Automatic sense prediction for implicit discourse relations in text

Emily Pitler, Annie Louis, Ani Nenkova
Computer and Information Science
University of Pennsylvania
Philadelphia, PA 19104, USA
epitler,lannie,nenkova@seas.upenn.edu

Abstract

We present a series of experiments on automatically identifying the sense of *implicit* discourse relations, i.e. relations that are not marked with a discourse connective such as "but" or "because". We work with a corpus of implicit relations present in newspaper text and report results on a test set that is representative of the naturally occurring distribution of senses. We use several linguistically informed features, including polarity tags, Levin verb classes, length of verb phrases, modality, context, and lexical features. In addition, we revisit past approaches using lexical pairs from unannotated text as features, explain some of their shortcomings and propose modifications. Our best combination of features outperforms the baseline from data intensive approaches by 4% for comparison and 16% for contingency.

1 Introduction

Implicit discourse relations abound in text and readers easily recover the sense of such relations during semantic interpretation. But automatic sense prediction for implicit relations is an outstanding challenge in discourse processing.

Discourse relations, such as causal and contrast relations, are often marked by explicit discourse connectives (also called cue words) such as "because" or "but". It is not uncommon, though, for a discourse relation to hold between two text spans without an explicit discourse connective, as the example below demonstrates:

(1) *The 101-year-old magazine has never had to woo advertisers with quite so much fervor before.*
[because] **It largely rested on its hard-to-fault demographics.**

In this paper we address the problem of automatic sense prediction for discourse relations in newspaper text. For our experiments, we use the Penn Discourse Treebank, the largest existing corpus of discourse annotations for both implicit and explicit relations. Our work is also informed by the long tradition of data intensive methods that rely on huge amounts of unannotated text rather than on manually tagged corpora (Marcu and Echihabi, 2001; Blair-Goldensohn et al., 2007).

In our analysis, we focus only on *implicit* discourse relations and clearly separate these from explicits. Explicit relations are easy to identify. The most general senses (comparison, contingency, temporal and expansion) can be disambiguated in explicit relations with 93% accuracy based solely on the discourse connective used to signal the relation (Pitler et al., 2008). So reporting results on explicit and implicit relations separately will allow for clearer tracking of progress.

In this paper we investigate the effectiveness of various features designed to capture lexical and semantic regularities for identifying the sense of implicit relations. Given two text spans, previous work has used the cross-product of the words in the spans as features. We examine the most informative word pair features and find that they are not the semantically-related pairs that researchers had hoped. We then introduce several other methods capturing the semantics of the spans (polarity features, semantic classes, tense, etc.) and evaluate their effectiveness. This is the first study which reports results on classifying naturally occurring implicit relations in text and uses the natural distribution of the various senses.

2 Related Work

Experiments on implicit and explicit relations Previous work has dealt with the prediction of discourse relation sense, but often for explicits and at the sentence level.

Soricut and Marcu (2003) address the task of

Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, pages 683–691,
Suntec, Singapore, 2-7 August 2009.

parsing discourse structures *within the same sentence*. They use the RST corpus (Carlson et al., 2001), which contains 385 Wall Street Journal articles annotated following the Rhetorical Structure Theory (Mann and Thompson, 1988). Many of the useful features, syntax in particular, exploit the fact that both arguments of the connective are found in the same sentence. Such features would not be applicable to the analysis of implicit relations that occur *intersententially*.

Wellner et al. (2006) used the GraphBank (Wolf and Gibson, 2005), which contains 105 Associated Press and 30 Wall Street Journal articles annotated with discourse relations. They achieve 81% accuracy in sense disambiguation on this corpus. However, GraphBank annotations do not differentiate between implicits and explicits, so it is difficult to verify success for implicit relations.

Experiments on artificial implicits Marcu and Echihabi (2001) proposed a method for cheap acquisition of training data for discourse relation sense prediction. Their idea is to use unambiguous patterns such as [Arg1, *but* Arg2.] to create synthetic examples of implicit relations. They delete the connective and use [Arg1, Arg2] as an example of an implicit relation.

The approach is tested using binary classification between relations on balanced data, a setting very different from that of any realistic application. For example, a question-answering application that needs to identify causal relations (i.e. as in Girju (2003)), must not only differentiate causal relations from comparison relations, but also from expansions, temporal relations, and possibly no relation at all. In addition, using equal numbers of examples of each type can be misleading because the distribution of relations is known to be skewed, with expansions occurring most frequently. Causal and comparison relations, which are most useful for applications, are less frequent. Because of this, the recall of the classification should be the primary metric of success, while the Marcu and Echihabi (2001) experiments report only accuracy.

Later work (Blair-Goldensohn et al., 2007; Sporleder and Lascarides, 2008) has discovered that the models learned do not perform as well on implicit relations as one might expect from the test accuracies on synthetic data.

3 Penn Discourse Treebank

For our experiments, we use the Penn Discourse Treebank (PDTB; Prasad et al., 2008), the largest available annotated corpora of discourse relations. The PDTB contains discourse annotations over the same 2,312 Wall Street Journal (WSJ) articles as the Penn Treebank.

For each explicit discourse connective (such as "but" or "so"), annotators identified the two text spans between which the relation holds and the sense of the relation.

The PDTB also provides information about *local* implicit relations. For each pair of adjacent sentences within the same paragraph, annotators selected the explicit discourse connective which best expressed the relation between the sentences and then assigned a sense to the relation. In Example (1) above, the annotators identified "because" as the most appropriate connective between the sentences, and then labeled the implicit discourse relation *Contingency*.

In the PDTB, explicit and implicit relations are clearly distinguished, allowing us to concentrate solely on the implicit relations.

As mentioned above, each implicit and explicit relation is annotated with a sense. The senses are arranged in a hierarchy, allowing for annotations as specific as *Contingency.Cause.reason*. In our experiments, we use only the top level of the sense annotations: Comparison, Contingency, Expansion, and Temporal. Using just these four relations allows us to be theory-neutral; while different frameworks (Hobbs, 1979; McKeown, 1985; Mann and Thompson, 1988; Knott and Sanders, 1998; Asher and Lascarides, 2003) include different relations of varying specificities, all of them include these four core relations, sometimes under different names.

Each relation in the PDTB takes *two arguments*. Example (1) can be seen as the predicate *Contingency* which takes the two sentences as arguments. For implicits, the span in the first sentence is called *Arg1* and the span in the following sentence is called *Arg2*.

4 Word pair features in prior work

Cross product of words Discourse connectives are the most reliable predictors of the semantic sense of the relation (Marcu, 2000; Pitler et al., 2008). However, in the absence of explicit markers, the most easily accessible features are the

words in the two text spans of the relation. Intuitively, one would expect that there is some relationship that holds between the words in the two arguments. Consider for example the following sentences:

The recent explosion of country funds mirrors the "closed-end fund mania" of the 1920s, Mr. Foot says, when narrowly focused funds grew wildly **popular**. *They fell into* **oblivion** *after the 1929 crash.*

The words "popular" and "oblivion" are almost antonyms, and one might hypothesize that their occurrence in the two text spans is what triggers the contrast relation between the sentences. Similarly, a pair of words such as *(rain, rot)* might be indicative of a causal relation. If this hypothesis is correct, pairs of words (w_1, w_2) such that w_1 appears in the first sentence and w_2 appears in the second sentence would be good features for identifying contrast relations.

Indeed, word pairs form the basic feature of most previous work on classifying implicit relations (Marcu and Echihabi, 2001; Blair-Goldensohn et al., 2007; Sporleder and Lascarides, 2008) or the simpler task of predicting which connective should be used to express a relation (Lapata and Lascarides, 2004).

Semantic relations vs. function word pairs If the hypothesis for word pair triggers of discourse relations were true, the analysis of unambiguous relations can be used to discover pairs of words with causal or contrastive relations holding between them. Yet, feature analysis has not been performed in prior studies to establish or refute this possibility.

At the same time, feature selection is always necessary for word pairs, which are numerous and lead to data sparsity problems. Here, we present a meta analysis of the feature selection work in three prior studies.

One approach for reducing the number of features follows the hypothesis of semantic relations between words. Marcu and Echihabi (2001) considered only nouns, verbs and and other cue phrases in word pairs. They found that even with millions of training examples, prediction results using all words were superior to those based on only pairs of non-function words. However, since the learning curve is steeper when function words were removed, they hypothesize that using only non-function words will outperform using all words once enough training data is available.

In a similar vein, Lapata and Lascarides (2004) used pairings of only verbs, nouns and adjectives for predicting which temporal connective is most suitable to express the relation between two given text spans. Verb pairs turned out to be one of the best features, but no useful information was obtained using nouns and adjectives.

Blair-Goldensohn et al. (2007) proposed several refinements of the word pair model. They show that *(i)* stemming, *(ii)* using a small fixed vocabulary size consisting of only the most frequent stems (which would tend to be dominated by function words) and *(iii)* a cutoff on the minimum frequency of a feature, all result in improved performance. They also report that filtering stopwords has a negative impact on the results.

Given these findings, we expect that pairs of function words are informative features helpful in predicting discourse relation sense. In our work that we describe next, we use feature selection to investigate the word pairs in detail.

5 Analysis of word pair features

For the analysis of word pair features, we use a large collection of automatically extracted explicit examples from the experiments in Blair-Goldensohn et al. (2007). The data, from now on referred to as TextRels, has explicit contrast and causal relations which were extracted from the English Gigaword Corpus (Graff, 2003) which contains over four million newswire articles.

The explicit cue phrase is removed from each example and the spans are treated as belonging to an implicit relation. Besides cause and contrast, the TextRels data include a no-relation category which consists of sentences from the same text that are separated by at least three other sentences.

To identify features useful for classifying comparison vs other relations, we chose a random sample of 5000 examples for Contrast and 5000 Other relations (2500 each of Cause and No-relation). For the complete set of 10,000 examples, word pair features were computed. After removing word pairs that appear less than 5 times, the remaining features were ranked by information gain using the MALLET toolkit[1].

Table 1 lists the word pairs with highest information gain for the Contrast vs. Other and Cause vs. Other classification tasks. All contain very frequent stop words, and interestingly for the Con-

[1] mallet.cs.umass.edu

trast vs. Other task, most of the word pairs contain discourse connectives.

This is certainly unexpected, given that word pairs were formed by deleting the discourse connectives from the sentences expressing Contrast. Word pairs containing "but" as one of their elements in fact signal the presence of a relation that is *not* Contrast.

Consider the example shown below:

The government says it has reached most isolated townships by now, *but because* roads are blocked, getting anything *but* basic food supplies to people remains difficult.

Following Marcu and Echihabi (2001), the pair [The government says it has reached most isolated townships by now, *but*] and [roads are blocked, getting anything *but* basic food supplies to people remains difficult.] is created as an example of the Cause relation. Because of examples like this, "but-but" is a very useful word pair feature indicating Cause, as the *but* would have been removed for the artifical Contrast examples. In fact, the top 17 features for classifying Contrast versus Other all contain the word "but", and are indications that the relation is Other.

These findings indicate an unexpected anomalous effect in the use of synthetic data. Since relations are created by removing connectives, if an unambiguous connective remains, its presence is a reliable indicator that the example should be classified as Other. Such features might work well and lead to high accuracy results in identifying synthetic implicit relations, but are unlikely to be useful in a realistic setting of actual implicits.

Comparison vs. Other			**Contingency vs. Other**		
the-but	s-but	the-in	the-and	in-the	the-of
of-but	for-but	but-but	said-said	to-of	the-a
in-but	was-but	it-but	a-and	a-the	of-the
to-but	that-but	the-it*	to-and	to-to	the-in
and-but	but-the	to-it*	and-and	the-the	in-in
a-but	he-but	said-in	to-the	of-and	a-of
said-but	they-but	of-in	in-and	in-of	s-and

Table 1: Word pairs with highest information gain.

Also note that the only two features predictive of the comparison class (indicated by * in Table 1): the-it and to-it, contain only function words rather than semantically related nonfunction words. This ranking explains the observations reported in Blair-Goldensohn et al. (2007) where removing stopwords degraded classifier performance and why using only nouns, verbs or adjectives (Marcu and Echihabi, 2001; Lapata and Lascarides, 2004) is not the best option[2].

6 Features for sense prediction of implicit discourse relations

The contrast between the "popular"/"oblivion" example we started with above can be analyzed in terms of lexical relations (near antonyms), but also could be explained by different polarities of the two words: "popular" is generally a positive word, while "oblivion" has negative connotations.

While we agree that the actual words in the arguments are quite useful, we also define several higher-level features corresponding to various semantic properties of the words. The words in the two text spans of a relation are taken from the gold-standard annotations in the PDTB.

Polarity Tags: We define features that represent the sentiment of the words in the two spans. Each word's polarity was assigned according to its entry in the Multi-perspective Question Answering Opinion Corpus (Wilson et al., 2005). In this resource, each sentiment word is annotated as positive, negative, both, or neutral. We use the number of negated and non-negated positive, negative, and neutral sentiment words in the two text spans as features. If a writer refers to something as "nice" in Arg1, that counts towards the positive sentiment count (*Arg1Positive*); "not nice" would count towards *Arg1NegatePositive*. A sentiment word is negated if a word with a General Inquirer (Stone et al., 1966) Negate tag precedes it. We also have features for the cross products of these polarities between Arg1 and Arg2.

We expected that these features could help Comparison examples especially. Consider the following example:

Executives at Time Inc. Magazine Co., a subsidiary of Time Warner, have said the joint venture with Mr. Lang wasn't a good one. The venture, formed in 1986, was supposed to be Time's low-cost, safe entry into women's magazines.

The word *good* is annotated with positive polarity, however it is negated. *Safe* is tagged as having positive polarity, so this opposition could indicate the Comparison relation between the two sentences.

Inquirer Tags: To get at the meanings of the spans, we look up what semantic categories each

[2]In addition, an informal inspection of 100 word pairs with high information gain for Contrast vs. Other (the longest word pairs were chosen, as those are more likely to be content words) found only six semantically opposed pairs.

word falls into according to the General Inquirer lexicon (Stone et al., 1966). The General Inquirer has classes for positive and negative polarity, as well as more fine-grained categories such as words related to virtue or vice. The Inquirer even contains a category called "Comp" that includes words that tend to indicate Comparison, such as "optimal", "other", "supreme", or "ultimate".

Several of the categories are complementary: Understatement versus Overstatement, Rise versus Fall, or Pleasure versus Pain. Pairs where one argument contains words that indicate Rise and the other argument indicates Fall might be good evidence for a Comparison relation.

The benefit of using these tags instead of just the word pairs is that we see more observations for each semantic class than for any particular word, reducing the data sparsity problem. For example, the pair *rose:fell* often indicates a Comparison relation when speaking about stocks. However, occasionally authors refer to stock prices as "jumping" rather than "rising". Since both *jump* and *rise* are members of the Rise class, new *jump* examples can be classified using past *rise* examples.

Development testing showed that including features for all words' tags was not useful, so we include the Inquirer tags of only the verbs in the two arguments and their cross-product. Just as for the polarity features, we include features for both each tag and its negation.

Money/Percent/Num: If two adjacent sentences both contain numbers, dollar amounts, or percentages, it is likely that a comparison relation might hold between the sentences. We included a feature for the count of numbers, percentages, and dollar amounts in Arg1 and Arg2. We also included the number of times each combination of number/percent/dollar occurs in Arg1 and Arg2. For example, if Arg1 mentions a percentage and Arg2 has two dollar amounts, the feature *Arg1Percent-Arg2Money* would have a count of 2. This feature is probably genre-dependent. Numbers and percentages often appear in financial texts but would be less frequent in other genres.

WSJ-LM: This feature represents the extent to which the words in the text spans are typical of each relation. For each sense, we created unigram and bigram language models over the implicit examples in the training set. We compute each example's probability according to each of these language models. The features are the ranks of the spans' likelihoods according to the various language models. For example, if of the unigram models, the most likely relation to generate this example was Contingency, then the example would include the feature *ContingencyUnigram1*. If the third most likely relation according to the bigram models was Expansion, then it would include the feature *ExpansionBigram3*.

Expl-LM: This feature ranks the text spans according to language models derived from the explicit examples in the TextRels corpus. However, the corpus contains only Cause, Contrast and No-relation, hence we expect the WSJ language models to be more helpful.

Verbs: These features include the number of pairs of verbs in Arg1 and Arg2 from the same verb class. Two verbs are from the same verb class if each of their highest Levin verb class (Levin, 1993) levels (in the LCS Database (Dorr, 2001)) are the same. The intuition behind this feature is that the more related the verbs, the more likely the relation is an Expansion.

The verb features also include the average length of verb phrases in each argument, as well as the cross product of this feature for the two arguments. We hypothesized that verb chunks that contain more words, such as "They [are allowed to proceed]" often contain rationales afterwards (signifying Contingency relations), while short verb phrases like "They proceed" might occur more often in Expansion or Temporal relations.

Our final verb features were the part of speech tags (gold-standard from the Penn Treebank) of the main verb. One would expect that Expansion would link sentences with the same tense, whereas Contingency and Temporal relations would contain verbs with different tenses.

First-Last, First3: The first and last words of a relation's arguments have been found to be particularly useful for predicting its sense (Wellner et al., 2006). Wellner et al. (2006) suggest that these words are such predictive features because they are often explicit discourse connectives. In our experiments on implicits, the first and last words are not connectives. However, some implicits have been found to be related by connective-like expressions which often appear in the beginning of the second argument. In the PDTB, these are annotated as alternatively lexicalized relations (AltLexes). To capture such effects, we included the first and last words of Arg1 as features, the first

and last words of Arg2, the pair of the first words of Arg1 and Arg2, and the pair of the last words. We also add two additional features which indicate the first three words of each argument.

Modality: Modal words, such as "can", "should", and "may", are often used to express conditional statements (i.e. "If I were a wealthy man, I wouldn't have to work hard.") thus signaling a Contingency relation. We include a feature for the presence or absence of modals in Arg1 and Arg2, features for specific modal words, and their cross-products.

Context: Some implicit relations appear immediately before or immediately after certain explicit relations far more often than one would expect due to chance (Pitler et al., 2008). We define a feature indicating if the immediately preceding (or following) relation was an explicit. If it was, we include the connective trigger of the relation and its sense as features. We use oracle annotations of the connective sense, however, most of the connectives are unambiguous.

One might expect a different distribution of relation types in the beginning versus further in the middle of a paragraph. We capture paragraph-position information using a feature which indicates if Arg1 begins a paragraph.

Word pairs Four variants of word pair models were used in our experiments. All the models were eventually tested on implicit examples from the PDTB, but the training set-up was varied.

Wordpairs-TextRels In this setting, we trained a model on word pairs derived from unannotated text (TextRels corpus).

Wordpairs-PDTBImpl Word pairs for training were formed from the cross product of words in the textual spans (*Arg1 x Arg2*) of the PDTB *implicit* relations.

Wordpairs-selected Here, only word pairs from *Wordpairs-PDTBImpl* with non-zero information gain on the TextRels corpus were retained.

Wordpairs-PDTBExpl In this case, the model was formed by using the word pairs from the *explicit* relations in the sections of the PDTB used for training.

7 Classification Results

For all experiments, we used sections 2-20 of the PDTB for training and sections 21-22 for testing. Sections 0-1 were used as a development set for feature design.

We ran four binary classification tasks to identify each of the main relations from the rest. As each of the relations besides Expansion are infrequent, we train using equal numbers of positive and negative examples of the target relation. The negative examples were chosen at random. We used all of sections 21 and 22 for testing, so the test set is representative of the natural distribution.

The training sets contained: *Comparison* (1927 positive, 1927 negative), *Contingency* (3500 each), *Expansion*[3] (6356 each), and *Temporal* (730 each).

The test set contained: 151 examples of *Comparison*, 291 examples of *Contingency*, 986 examples of *Expansion*, 82 examples of *Temporal*, and 13 examples of *No-relation*.

We used Naive Bayes, Maximum Entropy (MaxEnt), and AdaBoost (Freund and Schapire, 1996) classifiers implemented in MALLET.

7.1 Non-Wordpair Features

The performance using only our semantically informed features is shown in Table 7. Only the Naive Bayes classification results are given, as space is limited and MaxEnt and AdaBoost gave slightly lower accuracies overall.

The table lists the f-score for each of the target relations, with overall accuracy shown in brackets. Given that the experiments are run on natural distribution of the data, which are skewed towards Expansion relations, the f-score is the more important measure to track.

Our random baseline is the f-score one would achieve by randomly assigning classes in proportion to its true distribution in the test set. The best results for all four tasks are considerably higher than random prediction, but still low overall. Our features provide 6% to 18% absolute improvements in f-score over the baseline for each of the four tasks. The largest gain was in the Contingency versus Other prediction task. The least improvement was for distinguishing Expansion versus Other. However, since Expansion forms the largest class of relations, its f-score is still the highest overall. We discuss the results per relation class next.

Comparison We expected that polarity features would be especially helpful for identifying Com-

[3]The PDTB also contains annotations of entity relations, which most frameworks consider a subset of Expansion. Thus, we include relations annotated as EntRel as positive examples of Expansion.

Features	Comp. vs. Not	Cont. vs. Other	Exp. vs. Other	Temp. vs. Other	Four-way
Money/Percent/Num	19.04 (43.60)	18.78 (56.27)	22.01 (41.37)	10.40 (23.05)	(63.38)
Polarity Tags	16.63 (55.22)	19.82 (76.63)	71.29 (59.23)	11.12 (18.12)	(65.19)
WSJ-LM	18.04 (9.91)	0.00 (80.89)	0.00 (35.26)	10.22 (5.38)	(65.26)
Expl-LM	18.04 (9.91)	0.00 (80.89)	0.00 (35.26)	10.22 (5.38)	(65.26)
Verbs	18.55 (26.19)	36.59 (62.44)	59.36 (52.53)	12.61 (41.63)	(65.33)
First-Last, First3	21.01 (52.59)	36.75 (59.09)	63.22 (56.99)	15.93 (61.20)	(65.40)
Inquirer tags	17.37 (43.8)	15.76 (77.54)	70.21 (58.04)	11.56 (37.69)	(62.21)
Modality	17.70 (17.6)	21.83 (76.95)	15.38 (37.89)	11.17 (27.91)	(65.33)
Context	19.32 (56.66)	29.55 (67.42)	67.77 (57.85)	12.34 (55.22)	(64.01)
Random	9.91	19.11	64.74	5.38	

Table 2: f-score (accuracy) using different features; Naive Bayes.

parison relations. Surprisingly, polarity was actually one of the worst classes of features for Comparison, achieving an f-score of 16.33 (in contrast to using the first, last and first three words of the sentences as features, which leads to an f-score of 21.01). We examined the prevalence of positive-negative or negative-positive polarity pairs in our training set. 30% of the Comparison examples contain one of these opposite polarity pairs, while 31% of the Other examples contain an opposite polarity pair. To our knowledge, this is the first study to examine the prevalence of polarity words in the arguments of discourse relations in their natural distributions. Contrary to popular belief, Comparisons do not tend to have more opposite polarity pairs.

The two most useful classes of features for recognizing Comparison relations were the first, last and first three words in the sentence and the context features that indicate the presence of a paragraph boundary or of an explicit relation just before or just after the location of the hypothesized implicit relation (19.32 f-score).

Contingency The two best features for the Contingency vs. Other distinction were verb information (36.59 f-score) and first, last and first three words in the sentence (36.75 f-score). Context again was one of the features that led to improvement. This makes sense, as Pitler et al. (2008) found that implicit contingencies are often found immediately following explicit comparisons.

We were surprised that the polarity features were helpful for Contingency but not Comparison. Again we looked at the prevalence of opposite polarity pairs. While for Comparison versus Other there was not a significant difference, for Contingency there are quite a few more opposite polarity pairs (52%) than for not Contingency (41%).

The language model features were completely useless for distinguishing contingencies from other relations.

Expansion As Expansion is the majority class in the natural distribution, recall is less of a problem than precision. The features that help achieve the best f-score are all features that were found to be useful in identifying other relations.

Polarity tags, Inquirer tags and context were the best features for identifying expansions with f-scores around 70%.

Temporal Implicit temporal relations are relatively rare, making up only about 5% of our test set. Most temporal relations are explicitly marked with a connective like "when" or "after".

Yet again, the first and last words of the sentence turned out to be useful indicators for temporal relations (15.93 f-score). The importance of the first and last words for this distinction is clear. It derives from the fact that temporal implicits often contain words like "yesterday" or "Monday" at the end of the sentence. Context is the next most helpful feature for temporal relations.

7.2 Which word pairs help?

For Comparison and Contingency, we analyze the behavior of word pair features under several different settings. Specifically we want to address two important related questions raised in recent work by others: *(i)* is unannotated data from explicits useful for training models that disambiguate implicit discourse relations and *(ii)* are explicit and implicit relations intrinsically different from each other.

Wordpairs-TextRels is the worst approach. The best use of word pair features is Wordpairs-selected. This model gives 4% better absolute f-score for Comparison and 14% for Contingency over Wordpairs-TextRels. In this setting the TextRels data was used to choose the word pair features, but the probabilities for each feature were estimated using the training portion of the PDTB

Comp. vs. Other	
Wordpairs-TextRels	17.13 (46.62)
Wordpairs-PDTBExpl	19.39 (51.41)
Wordpairs-PDTBImpl	20.96 (42.55)
First-last, first3 *(best-non-wp)*	21.01 (52.59)
Best-non-wp + Wordpairs-selected	21.88 (56.40)
Wordpairs-selected	21.96 (56.59)
Cont. vs. Other	
Wordpairs-TextRels	31.10 (41.83)
Wordpairs-PDTBExpl	37.77 (56.73)
Wordpairs-PDTBImpl	43.79 (61.92)
Polarity, verbs, first-last, first3, modality, context *(best-non-wp)*	42.14 (66.64)
Wordpairs-selected	45.60 (67.10)
Best-non-wp + Wordpairs-selected	47.13 (67.30)
Expn. vs. Other	
Best-non-wp + wordpairs	62.39 (59.55)
Wordpairs-PDTBImpl	63.84 (60.28)
Polarity, inquirer tags, context *(best-non-wp)*	76.42 (63.62)
Temp. vs. Other	
First-last, first3 *(best-non-wp)*	15.93 (61.20)
Wordpairs-PDTBImpl	16.21 (61.98)
Best-non-wp + Wordpairs-PDTBImpl	16.76 (63.49)

Table 3: f-score (accuracy) of various feature sets; Naive Bayes.

implicit examples.

We also confirm that even within the PDTB, information from annotated explicit relations (Wordpairs-PDTBExpl) is not as helpful as information from annotated implicit relations (Wordpairs-PDTBImpl). The absolute difference in f-score between the two models is close to 2% for Comparison, and 6% for Contingency.

7.3 Best results

Adding other features to word pairs leads to improved performance for Contingency, Expansion and Temporal relations, but not for Comparison.

For contingency detection, the best combination of our features included polarity, verb information, first and last words, modality, context with Wordpairs-selected. This combination led to a definite improvement, reaching an f-score of 47.13 (16% absolute improvement in f-score over Wordpairs-TextRels).

For detecting expansions, the best combination of our features (polarity+Inquirer tags+context) outperformed Wordpairs-PDTBImpl by a wide margin, close to 13% absolute improvement (f-scores of 76.42 and 63.84 respectively).

7.4 Sequence Model of Discourse Relations

Our results from the previous section show that classification of implicits benefits from information about nearby relations, and so we expected improvements using a sequence model, rather than classifying each relation independently.

We trained a CRF classifier (Lafferty et al., 2001) over the sequence of implicit examples from all documents in sections 02 to 20. The test set is the same as used for the 2-way classifiers. We compare against a 6-way[4] Naive Bayes classifier. Only word pairs were used as features for both. Overall 6-way prediction accuracy is 43.27% for the Naive Bayes model and 44.58% for the CRF model.

8 Conclusion

We have presented the first study that predicts *implicit* discourse relations in a realistic setting (distinguishing a relation of interest from all others, where the relations occur in their natural distributions). Also unlike prior work, we separate the task from the easier task of explicit discourse prediction. Our experiments demonstrate that features developed to capture word polarity, verb classes and orientation, as well as some lexical features are strong indicators of the type of discourse relation.

We analyze word pair features used in prior work that were intended to capture such semantic oppositions. We show that the features in fact do not capture semantic relation but rather give information about function word co-occurrences. However, they are still a useful source of information for discourse relation prediction. The most beneficial application of such features is when they are selected from a large unannotated corpus of explicit relations, but then trained on manually annotated implicit relations.

Context, in terms of paragraph boundaries and nearby explicit relations, also proves to be useful for the prediction of implicit discourse relations. It is helpful when added as a feature in a standard, instance-by-instance learning model. A sequence model also leads to over 1% absolute improvement for the task.

9 Acknowledgments

This work was partially supported by NSF grants IIS-0803159, IIS-0705671 and IGERT 0504487. We would like to thank Sasha Blair-Goldensohn for providing us with the TextRels data and for the insightful discussion in the early stages of our work.

[4]the four main relations, EntRel, NoRel

References

N. Asher and A. Lascarides. 2003. *Logics of conversation*. Cambridge University Press.

S. Blair-Goldensohn, K.R. McKeown, and O.C. Rambow. 2007. Building and Refining Rhetorical-Semantic Relation Models. In *Proceedings of NAACL HLT*, pages 428–435.

L. Carlson, D. Marcu, and M.E. Okurowski. 2001. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*, pages 1–10.

B.J. Dorr. 2001. LCS Verb Database. *Technical Report Online Software Database, University of Maryland, College Park, MD*.

Y. Freund and R.E. Schapire. 1996. Experiments with a New Boosting Algorithm. In *Machine Learning: Proceedings of the Thirteenth International Conference*, pages 148–156.

R. Girju. 2003. Automatic detection of causal relations for Question Answering. In *Proceedings of the ACL 2003 workshop on Multilingual summarization and question answering-Volume 12*, pages 76–83.

D. Graff. 2003. English gigaword corpus. *Corpus number LDC2003T05, Linguistic Data Consortium, Philadelphia*.

J. Hobbs. 1979. Coherence and coreference. *Cognitive Science*, 3:67–90.

A. Knott and T. Sanders. 1998. The classification of coherence relations and their linguistic markers: An exploration of two languages. *Journal of Pragmatics*, 30(2):135–175.

J. Lafferty, A. McCallum, and F. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *International Conference on Machine Learning 2001*, pages 282–289.

M. Lapata and A. Lascarides. 2004. Inferring sentence-internal temporal relations. In *HLT-NAACL 2004: Main Proceedings*.

B. Levin. 1993. English Verb Classes and Alternations: A Preliminary Investigation. *Chicago, IL*.

W.C. Mann and S.A. Thompson. 1988. Rhetorical structure theory: Towards a functional theory of text organization. *Text*, 8.

D. Marcu and A. Echihabi. 2001. An unsupervised approach to recognizing discourse relations. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 368–375.

D. Marcu. 2000. *The Theory and Practice of Discourse and Summarization*. The MIT Press.

K. McKeown. 1985. *Text Generation: Using Discourse strategies and Focus Constraints to Generate Natural Language Text*. Cambridge University Press, Cambridge, England.

E. Pitler, M. Raghupathy, H. Mehta, A. Nenkova, A. Lee, and A. Joshi. 2008. Easily identifiable discourse relations. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING08), short paper*.

R. Soricut and D. Marcu. 2003. Sentence level discourse parsing using syntactic and lexical information. In *HLT-NAACL*.

C. Sporleder and A. Lascarides. 2008. Using automatically labelled examples to classify rhetorical relations: An assessment. *Natural Language Engineering*, 14:369–416.

P.J. Stone, J. Kirsh, and Cambridge Computer Associates. 1966. *The General Inquirer: A Computer Approach to Content Analysis*. MIT Press.

B. Wellner, J. Pustejovsky, C. Havasi, A. Rumshisky, and R. Sauri. 2006. Classification of discourse coherence relations: An exploratory study using multiple knowledge sources. In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*.

T. Wilson, J. Wiebe, and P. Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 347–354.

F. Wolf and E. Gibson. 2005. Representing discourse coherence: A corpus-based study. *Computational Linguistics*, 31(2):249–288.