A Shallow Discourse Parsing System Based On Maximum Entropy Model

Jia Sun, Peijia Li, Weiqun Xu, Yonghong Yan
The Key Laboratory of Speech Acoustics and Content Understanding
Institute of Acoustics, Chinese Academy of Sciences
No. 21 North 4th Ring West Road, Haidian District, 100190 Beijing, China
{sunjia,lipeijia,xuweiqun,yanyonghong}@hccl.ioa.ac.cn

Abstract

This paper describes our system for Shallow Discourse Parsing - the CoNLL 2015 Shared Task. We regard this as a classification task and build a cascaded system based on Maximum Entropy to identify the discourse connective, the spans of two arguments and the sense of the discourse connective. We trained the cascaded models with a variety of features such as lexical and syntactic features. We also report the results achieved by our team.

1 Introduction

Discourse parsing is one of the most challenging tasks in natural language processing (NLP) field. It focuses on parsing the structure of a piece of text into a set of discourse relations between inter sentences. There is considerable interest in discourse parsing, both as an end in itself and as an intermediate step in a variety of NLP applications like question answering (Verberne et al., 2007), text summarization (Louis et al., 2010), sentiment analysis and opinion mining (Somasundaran, 2010).

There are many approaches working on identifying the discourse relations and data-driven approaches are dominated. A number of pioneers take the discourse relations identification as a classification task (Marcu and Echihabi, 2002; Pitler et al., 2009; Duverle and Prendinger, 2009) by the construction of features like lexical, syntactic and constituent features. Some take the argument segmentation task as a semantic role detection task (Wellner and Pustejovsky, 2007) and a sequence labeling task (Ghosh et al., 2011). However, some of the previous research is based on different corpus, lacking an common evaluation data set. This has been addressed with the release of Penn Discourse Treebank (PDTB) 2.0 corpus (Prasad et al.,

2008) which provides detailed annotations about the discourse relations and argument spans addresses this problem. Besides, much research about discourse parsing working on the PDTB appears (Prasad et al., 2010; Lin et al., 2009) and they put more attention on the "harder" part - labelling the arguments. Lin (Lin et al., 2014) designed an end-to-end discourse parser with the PDTB including the explicit, implicit sense and the argument spans identification.

Shallow Discourse Parsing (Xue et al., 2015) is the CoNLL shared task this year[1] which takes a piece of newswire text as input and returns all the discourse relations in the form of a discourse connective (explicit or implicit) taking two arguments (which can be clauses, sentences, or multi-sentence segments) in JSON format. A relation will be parsed as correct if the explicit discourse connective (e.g., "because", "however") once it has, the spans of text that serve as the two arguments for each discourse connective and the sense (e.g., "Comparison") are all correct. The F1 score of the parser's performance is the evaluation metric.

In this paper, we describe our system details in Section 2, the evaluation result and subsequent experiments in Section 3. Finally, we draw some conclusions in Section 4.

2 Our System

2.1 Resources

The resources used in our system are as follows:
Labeled training and development data: The training and development (dev) data is derived from the PDTB 2.0 Section 2-21 and Section 22 in JSON format . There are 32535 relations and 1436 relations annotated in the training data and the dev data respectively. Table 1 shows the distribution of the four types in the data. There are

---

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

Type	Train data	Dev data
Explicit	14,727	680
Implicit	13,163	522
EntRel	4,133	215
AltLex	524	19
all	32,535	1,436

Table 1: Distribution of the four discourse relation types in the data sets.

Sense level 1	Sense level 2or3	Train	Dev
Temporal	A.P	1,277	78
	A.S	1,014	55
	Synchrony	499	100
Contingency	Cause.Reason	3,344	147
	Cause.Result	2,137	81
	Condition	1,197	52
Comparison	Contrast	4,714	257
	Concession	1,293	17
Expansion	Conjunction	7,817	310
	Instantiation	1,403	58
	Restatement	2,699	110
	Alternative	210	6
	A.C	241	6
	Exception	15	0
	EntRel	4,133	215

Table 2: Distribution of the 15 senses from the different data sets. A.P, A.S, A.C are the abbreviations of "Asynchronous.Precedence", "Asynchronous.Succession", "Alternative.Chosen alternative" respectively .

15 valid senses including the second-level "types" as well as a selected number of third-level "subtypes". Table 2 shows the distribution of the 15 senses in the data.

Test data: There are two test data sets. One is the blind set which contains 20,000 to 30,000 words of newswire text annotated following the PDTB annotation guidelines. The other test set is Section 23 of the PDTB which is used for comparison with previous work.

The connectives list: A list contains 100 discourse connectives in the PDTB and three syntactic categories form (Knott, 1996).

Opennlp-maxent: We used the open source package Opennlp-maxent[2] to construct the classification models.
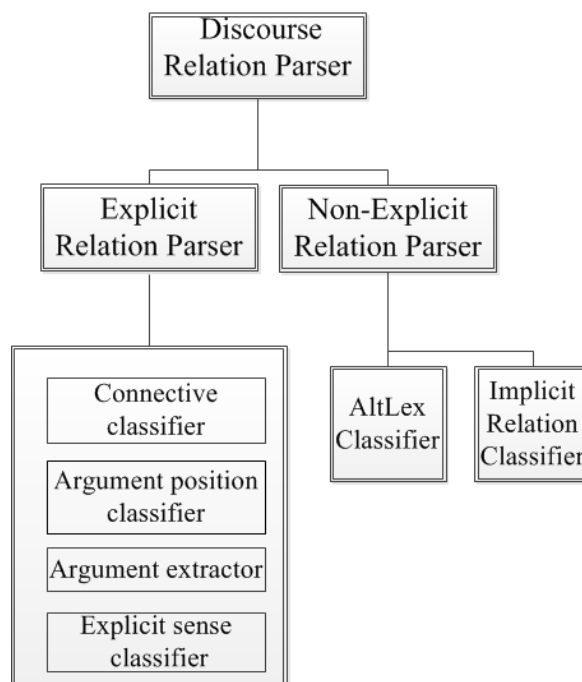
Figure 1: The structure of the system.

2.2 System overview and Features

Our system mainly follows the work of (Lin et al., 2014), which consists of two parts: the explicit relation parser and the non-explicit relation parser. The explicit relation parser is composed of the connective classifier, the argument position classifier, the argument extractor and the explicit sense classifier while the non-Explicit relation parser contains the AltLex classifier and the implicit classifier. The structure of our system is shown in Figure 1.

The set of features used in our system are listed in Table 3. All the features fall into four classes: lexical features, part-of-speech (POS) features, syntactic features and positional features.

- **Lexical features:** The lexical features (F1-F10) contain the connectives C, their contextual words and word-pair features (i.e., F7 (w_i, w_j) where w_i is a word from Arg1 and w_j is a word from Arg2) .

- **POS features:** F11-F17 belong to the POS features .

- **Syntactic features:** The syntactic features (F18-F26) include the connectives' syntactic category (F18): subordinating, coordinating, or discourse adverbial, the path of syntactic trees (F19, F20, F23), the number of siblings

Feature	Description
F1.	C string
F2.	the first word before C
F3.	the second word before C
F4.	F2 + C
F5.	F3 + C
F6.	C + the next word after C
F7.	word-pair
F8.	the first word of Arg2
F9.	the second word of Arg2
F10.	the third word of Arg2
F11.	the POS of C
F12.	the POS of F2
F13.	F11 + F12
F14.	the POS of the word after C
F15.	F11 + F14
F16.	the POS of F3
F17.	F11 + F16
F18.	the syntactic category of C
F19.	the path of C's parent to root
F20.	the compressed path of F19
F21.	the number of left siblings of C
F22.	the number of right siblings of C
F23.	the path of C's parent to N
F24.	whether the C's left sibling number is greater than 1
F25.	the constituent rules
F26.	the dependency rules
F27.	the relative position of N to C
F28.	the position of C in the sentence

Table 3: The features used in our system. "C" denotes the connectives. N means a current node in the constituent tree used in Section 2.3.2.

(F21, F22, F24), constituent rules (F25) and dependency rules (F26).

- **Position features:** F27 is the relative position in the syntactic tree structure (left, middle or right), while F28 is the connectives' positions in the sentence (start, middle or end).

2.3 Training

2.3.1 The Connective Classifier

All the 100 connectives that appeared in one discourse were extracted whether it functioned as a connective or not. We converted all upper case letters in connective to lower case ones.

The connective classifier decides whether a connective is functioned as a discourse connective.

The features used were F4, F6, F11-F15, F19-F20 in Table 3.

2.3.2 The Argument Labeller

Once the connective is identified, the argument labeller identifies the Arg1 and Arg2 spans of this instance. This is accomplished in two steps: (1) Classifying the locations of Arg1 by the Argument Position Classifier. (2) Labelling the spans of Arg1 and Arg2 by the Argument Extrator.

The Argument Position Classifier: Normally Arg2 immediately follows the connective while the position of Arg1 is uncertain. In this model, we classified the Arg1's locations into two classes: Arg1 was located within the same sentence of the connective (SS) or in the previous sentence of connective (PS) (Prasad et al., 2008).

We implemented this as a binary classification task. In this step, features F1-F5, F11, F13, F16-F17, F28 in Table 3 were adopted to train the model. After the position label of Arg1 was determined, the result was passed to the argument extractor.

The Argument Extractor: In this module, our classifier labelled the previous sentence as Arg1 immediately for the PS case. The argument spans for the SS case were extracted described as below.

- Classify each internal node N in the constituent tree as Arg1-node, Arg2-node, or None with features F1, F18, F21-F24, F27 in Table 3.

- Label a node as Arg1-node once its Arg1-node predicted probability is greater than 0.1 (which is tuned on the dev data set).

- Select only one Arg1-node and one Arg2-node in one instance with the maximal probability of the respective label.

- Extract the Arg1 and Arg2 spans by tree subtraction. If the Arg1 node is the ancestor of the Arg2 node, the span of Arg1 should be subtracted from the Arg2 span, and vice versa.

- Remove punctuation tokens and connectives out of the exact argument spans.

2.3.3 The Explicit Sense Classifier

After recognizing the discourse connective and its two arguments spans, the next step is to decide the

Data	Connective	Span	Sense	Parser
Dev	0.9152	0.2668	0.1367	0.1814
Test	0.9064	0.2336	0.1191	0.1505
Blind	0.826	0.2195	0.1232	0.1262

Table 4: The results F1 score obtained by our team

N	Non-Explicit	Overall
50	0.0683	0.1876
100	0.0683	0.1876
200	0.0614	0.1870
300	0.0603	0.1870
Baseline	0.0435	0.1814

Table 5: The F1 score in non-explict and overall parser when selecting features F7 using different N on dev data set.

sense of the connective. We trained this model using features F1, F4, F11 in Table 3. We picked the output whose maximal sense probability is greater than 0.45 which was experientially determined on dev data set.

2.3.4 The AltLex classifier

We extracted all adjacent sentence pairs within each paragraph and removed the pairs that were identified by the explicit relation parser. Then we trained the AltLex Classifier which decided whether the pairs were AltLex pairs and classified the senses with features F8-F10 in Table 3. The pairs labelled as non-AltLex relations were passed to the next implicit relation classifier.

2.3.5 The Implicit relation classifier

The implicit relation classifier classified the sense of each pair into one of the 15 valid senses or NoRel with F7, F25-F26 in Table 3. After predicting, we kept the implicit discourse relations whose maximal sense probability were greater than a threshold (0.25 in our case) which was determined on the dev data set .

3 Experiments and Results

There are two test data sets this year as described in Section 2.1 and the organizers reported the results on the two test data sets and the dev data set. The results of our system obtained are shown in Table 4. We ranked the 10th on every data set.

After the deadline of evaluation, we made some improvements in the module of implicit relation classifier inspired by (Lin et al., 2009). We selected the word-pair features (F7) while the experiments showed a little degradation in F1 score through selecting the constituent rules and the dependency rules (F25, F26) on the dev data set.

We computed the mutual information between each word-pair feature and the 15 valid senses and then selected the top N as the features. Table 5 shows the improvement of different N.

4 Conclusion

We divided the complex task of discourse parsing into a set of classification subtasks and glued them together. A variety of features, including lexical, part-of-speech, syntactic and positional feature were employed to train the baseline with open Maximum Entropy package, then the system was improved by setting probability-output threshold. We did not utilize any additional resources and only used the annotations the official provided. Our system ranked the 10th among seventeenth teams on the two test data sets.

References

David A Duverle and Helmut Prendinger. 2009. A novel discourse parser based on support vector machine classification. 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 665–673. Association for Computational Linguistics.

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

Alistair Knott. 1996. A data-driven methodology for motivating a set of coherence relations.

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: Volume 1-Volume 1*, pages 343–351. 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.

Annie Louis, Aravind Joshi, and Ani Nenkova. 2010. Discourse indicators for content selection in summarization. In *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 147–156. Association for Computational Linguistics.

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

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 *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May - 1 June 2008, Marrakech, Morocco*.

Rashmi Prasad, Aravind K. Joshi, and Bonnie L. Webber. 2010. Exploiting scope for shallow discourse parsing. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta*.

Swapna Somasundaran. 2010. *Discourse-level relations for Opinion Analysis*. Ph.D. thesis, University of Pittsburgh.

Suzan Verberne, Lou Boves, Nelleke Oostdijk, and Peter-Arno Coppen. 2007. Evaluating discourse-based answer extraction for why-question answering. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 735–736. ACM.

Ben Wellner and James Pustejovsky. 2007. Automatically identifying the arguments of discourse connectives. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Coumptational Natural Language Learning (EMNLP- CoNLL)*, pages 92–101.

Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Rashmi Prasad, Christopher Bryant, and Attapol Rutherford. 2015. The CoNLL-2015 shared task on shallow discourse parsing. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*.