Discovering Implicit Discourse Relations Through Brown Cluster Pair Representation and Coreference Patterns

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

Sentences form coherent relations in a discourse without discourse connectives more frequently than with connectives. Senses of these implicit discourse relations that hold between a sentence pair, however, are challenging to infer. Here, we employ Brown cluster pairs to represent discourse relation and incorporate coreference patterns to identify senses of implicit discourse relations in naturally occurring text. Our system improves the baseline performance by as much as 25%. Feature analyses suggest that Brown cluster pairs and coreference patterns can reveal many key linguistic characteristics of each type of discourse relation.

1 Introduction

Sentences must be pieced together logically in a discourse to form coherent text. Many discourse relations in the text are signaled explicitly through a closed set of discourse connectives. Simply disambiguating the meaning of discourse connectives can determine whether adjacent clauses are temporally or causally related (Pitler et al., 2008; Wellner et al., 2009). Discourse relations and their senses, however, can also be inferred by the reader even without discourse connectives. These implicit discourse relations in fact outnumber explicit discourse relations in naturally occurring text. Inferring types or senses of implicit discourse relations remains a key challenge in automatic discourse analysis.

A discourse parser requires many subcomponents which form a long pipeline. The implicit discourse relation discovery has been shown to be the main performance bottleneck of an end-to-end parser (Lin et al., 2010). It is also central to many applications such as automatic summarization and question-answering systems.

Existing systems, which make heavy use of word pairs, suffer from data sparsity problem as a word pair in the training data may not appear in the test data. A better representation of two adjacent sentences beyond word pairs could have a significant impact on predicting the sense of the discourse relation that holds between them. Data-driven theory-independent word classification such as Brown clustering should be able to provide a more compact word representation (Brown et al., 1992). Brown clustering algorithm induces a hierarchy of words in a large unannotated corpus based on word co-occurrences within the window. The induced hierarchy might give rise to features that we would otherwise miss. In this paper, we propose to use the cartesian product of Brown cluster assignment of the sentence pair as an alternative abstract word representation for building an implicit discourse relation classifier.

Through word-level semantic commonalities revealed by Brown clusters and entity-level relations revealed by coreference resolution, we might be able to paint a more complete picture of the discourse relation in question. Coreference resolution unveils the patterns of entity realization within the discourse, which might provide clues for the types of the discourse relations. The information about certain entities or mentions in one sentence should be carried over to the next sentence to form a coherent relation. It is possible that coreference chains and semantically-related predicates in the local context might show some patterns that characterize types of discourse relations. We hypothesize that coreferential rates and coreference patterns created by Brown clusters should help characterize different types of discourse relations.

Here, we introduce two novel sets of features for implicit discourse relation classification. Further, we investigate the effects of using Brown clusters as an alternative word representation and analyze the impactful features that arise from

Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 645–654,
Gothenburg, Sweden, April 26-30 2014. ©2014 Association for Computational Linguistics

	Number of instances	
	Implicit	Explicit
COMPARISON	2503 (15.11%)	5589 (33.73%)
CONTINGENCY	4255 (25.68%)	3741 (22.58%)
EXPANSION	8861 (53.48%)	72 (0.43%)
TEMPORAL	950 (5.73%)	3684 (33.73%)
Total	16569 (100%)	13086 (100%)

Table 1: The distribution of senses of implicit discourse relations is imbalanced.

Brown cluster pairs. We also study coreferential patterns in different types of discourse relations in addition to using them to boost the performance of our classifier. These two sets of features along with previously used features outperform the baseline systems by approximately 5% absolute across all categories and reveal many important characteristics of implicit discourse relations.

2 Sense annotation in Penn Discourse Treebank

The Penn Discourse Treebank (PDTB) is the largest corpus richly annotated with explicit and implicit discourse relations and their senses (Prasad et al., 2008). PDTB is drawn from Wall Street Journal articles with overlapping annotations with the Penn Treebank (Marcus et al., 1993). Each discourse relation contains the information about the extent of the arguments, which can be a sentence, a constituent, or an incontiguous span of text. Each discourse relation is also annotated with the sense of the relation that holds between the two arguments. In the case of implicit discourse relations, where the discourse connectives are absent, the most appropriate connective is annotated.

The senses are organized hierarchically. Our focus is on the top level senses because they are the four fundamental discourse relations that various discourse analytic theories seem to converge on (Mann and Thompson, 1988). The top level senses are COMPARISON, CONTINGENCY, EXPANSION, and TEMPORAL.

The explicit and implicit discourse relations almost orthogonally differ in their distributions of senses (Table 1). This difference has a few implications for studying implicit discourse relations and uses of discourse connectives (Patterson and Kehler, 2013). For example, TEMPORAL relations constitute only 5% of the implicit relations but 33% of the explicit relations because they might not be as natural to create without discourse connectives. On the other hand, EXPANSION relations might be more cleanly achieved without ones as indicated by its dominance in the implicit discourse relations. This imbalance in class distribution requires greater care in building statistical classifiers (Wang et al., 2012).

3 Experiment setup

We followed the setup of the previous studies for a fair comparison with the two baseline systems by Pitler et al. (2009) and Park and Cardie (2012). The task is formulated as four separate one-against-all binary classification problems: one for each top level sense of implicit discourse relations. In addition, we add one more classification task with which to test the system. We merge ENTREL with EXPANSION relations to follow the setup used by the two baseline systems. An argument pair is annotated with ENTREL in PDTB if an entity-based coherence and no other type of relation can be identified between the two arguments in the pair. In this study, we assume that the gold standard argument pairs are provided for each relation. Most argument pairs for implicit discourse relations are a pair of adjacent sentences or adjacent clauses separated by a semicolon and should be easily extracted.

The PDTB corpus is split into a training set, development set, and test set the same way as in the baseline systems. Sections 2 to 20 are used to train classifiers. Sections 0–1 are used for developing feature sets and tuning models. Section 21–22 are used for testing the systems.

The statistical models in the following experiments are from MALLET implementation (McCallum, 2002) and libSVM (Chang and Lin, 2011). For all five binary classification tasks, we try Balanced Winnow (Littlestone, 1988), Maximum Entropy, Naive Bayes, and Support Vector Machine. The parameters and the hyperparameters of each classifier are set to their default values. The code for our model along with the data matrices is available at `github.com/attapol/brown_coref_implicit`.

4 Features

Unlike the baseline systems, all of the features in the experiments use the output from automatic natural language processing tools. We use the Stanford CoreNLP suite to lemmatize and part-of-speech tag each word (Toutanova et al., 2003;

Toutanova and Manning, 2000), obtain the phrase structure and dependency parses for each sentence (De Marneffe et al., 2006; Klein and Manning, 2003), identify all named entities (Finkel et al., 2005), and resolve coreference (Raghunathan et al., 2010; Lee et al., 2011; Lee et al., 2013).

4.1 Features used in previous work

The baseline features consist of the following: First, last, and first 3 words, numerical expressions, time expressions, average verb phrase length, modality, General Inquirer tags, polarity, Levin verb classes, and production rules. These features are described in greater detail by Pitler et al. (2009).

4.2 Brown cluster pair features

To generate Brown cluster assignment pair features, we replace each word with its hard Brown cluster assignment. We used the Brown word clusters provided by MetaOptimize (Turian et al., 2010). 3,200 clusters were induced from RCV1 corpus, which contains about 63 million tokens from Reuters English newswire. Then we take the Cartesian product of the Brown cluster assignments of the words in Arg1 and the ones of the words in Arg2. For example, suppose Arg1 has two words $w_{1,1}, w_{1,2}$, Arg2 has three words $w_{2,1}, w_{2,2}, w_{2,3}$, and then $B(.)$ maps a word to its Brown cluster assignment. A word w_{ij} is replaced by its corresponding Brown cluster assignment $b_{ij} = B(w_{ij})$. The resulting word pair features are $(b_{1,1}, b_{2,1})$, $(b_{1,1}, b_{2,2})$, $(b_{1,1}, b_{2,3})$, $(b_{1,2}, b_{2,1})$, $(b_{1,2}, b_{2,2})$, and $(b_{1,2}, b_{2,3})$.

Therefore, this feature set can generate $O(3200^2)$ binary features. The feature set size is orders of magnitude smaller than using the actual words, which can generate $O(V^2)$ distinct binary features where V is the size of the vocabulary.

4.3 Coreference-based features

We want to take advantage of the semantics of the sentence pairs even more by considering how coreferential entities play out in the sentence pairs. We consider various inter-sentential coreference patterns to include as features and also to better describe each type of discourse relation with respect to its place in the coreference chain.

For compactness in explaining the following features, we define *similar words* to be the words assigned to the same Brown cluster.

Number of coreferential pairs: We count the number of inter-sentential coreferential pairs. We expect that EXPANSION relations should be more likely to have coreferential pairs because the detail or information about an entity mentioned in Arg1 should be expanded in Arg2. Therefore, entity sharing might be difficult to avoid.

Similar nouns and verbs: A binary feature indicating whether similar or coreferential nouns are the arguments of the similar predicates. Predicates and arguments are identified by dependency parses. We notice that sometimes the author uses synonyms while trying to expand on the previous predicates or entities. The words that indicate the common topics might be paraphrased, so exact string matching cannot detect whether the two arguments still focus on the same topic. This might be useful for identifying CONTINGENCY relations as they usually discuss two causally-related events that involve two seemingly unrelated agents and/or predicates.

Similar subject or main predicates: A binary feature indicating whether the main verbs of the two arguments have the same subjects or not and another binary feature indicating whether the main verbs are similar or not. For our purposes, the two subjects are said to be the same if they are coreferential or assigned to the same Brown cluster. We notice that COMPARISON relations usually have different subjects for the same main verbs and that TEMPORAL relations usually have the same subjects but different main verbs.

4.4 Feature selection and training sample reweighting

The nature of the task and the dataset poses at least two problems in creating a classifier. First, the classification task requires a large number of features, some of which are too rare and inconducive to parameter estimation. Second, the label distribution is highly imbalanced (Table 1) and this might degrade the performance of the classifiers (Japkowicz, 2000). Recently, Park and Cardie (2012) and Wang et al. (2012) addressed these problems directly by optimally select a subset of features and training samples. Unlike previous work, we do not discard any of data in the training set to balance the label distribution. Instead, we reweight the training samples in each class during parameter estimation such that the performance on the development set is maximized. In addition, the

	Current			Park and Cardie (2012)	Pitler et al. (2009)
	P	R	F_1	F_1	F_1
COMPARISON vs others	27.34	72.41	**39.70**	31.32	21.96
CONTINGENCY vs others	44.52	69.96	**54.42**	49.82	47.13
EXPANSION vs others	59.59	85.50	70.23	-	-
EXP+ENTREL vs others	69.26	95.92	**80.44**	79.22	76.42
TEMPORAL vs others	18.52	63.64	**28.69**	26.57	16.76

Table 2: Our classifier outperform the previous systems across all four tasks without the use of gold-standard parses and coreference resolution.

COMPARISON

Feature set	F_1	% change
All features	39.70	-
All excluding Brown cluster pairs	35.71	-10.05%
All excluding Production rules	37.27	-6.80%
All excluding First, last, and First 3	39.18	-1.40%
All excluding Polarity	39.39	-0.79%

CONTINGENCY

Feature set	F_1	% change
All	54.42	-
All excluding Brown cluster pairs	51.50	-5.37%
All excluding First, last, and First 3	53.56	-1.58%
All excluding Polarity	53.82	-1.10%
All excluding Coreference	53.92	-0.92%

EXPANSION

Feature set	F_1	% change
All	70.23	-
All excluding Brown cluster pairs	67.48	-3.92%
All excluding First, last, and First 3	69.43	-1.14%
All excluding Inquirer tags	69.73	-0.71%
All excluding Polarity	69.92	-0.44%

TEMPORAL

Feature set	F_1	% change
All	28.69	-
All excluding Brown cluster pairs	24.53	-14.50%
All excluding Production rules	26.51	-7.60%
All excluding First, last, and First 3	26.56	-7.42%
All excluding Polarity	27.42	-4.43%

Table 3: Ablation study: The four most impactful feature classes and their relative percentage changes are shown. Brown cluster pair features are the most impactful across all relation types.

number of occurrences for each feature must be greater than a cut-off, which is also tuned on the development set to yield the highest performance on the development set.

5 Results

Our experiments show that the Brown cluster and coreference features along with the features from the baseline systems improve the performance for all discourse relations (Table 2). Consistent with the results from previous work, the Naive Bayes classifier outperforms MaxEnt, Balanced Winnow, and Support Vector Machine across all tasks regardless of feature pruning criteria and training sample reweighting. A possible explanation is that the small dataset size in comparison with the large number of features might favor a generative model like Naive Bayes (Jordan and Ng, 2002). So we only report the performance from the Naive Bayes classifiers.

It is noteworthy that the baseline systems use the gold standard parses provided by the Penn Treebank, but ours does not because we would like to see how our system performs realistically in conjunction with other pre-processing tasks such as lemmatization, parsing, and coreference resolution. Nevertheless, our system still manages to outperform the baseline systems in all relations by a sizable margin.

Our preliminary results on implicit sense classification suggest that the Brown cluster word representation and coreference patterns might be indicative of the senses of the discourse relations, but we would like to know the extent of the impact of these novel feature sets when used in conjunction with other features. To this aim, we conduct an ablation study, where we exclude one of the feature sets at a time and then test the resulting classifier on the test set. We then rank each feature set by the relative percentage change in F_1 score when excluded from the classifier. The data split and experimental setup are identical to the ones described in the previous section but only with Naive Bayes classifiers.

The ablation study results imply that Brown cluster features are the most impactful feature set across all four types of implicit discourse relations. When ablated, Brown cluster features degrade the performance by the largest percentage compared to the other feature sets regardless of the relation types(Table 3). TEMPORAL relations ben-

efit the most from Brown cluster features. Without them, the F_1 score drops by 4.12 absolute or 14.50% relative to the system that uses all of the features.

6 Feature analysis

6.1 Brown cluster features

This feature set is inspired by the word pair features, which are known for its effectiveness in predicting senses of discourse relations between the two arguments. Marcu et al (2002), for instance, artificially generated the implicit discourse relations and used word pair features to perform the classification tasks. Those word pair features work well in this case because their artificially generated dataset is an order of magnitude larger than PDTB. Ideally, we would want to use the word pair features instead of word cluster features if we have enough data to fit the parameters. Consequently, other less sparse handcrafted features prove to be more effective than word pair features for the PDTB data (Pitler et al., 2009). We remedy the sparsity problem by clustering the words that are distributionally similar together and greatly reduce the number of features.

Since the ablation study is not fine-grained enough to spotlight the effectiveness of the individual features, we quantify the predictiveness of each feature by its mutual information. Under Naive Bayes conditional independence assumption, the mutual information between the features and the labels can be efficiently computed in a pairwise fashion. The mutual information between a binary feature X_i and class label Y is defined as:

$$I(X_i, Y) = \sum_{y} \sum_{x=0,1} \hat{p}(x, y) \log \frac{\hat{p}(x, y)}{\hat{p}(x)\hat{p}(y)}$$

$\hat{p}(\cdot)$ is the probability distribution function whose parameters are maximum likelihood estimates from the training set. We compute mutual information for all four one-vs-all classification tasks. The computation is done as part of the training pipeline in MALLET to ensure consistency in parameter estimation and smoothing techniques. We then rank the cluster pair features by mutual information. The results are compactly summarized in bipartite graphs shown in Figure 1, where each edge represents a cluster pair. Since mutual information itself does not indicate whether a feature is favored by one or the other label, we also verify the direction of the effects of each of the features included in the following analysis by comparing the class conditional parameters in the Naive Bayes model.

The most dominant features for COMPARISON classification are the pairs whose members are from the same Brown clusters. We can distinctly see this pattern from the bipartite graph because the nodes on each side are sorted alphabetically. The graph shows many parallel short edges, which suggest that many informative pairs consist of the same clusters. Some of the clusters that participate in such pair consist of named-entities from various categories such as airlines (*King, Bell, Virgin, Continental, ...*), and companies (*Thomson, Volkswagen, Telstra, Siemens*). Some of the pairs form a broad category such as political agents (*citizens, pilots, nationals, taxpayers*) and industries (*power, insurance, mining*). These parallel patterns in the graph demonstrate that implicit COMPARISON relations might be mainly characterized by juxtaposing and explicitly contrasting two different entities in two adjacent sentences.

Without the use of a named-entity recognition system, these Brown cluster pair features effectively act as features that detect whether the two arguments in the relation contain named-entities or nouns from the same categories or not. These more subtle named-entity-related features are cleanly discovered through replacing words with their data-driven Brown clusters without the need for additional layers of pre-processing.

If the words in one cluster semantically relates to the words in another cluster, the two clusters are more likely to become informative features for CONTINGENCY classification. For instance, technical terms in stock and trading (*weighted, Nikkei, composite, diffusion*) pair up with economic terms (*Trading, Interest, Demand, Production*). The cluster with *analysts* and *pundits* pairs up with the one that predominantly contains quantifiers (*actual, exact, ultimate, aggregate*). In addition to this pattern, we observed the same parallel pair pattern we found in COMPARISON classification. These results suggest that in establishing a CONTINGENCY relation implicitly the author might shape the sentences such that they have semantically related words if they do not mention named-entities of the same category.

Through Brown cluster pairs, we obtain features that detect a shift between generality and speci-

Figure 1: The bipartite graphs show the top 40 non-stopword Brown cluster pair features for all four classification tasks. Each node on the left and on the right represents word cluster from Arg1 and Arg2 respectively. We only show the clusters that appear fewer than six times in the top 3,000 pairs to exclude stopwords. Although the four tasks are interrelated, some of the highest mutual information features vary substantially across tasks.

ficity within the scope of the relation. For example, a cluster with industrial categories (*Electric, Motor, Life, Chemical, Automotive*) couples with specific brands or companies (*GM, Ford, Barrick, Anglo*). Or such a pair might simply reflects a shift in plurality e.g. *businesses - business* and *Analysts -analyst*. EXPANSION relations capture relations in which one argument provides a specification of the previous and relations in which one argument provides a generalization of the other. Thus, these shift detection features could help distinguish EXPANSION relations.

We found a few common coreference patterns of names in written English to be useful. First and last name are used in the first sentence to refer to a person who just enters the discourse. That person is referred to just by his/her title and last name in the following sentence. This pattern is found to be

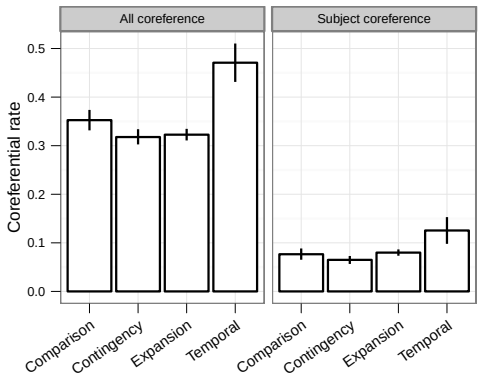

Figure 2: The coreferential rate for TEMPORAL relations is significantly higher than the other three relations ($p < 0.05$, corrected for multiple comparison).

informative for EXPANSION relations. For example, the edges (not shown in the graph due to lack of space) from the first name clusters to the title (*Mr, Mother, Judge, Dr*) cluster.

Time expressions constitutes the majority of the nodes in the bipartite graph for TEMPORAL relations. More strikingly, the specific dates (e.g. clusters that have positive integers smaller than 31) are more frequently found in Arg2 than Arg1 in implicit TEMPORAL relations. It is possible that TEMPORAL relations are more naturally expressed without a discourse connective if a time point is clearly specified in Arg2 but not in Arg1.

TEMPORAL relations might also be implicitly inferred through detecting a shift in quantities. We notice that clusters whose words indicate changes e.g. *increase, rose, loss* pair with number clusters. Sentences in which such pairs participate might be part of a narrative or a report where one expects a change over time. These changes conveyed by the sentences constitute a natural sequence of events that are temporally related but might not need explicit temporal expressions.

6.2 Coreference features

Coreference features are very effective given that they constitute a very small set compared to the other feature sets. In particular, excluding them from the model reduces F_1 scores for TEMPORAL and CONTINGENCY relations by approximately 1% relative to the system that uses all of the features. We found that the sentence pairs in these two types of relations have distinctive coreference patterns.

We count the number of pairs of arguments that are linked by a coreference chain for each type of relation. The coreference chains used in this study are detected automatically from the training set through Stanford CoreNLP suite (Raghunathan et al., 2010; Lee et al., 2011; Lee et al., 2013). TEMPORAL relations have a significantly higher coreferential rate than the other three relations ($p < 0.05$, pair-wise t-test corrected for multiple comparisons). The differences between COMPARISON, CONTINGENCY, and EXPANSION, however, are not statistically significant (Figure 2).

The choice to use or not to use a discourse connective is strongly motivated by linguistic features at the discourse levels (Patterson and Kehler, 2013). Additionally, it is very uncommon to have temporally-related sentences without using explicit discourse connectives. The difference in coreference patterns might be one of the factors that influence the choice of using a discourse connective to signal a TEMPORAL relation. If sentences are coreferentially linked, then it might be more natural to drop a discourse connective because the temporal ordering can be easily inferred without it. For example,

(1) Her story is partly one of personal downfall. [*previously*] She was an unstinting teacher who won laurels and inspired students... (WSJ0044)

The coreference chain between the two temporally-related sentences in (1) can easily be detected. Inserting *previously* as suggested by the annotation from the PDTB corpus does not add to the temporal coherence of the sentences and may be deemed unnecessary. But the presence of coreferential link alone might bias the inference toward TEMPORAL relation while CONTINGENCY might also be inferred.

Additionally, we count the number of pairs of arguments whose grammatical subjects are linked by a coreference chain to reveal the syntactic-coreferential patterns in different relation types. Although this specific pattern seems rare, more than eight percent of all relations have coreferential grammatical subjects. We observe the same statistically significant differences between TEMPORAL relations and the other three types of relations. More interestingly, the subject coreferential rate for CONTINGENCY relations is the lowest among the three categories ($p < 0.05$, pair-wise t-test corrected for multiple comparisons).

It is possible that coreferential subject patterns suggest temporal coherence between the two sentences without using an explicit discourse connective. CONTINGENCY relations, which can only indicate causal relationships when realized implicitly, impose the temporal ordering of events in the arguments; i.e. if Arg1 is causally related to Arg2, then the event described in Arg1 must temporally precede the one in Arg2. Therefore, CONTINGENCY and TEMPORAL can be highly confusable. To understand why this pattern might help distinguish these two types of relations, consider these examples:

(2) He also asserted that exact questions weren't replicated. [*Then*] When referred to the questions that match, he said it was coincidental. (WSJ0045)

(3) He also asserted that exact questions weren't replicated. When referred to the questions that match, she said it was coincidental.

When we switch out the coreferential subject for an arbitrary uncoreferential pronoun as we do in (3), we are more inclined to classify the relation as CONTINGENCY.

7 Related work

Word-pair features are known to work very well in predicting senses of discourse relations in an artificially generated corpus (Marcu and Echihabi, 2002). But when used with a realistic corpus, model parameter estimation suffers from data sparsity problem due to the small dataset size. Biran and McKeown (2013) attempts to solve this problem by aggregating word pairs and estimating weights from an unannotated corpus but only with limited success.

Recent efforts have focused on introducing meaning abstraction and semantic representation between the words in the sentence pair. Pitler et al. (2009) uses external lexicons to replace the one-hot word representation with semantic information such as word polarity and various verb classification based on specific theories (Stone et al., 1968; Levin, 1993). Park and Cardie (2012) selects an optimal subset of these features and establishes the strongest baseline to best of our knowledge.

Brown word clusters are hierarchical clusters induced by frequency of co-occurrences with other words (Brown et al., 1992). The strength of this word class induction method is that the words that are classified to the same clusters usually make an interpretable lexical class by the virtue of their distributional properties. This word representation has been used successfully to augment the performance of many NLP systems (Ritter et al., 2011; Turian et al., 2010).

Louis et al. (2010) uses multiple aspects of coreference as features to classify implicit discourse relations without much success while suggesting many aspects that are worth exploring. In a corpus study by Louis and Nenkova (2010), coreferential rates alone cannot explain all of the relations, and more complex coreference patterns have to be considered.

8 Conclusions

We present statistical classifiers for identifying senses of implicit discourse relations and introduce novel feature sets that exploit distributional similarity and coreference information. Our classifiers outperform the classifiers from previous work in all types of implicit discourse relations. Altogether these results present a stronger baseline for the future research endeavors in implicit discourse relations.

In addition to enhancing the performance of the classifier, Brown word cluster pair features disclose some of the new aspects of implicit discourse relations. The feature analysis confirms our hypothesis that cluster pair features work well because they encapsulate relevant word classes which constitute more complex informative features such as named-entity pairs of the same categories, semantically-related pairs, and pairs that indicate specificity-generality shift. At the discourse level, Brown clustering is superior to a one-hot word representation for identifying intersentential patterns and the interactions between words.

Coreference chains that traverse through the discourse in the text shed the light on different types of relations. The preliminary analysis shows that TEMPORAL relations have much higher inter-argument coreferential rates than the other three senses of relations. Focusing on only subject-coreferential rates, we observe that CONTINGENCY relations show the lowest coreferential rate. The coreference patterns differ substantially and meaningfully across discourse relations and deserve further exploration.

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.

Peter F Brown, Peter V deSouza, Robert L Mercer, Vincent J Della Pietra, and Jenifer C Lai. 1992. Class-based n -gram models of natural language. *Computational Linguistics*, 18(4):467–479, December.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Marie-Catherine De Marneffe, Bill MacCartney, Christopher D Manning, et al. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC*, volume 6, pages 449–454.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 363–370. Association for Computational Linguistics.

Nathalie Japkowicz. 2000. Learning from imbalanced data sets: a comparison of various strategies. In *AAAI workshop on learning from imbalanced data sets*, volume 68.

Michael Jordan and Andrew Ng. 2002. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Advances in neural information processing systems*, 14:841.

Dan Klein and Christopher D Manning. 2003. Accurate unlexicalized parsing. In *the 41st Annual Meeting*, pages 423–430, Morristown, NJ, USA. Association for Computational Linguistics.

Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2011. Stanford's multi-pass sieve coreference resolution system at the conll-2011 shared task. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 28–34. Association for Computational Linguistics.

Heeyoung Lee, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2013. Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics*.

Beth Levin. 1993. *English verb classes and alternations: A preliminary investigation*, volume 348. University of Chicago press Chicago.

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

Nick Littlestone. 1988. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine learning*, 2(4):285–318.

Annie Louis and Ani Nenkova. 2010. Creating local coherence: An empirical assessment. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 313–316. Association for Computational Linguistics.

Annie Louis, Aravind Joshi, Rashmi Prasad, and Ani Nenkova. 2010. Using entity features to classify implicit discourse relations. In *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 59–62. Association for Computational Linguistics.

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

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.

Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330.

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

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, 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.

Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher Manning. 2010. A multi-pass sieve for coreference resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 492–501, Stroudsburg, PA, USA. Association for Computational Linguistics.

Alan Ritter, Sam Clark, Oren Etzioni, et al. 2011. Named entity recognition in tweets: an experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1524–1534. Association for Computational Linguistics.

Philip Stone, Dexter C Dunphy, Marshall S Smith, and DM Ogilvie. 1968. The general inquirer: A computer approach to content analysis. *Journal of Regional Science*, 8(1).

Kristina Toutanova and Christopher D Manning. 2000. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics-Volume 13*, pages 63–70. Association for Computational Linguistics.

Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 173–180. Association for Computational Linguistics.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 384–394. Association for Computational Linguistics.

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.

Ben Wellner, James Pustejovsky, Catherine Havasi, Anna Rumshisky, and Roser Sauri. 2009. Classification of discourse coherence relations: An exploratory study using multiple knowledge sources. In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, pages 117–125. Association for Computational Linguistics.