Automatic Recognition and Classification on Chinese Discourse Connective

LI Yancui[1,2], SUN Jing[1], ZHOU Guodong[1,†]

1. Department of Computer Science and Technology, Soochow University, Suzhou 215006;

2. School of Information Engineering, Henan Institute of Science and Technology, Xinxiang 453003

Outline

- Introduction
- Related work
- Chinese Discourse Treebank
- Experiment of connective Recognition and classification
- Conclusion

Introduction

- Discourse analysis is the fundamental of many NLP applications, such as summarization, question-answering etc.
- Discourse relation is the main task of discourse analysis
- Motivation is explicit discourse relation
- When given connective the classification performance well

Related Work

- Corpus:
  - Chinese Complex Sentences
  - Tsinghua Chinese Treebank，TCT
  - Other(Adopt PDTB schema)
    - HIT-CDTB , Zhou Yuping , Huang H.H., Zhou L.J.
- Research
  - Hu Jinzhu's work on Chinese Complex Sentences
  - Tsinghua Chinese Treebank，TCT
    - Hong Luping explicit relation classification(2 class)
    - Li Yancui connective Recognition and classification
  - Other
    - Zhang Muyu, Huang H.H.

Chinese Discourse Treebank-Overall

- 1Pudong development and opening up is a cross-century project of promote Shanghai, building a modern economy, trade and financial canter. ||2 **Therefore**, there are a large number of new situations and new problems that have not previously been encountered. | 3 Pudong is **not** simply adopting "does a period of time, wait accumulation of experience then develop laws and regulations" approach to this. ||| 4 **But** learns lessons from developed countries and the Shenzhen Special Administrative Region. |||||5 Employ relevant experts and scholars at home and abroad. |||||6 Actively and timely formulate and launch the legal document. ||||7 **So** that economic activities can be incorporated into the legal system when they appeared. || 8 China's first drug procurement service canter of medical institutions, born in the Pudong New Area at the beginning of the last year, ||| 9 operated so far, ||||| 10 traded drugs more than one hundred million Yuan,|||| 11 have not been found a case of kickbacks." (chtb_0001)
- 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新情况、新问题。| 3**对此**，浦东**不是**简单的采取“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**借鉴发达国家和深圳等特区的经验教训，||||| 5聘请国内外有关专家学者，||||| 6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。|| 8去年初浦东新区诞生的中国第一家医疗机构药品采购服务中心，**正因为**一开始就比较规范,||| 9运转至今，
- ||||| 10成交药品一亿多元，|||| 11 没有发现一例回扣。（chtb_0001）

Chinese Discourse Treebank-Overall

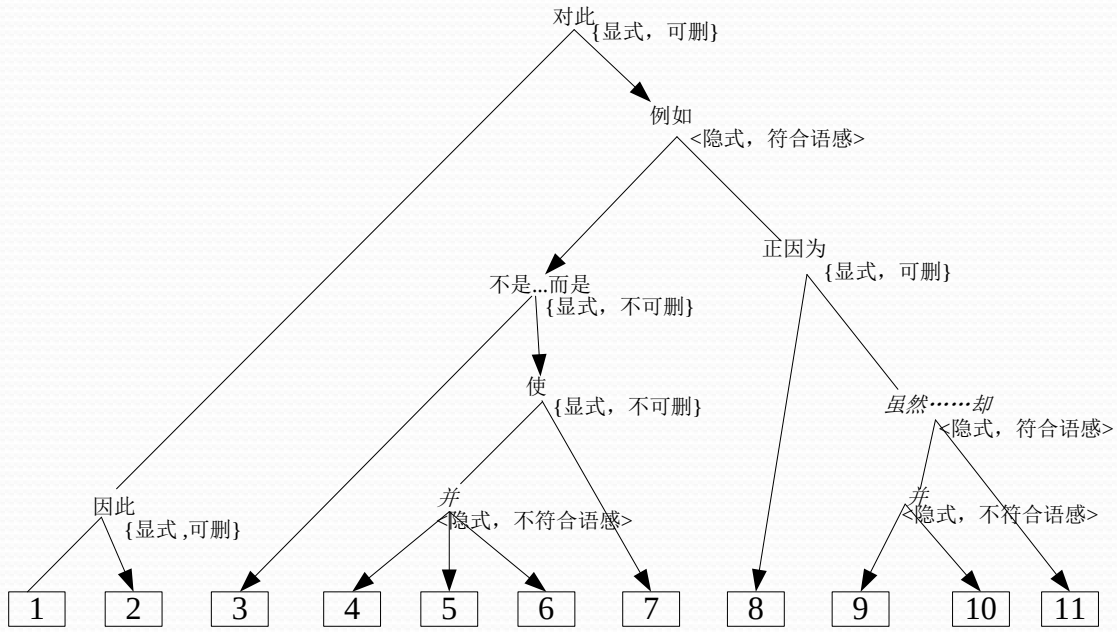

图1 例1的篇章结构树

Fig.1 Discourse structure tree of Example 1

Chinese Discourse Treebank-Overall

- 例2 <R ID="3" StructureType="逐层切分" ConnectiveType="显式关系" Layer="2" RelationNumber="单个关系" Connective="因此" RelationType ="因果关系" ConnectivePosition="37…38" ConnectiveAttribute="可删除" RoleLocation="normal" LanguageSense="true" Sentence="浦东…工程，|因此大量…问题。" SentencePosition="1…36|37…60" Center="2" ChildList="" ParentId="1" UseTime="51"/>
- Currently, the CDTB corpus consists of 500 newswire articles from Chinese Treebank, which are further divided into 2342 paragraphs, 10650 EDUs, 1812 are explicit relations,274 explicit connective words.

	Agreement	Kappa
EDU segmentation	91.7	0.91
Explicit or Implicit	94.7	0.81
Explicit connective identification	82.3	--

Chinese Discourse Treebank-Connective

- 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), 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)”
- From the part-of-speech perspective, connectives are not necessarily conjunctions. For example, 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)”. 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.
- 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.

Chinese Discourse Treebank-Connective

表1 CDTB中出现频率最多的显式连接词及次数
Table1 Most frequent connectives in CDTB

Connective	Frequency	Connective	frequency
并(and)	208	其中(among them)	154
也(also)	131	而(however)	70
但(but)	69	还(also)	68
使(so that)	56	以(in order to)	52
为(in order to)	49	同时(meanwhile)	46

About 30 connectives occurs more than 10 times
Almost half connectives occurs only once

Chinese Discourse Treebank-Connective

causality(1335)
- cause-result(686)
  - because...
- inference(38)
  - so that...
- hypothetical(70)
  - if...
- purpose(335)
  - in order to...
- condition(72)
  - only…
- background(134)
  - background...

transition(217)
- transition (200)
  - but...
- concessive(17)
  - altough...

coordination(4148)
- coordination(3503)
  - and...
- continue(517)
  - first...second...
- progressive(59)
  - in addition..
- selectional(10)
  - or...
- inverse(59)
  - compared with...

explanation(1617)
- explanation(911)
  - which including...
- summary-elaboration (234)
  - in a word...
- example(252)
  - e.g....
- evaluation (220)
  - evaluation ...

图2 基于连接词的关系分类及分布
Fig.2 Relation classification and distribution based on connectives

Chinese Discourse Treebank-Connective

表2 CDTB中部分可表示多种关系的连接词及次数
Table2 The connectives which can represent multi-relation in CDTB

连接词(connective)	表示关系及次数(relation and times)
并(and)	顺承关系;递进关系(1)、顺承关系(7)、解说关系(1)、并列关系(199) Continue; progressive(1), continue(7), explanation(1), coordination(199)
其中(among them)	解说关系(1)、总分关系(153) explanation(1),coordination(153)
也(also)	顺承关系(2)、并列关系(128)、并列关系;顺承关系(1) continue(2), coordination(128),coordination; continue(1)
而(however)	递进关系(6)、顺承关系(1)、转折关系(5)、对比关系(18)、并列关系(39)、因果关系(1) Progresive(6), continue(1),transition(5),inverse(18), coordination(199), causality (1)
但(but)	转折关系(66)、对比关系(3) transition(5),inverse(3)
还(also)	顺承关系(6)、并列关系(61)、选择关系(1) continue(6), coordination(199),selection(1)
使(so that)	因果关系(38)、目的关系(18) causality(38),purpose(18)
如(for example, if)	例证关系(17)、假设关系(11) Example(17), hypothetical(11)
又()	顺承关系(10)、并列关系(16) continue(10), coordination(16)
而且(also)	递进关系(1)、并列关系(19) Progressive(6), coordination(19)

Experiment

- Based on our CDTB corpus
- Using our previous work feature on TCT connective recognition and classification
- Using Natural Language Toolkit(NLTK) 3.0
- 10-corss validation

Experiment-Connective recognition

表3 是否为连接词识别正确率

Table 3 Connectives recognition accuracy

Corpus		自动句法树(Auto parse)			标准句法树(Standard parse)		
Our CDTB	Features	MaxEnt	Decision Tree	Bayes	MaxEnt	Decision Tree	Bayes
	Lexical	86.2	87.3	81.5	86.7	87.2	81.8
	Syntactic	80.9	82.2	80.3	84.3	84.7	82.4
	Lexical + Syntactic	86.6	88.1	81.9	87.9	88.9	83.7
	Lexica l+ Syntactic + Position	87.2	88.4	83.4	88.2	88.5	85.3
TCT	**Lexical + Syntactic + Position**	**91.2**	**92.1**	**88.1**	-	-	-

Experiment-Connective recognition

表4连接词识别的准确率、召回率和F1值

Table 4 Connectives recognition Precision、Recall and F1-measure

classifier	Type	Precision	Recall	F1-measure
Maxent	Auto parser	78.8	61.8	69.2
	Standard parser	78.9	61.8	69.3
Decision Tree	Auto parser	56.8	49.6	52.3
	Standard parser	58.9	48.5	52.7

Experiment-Connective classification

表5 给定连接词4大类别识别结果
Table 5 4 Categories of given connective classification results

类别 (Relation Type)	P	R	F1
因果类 (causality)	83.8	68.4	75.1
转折类 (transition)	78.5	59.6	67.0
并列类 (coordination)	82.5	93.6	87.7
解说类 (explanation)	89.7	82.8	85.9

表6自动连接词识别及4大类别识别结果
Table 6 4 Categories of connective recognition and classification results

类别 (Relation Type)	P	R	F1
因果类 (causality)	72.8	80.5	76.2
转折类 (transition)	73.2	70.8	71.2
并列类 (coordination)	64.7	95.8	77.2
解说类 (explanation)	82.5	86.7	84.5

Conclusion

- Using our CDTB, which include the connective annotation
- Using lexical, Syntactic , position features for discourse connective recognition and classification
- Connective recognition has less dependence on syntactic,F1 69.2%
- Connective classification accuracy is 95.7% when given connective list and 89.1% when automatic recognition connective. (has a great influence on the performance of connective recognition)
- Future work is improving the performance

Thanks and Question