Adapting Multilingual Parsing Models to Sinica Treebank

Liangye He, Derek F. Wong, Lidia S. Chao

Natural Language Processing & Portuguese-Chinese Machine Translation Lab
University of Macau
Macau SAR, China
wutianshui0515@gmail.com
{derekfw,lidiasc}@umac.mo

Abstract

This paper presents our work for participation in the 2012 CIPS-SIGHAN shared task of Traditional Chinese Parsing. We have adopted two multilingual parsing models – a factored model (Stanford Parser) and an unlexicalized model (Berkeley Parser) for parsing the Sinica Treebank. This paper also proposes a new Chinese unknown word model and integrates it into the Berkeley Parser. Our experiment gives the first result of adapting existing multilingual parsing models to the Sinica Treebank and shows that the parsing accuracy can be improved by our suggested approach.

1 Introduction

Work in syntactic parsing has developed substantial advanced Probabilistic Context-Free Grammar (PCFG) models (Collins, 2003; Klein and Manning, 2002; Charniak and Johnson, 2005; Petrov et al., 2006). The syntactic structures of English sentences can be well analyzed by utilizing these models. The highest traditional PARSEVAL F1 accuracy evaluation reported on English Parsing have already reached 92.4% (Fossum and Knight, 2009), which is very acceptable.

However, parsing Chinese still a tough task. Chinese varies from English in many linguistic aspects. That makes a big difference between the Chinese syntactic trees' structures and the English ones. For example, the Chinese syntactic tree is constructed flatter than the English one (Levy and Manning, 2003).

In this paper, we present our solution for the 2012 CIPS-SIGHAN shared task of Traditional Chinese parsing. We exploit two existing powerful parsing models – the factored model (Stanford Parser) and the unlexicalized model (Berkeley Parser), which have already shown their effectiveness in English, and adapt it to our task with necessary modification. First, in order to make use of Stanford Parser, we try to build a head propagation table of Traditional Chinese for the adaptation of the specific Traditional Chinese Corpus – Sinica Treebank (Chen et al., 2000). Second, we propose a new Chinese unknown word model to estimate the word emission probability, to improve the Traditional Chinese parsing performance for the Berkeley Parser.

2 Related Work

There have been several efforts to achieve high quality parsing results for Chinese by using varied parsing models (Bikel and Chiang, 2000; Levy and Manning, 2003; Petrov and Klein, 2007). Table 1 gives their respective performance.

We can see that the Berkeley Parser (Petrov and Klein, 2007) attained the state-of-the-art performance, around 83% PARSEVAL F1 measure on Penn Chinese Treebank (CTB) (Xue, 2002).

However, different corpus has different design criteria and annotation schema. As to our best knowledge, there is still no attempt to employ the existing parsing models to adapt to this Traditional Chinese Corpus. More work should be carried out to investigate what performances the

Proceedings of the Second CIPS-SIGHAN Joint Conference on Chinese Language Processing, pages 211–215,
Tianjin, China, 20-21 DEC. 2012

Work	Experimental Treebank	F1 Performance
Bikel and Chiang (2000)	CTB	76.7%
Levy and Manning (2003)	CTB	78.8%
Petrov and Klein (2007)	CTB	80.7%

Table 1: Previous Work on Parsing Chinese

mentioned existing sophisticated can get when utilized in different corpora.

2.1 The Sinica Treebank

In the 2012 CIPS-SIGHAN shared task of Traditional Chinese Parsing, the released training and testing datasets is extracted from the Sinica Treebank v3.0. The Sinica Treebank has some Traditional Chinese specific linguistic information annotated and is based on the Head-Driven Principle; each non-preterminal is made up of a Head and its modifiers. The phrasal type and the relations with other constituents are specified by the Head. For example, the traditional tree view of sentence *嘉珍和我住在同一條巷子* is shown in Figure 1:

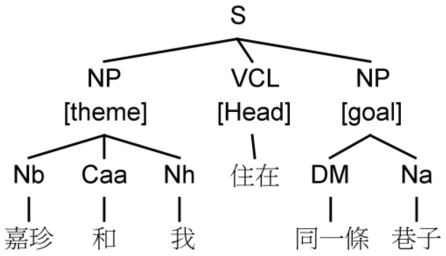

Figure 1: Part of the Sinica Treebank, each phrasal tag (in this case, *S*) is composed into Head and dependencies

3 Multilingual Parsing Models

In our experiments we will employ two multilingual statistical parsers – the Stanford Parser and the Berkeley Parser. We will describe the Stanford package and our modification in order to make this package adapt to the Sinica Treebank in Subsection 3.1. The Berkeley parser will be referred to in Section 3.2. In that Section we will also propose a new Chinese unknown word model.

3.1 The Stanford Parser

3.1.1 A Factored Model

A factored parser, which combine a high optimized unlexicalized parsing model (syntactic model) (Klein and Manning, 2003) and a dependency parser (semantic model) can be trained by the Stanford parser. The unlexicalized model produces a high optimized probabilistic context-free grammar, which adds some linguistically motivated annotation to both phrasal and Part-of-Speech tags to do disambiguation. In the lexical dependencies part, the information of direction, distance and valence between a constituent and its modifiers will be encoded into the dependency model. The probability of a tree is then calculated through the product of the probabilities that the syntactic model and the semantic model assign to that tree. Now the software package provides reinforcement for English, Chinese, Arabic, French and German.

3.1.2 Head Propagation Table for Sinica Treebank

In the newest version of Stanford parser, many languages are supported. In addition to using the default Chinese package[1], we have created the Sinica-specific extensions for Stanford parser. This package mainly contains a head propagation table, morphological features and some tuning of parser options for the Sinica Treebank.

In order to realize the rule binarization[2] for unlexicalized model and prepare the word-to-word affiliation for dependency model, the parser still needs to pick out the head child in the internal rule. Sinica Treebank indicates head information by adding some semantic label[3] to the phrasal tag, so we can build a head propagation table by traversing all the trees in the corpus.

[1] In the newest Stanford package, the default setting in Chinese Parsing is designed for CTB 5.0.

[2] See (Klein and Manning, 2003) for the explanation.

[3] We extract the head child which is tagged Head for the top phrasal tag

Parent	Direction	Priority List
S	left	*VP, VA, VA[+NEG], VA[+ASP], VA[+NEG,+ASP], VAC, VAC[+ASP], VB, VB[+ASP], VB[+DE], VB[+NEG], VC,VC[+ASP], VC[+NEG], VC[+DE], VC[+SPV], VC[+DE,+ASP], VCL,VD, VD[+NEG], VE, VE[+DE], VE[+NEG], VF, VG, VG[+DE], VG[+NEG], VH, VH[+D],VHC, VH[+ASP], VH[+NEG], VL, VK, VK[+ASP], VK[+DE], VK[+NEG], VI, VI[+ASP], VJ, VJ[+DE], VJ[+SPV], VJ[+NEG], V_11, V_12, V_2, V, S, NP, Na, Nb, Nc, Ndb, Ndc, Neqa, Neu, Ng, Nh, Nv, P,GP, DM, D, Dfa, A, Caa, Caa[P1], Caa[P2], Cab, Cbb*
VP	left	*VP, VA, VA[+NEG,+ASP], VA[+NEG], VA[+ASP], VAC, VAC[+SPV], VB, VB[+ASP], VB[+NEG], VC, VC[+NEG], VC[+DE], VC[+SPV], VCL, VCL[+NEG], VCL[+SPV], VD, VE, VE[+DE], VE[+NEG],VF, VG, VG[+NEG], VH, VH[+ASP], VH[+DE], VH[+NEG], VHC, VHC[+ASP], VHC[+SPV], VI, VJ, VJ[+DE], VJ[+NEG], VK, VK[+ASP], VK[+DE], VK[+NEG], VL, V_11, V_12, V_2, V, S, NP, Na, Nc, Ng, P, DM, D, Di, Dfa, Caa, Caa[P1], Caa[P2], Cab, Cbb,*
NP	left	*NP, N, Na, Nb, Nc, Ncd, Nd, Nda, Ndb, Ndc, Nde, Ndf, Nep, Neqa, Neqb, Neu, Nf, Nh, N・的, Nv, PP, P, GP, DE, DM, Caa, Caa[P1], Caa[P2], Cab*
GP	left	*VE, Ncd, Nes, Ng,P, GP, Caa, Caa[P1], Caa[P2]*
DM	left	*Neu, Nf, DM*

Table 2: The Head rules used for Sinica Treebank in the Stanford Parser

Table 2 gives our version of Traditional Chinese head propagation table. [4]

3.2 The Berkeley Parser

3.2.1 An Improved Unlexicalized Model

The Berkeley parser (Petrov et al. 2006; Petrov and Klein, 2007) enhanced the unlexicalized model which is adopted in the Stanford parser. In the grammar training phase, Berkeley parser use an automatic approach to realize the tree annotation which is analyzed and testified manually in Stanford's unlexicalized model; that is, iteratively rectify a raw X-bar grammar by repeatedly splitting and merging non-terminal symbols, with a reasonable smoothing. At first, the baseline X-bar grammar is obtained directly from the raw datasets by a binarization procedure. In each iteration, for splitting, the symbol could be split into subsymbols. This leads to a better parameter estimates for the probabilistic model. However, splitting will cause the overfitting problem. To solve this, the model will step into the merging and smoothing procedure. More details about the strategies of splitting, merging and smoothing, see (Petrov et al., 2006).

3.2.2 The Chinese Unknown Word Model

In parsing phase, if the unknown words belong to the categories of digit or date, the Berkeley Parser has some inbuilt ability to handle them. For words excluded these classes, the parser ignores character-level information and decide these words word categories only on the rare-word part-of-speech tag statistics. Let t denote the tag, and w denote the word. The model for estimation of the unknown word probability somehow can be written in this format:

$$P\ (w|t) \qquad (1)$$

In our work, we employ a more effective method, which is similar to but more detailed than the work of Huang et al. (2007), to compute the word emission probability to build up our

[4] We only show part of the head table which contains the main phrasal tags.

Model	PARSEVAL F1	POS Accuracy
Stanford-BA	45.20%	72.72%
Stanford-MOD	47.32%	72.92%
Berkeley-BA	49.60%	65.79%
Berkeley-MOD	50.42%	74.02%

Table 3: Experimental Results

new Chinese unknown word model. The geometric average[5] of the emission probability of the characters in the word is applied. We use c_k to denote *k*-th character in the word. Since some of the characters in w_i may not have appeared in any word tagged as t_i in that context in the training data, only characters that are mentioned in the context are included in the estimate of the geometric average then $P(c_k|t_{\mathrm{i}})$ is achieved:

$$P(w_i|t_i) = \sqrt[\Sigma\theta]{\prod_{c_k \epsilon w_{i,} P(c_k|t_{i_k})\neq 0} P(c_k|t_{i_k})^{\theta_k}} \quad (2)$$

Where:

$$n = |\{c_k \in w_i | P(c_k|t_{i_k}) \neq 0\}|$$

$$\theta_k = \exp(-dis(c_k))$$

In (2), we use θ_k to assign a weight to the emission probability of each character c_k. We will determine the head character and use an exponential function to represent the distance between the head character and other characters. In our experiment, we will use the first character and the last character as the head character respectively and try out which position in a Chinese word is most important.

4 Experiment

4.1 Experimental Setup

In our experiment, we divide the Sinica Treebank in 3 parts following the traditional supervised parsing experimental protocol: training (first 80%), development (second 10%) and test (remaining 10%). We systematically report the result with treebank transformed. Namely, we preprocess the treebank in order to turn each tree into the same format[6] as in Penn Treebank since mentioned constituency parsers only accept this format.

4.2 Evaluation Metrics

We use the standard labeled bracketed PARSEVAL metric (Black et al., 1991) for constituency evaluation, all the phrasal tags will be taken into account.[7] Besides, we also report the POS accuracy.

4.3 Experimental Results

For better description, we name the basic version of Stanford parser as *Stanford-BA* and the modified version with the Traditional Chinese head propagation table as *Stanford-MOD*. While *Berkeley-BA* and *Berkeley-MOD* represent for the basic Berkeley parser and the intensive one respectively. Table 3 gives their performance on parsing Traditional Chinese.

Coming to a comparing among these two parsers, Berkeley parser has better overall performance. The basic version of Berkeley parser, *Berkeley-BA*, beat *Stanford-BA* in 4.4%, scored 45.20% and 49.60% F1 respectively. For each model, our modification for adaptation also makes an improvement. After deploying the specific head propagation table, we got 2.12% and 0.2% improvement in constituent accuracy and POS accuracy respectively. While the *Berkeley-MOD* benefits from the new Chinese Unknown word model, the constituent F1 and POS accuracy reach to 50.42% and 74.02% respectively[8].

5 Conclusion and Future Work

In this paper, we reported our participation in the CIPS-SIGHAN-2012 Traditional Chinese Parsing Task. We employed two statistical parsing models designed in multilingual style and apply them to parse the Traditional Chinese. Each baseline results were given. We also make this

[5] As Huang et al. (2007) suggested, the geometric average is better than arithmetic average, but we do not testify it in this paper due to tight schedule.

[6] All the Semantic Role Labels are eliminated.

[7] While the official evaluation only takes S, VP, NP, GP, PP, XP, and DM into account.

[8] We use Berkeley-MOD for CIPS-SIGHAN 2012 Bakeoffs.

parser adapt to the Sinica Treebank. At first, For the Stanford Parser, we generated a head propagation table for Sinica Treebank. Besides, we also design a new Chinese unknown word model and integrate it into the Berkeley Parser. The result shows improvement over the base model.

However, after adapting those parsers to Traditional Chinese, we still find that probabilistic parsing was not efficient enough to provide accurate parsing result for Sinica Treebank compared to the work done in CTB. We still need to go deeper into the research of the corpus characteristics and the existing multilingual parsing models and make better adaptation.

Acknowledgments

This work was partially supported by the Research Committee of University of Macau under grant UL019B/09-Y3/EEE/LYP01/FST, and also supported by Science and Technology Development Fund of Macau under grant 057/2009/A2.

References

Bikel D. M. and Chiang D. 2000. Two Statistical Parsing Models Applied to the Chinese Treebank . *Second Chinese Language Processing Workshop*. 1–6.

Black E., Abney S., Flickenger S., Gdaniec C., Grishman C., Harrison P., Hindle D., Ingria R., Jelinek F., Klavans J. L., Liberman M. Y., Marcus M. P., Roukos S., Santorini B. and Strzalkowski T. 1991. A procedure for quantitatively comparing the syntactic coverage of English grammars. Proceedings of *the Workshop on Speech and Natural Language*. 306–311.

Charniak E. and Johnson M. 2005. Coarse-to-Fine n-Best Parsing and MaxEnt Discriminative Reranking. *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*. 173–180.

Collins M. 2003. Head-driven statistical models for natural language parsing. *Computational Linguistics*. 29:589–637.

Fossum V. and Knight K. 2009. Combining Constituent Parsers. Proceedings of *Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*. 253–256.

Huang C. R., Chen F. Y., Chen K.J., Gao Z. M., and Chen K. Y. 2000. Sinica Treebank: Design Criteria, Annotation Guidelines, and On-line Interface. *Second Chinese Language Processing Workshop*. 29–37.

Huang Z., Harper M., and Wang W. 2007. Mandarin Part-of-Speech Tagging and Discriminative Reranking. Proceedings of *the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*. 1093–1102.

Klein D. and Manning C. D. 2002. Fast exact inference with a factored model for natural language parsing. *Advances in Neural Information Processing Systems*. 15:3–10.

Klein D. and Manning C. D. 2003. Accurate Unlexicalized Parsing. Proceedings of *the 41st Annual Meeting of the Association for Computational Linguistics*. 423–430.

Levy R. and Manning C. D. 2003. Is it Harder to Parse Chinese, or the Chinese Treebank? Proceedings of *the 41st Annual Meeting of the Association for Computational Linguistics*. 439–446.

Petrov S., Barrett L., Thibaux R., and Klein D. 2006. Learning Accurate, Compact, and Interpretable Tree Annotation. Proceedings of *the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*. 433–440.

Petrov S and Klein D. 2007. Improved Inference for Unlexicalized Parsing. *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*. 404–411.

Xue N., Chiou F. D., and Palmer M. 2002. Building a large-scale annotated Chinese corpus. Proceedings of *the 19th International Conference on Computational linguistics-Volume 1*. 1–8.