Towards Semi-Supervised Classification of Discourse Relations using Feature Correlations

Hugo Hernault and **Danushka Bollegala** and **Mitsuru Ishizuka**

Graduate School of Information Science & Technology

The University of Tokyo

7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan

`hugo@mi.ci.i.u-tokyo.ac.jp`
`danushka@iba.t.u-tokyo.ac.jp`
`ishizuka@i.u-tokyo.ac.jp`

Abstract

Two of the main corpora available for training discourse relation classifiers are the RST Discourse Treebank (RST-DT) and the Penn Discourse Treebank (PDTB), which are both based on the Wall Street Journal corpus. Most recent work using discourse relation classifiers have employed fully-supervised methods on these corpora. However, certain discourse relations have little labeled data, causing low classification performance for their associated classes. In this paper, we attempt to tackle this problem by employing a semi-supervised method for discourse relation classification. The proposed method is based on the analysis of feature co-occurrences in unlabeled data. This information is then used as a basis to extend the feature vectors during training. The proposed method is evaluated on both RST-DT and PDTB, where it significantly outperformed baseline classifiers. We believe that the proposed method is a first step towards improving classification performance, particularly for discourse relations lacking annotated data.

1 Introduction

The RST Discourse Treebank (RST-DT) (Carlson et al., 2001), based on the Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) framework, and the Penn Discourse Treebank (PDTB) (Prasad et al., 2008), are two of the most widely-used corpora for training discourse relation classifiers. They are both based on the Wall Street Journal (WSJ) corpus, although there are substantial differences in the relation taxonomy used to annotate the corpus. These corpora have been used in most of the recent work employing discourse relation classifiers, which are based on fully-supervised machine learning approaches (duVerle and Prendinger, 2009; Pitler et al., 2009; Lin et al., 2009).

Still, when building a discourse relation classifier on either corpus, one is faced with the same practical issue: Certain relations are very prevalent, such as ELABORATION[N][S] (RST-DT), with more than 4000 instances, whereas other occur rarely, such as EVALUATION[N][N][1] (RST-DT), with three instances, or COMPARISON.PRAGMATIC CONCESSION (PDTB), with 12 instances. This lack of training data causes poor classification performance on the classes associated to these relations.

In this paper, we try to tackle this problem by using feature co-occurrence information, extracted from unlabeled data, as a way to inform the classifier when unseen features are found in test vectors. The advantage of the method is that it relies solely on unlabeled data, which is abundant, and cheap to collect.

The contributions of this paper are the following: First, we propose a semi-supervised method that exploits the abundant, freely-available unlabeled data, which is harvested for feature co-occurrence information, and used as a basis to extend feature vectors to help classification for cases where unknown features are found in test vectors. Second, the proposed method is evaluated on the RST-DT and PDTB corpus, where it significantly improves F-score when trained on moderately small datasets. For instance, when trained on a dataset with around 1000 instances, the proposed method increases the macro-average F-score up to 30%, compared to a baseline classifier.

2 Related Work

Since the release in 2002 of the RST-DT corpus, several fully-supervised discourse parsers have

---

[1] We use the notation [N] and [S] respectively to denote the nucleus and satellite in a RST discourse relation.

been built in the RST framework. In duVerle and Prendinger (2009), a discourse parser based on Support Vector Machines (SVM) (Vapnik, 1995) is proposed. Shallow lexical, syntactic and structural features, including 'dominance sets' (Soricut and Marcu, 2003) are used.

The unsupervised method of Marcu and Echihabi (2002) was the first to try to detect 'implicit' relations (i.e. relations not accompanied by a cue phrase, such as *'however'*, *'but'*), using word pairs extracted from two spans of text. Their method attempts to capture the difference of polarity in words.

Discourse relation classifiers have also been trained using PDTB. Pitler et al. (2008) performed a corpus study of the PDTB, and found that 'explicit' relations can be most of the times distinguished by their discourse connectives.

Lin et al. (2009) studied the problem of detecting implicit relations in PDTB. Their relational classifier is trained using features extracted from dependency paths, contextual information, word pairs and production rules in parse trees. For the same task, Pitler et al. (2009) also use word pairs, as well as several other types of features such as verb classes, modality, context, and lexical features.

In this paper, we are not aiming at defining novel features for improving performance in RST or PDTB relation classification. Instead we incorporate features that have already shown to be useful for discourse relation learning and explore the possibilities of using unlabeled data for this task.

3 Method

In this section, we describe a semi-supervised method for relation classification, based on feature vector extension. The extension process employs feature co-occurrence information. Co-occurrence information is useful in this context as, for instance, we might know that the word pair *(for, when)* is a good indicator of a TEMPORAL relation. Or, after analyzing a large body of unlabeled data, we might also notice that this word pair co-occurs often with the word 'run-up' placed at the end of a span of text. Suppose now that we have to classify a test instance containing the feature 'run-up', but not the word pair *(for, when)*. In this case, by using the co-occurrence information, we know that the instance has a chance of being a TEMPORAL relation. We first explain how to compute

a feature correlation matrix, using unlabeled data. In a second section, we show how to extend feature vectors in order to include co-occurrence information. Finally, we describe the features used in the discourse relation classifiers.

3.1 Feature Correlation Matrix

A training/test instance is represented using a d-dimensional feature vector $\mathbf{f} = [f_1, \ldots, f_d]^{\mathrm{T}}$, where $f_i \in \{0, 1\}$. We define a *feature correlation matrix*, C such that the (i, j)-th element of C, $C_{(i,j)} \in \{0, 1\}$ denotes the correlation between the two features f_i and f_j. If both f_i and f_j appear in a feature vector then we define them to be co-occurring. The number of different feature vectors in which f_i and f_j co-occur is used as a basis to compute $C_{(i,j)}$. Importantly, feature correlations can be calculated using only unlabeled data.

It is noteworthy that feature correlation matrices can be computed using any correlation measure. For the current task we use the χ^2-measure (Plackett, 1983) as the preferred correlation measure because of its simplicity. We create the feature correlation matrix C, such that, for all pairs of features (f_i, f_j),

$$C_{(i,j)} = \left\{ \begin{array}{l} 1 \text{ if } \chi_{i,j}^2 > c \\ 0 \text{ otherwise} \end{array} \right. . \qquad (1)$$

Here c is the critical value, which, for a confidence level of 0.05 and one degree of freedom, can be set to 3.84.

3.2 Feature Vector Extension

Once the feature correlation matrix is computed using unlabeled data as described in Section 3.1, we can use it to extend a feature vector during testing. One of the reasons explaining why a classifier might perform poorly on a test instance, is that there are features in the test instance that were not observed during training. Let us represent the feature vector corresponding to a test instance x by $\mathbf{f_x}$. Then, we use the feature correlation matrix to find the set of correlated features $F_c(f_i)$ of a particular feature f_i that occur in $\mathbf{f_x}$.

Specifically, for a feature $f_i \in \mathbf{f_x}$, $F'(f_i)$ consists of features f_j, where $C_{(i,j)} = 1$. We define the extended feature vector $\mathbf{f'_x}$ of $\mathbf{f_x}$ as the union of all the features that appear in $\mathbf{f_x}$ and $F_c(f_x)$. Since a discourse relation is defined between two spans of short texts (elementary discourse units), which are typically two clauses or sentences, a particular feature does not usually occur more than once

in a feature vector. Therefore, we introduced the proposed method in the context of binary valued features. However, the above mentioned discussion can be naturally extended to cover real-valued features.

3.3 Features

Figure 1 shows the parse tree for a sentence composed of two discourse units, which serve as arguments of a discourse relation we want to generate a feature vector from. Lexical heads have been calculated using the projection rules of Magerman (1995), and indicated between brackets. For each argument, surrounded by dots, is the minimal set of sub-parse trees containing strictly all the words of the argument.

We extract all possible lemmatized word pairs from the two arguments. Next, we extract from left and right argument separately, all production rules from the sub-parse trees. Finally, we encode in our features three nodes of the parse tree, which capture the local context at the connection point between the two arguments (Soricut and Marcu, 2003): The first node, which we call N_w, is the highest ancestor of the first argument's last word w, and is such that N_w's right-sibling is the ancestor of the second argument's first word. N_w's right-sibling node is called N_r. Finally, we call N_p the parent of N_w and N_r. For each node, we encode in the feature vector its part-of-speech (POS) and lexical head. For instance, in Figure 1, we have N_w = S(comment), N_r = SBAR(when), and N_p = VP(declined).

4 Experiments

It is worth noting that the proposed method is independent of any particular classification algorithm. As our goal is strictly to evaluate the relative benefit of employing the proposed method, we select a logistic regression classifier, for its simplicity. We used the multi-class logistic regression (maximum entropy model) implemented in Classias (Okazaki, 2009). Regularization parameters are set to their default value of one.

Unlabeled instances are created by selecting texts of the WSJ, and segmenting them into elementary discourse units (EDUs) using our sequential discourse segmenter (Hernault et al., 2010). As there is no segmentation tool for the PDTB framework, we assumed that feature correlation information taken from EDUs created using a RST

segmenter is also useful for extending feature vectors of PDTB relations.

Since we are interested in measuring the overall performance of a discourse relation classifier across all relation types, we use macro-averaged F-score as the preferred evaluation metric for this task. We train a multi-class logistic regression model without extending the feature vectors as a baseline method. This baseline is expected to show the effect of using the proposed feature extension approach for the task of discourse relation learning.

Experimental results on RST-DT and PDTB datasets are depicted in Figures 2 and 3. We observe that the proposed feature extension method outperforms the baseline for both RST-DT and PDTB datasets for the full range of training dataset sizes. However, the difference between the two methods decreases as we increase the amount of training data. Specifically, with 200 training instances, for RST-DT, the baseline method has a macro-averaged F-score of 0.079, whereas the the proposed method has a macro-averaged F-score of 0.159 (around 101% increase in F-score). For 1000 training instances, the F-score for RST-DT increases by 29.2%, from 0.143 to 0.185, while the F-score for PDTB increases by 27.9%, from 0.109 to 0.139. However, the difference between the two methods diminishes beyond 10000 training instances.
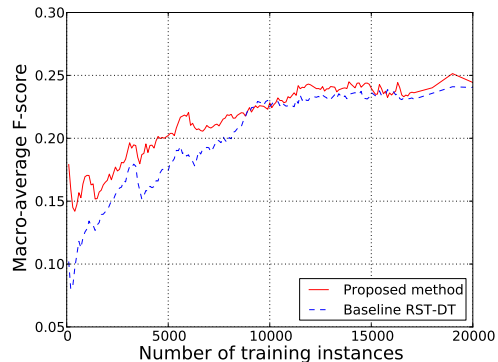

Figure 2: Macro-average F-score (RST-DT) as a function of the number of training instances used.

5 Conclusion

We presented a semi-supervised method for improving the performance of discourse relation classifiers. The proposed method is based on the analysis of co-occurrence information harvested from unlabeled data only. We evaluated
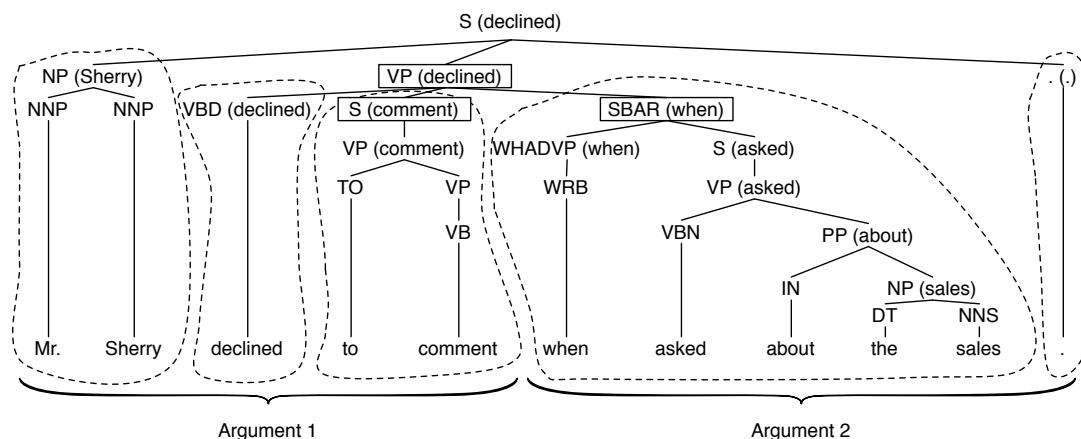
Figure 1: Two arguments of a discourse relation, and the minimum set of subtrees that contain them—lexical heads are indicated between brackets.

Figure 3: Macro-average F-score (PDTB) as a function of the number of training instances used.

the method on two of the most widely-used discourse corpora, RST-DT and PDTB. The method performs significantly better than a baseline classifier trained on the same features, especially when the number of labeled instances used for training is small. For instance, using 1000 training instances, we observed an increase of nearly 30% in macro-average F-score. This is an interesting perspective for improving classification performance of relations with little training data. In the future, we plan to improve the method by employing ranked co-occurrences. This way, only the most relevant correlated features can be selected during feature vector extension. Finally, we plan to investigate using larger amounts of unlabeled training data.

References

L. Carlson, D. Marcu, and M. E. Okurowski. 2001. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. *Proc. of Second SIGdial Workshop on Discourse and Dialogue-Volume 16*, pages 1–10.

D. A. duVerle and H. Prendinger. 2009. A novel discourse parser based on Support Vector Machine classification. In *Proc. of ACL'09*, pages 665–673.

H. Hernault, D. Bollegala, and M. Ishizuka. 2010. A sequential model for discourse segmentation. In *Proc. of CICLing'10*, pages 315–326.

Z. Lin, M-Y. Kan, and H. T. Ng. 2009. Recognizing implicit discourse relations in the Penn Discourse Treebank. In *Proc. of EMNLP'09*, pages 343–351.

D. M. Magerman. 1995. Statistical decision-tree models for parsing. *Proc. of ACL'95*, pages 276–283.

W. C. Mann and S. A. Thompson. 1988. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

D. Marcu and A. Echihabi. 2002. An unsupervised approach to recognizing discourse relations. In *Proc. of ACL'02*, pages 368–375.

N. Okazaki. 2009. Classias: A collection of machine-learning algorithms for classification.

E. Pitler, M. Raghupathy, H. Mehta, A. Nenkova, A. Lee, and A. Joshi. 2008. Easily identifiable discourse relations. In *Proc. of COLING'08 (Posters)*, pages 87–90.

E. Pitler, A. Louis, and A. Nenkova. 2009. Automatic sense prediction for implicit discourse relations in text. In *Proc. of ACL'09*, pages 683–691.

R. L. Plackett. 1983. Karl Pearson and the chi-squared test. *International Statistical Review*, 51(1):59–72.

R. Prasad, N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi, and B. Webber. 2008. The Penn Discourse Treebank 2.0. In *Proc. of LREC'08*.

R. Soricut and D. Marcu. 2003. Sentence level discourse parsing using syntactic and lexical information. *Proc. of NA-ACL'03*, 1:149–156.

V. N. Vapnik. 1995. *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc.