An Empirical Examination of Challenges in Chinese Parsing

Jonathan K. Kummerfeld[†] **Daniel Tse**[‡] **James R. Curran**[‡] **Dan Klein**[†]

[†]Computer Science Division
University of California, Berkeley
Berkeley, CA 94720, USA
{jkk,klein}@cs.berkeley.edu

[‡]School of Information Technology
University of Sydney
Sydney, NSW 2006, Australia
{dtse6695,james}@it.usyd.edu.au

Abstract

Aspects of Chinese syntax result in a distinctive mix of parsing challenges. However, the contribution of individual sources of error to overall difficulty is not well understood. We conduct a comprehensive automatic analysis of error types made by Chinese parsers, covering a broad range of error types for large sets of sentences, enabling the first empirical ranking of Chinese error types by their performance impact. We also investigate which error types are resolved by using gold part-of-speech tags, showing that improving Chinese tagging only addresses certain error types, leaving substantial outstanding challenges.

1 Introduction

A decade of Chinese parsing research, enabled by the Penn Chinese Treebank (PCTB; Xue et al., 2005), has seen Chinese parsing performance improve from 76.7 F_1 (Bikel and Chiang, 2000) to 84.1 F_1 (Qian and Liu, 2012). While recent advances have focused on understanding and reducing the errors that occur in segmentation and part-of-speech tagging (Qian and Liu, 2012; Jiang et al., 2009; Forst and Fang, 2009), a range of substantial issues remain that are purely syntactic.

Early work by Levy and Manning (2003) presented modifications to a parser motivated by a manual investigation of parsing errors. They noted substantial differences between Chinese and English parsing, attributing some of the differences to treebank annotation decisions and others to meaningful differences in syntax. Based on this analysis they considered how to modify their parser to capture the information necessary to model the syntax within the PCTB. However, their manual analysis was limited in scope, covering only part of the parser output, and was unable to characterize the relative impact of the issues they uncovered.

This paper presents a more comprehensive analysis of errors in Chinese parsing, building on the technique presented in Kummerfeld et al. (2012), which characterized the error behavior of English parsers by quantifying how often they make errors such as PP attachment and coordination scope. To accommodate error classes that are absent in English, we augment the system to recognize Chinese-specific parse errors.[1] We use the modified system to show the relative impact of different error types across a range of Chinese parsers.

To understand the impact of tagging errors on different error types, we performed a part-of-speech ablation experiment, in which particular confusions are introduced in isolation. By analyzing the distribution of errors in the system output with and without gold part-of-speech tags, we are able to isolate and quantify the error types that can be resolved by improvements in tagging accuracy.

Our analysis shows that improvements in tagging accuracy can only address a subset of the challenges of Chinese syntax. Further improvement in Chinese parsing performance will require research addressing other challenges, in particular, determining coordination scope.

2 Background

The closest previous work is the detailed manual analysis performed by Levy and Manning (2003). While their focus was on issues faced by their factored PCFG parser (Klein and Manning, 2003b), the error types they identified are general issues presented by Chinese syntax in the PCTB. They presented several Chinese error types that are rare or absent in English, including noun/verb ambiguity, NP-internal structure and coordination ambiguity due to *pro*-drop, suggesting that closing the English-Chinese parsing gap demands techniques

[1] The system described in this paper is available from http://code.google.com/p/berkeley-parser-analyser/

beyond those currently used for English. However, as noted in their final section, their manual analysis of parse errors in 100 sentences only covered a portion of a single parser's output, limiting the conclusions they could reach regarding the distribution of errors in Chinese parsing.

2.1 Automatic Error Analysis

Our analysis builds on Kummerfeld et al. (2012), which presented a system that automatically classifies English parse errors using a two stage process. First, the system finds the shortest path from the system output to the gold annotations, where each step in the path is a tree transformation, fixing at least one bracket error. Second, each transformation step is classified into one of several error types.

When directly applied to Chinese parser output, the system placed over 27% of the errors in the catch-all 'Other' type. Many of these errors clearly fall into one of a small set of error types, motivating an adaptation to Chinese syntax.

3 Adapting error analysis to Chinese

To adapt the Kummerfeld et al. (2012) system to Chinese, we developed a new version of the second stage of the system, which assigns an error category to each tree transformation step.

To characterize the errors the original system placed in the 'Other' category, we looked through one hundred sentences, identifying error types generated by Chinese syntax that the existing system did not account for. With these observations we were able to implement new rules to catch the previously missed cases, leading to the set shown in Table 1. To ensure the accuracy of our classifications, we alternated between refining the classification code and looking at affected classifications to identify issues. We also periodically changed the sentences from the development set we manually checked, to avoid over-fitting.

Where necessary, we also expanded the information available during classification. For example, we use the structure of the final gold standard tree when classifying errors that are a byproduct of sense disambiguation errors.

4 Chinese parsing errors

Table 1 presents the errors made by the Berkeley parser. Below we describe the error types that are either new in this analysis, have had their definition altered, or have an interesting distribution.[2]

Error Type	Brackets	*(% of total)*
NP-internal*	6019	*(22.70%)*
Coordination	2781	*(10.49%)*
Verb taking wrong args*	2310	*(8.71%)*
Unary	2262	*(8.53%)*
Modifier Attachment	1900	*(7.17%)*
One Word Span	1560	*(5.88%)*
Different label	1418	*(5.35%)*
Unary A-over-A	1208	*(4.56%)*
Wrong sense/bad attach*	1018	*(3.84%)*
Noun boundary error*	685	*(2.58%)*
VP Attachment	626	*(2.36%)*
Clause Attachment	542	*(2.04%)*
PP Attachment	514	*(1.94%)*
Split Verb Compound*	232	*(0.88%)*
Scope error*	143	*(0.54%)*
NP Attachment	109	*(0.41%)*
Other	3186	*(12.02%)*

Table 1: Errors made when parsing Chinese. Values are the number of bracket errors attributed to that error type. The values shown are for the Berkeley parser, evaluated on the development set. * indicates error types that were added or substantially changed as part of this work.

In all of our results we follow Kummerfeld et al. (2012), presenting the number of bracket errors (missing or extra) attributed to each error type. Bracket counts are more informative than a direct count of each error type, because the impact on EVALB F-score varies between errors, e.g. a single attachment error can cause 20 bracket errors, while a unary error causes only one.

NP-internal. (Figure 1a**).** Unlike the Penn Treebank (Marcus et al., 1993), the PCTB annotates some NP-internal structure. We assign this error type when a transformation involves words whose parts of speech in the gold tree are one of: CC, CD, DEG, ETC, JJ, NN, NR, NT and OD.

We investigated the errors that fall into the NP-internal category and found that 49% of the errors involved the creation or deletion of a single pretermianl phrasal bracket. These errors arise when a parser proposes a tree in which POS tags (for instance, JJ or NN) occur as siblings of phrasal tags (such as NP), a configuration used by the PCTB bracketing guidelines to indicate complementation as opposed to adjunction (Xue et al., 2005).

[2] For an explanation of the English error types, see Kummerfeld et al. (2012).

Verb taking wrong args. (Figure 1b). This error type arises when a verb (e.g. 扭转 *reverse*) is hypothesized to take an incorrect argument (布什 *Bush* instead of 地位 *position*). Note that this also covers some of the errors that Kummerfeld et al. (2012) classified as NP Attachment, changing the distribution for that type.

Unary. For mis-application of unary rules we separate out instances in which the two brackets in the production have the the same label (A-over-A). This cases is created when traces are eliminated, a standard step in evaluation. More than a third of unary errors made by the Berkeley parser are of the A-over-A type. This can be attributed to two factors: (i) the PCTB annotates non-local dependencies using traces, and (ii) Chinese syntax generates more traces than English syntax (Guo et al., 2007). However, for parsers that do not return traces they are a benign error.

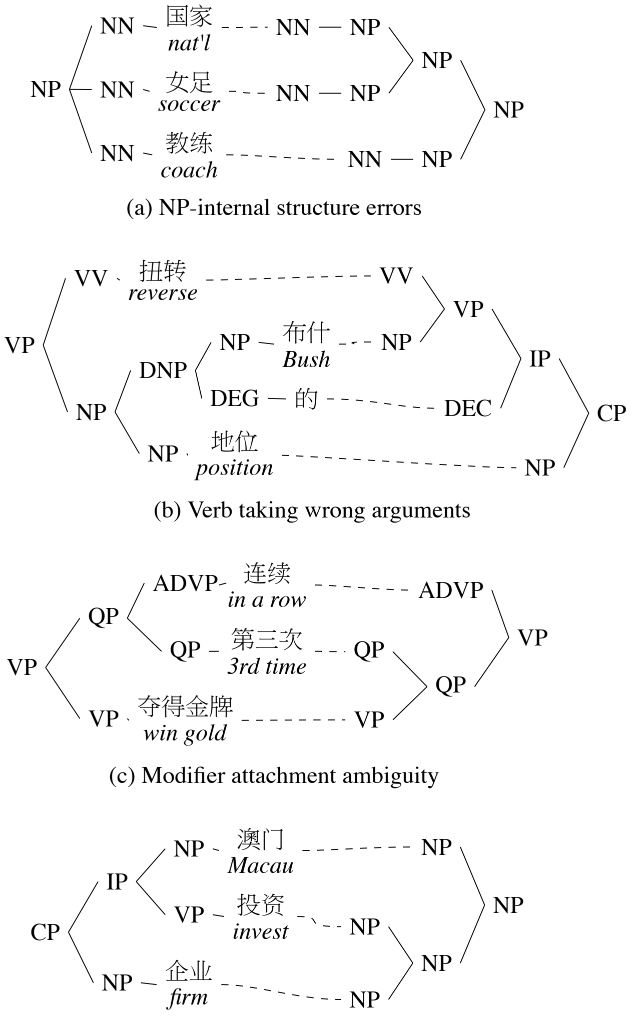
Modifier attachment. (Figure 1c). Incorrect modifier scope caused by modifier phrase attachment level. This is less frequent in Chinese than in English: while English VP modifiers occur in pre- and post-verbal positions, Chinese only allows pre-verbal modification.

Wrong sense/bad attach. (Figure 1d). This applies when the head word of a phrase receives the wrong POS, leading to an attachment error. This error type is common in Chinese because of POS fluidity, e.g. the well-known Chinese verb/noun ambiguity often causes mis-attachments that are classified as this error type.

In Figure 1d, the word 投资 *invest* has both noun and verb senses. While the gold standard interpretation is the relative clause *firms that Macau invests in*, the parser returned an NP interpretation *Macau investment firms*.

Noun boundary error. In this error type, a span is moved to a position where the POS tags of its new siblings all belong to the list of NP-internal structure tags which we identified above, reflecting the inclusion of additional material into an NP.

Split verb compound. The PCTB annotations recognize several Chinese verb compounding strategies, such as the serial verb construction (规划建设 *plan [and] build*) and the resultative construction (煮熟 *cook [until] done*), which join a bare verb to another lexical item. We introduce an error type specific to Chinese, in which such verb compounds are split, with the two halves of the compound placed in different phrases.



(a) NP-internal structure errors

(b) Verb taking wrong arguments

(c) Modifier attachment ambiguity

(d) Sense confusion

Figure 1: Prominent error types in Chinese parsing. The left tree is the gold structure; the right is the parser hypothesis.

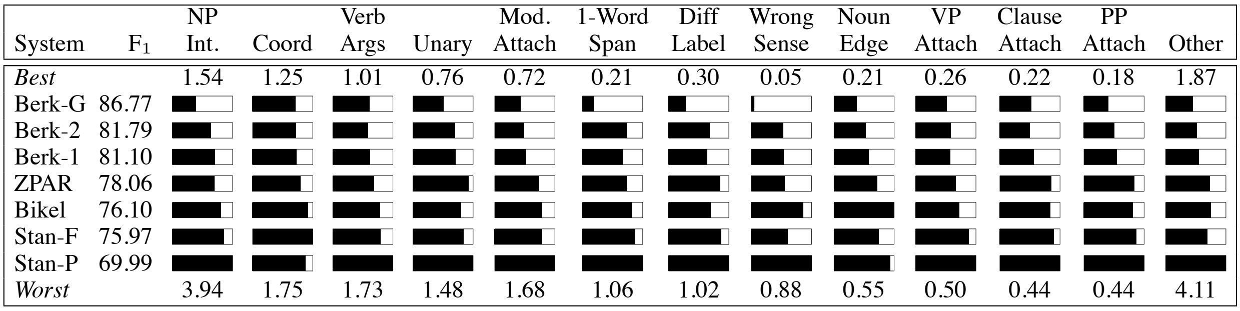
Scope error. These are cases in which a new span must be added to more closely bind a modifier phrase (ADVP, ADJP, and PP).

PP attachment. This error type is rare in Chinese, as adjunct PPs are pre-verbal. It does occur near coordinated VPs, where ambiguity arises about which of the conjuncts the PP has scope over. Whether this particular case is PP attachment or coordination is debatable; we follow Kummerfeld et al. (2012) and label it PP attachment.

4.1 Chinese-English comparison

It is difficult to directly compare error analysis results for Chinese and English parsing because of substantial changes in the classification method, and differences in treebank annotations.

As described in the previous section, the set of error categories considered for Chinese is very different to the set of categories for English. Even for some of the categories that were not substan-



System	F_1	NP Int.	Coord	Verb Args	Unary	Mod. Attach	1-Word Span	Diff Label	Wrong Sense	Noun Edge	VP Attach	Clause Attach	PP Attach	Other
Best		1.54	1.25	1.01	0.76	0.72	0.21	0.30	0.05	0.21	0.26	0.22	0.18	1.87
Berk-G	86.77													
Berk-2	81.79													
Berk-1	81.10													
ZPAR	78.06													
Bikel	76.10													
Stan-F	75.97													
Stan-P	69.99													
Worst		3.94	1.75	1.73	1.48	1.68	1.06	1.02	0.88	0.55	0.50	0.44	0.44	4.11

Table 2: Error breakdown for the development set of PCTB 6. The area filled in for each bar indicates the average number of bracket errors per sentence attributed to that error type, where an empty bar is no errors and a full bar has the value indicated in the bottom row. The parsers are: the Berkeley parser with gold POS tags as input (Berk-G), the Berkeley product parser with two grammars (Berk-2), the Berkeley parser (Berk-1), the parser of Zhang and Clark (2009) (ZPAR), the Bikel parser (Bikel), the Stanford Factored parser (Stan-F), and the Stanford Unlexicalized PCFG parser (Stan-P).

tially changed, errors may be classified differently because of cross-over between categories between two categories (e.g. between Verb taking wrong args and NP Attachment).

Differences in treebank annotations also present a challenge for cross-language error comparison. The most common error type in Chinese, NP-internal structure, is rare in the results of Kummerfeld et al. (2012), but the datasets are not comparable because the PTB has very limited NP-internal structure annotated. Further characterization of the impact of annotation differences on errors is beyond the scope of this paper.

Three conclusions that can be made are that (i) coordination is a major issue in both languages, (ii) PP attachment is a much greater problem in English, and (iii) a higher frequency of trace-generating syntax in Chinese compared to English poses substantial challenges.

5 Cross-parser analysis

The previous section described the error types and their distribution for a single Chinese parser. Here we confirm that these are general trends, by showing that the same pattern is observed for several different parsers on the PCTB 6 dev set.[3] We include results for a transition-based parser (ZPAR; Zhang and Clark, 2009), a split-merge PCFG parser (Petrov et al., 2006; Petrov and Klein, 2007; Petrov, 2010), a lexicalized parser (Bikel and Chiang, 2000), and a factored PCFG and dependency parser (Levy and Manning, 2003; Klein and Manning, 2003a,b). [4]

Comparing the two Stanford parsers in Table 2, the factored model provides clear improvements on sense disambiguation, but performs slightly worse on coordination.

The Berkeley product parser we include uses only two grammars because we found, in contrast to the English results (Petrov, 2010), that further grammars provided limited benefits. Comparing the performance with the standard Berkeley parser it seems that the diversity in the grammars only assists certain error types, with most of the improvement occurring in four of the categories, while there is no improvement, or a slight decrease, in five categories.

6 Tagging Error Impact

The challenge of accurate POS tagging in Chinese has been a major part of several recent papers (Qian and Liu, 2012; Jiang et al., 2009; Forst and Fang, 2009). The Berk-G row of Table 2 shows the performance of the Berkeley parser when given gold POS tags.[5] While the F_1 improvement is unsurprising, for the first time we can clearly show that the gains are only in a subset of the error types. In particular, tagging improvement will not help for two of the most significant challenges: coordination scope errors, and verb argument selection.

To see which tagging confusions contribute to which error reductions, we adapt the POS ablation approach of Tse and Curran (2012). We consider the POS tag pairs shown in Table 3. To isolate the effects of each confusion we start from the gold

[3] We use the standard data split suggested by the PCTB 6 file manifest. As a result, our results differ from those previously reported on other splits. All analysis is on the dev set, to avoid revealing specific information about the test set.

[4] These parsers represent a variety of parsing methods, though exclude some recently developed parsers that are not publicly available (Qian and Liu, 2012; Xiong et al., 2005).

[5] We used the Berkeley parser as it was the best of the parsers we considered. Note that the Berkeley parser occasionally prunes all of the parses that use the gold POS tags, and so returns the best available alternative. This leads to a POS accuracy of 99.35%, which is still well above the parser's standard POS accuracy of 93.66%.

Confused tags		*Errors*	Δ F$_1$
VV	NN	1055	-2.72
DEC	DEG	526	-1.72
JJ	NN	297	-0.57
NR	NN	320	-0.05

Table 3: The most frequently confused POS tag pairs. Each Δ F$_1$ is relative to Berk-G.

tags and introduce the output of the Stanford tagger whenever it returns one of the two tags being considered.[6] We then feed these "semi-gold" tags to the Berkeley parser, and run the fine-grained error analysis on its output.

VV/NN. This confusion has been consistently shown to be a major contributor to parsing errors (Levy and Manning, 2003; Tse and Curran, 2012; Qian and Liu, 2012), and we find a drop of over 2.7 F_1 when the output of the tagger is introduced. We found that while most error types have contributions from a range of POS confusions, verb/noun confusion was responsible for virtually all of the noun boundary errors corrected by using gold tags.

DEG/DEC. This confusion between the relativizer and subordinator senses of the particle 的 *de* is the primary source of improvements on modifier attachment when using gold tags.

NR/NN and JJ/NN. Despite their frequency, these confusions have little effect on parsing performance. Even within the NP-internal error type their impact is limited, and almost all of the errors do not change the logical form.

7 Conclusion

We have quantified the relative impacts of a comprehensive set of error types in Chinese parsing. Our analysis has also shown that while improvements in Chinese POS tagging can make a substantial difference for some error types, it will not address two high-frequency error types: incorrect verb argument attachment and coordination scope. The frequency of these two error types is also unimproved by the use of products of latent variable grammars. These observations suggest that resolving the core challenges of Chinese parsing will require new developments that suit the distinctive properties of Chinese syntax.

Acknowledgments

We would like to thank the anonymous reviewers for their helpful suggestions. This research was supported by a General Sir John Monash Fellowship to the first author, the Capital Markets CRC under ARC Discovery grant DP1097291, and the NSF under grant 0643742.

References

Daniel M. Bikel and David Chiang. 2000. Two Statistical Parsing Models Applied to the Chinese Treebank. In *Proceedings of the Second Chinese Language Processing Workshop*, pages 1–6. Hong Kong, China.

Martin Forst and Ji Fang. 2009. TBL-improved non-deterministic segmentation and POS tagging for a Chinese parser. In *Proceedings of the 12th Conference of the European Chapter of the ACL*, pages 264–272. Athens, Greece.

Yuqing Guo, Haifeng Wang, and Josef van Genabith. 2007. Recovering Non-Local Dependencies for Chinese. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 257–266. Prague, Czech Republic.

Wenbin Jiang, Liang Huang, and Qun Liu. 2009. Automatic Adaptation of Annotation Standards: Chinese Word Segmentation and POS Tagging – A Case Study. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, volume 1, pages 522–530. Suntec, Singapore.

Dan Klein and Christopher D. Manning. 2003a. Accurate Unlexicalized Parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 423–430. Sapporo, Japan.

Dan Klein and Christopher D. Manning. 2003b. Fast Exact Inference with a Factored Model for Natural Language Parsing. In *Advances in Neural Information Processing Systems 15*, pages 3–10. MIT Press, Cambridge, MA.

Jonathan K. Kummerfeld, David Hall, James R. Curran, and Dan Klein. 2012. Parser Showdown at the Wall Street Corral: An Empirical Investigation of Error Types in Parser Output. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1048–1059. Jeju Island, South Korea.

[6] We introduce errors to gold tags, rather than removing errors from automatic tags, isolating the effect of a single confusion by eliminating interaction between tagging decisions.

Roger Levy and Christopher Manning. 2003. Is it harder to parse Chinese, or the Chinese Treebank? In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 439–446. Sapporo, Japan.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Slav Petrov. 2010. Products of Random Latent Variable Grammars. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 19–27. Los Angeles, California.

Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. 2006. Learning Accurate, Compact, and Interpretable Tree Annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, pages 433–440. Sydney, Australia.

Slav Petrov and Dan Klein. 2007. 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*, pages 404–411. Rochester, New York, USA.

Xian Qian and Yang Liu. 2012. Joint Chinese word segmentation, POS tagging and parsing. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 501–511. Jeju Island, Korea.

Daniel Tse and James R. Curran. 2012. The Challenges of Parsing Chinese with Combinatory Categorial Grammar. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 295–304. Montréal, Canada.

Deyi Xiong, Shuanglong Li, Qun Liu, Shouxun Lin, and Yueliang Qian. 2005. Parsing the Penn Chinese Treebank with semantic knowledge. In *Proceedings of the Second international joint conference on Natural Language Processing*, pages 70–81. Jeju Island, Korea.

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.

Yue Zhang and Stephen Clark. 2009. Transition-Based Parsing of the Chinese Treebank using a Global Discriminative Model. In *Proceedings of the 11th International Conference on Parsing Technologies (IWPT'09)*, pages 162–171. Paris, France.