

Building Chinese Discourse Corpus with Connective-driven Dependency Tree Structure

Yancui Li^{1,2} Wenhe Feng² Jing Sun¹ Fang Kong¹ Guodong Zhou¹

¹Natural Language Processing Lab, School of Computer Science and Technology, Soochow University, Suzhou 215006, China

²Henan Institute of Science and Technology, Xinxiang 453003, China

{yancuili, wenhefeng}@gmail.com {20104027009, kongfang, gdzhou}@suda.edu.cn

1. Introduction

Background

- ✓ Discourse structure is fundamental to many text-based applications, such as summarization and question-answering
- ✓ Constructing discourse resources has been attracting more and more attention in recent years

Motivation

✓ The general notion of discourse structure mainly consists of discourse unit, connective, structure, relation and nuclearity. previous studies on discourse failed to fully express these kinds of information

➢ 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.

➢ 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

[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

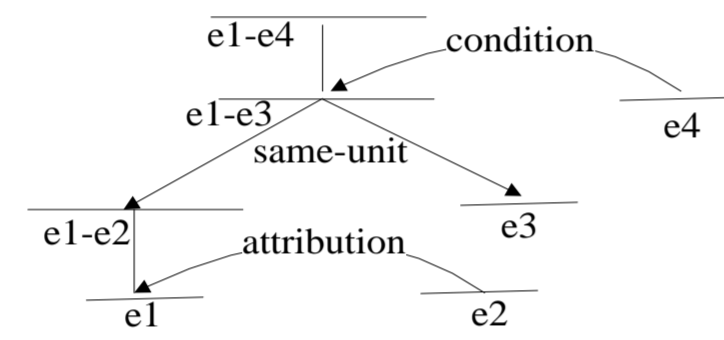


Figure 1: An example of discourse structure in RST Example (1): An example of the connective-argument scheme in PDTB

A) [Catching up with commercial competitors in retail banking and financial services will be difficult.]_{Arg1}, they argue, will be difficult, particularly if [market conditions turn sour.]_{Arg2}. (Contingency.Cause. 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)

✓ The special characteristics of Chinese discourse structure

➢ 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

➢ The connectives in Chinese omit much more frequently than those in English with about 82.0% vs. 54.5% in Zhou and Xue (2012)

➢ The difference in classifying Chinese discourse relations from English (Xing, 2001; Huang and Liao, 2011)

➢ The nucleus of a Chinese discourse relation is dynamically determined from the global meaning of a discourse

2. Related Work

English

✓ Rhetorical Structure Theory Discourse Treebank (RST-DT) (Carlson et al., 2003)

- 385 documents from the Wall Street Journal
- 18 relation classes with 78 finer grained rhetorical relations

✓ Penn Discourse Treebank (Prasad et al., 2008)

- 2159 documents from Penn Treebank-2
- 18459 explicit and 16224 implicit relations
- A three level hierarchy of relation

Chinese

✓ English Tradition

- PDTB: HIT-CDTB(525 documents), Zhou and Xue(98 documents), Sinica Treebank 3.1 (81 documents), CUHK Discourse Treebank for Chinese (890 documents), CTB5.0, only annotate explicit relation)
- RST: Zhejiang University financial comment corpus (97 documents)

✓ The corpus of Chinese complex sentences

- 600 thousand sentences
- Only for explicit connective

3. Connective-driven Dependency Tree

Example (3): CDT example from CTB

1浦东开发开放是一项振兴上海,建设现代化经济、贸易、金融中心的跨世纪工程, 2(因此)大量出现的是以前不曾遇到过的情况、新问题。 3(对此),浦东不是简单地采取“干一段时间,等积累了经验以后再制定法规条例”的做法, 4(而是)借鉴发达国家、深圳等特区的经验教训, 5<并且>聘请国内外有关专家学者, 6<并且>积极、及时地制定和推出法规性文件, 7(使)这些经济活动一出现就被纳入法制轨道。

Elementary Discourse Unit

✓ Leaf node

✓ Definition: From the syntactic structure perspective, an EDU should contain at least one predicate and express at least one proposition. From the functional perspective, an EDU should be related to other EDUs with some propositional function. From the morphological perspective, an EDU should be segmented by some punctuation, e.g. comma, semicolon and period

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)

Connective

✓ Non-leaf node

✓ With any discourse-like word or phrase as connective, depends on its meaning, e.g., “为 (in order to)” is a connective, while “为 (for)” is not

✓ To help determine implicit relations, two special strategies are proposed

- Whether or not explicit connectives can be deleted without changing the rhetorical relation of a discourse
- We cluster implicit connectives into two categories according to their language senses, either “good language intuition” or “bad language intuition”

“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.”

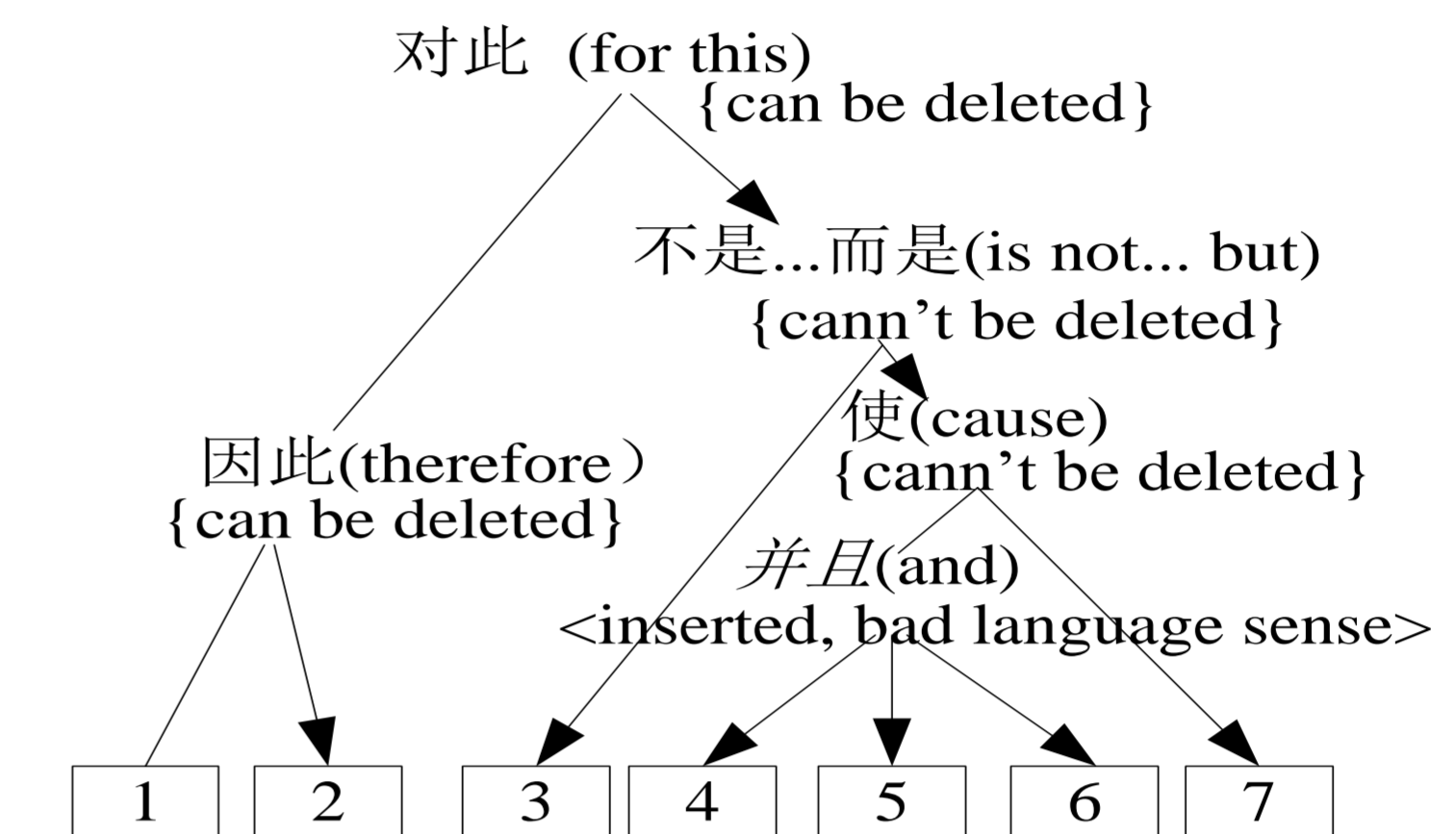


Figure 2: CDT representation of Example (3)

Discourse Structure

✓ The adoption of tree structure conforms to traditional Chinese discourse theories and practice

✓ the hierarchical structure of connectives indicates the hierarchical structure of discourse units

Discourse Relation

• We use the connective itself to express the discourse relation

• For the abstraction of discourse relations, we leave it in a later separate stage

• We give a set of relations, connective itself is the foundation of discourse relation, and the relation set can be adjusted dynamically according to the application requirements

Nucleus and Satellite

✓ Discourse relations may be either mononuclear or multi-nuclear

✓ We adopt the dependency grammar, and select the unit which can stand for the relationship with other discourse units in a discourse.

4. Chinese Discourse Treebank

Annotator Training

✓ 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

Tagging Strategies

✓ Employ a top-down strategy

Consistency

	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	--

Corpus Statistics

✓ 500 newswire articles from Chinese Treebank 6.0

✓ 2342 paragraphs (a CDT for one paragraph)

✓ 10650 EDUs with an average of 4.5 EDUs per tree

✓ 7310 relations, of which 1812 are explicit relations (24.8%) and 5498 are implicit relations (75.2%)

✓ 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)

✓ 3557 (48.7%) relations are mononuclear relations with 2110 nucleus ahead, while the remaining 3754 relations are multi-nuclear

✓ 282 connectives, among which 274 (140 can be deleted) appears as explicit connectives and 44 can be inserted in place of implicit connectives

5. Comparison with other Discourse Banks

	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 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 its attribute; mapping of connective to the set of discourse relations in a later stage
Structure	Complete tree	Partial tree, deduced by connective and its 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

6. Preliminary Experimentation

Classifier	Gold standard parse Automatic parse					
	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

7. Conclusions

• Propose a Connective-driven Dependency Tree (CDT) structure as a representation scheme for Chinese discourse structure

• Describe CDT in detail from various perspectives, such as EDU, connective, structure, relation and nuclearity

• Annotate 500 CDTB corpus guided by CDT scheme

• 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