A Conditional Random Field-based Traditional Chinese Base-Phrase Parser for SIGHAN Bake-off 2012 Evaluation

Yih-Ru Wang
National Chiao Tung University
Hsinchu, Taiwan, ROC.
yrwang@mail.nctu.edu.tw

Yuan-Fu Liao
National Taipei University of Technology, Taipei, Taiwan, ROC
yfliao@ntut.edu.tw

Abstract

This paper describes our system for the sub-task 1 of traditional Chinese Parsing of SIGHAN Bake-off 2012 evaluation. Since this research mainly focuses on speech recognition and synthesis applications, only base phrase chunking was implemented using three Conditional Random Field (CRF) modules, including word segmentation, POS tagging and base phrase chunking sub-systems. The official evaluation results show that the system achieved 0.5038 (0.7210/0.387) micro- and 0.5301 (0.7343/0.4147) macro-averaging F1 (precision/recall) rates on full sentence parsing task. However, if only the performance of base phrase chunking was considered, the F-measures may be around 0.70 and is somehow good enough for speech recognition and synthesis applications.

1 Introduction

For NLP researches, a semantic parser is used for mapping a natural-language sentence into a formal representation of its meaning. It usually first groups the elements in a sentence into words, phrases and clause and then tags each word, phrase and clause with a semantic label.

There are still many challenges in semantic parsing, but the intermediate results of the semantic parsing are already quite useful for speech recognition and text-to-speech applications. For example, word sequences information could be used to build the language model in automatic speech recognition (ASR), and the phrase and clause results can be used to further verify the recognition result. In text-to-speech system, boundary information of the words, phrases and clauses can be used to better predict the prosody of synthesis speech.

2 CRF-based Traditional Chinese Base-Phrase Chunking System

The block diagram of our system is shown in Fig. 1. There are five sub-systems including a text normalization, a word segmentation, a POS tagging, a compound word construction and a base-phrase chunking modules.
First of all, in Chinese, there are lots of canonical composition glyphs. The word construction sub-system canonical composition glyphs, or variant characters, were handled in a text-normalization sub-system. The other modules will be briefly described as follows:

### 2.1 Word Segmentation

The word segmentation sub-system is a CRF-based system. It follows the Zhan’s work (Zhan and Huang, 2006). The 6 tags, named $B1$, $B2$, $B3$, $M$, $E$ and $S$, were used to represent the activated function in CRF. The information using in feature template are

- $C_n$ : Unicode current character (Unicode plain-0 only).
- $B_n$ : radical of current character ("bushu", 部首)
- $SB_n$ : if $B_n = B_{n-1}$
- $WL_n$ : maximum length word in lexicon match to string including current character, the 87,000 lexicon from Sinica\(^1\) was used as the system internal lexicon, and a user-defined lexicon was allowed to define more words, and in most cases they will be named entities.
- $WT_n$ : tags of current characters in the maximum word length matched word in lexicon (indicate character position in word using $B1$, $B2$, $B3$, $M$, $E$, $S$).
- $D/E_n$ : whether the current character is a digit.
- $PM_n$ : whether the current character is a punctuation mark (PM).

The above features and the templates used in our system were commonly used in Chinese word segmentation task. It’s worth to mention that the radical of Chinese character was a useful feature for same OOV words. The top-n sequences of word segmentation sub-system were sent to the next sub-system.

---

\(^1\) [http://www.aclclp.org.tw/use_ced_c.php](http://www.aclclp.org.tw/use_ced_c.php)
The sub-system was trained by using the Sinica corpus, version 4.0\textsuperscript{2}. A lot of data was corrected in the database by using consistence check.

About more than 1\% of data in Sinica Corpus was corrected. The word unigram and unigram of Sinica corpus were first generated, and we find all the word-pairs were also combined into a single word in the corpus besides the words with POS “Nf” and “Neu”. There are about 10\% of the word-pairs can also be segmented into single words. Some word segmentation inconsistency were checked and corrected, like

(1) /民意代表(Na)/ and /民意(Na) 代表(Na)/ both appeared in the corpus,

(2) The word /長途(A) 電話(Na)/ are segmented in all the cases in corpus, but the word /長途電話(Na)/ was included in the Sinica lexicon.

In this case, the lexicon was modified.

(3) Most of the bound morphemes (prefixes, suffixes), named entities, compound words, idioms, abbreviations.

Some words, especially function words, were segmented into more than one segmentation and POS possibilities, like [就是(T), 就是 (SHI), 就是(Nc), 就是(D), 就是(Cbb)] and [真是(VG), 真是(D), 真(D) 是(SHI)], while these were not yet checked in our study.

The researchers have set a high standard for their significant works in developing the corpus, yet it is still impossible to ignore the words proposed by Andrew Rosenberg (2012): “The corpus is an invaluable resource in Spoken and Natural Language Processing. Consistent data sets have allowed for empirical evaluation of competing algorithms. …. However, despite dubbing these annotations as “gold-standard”. many corpora contain labeling errors and idiosyncrasies. The current view of the corpus as a static resource makes correction of errors and other modifications prohibitively difficult.” Hence, we hope to see the dynamic Chinese linguistic resources as soon as possible and the users of corpus could then contribute their error corrections.

Then, 9/10 of the corpus (about 1 million words) was used for training and 1/10 (about 120K words) was used as evaluation data. The F-measure of the word segmentation sub-system is 97.37\%. The difference of precision and recall rate was less than 0.1\%.

2.2 POS tagging

In our system, the top-N output sequences of the word segmentation were sent to the POS tagger. The possible POS types of each word should be the most effective feature for POS tagging. Since a lexicon was used in word segmentation sub-system, the possible POS’s of each lexical word was also store in the lexicon. The information using in feature template are

- \( PM_n \) : Unicode of the first character of current word when it is PM, or “X” if it is not PM.
- \( WL_n \) : word length of current word.
- \( LPOS_n \) : all possible POSs of current words if the word is in the internal and external lexicons, or “X” if it is not in the lexicons, i.e., for word “—”(one) can be “Cbb_Di_D_Neu”
- \( FC_n \) : first character of current word if the word is not in lexicon, or “X” if it is in lexicon.
- \( LC_n \) : last character if the word is not in lexicon, or “X” if it is in lexicon.

There are 47 types of POS in the system those are used in Sinica corpus version 4.0 as well.

The sub-system was also trained by the same corpus used in word segmentation sub-system. The accuracy of the POS tagging sub-system is 94.16\%. The recognition of 47 POS types was reasonable except noun type “Nv” due to its ambiguity.

In the basic system, the POS tagger will process the top-N sequences out from word segmentation. The log-likelihood of word segmentation and POS tagging were added and found the best output sequence.

The F-measure of word segmentation and recognition rate of POS tagger were usually used as the performance measures of a parsing system. In our study, we also check the effectiveness of our word segmentation and POS tagger sub-system in the speech recognition application. The above two sub-system was used in building the language model in ASR system. Sinica corpus, CIRB030\textsuperscript{3} and Taiwan Panorama Magazine\textsuperscript{4}, contain 380 million words totally, were parsed to build the trigram language model for speech recognizer. 60K words were used in the recognition

\textsuperscript{2} http://www.aclclp.org.tw/use_asbc_c.php
\textsuperscript{3} http://www.aclclp.org.tw/use_cir_c.php
\textsuperscript{4} http://www.aclclp.org.tw/use_gh_c.php
lexicon. The performance of the Mandarin speech recognizer was evaluated in the TCC-300 speech database. The out-of-vocabulary rate is 3.1% for 15479 words test data. Word error rate (WER) of the recognizer reduces to 13.4%. About 40% word error rate reduction was achieved comparing to the CRF-based word segmentation and POS tagger system we built from Bakeoff-2005 training database.

2.3 Compound word construction

The first compound word construction rule which was implemented in our system is the Determinative-Measure compound word. In Sinica Treebank, except the 47 POS types, one more POS tagger DM, Determinative-Measure compounds, was used. The following DM construction rules, which check the POS of word sequence, were used to construct the DM compound in the word sequence, recursively.

- Neu + Nf + Neu + !(Nf) ⇒ DM+ !(Nf)
- Neu + Neuq ⇒ Neu
- (Neu, Nes, Nep, Neqa, Neqb)+Nf ⇒ DM
- DM+(Nf, Neqb) ⇒ DM
- (Nep, Nes)+DM ⇒ DM
- Neu+(“大”/(da/, big), “小”/(xian/,small)) +Nf ⇒ DM

In “Chinese information processing issued by the Central Standards Bureau,” there are lots of rules for constructing traditional Chinese compound words. In our system, some of them were implemented. Those rules are listed in follows:

- 半 A 半 B,
- 一 A 一 B,
- 如 A 如 B.

ADAB, D is a character with POS Di,

AABB, AB is a lexical word with POS Vx, where A, B are single character.

2.4 Base-phrase chunking

In the base-phrase chunking sub-system, the POS sequence was the most useful feature in base-phrase chunking. Besides the POS and simplified POS, some character information of the word were also used.

- POS_n : POS of current word.
- SP_n : simplified POS of current word.

The types of POS was simplified from 47 to 13 categories, {A, C, D, DE, FW, I, N, P, PM, SHI, T, V}

- LW_n : word length of current word.
- SW1_n : set to 1 if word W_n is same as word W_{n-1}, 0 if otherwise.
- SW2_n : set to 1 if word W_n is same as word W_{n-2}, 0 if otherwise.
- FC_n : first character of current word.
- EC_n : last character of current word.

The templates used in the system were shown in Figure 2.

| POS n-gram          | POS_n, POS_{n-1}, POS_{n+1}, POS_{n+2}, (POS_{n-2} POS_{n-1} POS_{n}), (POS_{n} POS_{n+1} POS_{n+2}), (POS_{n-1} POS_{n+1} POS_{n+2}), (POS_{n} POS_{n+1} POS_{n+2}) |
|---------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|

| Simplified POS n-gram | SP_n, SP_{n-1}, SP_{n+1}, SP_{n+2}, (SP_n SP_{n+1} POS), (SP_{n-1} SP_{n+1} POS), (SP_{n-2} SP_{n+1} POS), (SP_{n+1} POS) |

| POS and word-length  | (POS_n, LC_n), (POS_{n-1}, LC_{n-1}), (POS_{n+1}, LC_{n+1}) |
|----------------------|-------------------------------------------------------------------|

| POS and first/last character  | (POS_n, FC_n), (POS_{n-1}, FC_{n-1}), (POS_{n+1}, FC_{n+1}), (POS_{n+1} LC_{n+1}) |

| Repeated word              | (LW_n, SW1_n), (LW_n, SW2_n) |

In the knowledge bases for semantic parsing, the lexical senses, like information in Wordnet, ..., etc, are important features for parsing (Mel’čuk, 1996; Shi and Mihalcea, 2005), however in our current system the lexical sense information is not considered yet. The activated
function of the BP chunking was set to 7 tags, \{ADVP, GP, NP, PP, S, VP, XDE(X • DE)\}.

Then, 9/10 of the Bakeoff-2012 Task-4 training corpus was used for training the base-phrase chunking module and 1/10 for was used as self-evaluation data. The result of the base-phrase chunking was shown in Table 1.

The Chinese parsing system as shown in Figure 1 was implemented by using the CRF++ package\(^9\). The base phrase tags, ADVP and XDE, were combined into XP as the Bakeoff-2012 result.

Since our system is also used as a front end of text-to-speech (TTS) system, usually the input is taken from books and released news. Fig. 3 shows partial parsing result. The context is from “Angels & Demons”, Dan Brown, 2000. The performance is acceptable for TTS application.

### 3 Evaluation Results on Traditional Chinese Parsing Sub-task 1

The system use for Bakeoff-2012 Traditional Chinese Parsing sub-task 1 is modified from the basic parser described in last section.

In the Bakeoff-2012 Traditional Chinese Parsing sub-task 1, the input sentences were segmented with gold standard word sequences. Thus, the basic system was modified to generate the n-best word sequences in POS tagging and compound word construction stages for this evaluation. The n-best word sequences satisfied with the defined principles, minimum edit-distance and maximum log-likelihood, in the test data set were returned as pre-processing word sequences. Finally, the n-best word sequences with their corresponding POS tags can be sent into base-phrase chunking module for getting the base-phrase chunking results.

The official evaluation report of our system for Traditional Chinese Parsing sub-Task 1 is shown in Fig. 4.


d| Task : Subtask1  
| --- | --- | --- | --- |
| Track : Closed  
| System : Single  
| Run : Run1  

[
| Part 1 | Overall Performance  
| --- | --- | --- | --- | --- | --- | --- |
| Micro-averaging Precision | 0.7215  
| Micro-averaging Recall | 0.387  
| Micro-averaging F1 | 0.5038  
| Macro-averaging Precision | 0.7343  
| Macro-averaging Recall | 0.4147  
| Macro-averaging F1 | 0.5301  

[
| Part 2 | Summary  
| --- | --- | --- | --- | --- | --- |
| Type | (#Truth) | (#Parser) | (%Ratio)  
| S | 1233 | 877 | 71.13  
| VP | 679 | 132 | 19.44  
| NP | 2974 | 902 | 30.33  
| GP | 26 | 15 | 57.69  
| PP | 96 | 12 | 12.5  
| XP | 0 | 0 | N/A  

Fig. 4. Official Bake-off 2012 test results of our base-phrase chunking system.

---

9 http://crfpp.googlecode.com/svn/trunk/doc/index.html

---

Fig. 3. Partial parsing result of “Angels & Demons”, Dan Brown, 2000.

In the speech applications, the accuracy of BP phrase still needs to be improved. Using more training data will be the most effective way to improve the BP chunking.
Basically, the evaluation results show that our system achieved 0.5038 (0.7210/0.387) micro-
and 0.5301 (0.7343/0.4147) macro-
averaging F1 (precision/recall) on full sentence parsing task. 

However, it is believed that the main reason for low recall rate is only base phrases were 
tagged in our system. Therefore, if only the per-
formance of base phrase chunking were consid-
ered, the F-mesures may be around 0.70. The results are somehow good enough for speech 
recognition and synthesis applications. 

Another possibility of performance degrada-
tion is that the number of (X‧DE) phrases in the 
training corpus is above 13% of total base 
phrases (In fact, 的/de/) should be one of the 
most frequently occurred words in traditional 
Chinese text). But, there is no (X‧DE) phrase in 
the evaluation data. It may be the reason why the 
performance of base phrase chunking was de-
genrate from 0.84 to 0.70. 

4 Conclusions

In this paper, a Tradition Chinese base phrase 
parser that considered only base phrase chunking 
was implemented. The official Bake-off 2012 
evaluation results on full sentence parsing task 
show that our system achieved 0.5038 
(0.7210/0.387) micro-
and 0.5301 (0.7343/ 
0.4147) macro-
averaging F1 (precision/recall) rates. However, if only the performance of base 
phrase chunking was considered, the F-mesures 
may be around 0.70. Therefore, the results are 
somehow good enough for speech recognition 
and synthesis applications. In the near future, 
word senses and semantic information in Word-
net database will be explored to improve the 
performance of our system. 

Acknowledgments

This work was supported by the National Sci-
ence Council, Taiwan, ROC, under the project 
with contract NSC 101-2221-E-009-149-MY2 
and 101-2221-E-027-129. 

References

Adam L. Berger, Stephen A. Della Pietra, and Vic-
cent J. Della Pietra. 1996. A maximum entropy ap-
proach to natural language processing. Compu-
tational Linguistics, 22(1):39-71. 

Andrew Rosenberg. 2012. Rethinking The Corpus: 
Moving towards Dynamic Linguistic Resources. In 
Proceedings of INTERSPEECH-2012, Portland, 
USA. 

Hai Zhao, Chang-Ning Huang and Mu Li. 2006. An 
Improved Chinese Word Segmentation System with 
Conditional Random Field. In Proceedings of the 
Fifth SIGHAN Workshop on Chinese Language 
Processing : 108-117. Sidney, Australia. 

Igor Mel’čuk. 1996. Lexical Functions in Lexicogra-
phy and Natural Language Processing, chapter Lexical Functions: A Tool for the Description of 
Lexical Relations in the Lexicon, Benjamins Publishing Corp. 

J. Lafferty, A. McCallum, and F. Pereira. 2001. Con-
ditional random fields: Probabilistic models for seg-
menting and labeling sequence data. In Pro-
ceedings of In Intl. Conf. on Machine Learning : 
282-289. 

L. Shi and R. Mihalcea. 2005. Putting pieces together: 
Combining FrameNet, VerbNet and WordNet for 
robust semantic parsing. In Proceedings of Com-
putational Linguistics and Intelligent Text Pro-
cessing: Sixth International Conference : 100–111, 
Mexico City, Mexico. 

V. Vapnik. 1995. The Nature of Statistical Learning 
Theory. Springer-Verlag, New York. 

Wenliang Chen, Yujie Zhang, and Hitoshi Isahara. 
2006. An empirical study of Chinese chunking. In 
Proceedings of the COLING/ACL 2006 Main 
Conference Poster Sessions : 97–104, Sydney, Aus-
tralia, July. Association for Computational Linguistics.