SoNLP-DP System for ConLL-2016 English Shallow Discourse Parsing

Fang Kong[1] **Sheng Li**[2] **Junhui Li**[1] **Muhua Zhu**[2] **Guodong Zhou**[1]
[1]Natural Language Processing Lab, Soochow University, China
{kongfang, lijunhui, gdzhou}@suda.edu.cn
[2] Alibaba Inc., Hangzhou, China
{lisheng.ls, muhua.zmh}@alibaba-inc.com

Abstract

This paper describes the submitted English shallow discourse parsing system from the natural language processing (NLP) group of Soochow university (SoNLP-DP) to the CoNLL-2016 shared task. Our System classifies discourse relations into explicit and non-explicit relations and uses a pipeline platform to conduct every subtask to form an end-to-end shallow discourse parser in the Penn Discourse Treebank (PDTB). Our system is evaluated on the CoNLL-2016 Shared Task closed track and achieves the 24.31% and 28.78% in F1-measure on the official blind test set and test set, respectively.

1 Introduction

Discourse parsing determines the internal structure of a text via identifying the discourse relations between its text units and plays an important role in natural language understanding that benefits a wide range of downstream natural language applications, such as coherence modeling (Barzilay and Lapata, 2005; Lin et al., 2011), text summarization (Lin et al., 2012), and statistical machine translation (Meyer and Webber, 2013).

As the largest discourse corpus, the Penn Discourse TreeBank (PDTB) corpus (Prasad et al., 2008) adds a layer of discourse annotations on the top of the Penn TreeBank (PTB) corpus (Marcus et al., 1993) and has been attracting more and more attention recently (Elwell and Baldridge, 2008; Pitler and Nenkova, 2009; Prasad et al., 2010; Ghosh et al., 2011; Kong et al., 2014; Lin et al., 2014). Different from another famous discourse corpus, the Rhetorical Structure Theory(RST) Treebank corpus(Carlson et al., 2001), the PDTB focuses on shallow discourse relations either lexically grounded in explicit discourse connectives or associated with sentential adjacency. This theory-neutral way makes no commitment to any kind of higher-level discourse structure and can work jointly with high-level topic and functional structuring (Webber et al., 2012) or hierarchial structuring (Asher and Lascarides, 2003).

Although much research work has been conducted for certain subtasks since the release of the PDTB corpus, there is still little work on constructing an end-to-end shallow discourse parser. The CoNLL 2016 shared task evaluates end-to-end shallow discourse parsing systems for determining and classifying both explicit and non-explicit discourse relations. A participant system needs to (1)locate all explicit (e.g., "because", "however", "and".) discourse connectives 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").

In this paper, we describe the system submission from the NLP group of Soochow university (SoNLP-DP). Our shallow discourse parser consists of multiple components in a pipeline architecture, including a connective classifier, argument labeler, explicit classifier, non-explicit classifier. Our system is evaluated on the CoNLL-2016 Shared Task closed track and achieves the 24.31% and 28.78% in F1-measure on the official blind test set and test set, respectively.

The remainder of this paper is organized as follows. Section 2 presents our shallow discourse parsing system. The experimental results are described in Section 3. Section 4 concludes the paper.

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 65–69,
Berlin, Germany, August 7-12, 2016.

2 System Architecture

In this section, after a quick overview of our system, we describe the details involved in implementing the end-to-end shallow discourse parser.

2.1 System Overview

A typical text consists of sentences glued together in a systematic way to form a coherent discourse. Referring to the PDTB, shallow discourse parsing focus on shallow discourse relations either lexically grounded in explicit discourse connectives or associated with sentential adjacency. Different from full discourse parsing, shallow discourse parsing transforms a piece of text into a set of discourse relations between two adjacent or non-adjacent discourse units, instead of connecting the relations hierarchically to one another to form a connected structure in the form of tree or graph.

Specifically, given a piece of text, the end-to-end shallow discourse parser returns a set of discourse relations in the form of a discourse connective (explicit or implicit) taking two arguments (clauses or sentences) with a discourse sense. That is, a complete end-to-end shallow discourse parser includes:

- connective identification, which identifies all connective candidates and labels them as whether they function as discourse connectives or not,
- argument labeling, which identifies the spans of text that serve as the two arguments for each discourse connective,
- explicit sense classification, which predicts the sense of the explicit discourse relations after achieving the connective and its arguments,
- non-explicit sense classification, for all adjacent sentence pairs within each paragraph without explicit discourse relations, which classify the given pair into EntRel, NoRel, or one of the Implicit/AltLex relation senses.

Figure 1 shows the components and the relations among them. Different from traditional approach (i.e., Lin et al. (2014)), considering the interaction between argument labeler and explicit sense classifier, co-occurrence relation between explicit and non-explicit discourse relations in a text, our system does not employ complete sequential pipeline framework.

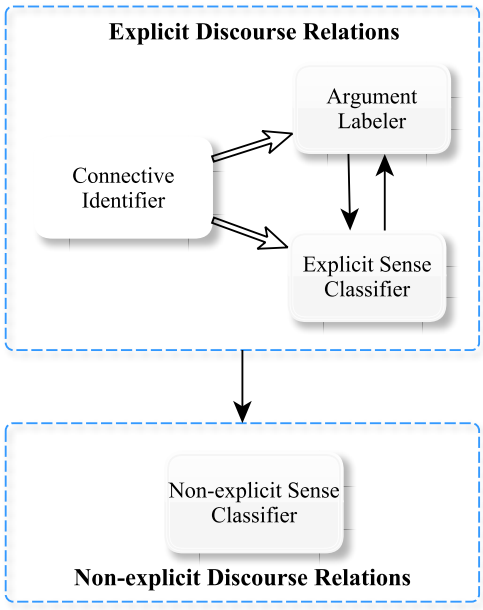

Figure 1: Framework of our end-to-end shallow discourse parser

2.2 Connective Identification

Our connective identifier works in two steps. First, the connective candidates are extracted from the given text referring to the PDTB. There are 100 types of discourse connectives defined in the PDTB. Then every connective candidate is checked whether it functions as a discourse connective.

Pitler and Nenkova (2009) showed that syntactic features extracted from constituent parse trees are very useful in disambiguating discourse connectives. Followed their work, Lin et al. (2014) found that a connective's context and part-of-speech (POS) are also helpful. Motivated by their work, we get a set of effective features, includes:

- Lexical: connective itself, POS of the connective, connective with its previous word, connective with its next word, the location of the connective in the sentence, i.e., start, middle and end of the sentence.
- Syntactic: the highest node in the parse tree that covers only the connective words (dominate node), the context of the dominate node [1], whether the right sibling contains a VP, the path from the parent node of the connective to the root of the parse tree.

[1] We use POS combination of the parent, left sibling and right sibling of the dominate node to represent the context. When no parent or siblings, it is marked NULL.

2.3 Argument Labeling

Argument labeler need to label the Arg1 and Arg2 spans for every connective determined by connective identifier. Following the work of Kong et al. (2014), we employ the constituent-based approach to argument labeling by first extracting the constituents from a parse tree are casted as argument candidates, then determining the role of every constituent as part of Arg1, Arg2, or NULL, and finally, merging all the constituents for Arg1 and Arg2 to obtain the Arg1 and Arg2 text spans respectively. Note that, we do not use ILP approach to do joint inference.

After extracting the argument candidates, a multi-category classifier is employed to determine the role of every argument candidate (i.e., Arg1, Arg2, or NULL) with features reflecting the properties of the connective, the candidate constituent and relationship between them. Features include,

- Connective related features: connective itself, its syntactic category, its sense class[2]
- Number of left/right siblings of the connective.
- The context of the constituent. We use POS combination of the constituent, its parent, left sibling and right sibling to represent the context. When there is no parent or siblings, it is marked NULL.
- The path from the parent node of the connective to the node of the constituent.
- The position of the constituent relative to the connective: left, right, or previous.

2.4 Explicit sense classification

After a discourse connective and its two arguments are identified, the sense classifier is proved to decide the sense that the relation conveys.

Although the same connective may carry different semantics under different contexts, only a few connectives are ambiguous (Pitler and Nenkova, 2009). Following the work of Lin et al. (2014), we introduce four features to train a sense classifier: the connective itself, its lower format, its POS and the combination of the previous word and the connective.

[2]In training stage, we extract the gold sense class from the annotated corpus. And in testing stage, the sense classification will be employed to get the automatic sense.

2.5 Non-explicit sense Classification

Referring to the PDTB, the non-explicit relations[3] are annotated for all adjacent sentence pairs within paragraphs. So non-explicit sense classification only considers the sense of every adjacent sentence pair within a paragraph without explicit discourse relations.

Our non-explicit sense classifier includes five traditional features:

Production rules: According to Lin et al. (2009), the syntactic structure of one argument may constrain the relation type and the syntactic structure of the other argument. Three features are introduced to denote the presence of syntactic productions in Arg1, Arg2 or both. Here, these production rules are extracted from the training data and the rules with frequency less than 5 are ignored.

Dependency rules: Similar with Production rules, three features denoting the presence of dependency productions in Arg1, Arg2 or both are also introduced in our system.

Fisrt/Last and First 3 words: This set of features include the first and last words of Arg1, the first and last words of Arg2, the pair of the first words of Arg1 and Arg2, the pair of the last words as features, and the first three words of each argument.

Word pairs: We include the Cartesian product of words in Arg1 and Arg2. We apply MI (Mutual Information) method to select top 500 word pairs.

Brown cluster pairs: We include the Cartesian product of the Brown cluster values of the words in Arg1 and Arg2. In our system, we take 3200 Brown clusters provided by CoNLL shared task.

Besides, we notice that not all adjacent sentences contain relation between them. Therfore, we view these adjacent sentences as NoRel relations like the PDTB.

3 Experimentation

We train our system on the corpora provided in the CoNLL-2016 Shared Task and evaluate our system on the CoNLL-2016 Shared Task closed track. All our classifiers are trained using the OpenNLP maximum entropy package[4] with the default pa-

[3]The PDTB provides annotation for Implicit relations, AltLex relations, entity transition (EntRel), and otherwise no relation (NoRel), which are lumped together as Non-Explicit relations.

[4]http://maxent.sourceforge.net/

rameters (i.e. without smoothing and with 100 iterations). We firstly report the official score on the CoNLL-2016 shared task on development, test and blind test sets. Then, the supplementary results provided by the shared task organizes are reported.

	Arg1&2	Conn	Parser
Dev	47.87	94.22	35.56
Test	41.68	94.71	28.78
Blind	36.19	91.62	**24.31**
Blind (Wang and Lan, 2015)	46.37	91.86	24.00

Table 1: the official F1 score of our system.

In Table 1, we present the official results of our system performances on the CoNLL-2016 development, test and blind test sets, respectively. In the blind test, our parser achieve a better result than the best system of last year (Wang and Lan, 2015).

		Arg1&2	Conn	Parser
Dev	Exp	46.37	94.22	42.97
	Non-Exp	49.51	-	27.54
Test	Exp	40.81	94.71	36.57
	Non-Exp	42.68	-	19.82
Blind	Exp	38.25	91.62	31.18
	Non-Exp	33.73	-	16.10

Table 2: the supplementary F1 score of our system.

In Table 2, we reported the supplementary results provided by the shared task organizes on the development, test and blind test sets. These additional experiments investigate the performance of our shallow discourse parsing for explicit and non-explicit relations separately. From the results, we can find that the sense classification for both explicit and non-explicit discourse relations are the biggest obstacles to the overall performance of discourse parsing.

Further, we reports all the official performance in Table 3 on the development, test and blind test set in detail. From the table, we observe:

- For argument recognition of explicit discourse relations, the performance of Arg2 is much better than that of Arg1 on all the three datasets. So the performance of *Arg1 & Arg2* recognition mainly depends on the performance of Arg1 recognition. With respect to non-explicit discourse relations, the performance gap of argument recognition on Arg1 and Arg2 is very small.
- With respect to explicit discourse relations, the sense classification works almost perfectly on development data. It also works well on the test and blind test sets. With respect to non-explicit discourse relations, the sense classification works much worse than that of explicit sense classification. The performance gap caused by non-explicit sense classification reaches 15% 16%.

4 Conclusion

We have presented the SoNLP-DP system from the NLP group of Soochow university that participated in the CoNLL-2016 shared task. Our system is evaluated on the CoNLL-2016 Shared Task closed track and achieves the 24.31% and 28.78% in F1-measure on the official blind test set and test set, respectively.

Acknowledgements

This research is supported by Key project 61333018 and 61331011 under the National Natural Science Foundation of China, Project 61472264, 61401295, 61305088 and 61402314 under the National Natural Science Foundation of China.

References

Nicholas Asher and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge Unversity Press.

Regina Barzilay and Mirella Lapata. 2005. Modeling local coherence: An entity-based approach. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2001. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Proceedings of Second SIGdial Workshop on Discourse and Dialogue*.

Robert Elwell and Jason Baldridge. 2008. Discourse connective argument identification with connective specific rankers. In *Second IEEE International Conference on Semantic Computing*.

Sucheta Ghosh, Richard Johansson, Giuseppe Riccardi, and Sara Tonelli. 2011. Shallow discourse parsing with conditional random fields. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*.

Fang Kong, Hwee Tou Ng, and Guodong Zhou. 2014. A constituent-based approach to argument labeling

		Dev			Test			Blind test		
		P	R	F1	P	R	F1	P	R	F1
Explicit	Connective	93.53	94.93	94.22	94.04	95.38	94.71	90.47	92.80	91.62
	Arg1	52.50	53.28	52.89	46.80	47.47	47.14	46.58	47.79	47.18
	Arg2	74.26	75.37	74.81	70.64	71.65	71.14	67.99	69.74	68.85
	Arg1 & Arg2	46.03	46.72	46.37	40.52	41.10	40.81	37.77	38.75	38.25
	Overall	42.84	43.10	42.97	36.81	36.33	36.57	31.61	30.76	31.18
Non-Explicit	Connective	-	-	-	-	-	-	-	-	-
	Arg1	47.35	75.85	58.31	40.55	70.79	51.56	29.71	72.93	42.22
	Arg2	48.54	77.75	59.77	40.55	70.79	51.56	31.55	77.44	44.83
	Arg1 & Arg2	40.21	64.41	49.51	33.56	58.59	42.68	23.74	58.27	33.73
	Overall	35.67	22.43	27.54	27.19	15.60	19.82	27.82	11.33	16.10
All	Connective	93.53	94.93	94.22	94.04	95.38	94.71	90.47	92.80	91.62
	Arg1	50.35	63.31	56.09	43.89	57.04	49.61	37.55	56.19	45.02
	Arg2	60.72	76.36	67.65	54.87	71.31	62.02	48.30	72.28	57.91
	Arg1 & Arg2	42.97	54.03	47.87	36.87	47.92	41.68	30.19	45.17	36.19
	Overall	39.86	32.10	35.56	33.07	25.48	28.78	30.36	20.26	24.31

Table 3: Official results (%) of our parser on development, test and blind test sets. Group *Explicit* indicates the performance with respect to explicit discourse relations; group *Non-Explicit* indicates the performance with respect to non-explicit discourse relations, and group *all* indicates the performance with respect to all discourse relations, including both explicit and non-explicit ones.

with joint inference in discourse parsing. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*.

Ziheng Lin, Min-Yen Kan, and Hwee Tou Ng. 2009. Recognizing implicit discourse relations in the Penn Discourse Treebank. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2011. Automatically evaluating text coherence using discourse relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*.

Ziheng Lin, Chang Liu, Hwee Tou Ng, and Min-Yen Kan. 2012. Combining coherence models and machine translation evaluation metrics for summarization evaluation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2014. A PDTB-styled end-to-end discourse parser. *Natural Language Engineering*, 20(2):151–184.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Thomas Meyer and Bonnie Webber. 2013. Implicitation of discourse connectives in (machine) translation. In *Proceedings of the Workshop on Discourse in Machine Translation*.

Emily Pitler and Ani Nenkova. 2009. Using syntax to disambiguate explicit discourse connectives in text. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the LREC 2008 Conference*.

Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2010. Exploiting scope for shallow discourse parsing. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*.

Jianxiang Wang and Man Lan. 2015. A refined end-to-end discourse parser. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning - Shared Task*, pages 17–24, Beijing, China, July. Association for Computational Linguistics.

Bonnie Webber, Marcus Egg, and Valia Kordoni. 2012. Discourse structure and language technology. *Natural Language Engineering*, 18(4):437–490, 10.