Building a Macro Chinese Discourse Treebank

Xiaomin Chu, Feng Jiang, Sheng Xu, Qiaoming Zhu
Natural Language Processing Lab
Soochow University
{xmchu, fjiang, sxu}@stu.suda.edu.cn, qmzhu@suda.edu.cn

Abstract

Discourse structure analysis is an important research topic in natural language processing. Discourse structure analysis not only helps to understand the discourse structure and semantics, but also provides strong support for deep applications of natural language processing, such as automatic summarization, statistical machine translation, question and answering, etc. At present, the analyses of discourse structure are mainly concentrated on the micro level, while the analyses on macro level are few. Therefore, this paper focuses on the construction of representation schema and corpus resources on the macro level of discourse structure. This paper puts forward a macro discourse structure framework and constructs the logical semantic structure and functional pragmatic structure respectively. On this basis, a macro Chinese discourse structure treebank is annotated, consisting of 147 Newswire articles. Preliminary experimental results show that the representation schema and corpus resource constructed in this paper can lay the foundation for further analysis of macro discourse structure.

Keywords: discourse structure, macro discourse structure corpus, logical semantic structure, functional pragmatic structure

1. Introduction

A discourse is not formed by independent and isolated discourse units, but by related and structure units. The task of discourse analysis is to segment sentences into elementary discourse units(EDUs) and recognize the relations among them to form a complete discourse structure. Due to the semantic integrity of discourse units, discourse relations and their well-formed structure, discourse informations have been applied to many natural language processing applications, such as information retrieval (Zou et al., 2014), automatic summarization (Ferreira et al., 2014; Cohan and Goharian, 2017), question and answering (Sadek and Meziane, 2016) and statistical machine translation (Guzmán et al., 2014). Previous research works have proven that discourse informations are beneficial to these NLP applications.

The advent of large-scale collections of annotated data shifted the research community of natural language processing. These corpora have accelerated the development efforts and energized the research community.

Generally speaking, there exist two hierarchical levels of discourse structures: micro level and macro level. At present, the analyses of discourse structure are mainly concentrated on the micro level, that is, the relations and structures between sentences or sentence groups. But the analyses on macro level are relatively few, that is, the relations and structures between paragraphs or documents.

Through the above analysis, it is obvious to realize that developing a macro discourse structure corpus is helpful to understand the overall discourse information and quite necessary for macro natural language processing tasks.

2. Macro Discourse Structure Framework

The overall discourse structure is relevant to the discourse genre and discourse pattern. Thus discourse structures vary if the genres are different. For example, news articles are commonly described in "summary-story" structure, and academic papers are consist of "abstract, introduction, related work, experimentation, conclusion", while court documents are recorded in the structure of "in what way, for what reason, where, according to what inference".

We focus on the news genre in this paper, and will expand the research scope to other discourse genres in future studies. We expand the discourse analysis from intra-paragraph to the overall discourse on the basis of original discourse structure analysis.

Inspired by Rhetorical Structure Theory (Mann and Thompson, 1987) and Macrostructure Theory (Van Dijk, 1980), we explore a macro discourse structure representation schema. Furthermore, we construct the logical semantic structure and functional pragmatic structure on the macro level of discourse analysis respectively. For each structure we define the structural elements such as leaf nodes, non-leaf nodes and edges pointing.

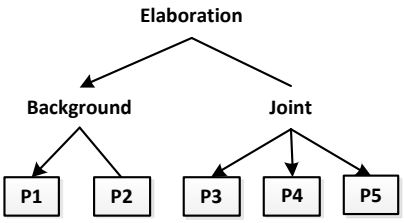

Figure 1: Logical semantic structure of chtb_0019.

Take the chtb_0019 for example, which is a typical news wire article from CTB 8.0 (Xue et al., 2002). There are five paragraphs in the news "Significant achievements in the construction of Ningbo Bonded Area", and the discourse logical semantic structure of this article is shown as Figure 1. Limited to the length of this paper, the full discourse text of this example is not included, please refer to the corpus CTB 8.0. The main contents of the five paragraphs respectively are: P1) Ningbo Bonded Area achieved fruitful results after three years of construction; P2) the basic situation of the Ningbo Bonded Area; P3) the situation of import

and export trade, warehouses, storage area, etc. P4) the situation of industrial processing projects and enterprises; P5) the situation of administrative services and information construction.

This news report is made up of 5 paragraphs(P1, P2, P3, P4 and P5). The paragraphs and paragraphs are connected by discourse relations. In the structure tree shown in Figure 1, leaf nodes represent paragraphs, and non-leaf nodes represent discourse relations. The edges connect the discourse units, while the arrows pointing to the primary discourse units. In this example, paragraph P1 points out the theme of the overall article. According to the direction of arrows in the discourse structure tree, we can quickly locate the most important part(P1) in this article.

From the discourse structure tree of this example, we can see that the analysis of discourse structure contributes to the understanding of the content and the theme of the discourse. Based on the macro discourse structure analysis, we can further enhance the performance of natural language processing applications, such as, information extraction base on the discourse structure, question answering system base on the discourse relations, and automatic summarization base on the primary-secondary relations, etc.

The detailed definitions of macro discourse structure are described as follows.

2.1. Leaf Nodes

Unlike the definition on micro level (the elementary units are treated as leaf nodes), we directly treat the paragraphs which are naturally segmented in the discourses as leaf nodes on the macro level. The natural segmentation of paragraphs are paragraphs segmented by the author's intention and the logical meaning of his writing. For example, there are five paragraphs in the news "Significant achievements in the construction of Ningbo Bonded Area" (chtb_0019), so we treat the five paragraphs(P1,P2,P3,P4,P5) as leaf nodes directly, and the discourse structure of this article is shown as Figure 1.

2.2. Non-leaf Nodes

Discourse relations connect discourse units, which are treated as non-leaf nodes in our macro discourse structure. In the representation scheme, we classify the discourse relations into three categories and fifteen subcategories, listed in Table 1.

Categories	Subcategories
Coordination	Joint, Sequence, Progression, Contrast, Supplement
Causality	Cause-Result, Result-Cause, Background, Behavior-Purpose, Purpose-Behavior
Elaboration	Elaboration, Summary, Evaluation, Statement-Illustration, Illustration-Statement

Table 1: Discourse Relations in our framework

As shown in Figure 1, the relations *Elaboration*, *Background*, and *Joint* are non-leaf nodes in the logical semantic structure tree of the chtb_0019. Specifically, P2 elaborates the background of "Ningbo Bonded Area", which forms a *Background* relation together with P1. P3, P4 and P5 elaborate the "fruitful results" achieved by Ningbo Bonded Area from three aspects respectively, and the three paragraphs form a *Joint* relation. The whole unit constituted by P3, P4 and P5 elaborate the whole unit constituted by P1 and P2, which form an *Elaboration* relation.

2.3. Edges pointing

A discourse relation generally includes two or more discourse units. These discourse units belong to the same relation layer. If one of the discourse units can generalize the intention and content of the relation layer it belongs to, and can connect to other layers on behalf of the relation layer, this discourse unit is a primary unit, while others are secondary ones. There are also some discourse relations, which have no primary and secondary differences between the discourse units they connected, so the discourse units are equally important. We define three types of primary-secondary relations: 1) primary-secondary(PS), the former unit is primary, and the latter unit is secondary; 2) secondary-primary(SP), the former unit is secondary, and the latter unit is primary; 3) equal importance(EI), the discourse units are equally important.

In the macro logical semantic discourse structure, we use the edges pointing to represent the primary-secondary relations. As shown in Figure 1, the arrows point to the primary units. P2 introduces the approval and development situation of "Ningbo Bonded Area", which is the background of the event "Ningbo Bonded Area achieved fruitful results" mentioned in P1. Obviously, P1 expresses more important semantic information, and therefore, this *Background* relation is a PS relation. In the *Joint* relation formed by P3, P4 and P5, the three paragraphs are equally important, so this *Joint* relation is an EI relation. In the *Elaboration* relation formed by P1-P2 and P3-P5, the elaborated unit P1-P2 is more import than the elaborate unit P3-P5, so this *Elaboration* relation is a PS relation. In the overall discourse, P1 can best express the discourse topic "Significant achievements in the construction of Ningbo Bonded Area" (also the discourse title), P1 is therefore the most important paragraph among all these discourse units. Based on the edges pointing of the logical semantic structure tree, readers can also get the same conclusion.

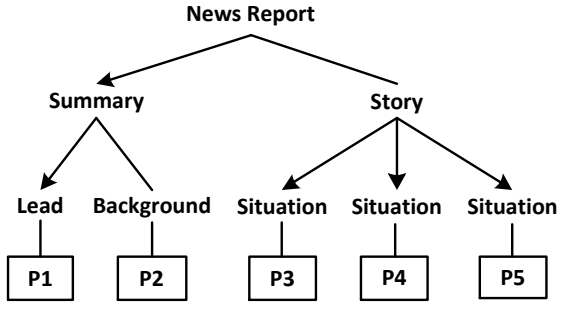

Figure 2: Functional Pragmatic Structure of chtb_0019.

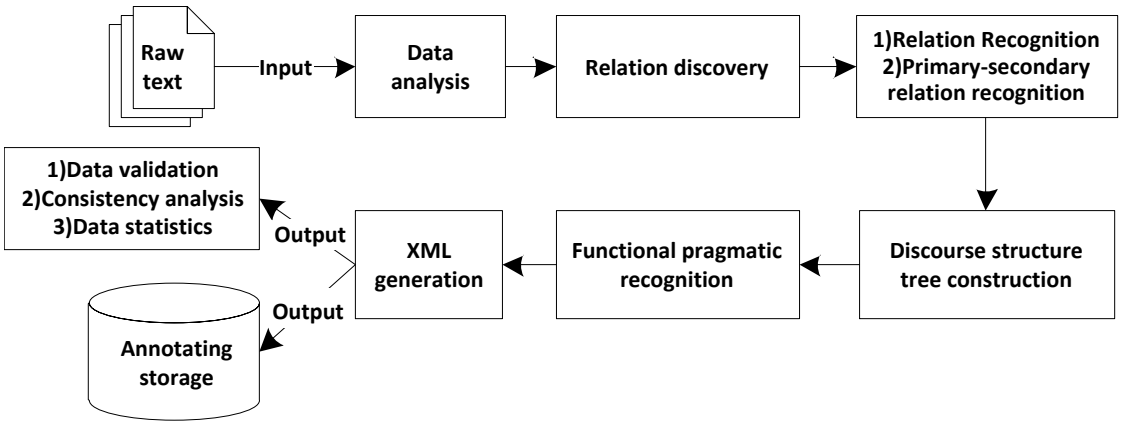

Figure 3: The processing flow of corpus annotating.

2.4. Macro Discourse Structure

In the representation schema we define, the discourse is organized as a tree structure, in which paragraphs appear in the leaf nodes and the discourse relations appear in the non-leaf nodes. The tree structure is an appropriate representation of discourse structure, which expresses the hierarchical relationship of the discourse. Essentially, the depth of the hierarchical structure indicates the depth of the corresponding discourse semantic.

2.5. Functional Pragmatic Structure

In addition to the logical semantic structure, we also define the functional pragmatic structure. Specifically, on the basis of the logical semantic structure, we add function pragmatic attribute to each node. This paper defines 18 functional types, including *News Report*, *Lead*, *Sub-Summary*, *Summary*, *Situation*, *Story*, *Cause*, *Sum-up*, *Result*, *Behavior*, *Purpose*, *Statement*, *Illustration*, *Background*, *Comment*, *Supplement*, *Contrast*, and *Progression*.

In the functional pragmatic structure of chtb_0019(as shown in Figure 2), the root node is *News Report*. P1 is the *Lead* of the article. P2 elaborates the background of P1, so its function is *Background*. P1 and P2 form the node of *Summary*. P3, P4 and P5 elaborate the *Summary* in detail, and their corresponding functions are *Situation*. The parent node of P3-P5 is defined as *Story*. Each node has its corresponding function in the article, and all the functional nodes constitute a complete article.

3. Annotation Task

Guided by the macro discourse structure framework defined above, we annotate a Macro Chinese Discourse Treebank (MCDTB) consisting of 147 Xinhua newswire articles on the top of Chinese Discourse Treebank (CDTB) (Li et al., 2014).

Because the discourse units are not isolated from the overall discourse, it's difficult to judge whether the discourse units are important or not and what relations are between the discourse units simply from the units themselves. It is necessary to have a comprehensive understanding of the overall article before the annotation work.

We have three annotators. We independently annotated articles based on an initial set of annotating guidelines, and then held discusses to compare results. At the meeting, we discussed the issue of segments, relations and primary-secondary relations, and analyzed the causes of differences. These exploratory sessions led to enhancements in the annotating guidelines and annotation quality. On the basis of the annotated logical semantic structure, we annotate the function of each node and complete the tagging of functional pragmatics. According to the contrast relation between semantic and pragmatic annotations, we summarize the transform rules from semantic to pragmatic, and construct a rule-based pragmatic transformation model by which can automatically tag functional pragmatics directly.

3.1. Annotation Strategy

We employ a combination of top-down and bottom-up strategy in the annotation work. **1) Top-down**: We determine the overall level first and then analysis goes on step by step to the individual discourse units. Such a top-down strategy can easily grasp the overall discourse structure, which consistent with the reading habit of human beings. **2) Bottom-up**: Meanwhile, we determine whether the lower discourse units need to be combined first according to the similarity of their forms and contents, and combine them together as a whole unit to contact with other parts. The annotation work shows our annotation strategy is effective.

3.2. Annotation process

Given a a piece of raw materials, we read and analyze the text first. A complete tree structure of the discourse is constructed, after the steps of relation discovery, discourse relation recognition, primary-secondary relation recognition, discourse structure tree construction. Then the annotation of functional pragmatics is automatically completed by the rule-based logical semantics to the functional pragmatics conversion program. The annotation results are saved in the form of XML. In order to ensure the consistency of the annotation, we verify the annotation results and calculate the consistency. After all the annotation work is completed, data statistics and analysis of the annotation results are car-

ried out. The specific process is shown in Figure 3)
We develop an annotation platform in order to enable annotators to construct discourse structures visually. Annotators annotate the discourse topic, lead, abstract, paragraph segmentations, paragraph topics, discourse relations, and primary-secondary relations for each discourse with the annotation platform. A complete discourse structure tree can be automatically generated and all annotation informations are saved in a XML file.

4. Quality Assurance

To ensure the quality of our corpus, we adopt the annotator consistency using agreement and kappa. Table 2 illustrates the annotator consistency in detail. We measure the agreement and kappa of discourse spans, primary-secondary relations and discourse relations. It's very difficult to achieve high consistency because the judgments of relation and structure are very subjective.
The method of consistency calculation used in this paper refers to the work of the corpus of RST (Marcu et al., 1999), and the appropriate adjustment is made according to the contents of our annotation.

Categories	Agreement	Kappa
Discourse Spans	88.54%	0.771
P-S Relations	80.67%	0.694
Discourse Relations	83.05%	0.556

Table 2: Annotating consistency (P-S Relations refers to primary-secondary relations in this table.)

5. Corpus Details

Our corpus consists of 147 newswire articles from Chinese Treebank 8.0. There are 648 paragraphs with 670 discourse relations annotated. There are 5.56 paragraphs and 624 Chinese characters in each article on average. Detailed statistical data are shown in Table 3 and Table 4.

Statistics Items	Value
Count of documents	147
Count of paragraphs	817
Amount of macro discourse relations	670
Average paragraphs (paragraphs/document)	5.56
Maximal of paragraphs	13
Minimal of paragraphs	2
Count of sentences	1,802
Average sentences (sentences/paragraph)	2.2
Count of characters	91,709
Average characters (characters /paragraph)	624

Table 3: Corpus basic statistic data

In terms of discourse relations, compared with the categories of *Coordination* and *Elaboration*, the amount of *Causality* relations is less and the data set is unbalanced. In terms of primary-secondary relations, compared with PS and EI relations, the amount of SP relation is very small, and the data set is quite unbalanced.

6. Preliminary Experiment

Based on the corpus we built, we can do the following analysis: discourse span segmentation, discourse relation recognition, primary-secondary relationship recognition, and discourse structure tree construction. In this section, we evaluated our annotated corpus with the task of recognition of primary-secondary relationship.
According to the characteristics of macro level primary-secondary relations and feature information used in the researches before, this paper adds the semantic information and takes the topic similarity as an important feature. The topic similarity refers to the semantic similarity between the discourse unit and the discourse topic. This paper puts forward two calculation methods of the topic similarity based on the word2vec (Mikolov et al., 2013) and LDA (Blei et al., 2003) respectively.
We conducted four sets of experiments to verify the feasibility of our proposed semantic similarity model and efficient. (S1) The first set of experiments used structural features, and the result was used as the benchmark system. (S2) The second groups added the LDA topic similarity as a semantic feature. (S3) The third groups added the word2vec topic similarity as a semantic feature. (S4) The fourth groups added the word2vec and LDA topic similarity both. Experimental results are shown in Table 5.

7. Discussion

Why don't we directly use the discourse relations defined in RST, but redefine new discourse relations? 1) the expressions in Chinese and English are different; 2) the expressions on micro level and macro level are different. For example, a paragraph not only discusses the background of the event, but also describes the circumstance of the event, then it cannot be determined whether it corresponds to *Circumstance* or *Background* in the RST definition.
These differences have been identified in previous studies, so there are many attempts at annotating corpora in both Chinese and English, such as Carlson et al. (2003), Yue (2008), and Li et al. (2014). Carlson et al. (2003) also discussed the annotation problem of macro-level. In the study of computational models, more and more researchers have constructed intra-sentential, multi-sentential, and multi-paragraph models separately to achieve higher performance (Joty et al., 2013; Wang et al., 2017).
The difficulties of annotating work: 1) the annotation processing is very subjectivity because of the different understandings among different annotators, so the consistency is not very high; 2) a lot of discussion is needed to achieve consistent understanding; 3) the structure framework and relation set have been repeatedly confirmed to give structure definition and annotation guidelines more clearly. Because of these reasons, the current annotation scale is not large enough, and our following research work will continue to expand the scale.

8. Conclusion

In this paper, we expand the discourse structure analysis from intra-paragraph to the overall discourse. We propose

Categories	PS	SP	EI	Subtotal	Percentage
Coordination	65	4	243	312	46.57%
Causality	94	8	7	109	16.27%
Elaboration	232	14	3	249	37.16%
Total	391	26	253	670	100.00%
Percentage	58.36%	3.88%	37.76%	100.00%	-

Table 4: Statistics of discourse relations and primary-secondary relations

Feature set	Accuracy	F-score
S1	81.96%	80.4%
S2	82.11%	80.5%
S3	82.26%	80.6%
S4	82.70%	81.1%

Table 5: Experimental results using different feature sets

a macro discourse structure representation scheme, and describe the scheme in detail. We also annotate a Marco Chinese Discourse Treebank consisted of 147 news wire articles based on the representation schema we defined. To evaluate our annotated corpus, we take the preliminary experiment with the task of recognition of primary-secondary relations. In the future work, we will enlarge the scale of the MCDTB corpus and explore the macro discourse structure computational models.

9. Acknowledgements

The work is supported by the National Natural Science Foundation of China (61773276, 61673290, 61331011) and Jiangsu Provincial Science and Technology Plan (BK20151222).

10. Bibliographical References

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Carlson, L., Marcu, D., and Okurowski, M. E. (2003). Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Current and new directions in discourse and dialogue*, pages 85–112. Springer.

Cohan, A. and Goharian, N. (2017). Scientific document summarization via citation contextualization and scientific discourse. *International Journal on Digital Libraries*, pages 1–17.

Ferreira, R., de Souza Cabral, L., Freitas, F., Lins, R. D., de França Silva, G., Simske, S. J., and Favaro, L. (2014). A multi-document summarization system based on statistics and linguistic treatment. *Expert Systems with Applications*, 41(13):5780–5787.

Guzmán, F., Joty, S., Màrquez, L., and Nakov, P. (2014). Using discourse structure improves machine translation evaluation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 687–698.

Joty, S. R., Carenini, G., Ng, R. T., and Mehdad, Y. (2013). Combining intra-and multi-sentential rhetorical parsing for document-level discourse analysis. In *ACL (1)*, pages 486–496.

Li, Y., Feng, W., Sun, J., Kong, F., and Zhou, G. (2014). Building chinese discourse corpus with connective-driven dependency tree structure. In *EMNLP*, pages 2105–2114. Citeseer.

Mann, W. C. and Thompson, S. A. (1987). Rhetorical structure theory: A theory of text organization (no. isi/rs-87-190). marina del rey. *CA: Information Sciences Institute*.

Marcu, D., Amorrortu, E., and Romera, M. (1999). Experiments in constructing a corpus of discourse trees. *Towards Standards and Tools for Discourse Tagging*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Sadek, J. and Meziane, F. (2016). A discourse-based approach for arabic question answering. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 16(2):11.

Van Dijk, T. A. (1980). *Macrostructures: An interdisciplinary study of global structures in discourse, interaction, and cognition*. Lawrence Erlbaum Associates.

Wang, Y., Li, S., and Wang, H. (2017). A two-stage parsing method for text-level discourse analysis. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 184–188.

Xue, N., Chiou, F.-D., and Palmer, M. (2002). Building a large-scale annotated chinese corpus. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–8. Association for Computational Linguistics.

Yue, M. (2008). Rhetorical structure annotation of chinese news commentaries. *Journal of Chinese Information Processing*, 4(002).

Zou, B., Zhou, G., and Zhu, Q. (2014). Negation focus identification with contextual discourse information. In *ACL (1)*, pages 522–530.