Improving Implicit Discourse Relation Recognition Through Feature Set Optimization

Joonsuk Park
Department of Computer Science
Cornell University
Ithaca, NY, USA
jpark@cs.cornell.edu

Claire Cardie
Department of Computer Science
Cornell University
Ithaca, NY, USA
cardie@cs.cornell.edu

Abstract

We provide a systematic study of previously proposed features for *implicit discourse relation identification*, identifying new feature *combinations* that optimize F_1-score. The resulting classifiers achieve the best F_1-scores to date for the four top-level discourse relation classes of the Penn Discourse Tree Bank: COMPARISON, CONTINGENCY, EXPANSION, and TEMPORAL. We further identify factors for feature extraction that can have a major impact on performance and determine that some features originally proposed for the task no longer provide performance gains in light of more powerful, recently discovered features. Our results constitute a new set of baselines for future studies of implicit discourse relation identification.

1 Introduction

The ability to recognize the discourse relations that exist between arbitrary text spans is crucial for understanding a given text. Indeed, a number of natural language processing (NLP) applications rely on it — e.g., question answering, text summarization, and textual entailment. Fortunately, *explicit discourse relations* — discourse relations marked by explicit connectives — have been shown to be easily identified by automatic means (Pitler et al., 2008): each such connective is generally strongly coupled with a particular relation. The connective "because", for example, serves as a prominent cue for the CONTINGENCY relation.

The identification of *implicit discourse relations* — where such connectives are absent — is much harder. It has been the subject of much recent research since the release of the Penn Discourse Treebank 2.0 (PDTB) (Prasad et al., 2008), which annotates relations between adjacent text spans in Wall Street Journal (WSJ) articles, while clearly distinguishing *implicit* from *explicit* discourse relations.[1] Recent studies, for example, explored the utility of various classes of features for the task, including linguistically informed features, context, constituent and dependency parse features, and features that encode entity information or rely on language models (Pitler et al., 2009; Lin et al., 2009; Louis et al., 2010; Zhou et al., 2010).

To date, however, there has not been a systematic study of combinations of these features for implicit discourse relation identification. In addition, the results of existing studies are often difficult to compare because of differences in data set creation, feature set choice, or experimental methodology.

This paper provides a systematic study of previously proposed features for implicit discourse relation identification and identifies feature combinations that optimize F_1-score using forward selection (John et al., 1994). We report the performance of our binary (one vs. rest) classifiers on the PDTB data set for its four top-level discourse relation classes: COMPARISON, CONTINGENCY, EXPANSION, and TEMPORAL. In each case, the resulting classifiers achieve the best F_1-scores for the PDTB to date. We

[1] Research on implicit discourse relation recognition prior to the release of the PDTB instead relied on synthetic data created by removing explicit connectives from explicit discourse relation instances (Marcu and Echihabi, 2002), but the trained classifiers do not perform as well on real-world data (Blair-Goldensohn et al., 2007).

Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL), pages 108–112,
Seoul, South Korea, 5-6 July 2012. ©2012 Association for Computational Linguistics

further identify factors for feature extraction that can have a major impact performance, including stemming and lexicon look-up. Finally, by documenting an easily replicable experimental methodology and making public the code for feature extraction[2], we hope to provide a new set of baselines for future studies of implicit discourse relation identification.

2 Data

The experiments are conducted on the PDTB (Prasad et al., 2008), which provides discourse relation annotations between adjacent text spans in WSJ articles. Each training and test instance represents one such pair of text spans and is classified in the PDTB w.r.t. its **relation type** and **relation sense**.

In the work reported here, we use the **relation type** to distinguish examples of *explicit* vs. *implicit* discourse relations. In particular, we consider all instances with a relation type other than *explicit* as implicit relations since they lack an explicit connective between the text spans. The **relation sense** determines the relation that exists between its text span *arguments* as one of: COMPARISON, CONTINGENCY, EXPANSION, and TEMPORAL. For example, the following shows an explicit CONTINGENCY relation between *argument1* (arg1) and **argument2** (arg2), denoted via the *connective* "because":

> *The federal government suspended sales of U.S. savings bonds because* **Congress hasn't listed the ceiling on government debt.**

The four relation senses comprise the target classes for our classifiers.

A notable feature of the PDTB is that the annotation is done on the same corpus as Penn Treebank (Marcus et al., 1993), which provides parse trees and part-of-speech (POS) tags. This enables the use of gold standard parse information for some features, e.g., the *production rules* feature, one of the most effective features proposed to date.

3 Features

Below are brief descriptions of features whose efficacy have been empirically determined in prior works[3], along with the rationales behind them:

Word Pairs (cross product of unigrams: arg1 × arg2) — A few of these word pairs may capture information revealing the discourse relation of the target spans. For instance, *rain-wet* can hint at CONTINGENCY.

First-Last-First3 (the first, last, and first three words of each argument) — The words in this range may be expressions that function as connectives for certain relations.

Polarity (the count of words in arg1 and arg2, respectively, that hold negated vs. non-negated positive, negative, and neutral sentiment) according to the MPQA corpus (Wilson et al., 2005)) — The change in sentiment from arg1 to arg2 could be a good indication of COMPARISON.

Inquirer Tags (negated and non-negated fine-grained semantic classification tags for the verbs in each argument and their cross product) — The tags are drawn from the General Inquirer Lexicon (Stone et al., 1966)[4], which provides word level relations that might be propagated to the target spans' discourse relation, e.g., rise:fall.

Verbs (count of pairs of verbs from arg1 and arg2 belonging to the same Levin English Verb Class (Levin and Somers, 1993)[5], the average lengths of verb phrases as well as their cross product, and the POS of the main verb from each argument) — Levin Verb classes provide a means of clustering verbs according to their meanings and behaviors. Also, longer verb phrases might correlate with CONTINGENCY, indicating a justification.

Modality (three features denoting the presence of modal verbs in arg1, arg2, or both) — Modal verbs often appear in CONTINGENCY relations.

Context (the connective and the sense of the immediately preceding and following relations (if explicit), and a feature denoting if arg1 starts a paragraph) — Certain relations co-occur.

Production Rules (three features denoting the presence of syntactic productions in arg1, arg2 or both, based on all pairs of parent-children nodes in the argument parse trees) — The syntactic structure of an argument can influence that of the other argument as

[2] These are available from http://www.joonsuk.org.

[3] Word Pairs (Marcu and Echihabi, 2002). First-Last-First3 (Wellner et al., 2006). Polarity, Verbs, Inquirer Tags, Modality, Context (Pitler et al., 2009). Production Rules (Lin et al., 2009).

[4] http://www.wjh.harvard.edu/ inquirer/inqdict.txt

[5] http://www-personal.umich.edu/ jlawler/levin.html

well as its relation type.

4 Experiments

We aim to identify the optimal subsets of the aforementioned features for each of the four top-level PDTB discourse relation senses: COMPARISON, CONTINGENCY, EXPANSION, and TEMPORAL. In order to provide a meaningful comparison with existing work, we carefully follow the experiment setup of Pitler et al. (2009), the origin of the majority of the features under consideration:

First, sections 0-2 and 21-22 of the PDTB are used as the validation and test set, respectively. Then, we randomly down-sample sections 2-20 to construct training sets for each of the classifiers, where each set has the same number of positive and negative instances with respect to the target relation. Since the composition of the corresponding training set has a noticeable impact on the classifier performance we select a down-sampled training set for each classifier through cross validation. All instances of non-explicit relation senses are used; the ENTREL type is considered as having the EXPANSION sense.[6]

Second, Naive Bayes is used not only to duplicate the Pitler et al. (2009) setting, but also because it equaled or outperformed other learning algorithms, such as SVM and MaxEnt, in preliminary experiments, while requiring a significantly shorter training time.[7]

Prior to the feature selection experiments, the best preprocessing methods for feature extraction are determined through cross validation. We consider simple lowercasing, Porter Stemming, PTB-style tokenization[8], and hand-crafted rules for matching tokens to entries in the polarity and General Inquirer lexicons.

Then, feature selection is performed via forward selection, in which we start with the single best-performing feature and, in each iteration, add the feature that improves the F_1-score the most, until no significant improvement can be made. Once the optimal feature set for each relation sense is determined by testing on the validation set, we retrain each classifier using the entire training set and report final performance on the test set.

5 Results and Analysis

Table 5 indicates the performance achieved by employing the feature set found to be optimal for each relation sense via forward selection, along with the performance of the individual features that constitute the ideal subset. The two bottom rows show the results reported in two previous papers with the most similar experiment methodology as ours. The notable efficacy of the *production rules* feature, yielding the best or the second best result across all relation senses w.r.t. both F_1-score and accuracy, confirms the finding of Zhou et al. (2010). In contrast to their work, however, combining existing features enhances the performance. Below, we discuss the primary observations gleaned from the experiments.

Word pairs as features. Starting with earlier works that proposed them as features (Marcu and Echihabi, 2002), some form of *word pairs* has generally been part of feature sets for implicit discourse relation recognition. According to our research, however, these features provide little or no additional gain, once other features are employed. This seems sensible, since we now have a clearer idea of the types of information important for the task and have developed a variety of feature types, each of which aims to represent these specific aspects of the discourse relation arguments. Thus, general features like *word pairs* may no longer have a role to play for implicit discourse relation identification.

Preprocessing. Preprocessing turned out to impact the classifier performance immensely, especially for features like *polarity* and *inquirer tags* that rely on information retrieved from a lexicon. For these features, if a match for a given word is not found in the lexicon, no information is passed on to the classifier.

As an example, consider the General Inquirer lexicon. Most of its verb entries are present tense singular in form; thus, without stemming, dictionary look up fails for a large portion of the verbs. In our case, the F_1-score increases by roughly 10% after stemming.

Further tuning is possible by a few hand-written

[6] Some prior work uses a different experimental setting. For instance, Zhou et al. (2010) only considers two of the non-explicit relations, namely *Implicit* and *NoRel*.

[7] We use classifiers from the nltk package (Bird, 2006).

[8] Stanford Parser (Klein and Manning, 2003).

Feature Type	COMP. vs Rest		CONT. vs Rest		EXP. vs Rest		TEMP. vs Rest	
	F_1	Acc.	F_1	Acc.	F_1	Acc.	F_1	Acc.
1. Polarity	16.49	46.82	28.47	61.39	64.20	56.80	13.58	50.69
2. First-Last-First3	22.54	53.05	37.64	66.71	62.27	56.40	15.24	51.81
3. Inquirer Tags	18.07	82.14	34.88	69.60	77.76	66.38	21.65	80.04
4. Verbs	18.05	55.29	23.61	78.33	68.33	58.37	18.11	58.44
5. Production Rules	30.04	75.84	47.80	71.90	77.64	69.60	20.96	63.36
Best Combination	2 & 4 & 5		2 & 4 & 5		1 & 3 & 4 & 5		1 & 3 & 5	
	31.32	74.66	49.82	72.09	79.22	69.14	26.57	79.32
Pitler '09 (Best)	21.96	56.59	47.13	67.30	76.42	63.62	16.76	63.49
Zhou '10 (Best)*	*31.79*	*58.22*	*47.16*	*48.96*	*70.11*	*54.54*	*20.30*	*55.48*

* *The experiments are conducted under a slightly different setting, as described in Section 4.*

Table 1: Summary of Classifier Performance. 4-way classifiers have been tested as well, but their performance is not as good as that of the binary classifiers shown here. One major difference is that it is harder to balance the number of instances across all the classes when training 4-way classifiers.

rules to guide lexicon lookup. The word *supplied*, for instance, becomes *suppli* after stemming, which still fails to match the lexicon entry *supply*, unless adjusted accordingly.

Binning. An additional finding regards features that capture numeric, rather than binary, information, such as *polarity*. Since this feature encodes the counts of each type of sentiment word (with respect to each argument and their cross product), and Naive Bayes can only interpret binary features, we first employed a binning mechanism with each bin covering a single value. For instance, if arg1 consists of three positive words, we included *arg1pos1*, *arg1pos2* and *arg1pos3* as features instead of just *arg1pos3*.

The rationale behind binning is that it captures the proximity of related instances. Imagine having three instances each with one, two, and three positive words in arg1, respectively. Without binning, the features added are simply *arg1pos1*, *arg1pos2*, *arg1pos3*, respectively. From the perspective of the classifier, the third instance is no more similar to the second instance than it is to the first instance, even though having three positive words is clearly closer to having two positive words than having one positive word. With binning, this proximity is captured by the fact that the first instance has just one feature in common with the third instance, whereas the second instance has two.

Binning, however, significantly degrades performance on most of the classification tasks. One possible explanation is that these features function as an abstraction of certain lexical patterns, rather than directly capturing similarities among instances of the same class.

6 Conclusion

We employ a simple greedy feature selection approach to identify subsets of known features for implicit discourse relation identification that yield the best performance to date w.r.t. F_1-score on the PDTB data set. We also identify aspects of feature set extraction and representation that are crucial for obtaining state-of-the-art performance. Possible future work includes evaluating the performance without using the gold standard parses. This will give a better idea of how the features that rely on parser output will perform on real-world data where no gold standard parsing information is available. In this way, we can ensure that findings in this area of research bring practical gains to the community.

Acknowledgments

We would like to thank Annie Louis and Yu Xu for helping us reimplement the systems from Louis et al. (2010) and Zhou et al. (2010), respectively. We also thank the anonymous reviewers for their helpful feedback. This work was supported in part by National Science Foundation Grants IIS-1111176 and IIS-0968450, and by a gift from Google.

References

Steven Bird. 2006. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, COLING-ACL '06, pages 69–72, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sasha Blair-Goldensohn, Kathleen McKeown, and Owen Rambow. 2007. Building and refining rhetorical-semantic relation models. In *HLT-NAACL*, pages 428–435.

G. John, R. Kohavi, and K. Pfleger. 1994. Irrelevant Features and the Subset Selection Problem. In W. Cohen and H. Hirsh, editors, *Proceedings of the Eleventh International Conference on Machine Learning*, pages 121–129. Morgan Kaufmann.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *IN PROCEEDINGS OF THE 41ST ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS*, pages 423–430.

Beth Levin and Harold Somers. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. University Of Chicago Press.

Ziheng Lin, Min-Yen Kan, and Hwee Tou Ng. 2009. Recognizing implicit discourse relations in the penn discourse treebank. In *EMNLP*, pages 343–351.

Annie Louis, Aravind K. Joshi, Rashmi Prasad, and Ani Nenkova. 2010. Using entity features to classify implicit discourse relations. In *SIGDIAL Conference*, pages 59–62.

Daniel Marcu and Abdessamad Echihabi. 2002. An unsupervised approach to recognizing discourse relations. In *ACL*, pages 368–375.

M. Marcus, M. Marcinkiewicz, and B. Santorini. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Emily Pitler, Mridhula Raghupathy, Hena Mehta, Ani Nenkovak, Alan Lee, and Aravind K. Joshi. 2008. Easily identifiable discourse relations. In *COLING (Posters)*, pages 87–90.

Emily Pitler, Annie Louis, and Ani Nenkova. 2009. Automatic sense prediction for implicit discourse relations in text. In *ACL/AFNLP*, pages 683–691.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The penn discourse treebank 2.0. In Bente Maegaard Joseph Mariani Jan Odjik Stelios Piperidis Daniel Tapias Nicoletta Calzolari (Conference Chair), Khalid Choukri, editor, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

P J Stone, D C Dunphy, M S Smith, and D M Ogilvie. 1966. *The General Inquirer: A Computer Approach to Content Analysis*, volume 08. MIT Press.

Ben Wellner, Lisa Ferro, Warren R. Greiff, and Lynette Hirschman. 2006. Reading comprehension tests for computer-based understanding evaluation. *Natural Language Engineering*, 12(4):305–334.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *HLT/EMNLP*.

Zhi-Min Zhou, Yu Xu, Zheng-Yu Niu, Man Lan, Jian Su, and Chew Lim Tan. 2010. Predicting discourse connectives for implicit discourse relation recognition. In *COLING (Posters)*, pages 1507–1514.