (described in the Subsection 3.1) as one of the features in this experiment. We then run the second experiment in order to observe the effect of named entities with the mentioned feature bundle. This set of features encoding some syntactic-level information may improve the overall classification performance like the same features facilitated a shallow discourse parsing task by Ghosh et al. (2011); in addition to the feature bundle, the named entities might reflect some information about distribution of discourse entities.

5 Experiments

We perform our experiments at two different stages: (1) we first draw a baseline using basic features from the previous work and a standard sentiment lexicon by Wilson et al. (2005b), then (2) we run further experiments to improve the baseline with additional features. Our goal is to investigate possible improvements using discourse features or some other features that may encode discourse information via shallow parsing.

The experiments are entirely run using conditional random fields, keeping the same settings for the three experiments. We used standard training technique for conditional random fields, as provided by the tool developers in the instruction manual. We use the CRF++ tool [2] for sequence labeling classification by Lafferty et al. (2001), with second-order Markov dependency between tags. Beside the individual specification of a feature in the feature description template, the features in various combinations are also represented. We used this tool because the output of CRF++ is compatible with CoNLL 2000 chunking shared task, and we view our task as an opinion expression chunking task. On the other hand, linear-chain CRFs for sequence labeling offer advantages over both generative models like HMMs and classifiers applied at each sequence position. Also Sha and Pereira (2003) claim that, as a single model, CRFs outperform other models for shallow parsing. We use conditional random fields to classify subjective (any of direct or expressive) and objective expressions. We encode the opinion expression spans by means of the IOB2 scheme Sang et al. (1999). In order to represent MPQA opinion expressions with IOB2 tags, we remove the expressions where the expression spans are overlapping expressions (i.e. an opinion expression span can be overlapped by another opinion expression span), though overlapping expressions are rare in MPQA [Johansson and Moschitti (2013)].

[2] (http://crfpp.sourceforge.net/)

Since the dataset is fairly small, we perform a 5-fold cross validation over the dataset to have a rough estimation of how accurately the predictive model will perform in practice. One round of cross-validation involves random multiple rounds to partition data into complementary subsets: the training set (75%), the validation set (10%) and the test set (15%). The results are averaged over the rounds. This multiple round partition is kept the same for all the experiments in this paper in order to make results comparable. Our training, validation and test sets are different from the respective sets used by Breck et al. (2007) and Johansson and Moschitti (2013).

5.1 Evaluation

We present all results using precision, recall and F1 measures. To compute precision and recall, we used two scoring schemes: exact and overlap-based scoring. A span is counted as exact-correct if its extent exactly coincides with one in the gold standard, whereas in overlap-based measures, a span is counted as correctly detected if it overlaps with a span in the gold standard. Note that all the partial measures are bounded below by the exact measures and above by the overlap-based measures. Further details on these scoring techniques are given in Johansson and Moschitti (2013).

The results are primarily compared using two metrics: micro-averages and macro-averages of precision, recall and F1 measures. In order to facilitate comparison between baseline and other experiments results we compute macro and micro averages of results from the 5-fold cross validation for each experiments.

5.2 Baseline

We construct our baseline with four features. three of them are linguistic features, viz. the current token, the lemma and the part-of-speech (PoS) tag of the token. The fourth one is the polarity value of the current token taken from a standard subjectivity lexicon maintained by Wilson et al. (2005b). The selection of baseline features is motivated by the work of Breck et al. (2007). The features are listed in the Table 1.

Features used to prepare the baseline.	
BF1.	Token (T)
BF2.	Lemma (L)
BF3.	PoS tag
BF4.	Polarity Values (POLV)

Table 1: Baseline Feature sets opinion expression labeling.

5.3 Experiment with Discourse Connectives & Arguments

In order to observe the effect of (explicit) discourse connective senses and their argument spans, we use conditional random fields with an extended set of features from shallow discourse structure by Ghosh (2012) on the top of the baseline features. In particular, we use one of the four explicit discourse connective senses (viz. Expansion, Contingency, Comparison and Temporal) and its two arguments with spans. We also use IOB2 tags with argument spans. In order to reduce the complexity of the classification, overlapping argument span tags are removed, which however are fairly small in amount. We illustrate the features used in this experiment in Table 2. The features viz. CONN, ARG1 and ARG2 are gold-labeled features, i.e. they are directly extracted from available PDTB annotation.

Features used to perform Expt. with discourse structure.	
E1F1.	Sense of Connective (CONN)
E1F2.	Arg1 Span (ARG1)
E1F3.	Arg2 Span (ARG2)
Additional features used	
BF1-BF4.	All baseline features

Table 2: Feature sets for opinion expression labeling with Shallow Discourse Structure Features.

5.4 Experiment with Named Entities (NEs) and syntax based features

In this experiment we used four new features on the top of baseline features, which are listed in Table 3. Apart from Named Entities (NE), a bundle of other features are used (IOB, L+I, BMV), which were previously used in a shallow discourse parsing task Ghosh et al. (2011). No member of this bundle feature-set from the shallow discourse parsing task directly provides information about discourse, but when used altogether these may reflect some discourse information. Among the bundle of features, IOB chain and Inflection provide morpho-syntactic information, whereas the lemma and boolean value of the main verb of the main clause provide lexical information.

We use the scripts provided for the CoNLL chunking shared task 2000 [3] to extract IOB chains. Besides, we use the Morpha tool by Minnen et al. (2001) to extract lemma and inflection for the tokens. The main verb of the main clause is extracted following the head rules by Yamada and

[3] (http://ilk.uvt.nl/team/sabine/homepage/software.html)

Matsumoto[4]. We used the Stanford Named Entity tagger by Finkel et al. (2005) to tag the named entities. This tagger is a three-class (viz. PERSON, ORGANISATION, LOCATION) tagger for English. The pre-trained models are trained both on CoNLL 2003 and MUC data, for the intersection of those class sets. NEs are included as a feature following the previous work by Stoyanov and Cardie (2008), where the authors show that information from NEs contribute to the entity relation structure in a discourse.

Features used to perform Expt. with NE & other features.	
E2F1.	Named Entities (NE)
E2F2.	IOB chain (IOB)
E2F3.	Lemma+Inflection (L+I)
E2F4.	Boolean feature for main verb of main clause (BMV)
Additional features used	
BF1-BF4.	All baseline features

Table 3: Feature sets for opinion expression labeling with NE & other features.

6 Results

We present the results obtained at different levels of fine-grained opinion mining. We attempt to compare some of the results with the respective results from Johansson and Moschitti (2013) in order to understand the trend of improvement of results over our baseline. The explored levels of fine-grained mining is demonstrated in the Table 4. We report here the interesting findings and comparisons from this level-wise studies.

All the systems (i.e. baseline, discourse-structure based and NE-syntax based systems) perform the worst for the polarity detection. This trend is the same with the system of Johansson and Moschitti (2013) (J&M). In the Table 5 we compare the macro-averages of our system to the system of J&M, in the case of polarity tagged expression classification, where the OSEs are removed, and the DSEs and ESEs are included but not distinguished. In this case NE and syntax based system performs better than the baseline and discourse structure based system, because may be NEs are better feature for polarity extraction than surface senses of discourse connectives. The recall of the J&M's system is balanced with precision therefore it performs better than the other systems.

L1. With Not Distinguished DSE+ESE+OSE+Polarity		
L2.Without OSE+Polarity	1.	ND(DSE+ESE)
	2.	(DSE+ESE)
L3. Without Polarity	1.	ND(DSE+ESE)+OSE
	2.	DSE+ESE+OSE
L4. Without OSE	1.	ND(DSE+ESE)+Polarity
	2.	(DSE+ESE)+Polarity
L5. With DSE+ESE+OSE+Polarity		

Table 4: The Explored Levels of Opinion Mining Results (ND: Not Distinguished).

Partial Metric			
Metrics	P	R	F1
J&M	0.547	0.456	0.497
Baseline	0.628	0.208	0.313
Discourse based	0.596	0.127	0.209
NE&Syntax based	0.658	0.228	0.339

Table 5: Results for identifying Polarity expressions without OSEs and with not distinguished DSEs and ESEs (Ref. level L4.1 in Tab. 4).

Overlap Metric			
Metrics	P	R	F1
J&M	0.834	0.75	0.79
Baseline	0.768	0.411	0.536
Discourse based	0.772	0.425	0.548
NE&Syntax based	0.733	0.321	0.447

Table 6: Results for identifying Subjective expressions without OSEs and polarity tags (Ref. level L3.2 in Fig. 4).

In the Table 6 we also compare another most relevant result by J&M with our macro average results at corresponding level, where the OSEs removed, the DSEs and ESEs included but not distinguished, and there is no polarity values. We observe that the system of J&M outperforms our systems. In this case the results of J&M's system is computed using 10-fold cross validation, whereas we used 5-fold cross validation, in addition to this the test data of J&M is not the same with our system. This comparisons make it clear that all the systems perform well with no polarity tags and perform worse for polarity tagged expression classification. J&M's system has a balanced precision and recall score wheres our systems suffer from low recall.

We view the experiment results with no distinguished DSEs, ESEs, OSEs and polarities (Level: L1 Fig. 4). The best results obtained with NEs and syntax-based feature set is highlighted in the Table 7. We observe that the exact macro-average scores obtained with shallow discourse structure feature classification outperforms our own baseline; NEs and syntax based feature fails to outperform that baseline, this is may be due to the fact that at this level the discourse structure provide more information than the NEs.

Table 10 shows that shallow discourse structure features do not provide any improvement to our baseline results at the level of L5 (Table 4). The

[4] The software can be downloaded from `http://www.jaist.ac.jp/\~h-yamada/`

Experiments	Averages	Exact Measures		
		P	R	F1
Baseline	macro avg	0.826	0.442	0.576
	micro avg	0.819	0.417	0.553
Expt with discourse struct.	macro avg	**0.833**	**0.459**	**0.592**
	micro avg	**0.830**	**0.425**	**0.562**
Expt with NE based features	macro avg	0.849	0.372	0.517
	micro avg	0.856	0.338	0.484

Table 7: Baseline & Other Experiment Results with not distinguished DSE+ESE+OSE+Polarities (L1).

NE & Syntax based Feature			
Metrics	P	R	F1
Before Feature optimisation	0.816	0.477	0.602
After Feature optimisation	0.886	0.477	0.620

Table 8: Exact Score comparison for identifying subjective expressions with NE based features with the best performing split before and after feature optimisation with test split.

Features	P	R	F1
Features in Isolation			
Baseline (B)	0.765	0.433	0.553
Named Entity (NE)	0.500	0.122	0.196
IOB_Chain (IOB)	0.428	0.100	0.162
Morph(L+INFL)	0.044	0.067	0.053
Hill-Climbing Feature Analysis			
B+NE	0.794	0.467	0.588
B+NE+IOB	0.824	0.467	0.596
B+NE+IOB+Morph	**0.816**	**0.477**	**0.602**
B+NE+IOB+Morph+BMV	0.875	0.431	0.577
Feature Ablation			
B+NE+IOB	0.824	0.467	0.596
B+NE+Morph	0.794	0.467	0.588
IOB+NE+Morph	0.285	0.067	0.108
B+IOB+Morph	0.750	0.400	0.522

Table 9: Feature Analysis Results with Single and Combined Features for the Expt with NE and syntax based feature set

reason behind this may be that the information provided by the current shallow discourse structure is falling short to achieve an improvement, whereas at this level the NE and syntax-based feature-set is useful to achieve better performance over the baseline scores. Results reported in Table 8 show a considerable improvement in the results over best performing split after the optimisation.

6.0.1 Feature Analysis

Our baseline feature set includes a small set of lexical and syntactic features, which convey the essential information needed to classify opinion expressions. We enrich this baseline set with some additional features, which better represent the position of opinion expressions and the respective boundaries, as well as the internal clause structure. Then, we carry out a selection step in order to identify only the feature combination that performs best in our parsing task.

We follow the hill-climbing (greedy) feature selection technique proposed by Caruana and Freitag (1994). In this optimization scheme, the best-performing set of features is selected on the basis of the best F1 "exact" score. Therefore, we increase the number of features at each step, and report the corresponding performance. In order to understand better the contribution of each feature and also to avoid sub-optimal solutions, we also run an ablation test by leaving out one feature in turn from the best-performing set. We use the development split to generate results for the feature analysis to find the best performing feature set, whereas the train split is used to build the model. Final results are generated using only the test split.

We run the hill climbing feature analysis on the best performing partition among the five partitions prepared for cross-validation. The results of our feature analysis are reported in Table 9. We do not report the scores having zero as F1-measure. We also run backward hill climbing technique, and the result is the same with forward hill climbing, because our feature set is fairly small in size. Therefore we do not report it in Table 9.

Both the feature-in-isolation procedure and the ablation test show that the bundle of baseline features is the best performing because it conveys the most essential information to classify any opinion expression. Apart from that, the named entity feature is the next most relevant feature, which carries the sufficient information on the position of a opinion expression, because an opinion expression starts frequently just after a NE occurrence. The named entity feature is more effective when integrated with information from the IOB chain, because the IOB chain feature conveys information on the span. The boolean value of the main verb in the main clause is not an important feature, probably because it conveys redundant information, therefore we do not use it any more.

We observe that the performance of the lemma increases if integrated with the inflection feature, while inflection in isolation scores a null Precision, Recall and F1. Therefore, we consider lemma and inflection together as a single feature (we call it Morph in Table 9). The best performing set includes three new features: named entities and the two features used in shallow discourse parsing, namely IOB chain and Morph.

Finally we compute the results with the test split

Experiments	Averages	Exact Measures		
		P	R	F1
Baseline	macro avg	0.671	0.361	0.468
	micro avg	0.659	0.337	0.446
Expt with discourse struct.	macro avg	0.656	0.330	0.436
	micro avg	0.638	0.317	0.423
Expt with NE based features	macro avg	**0.793**	**0.376**	**0.510**
	micro avg	**0.772**	**0.356**	**0.487**

Table 10: Baseline & Other Experiment Results with DSE+ESE+OSE+Polarities (L5).

given in Table 10. The best results obtained with NEs and syntax-based feature set is highlighted in the Table 10. We observe that the exact macro-average scores obtained with NE and syntax-based feature classification outperforms our own baseline.

7 Discussion

The classification result suffers from low recall values whereas the precision is considerably high. This is because the CRF classifier is being too conservative to tag subjectivity labels. We also analyze the result outputs from the experiment described in Section 5.4. We present here some interesting representative examples of mistakes done by the classifier.

The classifier is not able to tag a long opinion span like "*Neither Equus nor Tony Lama gave a reason* for the changed offer and Tony Lama couldn't be reached for comment". This may depend on the fact that the classifier may not get enough clue from the features on how many tokens to tag.

The use of shallow discourse structure was meant to facilitate the classification of opinion expression boundaries by exploiting information on argument spans. Some interesting cases observed while manually inspecting the problematic cases are the following (the italics strings in the examples are argument 1, while the bold parts mark argument 2, and the underlined tokens are discourse connectives according to PDTB annotations):

(a) an example with intra-sentential explicit relation:

(eg1) The White House said *Mr. Bush decided to grant duty-free status for 18 categories,* but **turned down such treatment for other types of watches, " because of the potential for material injury to watch producers located in the U.S. and the Virgin Islands.** [COMPARISON]

This sentence is annotated in both schemes, wholly in PDTB and partly in MPQA. There are MPQA expressions annotated both in Arg1 and in Arg2, as MPQA annotation implicitly makes use of the contrastive sense of "but". Our classifier performs well with these kind of sentences, where the relation is straightforward because no other deeper sense of the relations is implied. Problems arise when there is no MPQA annotation in one of the two arguments (i.e. the next example).

(b) an example with inter-sentential explicit relation:

(eg2) The White House said *President Bush has approved duty-free treatment for imports of certain types of watches that are n't produced in "significant quantities" in the U.S., the Virgin Islands and other U.S. possessions.* The action came in response to a petition filed by Timex Inc. for changes in the U.S. Generalised System of Preferences for imports from developing nations. Previously, **watch imports were denied such duty-free treatments.** [TEMPORAL]

In this case, only part of argument 1 (i.e. *has approved*) is annotated as subjective opinion expression, whereas no MPQA annotation is present in argument 2. Therefore in this case the features based on the discourse relation are not helpful. On the other hand, in this example the NE tags play a significant role in correctly locating the opinion expressions.

8 Conclusion

In this paper we investigated whether shallow discourse-level information improves the classification of subjective opinions. We chose two standard annotation schemes, viz. PDTB and MPQA, to analyze the interoperability of these schemes. Primarily we used a baseline using few linguistic features and polarity feature from a standard subjectivity lexicon by Wilson et al. (2005b). Then we performed another experiment using a set of syntax-based features from Ghosh et al. (2011) and named entities.

We found that both of the feature-sets succeed to improve the baseline considerably at various levels of fine-grained opinion mining. This is probably because the named entities tend to express the information on the opinion holder usually preceding an opinion expression. Also discourse-based features are useful, because they provide the meaning structural information on the text.

As a future work, we plan to enrich the feature-set with additional discourse level information. A constraint based approach could also be chosen to balance precision and recall.

References

Nicholas Asher, Farah Benamara, and Yannick Mathieu. 2009. Appraisal of opinion expressions in discourse. In *Lingvisticae Investigations*, volume 31(2), pages 279–292.

Eric Breck, Yejin Choi, and Claire Cardie. 2007. Identifying expressions of opinion in context. In *the 20th international joint conference on Artifical intelligence.* Morgan Kaufmann Publishers Inc.

R. Caruana and D. Freitag. 1994. Greedy attribute selection. In *11th International Conference in Machine Learning.* Morgan Kaufmann.

Yejin Choi and Claire Cardie. 2010. Hierarchical sequential learning for extracting opinions and their attributes. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 269–274.

Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. 2005. Identifying sources of opinions with conditional random fields and extraction patterns. In Association for Computational Linguistics, editor, *Human Language Technology and Empirical Methods in Natural Language Processing.*

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, pages 363–370.

Sucheta Ghosh, Richard Johansson, Giuseppe Riccardi, and Sara Tonelli. 2011. Shallow discourse parsing with conditional random fields. In *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP)*, Chiang Mai, Thailand.

Sucheta Ghosh. 2012. *End to End Discourse Parsing with Cascaded Structured Prediction.* Ph.D. thesis, University of Trento.

Vasileios Hatzivassiloglou and Kathleen R. McKeown. 1997. Predicting the semantic orientation of adjectives. In *eighth conference on European chapter of the Association for Computational Linguistics.*

B. Heerschop, Goossen F., Hogenboom A., Frasincar F., Kaymak U., and F. de Jong. 2011. Polarity analysis of texts using discourse structure. In ACM, editor, *20th ACM international conference on Information and knowledge management*, pages 1061–1070.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In ACM, editor, *the tenth ACM SIGKDD international conference on Knowledge discovery and data mining.*

Richard Johansson and Alessandro Moschitti. 2013. Relational features in fine-grained opinion analysis. *Computational Linguistics.*

Soo-Min Kim and Eduard Hovy. 2004. Determining the sentiment of opinions. In ACL, editor, *the 20th international conference on Computational Linguistics.*

Soo-Min Kim and Eduard Hovy. 2006. Automatic identification of pro and con reasons in online reviews. In *COLING-ACL-06 Poster Session*, pages 483–490.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceeding of the 18th International Conf. on Machine Learning.* Morgan Kaufmann.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a Large Annotated Corpus of English: the Penn Treebank. *Computational Linguistics*, 19(2):313–330.

G. Minnen, J. Carroll, and D. Pearce. 2001. Applied morphological processing of english. *Natural Language Engineering*, 7(3):207–223.

Livia Polanyi and Annie Zaenen. 2006. Contextual valence shifters. In Springer Netherlands, editor, *Computing attitude and affect in text: Theory and applications.*, pages 1–10.

Ana-Maria Popescu and Orena Etzioni. 2007. Extracting product features and opinions from reviews. In Springer London, editor, *Natural language processing and text mining.*

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse Treebank 2.0. In *Proceedings of the 6th International Conference on Languages Resources and Evaluations (LREC 2008)*, Marrakech, Morocco.

Tjong Kim Sang, F. Erik, and Jorn Veenstra. 1999. Representing text chunks. In *In Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 173–179.

Fei Sha and Fernando Pereira. 2003. Shallow parsing with conditional random fields. In *Proceedings of HLT/NAACL*, pages 213–220.

Swapna Somasundaran and Janyce Wiebe. 2009. Recognizing stances in online debates. In *Proc. of ACL-IJCNLP-09*, pages 226–234.

Veselin Stoyanov and Claire Cardie. 2008. Topic identification for fine-grained opinion analysis. In Association for Computational Linguistics, editor, *22nd International Conference on Computational Linguistics*, volume 1.

Oscar Täckström and Ryan McDonald. 2011. Semi-supervised latent variable models for sentence-level sentiment analysis. In *ACL-11*, pages 569– 574.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. In *Language Resources and Evaluation*, volume 39, pages 165–210.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005a. Recognizing contextual polarity in phrase-level sentiment analysis. In Association for Computational Linguistics, editor, *Human Language Technology and Empirical Methods in Natural Language Processing.*

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005b. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 347–354.

Cäcilia Zirn, Mathias Niepert, Heiner Stuckenschmidt, and Michael Strube. 2011. Fine-grained sentiment analysis with structural features. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 336–344, Chiang Mai, Thailand, November. Asian Federation of Natural Language Processing.