Proceedings of the Second SIGHAN Workshop on Chinese Language Processing, July 2003, pp. 63-70.

Chinese Lexical Analysis Using Hierarchical Hidden Markov Model

Hua-Ping ZHANG[1] **Qun LIU**[1,2] **Xue-Qi CHENG**[1] **Hao Zhang**[1] **Hong-Kui Yu**[1]
[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 paper presents a unified approach for Chinese lexical analysis using hierarchical hidden Markov model (HHMM), which aims to incorporate Chinese word segmentation, Part-Of-Speech tagging, disambiguation and unknown words recognition into a whole theoretical frame. A class-based HMM is applied in word segmentation, and in this level unknown words are treated in the same way as common words listed in the lexicon. Unknown words are recognized with reliability in role-based HMM. As for disambiguation, the authors bring forth an n-shortest-path strategy that, in the early stage, reserves top N segmentation results as candidates and covers more ambiguity. Various experiments show that each level in HHMM contributes to lexical analysis. An HHMM-based system ICTCLAS was accomplished. The recent official evaluation indicates that ICTCLAS is one of the best Chinese lexical analyzers. In a word, HHMM is effective to Chinese lexical analysis.

1 Introduction

Word is the independent and meaningful atom in natural language. Unlike English and Spanish, there is no delimiter to mark word boundaries and no explicit definition of words in some Asian languages. As for Chinese language processing, the fundamental task is word segmentation, which transforms Chinese character string into words sequence. It is prerequisite to POS tagger, parser and other deep processing, and the lexical result is the basis of further applications such as machine translation, information retrieval and information extraction.

Since the first system CDWS appeared in 1983, word segmentation has been researched intensively. Many solutions were proposed and could be broadly categorized into rules-based approaches that make use of linguistic knowledge and statistical approaches that train on corpus after machine learning. The classic rule-based approaches include maximum matching and shortest path (SP), which achieve the minimum number of segmented words. Zhang and Liu (2002) present an extended SP algorithm named "n-shortest paths". Some researchers introduce more complicated rules, such as error-driven learning (Hockenmaier and Brew, 1998) and parsing (Wu and Jiang, 1998). Rule is the only feasible way to segment words unless necessary resources such as large amount of corpus are available. With the development of hand-corrected resource, statistical approaches became more popular. The language models commonly applied are n-gram (Zhang and Liu, 2002; Gao et al., 2001), EM (Peng and Schuurmans, 2001), and channel noise model. As far as we know, however, there is yet neither purely rule-based system nor purely statistical one. It tends to tackle Chinese lexical problem with mixture of rules and statistical information. On one hand, trainable rules (Palmer, D. 1997) seem more adaptive and efficient in that rule-based approaches benefit from frequency of rule occurrence, on the other hand, statistical solutions employ rules to detect ambiguity, numeric expression, time and other named entities. Apart from the above approaches, we also notice some other promising ideas such as compression-based (Teahan et al., 2001), classifier-based (Xue and Susan, 2002) and self-supervised segmentation without lexicon. According to recent reports, word segmentation has achieved good result in precision, especially on texts that do not contain ambiguity or out-of-vocabulary words.

However, segmentation ambiguity and un-

known words[1] cause bottlenecks and greatly degrade performance in word segmentation. Ambiguous or unknown string is hard to be correctly segmented; at the same time, it also influences on segmenting its neighboring words. What's worse, ambiguity often occurs with unknown words. Take "克林顿对内塔尼亚胡说"(*Clinton said to Netanyahu*) as exemplification, "内塔尼亚胡"(*Netanyahu*) is unknown transliterated personal name, and both "对内" (*for home*) and "胡说" (*talk nonsense*) has two ambiguous segmentations: split into halves or not. Here, it's difficult to identify unknown word "内塔尼亚胡" because of the ambiguities, while disambiguation is also difficult to accomplish before unknown words detection. Therefore, the final lexical result is very likely to be "克林顿/对内/塔尼亚/胡说" instead of "克林顿/对/内塔尼亚胡/说".

Historically, much effort has been made in the two sub-problems of word segmentation. Almost all previous solutions (Chunyu et al. 2002; Zhang, 1998; Zheng, 1999) of disambiguation attempt to cover each possible case with trivial rules, while recently statistical approaches are applied in some special categories of ambiguity. For instance, vector space model was applied in combinational ambiguity (Luo et al. 2002). Concerning unknown word, we only need focus on unknown named entities, including personal name (PER), location name (LOC), and organization name (ORG). The motivation in named entity recognition is to utilize its components and contexts. Like word segmentation and disambiguation, the usual approach is to apply rules (Sun, 1993;Tan, 1999;Luo and Ji, 2001; Luo and Song, 2001). Recognition rules are summarized on name libraries or different linguistic phenomena. Compared with rules-based approach, machine learning from large corpus seems easy but better in performance. The statistical approaches proposed recently include hidden Markov model (Zhang and Liu, 2002; Zhang et al. 2002), agent-based (Ye, 2003), class-based trigram model (Sun et al., 2002).

After nearly 20 years of hard work, rapid progresses are made on word segmentation, disambiguation and unknown word recognition research individually. To the best of our knowledge, however, all the achievement has not ever, all the achievement has not integrated into a unified model with a general theoretical basis. In previous lexical analyzers, so-called word segmentation algorithm actually only employs on common words listed in the lexicon, while disambiguation and unknown word recognition have their own independent mechanism and become distinct processes from segmentation. Without scientific quantification, unknown words and disambiguation result could not compete with other segmentation candidates. In a word, previous work lacks a whole frame incorporating the different sub-tasks in lexical analysis, while there is also no consistent mechanism to evaluate various lexical results Therefore, previous lexical system is difficult achieve better performance on real texts that contain irregular character strings mentioned above.

This paper presents an HHMM-based approach for Chinese lexical analysis. It aims to utilize a general model to proceed all steps in lexical analysis, including word segmentation, disambiguation, unknown words recognition and part-of speech (POS) tagging. In the preprocessing, top n segmentation candidates covering the possible ambiguity are provided using n-shortest-path algorithm (Zhang and Liu, 2002). Then, simple unknown named entities like personal names and location names are identified on the candidate set using class-based HMM. Following that, a higher level of HMM could be employed on recognizing organization and other recursive named entity, which includes another simple unknown word. Unknown words recognized with credible probability are added to class-based HMM for word segmentation. In this level of HHMM, unknown words and ambiguity are treated in the same way as common words. POS tagging is the top level in HHMM. After HHMM based approach applied, Chinese lexical analysis system ICTCLAS achieves well in segmentation and POS tagging. The official evaluation, which was held by the National Foundation of 973 Plan of China, shows that ICTCLAS rank top and it is one of the best Chinese lexical analyzers.

The structure of this paper is as follows. The next section reviews HHMM and presents the framework of HHMM-based Chinese lexical analysis. Then we explain the class-based HMM for word segmentation. Next we detail role-based unknown words recognition and n-shortest-path disambiguation. The following section describes

1 We define unknown words to be those neitheri n-cluded in the core lexicon nor recognized through FSA.

various experiments designed to evaluate lexical analysis performance and contribution from different level in HHMM.

2 HHMM and Chinese lexical analysis

2.1 An overview of HHMM

Hidden Markov model (HMM, L.R. Rabiner, 1989) has become the method of choice for modeling stochastic processes and sequence in natural language processing, because HMM is very rich in mathematical structure and hence can form theoretical basis for use. However, compared with the sophisticated phenomena in natural language, traditional HMM seems hard to use due to the multiplicity of length scales and recursive nature of the sequences. Therefore Shai Fine et al (1998) proposed hierarchical hidden Markov model, which is a recursive and generalized HMM.

Based on Shai's work, we give a formal description of HHMM. An HHMM is specified by a six-tuple (S , O , Π, A, B, D), where D is the depth of levels, S and O are the finite set of states and the final output alphabet or intermediate output, and Π,A and B are the probabilities of the initial state, state transitions and emissions of symbol or intermediate output, respectively. The contrast between traditional HMM and HHMM lies in:

1) The state set S can be classified into different sub-sets according to its level. A state in S is annotated with q_i^d ($0<d\leq D$, $0<i\leq |S^d|$), where d is the level index, i is the state index and S^d is the set of state in level d. When d=D, q_i^d is called terminal state because its observation is symbols, or else, it is called internal state whose observation is from its child HMM in (d+1)th level.

2) Every internal state q^d ($0<d<D$) has its child states, which form an independent HMM. In the child HMM, the state transitive probabilities are

$$A(q^d)=(a_{ij}(q^d)), and\ a_{ij}(q^d)=P(q_j^{d+1}|q_i^{d+1}).$$

And the initial distribution vector is like $\Pi(q^d)=(\pi^d(q_i^{d+1}))=(P(q_i^{d+1}|q^d))$, where $P(q_i^{d+1}|q^d)$ is defined to be the probability that state q^d initially activates its child state q_i^{d+1}.

3) Only the bottom HMM can observe the symbols. The corresponding symbol emission probabilities are $B(q^D)=(b_k(q^D)))$, where $b_k(q^D)=P(o_k|q^D)$ and O_k is in symbol set. For the d (d<D) level HMM, state sequence in its child HMM could be viewed as its observation. The emission probabilities could be estimated as above.

All in all, HHMM includes D levels of HMM while each level is independent HMM. Moreover, each HMM only links with its parent and child. The whole parameters set of HHMM is denoted by

$$\lambda=\{(\lambda(q^d)\}_{d\in\{1,...,D\}}$$
$$=\{\{A(q^d)\}_{d\in\{1,...,D-1\}},\{\Pi(q^d)\}_{d\in\{1,...,D-1\}},\{B(q^D)\}\}.$$

Actually, HMM is the specific form of HHMM with D=1.

2.2 Framework of HHMM-based lexical analysis

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

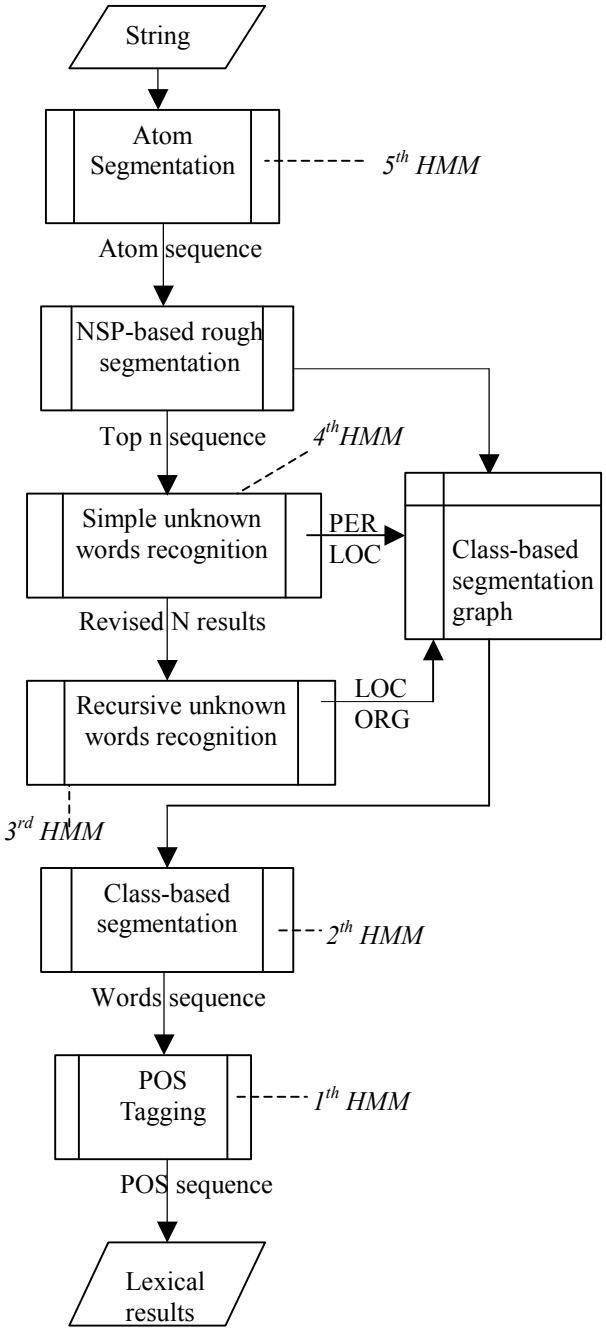

Figure 1. HHMM-based Chinese lexical analysis

detail of operation in that it's a simple application on the basis of HMM. POS tagging using HMM is also skipped because role tagging, which presented in section 5, is similar to it in nature. The other levels of HHMM will be provided in the next parts.

3 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.

$$c_i = \begin{cases} w_i & \text{iff } w_i \text{ is listed in the segmentation lexicon;} \\ \text{PER} & \text{iff } w_i \text{ is unlisted}^* \text{ personal name;} \\ \text{LOC} & \text{iff } w_i \text{ is unlisted location name;} \\ \text{ORG} & \text{iff } w_i \text{ is unlisted organization name;} \\ \text{TIME} & \text{iff } w_i \text{ is unlisted time expression;} \\ \text{NUM} & \text{iff } w_i \text{ is unlisted numeric expression;} \\ \text{STR} & \text{iff } w_i \text{ is unlisted symbol string;} \\ \text{BEG} & \text{iff beginning of a sentence} \\ \text{END} & \text{iff ending of a sentence} \\ \text{OTHER} & \text{otherwise.} \end{cases}$$

* "unlisted" is referred as being outside the lexicon

Figure 2: Class Definition of word w_i

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.

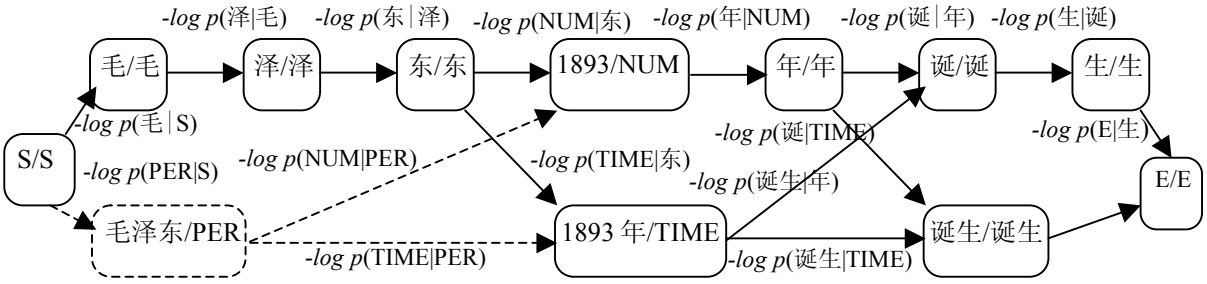

Figure3. Class-based word segmentation

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.

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

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

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 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.

4 NSP-based disambiguation strategy

Segmentation ambiguous error is made mainly because of improper decision in the earlier stage. For example, overlapping ambiguity in "结合/成/分子/时" (When combining into molecule) and combining ambiguity in "这/个/人/手/上/有/痣"(The person has naevi on his hand) are difficult to solve only in the initial stage of word segmentation. However, it's simple to find the correct result among the possible candidates in POS tagging or further processes. Therefore, the initial process should not make the final decision, but provide candidates covering the correct segmentation.

We take n-shortest-path (NSP, Zhang and Liu, 2002) algorithm as the disambiguation strategy. NSP, which selects n shortest paths, is an extension of Djikstra's algorithm. The motivation in disambiguation using NSP is covering more ambiguity with top n results in rough segmentation, which is the initial step in lexical analysis and produces candidate results.

Considering efficiency and performance, rough segmentation coverage, which is percentage of correct results, should be much higher while the average size of candidate set should be as small as possible. Compared with NSP, full segmentation, which produces all the possible segmentation paths, suffers from large amount of candidates, while other approaches lose so many correct results. As shown in Table 1, NSP-based rough segmentation enjoys two good properties: higher coverage and fewer candidates. In other word, NSP is effective strategy for disambiguation.

Approach	Max Size	AV Size	Coverage
MM	1	1	85.46%
SP	1	1	91.80%
ML	1	1	93.50%
FS	>3,424,507	>391.79	100.00%
NSP	8	5.82	99.92%

Table 1. Comparison between NSP and other approaches of rough segmentation

Note:

1) MM: maximum matching; SP: shortest path; ML: Maximum likelihood; FS: Full segmentation
2) Max size and AV size is the maximum and average size of segmentation candidate set, respectively;
3) Coverage=# of correctly segmented/# of sentence*100%
4) The size of testing set is 2 million Chinese characters.

5 Unknown words recognition using role-based HMM

The task includes: locating the boundary of a unknown word w_i, identifying the word class c_i, and computing the probability $p(w_i|c_i)$, which is required in class-based segmentation. Here, we introduce two levels of HMM to recognize simple and recursive unknown words on the rough segmentation set.

5.1 Role set for unknown words recognition

In the same way of class-based HMM for word segmentation, here we classify word class into various role according to its linguistic features shown in unknown words recognition. In table 2, we present a simplified role set for unknown personal name recognition. Role is similar as word class. Their difference is: a word has only a word class, but a word class has one role or more.

Role	Significance	Sample
A	Previous context	***来到***/于/洪/洋/的/家
B	Next context	黄/文/ ***摄***
C	Surname	***欧阳***/修
D	First token of 2-Hanzi given name	朱/***镕***/基/总理
E	Second token of 2-Hanzi* given name	朱/镕/***基***/总理
H	Suffix	王/***总***; 刘/***老***
L	Token in transliterated name	***蒙***/帕/蒂/·/梅/拉/费
Z	Remote context	***深切***/ 缅怀/邓/小/平
…		

Table 2. Simplified role set of personal names

* Hanzi: Chinese character

5.2 Role tagging and Recognizing Unknown words recognition

Given a word sequence W=(w_1,…w_n) , we could get its class result C=(c_1,…c_n). Now we could tag W with role R=(r_1,…r_n), where all roles are from the same set. Among all the roles sequence, we select the sequence $R^{\#}$ with the maximum probability as the final choice. Through the same induction detailed in section 3, we could get

$$R^{\#} = \arg\min_{R} \sum_{i=1}^{n} [-\ln p(c_i \mid r_i) - \ln p(r_i \mid r_{i-1})]$$

It is a tagging process and we make use of Viterbi algorithm (L.R.Rabiner, 1988) that selects the global optimum among all the state sequences. Here, tagging word class sequence“毛/泽/东/TIME/诞生” (Mao Ze-Dong was born in sometime.) with personal roles, we could get $R^{\#}$=“毛/C 泽/D 东/E TIME/B 诞生/Z” through Vitebi selection.

Unknown words are recognized through maximum pattern matching on role sequence. For instance, “C”, “D”,“E” is surname, first and second token of 2-Hanzi given name, respectively. So token sequence tagged with role “CDE” is likely to form a traditional Chinese personal name. Therefore, “毛泽东” will be recognized as a Chinese personal name according to its roles.

Let w_i be recognized unknown word and c_i be the word class, we estimate the probability $p(w_i|c_i)$ with the following formula:

$$p(w_i \mid c_i) = \prod_{j=0}^{k-1} p(c_{p+j} \mid r_{p+j}) \times \prod_{j=1}^{k-1} p(r_{p+j} \mid r_{p+j-1});$$

where w_i is made up of tokens from pth to (p+k-1)th.

Hence p(毛泽东|PER)=p(毛|C) p(泽|D) p(东|E) p(D|C)p(E|D). Finally unknown word “毛泽东” and p(毛泽东|PER) can be added into the class-based HMM, shown as dashed area in Figure 3.

5.3 Recursive unknown word recognition

Organization name like “周恩来和邓颖超纪念馆”(Memorial Hall of Zhou En-Lai and Deng Yun-Chao) and some sophisticated location name like “张自忠路”(Zhang Zi-Zhong Road) often include one or more unknown words. We call them “recursive unknown word”.

Our solution is: Firstly, recognizing non-recursive unknown words in the lower level of role-based HMM, then revising the word class sequence with the recognized results; next applying another role HMM to recognize the recursive ones. Take the original word class sequence “周/恩/来/和/邓/颖/超/纪念馆” as exemplification. In the first step, “周恩来” and “邓颖超” would be recognized as personal name. Then, the original class sequence could be replaced with “PER/和/PER/纪

念馆”. Based on the revised class result, the higher role-based HMM could recognize the recursive unknown word “周恩来和邓颖超纪念馆” as an organization name. Our method utilizes previous results and greatly reduces data sparseness.

The role training set is transformed from corpus tagged with POS. Zhang and Liu (2002) provided the algorithm for role data conversion, model training, named entity recognition and the other procedures in role-based HMM.

6 Experiments

An HHMM-based Chinese lexical system ICTCLAS was accomplished. The following experiments are performed on ICTCLAS.

As commonly used, we conduct our evaluations on terms of segmentation accuracy (SEG), accuracy of POS tagging (TAG1) with 24 tags, accuracy of POS tagging (TAG2) with 48 tags, precision of named entity recognition (P), recall of named entity recognition (R) and F-measure (F) that is weighted combination of P and R. They are calculated as following:

SEG= # of correctly segmented words/ # of words;
TAG1= # of correctly tagged 24-tag POS/ # of words;
TAG2= # of correctly tagged 48-tag POS/ # of words;
P= # of correct recognized NE/# of recognized NE;
R= # of correct recognized NE/# of NE×100%;

$F = \frac{R \times P \times (1+\beta^2)}{R + P \times \beta^2}$, here β is assigned with 1, and F is called F-1.

6.1 Chinese Lexical analysis and HHMM

On 1,108,049-word news corpus from the People ’s Daily, we conduct four experiments:

1) BASE: ICTCLAS with only class-based segmentation and POS tagging;
2) +PER: Adding role-based HMM for personal name recognition to BASE;
3) +LOC: Adding role-based HMM for location name recognition to +PER;
4) +ORG: Adding role-based HMM for location name recognition to +LOC.

Figure 4 gives the contrast among the four experiments in performance. It indicates that: firstly, every level in HHMM contributes to lexical analysis. For instance, SEG increases from 96.55% to 97.96% after personal HMM is added. If all levels of HMM are integrated, ICTCLAS achieves 98.25% SEG, 95.63% TAG1 and 93.38% TAG2. Secondly, low levels in HHMM benefits from the higher one. After organization recognition is applied, F-1 value of organization adds by 25.91%, furthermore, the performance of segmentation, POS tagging and recognition of personal and location name improves, too. It is because high level not only solves its own problem, but also helps the lower HMM filter improper candidate. For example, in the sentence “刘庄的水很甜”(The water in Liu village is sweet), “刘庄”(Liu village) is very likely to be incorrectly recognized as a personal name in +PER experiment. However, it will be revised as a location name in +LOC experiment.

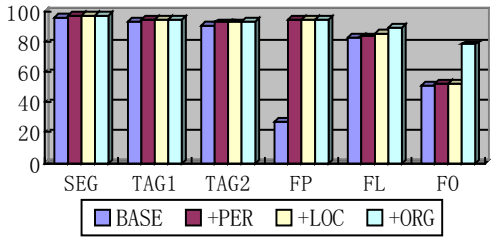

Figure 4. Contrast among 4 cases in performance
Note:
FP: F-1 value of personal name recognition;
FL: F-1 value of location name recognition;
FO: F-1 value of organization name recognition

6.2 Official evaluation on ICTCLAS

On July 6, 2002, ICTCLAS participated the official evaluation, which was held by the National Foundation of 973 Project of China. The open evaluation is conducted on real texts from six domains. The performance of ICTCLAS lists as Table 3.

Domain	Words	SEG	TAG1	RTAG
Sport	33,348	97.01%	86.77%	89.31%
Int. news	59,683	97.51%	88.55%	90.78%
Literature	20,524	96.40%	87.47%	90.59%
Law	14,668	98.44%	85.26%	86.59%
Theoretics	55,225	98.12%	87.29%	88.91%
Economics	24,765	97.80%	86.25%	88.16%
Total:	208,213	97.58%	87.32%	89.42%

Table 3. Official evaluation result of ICTCLAS
Note:
1) RTAG=TAG1/SEG*100%
2) The result about POS is not comparable because our tag set is greatly different from theirs.

Compared with other systems, ICTCLAS ranked top in the evaluation, and it is one of the

best Chinese lexical analyzer.

7 Conclusion

Our contributions are:

1) Applying HHMM to different lexical tasks, including word segmentation, POS tagging, unknown words recognition, and disambiguation.

2) Using class-based HMM for word segmentation, which integrates common words and unknown ones into a unified frame.

3) Proposing NSP strategy for segmentation disambiguation.

4) Bringing forth role-based HMM to recognize simple and recursive unknown words.

Various experiments show that each level in HHMM contributes to the final performance. Evaluation on ICTCLAS confirms that HHMM-based Chinese lexical analysis is effective.

8 Acknowledgements

The authors wish to thank Prof. Shiwen Yu of Peking University for the training corpus. And we acknowledge our debt to Gang Zou, Dr. Bin Wang, Dr. Jian Sun, Ji-Feng Li and other colleagues. Huaping Zhang would especially express gratitude to his graceful girl friend Feifei and her family for their encouragement during the hard work. We also thank three anonymous reviewers for their elaborate and helpful comments.

References

Andi Wu ,Zixin Jiang. *Word Segmentation in Sentence Analysis*. 1998 International Conference on Chinese Information Processing, Beijing, 1998. 169-180.

Chunyu Kit,Haihua Pan and Hongbiao Chen. *Learning Case based Knowledge for Disambiguating Chinese Word Segmentation: A preliminary study*. First SIGHAN Workshop attached with the 19th COLING, 2002.8, pp.63-70

Dai, Y., Khoo, C.S.G. and Loh, T.E. 1999. *A new statistical formula for Chinese text segmentation incorporating contextual information*. Proc ACM SIGIR99, pp. 82-89.

Gao Shan, Zhang Yan. *The Research on Integrated Chinese Word Segmentation and Labeling based on trigram statistical model*, proceeding of natural language understanding and machine translation, Beijing, Tsinghua University Press. 2001.116-122; (in Chinese)

Hockenmaier, J. and Brew, C. .1998. *Error-driven learning of Chinese word segmentation*. In J. Guo, K. T. Lua, and J. Xu, editors, 12th Pacific Conference on Language and Information, pp. 218-229, Singapore. Chinese and Oriental Languages Processing Society.

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

Luo Xiao, Sun, Maosong Benjamin K Tsou. *Covering Ambuguity Resolution in Chinese Word Segmentation Based on Contextual Information*. the 19th COLING, 2002.8, pp.598-604

Luo Z. and Song R. 2001. *Integrated and Fast Recognition of Proper Noun in Modern Chinese Word Segmentation*. Proceedings of International Conference on Chinese Computing 2001, Singapore, pp. 323-328.

Nianwen Xue and Susan P. Converse. *Combining Classifiers for Chinese Word Segmentation*, First SIGHAN Workshop attached with the 19th COLING, 2002.8, pp.63-70.

Palmer, D. 1997. *.A trainable rule-based algorithm for word segmentation.* Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL '97), Madrid, 1997.

Peng, F. and Schuurmans, D. (2001). *A hierarchical EM approach to word segmentation*. In 6th Natural Language Processing Pacific Rim Symposium (NLPRS-2001)

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

Sun J., Gao J. F., Zhang L., Zhou M Huang, C.N.2002. *Chinese Named Entity Identification Using Class-based Language Model*, Proc. of the 19th International Conference on Computational Linguistics, Taipei, pp 967-973

Sun M.S. (1993) English Transliteration Automatic Recognition. In "Computational Language Research and Development", L. W. Chen & Q. Yuan, ed., Beijing Institute of Linguistic Press.

Tan H. Y. (1999) *Chinese Place Automatic Recognition Research.* In "Proceedings of Computational Language ", C. N. Huang & Z.D. Dong, ed., Tsinghua Univ. Press, Beijing, China

Teahan, W. J. and Wen, Y. and McNab, R. and Witten I. H. 2001, *A Compression-based Algorithm for Chinese Word Segmentation.* In Comput. Ling., 26(3):375-393.

Xiao Luo, Maosong Sun, Benjamin K Tsou. *Covering Ambuguity Resolution in Chinese Word Segmentation Based on Contextual Information*. the 19th COLING, 2002.8, pp.598-604

Ye S.R, Chua T.S., Liu J. M. 2002. *An Agent-based Approach to Chinese Named Entity Recognition*, Proc. of the 19th International Conference on Computational Linguistics, Taipei, pp 1149-1155

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 COLING 2002