Discourse Relation Sense Classification Systems for CoNLL-2016 Shared Task

Ping Jian, Xiaohan She, Chenwei Zhang, Pengcheng Zhang, Jian Feng
School of Computer Science and Technology, Beijing Institute of Technology
pjian,xhshe,zcwhzy,pengchengzhang,zzhy@bit.edu.cn

Abstract

This paper reports the submitted discourse relation classification systems of the language information processing group of Beijing Institute of Technology (BIT) to the CoNLL-2016 shared task. In this work, discriminative methods were employed according to the different characteristics of English and Chinese discourse structures. Additionally, distributed representations were introduced to catch the deep semantic relations. Experiments shows their effectiveness on both English and Chinese tasks.

1. Introduction

In natural language processing (NLP), discourse parsing is the process of understanding the internal structure of a text and identifying the discourse relations in between its text unites (Lin et al., 2014). It is a recognized challenging task since deep semantic understanding and discourse wide global information even world knowledge are essential to achieve well acceptable solutions. According to alternative discourse structure theoretical frameworks, RST-DT Corpus (Carlson et al., 2003) provides the possibility of data-driven modeling for complete tree structure while PDTB (Prasad et al., 2008) offers a framework to predicting shallow discourse structures statistically in a "predicate-argument" style. Compared with RST-DT, PDTB is larger, so it draws more attentions in these years to support discourse parsing model verification.

In this situation, CoNLL launched Shallow Discourse Parsing Shared Task in the year 2015[1] and called for PDTB-styled individual discourse relations that are presented in a free text under an end-to-end paradigm (Xue et al., 2015). According to the annotation framework of PDTB, relations held between arguments can be either explicit or non-explicit. Non-explicit relations are further divided into implicit, EntRel and AltLex ones (Prasad et al., 2008). In CoNLL-2015 Shared Task, the PDTB senses were regularized into more reasonable 15 categories to facilitate machine learning (Xue et al., 2015). Participants were required to run their systems on a web-based evaluation platform and the systems should (1) locate the explicit discourse connectives (e.g., "because", "however") in the text, (2) identify the spans of text that serve as the two arguments for each discourse connective, and (3) predict the sense of the discourse relations (e.g., "Cause", "Condition", "Contrast").

This is the 2nd edition of the CoNLL Shared Task on Shallow Discourse Parsing this year. Besides the English PDTB-styled end-to-end paradigm, PDTB-styled Chinese end-to-end parsing is also involved (Xue et al., 2016). It is attributed to the annotation of discourse structures in Chinese texts, a PDTB-styled Chinese discourse Treebank (CDTB) (Zhou and Xue, 2012). Based on the adapted PDTB annotation scheme, discourse structures in CDTB own the same "predicate-argument" pattern and similar sense hierarchy.

The same as English discourse parsing, the CDTB sense in CoNLL-2016 Shared Task is also transferred. 8 categories for explicit and non-explicit relations are refactored: "Causation", "Conditional", "Conjunction", "Contrast", "Expansion", "Purpose", "Temporal" and "Progression".

In addition to the Chinese discourse parsing, CoNLL-2016 Shared Task also allows participants to do the supplementary task which is sense classification using gold standard argument pairs both in English and Chinese. It is proved that implicit sense discrimination is the most difficult subtask in discourse parsing, not only as an individual task but also as a key component in pipeline end-to-end system (Hong et al., 2012; Lin et al., 2014). Implicit discourse relation is also the most attended issue at the

[1] http://www.cs.brandeis.edu/~clp/conll15st/

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 158–163,
Berlin, Germany, August 7-12, 2016. ©2016 Association for Computational Linguistics

beginning of the release of PDTB (Pitler et al., 2008; Lin et al., 2009; Zhou et al., 2010; Prasad et al., 2010).

Due to the lack of effective structural semantic representation model, discourse relation sense disambiguation, which is a deep semantic analysis problem, is always conducted by modeling large scale shallow linguistic features. We can see that the named efficient features such as lexical and syntactic features (word co-occurrences, function words, phrase or dependency parses), partial shallow semantic features (co-reference patterns, semantic attribute of words, e.g., polarity) and a few dynamic features are adopted in existing works (Marcu and Echihabi, 2002; Pitler et al., 2008; Lin et al., 2009; Zhou et al., 2010; Prasad et al., 2010; Feng and Hirst, 2012; Rutherford and Xue, 2014). In response to the data scarcity problem, semi-supervised and unsupervised methods are explored for implicit relations inference in recent years (Hernault et al., 2011; Hong et al., 2012; Lan et al., 2013; Fisher and Simmons, 2015). Experiments demonstrate that these kinds of methods can acquire more stable statistical distribution via large scale unlabeled corpus hence achieve higher classification accuracy.

In this Shared Task, we focus on the supplementary task and submit both the English and Chinese discourse relation sense classification systems. According to the different characteristics of English and Chinese discourse structures, we examine rule-based and statistical discriminative classification approaches, conventional and distributed semantic representation models, as well as the expressiveness of extra resources.

The organization of this work is as follows. Section 2 presents our explicit relation classifiers. Section 3 gives the description of the non-explicit relation classification models in our system. Section 4 reports the preliminary experimental results on the training and development dataset, and the final results on two test datasets. Conclusions are provided in Section 5.

2. Explicit Discourse Relation Sense Classification

The explicit discourse relation refers to the relationship between two elementary discourse units which are connected by a discourse connective. As pointed in (Dinesh et al., 2005), the connective itself is a very good feature for sense discrimination, because only a few connectives are ambiguous. In CDTB, this phenomenon is more common.

Num. of different senses	Ratio of connectives	Frequency	Ratio of frequency
1	30%	623	3.4%
2	17%	1214	6.6%
3	14%	1025	5.6%
4	5%	1799	9.6%
more than 4	34%	13783	74.8%

Table 1: Distributions of the connectives and relation senses in the training set from PDTB

Num. of different senses	Ratio of connectives	Frequency	Ratio of frequency
1	64.9%	5525	66.6%
2	18.6%	1997	25.4%
3	12.4%	594	7.6%
4	4.1%	32	0.4%
more than 4	0%	0	0%

Table 2: Distributions of the connectives and relation senses in the training set from CDTB

Table 1 and Table 2 show the distributions of the connectives and the relation senses they acting in the PDTB training set and CDTB training set respectively. In training set extracted from PDTB, 30% connectives act as unique sense and these connectives appear 623 times totally in the set, which occupy only 3.4% in all of the tokens. Whereas, there are 64.9% connectives express unique sense in Chinese texts and their frequency achieves 2/3. On the whole, we can see that more than 92% connective tokens correspond to less than 3 relation senses in CDTB. On the contrary, nearly 85% connective tokens correspond to more than 3 relation senses in PDTB.

We further check the different senses' distribution of ambiguous connectives. There are 85% ambiguous connectives in Chinese texts tend to express one sense, and the reliability of this tendency is 90%. For example, the connective “不过” acts as two senses in the training set: “Contrast” and “Expansion”. But the number of “Contrast” samples is 430 while “Expansion” appears only 10 times.

In a word, compared with English, Chinese connectives present less sense perplexity when forming the discourse structures.

2.1 Explicit Relation Classification for English

We employ a SVM classifier to predict the sense of connectives in English task. Following the work of Lin et al. (2014), three features are introduced to train the classifier: the connective itself, its POS tag and the previous word.

Connective	Sense
不过	Contrast
并	Conjunction
但是	Contrast
…	…
通过	Causation

Table 3: Part of the connective-sense table used in Chinese connective sense classification

2.2 Explicit Relation Classification for Chinese

According to the analyses on the sense distribution of Chinese connectives, we prefer rule-based method to conduct explicit relation classification on CDTB.

We calculate the probability distribution of the discourse relation for each connective:

$$\mathrm{p}(s_j|c_i) = \frac{\mathrm{num}(s_j, c_i)}{\sum_{s\in S}\mathrm{num}(s, c_i)}$$

where $\mathrm{num}(s_j, c_i)$ is the number of connective c_i acting as sense s_j. Connectives are classified to the sense who has the maximum probability $\mathrm{p}(s_j|c_i)$ in the test set. It is safe in most cases because the majority of Chinese connectives tend to express unique relations sense. Table 3 shows a part of our connective-sense table.

As no extra resources were employed in above models, our explicit classification systems were conducted in the closed track.

3. Non-explicit Discourse Relation Sense Classification

The non-explicit discourse relation refers to the relationship expressed implicitly, lexicalized or entity-based inferred between abstract object units[2]. As a typical classification problem, we build a SVM classifier to predict the senses and put attentions on more efficient feature representations.

We employ three primary features which perform well in our preliminary study:

Polarity Tags: Polarity is always a useful feature when processing semantic problems. We count the number of positive, negative and neutral words in the given abstract units (which are called Arg1 and Arg2 in the following) as an intuitional feature for non-explicit relation disambiguation. All of content words' polarity is derived from Multi-perspective Question Answering Opinion Corpus (Wilson et al., 2005) in English, and HowNet[3] in Chinese.

Inquirer Tags: Verb is one of the most important components bearing the semantic information of a sentence. The General Inquirer lexicon (Stone et al., 1966) provides semantic categories of verbs and we sum the Inquirer tags of verbs appeared in Arg1 and Arg2 of English sentences. We prefer the General Inquirer lexicon rather than the provided VerbNet because the former has much more information when dealing with synsets.

Word Pairs: Extracting words respectively from Arg1 and Arg2 has been proved to be helpful for implicit discourse relation prediction (Pitler et al., 2009). But there is still disagreement on the use of the function words. Due to probable data sparseness, we ignore all of function words in both arguments and focus on only content words in our systems. Also as a way to release the sparseness, we use information gain to reduce the dimension of word pairs and keep more discriminative ones.

3.1 Distributed Representation in Implicit Relation Classification

To enhance the semantically expressing power of lexical features, distributed representation is introduced into our implicit relation prediction in different ways.

Simple Embedding: We generate embedding for each word pair by catenating the embedding of its member words one by one. The average of those word-pairs' embedding is brought to replace the one-hot representation of the word pair in the classification.

Huffman Tree-based Prediction: As one of the significant optimization methods in word embedding, hierarchical softmax (Mikolov et al., 2013) predicts the most probable word to co-occur with the corresponding context. All words appeared in the training set are stored in a Huffman tree, organized by word frequency. The Huffman tree which is demonstrated to take efficiency and overfitting issue into account is expected to be a more advanced structure to incorporate distributed representations. Furthermore, the Huffman tree takes the prior probabilities of the connective candidates into account via locating them at the different positions (depths) in the tree. It is expected to achieve better performance than simple embedding and SVM

[2] Because the EntRel and AltLex relations are incorporated into the implicit ones to induce an integrated disambiguation, we call all of them "implicit relations" in the following sections for simplicity.

[3] http://www.keenage.com/

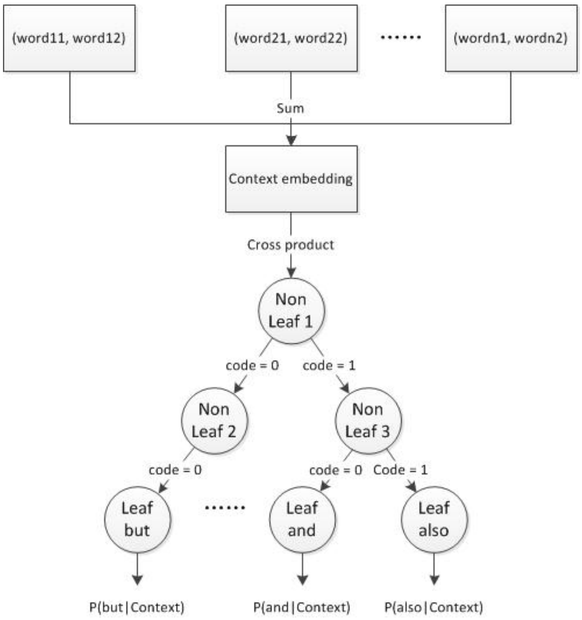

Figure 1: Huffman Tree-based prediction for implicit connectives

classifiers.

The original objective function in Huffman tree prediction is to calculate words' probability when the corresponding context is given. We set the context as content word pairs extracted from the arguments, and all of implicit discourse connectives are going to be predicted. The prediction process is illustrated in Figure 1.

In the Huffman Tree-based prediction, word pair vectors are summed to make context embedding. The posterior probability of each connective[4] predicted is put together to build a new feature for the SVM classifier.

We utilize a larger scale corpus "Central News Agency of Taiwan, English Service"[5] (CNA) to train the Huffman tree. All the explicit discourse relations are extracted from the corpus by pattern matching (Marcu and Echihabi, 2002) and the explicit connectives are dropped to make "pseudo-implicit" training samples.

3.2 Implicit Relation Classification models for English and Chinese

Including the conditional classification with one-hot representation, we build up three comparative models for English implicit relation task (Table 4) and two for Chinese task (Table 5). Since the inconsistency between the distributions of the "pseudo-implicit" and real implicit instances is more serious in Chinese, Huffman Tree-based Prediction is not conducted for Chinese task.

Learning method	Resources	Extra resources
SVM with One-Hot features	MPQA	General Inquirer lexicon
SVM with Simple Embedding	MPQA, word embeddings	General Inquirer lexicon
SVM with Simple Embedding+Huff. Tree Prediction	*MPQA, word embeddings*	*General Inquirer lexicon, CNA*

Table 4: Comparative models for English implicit relation prediction. The submitted model is in italic.

Learning method	Resources	Extra resources
SVM with One-Hot features	No	HowNet
SVM with Simple Embedding	*Word embeddings*	*HowHet*

Table 5: Comparative models for Chinese implicit relation prediction. The submitted model is in italic.

As the sparseness of word pairs is more severe in Chinese situation, a strategy of **Word Pairs Fuzzy Matching** is proposed: Based on the word embedding library, some word similarity groups are formed to ensure that the majority of arguments to be disambiguated contain discriminative word pairs.

4. Experiments

The same as the CoNLL-2015's task, participants are required to deploy their systems on the provided platform instead of submitting the output. The organizer also offers potentially useful linguistic resources for the closed track. In this section, the experimental results are presented and the experimental analyses are induced. All the systems are evaluated on TIRA evaluation platform (Potthast et al., 2014).

4.1 Explicit Relation Classification Experiments

Table 6 presents the English connective sense classification results conducted by SVM classifier. All the SVM classifiers utilized in our experiments were implemented by the LibSVM[6].

Unfortunately, we submitted a wrong edition of our system during the competition for technical reasons and the official outputs produced by this edition are also listed in Table 6 (*System submitted*).

Sense classification results for Chinese connectives are displayed in Table 7. For compari-

[4] For the sparseness issue, implicit connectives which appear more than 1% of all the implicit relation instances are considered and the dimension of the feature vector is 19 in our model.

[5] https://catalog.ldc.upenn.edu/LDC2011T07

[6] http://www.csie.ntu.edu.tw/~cjlin/libsvm/index.html

Method	Dev			Test			Blind test		
	P	R	F1	P	R	F1	P	R	F1
SVM classification	**89.89**	**62.06**	**73.43**	**87.10**	**53.47**	**66.26**	**75.44**	**61.87**	**67.98**
System submitted	*23.22*	*23.22*	*23.22*	*24.62*	*24.62*	*24.62*	*17.99*	*17.99*	*17.99*

Table 6: Explicit connective sense classification results for English. The system submitted is in italic.

Method	Dev			Test			Blind test		
	P	R	F1	P	R	F1	P	R	F1
Rule-based	***92.21***	***92.21***	***92.21***	***94.74***	***93.75***	***94.24***	***75.27***	***75.27***	***75.27***
SVM classification	71.43	71.43	71.43	79.17	79.17	79.17	45.94	45.94	45.94

Table 7: Explicit connective sense classification results for Chinese. The system submitted is in italic.

Method	Dev			Test			Blind test		
	P	R	F1	P	R	F1	P	R	F1
SVM with One-Hot features	16.56	16.56	16.56	15.89	15.89	15.89	18.22	18.22	18.22
SVM with Simple Embedding	17.09	17.09	17.09	16.39	16.39	16.39	18.99	18.99	18.99
SVM with Simple Embedding +Huff. Tree Prediction	***17.36***	***17.36***	***17.36***	***16.58***	***16.58***	***16.58***	***19.30***	***19.30***	***19.30***

Table 8: Implicit relation sense classification results for English. The system submitted is in italic.

Method	Dev			Test			Blind test		
	P	R	F1	P	R	F1	P	R	F1
SVM with One-Hot features	15.69	15.69	15.69	11.42	11.42	11.42	16.29	16.29	16.29
SVM with Simple Embedding	***21.90***	***21.90***	***21.90***	***21.73***	***21.73***	***21.73***	***18.11***	***18.11***	***18.11***

Table 9 Implicit relation sense classification results for Chinese. The system submitted is in italic.

son, we also conducted a typical SVM classifier in which two features are applied: the connective itself and its POS tag. Since the Chinese training data is much smaller and there are too many low-frequency connectives involved, only the connectives which appear more than 10 times are considered in the experiment. Because of the serious imbalance and small quantity of training samples, the SVM classifier gets a poor classification precisions. Whereas, the rule-based approach performs soundly and achieves acceptable results. It is simple, crude but practically effective in Chinese explicit relation classification.

4.2 Implicit Relation Classification Experiments

The implicit relation sense classification results for English and Chinese are listed in Table 8 and Table 9 respectively.

As we can see, although the overall performance of English system is not good enough, the results of Simple Embedding and Huffman Tree-based Prediction are always better than the One-Hot paradigm. The Huffman Tree Prediction outperforms the Simple Embedding slightly mainly because the training samples from CNA are seriously imbalance. A finer sifted corpus will be introduced in the future work to improve this work.

In Chinese experiments, the Simple Embedding with Word Pairs Fuzzy Matching gains significant improvement compared with the One-Hot paradigm, which means that the sparseness of word pairs is alleviated effectively.

5. Conclusion

In this paper we report our English and Chinese discourse relation classification systems which handle explicit and non-explicit relations separately. It is showed that the discourse devices usages and the patterns of the discourse organization are quite different from Chinese to English. Adaptations are required to access better performance when transfer typical methods designed for English to Chinese texts.

Implicit relation disambiguation is still the most challenge task in discourse analysis. Distributed representation is an effective manner to release the data sparseness and explores relatively deep semantics. However, delicate semantic models such as structural semantic models are still remain to be explored to capture the real deep semantics of the texts for more meaningful conclusions.

Acknowledgement

The authors would like to thank the organizers of CoNLL-2016 and the reviewers for their helpful suggestions. This research is supported by the National Natural Science Foundation of China (61202244, 61502259) and the National Basic Re-search Program of China (973 Program, 2013CB329303).

References

Lynn Carlson, Daniel Marcu, and Mary E. Okurowski. 2003. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In *Current Directions in Discourse and Dialogue*, pages 85-112.

Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2005. Attribution and the (non)-alignment of syntactic and discourse arguments of connectives. In *Proceedings of the ACL Workshop in Frontiers in Corpus Annotation*, pages 29-36.

Vanessa W. Feng and Graeme Hirst. 2012. Textlevel discourse parsing with rich linguistic features. In *Proceedings of ACL*, pages 60-68.

Robert Fisher and Reid Simmons. 2015. Spectral semi-supervised discourse relation classification. In *Proceedings of ACL-IJCNLP (Short Papers)*, pages 89-93.

Hugo Hernault, Danushka Bollegala, and Mitsuru Ishizuka. 2011. Semi-supervised discourse relation classification with structural learning. In *Proceedings of Computational Linguistics and Intelligent Text (CICLing)*, pages 340-352.

Yu Hong, Xiaopei Zhou, Tingting Che, Jianmin Yao, Guodong Zhou, and Qiaoming Zhu. 2012. Cross-Argument Inference for Implicit Discourse Relation Recognition. In *Proceedings of CIKM*, pages 295-304.

Man Lan, Yu Xu, and Zheng-Yu Niu. 2013. Leveraging synthetic discourse data via multi-task learning for implicit discourse relation recognition. In *Proceedings of ACL*, pages 476-485.

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

Ziheng Lin, Hwee T. Ng and Min-Yen Kan. 2014. A PDTB-Styled End-to-End Discourse Parser. *Natural Language Engineering*, 20(2): 151-184.

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

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, page 3111--3119.

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.

Emily Pitler, Annie Louis, and Ani Nenkova. 2009. Automatic sense prediction for implicit discourse relations in text. In *Proceedings of ACL-IJCNLP*, pages 683-691.

Emily Pitler, Mridhula Raghupathy, Hena Mehta, Ani Nenkova, Alan Lee, and Aravind Joshi. 2008. Easily identifiable discourse relations. In *Proceedings of Coling, Companion volume - Posters and Demonstrations*, pages 87-90.

Martin Potthast, Tim Gollub, Francisco Rangel, Paolo Rosso, Efstathios Stamatatos, and Benno Stein. 2014. Improving the Reproducibility of PAN's Shared Tasks: Plagiarism Detection, Author Identification, and Author Profiling. *The 5th International Conference of the CLEF Initiative*, pages 268–299.

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 *Proceedings of LREC*, pages 2961-2968.

Rashmi Prasad, Aravind K. Joshi, and Bonnie L. Webber. 2010. Realization of discourse relations by other means: Alternative lexicalizations. In *Proceedings of Coling*, pages 1023-1031.

Attapol T Rutherford and Nianwen Xue. 2014. Discovering implicit discourse relations through brown cluster pair representation and coreference patterns. In *Proceedings of EACL*, pages 645-654.

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

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of HLT-EMNLP*, pages 347-354.

Nianwen Xue, Hwee T. Ng, Sameer Pradhan, Rashmi Prasad, Christopher Bryant, and Attapol Rutherford. 2015. The conll-2015 shared task on shallow discourse parsing. In *Proceedings of CoNLL Shared Task*, pages 1-15.

Nianwen Xue, Hwee T. Ng, Sameer Pradhan, Attapol T. Rutherford, Bonnie Webber, Chuan Wang, and Hongmin Wang. 2016. The CoNLL-2016 shared task on multilingual shallow discourse parsing. In *Proceedings of CoNLL Shared Task*.

Yuping Zhou and Nianwen Xue. 2012. PDTB-style discourse annotation of Chinese text. In *Proceedings of ACL*, pages 69-77.

Zhi-Min Zhou, Yu Xu, Zheng-Yu Niu, Man Lan, Jian Su, and Chew-Lim Tan. 2010. Predicting discourse connectives for implicit discourse relation recognition. In *Proceedings of Coling*, pages 1507-1514.