Sentiment Polarity Classification with Low-level Discourse-based Features

Evgeny A. Stepanov, Giuseppe Riccardi
Signals and Interactive Systems Lab
Department of Information Engineering and Computer Science
University of Trento, Trento, TN, Italy
{evgeny.stepanov,giuseppe.riccardi}@unitn.it

Abstract

English. The poor state-of-the-art performances of discourse parsers prevent their application to downstream tasks. However, discourse parsing sub-tasks such as the detection of connectives and their sense classification have achieved satisfactory level of performance. In this paper we investigate the relevance of discourse connective features for tasks such as sentiment polarity classification. In the literature, discourse connectives are usually considered as modifiers of a polarity of a sentence or a word. In this paper we present experiments on using automatically extracted connectives and their senses as low-level features and as an approximation of a discourse structure for polarity classification of reviews. We demonstrate that, despite insignificant contributions to bag-of-words, the discourse-only models perform significantly above chance level.

Italiano. *Lo stato dell'arte degli analizzatori automatici del discorso impediscono la loro adozione nei contesti applicativi. Tuttavia, i sotto-processi automatici di analisi del discorso quali l'denticazione dei connettivi e la classicazione della loro etichetta semantica hanno comunque raggiunto un livello di prestazioni soddisfacente. In questo documento indaghiamo la rilevanza dei connettivi del discorso per i task come la classificazione della polarità dei sentimenti. In letteratura i connettivi del discorso sono comunemente considerati come modificatori della polarità di una frase o di una parola. In questo documento presentiamo alcuni esperimenti sull'estrazione automatica di connettivi, e relativi significati, e del loro utilizzo come caratteristiche di basso livello e come approssimazione della struttura di un discorso al fine di permettere la classificazione della polarità nelle recensioni. I connettivi del discorso assieme ai modelli bag-of-words permettono di ottenere risultati allo stato dell'arte e molto al di sopra dei modelli di base.*

1 Introduction

Discourse analysis has applications in many Natural Language Processing tasks; Webber et al. (2011) and Taboada and Mann (2006) among others list opinion mining, summarization, information extraction, essay scoring, etc. Availability of large discourse annotated resources such as Penn Discourse Treebank (PDTB) (Prasad et al., 2008a) and Rhetorical Structure Theory - Discourse Treebank (RST-DT) (Carlson et al., 2002) made it possible to develop statistical discourse parsers (e.g. (Marcu, 2000; Lin et al., 2014; Ghosh et al., 2011; Stepanov and Riccardi, 2013)). However, independent of the theory (RST or PDTB) the problem of end-to-end discourse parsing is far from being solved; thus, downstream application of these parsers yields mixed results.

In this paper we focus on PDTB approach to discourse parsing, which can be roughly partitioned into detection of discourse relations, extractions of their argument spans and sense classification. In CoNLL 2015 Shared Task on Shallow Discourse Parsing (Xue et al., 2015) the best system (Wang and Lan, 2015) achieved F_1 of 24 on the end-to-end parsing on a blind test set using strict evaluation that required exact match of all the spans and labels. Having such low end-to-end performances makes it difficult to apply PDTB-style discourse parsing to other NLP tasks. However, if we consider discourse parsing tasks individually, detection of discourse connectives and their classi-

Class	Type	Sub-Type
Comparison	*Contrast*	–
	Concession	–
Contingency	*Cause*	Reason Result
	Condition	–
Expansion	*Conjunction*	–
	Instantiation	–
	Restatement	–
	Alternative	– Chosen Alternative
	Exception	–
Temporal	*Synchronous*	–
	Asynchronous	Precedence Succession

Table 1: Simplified PDTB discourse relation sense hierarchy from CoNLL 2015 Shared Task.

fication into senses achieve high results: ≈ 90 for discourse connective detection and similarly ≈ 90 for connective sense classification (Stepanov et al., 2015). Thus, the output of these tasks could be used in other NLP applications.

Discourse connectives are essentially function words and phrases. Function word frequencies is a popular feature in NLP tasks such as authorship detection (Kestemont, 2014), and it has also been applied to sentiment polarity classification (Abbasi et al., 2008). Resolving connective usage and sense ambiguities (Section 2), they are potentially able to provide more refined features than simple function word counts. On the other hand, grouping connectives with respect to their senses yields more coarse features. In this paper we explore the utility of these features for sentiment polarity classification of movie reviews (Pang and Lee, 2004).

2 Discourse Connectives and Their Senses

In PDTB discourse relations are annotated using 3-level hierarchy of senses. The top level (level 1) senses are the most general: **Expansion**: one clause elaborates on the information given in another (e.g. 'and', 'in addition'); **Comparison**: there is a comparison or contrast between two clauses (e.g. 'but'); **Contingency**: there is a causal relationship between clauses (e.g. 'because'); and **Temporal**: two clauses are connected time-wise (e.g. 'before').

A relation signaled by a discourse connective is an *explicit* discourse relation. *Implicit* discourse relations between text segments (usually sentences), on the other hand, are inferred. The two classes are almost equally represented (53% vs. 47%). While detection of senses of *implicit* discourse relations is a hard problem (Lin et al., 2009; Xue et al., 2015); presence of a discourse connective in a sentence is sufficient for detection and classification of *explicit* discourse relations.

There are two levels of ambiguity present for a connective (Pitler and Nenkova, 2009): (1) it might be used to connect discourse units, or coordinate smaller constituents (e.g. 'and'); (2) some connectives might have different senses depending on usage (e.g. 'since' might signal causation or temporal relation). AddDiscourse tool was developed by (Pitler and Nenkova, 2009) to resolve these ambiguities. While using just connectives the 4-way sense classification accuracy of the tool is 93.67%, incorporating syntactic features raises performance to 94.15%; which is as good as the inter-annotator agreement on the same data (PDTB corpus - 94.00% (Prasad et al., 2008b)). Classification of discourse connectives into full depth of sense hierarchy also has an acceptable level of performance: 89.68% on PDTB development set of CoNLL 2015 Shared Task (Stepanov et al., 2015). For the Shared Task some senses were merged, and partial senses were disallowed (Xue et al., 2015); as a result, there are only 14 senses listed in Table 1. We classify discourse connectives identified by the addDiscourse tool further into this simplified hierarchy of senses.

3 Methodology

We test the utility of discourse connectives and their senses on sentiment polarity classification task. We follow the supervised machine approach and use SVM^{light} (Joachims, 1999) classifier with default parameter settings. A document is represented as a boolean vector of features (i.e. presence) and discourse-based features are added through vector fusion. Through out experiments 10-fold cross-validation is used, and results are reported as average accuracy, which is equivalent to micro- precision, recall, and F_1 for a binary classification where both classes are of interest.

3.1 Data Set

For the experiments we use the polarity dataset (v. 2.0) of (Pang and Lee, 2004), also known as Movie Reviews Data Set. The Data Set consists of 1,000 negative and 1,000 positive reviews extracted from the Internet Movie Database (IMDb).

3.2 Baseline Results

Using the 10-fold cross-validation split of (Pang and Lee, 2004), SVM unigram model achieves 86.25% average accuracy. Unlike the original paper, data set is used *as is*: no additional preprocessing such as frequency cut-off or prefixing the tokens following 'not', 'isn't', etc. till the first punctuation with 'NOT_' (Das and Chen, 2001) was used (same as (Stepanov and Riccardi, 2011)).

3.3 Representation of Discourse Connectives as Features

Function words were already used as features for polarity classification in (Abbasi et al., 2008), and the authors report that function words 'no' and 'if' tend to occur more frequently in negative reviews. Thus we experiment considering presence of connectives and their raw and normalized frequencies. Discourse connectives contain multi-word expressions (e.g. 'in_addition', 'on_the_other_hand', etc.), long-distance connective pairs (e.g. 'if_then', 'either_or'), and open class words (e.g. adverbs 'finally', 'similarly', etc.); and they are all treated as a single token.

Under these settings, we explore both the refinement and the generalization scenarios. In the refinement scenario discourse connective surface forms are appended with automatic Class (most general sense) or Sense decisions. and in the generalization scenario Class and Sense decisions replace the connective surface string. Consequently, we have 5 conditions, ordered from general to specific:

- **Class:** Class of a connective (one of 'Expansion', 'Comparison', 'Contingency', or 'Temporal');
- **Sense:** Sense of a connective from Table 1 (e.g. 'Temporal.Synchronous');
- **Surface:** Connective tokens (e.g. 'as');
- **Surface/Class:** Surface and Class tuple of a connective (e.g. 'as-Temporal' or 'as-Contingency');
- **Surface/Sense:** Surface and Sense tuple of a connective (e.g. 'as-Temporal.Synchronous' or 'as-Contingency.Cause.Reason');

In the following sections we evaluate these representations in isolation and fused into bag-of-words vectors.

Feature	B	R	N
BL: Chance	51.05		
Class	52.60	59.77	58.55
Sense	56.00	59.15	59.25
Surface	61.65	63.40	63.00
Surface/Class	61.35	**64.00**	63.15
Surface/Sense	61.20	63.65	62.70

Table 2: 10-fold cross-validation average accuracies for discourse connective as stand-alone features in comparison to the chance level baseline (*BL: Chance*). Results are reported for presence (B), raw (R) and normalized frequencies (N).

Additionally, our goal is to explore whether senses of explicit discourse relations alone can capture low-level discourse structure; and whether this low-level structure is beneficial for sentiment polarity classification. In order to approximate this, we use bigrams and trigrams of identified Classes and Senses. We introduce beginning and end of document tags to capture document initial and document final explicit relations. In this setting the presence of n-grams is considered, rather than the frequency. The setting is also evaluated in isolation and in fusion with bag-of-words.

4 Experiments and Results

In this section we present sentiment polarity classification experiments using discourse connective features under the settings defined in Section 3: (1) presence and frequencies as stand-alone features, (2) their effect on the bag-of-word model through vector fusion, and (3) effect of n-grams of Class and Senses in stand-alone and fusion settings.

4.1 Discourse Connectives as Stand-Alone Features

Table 2 presents the results of the experiments using discourse connectives as the only features. All the models, except Class presence (**B**), perform significantly above chance-level. Low performance of the Class-only model is expected, since there are only 4 Classes. As expected, the finer the features the better the performance. However, the Surface/Sense setting is lower than its more coarse version Surface/Class for all frequency count settings (statistically not significant). This is caused by Sense-level classifier's inferior performance, that often misses underrepresented senses of connectives.

Feature	B	R	N
BL: BoW	86.25		
Class	86.35	85.25	86.35
Sense	86.15	85.60	86.30
Surface	86.10	84.85	86.35
Surface/Class	85.95	84.90	86.35
Surface/Sense	85.85	84.90	86.35

Table 3: 10-fold cross-validation average accuracies for fusion of discourse connective features with bag-of-words (baseline: *BL: BoW*). Results are reported for presence (B), raw (R) and normalized frequencies (N).

The raw frequency counts perform better for all the representations, followed by normalized frequency counts. The boolean feature vector representation has the lowest performances. In the next Section we fuse these feature vectors with the boolean bag-of-words representation.

4.2 Fusion of Bag-of-Words and Discourse Connective Features

From the previous set of experiments we have observed that discourse connective only models perform above the chance level (even though much below the bag-of-words baseline). In order to investigate the effect of newly proposed discourse features, we fuse them with bag-of-words vectors (i.e. the baseline). The results of fusion are reported in Table 3.

From the results in Table 3 we can observe that the effect of feature fusion overall is insignificant. Raw frequency vectors generally have a negative effect on the performance. For boolean frequency vectors (i.e. presence), the more coarse features (Class and Sense) slightly improve the performance. For the normalized frequency count vectors, on the other hand, both more coarse and more refined features contribute to the performance. However, none of the improvements is statistically significant.

4.3 N-grams of Discourse Connective Senses

The results of experiments using n-grams of Class and Senses of connectives is reported in Table 4. The general observation is that increasing n-gram size has a positive effect on performance when discourse features are used stand-alone, and they are significantly above chance (except Class unigrams). The fusion of n-gram features and bag-of-word representation is also beneficial.

Feature	1	2	3
BL: Chance	51.05		
Class	52.60	58.35	59.55
Sense	56.00	57.40	58.80
BL: BoW	86.25		
BoW + Class	86.35	**86.85**	86.65
BoW + Sense	86.15	86.20	86.65

Table 4: 10-fold cross-validation average accuracies for discourse connective class and sense 1-3 grams and their fusion with bag-of-words. Only presence (boolean) of an n-gram is considered.

The best performing combination is fusion of bigrams of Classes and bag-of-words that achieves accuracy of 86.85. However, the improvement is statistically insignificant. But the fact that performances improve over the fusion of bag-of-words and frequency-based discourse connective vectors indicates that n-grams of explicit discourse relations are able to capture structures relevant for the sentiment polarity classification.

5 Conclusions

We have described experiments on using low-level discourse-based features for sentiment polarity classification. The general observations are (1) discourse connectives in isolation generally significantly outperform the chance baseline; and (2) using even the most general top-level senses provides performance gains. This is particularly notable due to the fact that discourse connective detection and relation sense classification do not generalize well across domains (Prasad et al., 2011).

Discourse connectives signal *explicit* discourse relations, which are only 53% of all discourse relations in PDTB. *Implicit* discourse relations (47%), which have the same senses, are much harder to deal with. Given the state of the art on *implicit* relation sense classification, detection and application of all the discourse relations is not yet possible. However, as indicated by the experiments on using n-grams of relation senses, even approximations can contribute.

Acknowledgments

The research leading to these results has received funding from the European Union – Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 610916 – SENSEI.

References

Ahmed Abbasi, Hsinchun Chen, and Arab Salem. 2008. Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. *ACM Transactions on Information Systems*, 26(3):12:1–12:34, June.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2002. RST Discourse Treebank (RST-DT) LDC2002T07.

Sanjiv Das and Mike Chen. 2001. Yahoo! for amazon: Extracting market sentiment from stock message boards. In *Proceedings of the 8th Asia Pacific Finance Association Annual Conference (APFA 2001)*.

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 2011)*.

Thorsten Joachims. 1999. Making large-scale svm learning practical. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*. MIT-Press.

Mike Kestemont. 2014. Function words in authorship attribution: From black magic to theory? In *The 3rd Workshop on Computational Linguistics for Literature (CLfL) @ EACL*, pages 59–66, Gothenburg, Sweden. ACL.

Ziheng Lin, Min-Yen Kan, and Hwee Tou Ng. 2009. Recognizing implicit discourse relations in the Penn Discourse Treebank. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP 2009)*.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2014. A PDTB-styled end-to-end discourse parser. *Natural Language Engineering*, 20:151 – 184.

Daniel Marcu. 2000. The rhetorical parsing of unrestricted texts: A surface-based approach. *Computational Linguistics*, 26:395–448.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the ACL*.

Emily Pitler and Ani Nenkova. 2009. Using syntax to disambiguate explicit discourse connectives in text. In *Proceedings of the ACL-IJCNLP Conference*, pages 13–16.

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

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008b. The penn discourse treebank 2.0. In *Proceedings of International Conference on Language Resources and Evaluation (LREC)*.

Rashmi Prasad, Susan McRoy, Nadya Frid, Aravind K. Joshi, and Hong Yu. 2011. The biomedical discourse relation bank. *BMC Bioinformatics*, 12:188.

Evgeny A. Stepanov and Giuseppe Riccardi. 2011. Detecting general opinions from customer surveys. In *IEEE ICDM Workshops (ICDMW) - Sentiment Elicitation from Natural Text for Information Retrieval and Extraction Workshop (SENTIRE)*, pages 115–122, Vancouver, BC, December. IEEE.

Evgeny A. Stepanov and Giuseppe Riccardi. 2013. Comparative evaluation of argument extraction algorithms in discourse relation parsing. In *The 13th International Conference on Parsing Technologies (IWPT 2013)*, pages 36–44, Nara, Japan, November.

Evgeny A. Stepanov, Giuseppe Riccardi, and Ali Orkan Bayer. 2015. The UniTN discourse parser in CoNLL 2015 shared task: Token-level sequence labeling with argument-specific models. In *The 19th SIGNLL Conference on Computational Natural Language Learning (CoNLL) - Shared Task*, pages 25–31, Beijing, China, July. ACL.

Maite Taboada and William C. Mann. 2006. Applications of rhetorical structure theory. *Discourse Studies*, 8(4):567–88.

Jianxiang Wang and Man Lan. 2015. A refined end-to-end discourse parser. In *Proceedings of the The SIGNLL Conference on Computational Natural Language Learning*, Beijing, China, July. ACL.

Bonnie L. Webber, Markus Egg, and Valia Kordoni. 2011. Discourse structure and language technology. *Natural Language Engineering*, pages 437–490.

Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Rashmi Prasad, Christopher Bryant, and Attapol Rutherford. 2015. The CoNLL-2015 Shared Task on Shallow Discourse Parsing. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning: Shared Task*.