Discourse and Subjectivity in Sentiment Analysis

Christopher Hidey
Department of Computer Science
Columbia University
New York, NY 10027
chidey@cs.columbia.edu

1 Introduction

Research in sentiment analysis has often been concerned with how the polarity of the words may appear to change but in fact reflect the expression of the author in an alternative manner (Pang and Lee, 2008). Simple lexical approaches fail to take into account the context in which the sentiment is expressed. Sarcasm, for example, may be used to express negative sentiment, but a lexical approach would identify mostly positive words. Negation is another aspect of sentiment where it is often difficult to identify the scope of where the negation applies. Lastly, the author of a text may express their opinion relatively to something that was previously written. This may take the form of expanding a new idea, juxtapose a good example with a negative one, or providing an explanation. Discourse structure indicates relations between clauses or sentences such as comparison and contrast. These relations are used by authors to help with coherence, the way a text is structured for readability. Identifying these relations can help a sentiment classifier to learn when the context has changed and how to represent sentiment differently.

Although comparison and contrast is the discourse relation most expected to indicate changes in sentiment, it is unclear what role temporal, causal, and expansion relations may play. Each of these discourse relations may appear explicitly, marked by a discourse connective such as "but", "then", "because", or "also." The primary objective of this project is to analyze the role of discourse relations and possible improvements in sentiment analysis over previous work. This research builds on existing work using linguistic connectives that may identify similar or contrasting sentiment. Previous work has focused on using explicit markers that identify discourse relations across sentences. However, there has not been much research on polarity accounting for implicit discourse relations or relations within sentences. Identification of these implicit relations is helped by the use of word pairs- cross products of words that are separated by a discourse marker. Identification of sentence polarity containing an explicit connective within the sentence may also be improved by using word pairs from sentences marked with polarity. The hypothesis is that using word pairs for within-sentence and implicit discourse relations can improve the identification of sentence and document level sentiment.

2 Related Work

Lately, there has been increased interest in taking advantage of properties of discourse to help identify sentiment. Zhou focuses on predicting sentiment analysis at the discourse level (Zhou, 2013). Lazaridou et al. have also created a model for unsupervised joint inference for discourse and sentiment using Bayesian networks (Lazaridou et al., 2013).

The current state-of-the-art for sentiment analysis at the sentence level uses recursive neural tensor networks (RNTNs) to model sentiment at the level of individual nodes in a parse tree (Socher et al., 2013). This approach provides an improved method for handling negation and scope at various tree levels. This method does not explicitly focus on discourse relations but some of these relations may be

correctly handled at the appropriate node in the tree. However, RNTNs require an annotated treebank to train and parsed data to test. As parsing and sentiment are often domain-specific, there is still a need for quickly creating new models.

Early work in sentiment analysis recognized the use of common discourse markers to identify adjectives with similar or contrasting polarity (Hatzivassiloglou and McKeown, 1997). Hatzivassiloglou and McKeown created a classifier that takes advantage of the linguistic intuition that adjectives that are conjoined with "and" have similar sentiment and adjectives conjoined with "but" have opposite sentiment. The authors created an initial set of adjectives manually annotated with positive or negative polarity from the Wall Street Journal Corpus. They used annotated pairs to train and test two classifiers: a logistic regression classifier and a rule-based classifier. Then they used the classifier to cluster the adjectives into one of two classes. Because these clusters have no sense of positive or negative, they assigned the positive class to the one with the highest average frequency. They found that it is possible to identify the polarity of these adjectives with high accuracy.

Recent work on sentiment analysis for Twitter has looked at the use of "lightweight" discourse features (Mukherjee and Bhattacharyya, 2012). Researchers focused on models for unstructured, noisy text because many lexical models are trained on structured text and perform poorly out of domain. They identify a list of discourse connectives and semantic operators which may affect the polarity of a clause and create an algorithm to harness this information and weight the polarity according to the discourse information. They used a lexicon based system and train a support vector machine (SVM) in a supervised framework. They also tested their model on structured text (travel reviews) and found that it performs well.

Some researchers created a model for document level sentiment using latent sentence subjectivity (Yessenalina et al., 2010). They state that when using only annotator rationales generated by human judges to support the document level sentiment that it is possible to obtain much higher accuracy than using the full document. Their claim is that this is analogous to only using subjective sentences. However, documents marked with sentiment that also have individual sentences annotated with subjectivity are difficult to obtain so they model the subjectivity as a latent variable using a latent structured SVM. At training time, they attempt to find the weights that maximize:

$$\hat{w} = arg\max_{w} \sum_{t} \max_{h} w^T f(y_t, x_t, h)$$

The feature vector f consists of polarity and subjectivity features ψ_{pol} and ψ_{subj} and they design these features to be orthogonal such that $\psi_{pol}^T \psi_{subj} = 0$. They also include transition features for subjectivity in ψ_{subj}.

Other researchers furthered this work by including discourse features (Trivedi and Eisenstein, 2013). They use explicit discourse connectives to identify when there is a change in polarity. For their model, they use a latent structured SVM to train a classifier on movie reviews, using sentence subjectivity as a latent variable. Similarly, the feature vector f is composed of several subsets of features: polarity and subjectivity features based on a bag-of-words model and they also include transition features based on whether certain discourse connectives are present.

The work of (Trivedi and Eisenstein, 2013) is most similar to this project. In addition to the features used for implicit markers across sentences, this work introduces features for identifying implicit discourse relations as well. Much of the work using discourse relations has focused on exploiting structure when an explicit marker is present. Implicit discourse relations are much more difficult to identify than the explicit relations (Pitler et al, 2009). However, the performance on identifying implicit relations has improved by making use of features other than lexical ones, using syntax, semantics, or features derived using distributional methods. For this project, additional features include the use of one of these distributional methods: word pairs, which have been used extensively in discourse analysis.

Furthermore, this work also builds on the work of (Mukherjee and Bhattacharyya, 2012). While they identified markers that affect polarity using linguistic intuition, this research develops methods to identify these connectives automatically. However, they identify markers that are likely to flip, increase, or decrease polarity while this research only considers

markers likely to flip the polarity and save the rest for future work.

One approach to identifying discourse relations involves the use of word pairs created from the cross product of words that span a known discourse connective. Early work derived these word pairs from training data (Marcu, 2001) for the Penn Discourse Tree Bank (PDTB) (Prasad et al, 2008). Each word pair was used as a separate feature in a classifier and improvements in identifying implicit relations were gained. However, this approach results in very sparse vectors used as features. The lack of adequate data for all possible pairs of words requires the model to make inferences it cannot.

Later work looked at using aggregated word pairs as features (Biran et al., 2011), (Biran and McKeown, 2013). Instead of using word pairs derived from a training set, researchers used the Gigaword corpus to create counts of pairs of words across each of the 102 explicit discourse markers listed by the PDTB and normalized the counts using TF-IDF. Then during training, when an explicit marker is not present, word pairs are created from the cross product of all subsequent words and the cosine similarity between this new vector and each of the 102 word pairs is used as a feature.

It should also be noted that although many argument spans for discourse relations occur entirely within a single sentence (intra-sentence), previous work in sentiment mostly focused on relations between sentences (inter-sentence).

3 Data

Two data sets were used for this research. The first is the Sentiment140 corpus (Go et al., 2009), a data set weakly marked with polarity. This corpus was created by extracting 1.6 million Tweets with emoticons. The authors identified several "positive emoticons" indicating some form of happiness, hypothesizing that this indicates positive sentiment being expressed. Similarly, they identified negative sentiment using emoticons of various emotional states indicating sadness.

Of these Tweets, 33% have one of the 102 discourse connectives from the PDTB. For each discourse marker a separate data set was created, stemming individual words and balancing the positive and negative instances by subsampling. For testing, either 5,000 Tweets or 50% of the data was set aside for a test set, whichever is less.

In all 44 discourse connectives occur in these Tweets. This is unsurprising, given the source of the data. Intuitively, "as a consequence" or "subsequently" are expected to appear rarely in such informal text.

The full distribution is given in appendix A.

The second data set consists of longer documents marked with polarity. This corpus is a set of movie reviews derived from the Internet Movie Database (IMDB) (Maas et al., 2011). These movie reviews range in length from one sentence to a few paragraphs. The dataset as previously used is split into positive and negative reviews, where a positive review has a rating between 7 and 10 inclusive and a negative review has a rating between 1 and 4 inclusive. There are 50,000 movie reviews marked with polarity, evenly balanced between positive and negative, with 25,000 set aside for testing. There are an additional 50,000 unmarked movie reviews for unsupervised learning which were not used for this research. Furthermore, 5,000 reviews were set aside for development and parameter tuning. Each review has a numerical identification number, and the development set consists of the last 2,500 numerically from each of the positive and negative reviews.

4 Methodology

As mentioned, previous work has not focused on machine learning approaches for determining the importance of individual connectives in determining sentiment. One possibility for determining the importance of these words is to attempt to identify where long-range context is more important in identifying sentiment than individual words. The cross product of words across a discourse connective provides a set of features that may correlate to changes in sentiment. If the discourse connective indicates a transition in polarity, then the distribution of words on either side of the connective might be different. Thus, if a classifier trained using word pairs as binary features significantly outperforms a classifier using only words as binary features, it may indicate a transition between different distributions.

Using the data set extracted from the Senti-

ment140 corpus, models were created for each discourse marker using two sets of features. The first set of features consists solely of bag-of-words binary features and the second set also includes word-pair binary features. These models were trained using stochastic gradient descent with hinge loss, which corresponds to a linear support vector machine. Hyperparameters for the model were determined by grid search and cross-validation using the training set. These hyperparameters include the learning rate, the number of epochs, and α, the constant that multiplies the regularization term.

Using the binomial cumulative distribution function, p-values can be calculated for the difference in performance on the test set between the two sets of features. Then each discourse connective can be ranked according to their p-values, which should indicate their importance in determining sentiment.

The results of these classifiers inform the feature selection for the document-level classifier. Using the latent structured SVM requires determining two sets of features: polarity features and subjectivity features. The subjectivity features are used to identify the subset of sentences in the document that are most subjective and most likely to indicate the polarity of the document. The polarity features contribute to the classification of individual sentences and the overall document classification.

Because the training of the SVM is an iterative process it requires some initialization for the hidden variables (the indices of the subjective sentences). Because the optimization function is non-convex, previous work (Yessenalina et al., 2010) found success in using OpinionFinder (Wilson et al., 2005) to set the initial sentences. Other practical considerations include the extraction of features from the movie reviews. Each lower-cased word, rather than lemma or stem, in a sentence is assigned as a binary feature for that sentence. Furthermore, each feature vector is normalized to be unit length so that longer sentences are not weighted more heavily than shorter sentences. Experiments were performed with both normalized and un-normalized vectors and normalization improved the results.

Moreover, discourse connectives between sentences were used to derive binary subjectivity features (Trivedi and Eisenstein, 2013). In all, 204 additional features were created. For each discourse marker, there is an indicator of whether the previous and current sentence have the same level of subjectivity ("shift" or "continue"). Thus, a feature such as "(but, shift)" is set to 1 if that connective is present and there is a change in subjectivity and 0 otherwise.

As additional features, several discourse-motivated subjectivity features were included. First, the word pairs created in previous research (Biran and McKeown, 2013) were used in a similar manner to the explicit discourse connectives. However, because some discourse connectives occur very infrequently, only 59 of them have a corresponding word pair model. Thus, this set of features includes 118 additional features. For each sentence, the cosine similarity is calculated between the current and previous sentence and the word pairs for each of the 59 discourse connectives. Then, a feature such as "(but,shift)" is set to the corresponding cosine score if there is a change in subjectivity and 0 otherwise.

Second, another set of features is used to compare subjective and objective sentences. Using the cosine similarity model trained on newswire data, the score from the current sentence is compared to the score from a cosine similarity model trained on the Sentiment140 data, under the assumption that if a pair of sentences is more similar to newswire then it may be objective, as news is. Word pair models were created from the Sentiment140 data using the separate discourse data sets and normalized using TF-IDF (where a "document" is a collection of word pairs for a discourse marker). Then 3 binary features were created for each discourse marker: if the Sentiment140 model is greater, if the newswire model is greater, and if they are equal (most likely only occurs when they are both 0).

Finally, two sets of polarity features were used in experiments: the binary classifiers with word-pair features created from Sentiment140 on each connective data set and the top K classifiers according to their p-values. This results in up to 88 additional features, two for each discourse connective present in Twitter. For each discourse connective there are two binary features indicating whether the classifier identifies the sentence as positive or negative polarity. Each feature is set to 1 if the discourse connective is present and polarity value is assigned, and 0 otherwise (thus only 1 of these features can be 1 for

any given sentence).

5 Results

Table 1 shows the relative performance of the two different classifiers on the Twitter data, ordered by p-value.

Connective	Unigram	Word Pair	p-value
then	0.587	0.645	5.018e-14
when	0.608	0.658	3.675e-13
or	0.652	0.685	9.106e-06
but	0.63	0.658	1.820e-05
as	0.636	0.664	8.076e-05
still	0.604	0.632	0.00024
yet	0.56	0.602	0.00054
now	0.651	0.669	0.00374
next	0.605	0.625	0.00936

Table 1: Accuracy of Connective Sentiment Models on Sentiment140

As expected several contrastive markers such as "but","yet", or "still" occur near the top of the list. However, temporal markers such as "when" or "then" occur at the very top. Although these markers may not intuitively indicate a change in sentiment, there is some justification for why these markers would appear. These temporal markers indicate a transition between events, and it may be the case that a positive event is followed by a negative event, or vice versa. These temporal changes are significant enough to warrant further study.

Also interesting is the fact that the model trained only on discourse connectives with "but" outperforms a bag-of-words model using only data without a discourse connective, even though this model was trained on significantly more data. This inspired further work to attempt to create an improved model on all data accounting for connectives. For training/tuning, all the training and testing sets from each of the discourse connectives were combined, respectively. The remaining Tweets without a discourse connective were split with 50% in training and 50% in testing. Two models were compared: one with bag-of-words and another incorporating word pairs. There is a separate set of word pair features for each connective, so those features would only fire if that connective is in the sentence and the word pair appears. The initial bag-of-words only model achieves 70.1% accuracy and the model with word pairs achieves 70.7% accuracy, so there is a small but measurable improvement on all data when accounting for discourse.

Table 2 shows the results of the feature sets for the document-level polarity classifier. The initial column includes the accuracy after one iteration and the final column shows the accuracy after 10 iterations.

Model	Initial	Final
unigram:	85.50	88.16
markers:	85.98	88.48
best features:	85.76	89.04

Table 2: Accuracy of Document Sentiment Models on IMDB

Previous work is used for baselines. One baseline (*unigrams*) includes the bag-of-words features for subjectivity and polarity (Yessenalina et al., 2010) and another baseline (*markers*) includes the additional binary subjectivity features that indicate when a discourse marker is present and the subjectivity of the current sentence is either the same or different from the previous sentence (Trivedi and Eisenstein, 2013).

The best features include the unigrams and markers and were also determined to include the implicit word pair features and the top 10 sentiment discourse models from Twitter. Including the subjective versus objective features did not results in an increase in accuracy. One issue may be data sparsity and the newswire model may score sentences higher simply because it was trained on significantly more data, meaning that the Twitter model would have scores of 0 much more frequently.

The improvement using the best features is statistically significant compared to the unigram-only features and the marker features because of the large test set of 25,000 movie reviews ($p < .01$ with the binomial CDF).

6 Conclusion

Overall, the results show that accounting for discourse yields improvements for polarity analysis both within and across sentences. The results of the Sentiment140 experiments show that there is poten-

tial for identifying transition words using word pairs. Furthermore, there is some indication that temporal markers may indicate a change of state from one event to another and thus a change in sentiment.

The experiments on the movie reviews data show less promise. As some of the discourse features require jointly determining both the discourse class and subjectivity of a pair of sentences there may be too many errors to have any real effect. However, even though the polarity features were created from data from a slightly different domain, there was still a positive effect for including them. Furthermore, including only the top 10 discourse connectives yielded an additional increase, lending some credence to the method for identifying the influence of the connectives.

Obtaining and modeling data for individual discourse connectives is relatively low resource and could be applied to other languages and domains. Extracting Twitter data requires knowledge of emoticons or other markers for polarity and creating connective-specific data sets requires knowledge of discourse connectives but the amount of data and time to model is not intensive.

6.1 Problems

The main problem encountered during these experiments was reproducibility. Although the code for the latent structured SVM was available (Yessenalina et al., 2010), the code and data for follow-on experiments was not (Trivedi and Eisenstein, 2013). Thus in order to run experiments, the IMDB corpus needed to be obtained separately and code written to extract features and output them in a format that the structured SVM could handle. Furthermore, the original code only includes training an SVM for one iteration. In order to take advantage of the subjective sentence selection, the code needs to 1) identify subjective sentences, 2) train weights for subjectivity and polarity using these sentences, and 3) use the new subjective weights for step 1. When only bag-of-words features are considered, the extracted features do not change. However, when features are dependent on the value of the hidden variable (subjective or objective), they need to be recalculated after every iteration as their value may change. Thus there were several possibilities for the introduction of errors during the feature extraction and training, which may have hindered reproducibility.

Second, there were other practical considerations that were not addressed in any of the previous work. It is possible for the model to identify no subjective sentences in a document. OpinionFinder identifies 996 of the 25,000 training documents as having no subjective sentences, which is significant and raises questions over how to handle these data points. Ignoring them completely is an option, but as the number of these data points may change between iterations, it is not possible to have a fair comparison across iterations. Another possibility is to include only the first sentence or last sentence or some random set of sentences. Ultimately for these experiments if this was the case, all sentences were included, so that the model could later determine which sentences are most subjective.

Furthermore, there are issues with the data set that should be addressed in future work. The first is the use of the weakly supervised data as a test set. Although this data may be a strong indicator of polarity, it is likely not perfect, and the experiments would benefit from having an annotated data set. However, the data would need to specifically include examples containing discourse connectives and if these are sampled naturally that means the data should be large enough to contain these examples with connectives, which would require significant time to annotate.

Additionally, Twitter is very noisy, filled with intentional and unintentional variations of words. There may be other discourse connectives that are not captured using the extraction method of exact matching. For example, “because” varies informally as “cause” or “cuz” and there may be other markers with variations. Some study could be done to determine other possible connectives or the data could be normalized so that these markers appear as they should.

6.2 Future Work

One issue with the word pairs approach is that it results in very sparse feature vectors. Since the only concern is the polarity of data rather than the discourse class, it may be possible to aggregate the different data sets together. The differences in contexts between “yet” and “but” may not matter for this task and all their examples could be combined together

for modeling. One possibility is a constrained clustering approach that requires the data points with "but" in them to remain in that class and not appear in the same class as "because" data points.

Alternatively, future work could involve finding a dense, low-rank approximation for these feature vectors. Since some word pairs occur very rarely (maybe only once), it may be possible to use word embeddings to predict the feature value for two words that have never appeared in a cross product (Mikolov et al., 2013).

Finally, it should be noted that this method is not limited to discourse connectives. Discourse connectives just provide a natural linguistic starting point for words that indicate transitions. This method could also be expanded to identify other words given some other heuristic for possible transition words, as it is time-consuming to create models for every word in the data set.

References

Or Biran, Samuel Brody, and Noemie Elhadad. 2011. *Putting it Simply: a Context-Aware Approach to Lexical Simplification*. Proceedings of ACL.

Or Biran and Kathleen McKeown. 2013. *Aggregated Word Pair Features for Implicit Discourse Relation Disambiguation*. In proceedings of ACL 2013, Sofia, Bulgaria.

Alec Go, Richa Bhayani, and Lei Huang. 2009. *Twitter Sentiment Classification using Distant Supervision*. Technical report, Stanford.

Weiwei Guo and Mona Diab. 2012. *Modeling Sentences in the Latent Space*. In Proceedings of ACL.

Vasileios Hatzivassiloglou and Kathleen R. McKeown. 1997. *Predicting the semantic orientation of adjectives*. In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics, 174-181.

Angeliki Lazaridou, Ivan Titov, and Caroline Sporleder. 2013. *A Bayesian Model for Joint Unsupervised Induction*. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 1630-1639.

Maas, Andrew L. and Daly, Raymond E. and Pham, Peter T. and Huang, Dan and Ng, Andrew Y. and Potts, Christopher. 2011. *Learning Word Vectors for Sentiment Analysis*. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pages 142-150.

Daniel Marcu and Abdessamad Echihabi 2001. *An Unsupervised Approach to Recognizing Discourse Relations*. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. *Distributed Representations of Words and Phrases and their Compositionality*. In Proceedings of NIPS.

Subhabrata Mukherjee and Pushpak Bhattacharyya. 2012. *Sentiment Analysis in Twitter with Lightweight Discourse Analysis*. Proceedings of COLING 2012: Technical Papers, pages 18471864.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval 2(1-2), pp. 1135.

Emily Pitler and Ani Nenkova. 2009. *Using Syntax to Disambiguate Explicit Discourse Connectives in Text*. Proceedings of the ACL-IJCNLP 2009 Conference Short Papers. Suntec, Singapore, 4 August 2009.

Emily Pitler, Annie Louis, and Ani Nenkova. 2009. *Automatic sense prediction for implicit discourse relations in text* Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, Suntec, Singapore, 4 August 2009.

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 Language Resources and Evaluation (LREC). Marrakech, Morocco.

Richard Socher, Alex Perelygin, Jean Y. Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng and Christopher Potts. 2013. *Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank*. In Proceedings of EMNLP.

Rakshit Trivedi and Jacob Eisenstein. 2013. *Discourse Connectors for Latent Subjectivity in Sentiment Analysis*. Proceedings of NAACL-HLT 2013, pages 808-813.

Theresa Wilson, Paul Hoffmann, Swapna Somasundaran, Jason Kessler, Janyce Wiebe, Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. 2005. *Opinionfinder: A system for subjectivity analysis*. In Proceedings of HLT-EMNLP: Interactive Demonstrations.

Ainur Yessenalina, Yisong Yue, and Claire Cardie. 2010. *Multi-Level structured models for Document Level Sentiment Classification*. In Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP).

Yudong Zhou. 2013. *Fine-grained Sentiment Analysis with Discourse Structure*. Master Thesis, Saarland University.

Appendices

A : Distribution of Tweets Containing a Discourse Connective

Connective	Positive Tweets	Negative Tweets	Total
after	5829	6531	12360
also	3018	2197	5215
although	425	500	925
and	104735	105164	209899
as	12313	10741	23054
at that time	23	30	53
at the same time	189	241	430
at the time	54	61	115
but	29926	53837	83763
earlier	655	966	1621
even though	211	330	541
first	5829	4020	9849
for	88476	70344	158820
further	90	140	230
in comparison	10	8	18
in fact	117	76	193
in particular	38	22	60
in return	22	9	31
in short	21	27	48
instead	874	1406	2280
in the end	110	99	209
in turn	8	14	22
later	3430	2112	5542
nevertheless	16	15	31
next	5829	5822	11651
nonetheless	33	10	43
now	22476	26915	49391
on the other hand	43	41	84
on the whole	14	15	29
or	11667	10085	21752
plus	613	670	1283
rather	848	1009	1857
regardless	58	35	93
second	1223	1127	2350
so	41042	52950	93992
soon	5124	4914	10038
still	8027	15704	23731
that is	2163	1587	3750
then	8409	11113	19522
third	225	259	484
though	4840	7106	11946
thus	85	89	174
what's more	3	3	6
when	12106	15573	27679
while	3366	3337	6703
yet	2961	4612	7573