Sequential Annotation and Chunking of Chinese Discourse Structure

Frances Yung **Kevin Duh** **Yuji Matsumoto**
Nara Institute of Science and Technology
8916-5 Takayama, Ikoma, Nara, 630-0192 Japan
{pikyufrances-y, kevinduh, matsu}@is.naist.jp

Abstract

We propose a linguistically driven approach to represent discourse relations in Chinese text as *sequences*. We observe that certain surface characteristics of Chinese texts, such as the order of clauses, are overt markers of discourse structures, yet existing annotation proposals adapted from formalism constructed for English do not fully incorporate these characteristics. We present an annotated resource consisting of 325 articles in the Chinese Treebank. In addition, using this annotation, we introduce a discourse chunker based on a cascade of classifiers and report 70% top-level discourse sense accuracy.

1 Introduction

Discourse relations refer to the relations between units of text at document level. As a key for language processing, they are used in tasks such as automatic summerization, sentiment analysis and text coherence assessment (Lin et al., 2011; Trivedi and Eisenstein, 2013; Yoshida et al., 2014). While discourse-annotated English resources are available, resources in other languages are limited. In this work, we present the linguistic motivation behind the Chinese discourse annotated corpus we constructed, and preliminary experiments on discourse chunking of Chinese.

1.1 Related Work

Major discourse annotated resources in English include the RST Treebank (Carlson et al., 2001) and the Penn Discourse Treebank (PDTB) (Prasad et al., 2008). The RST Treebank represents discourse relations in a tree structure, where a *satellite* text span is related to a *nucleus* text span. On the other hand, the Penn Discourse Treebank represents discourse structure in a predicate-argument-like structure, where discourse connectives (DCs) relates two text spans (*Arg1* and *Arg2*). Under this framerodk, covert discourse relations are represented by implicit DCs.

PDTB's annotation scheme is adapted by the recently released Chinese Discourse Treebank (CDTB) (Zhou and Xue, 2015). Other efforts to exploit Chinese discourse relations include cross-lingual annotation projection based on machine translation or word-aligned parallel corpus (Zhou et al., 2012; Li et al., 2014). Combinition of the RST and PDTB formalisms is also proposed. Zhou et al. (2014) adds the distinction of *satellite* and *nucleus* to PDTB-style annotation, and Li et al. (2014b) labels the connectives in an RST tree.

1.2 Motivation

Interpretation of discourse relations, as of other linguistic structures, is subject to the surface form of the text. We notice that Chinese discourse structures are expressed by certain surface features that do not exist in English.

First of all, Chinese sentences are sequences of clauses, typically separated by punctuations. Each clause can be considered a discourse argument. Above the clause level, Chinese sentences (marked by '。') are also units of discourse (Chu, 1998). When presented with texts where periods and commas are removed, native Chinese speakers disagree with where to restore them (Bittner, 2013). The actual sentence segmentation of the text thus represents the spans of discourse arguments intended by the writer and should be taken into account.

Secondly, it is well known that syntactical structure is presented by word order in Chinese - so is

Proceedings of the Eighth SIGHAN Workshop on Chinese Language Processing (SIGHAN-8), pages 1–6,
Beijing, China, July 30-31, 2015.

discourse. While the *Arg1* can occur before or after *Arg2* in English, arguments predominantly occur in fixed order in Chinese, depending on the logical relation. For example, the same concession relation can be expressed by both constructions (1) and (2) in English, but only construction (1) is acceptable in Chinese.

1. 虽然 (*suiran*, although) *Arg2* , *Arg1* .
2. *Arg1* , 虽然 (*suiran* ,although) *Arg2* .

According to Chinese linguistics, adjunct clauses and discourse adverbials always precede the main clauses (Gasde and Paul, 1996; Chu and Ji, 1999). The clauses are semantically arranged in a topic-comment sequence following the writer's conceptual mind (Tai, 1985; Bittner, 2013). When the arguments are not arranged in the standard order, the sense of the DC is altered. For example, when '虽然' (*suiran*, although' is used in construction (2), it represents an 'expansion' relation (Huang et al., 2014). Therefore, discourse relations should be defined given the order of the arguments.

Lastly, parallel DCs are frequent in Chinese discourse, yet usually either one DC of the pair occurs to signify the same relation (Zhou et al., 2014). For example, (3) and (4) are grammatical alternatives to (1).

3. 虽然 (*suiran*, although) *Arg1* , 但是 (*danshi*, but) *Arg2* .
4. *Arg1* , 但是 (*danshi*, but) *Arg2* .

Instead of viewing '虽然 (*suiran*, although) - 但是 (*danshi*, but)' as a pair of parallel DCs, they can be regarded individually as a forward-linking (fw-linking) DC and a backing linking (bw-linking) DC. A fw-linking DC relates its attached discourse unit to a later coming unit, while a bw-linking DC relates its attached discourse unit to a previous unit. Findings in linguistic studies also show that fw-linking DCs only link discourse units within the sentence boundary. On the other hand, bw-linking DCs can link a discourse unit to a preceding unit within or outside the sentence boundary, except when it is paired with a fw-linking DC (Eifring, 1995).

To summarize, in contrast with the ambiguous arguments in English, punctuations and limitations on DC usage explicitly mark certain discourse structure in Chinese. Section 2 illustrates the design of our annotation scheme driven by these constraints.

2 Sequential discourse annotation

We propose to follow the natural discourse chains in Chinese and annotate discourse structure as a sequence of alternating arguments and DCs. This section highlights the main differences of our scheme comparing with other frameworks.

2.1 Arguments

Each clause separated by punctuations except quotation marks is treated as a candidate argument. Clauses that do not function as discourse units are classified into 3 types - *attribution*, *optional punctuation* and *non-discourse adverbial*.

The main difference of our annotation scheme is that the the order of the arguments for each DC is defined by default. Since the arguments of a particular discourse relation occur in fixed order and are always adjacent, each argument is related to the immediately preceding argument by a bw-linking DC. In turn, the DC in the first clause of a sentence links the sentence to the previous one, preserving the 2 layer structure denoted by punctuations. An implicit bw-linking DC is inserted if the clause does not contain an explicit DC.

Another characteristic of our annotation is that 'parallel DCs' are annotated separately as one fw-linking DC and one bw-linking DC. Implicit bw-linking DCs are inserted , if possible, even the relation is already marked by a fw-linking DC in the previous argument [1]. In other words, duplicated annotation of one relation is allowed. This helps create more valid samples to capture various combinations of Chinese DCs. When an argument spans more than one discourse units, a fw-linking DC is used to mark the start of the span. Similarly, an implicit DC is inserted if necessary.

2.2 Connectives

There is a large variety of DCs in Chinese and their syntactical categories are controversial. Huang et al. (2014) reports a lexicon of 808 DCs, 359 of which found in the data. Since many DCs signal the same relation, we adopt a functionalist approach to label DC senses.

In this approach, a DC does not limit to any syntactical category. Annotators are asked to perform

[1] Temporal relations are often marked by one fw-linking DC alone and it is not acceptable to insert an implicit bw-linking DC. In this case, the 'redundant' tag is used.

a linguistic test by replacing a candidate expression with an unambiguous and preferably frequent DC of similar sense, which we call a 'main DC'. If the replacement is acceptable, then the expression is identified as a DC and the sense is categorized under the 'main DC'.

For example, '尤为' and '特别是' (*youwei, tebieshi*, in particular / especially) are categorized under '尤其 ' (*youqi*, in particular), if the annotator agrees that they are interchangeable in the context. The list of main DCs is not pre-defined but is constructed in the course of annotation. Based on the assigned 'main DC', each DC instant is categorized into the 4 main senses defined in PDTB: *contingency*, *comparison*, *temporal*, and *expansion*.

The discourse and syntactical limitations of the DCs are considered in the replaceability test. For example, the following pairs are not labeled the same 'main DC' even the signaled discourse relation is the same:

- Fw v.s. bw-linking DCs:
  虽然 (*suiran*, although), 但是 (*danshi*, but)
- *Cause-result* v.s. *result-cause* order:
  因为...所以... (*yinwei...suoyi...*, because... therefore...) and
  之所以...是因为... (*zhisuoyi...shiyinwei...*, the reason why...is because...) [2]
- Placed before v.s. after subject:
  却 (*Que* but) and 但是 (*danshi* but)

The list of 'main DCs' is not pre-defined but is constructed in the course of annotation; an expression is registered as another 'main DC' if it cannot be replaced. Note that expressions that are considered as 'alternative lexicalizations' in PDTB or CDTB are also categorized as explicit connectives, if they pass the replaceability test. Otherwise, an implicit DC, chosen from the list of 'main DCs', is inserted.

2.3 Annotation results

Materials of the corpus are raw texts of 325 articles (2353 sentences) from the Chinese Treebank (Bies et al., 2007) . Errors that affect the annotation process, namely punctuation errors that lead to wrong segmentation, have been corrected.

201 DCs are identified in our data, of which 66 are fw-linking DCs. The DCs are categorized into 73 'main DCs' and 22 have ambiguous senses (labelled with more than one 'main DCs'). The distribution of the tags is shown in Table 1. Note that some of the *'implicit'* relations we define belongs to *'explicit'* in other annotation schemes since 'double annotation' occurs in our annotation.

	CON	COM	TEM	EXP	total
Explicit	380	248	521	683	1832
Implicit	1551	446	164	3022	5183
	ADV	ATT	OPT		total
Non-discourse	630	783	336		1749

Table 1: Distribution of various tags in the annotated corpus (4 senses: CONtingency, COMparison, TEMporal, EXPansion; 3 types of non-discourse-unit segments: ATTRibution, OPTional punctuation, and non-discourse ADVerbial)

3 End-to-end discourse chunker

Our linguistically driven annotation of discourse structure takes the surface discourse features as ground truth. In particular, we define discourse relations based on default argument order and span. We demonstrate its learnability by building a discourse chunker in the form of a classifier cascade as used in English discourse parsing(Lin et al., 2010). Features are extracted from the default arguments of each relation. We evaluate the accuracy of each component and the overall accuracy of the final output, classifying up to the 4 main senses. The pipeline consists of 5 classifiers, as shown in Figure 1, each of which is trained with the relevant samples, e.g. only arguments annotated with explicit DCs are used to train the explicit DC classifier. 289 and 36 articles are used as training and testing data respectively.

Features include lexical and syntactical features (bag of words, bag of POS, word pairs and production rules) that have been used in classifying implicit English DCs (Pitler et al., 2009; Lin et al., 2010), and probability distribution of senses for explicit DC classification. The extraction of features is based on automatic parsing by the Stanford Parser (Levy and Manning, 2003). We also use the surrounding discourse relations as features, hypothesizing that certain relation sequences are more likely than others. The classifiers are trained by SVM with a linear kernel using the LIBSVM package(Chang and Lin, 2011).

[2] the 2 pairs are treated as 4 different DCs.

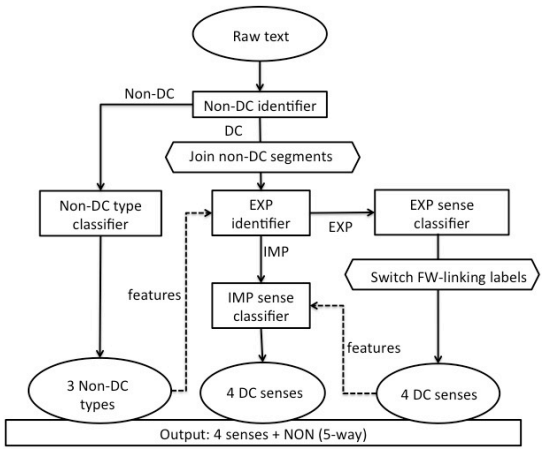

Figure 1: Cascade of discourse relation classifiers.

3.1 Results per component

Table 2 shows the accuracies of individual classifiers tested on relevant samples. Results based on predictions by the most frequent class are listed as baseline (BL). As expected, implicit relations (IMP) are much harder to classify than explicit relations (EXP). The classification result of non-discourse-unit segments (Non-dis or not) is similar to the preliminary report of Li et al. (2014b)(averaged F1 88.8%, accuracy 89.0%).

Step	classifiers	Test F1/Acc	BL F1/Acc
1	Non-dis or not	.91/.94	.44/.80
2	EXP identifier	.92/.93	.39/.65
3	EXP 4 senses	.90/.92	.15/.58
4	Non-dis 3 types	.86/.88	.17/.35
5	IMP 4 senses	.41/.61	.18/.58

Table 2: Accuracies of individual classifiers on 'gold' test samples. F1 is the average of the F1 for each class.

3.2 End-to-end evaluation

We run the classifiers from Steps 1-5. After Step 1, identified non-discourse-unit segments are joined as one argument and features are updated. The discourse context features are also updated after each step based on last classifier's output. The tag of a fw-linking DC is switched to the next segment, as a relation connecting the next segment to the current one. The current segment is thus passed to the implicit classifier, given that there is not any bw-linking DCs.

For applications that need discourse, it may not be necessary to distinguish between explicit and implicit relations. Thus, we combine the outputs of the explicit and implicit classifiers when evaluating the end-to-end outputs. Specifically, the pipeline outputs one of the 4 discourse senses or 'non-discourse-unit' across a segment boundary, while the reference can be more than one, since duplicated annotation is allowed. The system prediction is considered correct if it is included in the gold tag set. The combined outputs are evaluated in terms of accuracy.

Table 3 shows the classification accuracies evaluated by the above principle under different error propagation settings. For example, given gold identification of non-discourse segments (Step 1) and explicit DC classifier (Step 2), classification of the 4 main explicit sense reaches accuracy of 0.854, but is dropped to 0.800 if step 1 and step 2 are automatic [3]. It is observed that errors are generally propagated along the pipeline. Similar to the finding in English (Pitler et al., 2009), the discourse context as predicted by earlier classifiers does not affect the later steps - the results are the same based on gold or automatic outputs. The end-to-end accuracy of the proposed pipeline is 65.7% and the baseline (classify all as 'expansion') is 50.0%.

	Accuracies					
Step	non-dis or not 2-way	exp/imp /non-dis 3-way	explicit senses 4-way	non-dis types 3-way	implicit senses 4-way	over -all 5-way
4	Gold	Gold	Gold	Gold	.670	.706
3	Gold	Gold	Gold	.879	.670	.706
2	Gold	Gold	.854	.879	.670	.703
1	Gold	.888	.800	.865	.665	.697
-	.862	.847	.800	.836	.657	.657

Table 3: Accuracies at each stage under different error propagation settings.

Finally, we experimented with different variations of the pipeline, as shown in Table 4. The best result (70.1% accuracy), is obtained by classifying implicit DCs and non-discourse units in one step. For comaprison, Huang and Chen (2011) reports an accuracy of 88.28% on 4-way classification of inter-sentential discourse senses, and Huang and Chen (2012) reports an accuracy of 81.63% on 2-way classification of intra-sentential contingency vs comparison senses.

[3]Note that the results under the complete gold settings do not necessarily echo the results of the individual components, where duplicated outputs are counted individually.

Note that the result is much degraded if we train one 5-way classifier to classify all relations. This shows that explicit and implicit DCs ought to be treated separately, even though we do not concern about distinguishing them in the final output.

Pipeline variations	Overall 5-way acc.
steps 1-5	.657
combine steps 1-5	.549
switch steps 1 & 2	.697
switch steps 1 & 2 + combine steps 4&5	**.701**

Table 4: 5-way accuracies of modified pipelines

4 Conclusion

This work presents the annotation principles of our Chinese discourse corpus based on linguistics analysis. We propose to embrace the overt sequential features as ground truth discourse structures, and categorize DCs by their discourse functions. Based on the manually annotated corpus, we built and evaluate a classifier cascade that classifies explicit and implicit relations and the results support that our annotation is tractably learnable. The annotation is available at http://cl.naist.jp/nldata/zhendisco/.

References

Ann Bies, Martha Palmer, Justin Mott, and Colin Warner. 2007. English chinese translation treebank v 1.0.

Maria Bittner. 2013. Topic states in mandarin discourse. *Proceedings of the North American Conference on Chinese Linguistics*.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2001. Building a discourse-tagged corpus in the framework of rhetorical structure theory. *Proceedings of the SIGdial Workshop on Discourse and Dialogue*.

Chihchung Chang and Chihjen Lin. 2011. Libsvm : a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*.

Chauncey Chenghsi Chu and Zongren Ji. 1999. *A Cognitive-Functional Grammar of Mandarin Chinese*. Crane.

Chauncey Chenghsi Chu. 1998. *A discourse grammar of Mandarin Chine*. P. Lang.

Halvor Eifring. 1995. *Clause Combination in Chinese*. BRILL.

Horst-Dieter Gasde and Waltraud Paul. 1996. Funcional categories, topic prominence, and complex sentences in mandarin chinese. *Linguistics*, 34.

Hen-Hsen Huang and Hsin-Hsi Chen. 2011. Chinese discourse relation recognition. *Proceedings of the International Joint Conference on Natural Language Processings*.

Hen-Hsen Huang and Hsin-Hsi Chen. 2012. Contingency and comparison relation labeling and structure predictuion in chinese sentences. *Proceedings of the Annual Meeting of SIGDIAL*.

Hen-Hsen Huang, Tai-Wei Chang, Huan-Yuan Chen, and Hsin-Hsi Chen. 2014. Interpretation of chinese discourse connectives for explicit discourse relation recognition. *Proceedings of the International Conference on Computational Linguistics*.

Roger Levy and Christopher Manning. 2003. Is it harder to parse chinese, or the chinese treebank. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Junyi Jessy Li, Marine Carpuat, and Ani Nenkova. 2014. Cross-lingual discourse relation analysis: A corpus study and a semi-supervised classifacation system. *Proceedings of the International Conference on Computational Linguistics*.

Yancui Li, Wenhi Feng, Jing Sun, Fang Kong, and Guodong Zhou. 2014b. Building chinese discourse corpus with connective-driven dependency tree structure. *Proceedings of the Conference on Empirical Methods on Natural Language Processing*.

Ziheng Lin, Hwee Tou Ng, , and Min Yen Kan. 2010. A pdtb-styled end-to-end discourse parser. *Technical report, National University of Singapore*.

Ziheng Lin, Hwee Tou Ng, and Minyen Kan. 2011. Automatic evaluating text coherence using discourse relations. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Emily Pitler, Annie Louis, and Ani Nenkova. 2009. Automatic sense prediction for implicit discourse relations in text. *Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*.

Rashmi Prasad, Nikhit Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The penn discourse treebank 2.0. *Proceedings of the Language Resource and Evaluation Conference*.

James HY Tai. 1985. Temporal sequence and chinese word order. *Iconicity in Syntax*.

Rakshit Trivedi and Jacob Eisenstein. 2013. Discourse connectors for latent subjectivity in sentiment analysis. *Proceedings of the North American Chapter of the Association for Computational Linguistics*.

Yasuhisa Yoshida, Jun Suzuki, Tsutomu Hirao, and Masaaki Nagata. 2014. Dependency-based discourse parser for single-document summarization. *Proceedings of the Conference on Empirical Methods on Natural Language Processing*.

Yuping Zhou and Nianwen Xue. 2015. The chinese discourse treebank: a chinese corpus annotated with discourse relations. *Language Resources and Evaluation*, 49(2).

Lan Jun Zhou, Wei Gao, Binyang Li, Zhongyu Wei, and Kam-Fat Wong. 2012. Cross-lingual identification of ambiguous discourse connectives for resource-poor language. *Proceedings of the International Conference on Computational Linguistics*.

Lan Jun Zhou, Binyang Li, Zhongyu Wei, and Kam-Fai Wong. 2014. The cuhk discourse treebank for chinese: Annotating explicit discourse connectives for the chinese treebank. *Proceedings of the Language Resource and Evaluation Conference*.