Chinese Comma Disambiguation for Discourse Analysis

Yaqin Yang
Brandeis University
415 South Street
Waltham, MA 02453, USA
yaqin@brandeis.edu

Nianwen Xue
Brandeis University
415 South Street
Waltham, MA 02453, USA
xuen@brandeis.edu

Abstract

The Chinese comma signals the boundary of discourse units and also anchors discourse relations between adjacent text spans. In this work, we propose a discourse structure-oriented classification of the comma that can be automatically extracted from the Chinese Treebank based on syntactic patterns. We then experimented with two supervised learning methods that automatically disambiguate the Chinese comma based on this classification. The first method integrates comma classification into parsing, and the second method adopts a "post-processing" approach that extracts features from automatic parses to train a classifier. The experimental results show that the second approach compares favorably against the first approach.

1 Introduction

The Chinese comma, which looks graphically very similar to its English counterpart, is functionally quite different. It has attracted a significant amount of research that studied the problem from the viewpoint of natural language processing. For example, Jin et al (2004) and Li et al (2005) view the disambiguation of the Chinese comma as a way of breaking up long Chinese sentences into shorter ones to facilitate parsing. The idea is to split a long sentence into multiple comma-separated segments, parse them individually, and reconstruct the syntactic parse for the original sentence. Although both studies show a positive impact of this approach, comma disambiguation is viewed merely as a convenient tool to help achieve a more important goal. Xue and Yang (2011) point out that the very reason for the existence of these long Chinese sentences is because the Chinese comma is ambiguous and in some context, it identifies the boundary of a sentence just as a period, a question mark, or an exclamation mark does. The disambiguation of comma is viewed as a necessary step to detect sentence boundaries in Chinese and it can benefit a whole range of downstream NLP applications such as syntactic parsing and Machine Translation. In Machine Translation, for example, it is very typical for "one" Chinese sentence to be translated into multiple English sentences, with each comma-separated segment corresponding to one English sentence. In the present work, we expand this view and propose to look at the Chinese comma in the context of discourse analysis. The Chinese comma is viewed as a delimiter of elementary discourse units (EDUs), in the sense of the Rhetorical Structure Theory (Carlson et al., 2002; Mann et al., 1988). It is also considered to be the anchor of discourse relations, in the sense of the Penn Discourse Treebank (PDT) (Prasad et al., 2008). 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.

Developing a supervised or semi-supervised model of discourse segmentation would require ground truth annotated based on a well-established representation scheme, but as of right now no such annotation exists for Chinese to the best of our knowledge. However, syntactically annotated treebanks often contain important clues that can be used to infer discourse-level information. We present

Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 786–794,
Jeju, Republic of Korea, 8-14 July 2012.

a method of automatically deriving a preliminary form of discourse structure anchored by the Chinese comma from the Penn Chinese Treebank (CTB) (Xue et al., 2005), and using this information to train and test supervised models. This discourse information is formalized as a classification of the Chinese comma, with each class representing the boundary of an elementary discourse unit as well as the anchor of a coarse-grained discourse relation between the two discourse units that it delimits. We then develop two comma classification methods. In the first method, we replace the part-of-speech (POS) tag of each comma in the CTB with a derived discourse category and retrain a state-of-the-art Chinese parser on the relabeled data. We then evaluate how accurately the commas are classified in the parsing process. In the second method, we parse these sentences and extract lexical and syntactic information as features to predict these new discourse categories. The second approach gives us more control over what features to extract and our results show that it compares favorably against the first approach.

The rest of the paper is organized as follows. In Section 2, we present our approach to automatically extract discourse information from a syntactically annotated treebank and present our classification scheme. In Section 3, we describe our supervised learning methods and the features we extracted. Section 4 presents our experiment setup and experimental results. Related work is reviewed in Section 5. We conclude in Section 6.

2 Chinese comma classification

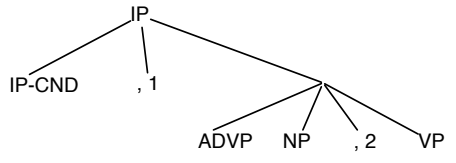
There are many ways to conceptualize the discourse structure of a text (Mann et al., 1988; Prasad et al., 2008), but there is more of a consensus among researchers about the fundamental building blocks of the discourse structure. For the Rhetorical Discourse Theory, the building blocks are Elementary Discourse Units (EDUs). For the PDT, the building blocks are abstract objects such as propositions, facts. Although they are phrased in different ways, syntactically these discourse units are generally realized as clauses or built on top of clauses. So the first step in building the discourse structure of a text is to identify these discourse units.

In Chinese, these elementary discourse units are generally delimited by the comma, but not all commas mark the boundaries of a discourse unit. In (1), for example, Comma [1] marks the boundary of a discourse unit while Comma [2] does not. This is reflected in its English translation: while the first comma corresponds to an English comma, the second comma is not translated at all, as it marks the boundary between a subject and its predicate, where no comma is needed in English. Disambiguating these two types of commas is thus an important first step in identifying elementary discourse units and building up the discourse structure of a text.

(1) 王翔 虽 年 过 半百，[1] 但
Wang Xiang although age over 50 , but
其 充沛 的 精力 和 敏捷 的
his abundant DE energy and quick DE
思维 ，[2] 给 人 一 个 挑战者
thinking , give people one CL challenger
的 印象 。
DE impression .

"Although Wang Xiang is over 50 years old, his abundant energy and quick thinking leave people the impression of a challenger."

Although to the best of our knowledge, no such discourse segmented data for Chinese exists in the public domain, this information can be extracted from the syntactic annotation of the CTB. In the syntactic annotation of the sentence, illustrated in (a), it is clear that while the first comma in the sentence marks the boundary of a clause, the second one marks the demarcation between the subject NP and the predicate VP and thus is not an indicator of a discourse boundary.



(a)

In addition to a binary distinction of whether a comma marks the boundary of a discourse unit, the CTB annotation also allows the extraction of a more elaborate classification of commas based on coordination and subordination relations of comma-separated clauses. This classification of the Chinese

comma can be viewed as a first approximation of the discourse relations anchored by the comma that can be refined later via a manual annotation process.

Based on the syntactic annotation in the CTB, we classify the Chinese comma into seven hierarchically organized categories, as illustrated in Figure 1. The first distinction is made between commas that indicate a discourse boundary (RELATION) and those that do not (OTHER). Commas that indicate discourse boundaries are further divided into commas that separate coordinated discourse units (COORD) vs commas that separate discourse units in a subordination relation (SUBORD). Based on the levels of embedding and the syntactic category of the coordinated structures, we define three different types of coordination (SB, IP_COORD and VP_COORD). We also define three types of subordination relations (ADJ, COMP, Sent_SBJ), based on the syntactic structure. As we will show below, each of the six relations has a clear syntactic pattern that can be exploited for their automatic detection.

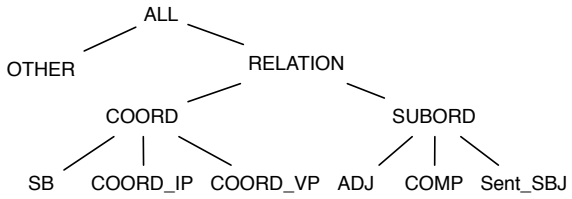


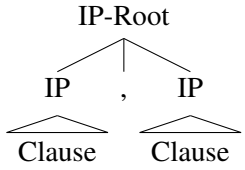
Figure 1: Comma classification

Sentence Boundary (SB): Following (Xue and Yang, 2011), we consider 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. This is illustrated in (2), where a Chinese sentence can be split into two independent shorter sentences at the comma. We view this comma to be a marker of the sentence boundary and it serves the same function as the unambiguous sentence boundary delimitors (periods, question marks, exclamation marks) in Chinese. The syntactic pattern that is used to infer this relation is illustrated in (b).

(2) 广东省 建立 了 自然
Guangdong province establish ASP natural
科学 基金 ，[3] 每年
science foundation , every year
投入 在一亿 元
investment at one hundred millioin yuan
以上 。
above .

"Natural Science Foundation is established in Guangdong Province. More than one hundred million yuan is invested every year."

(b)

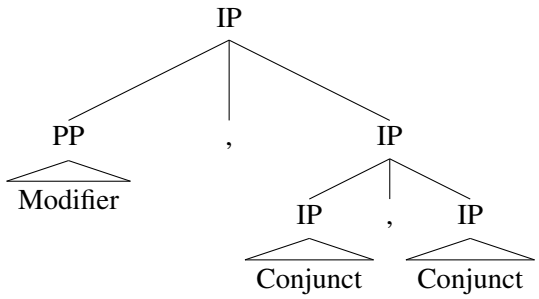


IP Coordination (IP_COORD): Coordinated IPs that are not the immediate children of the root IP are also considered to be discourse units and the commas linking them are labeled IP_COORD. Different from the sentence boundary cases, these coordinated IPs are often embedded in a larger structure. An example is given in (3) and its typical syntactic pattern is illustrated in (c).

(3) 据 陆仁法 介绍 ，[4]
According to Lu Renfa presentation ,
全国 税收 任务已
the whole country revenue goal already
超额 完成 ，[5] 总体
exceeding quota complete , overall
情况 比较 好。
situation fairly good .

"According to Lu Renfa, the national revenue goal is met and exceeded, and the overall situation is fairly good."

(c)



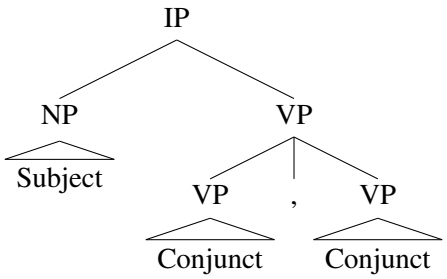
VP Coordination (VP_COORD): Coordinated VPs, when separated by the comma, are not semantically different from coordinated IPs. The only difference is that in the latter case, the coordinated VPs

share a subject, while coordinated IPs tend to have different subjects. Maintaining this distinction allow us to model subject (dis)continuity, which helps recover a subject when it is dropped, a prevalent phenomenon in Chinese. As shown in (4), the VPs in the text spans separated by Comma [6] have the same subject, thus the subject in the second VP is dropped. The syntactic pattern that allows us to extract this structure is given in (d).

(4) 中国 银行 是四大 国有
China Bank is four major state-owned
商业 银行之一 ，[6] 也 是
commercial bank one of these , also is
中国 的 主要 外汇 银行 。
China DE major foreign exchange bank .

"Bank of China is one of the four major state-owned commercial banks, and it is also China's major foreign exchange bank."

(d)

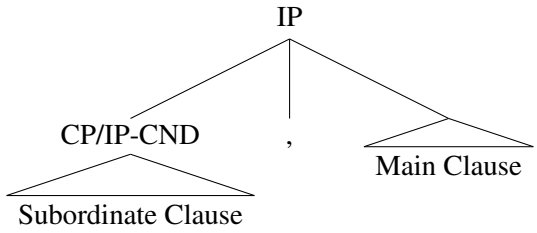


Adjunction (ADJ): Adjunction is one of three types of subordination relations we define. It holds between a subordinate clause and its main clause. The subordinate clause is normally introduced by a subordinating conjunction and it typically provides the cause, purpose, manner, or condition for the main clause. In the PDT terms, these subordinate conjunctions are discourse connectives that anchor a discourse relation between the subordinate clause and the main clause. In Chinese, with few exceptions, the subordinate clause comes before the main clause. (5) is an example of this relation.

(5) 若工程 发生 保险 责任 范围
if project happen insurance liability scope
内 的 自然 灾害 ，[7]
inside DE natural disaster ,
中保 财产 保险 公司
China Insurance property insurance company
将 按 规定 进行
will according to provision excecute
赔偿 。
compensation .

"If natural disasters within the scope of the insurance liability happen in the project, PICC Property Insurance Company will provide compensations according to the provisions."

(e)



(e) shows how (5) is represented in the syntactic structure in the CTB. Extracting this relation requires more than just the syntactic configuration between these two clauses. We also take advantage of the functional (dash) tags provided in the treebank. The functional tags are attached to the subordinate clause and they include CND (conditional), PRP (purpose or reason), MNR (manner), or ADV (other types of subordinate clauses that are adjuncts to the main clause).

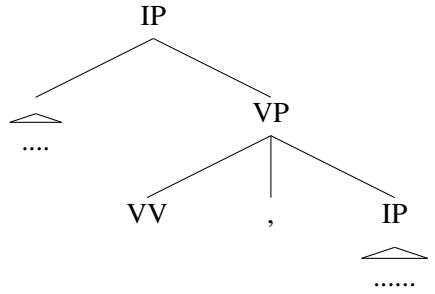
Complementation (COMP): When a comma separates a verb governor and its complement clause, this verb and its subject generally describe the attribution of the complement clause. Attribution is an important notion in discourse analysis in both the RST framework and in the PDT. An example of this is given in (6), and the syntactic pattern used to extract this relation is illustrated in (f).

(6) 该 公司 介绍 ，[8] 在未来 的
The company present , at future DE
五年 内 他们将 追加 投资
five year within they will additionally invest
九千万 美元 ，[9] 预计
ninety million U.S. dollars , estimate
年产值 可 达
annual output will reach
三亿 美元 。
three hundred million U.S. dollars .

"According to the the company's presentation, they will invest an additional ninety million

U.S. dollars in the next five years, and the estimated annual output will reach $ 300 million."

(f)

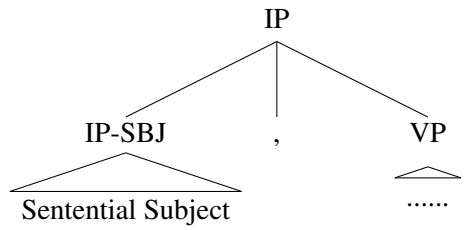


Sentential Subject (SBJ): This category is for commas that separate a sentential subject from its predicate VP. An example is given in (7) and the syntactic pattern used to extract this relation is illustrated in (g).

(7) 出口 快速 增长 , [10] 成为 推动
export rapid grow , become promote
经济 增长 的 重要 力量 。
economy growth DE important force .

"The rapid growth of export becomes an important force in promoting economic growth."

(g)



Others (OTHER): The remaining cases of comma receive the OTHER label, indicating they do not mark the boundary of a discourse segment.

Our proposed comma classification scheme serves the dual purpose of identifying elementary discourse units and at the same time detecting coarse-grained discourse relations anchored by the comma. The discourse relations identified in this manner by no means constitute the full discourse analysis of a text, they are, however, a good first approximation. The advantage of our approach is that we do not require manual discourse annotations, and all the information we need is automatically extracted from the syntactic annotation of the CTB and attached to instances of the comma in the corpus. This makes it possible for us to train supervised models to automatically classify the commas in any Chinese text.

3 Two comma classification methods

Given the gold standard parses, based on the syntactic patterns described in Section 2, we can map the POS tag of each comma instance in the CTB to one of the seven classes described in Section 2. Using this relabeled data as training data, we experimented with two automatic comma disambiguation methods. In the first method, we simply retrained the Berkeley parser (Petrov and Klein, 2007) on the relabeled data and computed how accurately the commas are labeled in a held-out test set. In the second method, we trained a Maximum Entropy classifier with the Mallet (McCallum et al., 2002) machine learning package to classify the commas. The features are extracted from the CTB data automatically parsed with the Berkeley parser. We implemented features described in (Xue and Yang, 2011), and also experimented with a set of new features as follows. In general, these new features are extracted from the two text spans surrounding the comma. Given a comma, we define the preceding text span as i span and the following text span as j span. We also collected a number of subject-predicate pairs from a large corpus that doesn't overlap with the CTB. We refer to this corpus as the auxiliary corpus.

Subject and Predicate features: We explored various combinations of the subject (sbj), predicate ($pred$) and object (obj) of the two spans. The subject of i span is represented as sbj_i, etc.

1. The existence of sbj_i, sbj_j, both, or neither.
2. The lemma of $pred_i$, the lemma of $pred_j$, the conjunction of sbj_i and $pred_j$, the conjunction of $pred_i$ and sbj_j
3. whether the conjunction of sbj_i and $pred_j$ occurs more than 2 times in the auxiliary corpus when j does not have a subject.
4. whether the conjunction of obj_i and $pred_j$ occurs more than 2 times in the auxiliary corpus when j does not have a subject
5. Whether the conjunction of $pred_i$ and sbj_j occurs more than 2 times in the auxiliary corpus when i does not have a subject.

Mutual Information features: Mutual information is intended to capture the association strength between the subject of a previous span and the predicate of the current span. We use Mutual Information

(Church and Hanks, 1989) as shown in Equation (1) and the frequency count computed based on the auxiliary corpus to measure such constraints.

$$MI = \log_2 \frac{\text{\# co-occur of S and P * corpus size}}{\text{\# S occur * \# P occur}} \quad (1)$$

1. The conjunction of sbj_i and $pred_j$ when j does not have a subject if their MIvalue is greater than -8.0, an empirically established threshold.
2. Whether obj_i and $pred_j$ has an MI value greater than 5.0 if j does not have a subject.
3. Whether the MI value of sbj_i and $pred_j$ is greater than 0.0, and they occur 2 times in the auxiliary corpus when j doesn't have a subject.
4. Whether the MI value of obj_i and $pred_j$ is greater than 0.0 and they occur 2 times in the auxiliary corpus when j doesn't have a subject.
5. Whether the MI value of $pred_i$ and sbj_j is greater than 0.0 and they occur more than 2 times in the auxiliary corpus when i does not have a subject.

Span features: We used span features to capture syntactic information, e.g. the comma separated spans are constituents in Tree (b) but not in Tree (d).

1. Whether i forms a single constituent, whether j forms a single constituent.
2. The conjunction and hierarchical relation of all constituent labels in i/j, if i/j does not form a single constituent. The conjunction of all constituent labels in both spans, if neither span form a single constituent.

Lexical features:

1. The first word in i if it is an adverb, the first word in j if it is an adverb.
2. The first word in i span if it is a coordinating conjunction, the first word in j if it is a coordinating conjunction.

4 Experiments

4.1 Datasets

We use the CTB 6.0 in our experiments and divide it into training, development and test sets using the data split recommended in the CTB 6.0 documentation, as shown in Table 1. There are 5436 commas in the test set, including 1327 commas that are sentence boundaries (SB), 539 commas that connect coordinated IPs (IP_COORD), 1173 commas that join coordinated VPs (VP_COORD), 379 commas that delimits a subordinate clause and its main clause (ADJ), 314 commas that anchor complementation relations (COMP), and 1625 commas that belong to the OTHER category.

4.2 Results

As mentioned in Section 3, we experimented with two comma classification methods. In the first method, we replace the part-of-speech (POS) tags of the commas with the seven classes defined in Section 2. We then retrain the Berkeley parser (Petrov and Klein, 2007) using the training set as presented in Table 1, parse the test set, and evaluate the comma classification accuracy.

In the second method, we use the relabeled commas as the gold-standard data to train a supervised classifier to automatically classify the commas. As shown in the previous section, syntactic structures are an important source of information for our classifier. For feature extraction purposes, the entire CTB6.0 is automatically parsed in a round-robin fashion. We divided CTB 6.0 into 10 portions, and parsed each portion with a model trained on other portions, using the Berkeley parser (Petrov and Klein, 2007). Measured by the ParsEval metric (Black et al., 1991), the parsing accuracy on the CTB test set stands at 83.29% (F-score), with a precision of 85.18% and a recall of 81.49%.

The results are presented in Table 2, which shows the overall accuracy of the two methods as well as the results for each individual category. As should be clear from Table 2, the results for the two methods are very comparable, with the second method performing modestly better than the first method.

4.2.1 Subject continuity

One of the goals for this classification scheme is to model subject continuity, which answers the question of how accurately we can predict whether two comma-separated text spans have the same subject or different subjects. When the two spans share the same subject, the comma belongs to the category VP_COORD. When they have different subjects, they belong to the categories IP_COORD or

Data	Train	Dev	Test
CTB-6.0	81-325, 400-454, 500-554 590-596, 600-885, 900 1001-1017, 1019, 1021-1035 1037-1043, 1045-1059,1062-1071 1073-1078, 1100-1117, 1130-1131 1133-1140, 1143-1147, 1149-1151 2000-2139, 2160-2164, 2181-2279 2311-2549, 2603-2774, 2820-3079	41-80 1120-1129 2140-2159 2280-2294 2550-2569 2775-2799 3080-3109	(1-40,901-931 newswire) (1018, 1020, 1036, 1044 1060-1061, 1072, 1118-1119, 1132 1141-1142, 1148 magazine) (2165-2180, 2295-2310 2570-2602, 2800-2819 3110-3145 broadcast news)

Table 1: CTB 6.0 data set division.

SB. When this question is meaningless, e.g., when one of the span does not even have a subject, the comma belongs to other categories. To evaluate the performance of our model on this problem, we recomputed the results by putting IP_COORD and SB in one category, putting VP_COORD in another category and the rest of the labels in a third category. The results are presented in Table 3.

4.2.2 The effect of genre

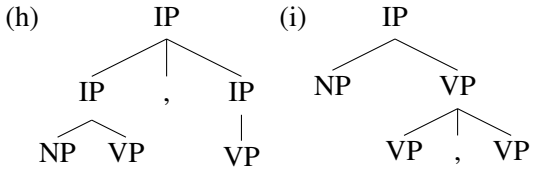
CTB 6.0 consists of data from three different genres, including newswire, magazine and broadcast news. Data genres may have very different characteristics. To evaluate how our model works on different genres, we train a model using training and development sets, and test the model on different genres as described in Table 1. The results on these three genres are presented in Table 4, and they shows a significant fluctuation across genres. Our model works the best on newswire, but not as good on broadcast news and magazine articles.

4.2.3 Comparison with prior work

(Xue and Yang, 2011) presented results on a binary classification of whether or not a comma marks a sentence boundary, while the present work addresses a multi-category classification problem aimed at identifying discourse segments and preliminary discourse relations anchored by the comma. However, since we also have a SB category, comparison is possible. For comparison purposes, we retrained our model on their data sets, and computed the results of SB vs other categories. The results are shown in Table 5. Our results are very comparable with (Xue and Yang, 2011) despite that we are performing a multicategory classification.

4.3 Error analysis

Even though our feature-based approach can theoretically "correct" parsing errors, meaning that a comma can in theory be classified correctly even if a sentence is incorrectly parsed, when examining the system output, errors in automatic parses often lead to errors in comma classification. A common parsing error is the confusion between Structures (h) and (i). If the subject of the text span after a comma is dropped as shown in (h), the parser often produces a VP coordination structure as shown in (i) and vice versa. This kind of parsing errors would lead to errors in our syntactic features and thus directly affect the accuracy of our model.



5 Related Work

There is a large body of work on discourse analysis in the field of Natural Language Processing. Most of the work, however, are on English. An unsupervised approach was proposed to recognize discourse relations in (Marcu and Echihabi, 2002), which extracts discourse relations that hold between arbitrary spans of text making use of cue phrases. Like the present work, a lot of research on discourse analysis is carried out at the sentence level. (Soricut and Marcu, 2003; Sporleder and Lapata, 2005; Polanyi et al., 2004). (Soricut and Marcu, 2003) and (Polanyi et al., 2004) implement models to perform discourse parsing, while (Sporleder and Lapata, 2005) introduces discourse chunking as an alternative to full-

Class	Metric	Method 1	Method 2
all	acc. (%)	71.5	72.9
SB	Prec. (%)	65.6	66.2
	Rec. (%)	71.7	73.1
	F. (%)	68.5	69.5
IP_COORD	Prec. (%)	53.3	56.0
	Rec. (%)	50.5	48.6
	F. (%)	52.0	52.0
VP_Coord	Prec. (%)	65.6	68.3
	Rec. (%)	76.3	78.2
	F. (%)	70.5	72.9
ADJ	Prec. (%)	66.9	66.8
	Rec. (%)	29.3	37.7
	F. (%)	40.8	48.2
Comp	Prec. (%)	88.3	91.2
	Rec. (%)	93.9	92.4
	F. (%)	91.0	91.8
SentSBJ	Prec. (%)	25.0	31.8
	Rec. (%)	6	10
	F. (%)	9.7	15.6
Other	Prec. (%)	86.9	85.6
	Rec. (%)	83.4	84.1
	F. (%)	85.1	84.8

Table 2: Overall accuracy of the two methods as well as the results for each individual category.

scale discourse parsing.

The emergence of linguistic corpora annotated with discourse structure such as the RST Discourse Treebank (Carlson et al., 2002) and PDT (Miltsakaki et al., 2004; Prasad et al., 2008) have changed the landscape of discourse analysis. More robust, data-driven models are starting to emerge.

Compared with English, much less work has been done in Chinese discourse analysis, presumably due to the lack of discourse resources in Chinese. (Huang and Chen, 2011) constructs a small corpus following the PDT annotation scheme and trains a statistical classifier to recognize discourse relations. Their work, however, is only concerned with discourse relations between adjacent sentences, thus side-stepping the hard problem of disambiguating the Chinese comma and analyzing intra-sentence discourse relations. To the best of our knowledge, our work is the first in attempting to disambiguating the Chinese comma as the first step in performing Chinese discourse analysis.

	Prec. (%)	Rec. (%)	F. (%)
VP_COORD	68.3	78.2	72.9
IP_COORD+SB	76.0	78.7	77.3
Other	89.0	80.2	84.4

Table 3: Subject continuity results based on Maximum Entropy model

Genre	NW	BN	MZ
Accuracy. (%)	79.1	73.6	67.7

Table 4: Results on different genres based on Maximum Entropy model

	Xue and Yang			our model		
(%)	p	r	f1	p	r	f1
Overall			89.2			88.7
EOS	64.7	76.4	70.1	63.0	77.9	69.7
NEOS	95.1	91.7	93.4	95.3	90.8	93.0

Table 5: Comparison of (Xue and Yang, 2011) and the present work based on Maximum Entropy model

6 Conclusions and future work

We proposed a approach to disambiguate the Chinese comma as a first step toward discourse analysis. Training and testing data are automatically derived from a syntactically annotated corpus. We presented two automatic comma disambiguation methods that perform comparably. In the first method, comma disambiguation is integrated into the parsing process while in the second method we train a supervised classifier to classify the Chinese comma, using features extracted from automatic parses. Much needs to be done in the area, but we believe our work provides insight into the intricacy and complexity of discourse analysis in Chinese.

Acknowledgment

This work is supported by the IIS Division of National Science Foundation via Grant No. 0910532 entitled "Richer Representations for Machine Translation". All views expressed in this paper are those of the authors and do not necessarily represent the view of the National Science Foundation.

References

L Carlson, D Marcu, M E Okurowski. 2002. *RST Discourse Treebank*. Linguistic Data Consortium 2002.

Caroline Sporleder, Mirella Lapata. 2005. *Discourse chunking and its application to sentence compression*. In Proceedings of HLT/EMNLP 2005.

Livia Polanyi, Chris Culy, Martin Van Den Berg, Gian Lorenzo Thione and David Ahn. 2004. *Sentential structure and discourse parsing*. In Proceeedings of the ACL 2004 Workshop on Discourse Annotation 2004.

Hen-Hsen Huang and Hsin-Hsi Chen. 2011. *Chinese Discourse Relation Recognition*. In Proceedings of the 5th International Joint Conference on Natural Language Processing 2011,pages 1442-1446.

Daniel Marcu and Abdessamad Echihabi. 2002. *An Unsupervised Approach to Recognizing Discourse Relations*. In Proceedings of the ACL, July 6-12, 2002, Philadelphia, PA, USA.

Radu Soricut and Daniel Marcu. 2003. *Sentence Level Discourse Parsing using Syntactic and Lexical Information*. In Proceedings of the ACL 2003.

Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi andBonnie Webber. 2004. *The Penn Discourse Treebank*. In Proceedings of LREC 2004.

Nianwen Xue and Yaqin Yang. 2011. *Chinese sentence segmentation as comma classification*. In Proceedings of ACL 2011.

Nianwen Xue, Fei Xia, Fu-Dong Chiou and Martha Palmer. 2005. *The Penn Chinese Treebank: Phrase Structure Annotation of a Large Corpus*. Natural Language Engineering, 11(2):207-238.

Slav Petrov and Dan Klein. 2007. *Improved Inferencing for Unlexicalized Parsing*. In Proceedings of HLT-NAACL 2007.

E. Black, S. Abney, D. Flickinger, C. Gdaniec, R. Grishman, P. Harrison, D. Hindle, R. Ingria, F. Jelinek, J. Klavans, M. Liberman, M. Marcus, S. Roukos, B. Santorini, and T. Strzalkowski. 1991. *A procedure for quantitively comparing the syntactic coverage of English grammars*. In Proceedings of the DARPA Speech and Natural Language Workshop, pages 306-311.

Mann, William C. and Sandra A. Thompson. 1988. *Rhetorical Structure Theory: Toward a functional theory of text organization*. Text 8 (3): 243-281.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. *The Penn Discourse Treebank 2.0.*. In Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008).

Meixun Jin, Mi-Young Kim, Dong-Il Kim, and Jong-Hyeok Lee. 2004. *Segmentation of Chinese Long Sentences Using Commas*. In Proceedings of the SIGHANN Workshop on Chinese Language Processing.

Xing Li, Chengqing Zong, and Rile Hu. 2005. *A Hierarchical Parsing Approach with Punctuation Processing for Long Sentence Sentences*. In Proceedings of the Second International Joint Conference on Natural Language Processing: Companion Volume including Posters/Demos and Tutorial Abstracts.

Andrew Kachites McCallum. 2002. *MALLET: A Machine Learning for Language Toolkit*. http://mallet.cs.umass.edu.

Church, K., and Hanks, P. 1989. *Word Association Norms, Mutual Information and Lexicography*. Association for Computational Linguistics, Vancouver , Canada