Building Chinese Discourse Corpus with Connective-driven Dependency Tree Structure

Yancui Li[1,2] **Wenhe Feng**[2] **Jing Sun**[1] **Fang Kong**[1] **Guodong Zhou**[1]
[1]Natural Language Processing Lab, School of Computer Science and Technology, Soochow University, Suzhou 215006, China
[2]Henan Institute of Science and Technology, Xinxiang 453003, China
{yancuili, wenhefeng}@gmail.com {20104027009, kongfang, gdzhou}@suda.edu.cn

Abstract

In this paper, we propose a Connective-driven Dependency Tree (CDT) scheme to represent the discourse rhetorical structure in Chinese language, with elementary discourse units as leaf nodes and connectives as non-leaf nodes, largely motivated by the Penn Discourse Treebank and the Rhetorical Structure Theory. In particular, connectives are employed to directly represent the hierarchy of the tree structure and the rhetorical relation of a discourse, while the nuclei of discourse units are globally determined with reference to the dependency theory. Guided by the CDT scheme, we manually annotate a Chinese Discourse Treebank (CDTB) of 500 documents. Preliminary evaluation justifies the appropriateness of the CDT scheme to Chinese discourse analysis and the usefulness of our manually annotated CDTB corpus.

1 Introduction

It is well-known that interpretation of a text requires understanding of its rhetorical relation hierarchy since discourse units rarely exist in isolation. Such discourse structure is fundamental to many text-based applications, such as summarization (Marcu, 2000) and question-answering (Verberne et al., 2007). Due to the wide and potential use of discourse structure, constructing discourse resources has been attracting more and more attention in recent years. In comparison with English, there are much fewer discourse resources for Chinese which largely restricts the researches in Chinese discourse analysis.

The general notion of discourse structure mainly consists of discourse unit, connective, structure, relation and nuclearity. However, previous studies on discourse failed to fully express these kinds of information. For example, the Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) represents a discourse as a tree with phrases or clauses as elementary discourse units (EDUs). However, RST ignores the importance of connectives to a great extent. Figure 1 gives an example tree structure with four EDUs (e1-e4). In comparison, Penn Discourse Treebank (PDTB) (Prasad et al., 2008) adopts the predicate-argument view of discourse relation, with discourse connective as predicate and two text spans as its arguments. Example (1) shows an explicit reason relation signaled by the discourse connective "particularly if" and an implicit result relation represented by the inserted discourse connective "so", with Arg1 in italics and Arg2 in bold. However, as a connective and its arguments are determined in a local contextual window, it is normally difficult to deduce a complete discourse structure from such a connective-argument scheme. In this sense, the PDTB at best only provides a partial solution to the discourse structure.

[Catching up with commercial competitors in retail banking and financial services,] e1 [they argue,] e2 [will be difficult,] e3 [particularly if market conditions turn sour.]e4

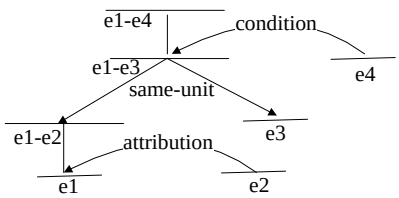

Figure 1: An example of discourse structure in RST

Example (1): An example of the connective-argument scheme in PDTB

A)[Catching up with commercial competitors in retail banking and financial services will be difficult]$_{Arg1}$, they argue, will be difficult, particularly **if** **[market**

Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 2105–2114, October 25-29, 2014, Doha, Qatar.

conditions turn sour] $_{Arg2}$. (Contingency.Condition. Hypothetical) (0616)

B) So much of the stuff poured into its Austin, Texas, offices *[that its mail rooms there simply stopped delivering it.]*$_{Arg1}$ (Implicit = so)**[Now, thousands of mailers, catalogs and sales pitches go straight into the trash.]**$_{Arg2}$ (Contingency.Cause. Result) (0989)

Obviously, both RST and PDTB have their own advantages and disadvantages in representing different characteristics of the discourse structure. In this paper, we attempt to propose a new scheme to Chinese discourse structure, adopt advantages of the tree structure from RST and connective from PDTB. Meanwhile, the special characteristics of Chinese discourse structure are well addressed.

First, it is difficult to define EDU in Chinese due to the frequent occurrence of the ellipsis of subjects, objects and predicates, and the lack of functional marks for EDU. Second, the connectives in Chinese omit much more frequently than those in English with about 82.0% vs. 54.5% in Zhou and Xue (2012). In Example (2), there are even no explicit connectives. Third, previous studies have shown the difference in classifying Chinese discourse relations from English (Xing, 2001; Huang and Liao, 2011). This suggests that the discourse relations defined for English (both RST and PDTB) are not readily suitable for Chinese. Finally, the nucleus of a Chinese discourse relation is normally not directly related to a particular relation type but should be dynamically determined from the global meaning of a discourse.

Example (2): An example of discourse with 4 EDUs

[据悉， 东莞 海关 共 接受
According to reports,Dongguan Customs total accept
企业 合同 备案 八千四百多份，]e1 [比 试点
company contract record 8400 plus class, than pilot
前 略 有 上升，]e2 [企业 反应 良好，]e3
before a slight increase, company responses well,
[普遍 表示 接受。]e4
generally acknowledge acceptance.

"[According to reports, Dongguan District Customs accepted more than 8400 records of company contracts,] e1 [a slight increase from before the pilot.]e2 [Companies responded well,]e3 [generally acknowledging acceptance.]e4"

In this paper, we present a Connective-driven Dependency Tree (CDT) discourse representation scheme, which takes advantage of both RST and PDTB, with elementary discourse units (limited to clauses) as leaf nodes and connectives as non-leaf nodes. Especially, we define EDU from three aspects, and employ the connective' level and semantic to indicate the rhetorical structure and the discourse relation. Besides, the nuclearity of discourse units in a discourse relation is decided on the overall discourse meaning. On the basis, we adopt the CDT scheme to annotate a certain scale corpus, called Chinese Discourse Treebank (CDTB) thereafter in this paper. Evaluation shows the appropriateness of the CDT scheme to Chinese discourse analysis.

The rest of this paper is organized as follows. Section 2 overviews related work. In Section 3, we present the CDT discourse representation scheme. In Section 4, we describe the annotation of the CDTB corpus. Section 5 compares CDTB with other major discourse corpora. Section 6 gives the experimental results on EDU recognition, the crucial step for discourse parsing. Finally, conclusion is given in section 7.

2 Related Work

In the past decade, several discourse corpora for English have emerged, with the Rhetorical Structure Theory Discourse Treebank (RST-DT) (Carlson et al., 2003) and the Penn Discourse Treebank (PDTB) (Prasad et al., 2008) most prevalent.

In the RST framework, a text is represented as a discourse tree, with non-overlapping text spans (either phrases or clauses) as leaves, and adjacent nodes are related through particular rhetorical relations to form a discourse sub-tree, which is then related to other adjacent nodes in the tree structure. According to RST, there are two types of discourse relations, mononuclear and multinuclear. Figure 1 shows an example of discourse tree representation, following the notational convention of RST. Among the four EDUs (e1-e4), e1 and e2 are connected by a mononuclear relation "attribution", where e1 is the nucleus, the span (e1-e2) and the EDU e3 are further connected by a multi-nuclear relation "same-unit", where they are equally salient. Annotated according to the RST framework, the RST-DT consists of 385 documents from the Wall Street Journal (WSJ). Besides, the original 24 discourse relations defined by Mann and Thompson (1988) are further divided into a set of 18 relation classes with 78 finer grained rhetorical relations in RST-DT.

As the largest discourse corpus so far, the Penn Discourse Treebank (PDTB) contains over one million words from WSJ. With EDUs limited to clauses, the PDTB adopts the predicate-

argument view of discourse relations, with connective as predicate and two text spans as its arguments. Example (1) shows two annotation tokens for the connective "particularly if" and "so". The current version of PDTB 2.0 annotates 40600 tokens, including 18459 explicit relations of 100 distinct types (e.g. "particularly if" and "if" are the same type) and 16224 implicit discourse relations of 102 distinct token types. Besides, PDTB provides a three level hierarchy of relation tags with the first level consisting of four major relation classes (Temporal, Contingency, Comparison, and Expansion), which are further divided into 16 types and 23 subtypes.

In comparison, there are few researches on Chinese discourse annotation (Xue, 2005a; Chen, 2006; Yue, 2008; Huang and Chen, 2011; Zhou and Xue, 2012), with no exception employing existing RST or PDTB frameworks. For example, Zhou and Xue (2012) use the PDTB annotation guidelines to annotate Chinese discourse with 98 files from Chinese Treebank (Xue et al., 2005b) of Xinhua newswire. In particular, they adopt a lexically grounded approach and make some adaptation based on the linguistic and statistical characteristics of Chinese text, with Arg1 and Arg2 defined semantically and the senses of discourse relations annotated besides connectives and their lexical alternatives. The agreement on relation types reaches 95.1% and the agreement on implicit relations with exact span match reaches 76.9%.

Instead, Chen (2006) and Yue (2008) use RST to annotate Chinese discourse. Chen (2006) selects comma as the segmentation signal of EDUs (in Example (2), "据悉(According to reports)" will be segmented as an EDU), and finds that RST fails to deal with some special features of Chinese. Yue (2008) manually annotates a set of 97 texts according to RST and shows the cross-lingual transferability of RST to Chinese. However, it also shows that EDUs in Chinese are much different from those in English, and many relation types in Chinese have no correspondence to English, and vice versa.

3 Connective-driven Dependency Tree

An appropriate representation scheme is fundamental to linguistic resource construction. With reference to various theories and representation scheme on the tree structure and nuclearity of RST, the connective, relation and discourse structure of Chinese complex sentence (Xing, 2001), the sentence-group theory (Cao, 1984), the connective treatment of PDTB, the conjunction dependent analysis (Feng and Ji, 2011) and the center theory of dependency grammar (Hays, 1964), we propose a new discourse representation scheme for Chinese, called Connective-driven Dependency Tree (CDT), with EDUs as leaf nodes and connectives as non-leaf nodes, to accommodate the special characteristics of the Chinese language in discourse structure.

For instance, Example (3) consists of 2 sentences, which is part of a paragraph from "chtb_0001", and its corresponding CDT representation is shown in Figure 2. Here, the number of "|" in Example (3) stands for the level of EDUs in CDT and the numbers marked in Figure 2 (such as 1, 2 etc.) distinguish EDUs. While an arrow points to the main EDU or main discourse unit (called nucleus), the combination of different EDUs can be considered as EDUs in a higher level and the new discourse units can thus be combined into higher-level units from bottom to up. In this way, the discourse structure can be expressed as a tree structure via bottom-up combination of EDUs.

Obviously, such discourse structure is constructed by two kinds of basic units, EDUs (leaf nodes) and connectives (non-leaf nodes). On the one hand, connectives can represent the discourse structure by its hierarchical level in the tree. The discourse structure is independent on the connective level essentially, rather than the reverse. On the other hand, connectives themselves can represent the discourse relation. This is why we call the scheme "Connective-driven". As for the abstract discourse relation, we can construct a set of discourse relations, mapping a connective to discourse relation, according to the users' specific requirements.

Example (3): CDT example from CTB

1 浦东 开发 开放 是 一项 振兴 上海，建设
Pudong development open up is a promote Shanghai, construct
现代化 经济、贸易、 金融 中心 的 跨世纪
modern economy, trade, financial century De cross-century
工程，‖2 (因此) 大量 出现的是 以前 不曾
project, therefore a large number arisen De previously never
遇到过的 新 情 况、新问题。| 3 (对此)， 浦东 {不是}
encounter DE new situation, new problem.To this, Pudong not
简单 的 采取 "干 一段 时间，等 积累 了
simply DE adopting "does a period time, wait accumulate Le
经验 以后再制定 法规 条例" 的 做法，‖4{而是}
experience after re-enactment laws regulations De approach,but
借鉴 发达 国家 和 深圳 等 特区 的
learn developed countries and Shenzhen etc. special zone DE
经验 教训，|||| 5<并且>聘请 国内外 有关专家
experience lesson, Invite at home and abroad revlant expert
学者，|||| 6<并且>积极、及时地 制定 和 推出

scholars, actively, timely DI formulate and issuing
法规性 文件，||| 7 {使} 这些 经济 活动 一 出现就 被
statutory file, make these economic activity as soon as appear bei
纳入 法制 轨道。
bring into legality track.

"1 Pudong's development and opening up is a century-spanning undertaking for vigorously promoting Shanghai and constructing a modern economic, trade, and financial center. ‖ 2 **Because of this,** new situations and new questions that have not been encountered before are emerging in great numbers. | 3 In response to this, Pudong is not simply adopting an approach of "work for a short time and then draw up laws and regulations only after experience has been accumulated."‖ 4 **Instead,** Pudong is taking advantage of the lessons from experience of developed countries and special regions such as Shenzhen, ||||5 by hiring appropriate domestic and foreign specialists and scholars, ||||6 actively and promptly formulating and issuing regulatory documents. ||| 7 So these economic activities are incorporated into the sphere of influence of the legal system as soon as they appear."

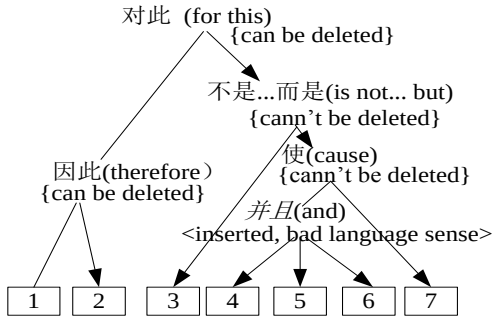

Figure 2: CDT representation of Example (3)

3.1 Elementary Discourse Unit

As the leaf nodes of CDT, EDUs are limited to clauses. In principle, EDUs play a crucial role to discourse analysis. Since from bottom-up discourse combination, EDUs are the start of discourse analysis, while from top-down discourse segmentation, they are the end of discourse analysis. Unfortunately, since there lacks obvious distinction between Chinese sentence structure and phrase structure, it is rather difficult to define Chinese EDU (clause). Till now, there is still no widely accepted definition in the Chinese linguistics community (Wang, 2010). Inspired by Li et al. (2013a), we give the definition of Chinese EDU from three perspectives. First, from the syntactic structure perspective, an EDU should contain at least one predicate and express at least one proposition. Second, from the functional perspective, an EDU should be related to other EDUs with some propositional function, i.e. not act as a grammatical element of other EDUs. Finally, from the morphological perspective, an EDU should be segmented by some punctuation, e.g. comma, semicolon and period. We use punctuation because there usually has a pause between clauses (EDUs), which can be shown in written commas, semicolons etc (Huang and Liao, 2011). Normally, it is easy to handle complex sentences and special sentence patterns (e.g. serial predicate sentences). For Example (4), A) is a single sentence with serial predicate; B) is complex sentence with two EDUs (clauses):

Example (4): EDU examples

A) He opened the door and went out. (single sentence, serial predicate, one EDU)
B) 1 He opened the door,| 2 **and** went out. (complex sentence, two EDUs)

Take as example, there exist 7 EDUs in Example (3), each marked with a number in front. According to our definition, the fragment "干一段时间，…法规条例"("work for a short time…has been accumulated") in EDU 3 is not segmented as a EDU since: 1) it acts as a grammatical element of other EDUs and has no direct relationship with other EDUs on propositional function; 2) it is marked by a pair of quotation marks and does not end with any punctuation. In contrast, the fragment "而是借鉴发达…法制轨道"("but learn developed…legality track.") is segment as 4 EDUs since it meets the three criteria in our EDU definition.

3.2 Connective

As non-leaf nodes in the CDT representation, connectives connect EDUs or discourse units. Thus, the main criterion of determining whether an expression is a connective is to check whether the two fragments it connects are EDUs (or discourse units). In our scheme, the list of explicit discourse connectives is judged by a data driven approach, i.e. with any discourse-like word or phrase marked as connective in the annotation practice, e.g. "因此(therefore)", "对此(to this)", "不是...而是...(is not...but....)", "使(so that), "正因为(just because)" in Example (3), "先...然后(first...then) ", "同时也(and at the same time)" in Example (5).

Example (5): Connective examples from CTB

A) 1<如果；只要>建筑公司进区，| 2 有关部门**先**送上这些法规性文件，‖ 3 **然后**有专门队伍进行监督检查。(chtb_0001)
1<If ; As long as>The construction company enters the region, |2 **first** the appropriate bureau delivers these regulatory documents,‖ 3 **Then** there is a specialized contingent that carries out a supervisory inspection.
B) 1 加工贸易…, 2 **同时 也** 是粤港澳台经贸合作的重要内容。(chtb_0031)

1 The processing trade …, | 2 **and at the same time** is important content in the economic and trade cooperation between Guangdong, Hong Kong, Macao and Taiwan.

It is worthy of mention that from the part-of-speech perspective, connectives are not necessarily conjunctions. For example, in Example (3) and (5), adverbs "先...然后(first…then)", verb phrases "不是…而是(is not…but)", and preposition phrases "对此(to this)" are determined as connectives. From the morphological perspective, a connective may contain more than one word, even discontinuous. As a common occurring phenomenon in Chinese discourse, there exist many paired Chinese connectives, e.g. "不是…而是 (is not…but)" in Figure 2. Even in some paired connectives, such as "因为…所以(because…so)", a word in a paired connective can appear independently as a connective. Please note that this may not be applied to other cases, e.g. "不是…而是 (is not…but)" as appeared in Example (3). Moreover, in many cases whether an expression is a connective or not depends on its meaning, e.g., "为 (in order to)" is a connective, while "为 (for)" is not. For the positional distribution, a connective may appear anywhere, i.e. in the beginning, middle, or the end of the first or second EDU. Example (3) and (5) show some of cases in different positions. The above characteristics pose special challenges on connective determination in Chinese language.

According to the appearance of a connective or not, a discourse relation can be either explicit or implicit. Previous studies have shown the difficulty of implicit relation recognition in English due to the omission of connectives (Pitler et al., 2009; Lin et al., 2009). This becomes even worse in Chinese since compared with the implicit ratio of 54.5% in English connectives, this ratio rises up to about 82% in Chinese (Zhou and Xue, 2012). It is worth noting that the majority of discourse relations in Chinese are implicit, so the insertion of a connective in an implicit position can significantly ease the understanding of the discourse. That is, a connective driven representation scheme is still applicable to a discourse with implicit connectives. To help determine implicit relations, two special strategies are proposed.

First, for each explicit connective, a decision is made whether or not it can be deleted without changing the rhetorical relation of a discourse. It should be emphasized that this constraint is largely semantic. The motivation behind the removal of explicit connectives is to enlarge implicit instances and help recognize implicit relations. As shown in Figure 2, we use the paired mark "()" to indicate that a connective can be deleted, e.g. connectives "(对此 to this)", "(因此 therefore)", "(正因此 just because)", and the paired mark "{}" to indicate that a connective cannot be deleted, e.g. connectives "{使 so that}", "{不是…而是 is not…but}".

Second, since a connective can be inserted to represent an implicit relation, our scheme tries to insert a connective which can be easily interpreted from the semantic perspective with little ambiguity into the most appropriate place. Most of the connective insertions for implicit relations occur between adjacent discourse spans. It is worth noting that not all implicit connectives are subjective to the language sense. To mark this difference, we cluster implicit connectives into two categories according to their language senses, either "good language intuition" or "bad language intuition". In our scheme, we use the paired mark "<>"to indicate inserted implicit connectives, e.g. connectives"<例如 e.g.>", "<却 but>" with "good language sense", connective "<并且 and>" with "bad language sense", as shown in Figure 2.

In some cases, it is possible that there exist several insertion options for an implicit connective due to the ambiguity in a discourse. For example, in Example (5A), connectives "如果 (if)" and "只要 (as long as)" are inserted into the first level to show the two discourse relation options. As far as this happens, connectives are inserted and ordered according to annotators' first intuition.

3.3 Discourse Structure

In Figure 2, the paragraph is organized as a tree structure, in which EDUs appear in the leaf nodes and the connectives appear in the non-leaf ones. The adoption of tree structure conforms to traditional Chinese discourse theories and practice. For example, a native Chinese speaker tends to determine the overall level boundary first and then the analysis goes on step by step to the individual clauses, when understanding a complex sentence. This process naturally forms a tree structure. Besides, tree structure is easier to formalize, compared with graph.

More specifically, the hierarchical structure of connectives indicates the hierarchical structure of discourse units. Apparently, discourse struc-

ture analysis can be viewed as hierarchical analysis of connectives, with hierarchical connective structure reflecting hierarchical combination of discourse units. Essentially, the discourse hierarchy indicates the correlation degrees of semantic relations in the discourse, the deeper tree level of two discourse units, the higher correlation degree of their semantic relation. Therefore, a discourse relation is the ultimate factor for the choice of hierarchical discourse structure. For a reference, please take Sentence 2 in Figure 2 as an example.

3.4 Discourse Relation

For discourse relation representation, a general approach is to assign an abstract relation type to a discourse relation directly, such as cause, conjunction, condition, purpose, etc, as done in RST-DT and PDTB. In our CDT scheme, we avoid to directly assign an abstract relation type to a discourse relation. Instead, we use the connective itself to express the discourse relation, as shown in Figure 2. In this way, the difficulty of pre-defining a set of acknowledged discourse relations and selecting an exact discourse relation can be avoided during the corpus annotation process. Since a Chinese discourse relation is largely controlled by connective (Xing, 2001), the key to determine a relation is to identify a suitable connective. Normally, most of relation annotations can easily map from connectives to abstract semantic classes of relations, if necessary, with the help of the discourse context. The majority of discourse relations in Chinese are implicit, but it makes sense to insist on a connective driven representation. With connective as a bridge, at least it makes discourse representation easier.

For the abstraction of discourse relations, we leave it in a later separate stage. Of course, there are cases where a connective may represent more than one discourse relation. For example, connective "而" can denotes the continuous relation "而 (especially)" and the transitional relation "而 (however)". Compared with annotating discourse relation directly, annotator's intuition is more accurate for specific connective. We don't object to label discourse relation, referring to the general work and Chinese analysis practice, give a set of relations (Figure 3), regarding it as connective's semantics, and then annotate the connective with it. In this way, we can obtain a general relation set and resolve the connective's polysemy problem. We believe that the connective itself is the foundation of discourse relation, and the relation set can be adjusted dynamically according to the application requirements.

Figure 3 shows a three-level set of discourse relations example. In the first level, this set contains four relations of causality, coordination, transition and explanation, which are further clustered into 17 sub-relations in the second level. For example, relation causality contains 6 sub-relations, i.e. cause-result, inference, hypothetical, purpose, condition and background. In the third level, the connectives are under each sub-relation. For example, cause-result relation can be represented by "because", 'therefore' etc. The numbers shown in the parentheses illustrate the distributions of different relations in our corpus. For example, there are 1335 causality relations in the first level, including 686 cause-result relations, 38 inference relations, 70 hypothetical relations, 335 purpose relations, 72 condition relations and 134 background relations.

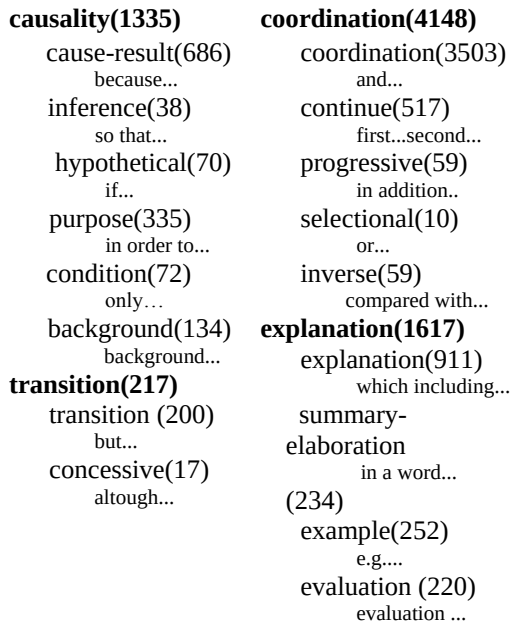

Figure 3: A three-level set of discourse relations

3.5 Nucleus and Satellite

Once discourse units are determined, adjacent spans are linked together via connectives to build a hierarchical structure. As stated above, discourse relations may be either mononuclear or multi-nuclear. A mononuclear relation holds between a nucleus and a satellite unit. Normally, the nucleus usually reflects the intention focus of the discourse and is thus more salient in the discourse structure, while the satellite usually represents supportive information for the nucleus. In comparison, a multi-nuclear relation usually

holds two or more discourse units of equal weight in the discourse structure.

For nucleus determination, we adopt the dependency grammar, and select the unit which can stand for the relationship with other discourse units in a discourse. As shown in Figure 2, on the first level, discourse relation "对此 (to this)" has the latter unit "浦东…法制轨道 (Pudong…as soon as they appear.)" as nucleus and the former unit "浦东…新问题 (Pudong …new problem)" as satellite, since the latter unit agrees with the main purpose of the discourse, which emphasizes some methods for the progress of Pudong. Moreover, since the combination of 4, 5 and 6 has the cause relation with 7, we choose 7 as nucleus because it can stand for the combination of 4, 5, 6 and 7, and has the selection relationship with 3.

4 Chinese Discourse Treebank

Given above the CDT scheme, we choose 500 Xinhua newswire documents from the Chinese Treebank (Xue et al., 2005b) in our Chinese Discourse Treebank (CDTB) annotation. In particular, we annotate one discourse tree for each paragraph.

In this section, we address the key issues with the CDTB annotation, such as annotator training, tagging strategies, corpus quality, along with the statistics of the CDTB corpus.

4.1 Annotator Training

The annotator team consists of a Ph.D. in Chinese linguistics as the supervisor (senior annotator) and four undergraduate students in Chinese linguistics as annotators (two pairs). The annotation is done in four phases. In the first phase, the annotators spend 3 months on learning the principles of CDT and the use of our developed discourse annotation tool. In the second phase, the annotators spend 2 months on independently annotating the same 50 documents (about 260 paraphrases), and another 2 months on cross-checking to resolve the difference and to revise the guidelines. In the third phase, the annotators spend 9 months on annotating the remaining 450 documents. In the final phase, the supervisor spends 3 months carefully proofread all 500 documents.

4.2 Tagging Strategies

In the CDTB annotation, we employ a top-down strategy. That is, we determine the overall level first and then the analysis goes on step by step to the individual EDUs. This strategy is adopted in our annotation tool. The advantages of the top-down strategy are three folds. First, such a strategy can easily grasp the whole discourse structure. This conforms to the global nature of discourse analysis. Second, due to the lack of clear difference between Chinese sentence and phrase structure, such a strategy can largely avoid the error propagation in Chinese EDU segmentation. Since in such a top-down strategy, EDU segmentation becomes an end question, and even if an EDU segmentation error happens, its impact is localized, i.e. with little impact on the whole discourse structure. Our annotation practice shows that such strategy is effective. Third, such a strategy accords with the cognitive of Chinese characteristics, and conforms to the mental process of Chinese discourse understanding (Huang and Liao, 2011). However, we do not exclude the bottom-up strategy. In some cases, on the cognitive psychological process, annotator is combine top-down and bottom-up strategies.

Take Example (3) as an example, an annotator first finds the first level, with the period at the end of sentence 1, and chooses discourse relation (either explicit or implicit), connective, and connective related information (e.g. whether can be added, deleted, and the language sense, etc.), nuclearity etc. Then, the annotator turns to sentence 1 and marks the second comma as level 2 with necessary information annotated, and goes on to sentence 2, recursively, until all EDUs are marked. In this way, a discourse tree with the CDT representation is constructed.

4.3 Quality Assurance

A number of steps are taken to ensure the quality of CDTB. These involve two tasks: checking the validity of the trees and tracking inter-annotator consistency.

4.3.1 Tree validation

We first manually check if a tree has a single root node and compare the tree with the document to check for missing sentence or fragments from the end of text. Then we check the attached information such as connectives, relations and nuclearity in the tree. We also check the tree with a tree traversal program to find the errors undetected by the manual validation process. Finally, all of the trees work successfully.

4.3.2 Consistency

To ensure the quality of CDTB, we adopt the inter-annotator consistency using Agreement and kappa on 60 documents (chtb0041-chtb

0100). Table 1 illustrates the inter-annotator consistency in details.

As shown in Table 1, we measure the agreement of EDU segmentation by determining whether punctuation (all period, comma etc. are considered) is treated as an EDU boundary. It shows that the agreement reaches 91.7% with Cohen's kappa value (Cohen, 1960) 0.91. This justifies the appropriateness of our EDU definition. Explicit or Implicit agreement 94.7% is calculate by the same EDU boundary (intersection) of two annotators. For the same explicit relation, the connective identification agreement is 82.3%, because this is strict measure when two annotators choose the same connective word. If we relax the measure to contain the same word, the agreement can reach 98%. For example, one annotate "也…并(also…and)", and the other annotate "并(and)" is wrong with our strict measure.

	Agreement	Kappa
EDU segmentation	91.7	0.91
Explicit or Implicit	94.7	0.81
Explicit connective identification	82.3	--
Implicit connective insertion	74.6	--
Mononuclear or Multinuclear	80.8	--
Nuclearity	82.4	--
Structure	77.4	--

Table 1: Inter-annotator consistency

It is not surprising that the agreement on implicit connective insertion with the same position and the same connective only reaches 74.6% since for some discourse relations, there may existing several connective alternatives. For example, both "so" and "therefore" can express the same causation relation. If we relax the constraint to the compatible connective, the agreement on implicit connective insertion can reach up to 84.5%.

Finally, it shows that the agreement on overall discourse structure (with the same connectives as non-leaf nodes, the same EDUs as leaf nodes) reaches 77.4%. This justifies the appropriateness of our CDT scheme, given the inherent ambiguity in Chinese discourse structure.

4.4 Corpus Statistics

Currently, the CDTB corpus consists of 500 newswire articles from Chinese Treebank, which are further divided into 2342 paragraphs with a CDT representation for one paragraph.

- For EDUs, CDTB contains 10650 EDUs with an average of 4.5 EDUs per tree. On average, there are 2 EDUs per sentence and 22 Chinese characters per EDU.
- For discourse relations, CDTB contains 7310 relations, of which 1812 are explicit relations (24.8%) and 5498 are implicit relations (75.2%). This indicates that implicit relations occur much more frequently in Chinese than in English, e.g. 75.2% in CDTB (Chinese) vs. ~50% in PDTB (English).
- With the deepest level of 9, most (98.5%) of discourse relations occur in level 1 (2342), level 2(2372), level 3(1532), level 4(712), and level 5(242). It also shows that 3557 (48.7%) relations are mononuclear relations with 2110 nucleus ahead, while the remaining 3754 relations are multi-nuclear. The numbers shown in the parentheses of Figure 3 illustrate the distributions of different relations. In comparison with the top 2 most frequently occurring relations in PDTB (English), i.e. the coordination and explanation relations, there exist 3503 (47.9%) and 911 instances respectively, with regard to the abstract relation set as shown in Figure 3.
- CDTB contains 282 connectives, among which 274 (140 can be deleted) appears as explicit connectives and 44 can be inserted in place of implicit connectives. Table 2 lists the top 10 frequent explicit connectives and implicit connectives.

Explicit connectives		Implicit connectives	
connectives	frequency	connectives	frequency
并(and)	208	因此(so)	368
其中(among them)	154	并(and)	354
也(also)	131	并且(and)	259
而(however)	70	例如(e.g)	140
但(but)	69	来(in order to)	68
还(also)	68	以(in order to)	61
使(so that)	56	然后(then)	55
以(in order to)	52	其中(among them)	48
为(in order to)	49	而(while)	47
同时(meanwhile)	46	因为(because)	32

Table 2: The most frequent connectives in CDTB

5 Comparison with other Discourse Banks

Table 3 compares the difference of CDTB with RST-DT and PDTB from various perspectives, such as EDU, connective, relation, structure and nuclearity.

	RST-DB	PDTB	CDTB
EDU	Clear defined; start of combination; one relation has two or more EDUs	Predicate-argument view; one relation has two arguments	Clear defined from three aspects; end of top-down segmentation; one relation has two or more EDUs
Connective	--	Mark explicit connectives and insert implicit connectives	Mark whether an connective can be deleted without changing the rhetorical relation; insert implicit connective with good intuition and bad intuition differentiated
Relation	Abstract set of relation types; annotate the relation types	Abstract set of relation types; annotate connective and relation type	Represent relation by connective; annotate connective and it's attribute; mapping of connective to the set of discourse relations in a later stage
Structure	Complete tree	Partial tree, deduced by connective and it's argument	Complete tree; top-down segmentation; structure can be represented by the connective hierarchy
Nuclearity	Determined by certain rhetorical relation	--	Determined by the global meaning of a discourse

Table 3: The comparison of RST-DT, PDTB and CDTB

6 Preliminary Experimentation

In order to evaluate the computability of CDTB, we give the experimental results on EDU recognition, which is crucial in discourse parsing. After excluding sentence end punctuations (such as period, question mark, and exclamatory mark), which are certainly EDU boundaries, there remains 7625 punctuations as EDU boundaries (positive instances) and 4876 punctuations as non-EDU boundaries (negative instances). With various features as adopted in Xue and Yang (2011) and Li et al. (2013b), Table 4 shows the performance of EDU recognition on the CDTB corpus with 10-fold cross validation.

Classifier	Gold standard parse			Automatic parse		
	Accuracy	F1(+)	F1(-)	Accuracy	F1(+)	F1(-)
MaxEnt	**90.6**	**91.1**	**90.5**	**89.0**	**90.3**	**87.2**
C45	90.2	90.5	90.1	88.7	90.0	87.7
NiveBayes	90.2	89.9	88.9	88.0	89.0	86.9

Table 4: Performance of EDUs recognition

As shown in Table 4, MaxEnt performs best, with accuracy up to 90.6% on gold standard parse tree, close to human agreement of 91.7%, and with accuracy up to 89% on automatic parse tree. This suggests the appropriateness of our definition of clause as EDU. Table 4 also gives the performance on both positive and negative instances. It shows better F1-measure on recognizing positive instances than negative instances.

7 Conclusions

In this paper, we propose a Connective-driven Dependency Tree (CDT) structure as a representation scheme for Chinese discourse structure. CDT takes advantage of both RST and PDTB, and well adapts to the special characteristics of Chinese discourse. In particular, we describe CDT in detail from various perspectives, such as EDU, connective, structure, relation and nuclearity. Given the CDT scheme, we annotate 500 documents in a top-down segmentation process to keep consistent with Chinese native's cognitive habit. Evaluation of the CDTB corpus on EDU recognition justifies the appropriateness of the CDT scheme to Chinese discourse structure and the usefulness of our CDTB corpus.

In the future work, we will focus on enlarging the scale of the corpus annotation and developing a complete Chinese discourse parser.

Acknowledgments

This research is supported by the Project 2012AA011102 under the National 863 High-Tech Program of China, by the National Natural Science Foundation of China, No.61331011, No.61273320.

The contact author of this paper, according to the meaning given to this role by Soochow University, is Guodong Zhou. The complete corpus is available for research purpose upon request.

Reference

Zheng Cao.1984. *Primary exploration on sentence group*. Zhejiang Education Press, Hangzhou,CN(in Chinese).

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2003. *Building a Discourse-tagged Corpus in the Framework of Rhetorical Structure Theory*. Springer Netherlands.

LiPing Chen. 2006. *English and Chinese discourse structure dimension theory and practice*. Ph.D. thesis, Shanghai international studies university doctoral dissertation.

Liping Chen. 2008. Chinese text structure annotation theory support，*Journal of Nanjing university of aeronautics and astronautics*, 10(3):69-71(in Chinese).

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20 (1): 37–46.

Wenhe Feng and Donghong Ji. 2011. Parallel structure analysis of the coordination structure and the controller status of connective. *Linguistic Sciences*, 2:168-181(in Chinese).

David G Hays. 1964. Dependency theory: formalism and some observations. *Language*, 40(4):511-525.

Hen-Hsen Huang and Hsin-Hsi Chen. 2011. Chinese Discourse Relation Recognition. In *Proceedings of 5th International Joint Conference on Natural Language Process*, pages 1442–1446, Chiang Mai, Thailand, November 2011.

Borong Huang and Xudong Liao. 2011. *Morden Chinese* (volume two, updated 5th edition). Higher Education Press. Beijing,CN(in Chinese).

Yancui Li, Wenhe Feng, and Guodong Zhou. 2013a. Elementary discourse unit in Chinese discourse structure analysis. In *Chinese Lexical Semantics*, pages 186-198, Wuhan, China, Springer Berlin Heidelberg.

Yancui Li, Wenhe Feng, and Guodong Zhou et al. 2013b. Research of Chinese Clause Identification Based on Comma. *Acta Scientiarum Naturalium Universitatis Pekinensis*, 49(1):7-14 (in Chinese with English abstract).

Ziheng Lin, Min-Yan 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*, pages 343–351, Singapore, 6-7 August 2009.

Daniel Marcu. 2000. *The theory and practice of discourse parsing and summarization*. MIT Press.

William Mann and Sandra Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243-281.

Emily Pitler, Annie Louis, and Ani Nenkova. Automatic sense prediction for implicit discourse relations in text. 2009. In *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP*, pages 683–691, Suntec, Singapore, 2-7 August 2009.

Rashmi Prasad, Nikhil Dinesh, and Lee et al. 2008. The Penn Discourse TreeBank 2.0. In *Processdings of Sixth International Conference on Language Resources and Evaluation (LREC)*, pages 2961-2968, Marrakech, Morocco.

Susan Verberne, Lou Boves, Nelleke Oostdijk, and Peter-Arno Coppen. 2007. Discourse-based answering of why-questions. *Traitement Automatique des Langues, special issue on Computational Approaches to Discourse and Document Processing*, 47(2):21-41.

Wenge Wang. 2010. The Current Research Situation of the Clause in Modern Chinese. *Chinese Language Learning*, (1): 67–76(in Chinese).

Fuyi Xing. 2003. *Research of Chinese complex sentence*. The Commercial Press, Beijing, CN (in Chinese).

Nianwen Xue. 2005a. Annotating the Discourse Connectives in the Chinese Treebank. In *Proceedings of the ACL Workshop on Frontiers in Corpus Annotation*, pages 84-91, Ann Arbor, Michigan.

Nianwen Xue, Fei Xia, Fu-Dong Chiou, and Martha Palmer. 2005b. The Penn Chinese Treebank: Phrase Structure Annotation of a Large Corpus. *Natural Language Engineering*, 11(2): 207-238.

Nianwen Xue and Yaqin Yang. 2011. Chinese sentence segmentation as comma classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:Human Language Technologies*, pages 631-635, Portland, Oregon,USA, June 2011.

Ming Yue. 2008. Rhetorical Structure Annotation of Chinese News Commentaries. *Journal of Chinese Information Processing*, 22(4): 19-23(in Chinese with English abstract).

Yuping Zhou and Nianwen Xue. 2012. PDTB-style discourse annotation of Chinese text. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 69-77, Jeju, Republic of Korea, 8-14 July 2012.