Chinese Comma Disambiguation on K-best Parse Trees

Fang Kong and Guodong Zhou

School of Computer Science and Technology, Soochow University, China

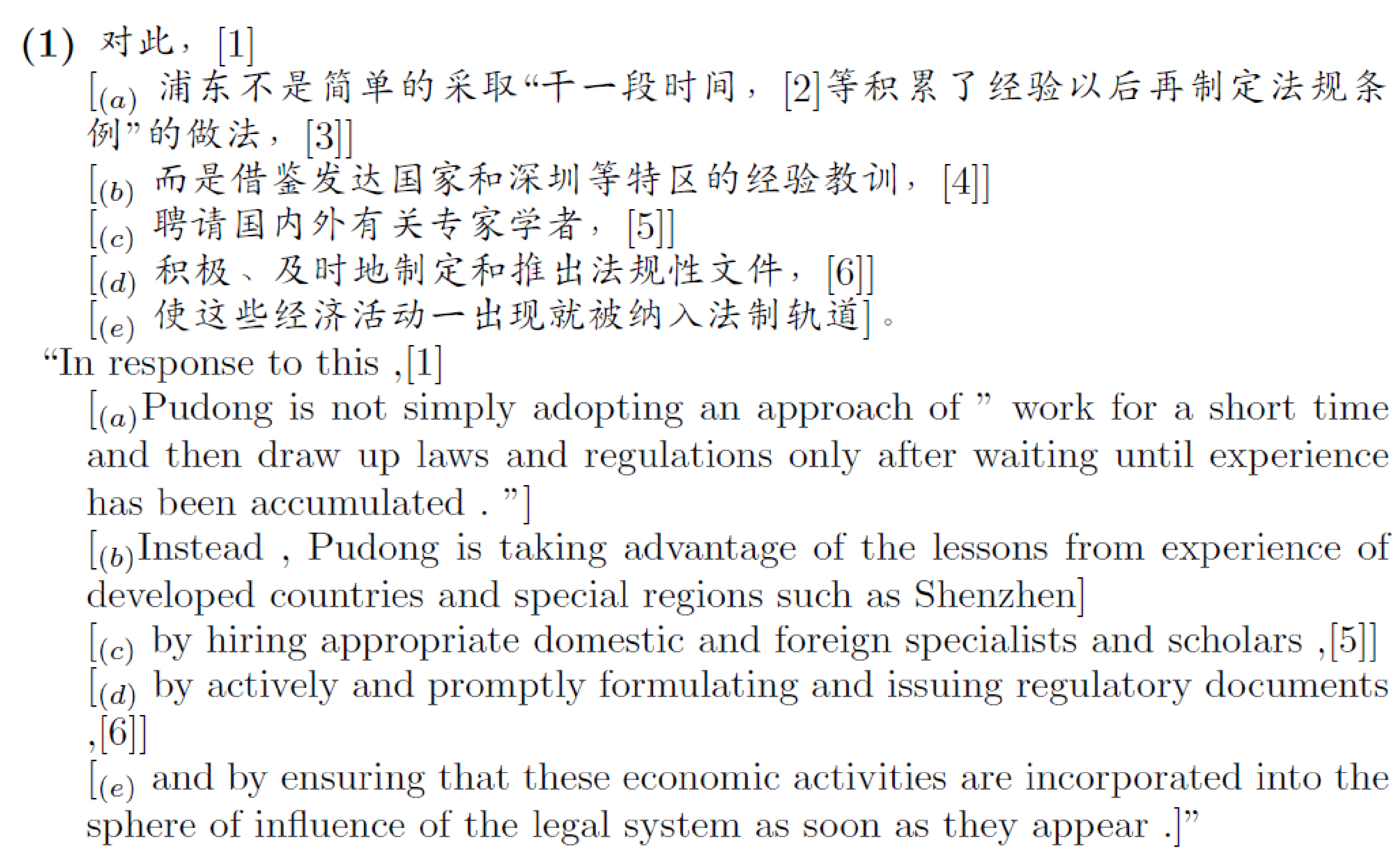
Outline

- Introduction
- Chinese Comma Classification
- Baseline System: A maximum entropy approach
- Refined System: K-best combination approach
- Experiments
- Conclusion

Introduction

- Chinese commas
  - The most common form of punctuation
  - Function quite different from its English counterpart
    - not only function similarly as the English periods
    - but also
      - act as the boundary of sentences
      - signal the boundary of discourse units and anchor discourse relations between text spans

Introduction



(1) 对此，[1]

$[_{(a)}$ 浦东不是简单的采取“干一段时间，[2]等积累了经验以后再制定法规条例”的做法，[3]]

$[_{(b)}$ 而是借鉴发达国家和深圳等特区的经验教训，[4]]

$[_{(c)}$ 聘请国内外有关专家学者，[5]]

$[_{(d)}$ 积极、及时地制定和推出法规性文件，[6]]

$[_{(e)}$ 使这些经济活动一出现就被纳入法制轨道]。

“In response to this ,[1]

$[_{(a)}$Pudong is not simply adopting an approach of ” work for a short time and then draw up laws and regulations only after waiting until experience has been accumulated . ”]

$[_{(b)}$Instead , Pudong is taking advantage of the lessons from experience of developed countries and special regions such as Shenzhen]

$[_{(c)}$ by hiring appropriate domestic and foreign specialists and scholars ,[5]]

$[_{(d)}$ by actively and promptly formulating and issuing regulatory documents ,[6]]

$[_{(e)}$ and by ensuring that these economic activities are incorporated into the sphere of influence of the legal system as soon as they appear .]”

Introduction

- Chinese comma Disambiguation
  - Classify the Chinese commas into multiple categories based on their functions

    --- syntactic patterns
  - Disambiguate the Chinese commas automatically

Introduction

- Related work
  - from the perspective of sentence segmentation
    - Syntactic parsing for long sentences
      - Jin et al. (2004), Li et al.(2005): view this task as a part of a “divide-and-conquer” strategy to syntactic parsing
    - Serving for some NLP applications
      - Xue and Yang (2011): view this task as the detection of loosely coordinated clauses separated by commas and simplify some downstream tasks such as SMT
      - Kong and Zhou (2013): employ this task to improve the detection of Chinese clauses, and improve the performance of Chinese empty category recovery furtherly .

Introduction

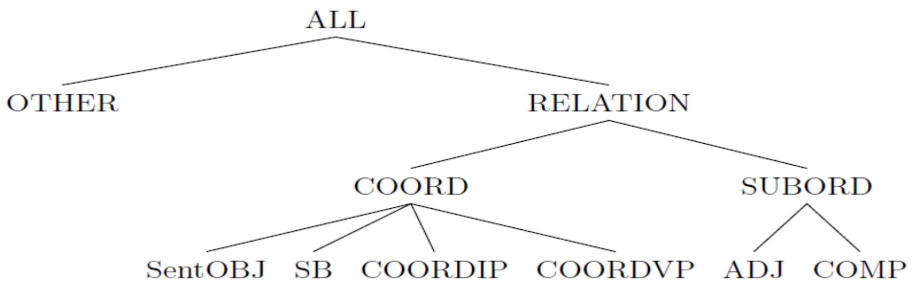
- Related work
  - from the perspective of discourse analysis
    - View some Chinese commas as a delimiter of elementary discourse units(EDUs)
    - Cast the EDUs identification, the first step in building up the discourse structure of Chinese text, as Chinese comma disambiguation
      - Yang and Xue(2012) proposed a discourse structure-oriented classification of the Chinese commas
      - Xu et al.(2013) also proposed a Chinese comma classification scheme

Introduction

- Work of this paper
  - Classify the Chinese commas into seven categories based on syntactic patterns and annotate a Chinese comma corpus which adds a layer of annotation to the manually-parsed sentence in the CTB6.0 corpus
  - Propose a machine learning approach to Chinese comma disambiguation
  - Employ a joint approach based on K-best parse trees to reduce the dependent on syntactic parsing

Chinese comma classification

- Seven categories
  - SB, sentence boundary. The loosely coordinated IPs that are the immediate children of the root IP to be independent sentences.
  - COORDIP, coordinated IPs that are not the immediate children of the root IP.
  - COORDVP, coordinated VPs, when separated by the comma.
  - SentOBJ, links two coordinated IPs in the object phrase.
  - COMP, separates a verb governor and its complement clause.
  - ADJ, links a subordinate clause with its main clause.
  - OTHER, the remaining cases of comma.



Chinese comma classification

- Chinese comma corpus
  - adding a layer of comma annotation in the CTB6
  - semi-automatic way (human adjust after rule-based approach)

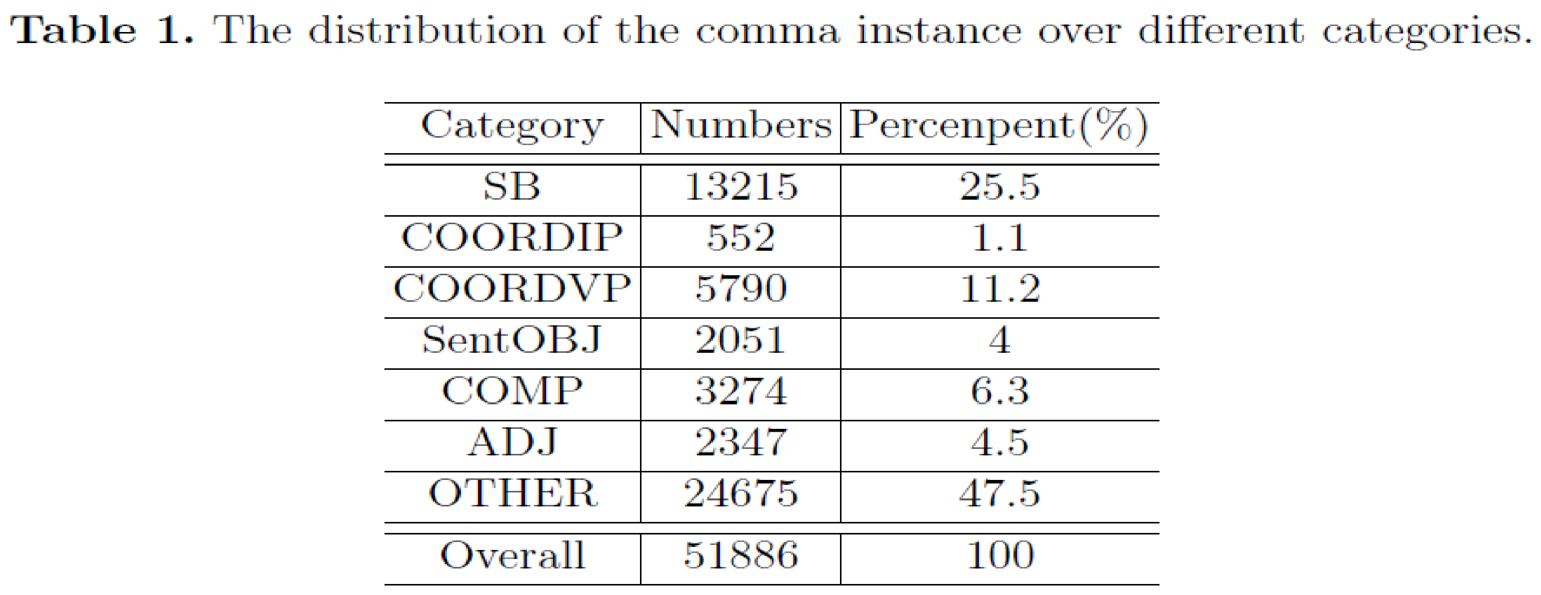
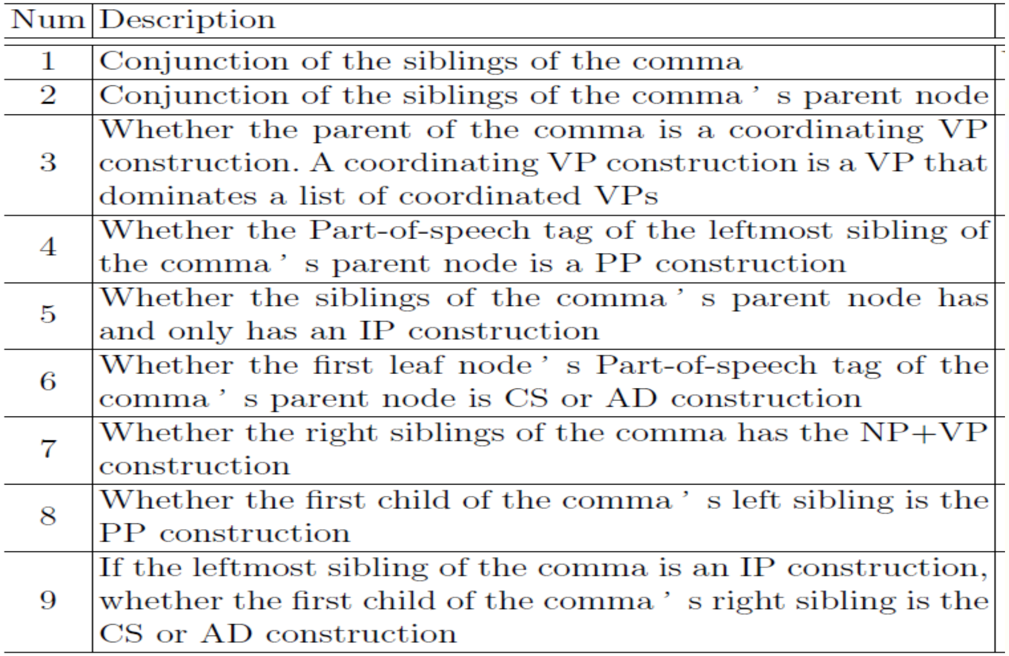


Table 1. The distribution of the comma instance over different categories.

Category	Numbers	Percenpent(%)
SB	13215	25.5
COORDIP	552	1.1
COORDVP	5790	11.2
SentOBJ	2051	4
COMP	3274	6.3
ADJ	2347	4.5
OTHER	24675	47.5
Overall	51886	100

Baseline system: A maximum entropy approach

- Cast this task as a multiple classification problem
- Feature set:
  - All the features from Xue and Yang(2011)
  - Additional features: reflect the properties of the context where current comma occurs



Num	Description
1	Conjunction of the siblings of the comma
2	Conjunction of the siblings of the comma ’ s parent node
3	Whether the parent of the comma is a coordinating VP construction. A coordinating VP construction is a VP that dominates a list of coordinated VPs
4	Whether the Part-of-speech tag of the leftmost sibling of the comma ’ s parent node is a PP construction
5	Whether the siblings of the comma ’ s parent node has and only has an IP construction
6	Whether the first leaf node ’ s Part-of-speech tag of the comma ’ s parent node is CS or AD construction
7	Whether the right siblings of the comma has the NP+VP construction
8	Whether the first child of the comma ’ s left sibling is the PP construction
9	If the leftmost sibling of the comma is an IP construction, whether the first child of the comma ’ s right sibling is the CS or AD construction

Refined system: K-best combination approach

- **Problem:** heavily depend on the performance of syntactic parser.
- **Solution:**
  - Using the general framework of re-ranking, joint Chinese comma disambiguation with the selection of the best parse tree
    - Allows uncertainty about syntactic parsing to be carried forward through a K-best list
    - A reliable comma disambiguation system, to a certain extent, can reflect qualities of syntactic parse trees
  - Given a sentence *s*, a joint parsing model is defined over a comma *c* and a parse tree *t* in a log-linear way:

P(t|s) is returned by a probabilistic syntactic parsing model

P(c|t,s) is returned by a probabilistic comma classifier.

α is a balance factor.

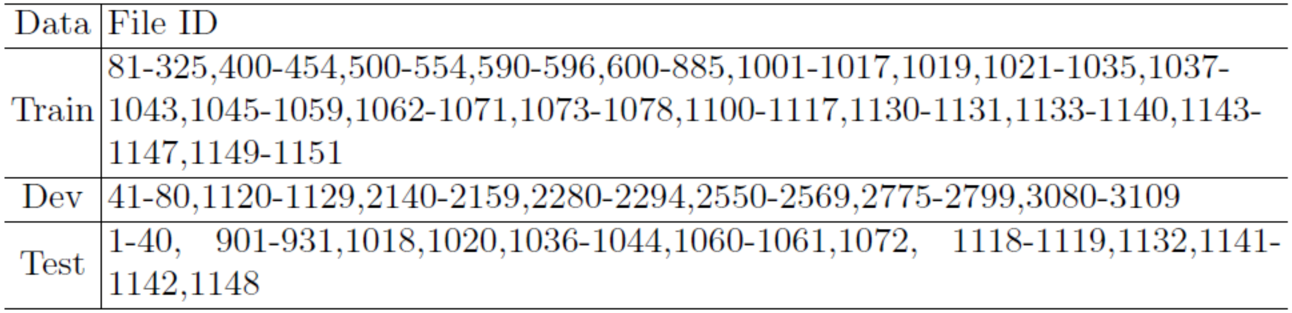
$$Score(c,t|s) = (1-\alpha)\log P(c|t,s) + \alpha \log P(t|s)$$

In our approach, *P(t|s)* is calculated as the product of all involved decisions' probabilities in the syntactic parsing model, and P(c|t,s) is calculated as the product of all the commas' probabilities in a sentence.

Experimentation

- **Experimental settings:**
  - Data set division:
  - Mallet machine learning package with the default parameters
  - Berkeley parser is used to generate top-best and 50-best parse trees

Table 3. CTB 6 Data set division.



Data	File ID
Train	81-325,400-454,500-554,590-596,600-885,1001-1017,1019,1021-1035,1037-1043,1045-1059,1062-1071,1073-1078,1100-1117,1130-1131,1133-1140,1143-1147,1149-1151
Dev	41-80,1120-1129,2140-2159,2280-2294,2550-2569,2775-2799,3080-3109
Test	1-40, 901-931,1018,1020,1036-1044,1060-1061,1072, 1118-1119,1132,1141-1142,1148

Experimentation

- **Results**

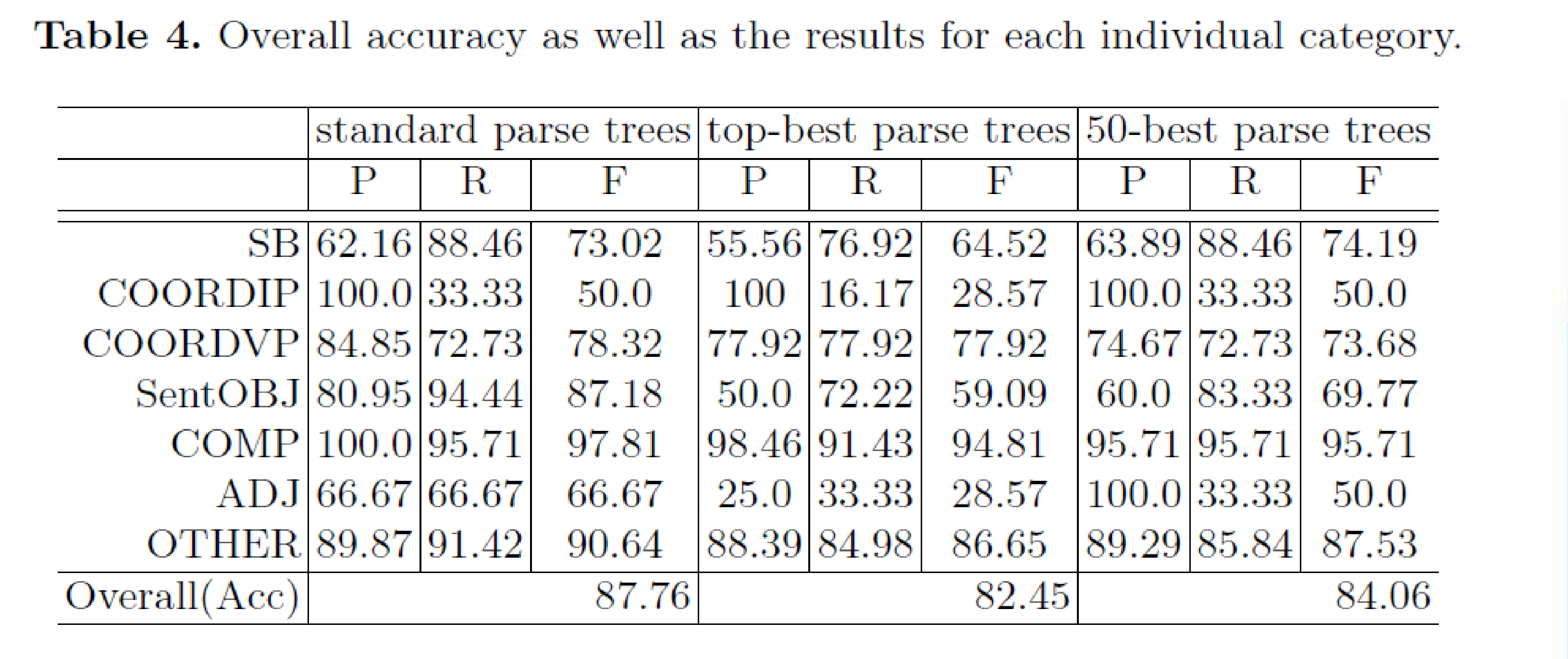


Table 4. Overall accuracy as well as the results for each individual category.

	standard parse trees			top-best parse trees			50-best parse trees		
	P	R	F	P	R	F	P	R	F
SB	62.16	88.46	73.02	55.56	76.92	64.52	63.89	88.46	74.19
COORDIP	100.0	33.33	50.0	100	16.17	28.57	100.0	33.33	50.0
COORDVP	84.85	72.73	78.32	77.92	77.92	77.92	74.67	72.73	73.68
SentOBJ	80.95	94.44	87.18	50.0	72.22	59.09	60.0	83.33	69.77
COMP	100.0	95.71	97.81	98.46	91.43	94.81	95.71	95.71	95.71
ADJ	66.67	66.67	66.67	25.0	33.33	28.57	100.0	33.33	50.0
OTHER	89.87	91.42	90.64	88.39	84.98	86.65	89.29	85.84	87.53
Overall(Acc)			87.76			82.45			84.06

Conclusion

- Based on syntactic patterns, we classify the Chinese commas into seven categories and annotate a Chinese comma corpus adding a layer of annotation in the CTB 6.0 corpus.
- Using this annotated corpus, we propose a approach to disambiguate the Chinese commas as a first step toward discourse analysis.
- In order to reduce the dependent on syntactic parsing, a joint mechanism based on K-best parse trees is proposed. Experiment results show the effectiveness of our joint approach.

Thanks for your attention!