Proceedings of the Second SIGHAN Workshop on Chinese Language Processing, July 2003, pp. 184-187.

HHMM-based Chinese Lexical Analyzer ICTCLAS*

Hua-Ping ZHANG[1] **Hong-Kui Yu**[1] **De-Yi Xiong**[1] **Qun LIU**[1,2]
[1]Inst. of Computing Tech., The Chinese Academy of Science, Beijing, 100080 CHINA
[2]Inst. of Computational Linguistics, Peking University, Beijing, 100871 CHINA
Email: zhanghp@software.ict.ac.cn

Abstract

This document presents the results from Inst. of Computing Tech., CAS in the ACŁ SIGHAN-sponsored First International Chinese Word Segmentation Bakeoff. The authors introduce the unified HHMM-based frame of our Chinese lexical analyzer ICTCLAS and explain the operation of the six tracks. Then provide the evaluation results and give more analysis. Evaluation on ICTCLAS shows that its performance is competitive. Compared with other system, ICTCLAS has ranked top both in CTB and PK closed track. In PK open track, it ranks second position. ICTCLAS BIG5 version was transformed from GB version only in two days; however, it achieved well in two BIG5 closed tracks. Through the first bakeoff, we could learn more about the development in Chinese word segmentation and become more confident on our HHMM-based approach. At the same time, we really find our problems during the evaluation. The bakeoff is interesting and helpful.

1 Introduction

ICT (Institute of Computing Technology, Chinese Academy of Sciences) participated the First International Chinese Word Segmentation Bakeoff. We have taken six tracks: Academia Sinica closed (ASc), U. Penn Chinese Tree Bank open and closed(CTBo,c), Hong Kong CityU closed (HKc), Peking University open and closed(PKo,c).

The structure of this document is as follows. The next section presents the HHMM-based framework of ICTCLAS. Next we detail the operation of six tracks. The following section provides evaluation result and gives further analysis.

2 HHMM-based Chinese lexical analysis

2.1 ICTCLAS Framework

As illustrated in Figure 1, HHMM-based Chinese lexical analysis comprises five levels: atom segmentation, simple and recursive unknown words recognition, class-based segmentation and POS tagging. In the whole frame, class-based segmentation graph, which is a directed graph designed for word segmentation, is an essential intermediate data structure that links disambiguation, unknown words recognition with word segmentation and POS tagging.

Atom segmentation, the bottom level of HHMM, is an initial step. Here, atom is defined to be the minimal segmentation unit that cannot be split in any stage. The atom consists of Chinese character, punctuation, symbol string, numeric expression and other non-Chinese char string. Any word is made up of an atom or more. Atom segmentation is to segment original text into atom sequence and it provides pure and simple source for its parent HMM. For instance, a sentence like "2002.9，ICTCLAS 的自由源码开始发布" (The free source codes of ICTCLAS was distributed in September, 2002) would be segmented as atom sequence "2002.9/，/ICTCLAS/的/自/由/源/码/开/始/发/布/". In this HMM, the original symbol is observation while the atom is state. We skip the detail of operation in that it's a simple application on the basis of HMM. POS tagging and role tagging using Viterbi are also skipped because they are classic application of HMM. Because of paper length limit, unknown words recognition is omitted. Our previous papers (Zhang et al. 2003) gave more

* ICTCLAS is the abbreviation for "Inst. of Computing Tech., Chinese Lexical Analysis System.". We published ICTCLAS as free software. The full source code and document of ICTCLAS is available at no cost for non-commercial use. Welcome researchers and technical users download ICTCLAS from Open Platform of Chinese NLP (www.nlp.org.cn).

explanation.

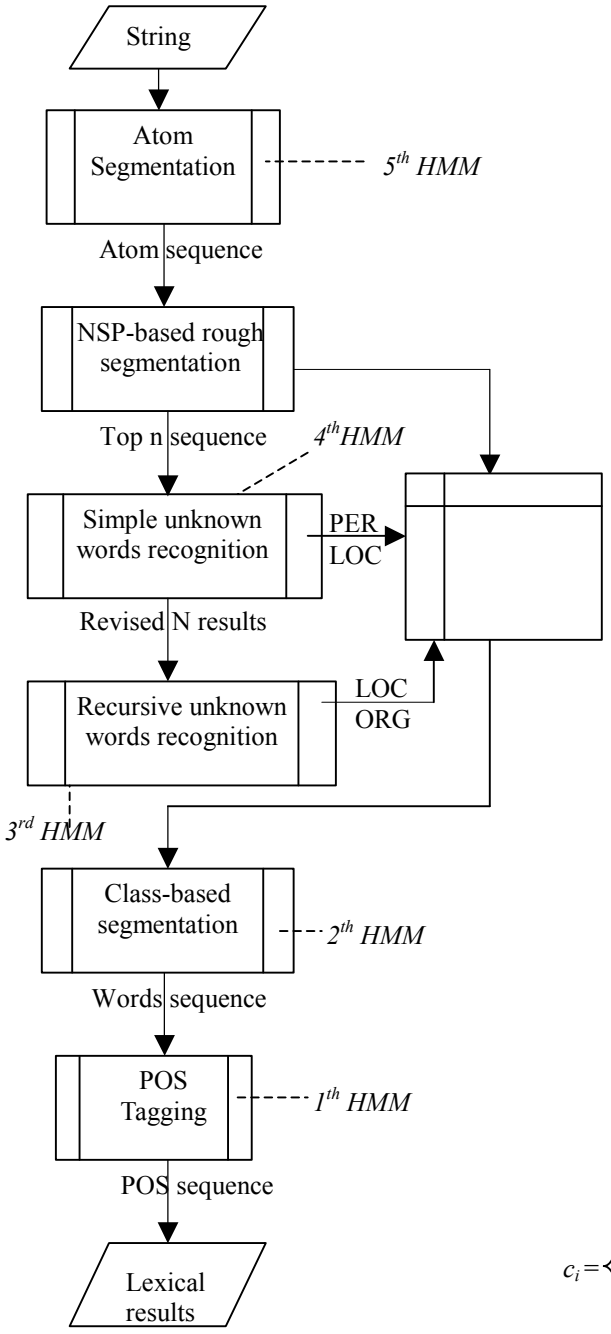

Figure 1. HHMM-based Chinese lexical analysis

2.2 Class-based HMM for word segmentation

We apply to word segmentation class-based HMM, which is a generalized approach covering both common words and unknown words.

Given a word w_i, class c_i is defined in Figure 2. Suppose |LEX| to be the lexicon size, then the total number of word classes is |LEX|+9.

Given the atom sequence A=(a_1,…a_n), let W=(w_1,…w_m) be the words sequence, C= (c_1,…c_m) be a corresponding class sequence of W, and $W^{\#}$ be the choice of word segmentation with the maximized probability, respectively. Then, we could get:

$$W^{\#} = \arg\max_{W} P(W|A) = \arg\max_{W} P(W,A)/P(A)$$

For a specific atom sequence A, P(A) is a constant and P(W,A)= P(W). So,

$$W^{\#} = \arg\max_{W} P(W)$$

On the basis of Baye's Theorem, it can be induced that:

$$W^{\#} = \arg\max_{W} P(W|C)P(C)$$

$W^{\#}$ can be found with another level of HMM if class c_i is viewed as state while word w_i is output. Therefore:

$$W^{\#} \approx \arg\max_{w_1 w_2 \ldots w_m} \prod_{i=1}^{m} p(w_i \mid c_i)\, p(c_i \mid c_{i-1}),$$

where c_0 is begin of sentence.

For convenience, we often use the negative log probability instead of the proper form. That is:

$$W^{\#} = \arg\min_{W} \sum_{i=1}^{m} [-\ln p(w_i \mid c_i) - \ln p(c_i \mid c_{i-1})]$$

$c_i =$
- w_i iff w_i is listed in the segmentation lexicon;
- PER iff w_i is unlisted* personal name;
- LOC iff w_i is unlisted location name;
- ORG iff w_i is unlisted organization name;
- TIME iff w_i is unlisted time expression;
- NUM iff w_i is unlisted numeric expression;
- STR iff w_i is unlisted symbol string;
- BEG iff beginning of a sentence
- END iff ending of a sentence
- OTHER otherwise.

* "unlisted" is referred as being outside the lexicon

Figure 2: Class Definition of word w_i

According to the word class definition, if w_i is listed in lexicon, then c_i is w_i, and $p(w_i|c_i)$ is equal to 1.0. Otherwise, $p(w_i|c_i)$ is probability that class

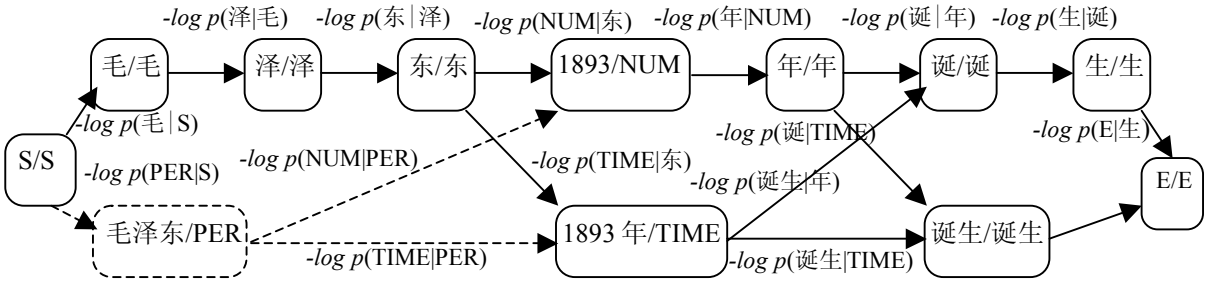

Figure3. Class-based word segmentation

Note:

1. The original sentence is "毛泽东 1893 年诞生" (Mao Ze-Dong was born in the year of 1893). Its atom sequence is "毛/泽/东/1893/年/诞/生/" after atom segmentation;
2. The node format is "word/class" (w_i / c_i) and the weight on the node is $-log\ p\ (w_i \mid c_i)$;
3. Weight on the directed edge is $-\log p\ (c_i \mid c_{i-1})$;
4. "毛泽东" (Mao Ze-Dong) is personal name outside the lexicon.The node " 毛泽东/PER" and the related edges with dash line is inserted after unknown words recognition.

c_i initially activates w_i, and it could be estimated in its child HMM for unknown words recognition.

As demonstrated in Figure 3, we provide the process of class-based word segmentation on "毛泽东 1893 年诞生" (Mao Ze-Dong was born in the year of 1893). The significance of our method is: it covers the possible ambiguity. Moreover, unknown words, which are recognized in the following steps, can be added into the segmentation graph and proceeded as any other common words.

After transformation through class-based HMM, word segmentation becomes single-source shortest paths problem. Hence the best choice $W^{\#}$ of word segmentation is easy to find using Djikstra's algorithm.

3 Tracks

Here, we would introduce the operation of some different track.

3.1 Closed Tracks

We participate all the closed tracks. As for each closed track, we first extracted all the common words and tokens that appear in the training corpus. Then build the segmentation core lexicons with the words. Those named entity words are classified into different named entities: numeric and time expression, personal names, location names, and transliterated names. According to named entities in the given corpus, we could train both class-based segmentation HMM and role-based HMM model for unknown word recognition. Therefore, the whole lexical system including unknown word detection is accomplished as shown in Figure 1.

3.2 Open Tracks

We only participate GB code open tracks. Actually, open track is similar to closed one. The only difference is the size of training data set. In Peking University open track, ICTCLAS is trained on six-month news corpus that is 5 months more than closed track. The entire corpus is also from Peking University. Except for the additional corpus, we have not employed any other special libraries or other resources.

As for CTB open track, we find that it cannot benefit from that 5 month PKU corpus. Actually, PKU standard is very different from CTB one though they seemed similar. Core lexicon extracted from Peking corpus degraded the performance on CTB testing data. Except for some named entity corpus, we could not get any more sources related to CTB standard. Therefore, CTB open track is operated in the similar way as closed track.

3.3 BIG5-coded Tracks

Before the bakeoff, BIG5-coded word segmentation has never been researched in our institute. Besides the character code, common words and sentence styles are greatly different in China mainland and Taiwan or Hong Kong. Because of time limitation, we have only spent two days on transforming our GB-coded ICTCLAS to BIG5-coded lexical analyzer. For each BIG5 closed, we extracted a BIG5-coded core lexicon. Then, the

BIG5 version ICTCLAS could work properly. The core source code is same as GB version.

4 Evaluation result

Track	ASc	CTBc	CTBo	HKc	PKc	PKo
Participant Number	6	6	7	4	10	8
Corpus Size (bytes)	38,882	125,248	125,248	114,384	56,254	56,254
True Word count	11,985	39,922	39,922	34,955	17,194	17,194
Test Word count	12,360	40,460	40,426	37,274	17,582	17,563
Insertions	434	1,789	1,755	2,439	485	458
Deletions	59	1,251	1,251	120	97	89
Substitutions	506	3,281	3,262	2,291	562	539
Nchange	999	6,321	6,268	4,850	1,144	1,086
Recall (Rank)	0.953 (3)	0.886 (2)	0.887 (4)	0.931 (3)	0.962 (1)	0.963 (1)
Precision (Rank)	0.924 (5)	0.875 (1)	0.876 (4)	0.873 (3)	0.940 (3)	0.943 (2)
F measure (Rank)	0.938 (5)	0.881 (1)	0.881 (4)	0.901 (3)	0.951 (1)	0.953 (2)
OOV rate	0.022	0.181	0.181	0.071	0.069	0.069
OOV Recall (Rank)	0.178 (5)	0.705 (1)	0.707 (5)	0.243 (4)	0.724 (2)	0.743 (2)
IV Recall(Rank)	0.970 (3)	0.927 (5)	0.927 (5)	0.984 (1)	0.979 (2)	0.980 (1)
***Time Cost (s)**	3.92	10.57	10.62	7.11	5.18	5.53
****Speed (bytes/s)**	9,919	11,849	11,794	16,088	10,860	10,173

Table 1. Evaluation result of ICTCLAS in the First International Chinese Word Segmentation Bakeoff

*Time Cost: CPU: Pentium 4, 1.6GHz; Main Memory: 192M

**Speed=Corpus Size / Time cost * 1000

Compared with other systems, ICTCLAS especially GB-coded version is competitive. In both GB-coded closed tracks, ICTCLAS ranked top. ICTCLAS also rank second position in Peking open track. Because of the lack of resources, CTB open track is almost as same as CTB closed track. The final performance in BIG5 track is not very good. As a preliminary BIG-coded system, however, we are satisfied with the result.

As is shown in Table 1, It could also be concluded that class-based segmentation HMM is effective. Excepted for CTB, IV Recall is over 97%.

5 Conclusion

Through the first bakeoff, we have learn more about the development in Chinese word segmentation and become more confident on our HHMM-based approach. At the same time, we really find our problems during the evaluation. The bakeoff is interesting and helpful. We look forward to participate forthcoming bakeoff.

6 Acknowledgements

The authors would like to thank Prof. Shiwen Yu of Peking University for the Peking corpus. And we acknowledge our debt to Gang Zou, Dr. Bin Wang, Dr. Jian Sun, Ji-Feng Li, Hao Zhang and other colleagues. Huaping Zhang would especially express gratitude to his graceful girl friend Feifei and her family for their encouragement. We also thank Richard Sproat, Qing Ma, Fei Xia and other SIGHAN colleagues for their elaborate organization and enthusiastic help in the First International Chinese Word Segmentation Bakeoff.

References

Lawrence. R.Rabiner.1989. *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition*. Proceedings of IEEE 77(2): pp.257-286.

Shai Fine, Yoram Singer, and Naftali Tishby.1998. *The hierarchical Hidden Markov Model: Analysis and applications*. Machine Learning, 32:41

Zhang Hua-Ping, Liu Qun. *Model of Chinese Words Rough Segmentation Based on N-Shortest-Paths Method*. Journal of Chinese information processing, 2002,16(5):1-7 (in Chinese)

ZHANG Hua-Ping, LIU Qun, Zhang Hao and Cheng Xue-Qi. 2002. *Automatic Recognition of Chinese Unknown Words Recognition*. Proc. of First SigHan attached on COLING 2002

ZHANG Hua-Ping, LIU Qun, YU Hong-Kui, CHENG Xue-Qi, BAI Shuo. *Chinese Named Entity Recognition Using Role Model*. International Journal of Computational Linguistics and Chinese language processing, 2003,Vol. 8 (2)