An Annotation System for Development of Chinese Discourse Corpus

Hen-Hsen Huang, Hsin-Hsi Chen
Department of Computer Science and Information Engineering,
National Taiwan University, Taipei, Taiwan
hhhuang@nlg.csie.ntu.edu.tw, hhchen@csie.ntu.edu.tw

ABSTRACT

Well-annotated discourse corpora facilitate the discourse researches. Unlike English, the Chinese discourse corpus is not widely available yet. In this paper, we present a web-based annotation system to develop a Chinese discourse corpus with much finer annotation. We first review our previous corpora from the practical point of view, then propose a flexible annotation framework, and finally demonstrate the web-based annotation system. Under the proposed annotation scheme, both the explicit and the implicit discourse relations occurring on various linguistic levels will be captured and labelled with three-level PDTB tags. Besides, the sentiment information of each instance is also annotated for advanced study.

輔助中文語篇語料庫開發的標記系統

標記詳細語篇資訊的語料庫，對於語篇研究有很大的幫助。在英文語言處理，目前已有公眾可以取得的質量良好語篇語料庫。相較之下，中文領域尚未有這樣的公開資源。語篇標記的工作需要投入相當的人力和時間，爲了提高工作效率，我們開發了一套系統，透過網頁介面，可以對中文語料標記詳細的語篇資訊。在本文中，我們首先回顧過去標記的成果，指出根據中文的語言特性，需要特別考量的要點。針對這些要點，提出了一套高度彈性的框架。在這套框架下，標記者將圈選出外顯或內隱、句內或跨句等各式各樣的語篇關係，並且標上PDTB的三階語篇關係標籤。此外，每一個語篇實例的情緒資訊也一併標記，作爲將來進階研究之用。

KEYWORDS : Chinese Discourse Analysis, Corpus Annotation, Corpus Linguistics, Sentiment Analysis
KEYWORDS IN CHINESE : 中文語篇分析, 語料標記, 語料庫語言學, 情緒分析

Proceedings of COLING 2012: Demonstration Papers, pages 223–230,
COLING 2012, Mumbai, December 2012.

1 Introduction

The study of discourse analysis attracts a lot of attention in recent years. The release of the well-annotated datasets such as the Rhetorical Structure Theory Discourse Treebank (RST-DT) (Carlson et al., 2002) and the Penn Discourse TreeBank (PDTB) (Prasad et al., 2008) facilitate the discourse researches. Many related subtopics such as discourse segmentation and discourse relation recognition grow rapidly. Discourse corpus becomes the essential component for the researches.

Both the RST-DT and the PDTB are annotated on Wall Street Journal articles from the Penn Discourse Treebank that are written in English. In Chinese, no discourse corpus is widely available yet. To investigate the Chinese discourse analysis, research groups independently developed the discourse corpora for their needs. We annotated two corpora based on the Sinica Treebank for Chinese discourse relation recognition (Huang and Chen, 2011; 2012). At present, Zhou and Xue (2012) are annotating the Penn Chinese Treebank with the PDTB-style scheme.

English and Chinese natives have their own written styles. Chen (1994) showed that the number of sentence terminators (period, question and exclamation marks) is a little larger than segment separators (comma and semicolon) in English. In contrast, the segment separators outnumber the sentence terminators in Chinese with the ratio 7:2 (Chen, 1994). It results in many segments in Chinese sentences. Analyses of documents randomly sampling from Sinica Chinese Treebank (Huang et al., 2000) show the distribution of the number of segments in Chinese sentences is 1 segment (12.18%), 2 segments (18.35%), 3 segments (20.15%), 4 segments (15.72%), 5 segments (12.91%), 6-10 segments (17.72%), and more than 10 segments (2.97%). Long sentences tend to have more complex structural relationships and thus make Chinese discourse annotation challenging.

For our previous two discourse annotation work (Huang and Chen, 2011; 2012), different annotation schemes were used. One corpus was annotated on the sentence level with the PDTB four-class tags. Another corpus was annotated on the clause level with the Contingency and the Comparison relations from the PDTB four-class tags. In this paper, we consider the specific written style of Chinese sentences and propose a flexible annotation scheme to develop a new Chinese discourse corpus.

In this corpus, the three level discourse relation tags from the PDTB 2.0 are fully used (Prasad et al., 2007). The discourse units can be on various levels. An argument of a discourse pair can be as short as a clause and as long as several sentences. In addition, the nested discourse pairs are annotated in our scheme. For example, the sentence (S1) is a Chinese sentence that consists of three clauses. As illustrated in Figure 1, (S1) forms a Comparison discourse pair on the top level, and it contains a nested Contingency discourse pair. We annotate not only the discourse relations, but also the sentiment information of each discourse pair and its two arguments. As shown in Figure 1, the polarity of the first clause is positive, the polarity of the fragments that consist of the last two clauses is negative, and finally the whole statement (S1) constitute a polarity of negative. Such information is valuable for the study of the correlations between discourse relation and sentiment analysis.

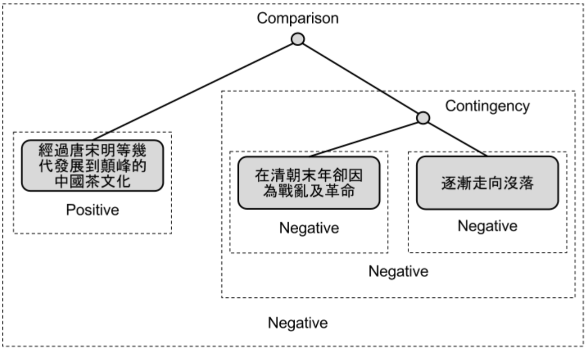


FIGURE 1 - Discourse structure and sentiment polarities of (S1).

(S1) 經過唐宋明等幾代發展到顛峰的中國茶文化 (‘After several dynasties such as Tang, Song, and Ming, the Chinese tea culture developed to the peak’)，在清朝末年卻因爲戰亂及革命 (‘however, because of war and revolution at the end of the Qing Dynasty’)，逐漸走向沒落 (‘gradually declined’)。

Constructing a well-annotated corpus with adequate amounts of data is not a trivial task. Various considerations and design processes should be involved. In this paper, we aim to share our experience of developing Chinese discourse corpora and introduce the approaches to facilitate the annotation work with a web-based system.

The rest of this paper is organized as follows. In Section 2, our previous two Chinese discourse corpora, which are annotated on the inter-sentential level and the intra-sentential level, respectively, are analyzed. Consideration and the annotation plan of the Chinese discourse corpus are described in Section 3. The design and its current status are given in Section 4. Finally, we conclude this paper in the last section.

2 Two Pilot Chinese Discourse Corpora

Two pilot Chinese discourse corpora were developed on the Sinica Treebank 3.1 (Huang et al., 2000), which is a traditional Chinese Treebank based on the Academic Sinica Balanced Corpus (Huang and Chen, 1992). To tackle the issue of Chinese discourse recognition, a moderate-sized corpus with the fundamental discourse relation was tagged as our first Chinese discourse corpus (Huang and Chen, 2011). For each article, annotators tag the discourse relation between every two successive sentences with one of the PDTB top four classes: *Temporal*, *Contingency*, *Comparison*, and *Expansion*. These four classes are the top level tags in the PDTB tagging system.

The downside of this corpus is that only a few Comparison and Contingency relations are labelled. After analysis, we find the Contingency and the Comparison relations tend to occur within a sentence, especially the Contingency relation. Since we annotate the relations on the inter-sentence level only, such instances are missing. Besides, the nested relations shown in Figure 1 are also completely missing in this corpus because only the relations between every two successive sentences are labelled.

To study the Contingency and the Comparison relations occurring in sentences and their nested structure, an intra-sentential corpus was constructed as our second corpus (Huang and Chen, 2012). The discourse unit in this corpus is clause, which is defined as a sequence of words in a sentence that are delimited by commas ('，'). Annotators decide the structure of a sentence and tag the relations between every successive clause in the sentence. To simplify the annotation work, only the sentences that consist of two, three, and four clauses are selected.

3 More Practical Considerations

To annotate a Chinese discourse corpus, we should tackle some practical issues. Firstly, the unit of a discourse argument is not regular. As mentioned in Section 1, an argument of a discourse pair may be as short as a clause, and may also be as long as several sentences. The more vexing case is the nested discourse relations illustrated in Figure 1. Annotators have to determine the correct boundary of arguments. That is important for training and testing discourse parsers.

Secondly, discourse markers are important clues for labelling discourse relations. In English, the explicit discourse markers are defined as three grammatical classes of connectives, including subordinating conjunctions, coordinating conjunctions, and adverbial connectives (Prasad et al., 2008). These words can be automatically extracted using a syntactic parser or a POS tagger. However, it is not clear what the Chinese discourse markers are. Cheng and Tian (1989) suggested a dictionary of Chinese discourse markers, which consist of many words including connectives and various parts of speech such as adverbs, verbs, prepositions, and time nouns.

Detecting the Chinese discourse markers automatically is not trivial. Wrong segmentation is prone to result in the less accurate marker detection. Besides, some words in a discourse marker dictionary are general function words that can be used in other purposes rather than discourse relation marking only. For example, the word 或 ("or") can be used as a discourse marker of the Expansion relation and a correlative conjunction. Thus, to disambiguate if a word is used as a discourse marker is necessary. Furthermore, the vocabulary of Chinese discourse markers is not a closed set. The explicit discourse markers are labelled by annotators on the character level.

Thirdly, veridicality is a property of a discourse relation that specifies whether both arguments of a discourse pair are truth or not (Hutchinson, 2004). In the three-level PTDB tagging scheme, the veridicality will be distinguished in different tags. For example, the tag CONTINGENCY:Condition:unreal-past indicates a discourse pair where the second argument of the pair did not occur in the past and the first argument denotes what the effect would have been if the second argument had occurred. By labelling the data with the full PDTB tagging scheme, the veridical information of the discourse pairs are naturally labelled at the same time.

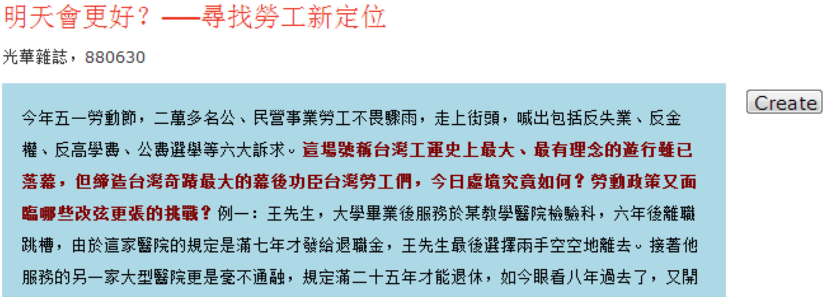
Fourthly, sentiment polarity is another property of a discourse relation that indicates the sentiment transition between the two arguments of a discourse pair (Hutchinson, 2004).

Such information will help us realize the correlations between discourse relations and sentiment polarities.

4 An Annotation Framework

A flexible interface that allows annotators to label a variety of discourse relations with detailed information is proposed. An annotator first signs in to the online annotation system, and a list of articles that are assigned to the annotator are given. The annotator labels the articles one by one. As shown in Figure 2, the annotator selects the clauses that form a discourse pair in the text if it is found. The selected clauses will be denoted in the bold and red font. The annotator clicks the button “Create” when all the clauses belonging to this discourse pair are selected, and then the advanced annotation window will be popped up.

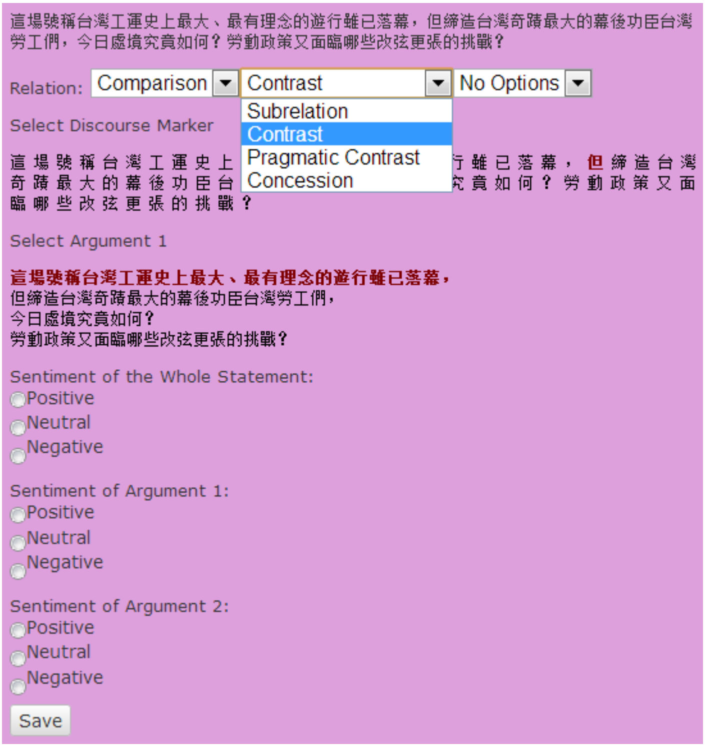
As shown in Figure 3, the discourse relation, the discourse marker, the boundaries of arguments, and the sentiment polarities of the two arguments and the entire discourse pairs are labelled in the pop-up window. The entire selected discourse pair is present in the top of the pop-up window. The following is the drop-down selection lists that correspond to the three levels of hierarchical discourse relation tags used in the PDTB. The next part is about to highlight the discourse markers from the text. As mentioned in Section 3, the annotator highlights the discourse marker on the character level. The annotator can select multiple characters for the phrase or the pairwise discourse marker such as “因爲 …, 所以…” (“Because ..., so ...”). The implicit discourse relation is distinguished if no discourse marker is highlighted. And then, the annotator splits the first argument and the second argument by selecting the clauses belonging to the first argument. The rest clauses are regarded as the second argument. The last part of annotation is labeling the sentiment information. There are three types of sentiment polarity, i.e., positive, neutral, and negative. The polarities of the whole discourse pair and both of its two arguments will be labelled. The annotator is asked to judge the sentiment polarities on the pragmatic level. That is, the sentiment polarity of the text is not determined by the surface semantics, but by its real meaning. The annotator submits the annotation by clicking the button ‘Save’ and continues to look up another discourse pair in the article. The nested relations can be annotated in this interface by choosing the repeated clauses in different rounds.


明天會更好？——尋找勞工新定位

光華雜誌，880630

今年五一勞動節，二萬多名公、民營事業勞工不畏驟雨，走上街頭，喊出包括反失業、反金權、反高學費、公費選舉等六大訴求。**這場號稱台灣工運史上最大、最有理念的遊行雖已落幕，但締造台灣奇蹟最大的幕後功臣台灣勞工們，今日處境究竟如何？勞動政策又面臨哪些改弦更張的挑戰？**例一：王先生，大學畢業後服務於某教學醫院檢驗科，六年後離職跳槽，由於這家醫院的規定是滿七年才發給退職金，王先生最後選擇兩手空空地離去。接著他服務的另一家大型醫院更是毫不通融，規定滿二十五年才能退休，如今眼看八年過去了，又開

FIGURE 2 – Choosing a discourse pair on the web-based online system



這場號稱台灣工運史上最大、最有理念的遊行雖已落幕，但締造台灣奇蹟最大的幕後功臣台灣勞工們，今日處境究竟如何？勞動政策又面臨哪些改弦更張的挑戰？

Relation: Comparison ▾ Contrast ▾ No Options ▾

Select Discourse Marker

這場號稱台灣工運史上 行雖已落幕，但締造台灣
奇蹟最大的幕後功臣台 究竟如何？勞動政策又面
臨哪些改弦更張的挑戰？

Select Argument 1

這場號稱台灣工運史上最大、最有理念的遊行雖已落幕，
但締造台灣奇蹟最大的幕後功臣台灣勞工們，
今日處境究竟如何？
勞動政策又面臨哪些改弦更張的挑戰？

Sentiment of the Whole Statement:
○Positive
○Neutral
○Negative

Sentiment of Argument 1:
○Positive
○Neutral
○Negative

Sentiment of Argument 2:
○Positive
○Neutral
○Negative

FIGURE 3 – Labeling the information for the chosen discourse pair with the web-based online system

Conclusion

Discourse corpus is indispensable for the study of discourse analysis. In this paper, we address some considerations specific to Chinese language. A flexible annotation framework is proposed to cover a variety of discourse relations and determine the argument boundary of a relation. Furthermore, the sentiment polarities are also annotated on the discourse pairs and their arguments. Such a corpus is helpful for the exploration of the areas of Chinese discourse processing and sentiment analysis. The cost of the detailed annotation is much higher and the annotation task is time-consuming. In order to facilitate the complicated annotation work, we demonstrate a web-based system that supports annotators to do the work fast and accurately.

Acknowledgments

This research was partially supported by Excellent Research Projects of National Taiwan University under contract 101R890858 and 2012 Google Research Award.

References

Carlson, L., Marcu, D., and Okurowski, M.E. (2002). *RST Discourse Treebank*. Linguistic Data Consortium, Philadelphia.

Chen, H.H. (1994). The Contextual Analysis of Chinese Sentences with Punctuation Marks. *Literal and Linguistic Computing*, Oxford University Press, 9(4): 281-289.

Cheng, X. and Tian, X. (1989). Xian dai Han yu (現代漢語), San lian shu dian (三聯書店), Hong Kong.

Huang, C. R., Chen, F. Y., Chen, K. J., Gao, Z. M., and Chen, K. Y. (2000). Sinica Treebank: design, criteria, annotation guidelines, and on-line interface. In *the 2nd Chinese Language Processing Workshop*, pages 29-37, Hong Kong, China.

Huang, C.R. and Chen, K.J., (1992). A Chinese corpus for linguistics research. In *the 14th International Conference on Computational Linguistics* (*COLING-92*), pages 1214-1217, Nantes, France.

Huang, H.H. and Chen, H.H. (2011). Chinese discourse relation recognition. In *the 5th International Joint Conference on Natural Language Processing* (*IJCNLP 2011*). pages 1442-1446, Chiang Mai, Thailand.

Huang, H.H. and Chen, H.H. (2012). Contingency and comparison relation labeling and structure prediction in Chinese sentences. In *the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue* (*SIGDIAL 2012*), pages 261-269, Seoul, South Korea.

Hutchinson, B. (2004). Acquiring the meaning of discourse markers. In *the 42nd Annual Meeting of the Association for Computational Linguistics* (*ACL 2004*), pages 684-691, Barcelona, Spain.

Prasad, R., Miltsakaki, E., Dinesh, N., Lee, A., Joshi, A., Robaldo, L., and Webber, B. (2007). *The Penn Discourse Treebank 2.0 Annotation Manual*. The PDTB Research Group.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., and Webber, B. (2008). The Penn Discourse TreeBank 2.0. In *the 6th Language Resources and Evaluation Conference* (*LREC 2008*), pages 2961-2968, Marrakech, Morocco.

Zhou, Y. & Xue, N. (2012). PDTB-style discourse annotation of Chinese text. In *the 50th Annual Meeting of the Association for Computational Linguistics* (*ACL 2012*), pages 69-77, Jeju, South Korea.