# Chinese Comma Disambiguation on K-best Parse Trees

Fang Kong and Guodong Zhou

School of Computer Science and Technology
Soochow University, China
{kongfang,gdzhou}@suda.edu.cn

**Abstract.** Chinese comma disambiguation plays key role in many natural language processing (NLP) tasks. This paper proposes a joint approach combining K-best parse trees to Chinese comma disambiguation to reduce the dependent on syntactic parsing. Experimental results on a Chinese comma corpus show that the proposed approach significantly outperform the baseline system. To our best knowledge, this is the first work improving the performance of Chinese comma disambiguation on K-best parse trees. Moreover, we release a Chinese comma corpus which adds a layer of annotation to the manually-parsed sentences in the CTB (Chinese Treebank) 6.0 corpus.

**Keywords:** Chinese Comma Disambiguation, Discourse Analysis, Sentence Segmentation, K-best parse trees.

## 1 Introduction

The Chinese commas function quite different from its English counterpart. On one hand, they can signal the sentence boundary. Similar to English, Chinese also uses periods, question marks, and exclamation marks to indicate sentence boundaries. Where these punctuation marks exist, sentence boundaries can be determined unambiguously. The difference is that the Chinese commas, as the most common form of punctuation, can also function similarly as the English periods in many kinds of contexts. On the other hand, the Chinese commas can also signal the boundary of discourse units and anchor discourse relations between text spans. Observing the CTB 6.0 corpus, we find that implicit discourse relations occupy more than 75% and much of them are anchored by the Chinese commas.

In recent years, Chinese comma disambiguation has attracted increasing attention due to its importance in many NLP tasks, such as sentence segmentation ([4,6,9,5]) and discourse analysis ([10,8]). Previous work has achieved reasonable success, they also found that the performance of Chinese comma disambiguation heavily depended on the performance of syntactic parser. In this paper, we 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 sentences in the CTB 6.0 corpus. Using the annotated corpus,

C. Zong et al. (Eds.): NLPCC 2014, CCIS 496, pp. 13–22, 2014.

a machine learning approach to Chinese comma disambiguation is proposed. Finally, a joint approach based on K-best parse trees is employed to reduce the dependent on syntactic parsing. Experiments on our Chinese comma corpus show that our joint approach can significantly improve the performance of Chinese comma disambiguation with automatic parse trees.

The rest of this paper is organized as follows. Section 2 introduces the related work from syntactic parsing and discourse analysis. In Section 3, we present our Chinese comma classification scheme and briefly overview our annotated Chinese comma corpus. Section 4 describes a machine learning approach to Chinese comma disambiguation as a baseline. In Section 5, a joint approach combining k-best syntactic parse trees is proposed. Section 6 presents the experiments and results. Finally, we give the conclusion and further work in Section 7.

## 2 Related Work

The Chinese comma can not only function similarly as the English periods, but also act as the boundary of sentences or discourse units. Currently, many research work about Chinese comma disambiguation has been conducted from the perspective of sentence segmentation and discourse analysis.

For Chinese sentence segmentation, the related work can be classified into two categories: in the context of syntactic parsing for long sentences, and serving for some NLP applications such as machine translation and empty category recovery. The representative work includes: Jin et al. [4] and Li et al.[6] view Chinese comma disambiguation as a part of a "divide-and-conquer" strategy to syntactic parsing. Long sentences are split into shorter sentence segments on commas before they are parsed, and the syntactic parses for the shorter sentence segments are then assembled into the syntactic parse for the original sentence. Xue and Yang [9] view Chinese comma disambiguation as the detection of loosely coordinated clauses separated by commas, which are syntactically and semantically complete on their own and do not have a close syntactic relation with one another. In this way, some downstream tasks such as parsing and Machine Translation can be simplified. Kong and zhou [5] employ a comma disambiguation method to improve syntactic parsing and help determine clauses in Chinese. Based on the detected clauses, a clause-level hybrid approach is proposed to address specific problems in Chinese empty category recovery and achieves significant performance improvement.

For discourse analysis, related work find that the Chinese comma can be further viewed as a delimiter of elementary discourse units (EDUs) and the anchor of discourse relations. Disambiguating the comma is thus necessary for the purpose of discourse segmentation, the identification of EDUs, a first step in building up the discourse structure of a Chinese text. The representative work includes: Yang and Xue [10] propose a discourse structure-oriented classification of the comma and conduct experiments with two supervised learning methods that automatically disambiguate the Chinese comma based on this classification. Motivated by the work of Yang and Xue, Xu et al. [8] also divide the Chinese

commas into seven categories based on syntactic patterns and propose three different machine learning methods to automatically disambiguate the Chinese commas.

All the previous work shows that the performance of Chinese comma disambiguation heavily depends on the performance of syntactic parser. Similar to the work of Yang and Xue [10] and Xue et al. [8], in this paper, we also classify the Chinese commas into seven categories based on syntactic patterns. Then a traditional machine learning approach will be employed to do Chinese comma disambiguation automatically. Based on this baseline system, we focus on improving the performance of Chinese comma disambiguation on K-best parse trees.

## 3 Chinese Comma Classification

Just as Zhou and Xue [11] noted, despite similarities in discourse features between Chinese and English, there are differences that have a significant impact on how discourse relations could be best annotated. For example, as illustrated in (1), there are six commas. In its corresponding English translation, we only can find the first,fifth and sixth commas. The second comma does not mark the boundary of discourse unit and is not translated, the third one corresponds to an English period. And the fourth comma marks the boundary of discourse unit but not translated in English. In fact, the Chinese sentence can be split into five discourse units marked with (a)–(e).

**(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 .]”

Figure 1 and Figure 2 show the corresponding syntactic parse tree and discourse parse tree, respectively. From the figures, we can find that different commas with different syntactic patterns can function differently in discourse modeling.

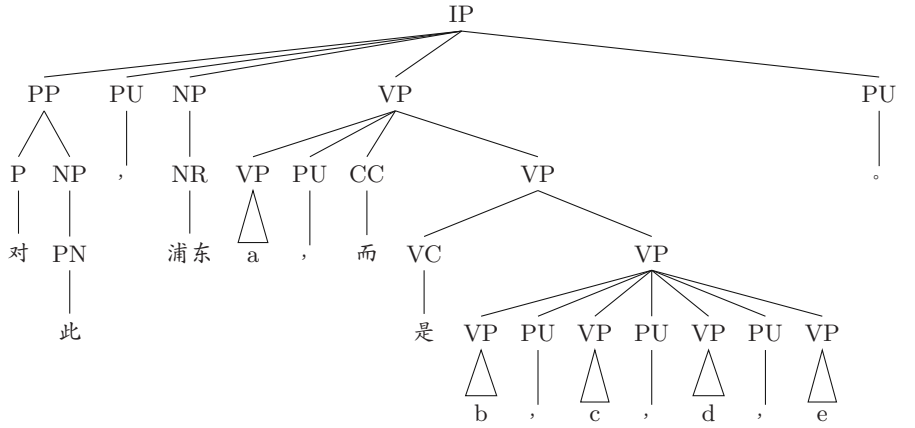


**Fig. 1.** Syntactic parse tree corresponding to Example (1)

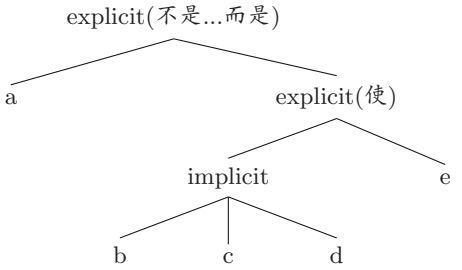


**Fig. 2.** Discourse parse tree corresponding to Example (1)

Referring the work of Yang and Xue [10] and Xu et al. [8], we classify the Chinese comma into seven hierarchically organized categories based on the syntactic patterns. Figure 3 shows our classification scheme. The description of the categories are following:

- SB, sentence boundary. The loosely coordinated IPs that are the immediate children of the root IP to be independent sentences, and the commas separating them to be delimiters of sentence boundary.
- COORDIP, coordinated IPs that are not the immediate children of the root IP are also considered to be discourse units and the commas linking them are labeled COORDIP.
- COORDVP, coordinated VPs, when separated by the comma, are not semantically different from coordinated IPs. The only difference is that the coordinated VPs share a subject.

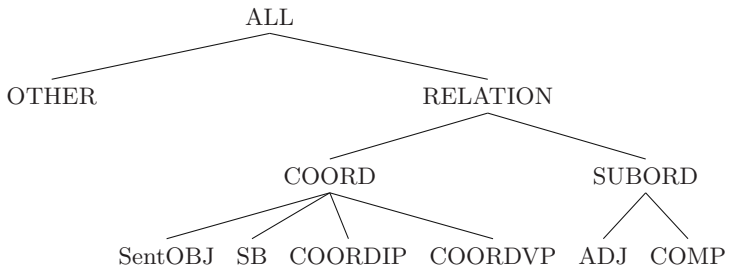


**Fig. 3.** Chinese Comma Classification

- SentOBJ, this category is for commas linking two coordinated IPs in the object phrase.
- COMP, when a comma separates a verb governor and its complement clause, this verb and its subject generally describe the attribution of the complement clause.
- ADJ, this category is for commas linking a subordinate clause with its main clause.
- OTHER, the remaining cases of comma.

Knowing the above Chinese comma classification scheme, all the commas in the CTB can be mapped to one of the seven classes based on the syntactic patterns. Using semi-automatic way (i.e., human adjust after rule-based approach), we build a Chinese comma corpus adding a layer of comma annotation in the CTB 6 corpus. Table 1 shows the distribution of all the comma instances over the seven categories.

**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 |

## 4 Baseline System: A Maximum Entropy Approach

After the commas are labeled, we have basically turned comma disambiguation into multiple classification problem. We trained a Maximum Entropy classifier with the Mellet machine learning package[1] to classify the Chinese commas.

[1] `http://mallet.cs.umass.edu`

The features are extracted from gold standard parse trees and automatical parse trees, respectively. We implement features described in Xue and Yang [9], and also introduce a set of new features. Table 2 lists the new features employed in Chinese comma classification, which reflect the properties of the context where current comma occurs. The third column of Table 2 shows the features corresponding to Figure 1, considering the fourth comma between span $c$ and $d$ as the current comma in question.

**Table 2.** New features employed in comma classification

| Num | Description | Example |
|---|---|---|
| 1 | Conjunction of the siblings of the comma | VP+VP-PU-VP-PU-VP |
| 2 | Conjunction of the siblings of the comma ' s parent node | VC-VP |
| 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 | True |
| 4 | Whether the Part-of-speech tag of the leftmost sibling of the comma ' s parent node is a PP construction | False |
| 5 | Whether the siblings of the comma ' s parent node has and only has an IP construction | False |
| 6 | Whether the first leaf node ' s Part-of-speech tag of the comma ' s parent node is CS or AD construction | False |
| 7 | Whether the right siblings of the comma has the NP+VP construction | False |
| 8 | Whether the first child of the comma ' s left sibling is the PP construction | False |
| 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 | False |

## 5 Refined System: K-best Combination Approach

Note that most of features employed in our baseline system are extracted from syntactic parse trees. Consistent with previous work on Chinese comma disambiguation, the performance of our baseline system should heavily depend on the performance of syntactic parser. In this section, we will propose a k-best combination approach to address this problem.

As well known, parsing re-ranking has been shown to be an effective technique to improve parsing performance [2,1,3]. This technique uses a set of linguistic features to re-rank the k-best output on the forest level or tree level. Motivated by these work, using the general framework of re-ranking, we joint Chinese comma disambiguation with the selection of the best parse tree. The idea behind this approach is that it allows uncertainty about syntactic parsing to be carried forward through a K-best list, and that 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:

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

while $P(t|s)$ is returned by a probabilistic syntactic parsing model, and $P(c|t,s)$ is returned by a probabilistic comma classifier. In our K-best combination 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. Here, the parameter $\alpha$ is a balance factor indicating the importance of the comma disambiguation model.

In particular, $(c^*, t^*)$ with maximal $Score(c,t|s)$ is selected as the final syntactic parsing tree and the comma disambiguation result.

## 6 Experimentation

### 6.1 Experimental Settings

We use the CTB 6.0 in our experiments and divide it into training, development and test sets, as shown in Table 3. All our classifiers are trained using the the Mellet machine learning package[2] with the default parameters (i.e. without smoothing and with 100 iterations). Under the automatic setting, the Berkeley parser [7] is used to generate top-best parse trees and 50-best parse trees, respectively.

**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 |

### 6.2 Results

Table 4 lists the results under three different settings: using gold standard parse trees, using top-best parse trees, and using 50-best parse trees.

The second column shows the results of our comma disambiguation system using gold standard parse trees. From the results, we can find that our baseline system performs best on the category COMP. Both precision and recall are more than 95%, and the F-score is 97.81%. While on the category COORDIP, our system achieves only 50.0% in F1-measure much due to the poor recall. The overall accuracy of our comma disambiguation system using gold parse trees is 87.76%.

[2] `http://mallet.cs.umass.edu`

The third column shows the performs of our comma disambiguation system using top-best automatic parse trees. For the category COMP, the system also achieves satisfactory results. But for the category COORDIP and the category ADJ, our system only achieves about 28% in F1-measure. In comparison with using gold standard parse trees, the performance of every category is reduced. Especially for the category SentOBJ and the category ADJ, the F-score reduced by about 28% and 38%, respectively. The overall accuracy reduced about 5%.

The fourth column shows the results of our comma disambiguation system joint with 50-best parse trees. In comparison with using top-best parse trees, we can find that the refined system can achieve better performance on all the categories except the category COORDVP. Except the category COMP and OTHER, the improvement on every category is larger than 10% in F1-measure. And the overall accuracy of the refined system improves about 1.5% comparing with using top-best parse trees.

Although our refined system reduces the performance gap between using automatic parse trees and using gold parse trees by about 30%, it still lags behind using gold standard parse trees about 3.7% in overall accuracy. This suggests that there exists some room in the performance improvement for the joint mechanism with K-best parse trees.

**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 |

## 7 Conclusion and Future Work

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.

In our future work, we will find more effective joint inference mechanism to improve the performance of Chinese comma disambiguation.

**Acknowledgements.** This research is supported by Key project 61333018 and 61331011 under the National Natural Science Foundation of China, Project 6127320 and 61472264 under the National Natural Science Foundation of China, Project 2012AA011102 under the National 863 Program of China, and Project 11KJA520003 under the Natural Science Major Fundamental Research Program of the Jiangsu Higher Education Institutions.

## References

1. Charniak, E., Johnson, M.: Coarse-to-fine n-best parsing and maxent discriminative reranking. In: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005), pp. 173–180. Association for Computational Linguistics, Ann Arbor (2005), http://www.aclweb.org/anthology/P05-1022
2. Collins, M., Duffy, N.: New ranking algorithms for parsing and tagging: Kernels over discrete structures, and the voted perceptron. In: Proceedings of 40th Annual Meeting of the Association for Computational Linguistics, pp. 263–270. Association for Computational Linguistics, Philadelphia (2002), http://www.aclweb.org/anthology/P02-1034
3. Huang, L.: Forest reranking: Discriminative parsing with non-local features. In: Proceedings of ACL 2008: HLT, pp. 586–594. Association for Computational Linguistics, Columbus (2008), http://www.aclweb.org/anthology/P/P08/P08-1067
4. Jin, M., Kim, M.Y., Kim, D., Lee, J.H.: Segmentation of chinese long sentences using commas. In: Streiter, O., Lu, Q. (eds.) ACL SIGHAN Workshop 2004, pp. 1–8. Association for Computational Linguistics, Barcelona (2004)
5. Kong, F., Zhou, G.: A clause-level hybrid approach to chinese empty element recovery. In: IJCAI. IJCAI/AAAI (2013)
6. Li, X., Zong, C., Hu, R.: A hierarchical parsing approach with punctuation processing for long chinese sentences. In: Proceeding of the Second International Joint COnference on Natural Language Processing: Companion Volume to the Proceedings of Conference Including Posters/Demons and Tutorial Abstracts, pp. 17–24 (2005), http://anthology.aclweb.org/I/I05/I05-2002
7. Petrov, S., Klein, D.: Improved inference for unlexicalized parsing. In: Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference, pp. 404–411. Association for Computational Linguistics, Rochester (2007), http://www.aclweb.org/anthology/N/N07/N07-1051
8. Xu, S., Li, P.: Recognizing chinese elementary discourse unit on comma. In: IALP, pp. 3–6. IEEE (2013)
9. Xue, N., Yang, Y.: Chinese sentence segmentation as comma classification. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pp. 631–635. Association for Computational Linguistics, Portland (2011), http://www.aclweb.org/anthology/P11-2111

10. Yang, Y., Xue, N.: Chinese comma disambiguation for discourse analysis. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 786–794. Association for Computational Linguistics, Jeju (2012), http://www.aclweb.org/anthology/P12-1083
11. Zhou, Y., Xue, N.: Pdtb-style discourse annotation of chinese text. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 69–77. Association for Computational Linguistics, Jeju Island (2012), http://www.aclweb.org/anthology/P12-1008