Is it harder to parse Chinese, or the Chinese Treebank?

Roger Levy
Department of Linguistics
Stanford University
rog@stanford.edu

Christopher Manning
Department of Computer Science
Stanford University
manning@cs.stanford.edu

Abstract

We present a detailed investigation of the challenges posed when applying parsing models developed against English corpora to Chinese. We develop a factored-model statistical parser for the Penn Chinese Treebank, showing the implications of gross statistical differences between WSJ and Chinese Treebanks for the most general methods of parser adaptation. We then provide a detailed analysis of the major sources of statistical parse errors for this corpus, showing their causes and relative frequencies, and show that while some types of errors are due to difficult ambiguities inherent in Chinese grammar, others arise due to treebank annotation practices. We show how each type of error can be addressed with simple, targeted changes to the independence assumptions of the maximum likelihood-estimated PCFG factor of the parsing model, which raises our F1 from 80.7% to 82.6% on our development set, and achieves parse accuracy close to the best published figures for Chinese parsing.

1 Background

Even narrow-coverage context-free natural language grammars produce explosive ambiguity (Church and Patil, 1982). Today's treebank-derived broad-coverage CFGs generate even more, some of it genuine linguistic ambiguity and some of it artificial (see (Krotov et al., 1998) and (Johnson, 1998) for discussion). Corpus-based statistical parsing is a leading technique to deal with this extreme ambiguity; the vast bulk of work in this field has been done in English, using the Wall Street Journal section of the English Penn Treebank (ETB). State-of-the-art statistical parsing techniques now handle most ambiguity adequately; the best statistical parsers for the ETB are now at roughly 90% labeled bracketing accuracy (Charniak, 2000; Collins, 2000). The remaining difficult-to-resolve ambiguities are fairly well-understood for English—perhaps the best-known are flat versus embedded adjunction structures (see (Johnson, 1998) for discussion) and NP-conjunct versus flat NP coordinations—but are hardly analyzed at all for any other language. More recently, however, a wider variety of parsed corpora has become available in other languages. We take advantage of the recently released Penn Chinese Treebank (version 2.0, abbreviated here as CTB; (Xue et al., 2002)) to address these questions for Chinese, a language with less morphology and more mixed headedness than English. Chinese, as we will show, has a rather different set of salient ambiguities from the perspective of statistical parsing. This section provides background on relevant linguistic differences between Chinese and English, and on relevant tree-structural differences between the two treebanks.

1.1 Linguistic differences between English and Chinese

Chinese and English are both isolating languages: they rely primarily on relatively rigid phrase structure rather than rich morphological information to encode functional relations between elements. For purposes of statistical parsing, three salient differences distinguish the two languages. First, Chinese makes less use of function words and morphology than English: determinerless nouns are more widespread, plural marking is restricted and rare, and verbs appear in a unique form with few supporting function words. Second, whereas English is largely left-headed and right-branching, Chinese is more mixed: most categories are right-headed, but verbal and prepositional complements follow their heads (Figure 2). Significantly, this means that attachment ambiguity among a verb's complements, a major source of parsing ambiguity in English, is rare in Chinese. The third major difference is subject *pro*-drop—the null realization of uncontrolled pronominal subjects—which is widespread in Chinese, but rare in English. This creates ambiguities between parses of subjectless structures as IP (equivalent to ETB's S) or as VP, and between interpretations of preverbal NPs as NP adjuncts or as subjects.

1.2 Tree-Structural Differences between English and Chinese Treebanks

The CTB consists of 325 newswire articles; 291 are on economic topics, 34 on politics and culture. Past work on CTB parsing (Bikel and Chiang, 2000; Chiang and Bikel, 2002) has used articles 1–270 for training, 301–325 for development, and 271–300 for testing. We found, however, that this development set was uncharacteristic of the corpus as a whole and not ideal for development. As an extreme example, the word 分 appears in it 28 times as a measure word, meaning 'point', twice as an adverb, and once as a verb; in the rest of the corpus it appears eight times as a verb, once as an adverb, and never as a measure word. There turns out to be a high concentration of articles on non-economic topics in 301–325 (the problem with 分 arising from sports articles). Therefore, for this paper we set aside articles 1–25 for development and used 26–270 as training during development. During development, we found the difference in parse accuracy for 1–25 and 301–325 to range around a remarkable 10%.

Whereas ETB annotation strongly reflects late 1970s mainstream transformational grammar, CTB annotation draws primarily on Government-Binding (GB) theory from the 1980s. GB differs from the former in two major respects: first, it rigidly requires phrasal projection of all lexical categories; second, it sharply distinguishes between levels of adjunction and complementation. Both these differences are noticeable when comparing treebanks. The first difference, projection of phrasal categories, is particularly prominent within NPs: CTB adjective-noun modification, for example, is always at the level of ADJP and NP, whereas in English it can be a direct rewrite of NP to JJ and NN tags (Figure 1). The second difference, distinction of adjunction and complementation levels, has been made only for VP (Figure 2),

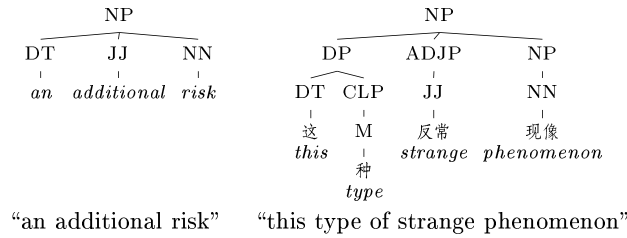

Figure 1: Noun modification in English and Chinese Treebanks

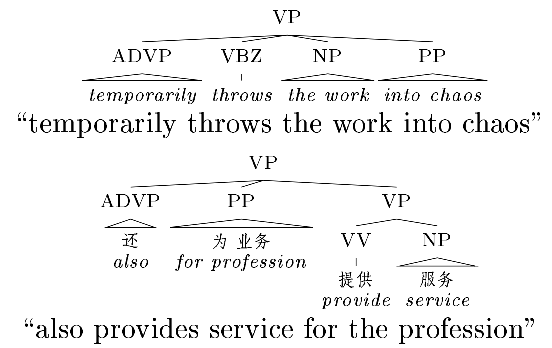

Figure 2: VP adjunction and complementation

consistent with the headedness issues described in Section 1.1.

The rigid requirement of phrasal category projection is also manifest elsewhere: all Chinese prenominal relative-clause equivalents have a level of CP annotation, equivalent to ETB's SBAR, containing a null WH-NP, even though Chinese has no relative pronouns (the overt prenominal modification marker, 的, introduces another level of CP annotation when present, as seen in Figure 3, but in this case the unary CP is compressed under standard pre-parsing tree transformations).[1] In the corresponding case for the ETB, reduced relative clauses are annotated as VPs.

These annotation practices have a strong effect on the gross statistics of the CTB after standard tree normalizations for parsing. The CTB has far fewer rule types than an ETB of equivalent size, and has a considerably lower branching factor. In particular a far higher proportion of

[1] Standard tree normalizations are: the removal of empty nodes and nodes dominating no non-empty terminals, and the subsequent removal of A over A unaries.

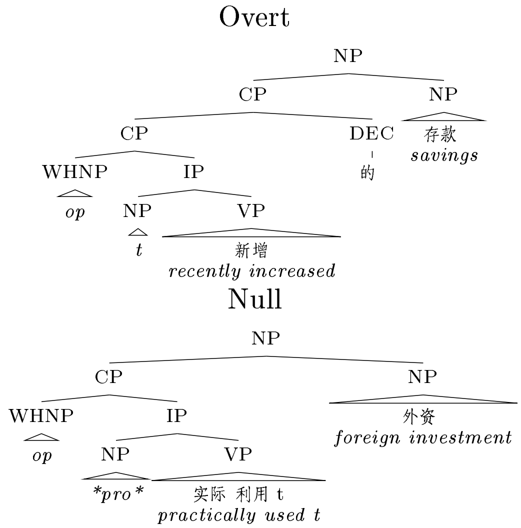

Figure 3: Prenominal modification with overt and null markers

	Rules	UnRu	BF	UnTok
WSJ-full	17023	252	2.17	24%
WSJ-small	3922	127	2.16	24%
Chinese	1797	52	1.87	41%

Table 1: Gross statistical differences between ETB and CTB. Rules and UnRu are the number of rule types and unary rule types respectively; BF is average Branching Factor and UnTok is percentage unary of local tree tokens.

	WSJ-small			Chinese		
	UA	PA	GPA	UA	PA	GPA
LP	79.4	83.3	87.7	76.4	78.7	81.8
LR	74.5	81.7	86.4	68.8	72.0	75.6
F1	76.8	82.5	87.1	72.4	75.2	78.3

Table 2: Baseline performance on training data. UA: unannotated labels; PA: parent-annotated; GPA: parent- and grandparent-annotated. LP is labeled precision; LR is labeled recall; F1 is the harmonic mean of LP and LR.

CTB rewrite rules are unary (Table 1).[2] This is consonant with the behavior of simple PCFGs on *training* data, as shown in Table 2. Parent and grandparent annotation (Johnson, 1998) has a much stronger effect on training-data parsing for ETB than for CTB. We believe that the greater precision/recall split seen here for CTB is also due to its lower branching factor.

2 Parsing model

We use the factored parsing model of (Klein and Manning, 2002). Parsing in this model involves combining two independent parses: one of a non-lexicalized, maximum likelihood-estimated (MLE) PCFG model and another of a constituent-free dependency parse. In addition to simplifying the parameterization of the parsing model and maintaining exactness, this model offers the prospect of increased flexibility in tuning the individual parse models. In particular, linguistic generalizations corresponding to category refinements are easily implemented via category-splitting in the PCFG model, without concern for affecting the dependency model.

In adapting this parsing model to Chinese, we have retained unchanged the dependency model developed for English; the model backs off to tags, and backoff parameters remain the same.[3] In all cases, test input to the parser was segmented but untagged.[4] Our focus in parser development has been to refine the PCFG model via stepwise refinements informed by major observed ambiguity classes. We illustrate that each of these refinements can effectively be viewed as an amendment to the independence assumptions made by a simple PCFG.

3 PCFG development

The simplest systematic augmentations to basic PCFG models are the inclusion of various types of contextual information in the structure of individual node labels. In principle, any contextual information could be used, but in practice two types are most heavily relied on: (i) information highly local to the enhanced node; and (ii) a unique preterminal/terminal pair identified as the *head* of the node. These practices have correlates in contemporary linguistic theory as principles of *locality* and *lex-*

[2] WSJ-small is a randomly selected tenth of the full English Wall Street Journal corpus.

[3] An algorithm to determine the *head* daughter of every non-terminal node is necessary for the dependency model and for grammar markovization (Collins, 1999), and since the CTB and ETB have different grammars, we did write a simple headfinder for the CTB grammar.

[4] For unknown words we estimated P(word|tag) based on the first character of the word.

icalism (Sag and Wasow, 1999). For headship, the choices of node enhancement strategy are fairly limited; for enrichment by local context, there are far more choices. Of the simplest local-context enrichment strategies, the one that has proven effective on a systematic basis involves *parent annotation*; (Johnson, 1998) showed that when uniformly applied, it considerably improved WSJ Treebank parsing. Uniform enhancement by other local context, such as sisters, daughters, or cousins, quickly leads to unacceptable sparseness under MLE.

To begin development, we tested the interaction of complete parent and/or grandparent annotation with PCFG markovization (see (Collins, 1999; Charniak, 2000) for discussion). The indications for the utility of parent annotation in CTB parsing are mixed. The CTB is smaller and thus more susceptible to grammar fragmentation, but it is also less flat (see Table 1). We found that first-order markovization was superior to zero-order, second-order, and unmarkovized PCFGs for all levels of ancestor annotation, and that within first-order markovization parent annotation was slightly superior to no annotation, with grandparent annotation decidedly worse.

4 Error analysis for parser development

Keeping in mind that less fragmented grammars are more robust to further category-splitting, we systematically investigated the major sources of error for the factored model with an unannotated first-order markov PCFG grammar whose only enrichment of CTB annotation was a refinement of punctuation tags along ETB lines, which achieved an F1 of 80.7%. To assess the major sources of parsing difficulty for Chinese, we tabulated frequencies of major types of parsing errors in a 100-sentence subset of our development set. Table 3 gives a breakdown of the major error types found; Figure 4 gives examples of unfamiliar major error types.[5] In the next section we describe each major error type, analyze its causes, and suggest simple PCFG enhancements that can be used to address it.

Type		Count
Flat as multilevel	VP	6
	IP	1
NP-NP Mod	+	13
	−	26
Prenom Mod	+	5
	−	5
Coord-attach	HV	10
	LV	16
	HN	7
	LN	0
Adjunct attach	IP/VP	7
	Other	3
mistagging[6]	N/V	17
	V/N	5
	Other	14

Table 3: Frequency of parse error types.

4.1 Analysis by error type and PCFG-enrichment fixes

Multilevel VP adjunction errors (Figure 5) are common in models without parent annotation, although even with parent annotation the presence of VP coordination would give multilevel VP adjunction nonzero probability. We address this error by taking advantage of the CTB's principled VP annotation practices, marking adjunction, complementation, and coordination VP levels, which builds the flat adjunction constraint back into the structure of the head daughter.

NP-NP modification, depicted in NNM+ in Figure 4, was the most common error seen; the greater prevalence of false positives is likely a result of the overall PCFG parsing preference for flatter structures. This type of parse ambiguity is grounded in the semantic ambiguity of compound noun interpretation. This semantic ambiguity exists in English as well, as in the

[5] Key to Table 3 and Figure 4:

NNM{+/−}	NP-NP modification false positive/negative
PNM{+/−}	(non-NP) prenominal mod. false positive/negative
CRD{H,L}{V,N}	incorrect {high/low} coordination attachment of righthand {verbal/nominal} material
Adj X/Y	incorrect adjunction into X; correct site was Y
mistag X/Y	category Y mistagged as X

[6] Note that only mistaggings leading to constituent-level parse errors were tallied.

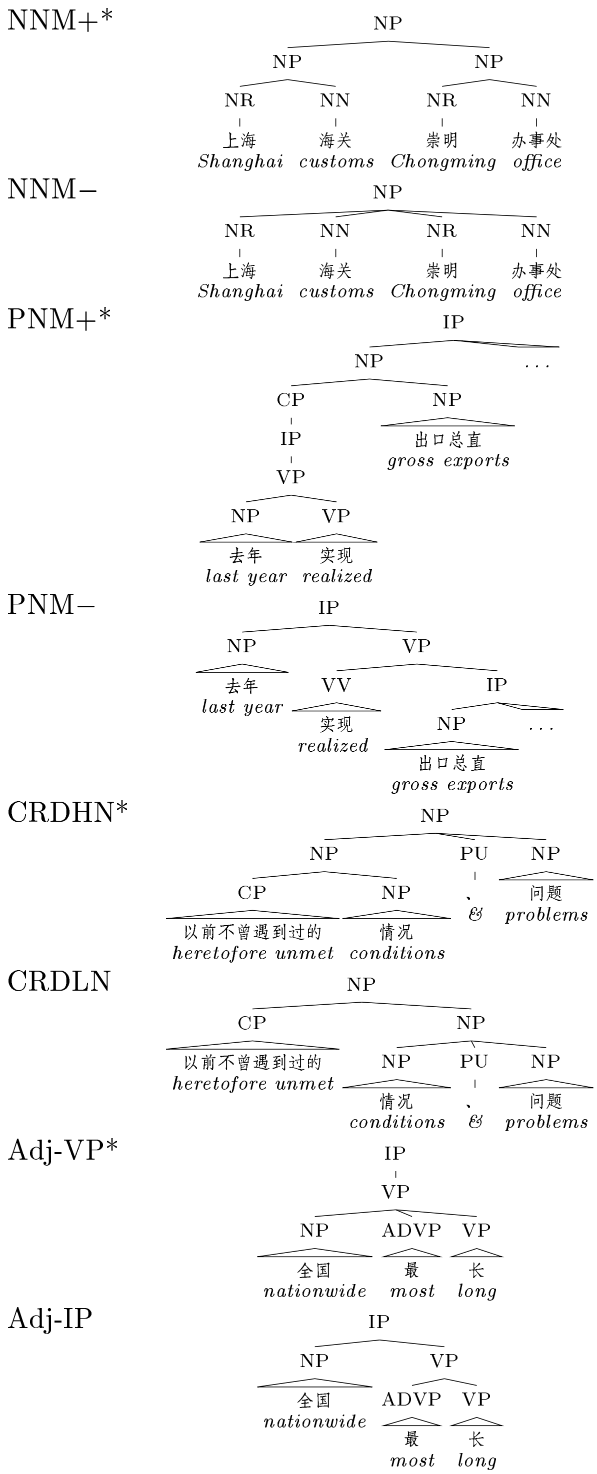

Figure 4: Major parse ambiguities. Starred examples are correct in corpus; alternates are parse errors.

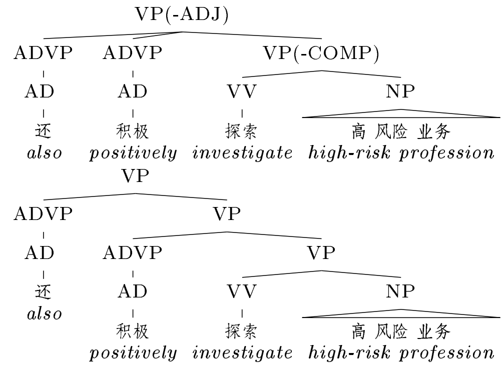

Figure 5: Flat (corpus) versus multilevel (incorrect-parse) adjunction. Parenthesized material is category-modification.

ETB string *commodity speculator Richard Dennis*, but these structures are typically bracketed flat in the ETB, underspecifying the semantic relations relative to the CTB. In CTB parsing, this type of ambiguity is difficult to resolve; different compound NP parses differ in dependency structure, so the dependency model resolves errors when word frequencies are large enough to be reliable, but this is often not possible. We found that the internal distributions of (i) NP modifiers of NPs and (ii) left-modified NPs both differ from the internal distribution of NPs in general; we take advantage of this in the PCFG model by marking both types (i) and (ii), which reduces the bias against NP-NP modification in compound NPs.

Prenominal modification errors, illustrated in PNM of Figure 4, are rather infrequent, despite the natural parallel with PP attachment ambiguity in English. Due to the highly articulated structure of prenominal modifiers, it seems difficult to address this problem directly; one measure we found somewhat successful is to mark IP daughters of prenominal modification.

Coordination scope errors occured in two major varieties: those where the misattached right conjunct is verbal (a VP or IP), and those where it is nominal—the latter case is illustrated in CRDHN and CRDLN in Figure 4.[7] The equiv-

[7] Chinese verbal coordination is generally marked with commas, whereas nominal coordination is marked with conjoiners or the mostly noun-conjoining punctuation mark "、".

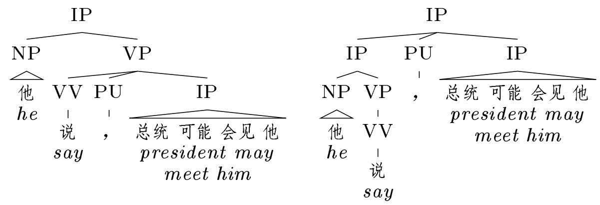

Figure 6: Ambiguity between communication verb subcategorization frame (left; corpus) and high coordination attachment (right; incorrect parse).

ocal majority of low over high verbal attachment errors contrasts qualitatively with ETB parsing, where low attachment is more common and parsers tend to err toward high attachment. There are two major sources of ambiguous attachment sites: (i) any VP can be parsed as an IP plus a unary IP→VP, so due to *pro*-drop any VP coordination is ambiguous with a higher IP coordination; (ii) VPs are multilevel, giving rise to ambiguities of scope over adjuncts. It seems that (i) is a difficult problem; in some cases, certain "discourse-level" adverbs such as 然而/*however* and 尤其是/*especially* prefer IP modification and are thus strong indicators of high attachment. To capture this we mark those adverbs possessing an IP grandparent. We address (ii) to some extent by marking VPs as adjunction or complementation structures, as shown before in Figure 5; in training data, only like-type VPs are coordinated.

With nominal coordination scope errors, the situation is different: we found no false low attachments. False high scopings can be reduced by marking NP conjuncts. (Charniak, 2000) claims that a similar strategy proved effective for WSJ parsing.

A related error arises from the introduction of IPs by communication verbs and commas, as in Figure 6. Only a few verbs in our training set take IPs this way, so we address this ambiguity with subcategorization annotation (Collins, 1999), marking VVs possessing IP sisters.

Most adjunction errors, such as into IP rather than VP, are in principle semantically impotent, since in both cases they are associated with the same verbal head. In practice, however, many adjuncts are NPs, and ambiguous adjunctions into IP are superficially indistinguishable from subjects due to *pro*-drop. NP adjuncts into VP are not ambiguous in this way, and from inspection the annotation practice of the Chinese Treebank appears to be to put NP adjuncts into VP unless they are followed by an overt subject, or otherwise distinguished as IP-level (e.g., have scope over a clear IP coordination). This could be dealt with in PCFG annotation in two ways. One is to retain subject (and/or non-subject) functional marking from the CTB. The other is to mark VP adjuncts. In practice, we found that the former hurt performance whereas the latter helped somewhat.

4.2 Tagging mistakes relevant to parsing

The strongest mistagging tendency was to tag verbs (VV) as common nouns (NN). Upon manual examination, the asymmetry of N-as-V and V-as-N mistagging frequency seems in line with the global prior over POS tags; the overall N:V ratio in the corpus is 2.5:1. We briefly explain here why N/V ambiguity is a hard problem in Chinese. All natural languages possess derivational means by which roots can switch between nominal and verbal categories. When there is overt morphological marking for these processes, as is always the case in Russian or German (and also as with Chinese and English suffixes such as -化/-*ify*) there is no ambiguity. But the sparse morphology of English and Chinese means that frequently there *is* noun/verb ambiguity at the word level. For English, most cases of this ambiguity can be resolved by the linguist on the basis of *paradigmatic* substitution in static context: whether an instance of the word *raise*, for example, is a noun or a verb can be quickly determined by checking whether *raised* can be substituted in the same context. Chinese, on the other hand, has no morphological paradigms, so any test to determine the part of speech must be made with *syntagmatic* substitution: whether the word can take an adverbial modifier or a prenominal modifier, for example.

In both languages, there are borderline cases, but they are handled differently by the respective treebanks. In English, gerunds (VBG) have both nominal and verbal properties. In the English Treebank, they have a single POS tag, but their distribution overlaps with both nouns and verbs, so that for example they can head both NPs and VPs. In Chinese, on the other hand, the tag assigned to N/V-ambiguous words is determined by the external context, specifically their maximal phrase: with the exception of domination by FRAG, all nouns are immediately dominated by NP, and all verbs by VP. To test the impact of Chinese Treebank N/V tagging practices, we tried training the parser with NN and VV training tags merged. This yielded a 5.4% drop in F1 for the vanilla PCFG, and a much smaller drop of 1.7% for the refined model, suggesting at first glance that context plus correct independence assumptions can compensate for most of the distributional information gained from N/V tag priors. But we also tried the same experiment with the ETB using the small training set and a vanilla PCFG, and found, remarkably, practically no effect: precision *increased* by 0.06%, and recall decreased by 0.21%. Although this result calls for further investigation, we tentatively conclude that in English, for most N/V-ambiguous tokens, morphology and POS prior contribute essentially nothing that cannot be adduced from the token's surrounding function-word context; whereas in Chinese, it seems that the lack of function words puts a much greater burden on prior knowledge of an N/V-ambiguous word's distributional behavior.

4.3 Further enhancements

When we included all the PCFG enhancements listed in this section, we achieved an increase of 1.4% in F1; interestingly, precision actually decreased by 0.4% whereas recall increased by 3.2%. Our PCFG enhancements were most effective at reducing NP-NP modification error, incorrect recursive bracketings, and IP/VP attachment errors. They were less effective at improving coordination attachment resolution and prenominal modification.

Although they were not directly identified as solutions to common types of errors, we identified several more PCFG refinements that reflect linguistically motivated generalizations and improve parsing performance. Generalizing from the specific error classes analyzed in the previous section, a key problem in Chinese parsing seems to be separating nonequivalent classes of IP. Two major classes of ambiguity are involved in IP membership: (i) the presence or absence in IP of subjects and adjuncts; and (ii) coordinate attachment of verbal material. We found that these ambiguities were most effectively dealt with by marking root IPs as well as those in certain sister contexts. Again in this case, CTB annotation practices introduce category conflations that, when reified in a PCFG, lead to false independence assumptions. For example, the BA marker is descriptively used in Chinese to preverbalize objects. Its syntax, however, is controversial (Bender, 2001). In the CTB, BA heads a VP and always has a unique sister IP; but that IP essentially always rewrites as NP VP. With these refinements of IP we achieved another 0.4% in error reduction, for a final development set figure of 82.1% LP, 83.1% LR, and 82.6% F1. We then ran the same model on the test set used in previous work (Bikel and Chiang, 2000; Chiang and Bikel, 2002). Results are shown in Table 4.

	$\leq$ 40 words		
	LP	LR	F1
Bikel & Chiang 2000	77.2	76.2	76.7
Present work	78.4	79.2	78.8
Chiang & Bikel 2002	81.1	78.8	79.9

Table 4: Test set parse performance

5 Results and Discussion

The trends we obtained are different enough from previous work to merit discussion. As shown in Table 4, previous work on CTB parsing consistently achieved higher results on precision than on recall. This is consonant with our initial experiments in CTB PCFG parsing: on the development set, a vanilla PCFG showed a 7% precision/recall split in favor of precision (the split for our small WSJ training set is 4.2% in the same direction). We suspect that this is due to the low branching factor in the CTB, which increases the potential reward from the parser's perspective for picking flatter structures. Simply splitting punctuation along the lines of English, combined with PCFG markovization and the introduction of a dependency model factor, reduced the LP/LR split to 1.1%. From there to our final model, nearly all improvement was in recall: precision improved by 0.3% to the final figure, whereas recall jumped by 3.4%. We interpret these results as indicating that we have unlocked a heretofore undiscovered space of independence-assumption refinements for CTB parsing, suggesting that there is still considerable room for improvement in CTB parsing even with a small (90,000-word) training set; a parser-combining model such as that proposed in (Henderson and Brill, 1999), for example, might be effective here.

This is an encouraging result for the use of detailed error analysis followed by focused tree-structure enhancements to improved parser performance. However, we found two limitations to our methodology. First, some important and addressable error types are relatively rare in Treebank data. The 100-sentence chunk of development data we chose to analyze simply did not contain any instances of BA, discussed above in Section 4.3, but errors involving BA occurred three times elsewhere in the development set. Second, some common error types are not the result of simple and easily fixable shortcomings in independence assumptions. In particular, we found that coordination scoping ambiguity and N/V tag ambiguity are major sources of relatively catastrophic error for our parser. Interestingly, coordination scope ambiguity is recognized as perhaps the most recalcitrant problem in ETB parsing, while many cases of N/V ambiguity are particularly difficult points of linguistic analysis for Chinese, as discussed in Section 4.2.

For the future, we believe that there is still room for considerable improvement in CTB parsing under our model. In addition to further PCFG refinements, tuning the dependency model may lead to improved performance. We found that head-dependent distances in the CTB are larger than in the ETB, consistent with the greater degree of center-embedding resulting from the mixed headedness of Chinese, and suggesting that a dependency model developed for English may not be optimal for Chinese. Since NP is right-headed while VP and IP are left-headed, an improved dependency model may be the best place to address at least one of the key problems we have identified for CTB parsing.

6 Acknowledgements

We are grateful to Dan Klein for valuable input, and for the parser implementation used here. This paper is based on research supported by the Advanced Research and Development Activity (ARDA)'s Advanced Question Answering for Intelligence (AQUAINT) Program.

References

Emily Bender. 2001. The syntax of Mandarin *ba*: Reconsidering the verbal analysis. *Journal of East Asian Linguistics*, 9(2):105–145.

Daniel 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.

Eugene Charniak. 2000. A maximum-entropy-inspired parser. In *Proceedings of NAACL*.

David Chiang and Daniel Bikel. 2002. Recovering latent information in treebanks. In *Proceedings of COLING-2002*, pages 183–189.

Ken Church and Ramish Patil. 1982. Coping with syntactic ambiguity or how to put the block in the box on the table. *American Journal of Computational Linguistics*, 8.

Michael Collins. 1999. *Head-Driven Statistical Models for Natural Language Parsing*. Ph.D. thesis, U. Penn.

Michael Collins. 2000. Discriminative reranking for natural language parsing. In *Proceedings of ICML*, pages 175–182. Morgan Kaufmann, San Francisco, CA.

John C. Henderson and Eric Brill. 1999. Exploiting diversity in natural language processing: Combining parsers. In *Proceedings of EM-NLP*.

Mark Johnson. 1998. PCFG models of linguistic tree representations. *Computational Linguistics*, 24(4):613–632.

Dan Klein and Christopher D. Manning. 2002. Fast exact inference with a factored model for natural language parsing. In *Proceedings of NIPS*.

Alexander Krotov, Mark Hepple, Robert Gaizauskas, and Yorick Wilks. 1998. Compacting the Penn Treebank grammar. In *Proceedings of ACL-COLING*, pages 699–703.

Ivan A. Sag and Thomas Wasow. 1999. *Syntactic Theory: A Formal Introduction*. CUP.

Nianwen Xue, Fu-Dong Chiou, and Martha Palmer. 2002. Building a large-scale annotated Chinese corpus. In *Proceedings of COLING*.