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.