Parsing TCT with Split Conjunction Categories

Li Dongchen
Key Laboratory of Machine Perception and Intelligence,
Speech and Hearing Research Center
Peking University, China
lidc@cis.pku.edu.cn

Wu Xihong
Key Laboratory of Machine Perception and Intelligence,
Speech and Hearing Research Center
Peking University, China
wxh@cis.pku.edu.cn

Abstract

We demonstrate that an unlexicalized PCFG with refined conjunction categories can parse much more accurately than previously shown, by making use of simple, linguistically motivated state splits, which break down false independence assumptions latent in a vanilla treebank grammar and reflect the Chinese idiosyncratic grammatical property. Indeed, its performance is the best result in the 3nd Chinese Parsing Evaluation of single model. This result has showed that refine the function words to represent Chinese subcat frame is a good method. An unlexicalized PCFG is much more compact, easier to replicate, and easier to interpret than more complex lexical models, and the parsing algorithms are simpler, more widely understood, of lower asymptotic complexity, and easier to optimize.

1 Introduction

In recent years, most research on parsing has focused on English and parsing on English has reported good performance (Charniak 2000, Collins 1999, Petrov 2006, 2008). However, parsing accuracy on Chinese is generally significantly inferior.

According to the first and second Chinese parsing evaluations (CIPS-ParsEval-2009(Qiang Zhou, 2009) and CIPS-SIGHAN-ParsEval-2010((Qiang Zhou, 2010)), the evaluation results in the Chinese clause and sentence levels show that the complex sentence parsing is still a big challenge for the Chinese language.

Other work has also investigated aspects of automatic grammar refinement, for example, Chiang and Bikel (2002) learn annotations such as head rules in a constrained declarative language for tree-adjoining grammars.

Probabilistic context-free grammars (PCFGs) underlie most high-performance parsers in one way or another (Collins, 1999; Charniak, 2000; Charniak and Johnson, 2005). However, as demonstrated in Charniak (1996) and Klein and Manning (2003), a PCFG which simply takes the empirical rules and probabilities off of a treebank does not perform well.

In this paper, we investigate the learning of a grammar consistent with a treebank at the level of evaluation symbols (such as NP, VP, etc.)

Klein and Manning (2003) addressed this question from a linguistic perspective, starting with a Markov grammar and manually splitting symbols in response to observed linguistic trends in the data. For example, the symbol NP might be split into the subsymbol NPˆS in subject position and the subsymbol NPˆVP in object position.

Matsuzaki et al. (2005), Prescher (2005), Petrov (2006) induce splits in a fully automatic fashion.

Klein (2003) parses with a well-engineered grammar (as supplied for English). It is fast, accurate, requires much less memory, and in many real-world uses, lexical preferences are unavailable or inaccurate across domains or genres and the unlexicalized parser will perform just as well as a lexicalized parser. However, the factored parser will sometimes provide greater accuracy on English through knowledge of lexical dependencies. Moreover, it is considerably better than the PCFG parser alone for most other languages (with less rigid word order), including German, Chinese, and Arabic.

Automatically split-merge approach is 4% higher than manual unlexicalized parsing in English. However, this may not be the case in Chinese due to the idiosyncratic property and spe-

Proceedings of the Second CIPS-SIGHAN Joint Conference on Chinese Language Processing, pages 174–178,
Tianjin, China, 20-21 DEC. 2012

cialized annotation style in Chinese Penn Treebank. With carefully engineered split from linguistic perspective and automatically split approach, we achieve a relatively accuracy interpretable parser.

Incorporating language-dependent idiosyncratic property improved performance on many languages. As for Chinese parsing, there is still a long way to go.

High-performance parsers on English have employed linguistically-motivated features. (Collins 1998) and (Charniak 2000) make use of lexicalized nonterminals, which allows lexical items' idiosyncratic parsing preferences to be modeled, but the preferences between head words and modifiers are language-dependent. Furthermore, model in (Collins 1998) include distance measure, subcat frame features and wh-movement, which are all tightly interrelated to particular language. (Charniak 1997) uses a scheme of clustering the head words like that in (Pereira, Tishby 1993).

There have been some attempts to adapt parsers developed for English to Chinese.

Adapting lexicalized parsers to other languages is not a trivial task as it requires at least the specification of head rules, and has had limited success. (Bikel, 2000) has transplanted lexicalized parsing to Chinese and the results on English and Chinese are far from equal. Adapting unlexicalized parsers appears to be equally difficult: (Levy and Manning, 2003) adapt the unlexicalized parser of (Klein and Manning, 2003) to Chinese. Automatically splitting grammars like the one of Matsuzaki it al. (2005) and Petrov et al. (2006) require a Treebank not additionally hand tailored to English. (Petro, 2007) exhibited a very accurate category split-and-merge approach without any language dependent modifications. This automatically inducing latent structure generalizes well across language boundaries and results in state of the art performance for Chinese.

All above are probabilistic methods on the utility of PCFGs, but the same situation is in other grammar systems. SPATTER parser based on decision-tree learning techniques in Magerman (1995) highly involves special characters of words. 30 binary questions represent 30 different binary partitions of the word vocabulary, and these questions are defined such that it is possible to identify each word by asking all 30 questions. Bikel (2000) adapts stochastic TAG model on English (Chiang, 2000) to Chinese and report Label Precision below 75%.

2 Linguistic Character of Chinese

Chinese is language with less morphology and more mixed headedness than English. As Levy and Manning (2003) showed, Chinese has a rather different set of salient ambiguities from the perspective of statistical parsing

Although basic linguistic discipline is quite the same between English and Chinese, There are salient differences which distinguish the two languages for purposes of statistical parsing. Chinese makes less use of morphology than English; whereas English is largely left-headed and right-branching, Chinese is more mixed.

Furthermore, the best-performing lexicalized PCFGs have increasingly made use of subcategorization. Charniak (2000) shows the value his parser gains from parent annotation of nodes. Collins (1999) uses a range of linguistically motivated and carefully hand-engineered subcategorizations to break down wrong context-freedom assumptions of the naive Penn treebank covering PCFG. Subcategorization is proven to be important whereas subcategorization is tightly relevant to function word, especially in Chinese.

3 Lexicalized Approach Is Incompetent

Although morphology variation is not explicit in Chinese, some function words around verbs distinguish their head verbal word tense. A straightforward way of incorporating this distinction is substitute Part-Of-Speech tag of function word to the word itself, similar to Hindle and Rooth's demonstration from PP attachment.

However, several results have brought into question how large a role lexicalization plays in such parsers. Johnson (1998) showed that the performance of an unlexicalized PCFG over the Penn Treebank could be improved enormously simply by annotating each node by its parent category. Klein and Manning (2003) exploited the capacity of an unlexicalized PCFG and affirmed the value of linguistic analysis for feature discovery. An unlexicalized PCFG is easier to interpret reason about, and improve than the more complex lexicalized models. The grammar representation is much more compact, and has much smaller grammar constants. We take this as a reflection of the fundamental sparseness of the lexical dependency information available in the Penn Treebank. As a speech person would say, one million words of training data just isn't enough. Even for topics central to the treebank's Wall Street Journal text, such as stocks, many very plausible dependencies occur only once, for example stocks stabilized, while many others occur not at all, for example stocks skyrocketed.2(2This observation motivates various class- or similarity based approaches to combating sparseness, and this remains a promising avenue of work, but success in this area has proven somewhat elusive, and, at any rate, current lexicalized PCFGs do simply use exact word matches if available, and interpolate with syntactic category-based estimates when they are not.) This is equally true for function word.

We do not want to argue that lexical selection is not a worthwhile component of a state-of the-art parser, though perhaps its usage method should be carefully tuned.

In this paper, we describe simple, linguistically motivated annotations which do much to close the gap between Chinese and English parsing models.

4 Tag Splitting Approach is Appropriate Here

The idea that part-of-speech tags are not fine-grained enough to abstract away from specific-word behavior is a cornerstone of lexicalization. Klein (2003) claimed the English Penn tag set conflates various grammatical distinctions that are commonly made in traditional and generative grammar, and brought performance improvement by part-of-tag splitting.

Just as the case in English Treebank, The Chinese Treebank tag set is sometimes too coarse to capture syntactic structure distinction. The Chinese Penn tag set conflates various grammatical distinctions that are commonly made in traditional and generative grammar. Thus a parser could hope to refine some tag to get useful information.

Some tags are too coarse to capture traditional grammatical distinctions. For example, coordinating conjunctions and subordinating conjunctions are collapsed to the unique tag "c". Furthermore, coordinating conjunctions ("和", "与", "而", "并且", "既", "不单是", "乃至于", "不论") all get the tag "c" in Tsinghua Chinese Treebank, However, there are exclusively noun-modifying conjunctions ("及", "兼"), exclusively verb-modifying conjunctions ("并且"), predominantly noun-modifying and subordinately verb-modifying ones ("不止", "甚至"), predominantly verb-modifying and subordinately IP-modifying ones ("也"), and so on.

Many of these distinctions are captured by parent-annotation (noun-modifying conjunctions occur under NP, verb-modifying conjunctions occur under VP and IP-modifying conjunctions occurs under CP), some are captured by grandparent-annotation (verb-modifying CS occur with grandparent VP and parent ADVP, IP-modifying CS occur with grandparent CP and parent ADVP). But some are not (both subordinating conjunctions and complementizers appear under SBAR). What is more, the grammatical relation tag has something to do with particular function word tag, and its mapping is complicated. Thus we hope to get value from subcategorized tags for specific lexemes.

5 Hierarchical Category Refinement of Function Words

Function word is a mine full of linguistic discriminative treasure, whereas the way how its power should be exploited does matters. We presented a flexible approach which refines the function words in a hierarchy fashion where the hierarchy layers provide different granularity of specificity. We expect to compare the utility of different granularity in the hierarchy and select the most effective layer.

As in Zhou (2004), every Chinese sentence in Tsinghua Chinese Treebank is annotated with a complete parse tree, where each non-terminal constituent is assigned with two tags. One is the syntactic constituent tag, which describes its external functional relation with other constituents in the parse tree. The other is the grammatical relation tag, which describes the internal structural relation of its sub-components. These two tag sets consist of 16 and 27 tags respectively. They form an integrated annotation for the syntactic constituent in a parse tree through top-down and bottom-up descriptions.

In all function words, conjunction stand out to be essential helpful in predicting the syntactic structure and syntactic label. The refinement of conjunction words category is beneficial both to labeling the syntactic constituent tag and to labeling the grammatical relation tag.

The most obvious distinction among conjunctions is

First we split off conjunctions with the Distinguishment whether they are structural conjunctions or logical conjunctions. We refer structural conjunctions to the conjunctions which conjunct two nominal phrases. If a structural conjunction is deleted from a sentence, the sentence will be illegal in accordance to Chinese grammar. On the other hand, logical conjunctions refer to the conjunctions which conjunctions two verbal phrases.

In structural conjunctions, there are two major subcategories. The first one is coordination conjunctions which can be deeply divided into attachment conjunctions and selection conjunctions. Attachment conjunctions may represent correspondence, range or enlargement, while selection conjunctions represent the “or” relation, whether before the former option or the latter option.

Logical conjunctions are the ones representing logic coordination, transition, preference, progression, condition, cause and effect or purpose. Note that almost all the logical conjunctions can be divided by whether they are modifying the former clause or the latter clause. For example, the conjunctions representing cause and effect contains “because” and “so”, where “because” should be modifying the cause, and “so” should be modifying the effect. The condition conjunctions are relatively complicated and divided separately.

6 Experimental Setup

We ran experiments on TCT. The training and test data set splits are described in Table below.

Treebank	Train Dataset	Develop Dataset	Test Dataset
TCT(Qiang Zhou, 2004)	16000 sentences	800 sentences	758 sentences

Table 1. Experiment DataSet Setup

Tsinghua Chinese Treebank is a 1,000,000 words Chinese treebank covering a balanced collection of journalistic, literary, academic, and other documents.

For our model, input trees were annotated or transformed to refine the conjunction word categories. Given a set of transformed trees, we viewed the local trees as grammar rewrite rules in the standard way, and used smoothed maximum-likelihood estimates for rule probabilities.

To parse the grammar, we used an array-based Java implementation of a generalized CKY scheme and automatically split and merge approach in Petrov (2006).

7 Final Results

We took the final model and used it to parse the specified test set in the 3nd Chinese Parsing Evaluation which contains 1000 sentences, and achieved the best precision, recall and F-measure. Because out model employed no lexical information, it is time and space efficient.

Table 1 Final results

SC_F1	ULC_P	ULC_R	ULC_F1
92.50%	87.44%	87.43%	87.44%

Table 2. Experiment Results of SC and ULC

NoCross_P	LC_P	LC_R	LC_F1
87.44%	78.01%	78.00%	78.01%

Table 2. Experiment Results of SC and ULC

Tot4_LC_P	Tot4_LC_R	Tot4_LC_F1
76.81%	76.66%	76.74%

Table 2. Experiment Results of SC and ULC

Where LR = label recall, LP = label precision, F1 = F-measure, EX = exact match, AC = average crossing, NC = no crossing, 2C = 2 or less crossing.

8 Conclusion

The advantages of unlexicalized grammars with refined function word categories are clear enough – easy to devise, easy to estimate, easy to parse with, and time- and space-efficient.
Here, we have shown that, surprisingly, simply refining the conjunction categories in a compact unlexicalized PCFG can parse accurately.

Acknowledgements
This research is supported in part by the National Basic Research Program of China (No.2013CB329304) and the Key Program of National Social Science Foundation of China (No. 12&ZD119).

References

Zhou Qiang. Annotation Scheme for Chinese Treebank. Journal of Chinese Information Processing. (2004)

Petrov, Klein, 2006. Learning Accurate, Compact, and Interpretable Tree Annotation , in ACL’ 06.

N. Xue, F.-D. Chiou, and M. Palmer. Building a large scale annotated Chinese corpus. In COLING ’02, 2002.

Qiang Zhou. Chinese Treebank Annotation Scheme. Journal of Chinese Information, 18(4), p1-8. (2004)

Qiang Zhou, Yuemei Li. Evaluation report of CIPS-ParsEval-2009. In Proc. of First Workshop on Chinese Syntactic Parsing Evaluation, Beijing China, Nov. 2009. pIII—XIII. (2009)

Qiang Zhou, Jingbo Zhu. Chinese Syntactic Parsing Evaluation. Proc. of CIPS-SIGHAN Joint Conference on Chinese Language Processing (CLP-2010), Beijing, August 2010, pp 286-295. (2010)

E. Charniak and M. Johnson. 2005. Coarse-to-fine n-best parsing and maxent discriminative reranking. In ACL'05, p. 173–180.

E. Charniak. 2000. A maximum–entropy–inspired parser. In NAACL '00, p. 132–139.

D. Chiang and D. Bikel. 2002. Recovering latent information in treebanks. In Computational Linguistics.

M. Collins. 1999. Head-Driven Statistical Models for Natural Language Parsing. Ph.D. thesis, U. of Pennsylvania.

M. Johnson. 1998. PCFG models of linguistic tree representations. Computational Linguistics, 24:613–632.

D. Klein and C. Manning. 2003. Accurate unlexicalized parsing. ACL '03, p. 423–430.

T. Matsuzaki, Y. Miyao, and J. Tsujii. 2005. Probabilistic CFG with latent annotations. In ACL '05, p. 75–82.

D. Prescher. 2005. Inducing head-driven PCFGs with latent heads: Refining a tree-bank grammar for parsing. In ECML'05.

S. Sekine and M. J. Collins. 1997. EVALB bracket scoring program. http://nlp.cs.nyu.edu/evalb/.

E. Charniak, S. Goldwater, and M. Johnson. 1998. Edge-based best-first chart parsing. 6th Wkshop on Very Large Corpora.

E. Charniak,M. Johnson, et al. 2006. Multi-level coarse-to-fine PCFG Parsing. In HLT-NAACL '06.

Z. Chi. 1999. Statistical properties of probabilistic context-free grammars. In Computational Linguistics.

M. Collins. 1999. Head-Driven Statistical Models for Natural Language Parsing. Ph.D. thesis, U. of Pennsylvania.

D. Gildea. 2001. Corpus variation and parser performance.EMNLP '01, pages 167–202.

R. Levy and C. Manning. 2003. Is it harder to parse Chinese, or the Chinese treebank? In ACL '03, pages 439–446.

M. Marcus, B. Santorini, and M. Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. In Computational Linguistics.

W. Skut, B.Krenn, T. Brants, and H. Uszkoreit. 1997. An annotation scheme for free word order languages. In Conference on Applied Natural Language Processing.

H. Sun and D. Jurafsky. 2004. Shallow semantic parsing of Chinese. In HLT-NAACL '04, pages 249–256.

K. Vijay-Shanker and A. Joshi. 1985. Some computational properties of Tree Adjoining Grammars. In ACL '85.

N. Xue, F.-D. Chiou, and M. Palmer. 2002. Building a large scale annotated Chinese corpus. In COLING '02.