Research on Discourse Parsing: from the Dependency View

Sujian li and **Liang Wang** and **An Yang** and **Yi Cheng** and **Zhenwen Li**
Key Laboratory of Computational Linguistics(MOE)
Department of Computer Science, Peking University
lisujian@pku.edu.cn

Discourse parsing aims to comprehensively acquire the logical structure of the whole text which may be helpful to some downstream applications such as summarization, reading comprehension, QA and so on. One important issue behind discourse parsing is the representation of discourse structure. Up to now, many discourse structures have been proposed (Mann and Thompson, 1987; Lascarides and Asher, 2008; Prasad et al., 2008), and the correponding parsing methods are designed (Soricut and Marcu, 2003; Joty et al., 2012; Feng and Hirst, 2012; Hernault et al., 2010; Zhou et al., 2010; Wang et al., 2012; Lan et al., 2013; Liu and Li, 2016), promoting the development of discourse research. In this paper, we mainly introduce our recent discourse research and its preliminary application from the dependency view.

First, as about discourse structure, we present why we choose to use the dependency strucutre. So far, there are two well known discourse representations which are widely researched in the field of natural language processing. One is PDTB and the other is RST. PDTB adopts the representation of one predicate and two arguments by taking an implicit or explicit connective as a predicate of two sentences. In PDTB, usually two adjacent sentences are selected and independently analyzed their logical relations which exhibit a flat and shallow discourse structure without knowing a wider context. RST posits a hierarchical tree structure. In a RST tree for a text, the leaves correspond to contiguous text spans called Elementary Discourse Units (EDUs). The adjacent EDUs are combined into a larger text span by rhetorical relations until the whole text constitutes a tree. This kind of tree exhibits a relatively global and deep discourse structure, and the corresponding parsing task is more challenging. With such a generative tree structure for a text, we have two problems. On one hand, it is difficult to generalize the meaning of interior text spans and design a set of production rules as in syntactic parsing, as there are no determinate generative rules for the interior text spans. On the other hand, it is not easy to keep the consistency of relations at different levels. For example, the relation "Expansion" may occur between two EDUs or between two paragraphs.

To solve these problems, we propose to use the discourse dependency structure which only consider the relations between EDUs (Li et al., 2014a). Then we can analyze the relations between EDUs directly, without worrying about any interior text spans. Without interior nodes, Dependency trees contain much fewer nodes and on average their annotation is simpler than RST trees. In addition, dependency structures can deal with non-projective relations, while constituency-based models need the addition of complex mechanisms like transformations, movements and so on. For a discourse dependency tree, it consists of EDUs which are linked by the binary, asymmetrical relations called dependency relations. A dependency relation holds between a subordinate EDU called the dependent, and another EDU on which it depends called the head. Each EDU has one and only one head. Thus, the dependency structure can be seen as a set of head-dependent links, labeled by functional relations.

The next problem is how to get a discourse dependency corpus. We adopt two kinds of methods. The first conversion method is simple and straightforward (Li et al., 2014a). We directly convert RST-DT into a discourse dependency corpus. In RST-DT, there are a total of 110 fine-grained relations which are categorized into 18 classes. One kind of relations is mononuclear and contain a nucleus and a satellite span. The kind of relations is multinuclear and contain two or more equally important nucleus spans. We recursively convert the n-ary RST trees to binary trees through adding

Proceedings of the 2nd International Workshop on Discourse Processing, pages 1–3

a new node for the latter n-1 nodes. Then we convert the binarized RST trees to dependency trees by pointing from a nucleus EDU to a satellite EDU. Through conversion, there may exist some conversion errors. In such cases, we hope to manually annotate a dependency corpus from scratch. Compared with conversion method, manual annotation is very costly. We also hope to construct a high quality and cost-effective corpus. Here we choose scientific abstracts as raw text, as scientific abstracts are usually composed of one passage with strong logics. 5 annotators are recruited after a test annotation, the annotation process lasts about 6 months, and the corpus SciDTB is finally constructed (Yang and Li, 2018). There are 17 coarse-grained relations and 26 fine-grained relations. SciDTB contains 798 unique abstracts and 18,978 discoure relations. 3% of all relations are non-projective.

Further, we hope to construct a Chinese discourse dependency corpus with the help of the English discourse corpus or other Chinese discourse corpus available. For the first attempt, we design one simple and efficient method to conduct zero-shot Chinese text-level dependency parsing through leveraging English discourse data and parsing techniques (Cheng and Li, 2019). This is motivated by the observation that the logical organization of a text is similar at the macro discourse level regardless of languages, in spite of some lexical or grammatical differences. Based on the observation, we conduct the Chinese-English mapping from the sentence and elementary discourse unit (EDU) levels using the machine translation techniques, and then return the parsing results of the corresponding English translations as the discourse structure of the Chinese text. This method can automatically conduct Chinese discourse parsing, with no need of a large scale of Chinese labeled data.

We also explore another possible way to integrate different Chinese discourse corpora available under the same dependency framework to form a much larger discourse treebank. Here three Chinese discourse corpora, HIT-CDTB (Zhang et al., 2014), CDTB (Li et al., 2014b) and SciCDTB (Cheng and Li, 2019), are chosen. HIT-CDTB adopts the predicate-argument structure similar to PDTB, with a connective as predicate and two text spans as arguments. Following the rhetorical structure theory(RST), CDTB use a hierarchical tree to represent the inner structure of each text, with EDUs as its leaves and connectives as intermediate nodes. SciCDTB is a small-scale DDS corpus composed of 108 scientific abstracts. The primary obstacle of unifying these corpora is inconsistency of the representation schemes, such as granularity of EDU and definition of relation types. Besides, the predicate-argument structure of HIT-CDTB leads to the problem that some discourse relations between adjacent text spans are absent. To tackle the problems, we redefine granularity of EDU, conduct mapping among different relation sets, and design semi-automatic methods to convert other discourse structures into DDS. On the unified dataset, we also implement several discourse dependency parsers and explore how the data can be leveraged to improve parsing performance.

Finally, our discourse research aims to improve some text applications and we conduct some preliminary research on summarization (Li et al., 2020). We chose to use Elementary Discourse Unit (EDU) as the summarization unit, which is first proposed from Rhetorical Structure Theory (**?**) and defined as a clause. The finer granularity makes EDU more suitable than sentence to be the basic summary composition unit (Li et al., 2016). At the same time, benefited from the development of EDU segmentation techniques, which can achieve a high accuracy of 94% (Wang et al., 2018), it is feasible to automatically obtain EDUs from the text. Next, to well handle the problem of composing EDUs into an informative and fluent summary, we propose a summarization method *EDUSum* that first designs an EDU selection model to extract and group informative EDUs and an EDU fusion model to fuse the EDUs in each group into one sentence. We also design the reinforcement learning mechanism to use EDU fusion results to reward the EDU selection action, boosting the final summarization performance. We applied *EDUSum* on *CNN/Daily Mail* and found that similar EDUs can be grouped to generate more informative summaries compared to using sentences as the basic selection unit. We will further seek new methods to exploit more discourse information including the dependency tree structure and relations into summarization.

In conclusion, we summarize some of our discourse research from the dependency view which may reduce the difficulty of discourse parsing. Based on our research experience, we found that both EDU segmentation and tree structure identification can reach a relatively satisfying performance.

However, discourse relation recognition is still far from satisfactory. In future work, we will focus on researching the identification of discourse relations and how to use discourse to improve more text applications.

Acknowledgments

This work was partially supported by National Key R&D Project (2019YFB1704002) and National Natural Science Foundation of China (61876009).

References

Yi Cheng and Sujian Li. 2019. Zero-shot chinese discourse dependency parsing via cross-lingual mapping. In *Proceedings of the 1st Workshop on Discourse Structure in Neural NLG*, pages 24–29, Tokyo, Japan. Association for Computational Linguistics.

Vanessa Wei Feng and Graeme Hirst. 2012. Text-level discourse parsing with rich linguistic features. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 60–68.

Hugo Hernault, Helmut Prendinger, David A du Verle, and Mitsuru Ishizuka. 2010. Hilda: A discourse parser using support vector machine classification. *Dialogue & Discourse*, 1(3):1–33.

Shafiq Joty, Giuseppe Carenini, and Raymond Ng. 2012. A novel discriminative framework for sentence-level discourse analysis. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 904–915.

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 the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 476–485.

Alex Lascarides and Nicholas Asher. 2008. Segmented discourse representation theory: Dynamic semantics with discourse structure. In *Computing meaning*, pages 87–124. Springer.

Junyi Jessy Li, Kapil Thadani, and Amanda Stent. 2016. The role of discourse units in near-extractive summarization. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 137–147, Los Angeles. Association for Computational Linguistics.

Sujian Li, Liang Wang, Ziqiang Cao, and Wenjie Li. 2014a. Text-level discourse dependency parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25–35.

Yancui Li, Wenhe Feng, Jing Sun, Fang Kong, and Guodong Zhou. 2014b. Building chinese discourse corpus with connective-driven dependency tree structure. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 2105–2114.

Zhenwen Li, Wenhao Wu, and Sujian Li. 2020. Composing elementary discourse units in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6191–6196.

Yang Liu and Sujian Li. 2016. Recognizing implicit discourse relations via repeated reading: Neural networks with multi-level attention. *arXiv preprint arXiv:1609.06380*.

William C Mann and Sandra A Thompson. 1987. *Rhetorical structure theory: A theory of text organization*. University of Southern California, Information Sciences Institute Los Angeles.

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.

Radu Soricut and Daniel Marcu. 2003. Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 228–235.

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.

Yizhong Wang, Sujian Li, and Jingfeng Yang. 2018. Toward fast and accurate neural discourse segmentation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 962–967, Brussels, Belgium. Association for Computational Linguistics.

An Yang and Sujian Li. 2018. Scidtb: Discourse dependency treebank for scientific abstracts. *arXiv preprint arXiv:1806.03653*.

Muyu Zhang, Bing Qin, and Ting Liu. 2014. Chinese discourse relation semantic taxonomy and annotation. *Journal of Chinese Information Processing*, 28:26–28.

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 the 23rd International Conference on Computational Linguistics: Posters*, pages 1507–1514. Association for Computational Linguistics.