The CUHK Discourse TreeBank for Chinese: Annotating Explicit Discourse Connectives for the Chinese TreeBank

Lanjun Zhou[1], Binyang Li[1,4], Zhongyu Wei[1], Kam-Fai Wong[1,2,3]

[1] Department of SEEM, The Chinese University of Hong Kong
[2] MoE Key Laboratory of High Confidence Software Technologies, China
[3] Shenzhen Research Institute, The Chinese University of Hong Kong
[4] University of International Relations, Beijing, China
{ljzhou,byli,zywei,kfwong}@se.cuhk.edu.hk

Abstract

The lack of open discourse corpus for Chinese brings limitations for many natural language processing tasks. In this work, we present the first open discourse treebank for Chinese, namely, the Discourse Treebank for Chinese (DTBC). At the current stage, we annotated explicit intra-sentence discourse connectives, their corresponding arguments and senses for all 890 documents of the Chinese Treebank 5. We started by analysing the characteristics of discourse annotation for Chinese, adapted the annotation scheme of Penn Discourse Treebank 2 (PDTB2) to Chinese language while maintaining the compatibility as far as possible. We made adjustments to 3 essential aspects according to the previous study of Chinese linguistics. They are sense hierarchy, argument scope and semantics of arguments. Agreement study showed that our annotation scheme could achieve highly reliable results.

Keywords: Discourse Annotation, Chinese Discourse, Explicit Discourse Connectives

1. Introduction

Discourse analysis raises issues about semantics, and especially the nature of the coherence and cohesion of texts. As to part-of-speech tagging and syntactic parsing, discourse analysis is one of the fundamental unsolved problems in computational linguistics. Recently, more and more research proves that discourse information is crucial for many natural language processing tasks. For instance, automatic summarization (Spärck Jones, 2007), text generation (McKeown, 1992), sentence compression (Sporleder and Lapata, 2005), information extraction (Patwardhan and Riloff, 2007), sentiment analysis (Somasundaran et al., 2009; Zhou et al., 2011), paraphrasing (Regneri and Wang, 2012) and question answering (Verberne et al., 2007) etc.

Currently, there are mainly two framework for discourse annotation, the Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) and the Penn Discourse Treebank 2 (PDTB2) (Prasad et al., 2008a). RST was designed for text generation initially. It provides an explanation of coherence for carefully prepared texts. RST defines a set of rhetorical relations between elementary discourse units (EDUs). In addition, RST defines the semantics of each EDU by grouping the relations into 3 categories: presentational relations, subject matter relations and multinuclear relations. PDTB2 is lexically grounded by assuming discourse relations are anchored by explicit or implicit discourse connectives. Carlson et al. (2003) reported that the inter-annotator agreement for relation identification in RST-Discourse Treebank was less than 80%. The agreement could be higher (less than 83%) by grouping similar relations. For PDTB, Miltsakaki et al. (2004) showed that over 90% of overall agreement for explicit connectives was achieved.

Although similar work for Chinese has been reported in Xue (2005), Huang and Chen (2011) and Zhou and Xue (2012), their data are still not publicly available. In this work, we present the first open discourse treebank for Chinese – the Discourse Treebank for Chinese (DTBC). DTBC aims to annotate discourse relations for the Chinese Treebank 5 (CTB5) (Xue et al., 2005). The original DTBC corpus which followed the PDTB annotation scheme was introduced in Zhou et al. (2012). In this paper, we refine the annotation scheme according to the characteristics of Chinese and re-annotate all the data. Huang and Chen (2011) constructed a Chinese discourse corpus with 81 articles. They adopted the top level senses from PDTB sense hierarchy and focused on the annotation of inter-sentential discourse relations. Their annotation results were seriously imbalanced (Over 85% of the annotated relations were EXPANSION). Since we are dealing with similar genre of text (i.e., Chinese news reports), similar results could be expected for inter-sentential relations. Because intra-sentential discourse information is complementary to the inter-sentential discourse information, we focused on the annotation of intra-sentential discourse relations in the current version of Discourse Treebank for Chinese (DTBC). In addition to the similarities and differences between English and Chinese discussed in the previous work (Zhou and Xue, 2012), we further modify/add 3 necessary aspects of the PDTB2 annotation scheme for Chinese (see Section 3 for details):

Sense hierarchy. We added 3 *type* level senses (i.e., CONTINGENCY.Inference, CONTINGENCY.Purpose and EXPANSION.Background) and 2 *subtype* level senses (i.e., EXPANSION.Conjunction.parallel and EXPANSION.Conjunction.progressive) while keeping the compatibility to PDTB2.

Argument scope. The frequently appearing structure "NP 的 (DE) VP" are regarded as nominalizations. Accordingly, "NP 的 (DE) VP" can be annotated as arguments.

Nucleus/Satellite of arguments. We adopt a consistent definition of `Arg1` and `Arg2` on *type* level senses without losing any information comparing to PDTB2. Furthermore, inspired by the idea of *nucleus* and *satellite* of RST, we define the semantics of each argument in our sense hierarchy.

According to the annotation results, highly reliable results could be achieved by adopting our annotation scheme. Furthermore, although parallel connectives (e.g., '*因为 (because)... 所以 (as a result)...*') were intuitively very common in Chinese, we found that only about 15% of the annotated relations were anchored by parallel connectives. This observation was very different from the work of Zhou and Xue (2012).

The remainder of this paper is organized as follows: Section 2 gives the related work and illustrate the differences between this work and the previous work. Section 3 describes the annotation scheme of DTBC in detail. Section 4 presents the agreement study and annotation statistics. Section 5 concludes this paper.

2. Related Work

Most of the discourse annotation work were based on the Treebanks (e.g., Penn Treebank II (Marcus et al., 1993)) because syntactic information was proven critical in recognizing both intra- and inter-sentential discourse relations (Soricut and Marcu, 2003; Duverle and Prendinger, 2009).

For English, there are mainly two corpora: (1) RST Discourse Treebank (RST-DT) (Carlson et al., 2003) following the RST framework and (2) Penn Discourse Treebank (PDTB) (Miltsakaki et al., 2004; Miltsakaki et al., 2008; Prasad et al., 2008a) utilizing a lexically-grounded approach. Based on the RST framework, corpora for other languages such as Spanish (da Cunha et al., 2011), Hindi (Prasad et al., 2008b), etc. were annotated. Based on the PDTB scheme, annotated data for Modern Standard Arabic (Al-Saif and Markert, 2010), Czech (Mladová et al., 2008), Turkish (Zeyrek and Webber, 2008; Zeyrek et al., 2009; Zeyrek et al., 2010), etc. were developed.

As far as we know, there is no open discourse treebank in Chinese. Xue proposed the Chinese Discourse Treebank (CDTB) Project (Xue, 2005). In their work, the issues arisen from their annotation work such as characteristics of Chinese discourse connectives, definition of arguments, scope of arguments and sense disambiguation were discussed and they argued that determining the argument scope was the most challenging part of the annotation. Their work did not include the adaptation of sense hierarchy and detailed semantic definition of arguments. But even more important, their annotated corpus was never published.

Huang and Chen (2011) constructed a Chinese discourse corpus with 81 articles. They adopted the top level senses from PDTB sense hierarchy and focused on the annotation of inter-sentential discourse relations. The annotation results were seriously imbalanced. For instance, about 85% of the annotated relations were PDTB.EXPANSION while only 3% of them were PDTB.CONTINGENCY. Since we are dealing with similar genre of text (i.e., Chinese news reports) in this work, similar results could be expected for inter-sentential relations. Because intra-sentential discourse information is complementary to the inter-sentential discourse information, we focus on the annotation of intra-sentential discourse relations in the current version of Discourse Treebank for Chinese (DTBC). Annotation study of DTBC (See Chapter 5, Section 5.3) showed that although PDTB.EXPANSION still accounted for the largest proportion (57%), each of the other discourse relations accounted for more than 10% of the annotated relations. Accordingly, there were great differences in the distributions of inter-sentential and intra-sentential discourse relations.

Zhou and Xue (2012) presented a PDTB-style discourse corpus for Chinese. They also discussed the key characteristics of Chinese text which differs from English, e.g., the parallel connectives, comma-delimited intra-sentential implicit relations etc. Their data set contains 98 documents from the Chinese Treebank (Xue et al., 2005) with annotated explicit and implicit relations. However, their data was relatively small in size and publicly unavailable.

Although the above work for Chinese has been reported, their data sets are all relatively small comparing to the annotation work on other languages. Their data are all not publicly available. Furthermore, the previous work ignored three key problems during the annotation work: (1) The adaptation of the sense hierarchy from PDTB to DTBC; (2) Annotating "NP 的 (DE) VP" structure that appears very frequently in Chinese when annotating arguments. (3) Defining the semantics of arguments by introducing the idea of *nucleus/satellite* of RST to DTBC. We will show how we deal with these problems in this work.

3. Annotation Scheme

The annotation scheme adopted in this work followed the settings of PDTB2 as far as possible for compatibility between DTBC and PDTB2. Previous work (Xue, 2005; Zhou and Xue, 2012) discussed the linguistic characteristics of Chinese which may affect the annotation process, we followed their observations and made additional but essential modifications in our annotation work. Specifically, we followed the definition of compound sentences based on the previous study on Chinese linguistics (Xing, 2000; Wang et al., 2006). We modified the sense hierarchy of PDTB2. In addition, we regard the frequently appeared structure "NP 的 (DE) VP" as nominalizations in Chinese. Moreover, we defined the semantics of arguments on *type* level of the sense hierarchy and also integrated the idea of *nucleus* and *satellite* defined in RST (Mann and Thompson, 1988) to our annotation scheme. The semantics of arguments are useful information for many natural language processing tasks (automatic summarization (Spärck Jones, 2007), sentiment analysis (Zhou et al., 2011) etc.) which consider the importance of the different segments of texts.

3.1. Discourse Connectives

Explicit connectives in PDTB2 mainly includes subordinating conjunctions, coordinating conjunctions and adverbials (ADVP and PP) with some special cases (e.g., modified connectives, parallel connectives etc.). A subordinating conjunction joins a subordinate clause to a main clause while a coordinating conjunction joins clauses with equal emphasis. Similar settings could be applied to Chinese.

Xue (2005) gave detailed examples to show the characteristics of Chinese discourse connectives and their senses. Zhou and Xue (2012) argued that one of the key differences between English and Chinese was that parallel connectives was pre-dominant in Chinese (e.g., '*因为 (because)... 所以 (as a result)...*', '*虽然 (although)... 但是 (but)...*' etc.). According to our annotation result (See Section 5), about 15% of the annotated discourse relations were triggered by parallel connectives. If we discontinuously annotated the parallel connectives as PDTB2 did, it would result in large number of repetitions. Hence, the parallel connectives were also annotated continuously in DTBC as Zhou and Xue (2012) did.

Another important characteristic of Chinese is that, the constitution of parallel connectives is very flexible. Parallel connectives could be composed of (1) a subordinating conjunction and a coordinating conjunction. For example '*虽然 (although)... 但是 (but)...*', '*虽然 (although)*' is a subordinating conjunction while '*但是 (but)*' is a coordinating conjunction; (2) Paired subordinate conjunctions; (3) Paired coordinate conjunctions. Furthermore, each part of the parallel connectives could be a constitution of multiple nonadjacent words which would result in multiple repetitions if they are not annotated continuously. For instance '在... 的同时... 又...' (when) etc.

Interestingly, the semantic characteristics of parallel connectives match the definition of *nucleus*&*satellite* in RST (Mann and Thompson, 1988) well. For instance, '*虽然 (although)*' indicates the *satellite* and '*但是 (but)*' signals the *nucleus* segment in parallel connective '*虽然 (although)... 但是 (but)...*'. Hence, for discourse relations triggered by parallel connectives, the semantic type of arguments could usually be identified easily. The details will be discussed in Section 3.3.

3.2. Senses Hierarchy

To be compatible with PDTB2 as far as possible, we adopted the same top level semantic classes: "TEMPORAL", "CONTINGENCY", "COMPARISON" and "EXPANSION". However, we made modifications for Chinese on the *type* level and *subtype* level according to the previous study of Xing (2000) and Wang et al. (2006). In the current stage, we do not distinguish the pragmatic using of connectives and some of the *subtypes* (e.g., the *subtypes* of COMPARISON.Concession).

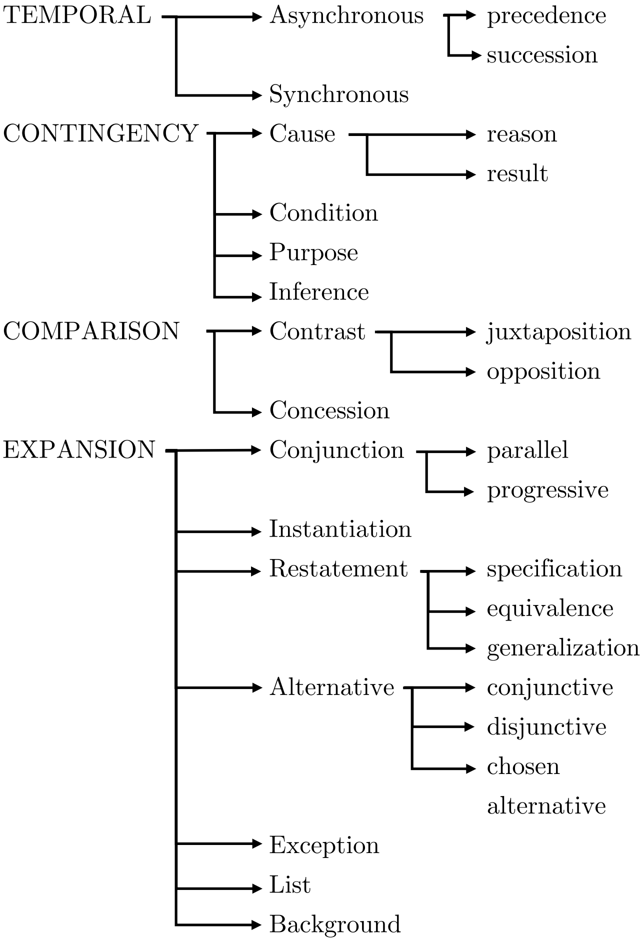

Figure 1: Sense Hierarchy of DTBC

We made modifications on the PDTB2 sense hierarchy based on the previous study on Chinese linguistics (Xing, 2000). Specifically, we added 3 *type* level senses (i.e., CONTINGENCY.Inference, CONTINGENCY.Purpose and EXPANSION.Background) and 2 *subtype* level senses (i.e., EXPANSION.Conjunction.parallel and EXPANSION.Conjunction.progressive) (See Figure 1). Actually, the newly added *subtype* level senses are defined as *type* level senses in Xing (2000). However, to keep the compatibility with PDTB, we define these senses to be the subtypes of EXPANSION.Conjunction.

CONTINGENCY.**Inference**: This *type* is used when `Arg1` provides the premises and `Arg2` expresses the conclusion based on the factual information of `Arg1`. Unlike CONTINGENCY.Cause, CONTINGENCY.Inference emphasize the justification process from `Arg1` to `Arg2` is fact-based. CONTINGENCY.Inference is very similar to the CONTINGENCY.pragmatic cause of PDTB2 except that there are specific connectives in Chinese anchoring this relation. Typical connectives of this sense are " 既然 (since)... 就 (hence)...", "... 可见 (as a result)..." etc. For example,

(2) **既然** [土耳其 在 塞岛 北部 保持着
since Turkey at Cyprus island north maintain
强大 的 军事 力量]$_{Arg1}$，[南部 希族人
strong DE military power , south Hellenes
加强 自己 的 防务 也 是 理所当然]$_{Arg2}$ 。
reinforce self 's defence also is of course .
(chtb_0812)

"Since Turkey maintains a strong military power at the north of Cyprus island, it is natural that the southern Hellenes reinforce their military power."

CONTINGENCY.**Purpose**: This *type* is used when Arg2 gives the intended situation and Arg1 provides the purpose of Arg2. Typical connectives of this sense are "... 以便 (in order to)...", "... 以免 (in order not to)...", "... 为了 (in order to)..." etc. For example,

(3) [进出口银行 决定 先 在
The Export-Import Bank of China decide first at
日本 取得 信用 评级 是 为 进入
Japan obtain credit rating is for enter
国际 资本 市场 融资 创造 作
international capital market financing create do
准备]$_{Arg1}$ ，**以便** [扩大 资金 来源 ，
preparation, in order to expand funds source,
支持 中国 机电 产品 和
support Chinese machinery products and
成套 设备 出口]$_{Arg2}$。
complete sets equipment export .
(chtb_0010)

"In order to expand the source of funds and support the export of Chinese machinery products and complete sets of equipment, the Export-Import Bank of China decides to obtain credit rating in Japan first so as to prepare for entering the international capital market."

EXPANSION.**Background**: This *type* is used when Arg1 gives the background information of Arg2. In RST, it is defined as: Arg1 increases the ability of reader (hearer) to comprehend an element in Arg2. Typical connectives of this sense are " 随着 (with)...", " 在 (in)... 下...(below)" etc. It appears frequently in CTB5 texts. For example,

(4) 据 了解 ，近 几 年 ，**随着**
According to reports , recent several years , **with**
[中国 经济 的 不断 发展 和
China economy DE continuous development and
对外开放 的 不断 深入]$_{Arg1}$ ，
opening up DE continuous deep ,
[外商 来 华 投资 热情 很
foreign investor come China invest passion very
高 ，投资 项目 和 金额 增长
high , investment project and funds increase
十分 迅速]$_{Arg2}$ 。
very quickly .
(chtb_0006)

"According to reports, in recent years, with the continuous development of economy and wider opening up, the passion of the foreign investor is very high and the number of projects as well as investment funds increase rapidly"

EXPANSION.conjunction.**parallel**: This subtype applies when Arg1 and Arg2 are parallel connected with the equal emphasis except the ways described by EXPANSION.alternative. Typical connectives of this sense is " 既 (as well as)... 又/也 (as well)...", " 一边 (while)... 一边 (while)...", " 又 (as well)... 又 (as well)..." etc. For example,

(5) 大家 **既** [为 他们 惋惜]$_{Arg1}$ ， **又**
Everyone **as well as** for them pity , **as well**
[觉得 这 是 一个 非常 值得 忧虑 的
feel this is a very deserve worried DE
问题]$_{Arg2}$ 。
problem .
(chtb_0206)

"Every one feels pity for them, and thinks that this is a problem worthy of concern as well."

EXPANSION.conjunction.**progressive**: This subtype applies when Arg1 and Arg2 are parallel connected with the different emphasis. We set the argument emphasized more than the other as Arg2 in this work. Typical connectives of this sense is " 不但 (not only)... 而且 (but also, furthermore)...", " 尚且 (even)... 何况 (let alone)...", "... 更 (even more)..." etc. For example,

(6) 关于 香港 回归 中国 后 的
About Hong Kong return China after DE
国际 金融 地位 问题 ，
international finance position question ,
戴相龙 强调 ，香港 的
Dai Xianglong emphasize , Hong Kong 's
国际 金融 地位 **不但** [能够
international finance position **not only** able
维持]$_{Arg1}$ ， **而且** [还 会 得到
maintain , **but also** also will obtain
加强]$_{Arg2}$ 。
strengthened .
(chtb_0900)

"Regarding to the question of international finance position of Hong Kong after returning to China, Dai Xianglong emphasized, that the international financial position of Hong Kong would not only be able to maintain, but would also be strengthened."

3.3. Arguments

In PDTB2, the annotated arguments should be clauses by principle while some special cases (e.g., *VP coordinations*, *nominalizations* and *anaphoric expressions*

	`Arg1`	`Arg2`
CONTINGENCY.Cause.reason	result	cause
CONTINGENCY.Cause.result	cause	result

Table 1: The semantics of `Arg1` and `Arg2` for CONTINGENCY.Cause.reason and CONTINGENCY.Cause.result in PDTB2. DTBC defines the semantics of arguments on *type* level instead of on *subtype* level as PDTB2 did.

etc.) were also annotated. For judging compound sentences and sentence constituents, we follow the settings of widely accepted *Modern Chinese* 现代汉语（重排本）(Wang et al., 2006). Previous work(Xue, 2005; Zhou and Xue, 2012) studied the similarities and differences of discourse annotation between English and Chinese. However, they missed a very important structure in Chinese, namely, nominalizations which appears very frequently in Chinese texts.

Note that judgement of nominalizations is still a highly controversial issue in Chinese. For instance, one of the most basic problem in Chinese linguistics is how to analysis "VP" in "NP 的 (DE) VP" structure (See Examples 7, 8). As almost all verb in Chinese could be used in this structure, more over this structure appears very common in Chinese texts. Unfortunately, this problem is a highly controversial issue in Chinese linguistics. To simplify the problem of judging nominalizations and avoid disputes, we set "VP"s in the "NP 的 (DE) VP" structure as nominalizations. Recall example (4), we annotated `Arg1` since it consists of two "NP 的 (DE) VP" structures.

(7) 生活 水平 的 提高
Living standard DE improve

(8) 这本 书 的 出版
This book DE publish

The semantic definitions of `Arg1` and `Arg2` in PDTB2 are sometimes tricky. For example, refer to Table 1, in PDTB2, `Arg1` describes the result based on the cause of `Arg2` in case 1. But in case 2, `Arg1` of CONTINGENCY.Cause.result describes the cause and `Arg2` gies the result. We found that a consistent definition of `Arg1` and `Arg2` on *type* level will simplify the annotation process without losing any information. Table 2 gives a brief summary of the semantics of arguments in DTBC. It is worthy noting that we also define the *nucleus* of each sense based on RST.

3.4. Annotation Process

The objective of DTBC is to add discourse layer to all 890 documents (8,771 sentences) in CTB5. The annotation work was shared by two trained Chinese native speakers. However, in consideration of the difficulty during the annotation process, the annotators were asked to annotate the first 1000 (about 100 documents) sentences to get familiar with the annotation scheme and the agreement study. The rest of the sentences (7,771 sentences) will be annotated by only one of the annotators. We developed a web-based annotation tool (DTBC Annotation Tool) can be accessed anywhere and monitor the annotation process anytime.

4. Results

Currently, we have already finished the annotation work for all 890 documents from CTB5. The size of DTBC is much larger than the previous work on Chinese (Huang and Chen (2011) annotated 81 documents; Zhou and Xue (2012) annotated 98 documents). Based on the DTBC corpus, we can then analyze the basic characteristics of intra-sentential discourse for Chinese.

4.1. Agreement Study

The result of inter-annotator agreement is shown in Table 3.The agreement study is carried out for the first 1,000 sentences of DTBC. According to our connective lexicon, there are 1,717 potential connectives in the first 1,000 sentences. Since we only annotated intra-sentence discourse connectives at the current stage, the inter-sentential discourse connectives are considered as non-discourse connectives. The inter-annotator agreement for discourse usage and sense annotation is more than 90% ($kappa \geq 0.80$) which is highly reliable. The disagreements in sense annotation mainly come from connectives annotated by only one annotator or *ambiguous* connectives (e.g., "同时", " 而" etc.).

The agreement on argument order is almost 1.0 ($kappa = 0.98$). That means it is often easy for the annotators to identify the semantics of `Arg1` and `Arg2` if the senses are already determined. However, as pointed out in (Xue, 2005), the scope of arguments to a discourse connective are the most challenging part during the annotation work. According to the results, the agreement would be lower than 0.60 if the annotated argument scopes from the two annotators were fully matched. As a result, we relaxed the requirements of full match to partial match with overlap proportions (OP). Refer to Table 3, the agreement are 0.73 and 0.61 for *lenient* and *strict*, respectively.

4.2. Corpus statistics

Refer to Table 4, the appearance of EXPANSION in DTBC is pre-dominant as the overwhelming majority of the CTB5 documents are news reports. 52% of the annotated senses belongs to EXPANSION comparing to only 33% in PDTB2. On the other hand, the appearance of TEMPORAL and COMPARISON in DTBC is significantly less than in PDTB2.

We are also interested in which *type* of senses appears most frequently in DTBC. Table 4 shows that, EXPANSION.Conjunction accounts for 37% of all senses annotated while the other 4 in top 5 accounts for about 10%. Note that the top 5 most frequent senses accounts for 72% of all senses annotated. If we expand top-5 to top-10, the proportion increases to 96%. Hence, the appearance of senses other than top-10 is

	`Arg1` / Span	`Arg2` / Other Span	*nucleus*
TEMPORAL.Asynchronous	situation happens firstly	situation happens secondly	Both
TEMPORAL.Synchronous	temporal overlapping situation 1	temporal overlapping situation 2	Both
CONTINGENCY.Cause	reason	result	`Arg2`
CONTINGENCY.Condition	condition	consequence	`Arg2`
CONTINGENCY.Purpose	purpose	intended situation	`Arg2`
CONTINGENCY.Inference	premise	conclusion	`Arg2`
COMPARISON.Contrast	one alternate with shared property	the other alternate with shared property	Both
COMPARISON.Concession	inconsistent situation affirmed by author	situation affirmed by author	`Arg2`
EXPANSION.Conjunction	an item	another item	Both
EXPANSION.Instantiation	a situation	instances to describ the situation	`Arg1`
EXPANSION.Restatement	a situation	a reexpression of the situ -ation	`Arg1`
EXPANSION.Alternative	one alternate	the other alternate	Both
EXPANSION.Exception	An general situation	an exception of `Arg1`	`Arg1`
EXPANSION.List	an item	an next item	Both
EXPANSION.Background	text for facilitating understanding	text whose understanding is being facilitated	`Arg2`

Table 2: The semantics of `Arg1` and `Arg2` in DTBC. 'Both' means the sense is a *multinuclear* discourse relation according to Mann and Thompson (1988)

Discourse usage (1,717 potential connectives)	
Agreement	0.97
Kappa	0.88
Senses (274 senses, *class* level)	
Agreement	0.91
Kappa	0.84
Argument order	
Agreement	0.99
Kappa	0.98
Argument scope - *lenient* ($OP \geq 50\%$)	
Agreement	0.73
Kappa	0.63
Argument scope - *strict* ($OP \geq 80\%$)	
Agreement	0.61
Kappa	0.47

Table 3: Inter-annotator agreement study for DTBC on *class* level. In the evaluation of argument scope, *lenient* means the overlap proportion $OP \geq 50\%$ while strict means the overlap proportion $OP \geq 80\%$ for both `Arg1` and `Arg2`

	DTBC	PDTB2
TEMPORAL	14%	19%
CONTINGENCY	23%	19%
COMPARISON	11%	29%
EXPANSION	52%	33%

Table 4: Distributions of *class* level senses in the first 400 documents of DTBC and PDTB2.

EXPANSION.Conjunction	37%
CONTINGENCY.Cause	10%
TEMPORAL.Synchronous	9%
EXPANSION.Specification	8%
CONTINGENCY.Purpose	8%
COMPARISON.Contrast	6%
TEMPORAL.Asynchronous	5%
CONTINGENCY.Condition	5%
COMPARISON.Concession	5%
EXPANSION.Background	3%
Top-10	96%

Table 5: Top-10 *type* level senses in the first 400 documents of DTBC. The percentages of the relations are calculated based on the overall annotation.

very rare in DTBC. This result provides important statistics for discourse analysis on DTBC.

5. Conclusions and Future Work

In this work, we have presented a practical discourse annotation scheme and the first open discourse treebank for Chinese (Discourse Treebank for Chinese). The annotation scheme followed the basic principles of PDTB2 as far as possible and at the same time integrates the characteristics of Chinese language. Specifically, we modified the sense hierarchy, improved the definition of argument scope and semantically defined the *nucleus*/*satellite* of arguments based on RST. More importantly, the scheme of DTBC is reliable during the annotation process and also compatible with PDTB2 on *class* level and most part of *type* level. The annotation results showed that the inter-annotator agreement were over 90% on discourse connective identification and over 85% on sense annotation. DTBC will be an invaluable linguistic resource for future research in Chinese discourse.

In the future, we are planning to (1) annotate the inter-sentence level discourse relations for DTBC; (2) include other literary form (e.g., essays and fictions etc.) to the annotation set; (3) Exploring supervised methods for Chinese discourse classification.

6. Acknowledgements

This research is partially supported by General Research Fund of Hong Kong (No. 417112) and Shenzhen Fundamental Research Program (JCYJ20130401172046450)

7. References

Amal Al-Saif and Katja Markert. 2010. The leeds arabic discourse treebank: Annotating discourse connectives for arabic. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 2046–2053, Valletta, Malta, may. European Language Resources Association (ELRA).

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2003. Building a discourse-tagged corpus in the framework of rhetorical structure theory. *Current and New Directions in Discourse and Dialogue*, pages 85–112.

Iria da Cunha, Juan Manuel Torres Moreno, and Gerardo Sierra. 2011. On the development of the rst spanish treebank. In *Linguistic Annotation Workshop*, pages 1–10.

D.A. Duverle and H. Prendinger. 2009. A novel discourse parser based on support vector machine classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2*, pages 665–673. Association for Computational Linguistics.

Hen-Hsen Huang and Hsin-Hsi Chen. 2011. Chinese discourse relation recognition. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 1442–1446, Chiang Mai, Thailand, November. Asian Federation of Natural Language Processing.

W.C. Mann and S.A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330.

K.R. McKeown. 1992. *Text generation: using discourse strategies and focus constraints to generate natural language text.* Cambridge Univ Pr.

E. Miltsakaki, R. Prasad, A. Joshi, and B. Webber. 2004. The penn discourse treebank. In *Proceedings of the 4th International Conference on Language Resources and Evaluation.* Citeseer.

Eleni Miltsakaki, Livio Robaldo, Alan Lee, and Aravind Joshi. 2008. Sense annotation in the penn discourse treebank. In *Computational Linguistics and Intelligent Text Processing*, pages 275–286. Springer.

Lucie Mladová, Sarka Zikanova, and Eva Hajicová. 2008. From sentence to discourse: Building an annotation scheme for discourse based on prague dependency treebank. In *LREC*.

Siddharth Patwardhan and Ellen Riloff. 2007. Effective information extraction with semantic affinity patterns and relevant regions. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 717–727, Prague, Czech Republic, June. Association for Computational Linguistics.

R. Prasad, N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi, and B. Webber. 2008a. The penn discourse treebank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, pages 2961–2968. Citeseer.

R. Prasad, S. Husain, D.M. Sharma, and A. Joshi. 2008b. Towards an annotated corpus of discourse relations in hindi. *Proceedings of IJCNLP-2008.*

Michaela Regneri and Rui Wang. 2012. Using discourse information for paraphrase extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 916–927. Association for Computational Linguistics.

S. Somasundaran, G. Namata, J. Wiebe, and L. Getoor. 2009. Supervised and unsupervised methods in employing discourse relations for improving opinion polarity classification. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 170–179. Association for Computational Linguistics.

R. Soricut and D. Marcu. 2003. Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 149–156. Association for Computational Linguistics.

K. Spärck Jones. 2007. Automatic summarising: The state of the art. *Information Processing & Management*, 43(6):1449–1481.

Caroline Sporleder and Mirella Lapata. 2005. Discourse chunking and its application to sentence compression. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 257–264. Association for Computational Linguistics.

Suzan Verberne, Lou Boves, Nelleke Oostdijk, and Peter-Arno Coppen. 2007. Evaluating discourse-based answer extraction for why-question answering.

In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 735–736. ACM.

Lijia Wang, Jianming Lu, Huaiqing Fu, Zhen Ma, Peicheng Su, Dexi Zhu, and Tao Lin. 2006. *Modern Chinese(现代汉语)*. The Commercial Press (商务印书馆), China.

Fuyi Xing. 2000. *Research of compound sentences for Chinese(汉语复句研究)*. The Commercial Press(商务印书馆), China.

N. Xue, F. Xia, F.D. Chiou, and M. Palmer. 2005. The penn chinese treebank: Phrase structure annotation of a large corpus. *Natural Language Engineering*, 11(02):207–238.

N. Xue. 2005. Annotating discourse connectives in the chinese treebank. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, pages 84–91. Association for Computational Linguistics.

Deniz Zeyrek and Bonnie L Webber. 2008. A discourse resource for turkish: Annotating discourse connectives in the metu corpus. In *IJCNLP*, pages 65–72. Association for Computational Linguistics.

Deniz Zeyrek, Ümit Turan, Cem Bozsahin, Ruket Çakici, Ayişği Sevdik-Çalli, İşin Demirşahin, Berfin Aktaş, Ihsan Yalçinkaya, and Hale Ögel. 2009. Annotating subordinators in the turkish discourse bank. In *Proceedings of the Third Linguistic Annotation Workshop*, pages 44–47. Association for Computational Linguistics.

Deniz Zeyrek, Işin Demirşahin, Ayişiği Sevdik-Çalli, Hale Ögel Balaban, İhsan Yalçinkaya, and Ümit Deniz Turan. 2010. The annotation scheme of the turkish discourse bank and an evaluation of inconsistent annotations. In *Proceedings of the Fourth Linguistic Annotation Workshop*, pages 282–289. Association for Computational Linguistics.

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 (Volume 1: Long Papers)*, pages 69–77, Jeju Island, Korea, July. Association for Computational Linguistics.

L. Zhou, B. Li, W. Gao, Z. Wei, and K.F. Wong. 2011. Unsupervised discovery of discourse relations for eliminating intra-sentence polarity ambiguities. In *Proceedings of the 2011 Conference on Empirical methods in natural language processing*, pages 162–171. Association for Computational Linguistics.

Lanjun Zhou, Wei Gao, Binyang Li, Zhongyu Wei, and Kam-Fai Wong. 2012. Cross-lingual identification of ambiguous discourse connectives for resource-poor language. In *Proceedings of COLING 2012: Posters*, pages 1409–1418, Mumbai, India, December. The COLING 2012 Organizing Committee.