Improving the Inference of Implicit Discourse Relations via Classifying Explicit Discourse Connectives

Attapol T. Rutherford
Department of Computer Science
Brandeis University
Waltham, MA 02453, USA
tet@brandeis.edu

Nianwen Xue
Department of Computer Science
Brandeis University
Waltham, MA 02453, USA
xuen@brandeis.edu

Abstract

Discourse relation classification is an important component for automatic discourse parsing and natural language understanding. The performance bottleneck of a discourse parser comes from implicit discourse relations, whose discourse connectives are not overtly present. Explicit discourse connectives can potentially be exploited to collect more training data to collect more data and boost the performance. However, using them indiscriminately has been shown to hurt the performance because not all discourse connectives can be dropped arbitrarily. Based on this insight, we investigate the interaction between discourse connectives and the discourse relations and propose the criteria for selecting the discourse connectives that can be dropped independently of the context without changing the interpretation of the discourse. Extra training data collected only by the freely omissible connectives improve the performance of the system without additional features.

1 Introduction

The analysis of discourse-level structure has received increasing attention from the field in recent years (Feng and Hirst, 2012; Patterson and Kehler, 2013; Li et al., 2014). Discourse-level analysis is typically concerned with relations between clauses and sentences, linguistic units that go beyond sentence boundaries. There are a few conceptions of the discourse structure representation of a text such as a tree (Mann and Thompson, 1988), or a graph (Wolf et al., 2005). In the work we describe here, we adopt the view of the Penn Discourse Treebank (PDTB) (Prasad et al., 2008), which views a text as a series of local discourse relations, each of which consists of a discourse connective as a predicate taking two arguments. Syntactically, these two arguments are typically realized as clauses or sentences. The discourse connective (underlined) can either be *explicit*, as in (1), or *implicit*, as in (2):

(1) [The city's Campaign Finance Board has refused to pay Mr Dinkins $95,142 in matching funds]$_{Arg1}$ <u>because</u> [his campaign records are incomplete]$_{Arg2}$.

(2) [So much of the stuff poured into its Austin, Texas, offices that its mail rooms there simply stopped delivering it]$_{Arg1}$. <u>Implicit=so</u> [Now, thousands of mailers, catalogs and sales pitches go straight into the trash]$_{Arg2}$.

Determining the sense of an explicit discourse relation such as (1) is straightforward since "because" is a strong indicator that the relation between the two arguments is CONTINGENCY.CAUSE. This task effectively amounts to disambiguating the sense of discourse connective, which can be done with high accuracy (Pitler et al., 2008).

However, in the absence of an explicit discourse connective, inferring the sense of a discourse relation has proved to a very challenging task (Park and Cardie, 2012; Rutherford and Xue, 2014). The sense is no longer localized on one or two discourse connectives and must now be inferred solely based on its two textual arguments. Given the limited amount of annotated data in comparison to the number of features needed, the process of building a classifier is plagued by the data sparsity problem (Li and Nenkova, 2014). As a result, the classification accuracy of implicit discourse relations remains much

Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL, pages 799–808,
Denver, Colorado, May 31 – June 5, 2015.

lower than that of explicit discourse relations (Pitler et al., 2008).

One potential method for reducing the data sparsity problem is through a distantly supervised learning paradigm, which is the direction we take in this work. Distant supervision approaches make use of prior knowledge or heuristics to cheaply obtain *weakly labeled data*, which potentially contain a small number of false labels. Weakly labeled data can be collected from unannotated data and incorporated in the model training process to supplement manually labeled data. For our task, we can collect instances of explicit discourse relations from unannotated data by some simple heuristics. After dropping the discourse connectives, we should be able to treat them as additional implicit discourse relations.

The approach assumes that when the discourse connective is omitted, the discourse relation remains the same, which is a popular assumption in discourse analysis (Fraser, 2006; Schourup, 1999). This assumption turns out to be too strong in many cases as illustrated in (3):

(3) [I want to go home for the holiday]$_{Arg1}$. Nonetheless, [I will book a flight to Hawaii]$_{Arg2}$.

If "Nonetheless" is dropped in (3), one can no longer infer the COMPARISON relation. Instead, one would naturally infer a CONTINGENCY relation. Dropping the connective and adding the relation as a training sample adds noise to the training set and can only hurt the performance. In addition, certain types of explicit discourse relations have no corresponding implicit discourse relations. For example, discourse relations of the type CONTINGENY.CONDITION are almost always expressed with an explicit discourse connective and do not exist in implicit relations. We believe this also explains the lack of success in previous attempts to boost the performance of implicit discourse relation detection with this approach. (Biran and McKeown, 2013; Pitler et al., 2009). This suggests that in order for this approach to work, we need to identify instances of explicit discourse relations that closely match the characteristics of implicit discourse relations.

In this paper, we propose two criteria for selecting such explicit discourse relation instances: *omission rate* and *context differential*. Our selection criteria first classify discourse connectives by their distributional properties and suggest that not all discourse connectives are truly optional and not all implicit and explicit discourse relations are equivalent, contrary to commonly held beliefs in previous studies of discourse connectives. We show that only the freely omissible discourse connectives gather additional training instances that lead to significant performance gain against a strong baseline. Our approach improves the performance of implicit discourse relations without additional feature engineering in many settings and opens doors to more sophisticated models that require more training data.

The rest of the paper is structured as follows. In Section 2, we describe the discourse connective selection criteria. In Section 3, we present our discourse connective classification method and experimental results that demonstrate its impact on inferring implicit discourse relations. We discuss related work and conclude our findings in Section 4 and 5 respectively.

2 Discourse Connective Classification and Discourse Relation Extraction

2.1 Datasets used for selection

We use two datasets for the purposes of extracting and selecting weakly labeled explicit discourse relation instances: the Penn Discourse Treebank 2.0 (Prasad et al., 2008) and the English Gigaword corpus version 3 (Graff et al., 2007).

The Penn Discourse Treebank (PDTB) is the largest manually annotated corpus of discourse relations on top of one million word tokens from the Wall Street Journal (Prasad et al., 2008; Prasad et al., 2007). Each discourse relation in the PDTB is annotated with a semantic sense in the PDTB sense hierarchy, which has three levels: CLASS, TYPE and SUBTYPE. In this work, we are primarily concerned with the four top-level CLASS senses: EXPANSION, COMPARISON, CONTINGENCY, and TEMPORAL. The distribution of top-level senses of implicit discourse relations is shown in Table 2. The spans of text that participate in the discourse relation are also explicitly annotated. These are called ARG1 or ARG2, depending on its relationship with the discourse connective.

The PDTB is our corpus of choice for its lexical

groundedness. The existence of a discourse relation must be linked or grounded to a discourse connective. More importantly, this applies to not only explicit discourse connectives that occur naturally as part of the text but also to implicit discourse relations where a discourse connective is added by annotators during the annotation process. This is crucial to the work reported here in that it allows us to compare the distribution of the same connective in explicit and implicit discourse relations. In the next subsection, we will explain in detail how we compute the comparison measures and apply them to the selection of explicit discourse connectives that can be used for collecting good weakly labeled data.

We use the Gigaword corpus, a large unannotated newswire corpus, to extract and select instances of explicit discourse discourse relations to supplement the manually annotated instances from the PDTB. The Gigaword corpus is used for its large size of 2.9 billion words and its similarity to the Wall Street Journal data from the PDTB. The source of the corpus is drawn from six distinct international sources of English newswire dating from 1994 - 2006. We use this corpus to extract weakly labeled data for the experiment.

2.2 Discourse relation extraction pattern

We extract instances of explicit discourse relations from the Gigaword Corpus that have the same patterns as the implicit discourse relations in the PDTB, using simple regular expressions. We first sentence-segment the Gigaword Corpus using the NLTK sentence segmenter (Bird, 2006). We then write a set of rules to prevent some common erroneous cases such as *because* vs *because of* from being included.

If a discourse connective is a subordinating conjunction, then we use the following pattern:

```
(Clause 1) (connective) (clause 2).
```

Clause 1 and capitalized clause 2 are then used as *Arg1* and *Arg2* respectively.

If a discourse connective is a coordinating conjunction or discourse adverbial, we use the following pattern:

```
(Sentence 1). (Connective),(clause 2).
```

Sentence 1 and Clause 2 with the first word capitalized are used as *Arg1* and *Arg2* respectively.

Although there are obviously many other syntactic patterns associated with explicit discourse connectives, we use these two patterns because these are the only patterns that are also observed in the implicit discourse relations. We want to select instances of explicit discourse relations that match the argument patterns of implicit discourse relations as much as possible. As restrictive as this may seem, these two patterns along with the set of rules allow us to extract more than 200,000 relation instances from the Gigaword corpus, so the coverage is not an issue.

2.3 Discourse connective selection and classification criteria

We hypothesize that connectives that are omitted often and in a way that is insensitive to the semantic context are our ideal candidates for extracting good weakly labeled data. We call this type of connectives *freely omissible discourse connectives*. To search for this class of connectives, we need to characterize connectives by the rate at which they are omitted and by the similarity between their context, in this case their arguments, in explicit and implicit discourse relations. This is possible because implicit discourse connectives are inserted during annotation in the PDTB. For each discourse connective, we can compute *omission rate* and *context differential* from annotated explicit and implicit discourse relation instances in the PDTB and use those measures to classify and select discourse connectives.

2.3.1 Omission rate (OR)

We use *omission rates* (OR) to measure the level of optionality of a discourse connective. The omission rate of a type of discourse connective (DC) is defined as:

$$\frac{\text{\# occurrences of DC in implicit relations}}{\text{\# total occurrences of DC}}$$

Our intuition is that the discourse connectives that have a high level of omission rate are more suitable as supplemental training data to infer the sense of implicit discourse relations.

2.3.2 Context differential

The omission of a freely omissible discourse connective should also be context-independent. If the omission of a discourse connective leads to a different interpretation of the discourse relation, this means that the explicit and implicit discourse relations bound by this discourse connective are not equivalent, and the explicit discourse relation instance cannot be used to help infer the sense of the implicit discourse relation. Conversely, if the contexts for the discourse connective in explicit and implicit discourse relations do not significantly differ, then the explicit discourse relation instance can be used as weakly labeled data.

To capture this intuition, we must quantify the context differential of explicit and implicit discourse relations for each discourse connective. We represent the

semantic context of a discourse connective through a unigram distribution over words in its two arguments, with Arg1 and Arg2 combined. We use Jensen-Shannon Divergence (JSD) as a metric for measuring the difference between the contexts of a discourse connective in implicit and explicit discourse relations. Computing a context differential of the discourse connective therefore involves fitting a unigram distribution from all implicit discourse relations bound by that discourse connective and fitting another from all explicit discourse relations bound by the same discourse connective. We choose this method because it has been shown to be exceptionally effective in capturing similarities of discourse connectives (Hutchinson, 2005) and statistical language analysis in general (Lee, 2001; Ljubesic et al., 2008).

The Jensen-Shannon Divergence (JSD) metric for difference between P_o, the semantic environments (unigram distribution of words in Arg1 and Arg2 combined) in implicit discourse relations, and P_r, the semantic environments in explicit discourse relations, is defined as:

$$JSD(P_o||P_r) = \frac{1}{2}D(P_o||M) + \frac{1}{2}D(P_r||M)$$

where $M = \frac{1}{2}(P_o + P_r)$ is a mixture of the two distributions and $D(.||.)$ is Kullback-Leibler divergence function for discrete probability distributions:

$$D(P||Q) = \sum_i \ln(\frac{P(i)}{Q(i)})P(i)$$

2.4 Discourse Connective Classification

Using the two metrics, we can classify discourse connectives into the following classes:

1. Freely omissible: High OR and low JSD
2. Omissible: Low non-zero OR and low JSD.
3. Alternating I: High OR and high JSD.
4. Alternating II: Low non-zero OR and high JSD.
5. Non-omissible: Zero OR. JSD cannot be computed because the connectives are never found in any implicit discourse relations.

Classifying the connectives into these classes allow us to empirically investigate which explicit discourse relations are useful as supplemental training data for determining the sense of implicit discourse relations. We discuss each type of connectives below.

2.4.1 Freely omissible discourse connectives

These are connectives whose usage in implicit and explicit discourse relations is indistinguishable and therefore suitable as a source of supplemental training data. These connectives are defined as having high omission rate and low context differential. This definition implies that the omission is frequent and insensitive to the context. "Because" and "in particular" in (4) and (5) are such connectives. Dropping them has minimal impact on the understanding the discourse relation between their two arguments and one might argue they even make the sentences sound more natural.

(4) We cleared up questions and inconsistencies very quickly because the people who had the skills and perspective required to resolve them were part of the task team. (WSJ0562)

(5) Both companies are conservative marketers that rely on extensive market research. P&G, in particular, rarely rolls out a product nationally before extensive test-marketing. (WSJ0589)

2.4.2 Omissible discourse connectives

They are connectives whose usage in implicit and explicit discourse relations is indistinguishable, yet they are not often omitted because the discourse relation might be hard to interpret without them. These connectives are defined as having low omission rate and low context differential. For example,

(6) Such problems will require considerable skill to resolve. However, neither Mr. Baum nor Mr. Harper has much international experience. (WSJ0109)

One can infer from the discourse that the problems require international experience, but Mr. Baum and Mr. Harper don't have that experience even without the discourse connective "however". In other words, the truth value of this proposition is not affected by the presence or absence of this discourse connective. The sentence might sound a bit less natural, and the discourse relation seems a bit more difficult to infer if "however" is omitted.

2.4.3 Alternating discourse connectives

They are connectives whose usage in implicit and explicit discourse relations is substantially different and they are defined as having high context differential. Having high context differential means that the two arguments of an explicit discourse connective differ substantially from those of an implicit discourse. An example of such discourse connectives is "nevertheless" in (7). If the discourse connective is dropped, one might infer EXPANSION or CONTINGENCY relation instead of COMPARISON indicated by the connective.

(7) Plant Genetic's success in creating genetically engineered male steriles doesn't automatically mean it would be simple to create hybrids in all crops. Nevertheless, he said, he is negotiating with Plant Genetic to acquire the technology to try breeding hybrid cotton. (WSJ0209)

We hypothesize that this type of explicit discourse relations would not be useful as extra training instances for inferring implicit discourse relations because they will only add noise to the training set.

2.4.4 Non-omissible discourse connectives

They are defined as discourse connectives whose omission rate is close to zero as they are never found in implicit discourse relations. For example, conditionals can not be easily expressed without the use of an explicit discourse connective like "if". We hypothesize that instances of explicit discourse relations with such discourse connectives would not be useful as additional training data for inferring implicit discourse relations because they represent discourse relation senses that do not exist in the implicit discourse relations.

3 Experiments

3.1 Partitioning the discourse connectives

We only include the discourse connectives that appear in both explicit and implicit discourse connectives in the PDTB to make the comparison and classification possible. As a result, we only analyze 69 out of 134 connectives for the purpose of classification. We also leave out 15 connectives whose most frequent sense acccounts for less than 90% of their instances. For example, *since* can indicate a TEMPORAL sense or a CONTINGENCY sense of almost equal chance, so it is not readily useful for gathering weakly labeled data. Ultimately, we have 54 connectives as our candidates for freely omissible discourse connectives.

We first classify the discourse connectives based on their omission rates and context differentials as discussed in the previous section and partition all of the explicit discourse connective instances based on this classification. The distributions of omission rates and context differentials show substantial amount of variation among different connectives. Many connectives are rarely omitted and naturally form its own class of non-omissible discourse connectives (Figure 1). We run the agglomerative hierchical clustering algorithm using Euclidean distance on the rest of the connectives to divide them into two groups: high omission and low omission rates. The boundary between the two groups is around 0.65.

The distribution of discourse connectives with respect to the context differential suggests two distinct groups across the two corpora (Figure 2). The analysis only includes connectives that are omitted at least twenty times in the PDTB corpus, so that JSD can be computed. The hierarchical clustering algorithm divides the connectives into two groups with the boundary at around 0.32, as should be apparent from the histogram. The JSD's computed from the explicit discourse relations from the two

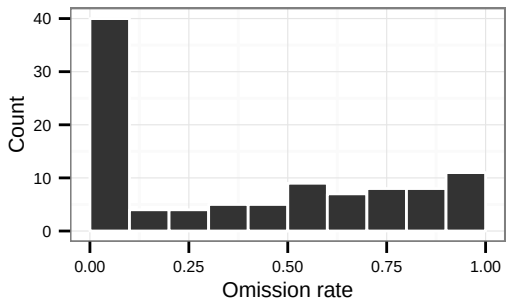

Figure 1: Omission rates of the discourse connective types vary drastically, suggesting that connectives vary in their optionality. Some connectives are never omitted.

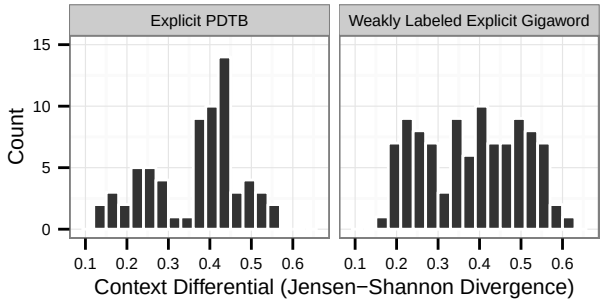

Figure 2: The distributions of Jensen-Shannon Divergence from both corpora shows two potential distinct clusters of discourse connectives.

corpora are highly correlated ($\rho = 0.80$, $p < 0.05$), so we can safely use the Gigaword corpus for the analysis and evaluation.

The omission rate boundary and context differential boundary together classify the discourse connectives into four classes in addition to the non-omissible connectives. When plotted against each other, omission rates and context differential together group the discourse connectives nicely into clusters (Figure 3). For the purpose of evaluation, we combine Alternating I and II into one class because each individual class is too sparse on its own. The complete discourse connective classification result is displayed in Table 1.

Sense	Train	Dev	Test
Comparison	1855	189	145
Contingency	3235	281	273
Expansion	6673	638	538
Temporal	582	48	55
Total	12345	1156	1011

Table 2: The distribution of senses of implicit discourse relations in the PDTB

Class Name	OR	JSD	Connectives
Alternating I	High	High	further, in sum, in the end, overall, similarly, whereas
Alternating II	Low	High	earlier, in turn, nevertheless, on the other hand, ultimately
Freely Omissible	High	Low	accordingly, as a result, because, by comparison, by contrast, consequently, for example, for instance, furthermore, in fact, in other words, in particular, in short, indeed, previously, rather, so, specifically, therefore,
Omissible	Low	Low	also, although, and, as, but, however, in addition, instead, meanwhile, moreover, rather, since, then, thus, while
Non-omissible	zero	NA	as long as, if, nor, now that, once, otherwise, unless, until

Table 1: Classification of discourse connectives based on omission rate (OR) and Jensen-Shannon Divergence context differential (JSD).

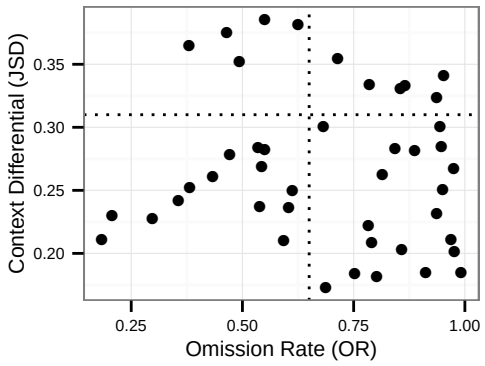

Figure 3: The scattergram of the discourse connectives suggest three distinct classes. Each dot represents a discourse connective.

3.2 Evaluation results

We formulate the implicit relation classification task as a 4-way classification task in a departure from previous practice where the task is usually set up as four *one vs other* binary classification tasks so that the effect of adding the distant supervision from the weakly labeled data can be more easily studied. We also believe this setup is more natural in realistic settings. Each classification instance consists of the two arguments of an implicit discourse relation, typically adjacent pairs of sentences in a text. The distribution of the sense labels is shown in Table 2. We follow the data split used in previous work for a consistent comparison (Rutherford and Xue, 2014). The PDTB corpus is split into a training set, development set, and test set. Sections 2 to 20 are used to train classifiers. Sections 0 and 1 are used for developing feature sets and tuning models. Section 21 and 22 are used for testing the systems.

To evaluate our method for selecting explicit discourse relation instances, we extract weakly labeled discourse relations from the Gigaword corpus for each class of discourse connective such that the discourse connectives are equally represented within the class. We train and test Maximum Entropy classifiers by adding varying number (1000, 2000, ..., 20000) of randomly selected explicit discourse discourse relation instances to the manually annotated implicit discourse relations in the PDTB as training data. We do this for each class of discourse connectives as presented in Table 1. We perform 30 trials of this experiment and compute average accuracy rates to smooth out the variation from random shuffling of the weakly labeled data.

The statistical models used in this study are from the MALLET implementation with its default setting (McCallum, 2002). Features used in all experiments are taken from the state-of-the-art implicit discourse relation classification system (Rutherford and Xue, 2014). The feature set consists of combinations of various lexical features, production rules, and Brown cluster pairs. These features are described in greater detail by Pitler et al. (2009) and Rutherford and Xue (2014).

Instance reweighting is required when using weakly labeled data because the training set no longer represents the natural distribution of the labels. We reweight each instance such that the sums of the weights of all the instances of the same label are equal. More precisely, if an instance i is from class j, then the weight for the instance w_{ij} is equal to the inverse proportion of class j:

$$w_{ij} = \frac{\text{Number of total instances}}{\text{Size of class } j \cdot \text{Number of classes}} = \frac{\sum_{j'}^{k} c_{j'}}{c_j \cdot k} = \frac{n}{c_j \cdot k}$$

where c_j is the total number of instances from class j and k is the number of classes in the dataset of size n. It is trivial to show that the sum of the weights for all instances from class j is exactly $\frac{n}{k}$ for all classes.

The impact of different classes of weakly labeled explicit discourse connective relations is illustrated in Figure 4. The results show that expicit discourse relations with freely omissible discourse connectives (high OR and low JSD) improve the performance on the standard test set and outperform the other classes of discourse connectives and the naive approach where all of the discourse

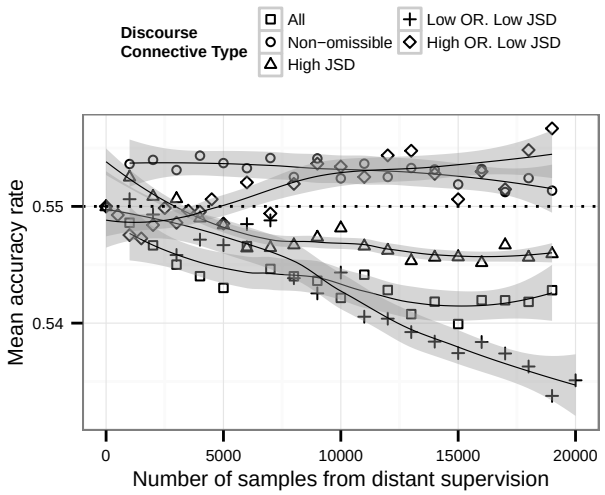

Figure 4: Discourse connectives with high omission rates and low context differentials lead to highest performance boost over the state-of-the-art baseline (dotted line). Each point is an average over multiple trials. The solid lines are LOESS smoothing curves.

connectives are used. In addition, it shows that on average, the system with weakly labeled data from freely omissible discourse connectives continues to rise as we increase the number of samples unlike the other classes of discourse connectives, which show the opposite trend. This suggests that discourse connectives must have both high omission rates and low context differential between implicit and explicit use of the connectives in order to be helpful to the inference of implicit discourse relations.

Table 3 presents results that show, overall, our best performing system, the one using distant supervision from freely omissible discourse connectives, raises the accuracy rate from 0.550 to 0.571 ($p < 0.05$; bootstrap test) and the macro-average F_1 score from 0.384 to 0.405. We achieve such performance after we tune the subset of weakly labeled data to maximize the performance on the development set. Our distant supervision approach improves the performance by adding more weakly labeled data and no additional features.

For a more direct comparison with previous results, we also replicated the state-of-the-art system described in Rutherford and Xue (2014), who follows the practice of the first work on this topic (Pitler et al., 2009) in setting up the task as four binary one vs. other classifiers. The results are presented in Table 4. The results show that the extra data extracted from the Gigaword Corpus is particularly helpful for minority classes such as *Comparison vs. Others* and *Temporal vs Others*, where our current system significantly outperforms that of Rutherford and Xue (2014). Interestingly, the *Expansion vs. Others* classifier

		Baseline features	Baseline + extra data
Expansion	Precision	0.608	0.614
	Recall	0.751	0.788
	F_1	0.672	0.691
Comparison	Precision	0.398	0.449
	Recall	0.228	0.276
	F_1	0.290	0.342
Contingency	Precision	0.465	0.493
	Recall	0.418	0.396
	F_1	0.440	0.439
Temporal	Precision	0.263	0.385
	Recall	0.091	0.091
	F_1	0.135	0.147
Accuracy		0.550	**0.571**
Macro-Average F_1		0.384	**0.405**

Table 3: Our current 4-way classification system outperforms the baseline overall. The difference in accuracy is statistically significant ($p < 0.05$; bootstrap test).

	R&X (2014)	Baseline + extra data	Baseline
Comparison vs Others	0.397	**0.410**	0.380
Contingency vs Others	0.544	0.538	0.539
Expansion vs Others	0.702	0.694	0.679
Temporal vs Others	0.287	**0.333**	0.246

Table 4: The performance of our approach on the binary classification task formulation.

did not improve as the *Expansion* class in the four-way classification (Table 3).

3.3 Just how good is the weakly labeled data?

We performed additional experiments to get a sense of just how good the weakly labeled data extracted from an unlabeled corpus are. Table 5 presents four-way classification results using just the weakly labeled data from the Gigaword Corpus. The results show that the same trend holds when the implicit relations from the PDTB are not included in the training process. The freely omissible discourse connectives achieves the accuracy rate of 0.505, which is significantly higher than the other classes, but they are weaker than the manually labeled data, which achieves the accuracy rate of 0.550 for the same number of training instances.

Weakly labeled data are not perfectly equivalent to the true implicit discourse relations, but they do provide strong enough additional signal. Figure 5 presents experimental results that compare the impact of weakly labeled data from Gigaword Corpus vs gold standard data from the PDTB for the freely omissible class. The mean accuracy rates from the PDTB data are significantly higher than those from the Gigaword Corpus (p <0.05; t-test

Class	Gigaword only	Gigaword + Implicit PDTB
Freely omissible	**0.505**	**0.571**
Omissible	0.313	0.527
Alternating I + II	0.399	0.546
Non-Omissible	0.449	0.554
All of above	0.490	0.547

Table 5: The accuracy rates for the freely omissible class are higher than the ones for the other classes both when using the Gigaword data alone and when using it in conjunction with the implicit relations in the PDTB.

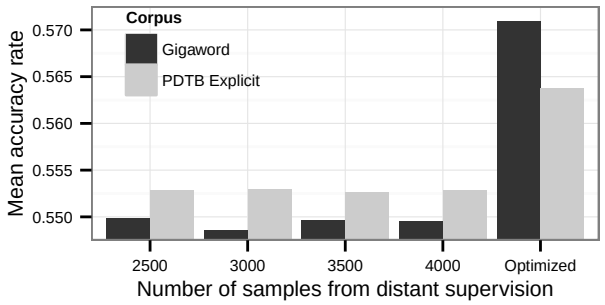

Figure 5: The PDTB corpus leads to more improvement for the same amount of the data. However, Gigaword corpus achieves significantly better performance ($p < 0.05$; bootstrap test) when both models are tuned on the developement set.

and bootstrap test) for the same number of training instances combined with the implicit discourse relations. However, when the number of introduced weakly labeled data exceeds a certain threshold of around 12,000 instances, the performance of the Gigaword corpus rises significantly above the baseline and the explicit PDTB (Figure 4).

The relative superiority of our approach derives precisely from the two selection criteria that we propose. The performance gain does not come from the fact that freely omissible discourse connectives have better coverage of all four senses (Table 6). When all classes are combined equally, the system performs worse as we add more samples although all four senses are covered. The coverage of all four senses is not sufficient for a class of discourse connectives to boost the performance. The two selection criteria are both necessary for the success of this paradigm.

4 Related work

Previous work on implicit discourse relation classification have focused on supervised learning approaches (Lin et al., 2010; Rutherford and Xue, 2014), and the distantly supervised approach using explicit discourse relations

Class	Sense			
	Comp.	Cont.	Exp.	Temp.
Freely omissible	2	6	10	1
Omissible	4	2	5	3
Alternating I	1	0	5	0
Alternating II	2	0	0	3
Non-omissible	0	3	3	2

Table 6: The sense distribution by connective class.

has not shown satisfactory results (Pitler et al., 2009; Park and Cardie, 2012; Wang et al., 2012; Sporleder and Lascarides, 2008) Explicit discourse relations have been used to remedy the sparsity problem or gain extra features with limited success (Biran and McKeown, 2013; Pitler et al., 2009). Our heuristics for extracting discourse relations has been explored in the unsupervised setting (Marcu and Echihabi, 2002), but it has never been evaluated on the gold standard data to show its true efficacy. Our distant supervision approach chooses only certain types of discourse connectives to extract weakly labeled data and is the first of its kind to improve the performance in this task tested on the manually annotated data.

Distant supervision approaches have recently been explored in the context of natural language processing due to the recent capability to process large amount of data. These approaches are known to be particularly useful for relation extraction tasks because training data provided do not suffice for the task and are difficult to obtain (Riloff et al., 1999; Yao et al., 2010). For example, Mintz et al. (2009) acquire a large amount of weakly labeled data based on the Freebase knowledge base and improves the performance of relation extraction. Distantly supervised learning has also recently been demonstrated to be useful for text classification problems (Speriosu et al., 2011; Marchetti-Bowick and Chambers, 2012). For example, Thamrongrattanarit et al. (2013) use simple heuristics to gather weakly labeled data to perform text classification with no manually annotated training data.

Discourse connectives have been studied and classified based on their syntactic properties such subordinating conjunction, adverbials, etc. (Fraser, 2006; Fraser, 1996). While providing a useful insight into how discourse connectives fit into utterances, the syntactic classification does not seem suitable for selecting useful discourse connectives for our purposes of distant supervision for our task.

5 Conclusion and Future Directions

We propose two selection criteria for discourse connectives that can be used to gather weakly labeled data for implicit discourse relation classifiers and improve the performance of the state-of-the-art system without further feature engineering. As part of this goal, we classify dis-

course connectives based on their distributional semantic properties and found that certain classes of discourse connectives cannot be omitted in every context, which plague the weakly labeled data used in previous studies. Our discourse connective classification allows for the better selection of data points for distant supervision.

More importantly, this work presents a new direction in distantly supervised learning paradigm for implicit discourse relation classification. This virtual dramatic increase in the training set size allows for more feature engineering and more sophisticated models. Implicit discourse relation classification is now no longer limited to strictly supervised learning approaches.

Acknowledgments

This work was funded partially by the National Science Foundation via Grant No. 0910532 entitled "Richer Representations for Machine Translation". All views expressed in this paper are those of the authors and do not necessarily represent the view of the National Science Foundation. We also would like to thank Karl Pichotta and Gary Patterson for feedback on the manuscript and the three anonymous reviewers for their suggestions and comments.

References

Or Biran and Kathleen McKeown. 2013. Aggregated word pair features for implicit discourse relation disambiguation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 69–73. The Association for Computational Linguistics.

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

Vanessa Wei Feng and Graeme Hirst. 2012. Text-level discourse parsing with rich linguistic features. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 60–68. Association for Computational Linguistics.

Bruce Fraser. 1996. Pragmatic markers. *Pragmatics*, 6:167–190.

Bruce Fraser. 2006. Towards a theory of discourse markers. *Approaches to discourse particles*, 1:189–204.

D Graff, J Kong, K Chen, and K Maeda. 2007. English gigaword third edition, 2007, ldc 2007t07.

Ben Hutchinson. 2005. Modelling the similarity of discourse connectives. In *Proceedings of the the 27th Annual Meeting of the Cognitive Science Society (CogSci2005)*.

Lillian Lee. 2001. On the effectiveness of the skew divergence for statistical language analysis. In *Artificial Intelligence and Statistics*, volume 2001, pages 65–72.

Junyi Jessy Li and Ani Nenkova. 2014. Reducing sparsity improves the recognition of implicit discourse relations. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 199–207, Philadelphia, PA, U.S.A., June. Association for Computational Linguistics.

Jiwei Li, Rumeng Li, and Eduard Hovy. 2014. Recursive deep models for discourse parsing. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2061–2069.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2010. A PDTB-Styled End-to-End Discourse Parser. *arXiv.org*, November.

N Ljubesic, Damir Boras, Nikola Bakaric, and Jasmina Njavro. 2008. Comparing measures of semantic similarity. In *Information Technology Interfaces, 2008. ITI 2008. 30th International Conference on*, pages 675–682. IEEE.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Micol Marchetti-Bowick and Nathanael Chambers. 2012. Learning for microblogs with distant supervision: Political forecasting with twitter. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 603–612. Association for Computational Linguistics.

Daniel Marcu and Abdessamad Echihabi. 2002. An unsupervised approach to recognizing discourse relations. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 368–375. Association for Computational Linguistics.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. http://www.cs.umass.edu/ mccallum/mallet.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 1003–1011. Association for Computational Linguistics.

Joonsuk Park and Claire Cardie. 2012. Improving implicit discourse relation recognition through feature set optimization. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 108–112. Association for Computational Linguistics.

Gary Patterson and Andrew Kehler. 2013. Predicting the presence of discourse connectives. In *Proceedings*

of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics.

Emily Pitler, Mridhula Raghupathy, Hena Mehta, Ani Nenkova, Alan Lee, and Aravind K Joshi. 2008. Easily identifiable discourse relations. *Technical Reports (CIS)*, page 884.

Emily Pitler, Annie Louis, and Ani Nenkova. 2009. Automatic sense prediction for implicit discourse relations in text. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 683–691. Association for Computational Linguistics.

Rashmi Prasad, Eleni Miltsakaki, Nikhil Dinesh, Alan Lee, Aravind Joshi, Livio Robaldo, and Bonnie L Webber. 2007. The penn discourse treebank 2.0 annotation manual.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K Joshi, and Bonnie L Webber. 2008. The penn discourse treebank 2.0. In *LREC*. Citeseer.

Ellen Riloff, Rosie Jones, et al. 1999. Learning dictionaries for information extraction by multi-level bootstrapping. In *AAAI/IAAI*, pages 474–479.

Attapol T. Rutherford and Nianwen Xue. 2014. Discovering implicit discourse relations through brown cluster pair representation and coreference patterns. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2014)*, Gothenburg, Sweden, April.

Lawrence Schourup. 1999. Discourse markers. *Lingua*, 107(3):227–265.

Michael Speriosu, Nikita Sudan, Sid Upadhyay, and Jason Baldridge. 2011. Twitter polarity classification with label propagation over lexical links and the follower graph. In *Proceedings of the First workshop on Unsupervised Learning in NLP*, pages 53–63. Association for Computational Linguistics.

Caroline Sporleder and Alex Lascarides. 2008. Using automatically labelled examples to classify rhetorical relations: An assessment. *Natural Language Engineering*, 14(03):369–416.

Attapol Thamrongrattanarit, Colin Pollock, Benjamin Goldenberg, and Jason Fennell. 2013. A distant supervision approach for identifying perspectives in unstructured user-generated text. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 922–926, Nagoya, Japan, October. Asian Federation of Natural Language Processing.

Xun Wang, Sujian Li, Jiwei Li, and Wenjie Li. 2012. Implicit discourse relation recognition by selecting typical training examples. In *Proceedings of COLING 2012*, pages 2757–2772, Mumbai, India, December. The COLING 2012 Organizing Committee.

Florian Wolf, Edward Gibson, Amy Fisher, and Meredith Knight. 2005. The discourse graphbank: A database of texts annotated with coherence relations. *Linguistic Data Consortium*.

Limin Yao, Sebastian Riedel, and Andrew McCallum. 2010. Collective cross-document relation extraction without labelled data. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1013–1023. Association for Computational Linguistics.