Chinese Named Entity Identification Using Class-based Language Model[1]

Jian Sun*, Jianfeng Gao#, Lei Zhang, Ming Zhou#, Changning Huang#**

* Beijing University of Posts & Telecommunications, China, jiansun_china@hotmail.com
# Microsoft Research Asia, {jfgao, mingzhou, cnhuang}@microsoft.com
** Tsinghua University, China

Abstract

We consider here the problem of Chinese named entity (NE) identification using statistical language model(LM). In this research, word segmentation and NE identification have been integrated into a unified framework that consists of several class-based language models. We also adopt a hierarchical structure for one of the LMs so that the nested entities in organization names can be identified. The evaluation on a large test set shows consistent improvements. Our experiments further demonstrate the improvement after seamlessly integrating with linguistic heuristic information, cache-based model and NE abbreviation identification.

1 Introduction

NE identification is the key technique in many applications such as information extraction, question answering, machine translation and so on. English NE identification has achieved a great success. However, for Chinese, NE identification is very different. There is no space to mark the word boundary and no standard definition of words in Chinese. The Chinese NE identification and word segmentation are interactional in nature.

This paper presents a unified approach that integrates these two steps together using a class-based LM, and apply Viterbi search to select the global optimal solution. The class-based LM consists of two sub-models, namely the *context model* and the *entity model*. The *context model* estimates the probability of generating a NE given a certain context, and the *entity model* estimates the probability of a sequence of Chinese characters given a certain kind of NE. In this study, we are interested in three kinds of Chinese NE that are most commonly used, namely person name (PER), location name (LOC) and organization name (ORG). We have also adopted a variety of approaches to improving the LM. In addition, a hierarchical structure for organization LM is employed so that the nested PER, LOC in ORG can be identified.

The evaluation is conducted on a large test set in which NEs have been manually tagged. The experiment result shows consistent improvements over existing methods. Our experiments further demonstrate the improvement after integrating with linguistic heuristic information, cache-based model and NE abbreviation identification. The precision of PER, LOC, ORG on the test set is 79.86%, 80.88%, 76.63%, respectively; and the recall is 87.29%, 82.46%, 56.54%, respectively.

2 Related Work

Recently, research on English NE identification has been focused on the machine-learning approaches, including hidden Markov model (HMM), maximum entropy model, decision tree and transformation-based learning, etc. (Bikel *et al*, 1997; Borthwick *et al*, 1999; Sekine *et al*, 1998). Some systems have been applied to real application.

Research on Chinese NE identification is, however, still at its early stage. Some researches apply methods of English NE identification to Chinese. Yu *et al* (1997) applied the HMM approach where the NE identification is formulated as a tagging

[1] This work was done while the author was visiting Microsoft Research Asia

problem using Viterbi algorithm. In general, current approaches to NE identification (e.g. Chen, 1997) usually contain two separate steps: word segmentation and NE identification. The word segmentation error will definitely lead to errors in the NE identification results. Zhang (2001) put forward class-based LM for Chinese NE identification. We further develop this idea with some new features, which leads to a new framework. In this framework, we integrate Chinese word segmentation and NE identification into a unified framework using a class-based language model (LM).

3 Class-based LM for NE Identification

The *n*-gram LM is a stochastic model which predicts the next word given the previous *n-1* words by estimating the conditional probability $P(w_n|w_1 \ldots w_{n-1})$. In practice, trigram approximation $P(w_i|w_{i-2}w_{i-1})$ is widely used, assuming that the word w_i depends only on two preceding words w_{i-2} and w_{i-1}. Brown et al (1992) put forward and discussed n-gram models based on classes of words. In this section, we will describe how to use class-based trigram model for NE identification. Each kind of NE (including PER, LOC and ORG) is defined as a class in the model. In addition, we differentiate the transliterated person name (FN) from the Chinese person name since they have different constitution patterns. The four classes of NE used in our model are shown in Table 1. All other words are also defined as individual classes themselves (i.e. one word as one class). Consequently, there are $|V|$+4 classes in our model, where $|V|$ is the size of vocabulary.

Table 1: Classes defined in class-based model

Tag	Description
PN	Chinese person name
FN	Transliterated person name
LN	Location name
ON	Organization name

3.1 The Language Modeling

3.1.1 Formulation

Given a Chinese character sequence $S = s_1 \ldots s_n$, the task of Chinese NE identification is to find the optimal class sequence $C^* = c_1 \ldots c_m$ ($m <= n$) that maximizes the probability $P(C|S)$. It can be expressed in the equation (1) and we call it class-based model.

$$C^* = \arg\max_C P(C \mid S) = \arg\max_C P(C) \times P(S \mid C) \quad (1)$$

The class-based model consists of two sub-models: the context model $P(C)$ and the entity model $P(S|C)$. The context model indicates the probability of generating a NE class given a (previous) context. $P(C)$ is a priori probability, which is computed according to Equation (2):

$$P(C) \cong \prod_{i=1}^{m} P(c_i \mid c_{i-2} c_{i-1}) \quad (2)$$

$P(C)$ can be estimated using a NE labeled corpus. The entity model can be parameterized by Equation (3):

$$\begin{aligned} & P(S \mid C) = P(s_1 \ldots s_n \mid c_1 \ldots c_m) \\ & \cong P([s_1 \ldots s_{c_1-end}] \ldots [s_{c_m-start} \ldots s_n] \mid c_1 \ldots c_m) \\ & \cong \prod_{j=1}^{m} P([s_{c_j-start} \ldots s_{c_j-end}] \mid c_j) \end{aligned} \quad (3)$$

The entity model estimates the generative probability of the Chinese character sequence in square bracket pair (i.e. starting from c_j-*start* to c_j-*end*) given the specific NE class.

For different class, we define the different entity model.

For the class of PER (including PN and FN), the entity model is a *character-based* trigram model as shown in Equation (4).

$$P([s_{c_j-start} \ldots s_{c_j-end}] \mid c_j = PER) = \prod_{k=c_j-start}^{c_j-end} P(s_k \mid s_{k-2}, s_{k-1}, c_j = PER) \quad (4)$$

where s can be any characters occurred in a person name. For example, the generative probability of character sequence 李大鹏 (Li Dapeng) is much larger than that of 许多年 (many years) given the PER since 李 is a commonly used family name, and 大 and 鹏 are commonly used first names. The probabilities can be estimated with the person name list.

For the class of LOC, the entity model is a *word-based* trigram model as shown in Equation (5).

$$
\begin{aligned}
&P([s_{c_j-start}\ldots s_{c_j-end}]\,|\,c_j = LOC)\\
&\approx \max_W P(w_1\ldots w_l\,|\,c_j = LOC)\\
&= \max_W [\prod_{k=1}^{l} P(w_k\,|\,w_{k-2}w_{k-1}, c_j = LOC)]
\end{aligned}
\tag{5}
$$

where $W = w_1\ldots w_l$ is possible segmentation result of character sequence $s_{c_j-start}\ldots s_{c_j-end}$.

For the class of ORG, the construction is much more complicated because an ORG often contain PER and/or LOC. For example, the ORG "中国国际航空公司" (Air China Corporation) contains the LOC "中国" (China). It is beneficial to such applications as question answering, information extraction and so on if nested NE can be identified as well. In order to identify the nested PER, LOC in ORG[2], we adopted class-based LMs for ORG further, in which there are three sub models, one is the class generative model, and the others are entity model: person name model and location name model in ORG. Therefore, the entity model of ORG is shown in Equation (6) which is almost same as Equation (1).

$$
\begin{aligned}
&P([s_{c_j-start}\ldots s_{c_j-end}]\,|\,c_j = ORG)\\
&\cong \max_{C'}[P(C'|\,c_j)P([s_{c_j-start}..s_{c_j-end}]\,|\,C', c_j = ORG)]\\
&= \max_{C'}\begin{bmatrix} P(c'_1..c'_k\,|\,c_j = ORG)\\ \times P([s_{c_j-start}..s_{c_j-end}]\,|\,c'_1..c'_k, c_j = ORG)\end{bmatrix}\\
&\cong \max_{C'}\begin{bmatrix} \prod_{i=1}^{k} P(c'_i|\,c'_{i-2}c'_{i-1}, c_j = ORG)\\ \times \prod_{i=1}^{k} P([s_{c'_i-start}..s_{c'_i-end}]\,|\,c'_i, c_j = ORG)\end{bmatrix}
\end{aligned}
\tag{6}
$$

where $C' = c'_1\ldots c'_k$ is the sequence of class corresponding to the Chinese character sequence.

In addition, if c_j is a normal word,

$$P([s_{c_j-start}\ldots s_{c_j-end}]\,|\,c_j) = 1. \tag{7}$$

Based on the context model and entity models, we can compute the probability *P(C|S)* and can get the optimal class sequence. The Chinese PER and transliterated PER share the same context class model when computing the probability.

3.1.2 Models Estimation

As discussed in 3.1.1, there are two kinds of probabilities to be estimated: *P(C)* and *P(S|C)*. Both probabilities are estimated using Maximum Likelihood Estimation (MLE) with the annotated training corpus.

The parser NLPWin[3] was used to tag the training corpus. As a result, the corpus was annotated with NE marks. Four lists were extracted from the annotated corpus and each list corresponds one NE class. The context model *P(C)* was trained with the annotated corpus and the four entity models were trained with corresponding NE lists. The Figure 1 shows the training process. (Begin of sentence (BOS) and end of sentence (EOS) is added)

NLPWin Tagged Sentence	*<LOC>美国</LOC>总统<PER>布什</PER>乘坐<ORG>中国国际航空公司</ORG>的航班到达<LOC>中国<LOC>*
Context Class	*BOS LN 总统 PN 乘坐 ON 的 航班 到达 LN EOS*
LN list	*美国* *中国*
FN list	*布 什*
ON list	*中国 国际 航空 公司*
ON Class list	*LN 国际 航空 公司*
Corresponding English Sentence	*<LOC>U.S.</LOC>president <PER>Bush</PER> arrived in <LOC> P.R. China </LOC> by flight No.1 of <ORG>Air China Corp.</ORG>*

Figure 1: Example of Training Process

3.1.3 Decoder

Given a sequence of Chinese characters, the decoding process consists of the following three steps:

Step 1: All possible word segmentations are generated using a Chinese lexicon containing 120,050 entries. The lexicon is only used for segmentation and there is no NE tag in it even if one word is PER, LOC or

[2] For simplification, only nested person, location names are identified in organization. The nested person in location is not identified because of low frequency

[3] NLPWin system is a natural language processing system developed by Microsoft Research.

ORG. For example, 北京 (Beijing) is not tagged as LOC.

Step 2: NE candidates are generated from any one or more segmented character strings and the corresponding generative probability for each candidate is computed using entity models described in Equation (4)–(7).

Step 3: Viterbi search is used to select hypothesis with the highest probability as the best output. Furthermore, in order to identify nested named entities, two-pass Viterbi search is adopted. The inner Viterbi search is corresponding to Equation (6) and the outer one corresponding to Equation (1). After the two-pass searches, the word segmentation and the named entities (including nested ones) can be obtained.

3.2 Improvement

There are some problems with the framework of NE identification using the class-based LM. First, redundant candidates NEs are generated in the decoding process, which results in very large search space. The second problem is that data sparseness will seriously influence the performance. Finally, the abbreviation of NEs cannot be handled effectively. In the following three subsections, we provide solutions to the three problems mentioned above.

3.2.1 Heuristic Information

In order to overcome the redundant candidate generation problem, the heuristic information is introduced into the class-based LM. The following resources were used: (1) Chinese family name list, containing 373 entries (e.g. 张 (Zhang), 王 (Wang)); (2) transliterated name character list, containing 618 characters (e.g. 什 (shi), 顿 (dun)); and (3) ORG keyword list, containing 1,355 entries (e.g. 大学 (university), 公司(corporation)).

The heuristic information is used to constrain the generation of NE candidates. For PER (PN), only PER candidates beginning with the family name is considered. For PER (FN), a candidate is generated only if all its composing character belongs to the transliterated name character list. For ORG, a candidate is excluded if it does not contain one ORG keyword.

Here, we do not utilize the LOC keyword to generate LOC candidate because of the fact that many LOC do not end with keywords.

3.2.2 Cache Model

The cache entity model can address the data sparseness problem by adjusting the parameters continually as NE identification proceeds. The basic idea is to accumulate Chinese character or word n-gram so far appeared in the document and use them to create a local dynamic entity model such as $P_{bicache}(w_i \mid w_{i-1})$ and $P_{unicache}(w_i)$. We can interpolate the cache entity model with the static entity LM $P_{static}(w_i \mid w_1 ... w_{i-2} w_{i-1})$:

$$\begin{aligned} & P_{cache}(w_i \mid w_1 .... w_{i-2}, w_{i-1}) \\ & = \lambda_1 P_{unicache}(w_i) + \lambda_2 P_{bicache}(w_i \mid w_{i-1}) \\ & + (1-\lambda_1-\lambda_2) P_{static}(w_i \mid w_1 .... w_{i-1}) \end{aligned} \quad (8)$$

where $\lambda_1, \lambda_2 \in [0,1]$ are interpolation weight that is determined on the held-out data set.

3.2.3 Dealing with Abbreviation

We found that many errors result from the occurrence of abbreviation of person, location, and organization. Therefore, different strategies are adopted to deal with abbreviations for different kinds of NEs. For PER, if Chinese surname is followed by the title, then this surname is tagged as PER. For example, 左校长 (President Zuo) is tagged as <PER>左</PER> 校长. For LOC, if at least two location abbreviations occur consecutive, the individual location abbreviation is tagged as LOC. For example, 中日关系 (Sino-Japan relation) is tagged as <LOC> 中 </LOC><LOC>日</LOC> 关系. For ORG, if organization abbreviation is followed by LOC, which is again followed by organization keyword, the three units are tagged as one ORG. For example, 中共北京市委 (Chinese Communist Party Committee of Beijing) is tagged as <ORG>中共<LOC>北京</LOC> 市委 </ORG>. At present, we collected 112 organization abbreviations and 18 location abbreviations.

4 Experiments

4.1 Evaluation Metric

We conduct evaluations in terms of precision (P) and recall (R).

$$P = \frac{number\ of\ correctly\ identified\ NE}{number\ of\ identified\ NE} \quad (9)$$

$$R = \frac{number\ of\ correct\ identified\ NE}{number\ of\ all\ NE} \quad (10)$$

We also used the F-measure, which is defined as a weighted combination of precision and recall as Equation (11):

$$F = \frac{(\beta^2 + 1) \times P \times R}{(\beta^2 \times P) + R} \quad (11)$$

where β is the relative weight of precision and recall.

There are two differences between MET evaluation and ours. First, we include nested NE in our evaluation whereas MET does not. Second, in our evaluation, only NEs with correct boundary and type label are considered the correct identifications. In MET, the evaluation is somewhat flexible. For example, a NE may be identified partially correctly if the label is correct but the boundary is wrongly detected.

4.2 Data Sets

The training text corpus contains data from People's Daily (Jan.-Jun.1998). It contains 357,544 sentences (about 9,200,000 Chinese characters). This corpus includes 104,487 Chinese PER, 51,708 transliterated PER, 218,904 LOC, and 87,391 ORG. These data was obtained after this corpus was parsed with NLPWin.

We built the wide coverage test data according to the guidelines[4] that are just same as those of 1999 IEER. The test set (as shown in Table 2) contains half a million Chinese characters; it is a balanced test set covering 11 domains. The test set contains 11,844 sentences, 49.84% of the sentences contain at least one NE. The number of characters in NE accounts for 8.448% in all Chinese characters.

We can see that the test data is much larger than the MET test data and IEER data

[4] The difference between IEER's guidelines and ours is that the nested person and location name in organization are tagged in our guidelines.

Table 2: Statistics of Open-Test

ID	Domain	Number of NE Tokens			Size
		PER	LOC	ORG	
1	Army	65	202	25	19k
2	Computer	75	156	171	59k
3	Culture	548	639	85	138k
4	Economy	160	824	363	108k
5	Entertainment	672	575	139	104k
6	Literature	464	707	122	96k
7	Nation	448	1193	250	101k
8	People	1147	912	403	116k
9	Politics	525	1148	218	122k
10	Science	155	204	87	60k
11	Sports	743	1198	628	114k
	Total	5002	7758	2491	1037k

4.3 Training Data Preparation

The training data produced by NLPWin has some noise due to two reasons. First, the NE guideline used by NLPWin is different from the one we used. For example, in NLPWin, 北京市(Beijing City) is tagged as <LOC>北京</LOC> 市, whereas 北京市 should be LOC in our definition. Second, there are some errors in NLPWin results. We utilized 18 rules to correct the frequent errors. The following shows some examples.

LN + Location Key	→	*LN*
*LN + 驻 + LN + **馆*	→	*ON*
(中|美|英|法...) 国人	→	*LN + 人*

The Table 4 shows the quality of our training corpus.

Table 4 Quality of Training Corpus

NE	P (%)	R (%)	F (%)
PER	61.05	75.26	67.42
LOC	78.14	71.57	74.71
ORG	68.29	31.50	43.11
Total	70.07	66.08	68.02

4.4 Experiments

We conduct incrementally the following four experiments:

(1) Class-based LM, we view the results as baseline performance;
(2) Integrating heuristic information into (1);
(3) Integrating Cache-based LM with (2);
(4) Integrating NE abbreviation processing with (3).

4.4.1 Class-based LM (Baseline)

Based on the basic class-based models estimated with the training data, we can get the baseline performance, as is shown in Table 5. Comparing Table 4 and Table 5, we found that the performance of baseline is better than the quality of training data.

Table 5 Baseline Performance

NE	P (%)	R (%)	F (%)
PER	65.70	84.37	73.87
LOC	82.73	76.03	79.24
ORG	56.55	38.56	45.86
Total	72.61	72.44	72.53

4.4.2 Integrating Heuristic Information

In this part, we want to see the effects of using heuristic information. The results are shown in Table 6. In experiments, we found that by integrating the heuristic information, we not only achieved more efficient decoding, but also obtained higher NE identification precision. For example, the precision of PER increases from 65.70% to 77.63%, and precision of ORG increases from 56.55% to 81.23%. The reason is that adopting heuristic information reduces the noise influence.

However, we noticed that the recall of PER and LOC decreased a bit. There are two reasons. First, organization names without organization ending keywords were not marked as ORG. Second, Chinese names without surnames were also missed.

Table 6 Results of Heuristic Information Integrated into the Class-based LM

NE	P (%)	R (%)	F (%)
PER	77.63	80.89	79.23
LOC	80.05	80.80	80.42
ORG	81.23	36.65	50.51
Total	79.26	73.41	76.23

4.4.3 Integrating Cache-based LM

Table 7 shows the evaluation results after cache-based LM was integrated. From Table 6 and Table 7, we found that almost all the precision and recall of PER, LOC, ORG have obtained slight improvements.

Table 7 Results of our system

NE	P (%)	R (%)	F (%)
PER	79.12	82.06	80.57
LOC	80.11	81.27	80.69
ORG	79.71	39.89	53.17
Total	79.72	74.58	77.06

4.4.4 Integrating with NE Abbreviation Processing

In this experiment, we integrated with NE abbreviation processing. As shown in Table 8, the experiment result indicates that the recall of PER, LOC, ORG increased from 82.06%, 81.27%, 36.65% to 87.29%, 82.46%, 56.54%, respectively.

Table 8 Results of our system

NE	P (%)	R (%)	F (%)
PER	79.86	87.29	83.41
LOC	80.88	82.46	81.66
ORG	76.63	56.54	65.07
Total	79.99	79.68	79.83

4.4.5 Summary

From above data, we observed that (1) the class based SLM performs better than the training data automatically produced with the parser; (2) the distinct improvements is achieved by using heuristic information; (3) Furthermore, our method of dealing with abbreviation increases the recall of NEs.

In addition, the cache-based LM increases the performance not so much. The reason is as follows: The cache-based LM is based on the hypothesis that a word used in the recent past is much likely either to be used soon than its overall frequency in the language or a 3-gram model would suggest (Kuhn, 1990). However, we found that the same NE often varies its morpheme in the same document. For example, the same NE 中共北京市委 (Chinese Communist Party Committee of Beijing),北京市委 (Committee of Beijing City), 市委 (Committee) occur in order.

Furthermore, we notice that the segmentation dictionary has an important impact on the performance of NE identification. We do not think it is better if more words are added into dictionary. For example, because *中国人(Chinese)* is in our

dictionary, there is much possibility that *中国 (China)* in *中国人* is missed identified.

5 Evaluation with MET2 and IEER Test Data

We also evaluated on the MET2 test data and IEER test data. The results are shown in Table 9. The results on MET2 are lower than the highest report of MUC7 (PER: Precision 66%, Recall 92%; LOC: Precision 89%, Recall 91%; ORG: Precision 89%, Recall 88%, http://www.itl.nist.gov). We speculate the reasons for this in the following. The main reason is that our class-based LM was estimated with a general domain corpus, which is quite different from the domain of MUC data. Moreover, we didn't use a NE dictionary. Another reason is that our NE definitions are slightly different from MET2.

Table 9 Results on MET2 and IEER

NE	MET2 Data			IEER Data		
	P (%)	R (%)	F (%)	P (%)	R (%)	F (%)
PER	65.86	94.25	77.54	79.38	84.43	81.83
LOC	77.42	89.60	83.07	79.09	80.18	79.63
ORG	88.47	75.33	81.38	88.03	62.30	72.96
Total	77.89	86.09	81.79	80.82	76.78	78.75

6 Conclusions & Future work

In this research, Chinese word segmentation and NE identification has been integrated into a framework using class-based language models (LM). We adopted a hierarchical structure in ORG model so that the nested entities in organization names can be identified. Another characteristic is that our NE identification do not utilize NE dictionary when decoding.

The evaluation on a large test set shows consistent improvements. The integration of heuristic information improves the precision and recall of our system. The cache-based LM increases the recall of NE identification to some extent. Moreover, some rules dealing with abbreviations of NEs have increased dramatically the performance. The precision of PER, LOC, ORG on the test set is 79.86%, 80.88%, 76.63%, respectively; and the recall is 87.29%, 82.46%, 56.54%, respectively.

In our future work, we will be focusing more on NE coreference using language model. Second, we intend to extend our model to include the part-of-speech tagging model to improve the performance. At present, the class-based LM is based on the general domain and we may need to fine-tune the model for a specific domain.

ACKNOWLEDGEMENT

I would like to thank Ming Zhou, Jianfeng Gao, Changning Huang, Andi Wu, Hang Li and other colleagues from Microsoft Research for their help. And I want to thank especially Lei Zhang from Tsinghua University for his help in developing the ideas.

References

Borthwick. A. (1999) *A Maximum Entropy Approach to Named Entity Recognition.* PhD Dissertation

Bikel D., Schwarta R., Weischedel. R. (1997) *An algorithm that learns what's in a name*. Machine Learning 34, pp. 211-231

Brown, P. F., DellaPietra, V. J., deSouza, P. V., Lai, J. C., and Mercer, R. L. (1992). *Class-based n-gram models of natural language*. Computational Linguistics, 18(4):468--479.

Chinchor. N. (1997) *MUC-7 Named Entity Task Definition Version 3.5*. Available by from ftp.muc.saic.com/pub/MUC/MUC7-guidelines

Chen H.H., Ding Y.W., Tsai S.C. and Bian G.W. (1997) *Description of the NTU System Used for MET2*

Gao J.F., Goodman J., Li M.J., Lee K.F. (2001) *Toward a unified Approach to Statistical Language Modeling for Chinese.* To appear in ACM Transaction on Asian Language Processing

Kuhn R., Mori. R.D. (1990) *A Cache-Based Natural Language Model for Speech Recognition*. IEEE Transaction on Pattern Analysis and Machine Intelligence.Vol.12. No. 6. pp 570-583

Mikheev A., Grover C. and Moens M. (1997) *Description of the LTG System Used for MUC-7*

Sekine S., Grishman R. and Shinou H. (1998), "*A decision tree method for finding and classifying names in Japanese texts*", Proceedings of the Sixth Workshop on Very Large Corpora, Canada

Yu S.H., Bai S.H. and Wu P. (1997) *Description of the Kent Ridge Digital Labs System Used for MUC-7*

Zhang L. (2001) *Study on Chinese Proofreading Oriented Language Modeling*, PhD Dissertation