Using N-best Lists for Named Entity Recognition from Chinese Speech

Lufeng ZHAI*, Pascale FUNG*, Richard SCHWARTZ†, Marine CARPUAT‡, Dekai WU‡

* HKUST
Human Language Technology Center
Electrical & Electronic Engineering
University of Science and Technology
Clear Water Bay, Hong Kong
{lfzhai,pascale}@ee.ust.hk

† BBN Technologies
9861 Broken Land Parkway
Columbia, MD 21046
U.S.A
schwartz@bbn.com

‡ HKUST
Human Language Technology Center
Department of Computer Science
University of Science and Technology
Clear Water Bay, Hong Kong
{marine,dekai}@cs.ust.hk

Abstract

We present the first known result for named entity recognition (NER) in realistic large-vocabulary spoken Chinese. We establish this result by applying a maximum entropy model, currently the single best known approach for textual Chinese NER, to the recognition output of the BBN LVCSR system on Chinese Broadcast News utterances. Our results support the claim that transferring NER approaches from text to spoken language is a significantly more difficult task for Chinese than for English. We propose re-segmenting the ASR hypotheses as well as applying post-classification to improve the performance. Finally, we introduce a method of using *n*-best hypotheses that yields a small but nevertheless useful improvement NER accuracy. We use acoustic, phonetic, language model, NER and other scores as confidence measure. Experimental results show an average of 6.7% relative improvement in precision and 1.7% relative improvement in F-measure.

1. Introduction

Named Entity Recognition (NER) is the first step for many tasks in the fields of natural language processing and information retrieval. It is a designated task in a number of conferences, including the Message Understanding Conference (MUC), the Information Retrieval and Extraction Conference (IREX), the Conferences on Natural Language Learning (CoNLL) and the recent Automatic Content Extraction Conference (ACE).

There has been a considerable amount of work on English NER yielding good performance (Tjong Kim Sang *et al.* 2002, 2003; Cucerzan & Yarowsky 1999; Wu *et al.* 2003). However, Chinese NER is more difficult, especially on speech output, due to two reasons. First, Chinese has a large number of homonyms and the vocabulary used in Chinese person names is an open set so more characters/words are unseen in the training data. Second, there is no standard definition of Chinese words. Word segmentation errors made by recognizers may lead to NER errors. Previous work on Chinese textual NER includes Jing *et al.* (2003) and Sun *et al.* (2003) but there has been no published work on NER in spoken Chinese.

Named Entity Recognition for speech is more difficult than for text, since the most reliable features for textual NER (punctuation, capitalization, and syntactic patterns) are often not available in speech output. NER on automatically recognized broadcast news was first conducted by MITRE in 1997, and was subsequently added to Hub-4 evaluation as a task. Palmer *et al.* (1999) used error modeling, and Horlock & King (2003) proposed discriminative training to handle NER errors; both used a hidden Markov model (HMM). Miller *et al.* (1999) also reported results in English speech NER using an HMM model. In a NIST 1999 evaluation, it was found that NER errors on speech arise from a combination of ASR errors and errors of the underlying NER system.

In this work, we investigate whether the NIST finding holds for Chinese speech NER as well. We present the first known result for recognizing named entities in realistic large-vocabulary spoken Chinese. We propose to use the best-known model for Chinese textual NER—a maximum entropy model—on Chinese speech NER. We also propose using re-segmentation and post-classification to improve this model. Finally, we propose to integrate the ASR and NER components to optimize NER performance by making use of the *n*-best ASR output.

2. A Spoken Chinese NER Model

2.1 LVCSR output

We use the ASR output from BBN's Byblos system on broadcast news data from the Xinhua News Agency,

which has 1046 sentences. This system has a character error rate of 7%. We had manually annotated them with named entities as an evaluation set according to the PFR corpus annotation guideline (PFR 2001).

2.2 A maximum-entropy NER model with post-classification

To establish a baseline spoken Chinese NER model, we selected a maximum entropy (MaxEnt) approach since this is currently the single most accurate approach known for recognizing named entities in text (Tjong Kim Sang *et al.*, 2002, 2003, Jing *et al.*, 2003)[1]. In the CoNLL 2003 NER evaluation, 5 out of 16 systems use MaxEnt models and the top 3 results for English and top 2 results for German were obtained by systems that use MaxEnt.

Natural language can be viewed as a stochastic process. We can use $p(y|x)$ to denote the probability distribution of what we try to predict y (.e.g. part-of-speech tag, Named Entity tag) conditioned on what we observe x (e.g. previous POS or the actual word). The Maximum Entropy principle can be stated as follows: given some set of constrains from observations, find the most uniform probability distribution (Maximum Entropy) $p(y|x)$ that satisfies these constrains:

$$y^* = \arg\max_{yi} P(y_i \mid x_i)$$

$$P(y_i \mid x_i) = \frac{1}{Z(x_i)} \exp(\sum_{j=0}^{m} \lambda_j \cdot f_j(x_i, y_i))$$

$$Z(x_i) = \sum_{k=0}^{l} \exp(\sum_{j=0}^{m} \lambda_j \cdot f_j(x_i, y_k))$$

In the above equations, $f_j(x_i,y_k)$ is a binary valued feature function, and λ_j is a weight that indicates how important feature f_j is for the model. $Z(x_i)$ is a normalization factor. We estimate the weights using the improved iterative scaling (IIS) algorithm.

For our task, we first compare a character-based MaxEnt model to a word-based model. Since recognition errors also lead to segmentation errors which in turn have an adverse effect on the NER performance, we experiment with disregarding the word boundaries in the ASR hypothesis and instead re-segment using a MaxEnt segmenter. We also compare an approach of one-pass identification/classification to a two-pass approach where the identified NE candidates are classified later. In addition, we propose a hybrid approach of using one-pass identification/classification results, discarding the extracted NE tags, and re-classifying the extracted NE in a second pass.

[1] We exclude from the present focus the slight improvements that are usually possible to obtain by combination of multiple models, usually through ad hoc methods such as voting.

2.3 Experimental setup

We use two annotated corpora for training. One is a corpus of People's Daily newspaper from January 1998, annotated by the Institute of Computational Linguistics of Beijing University (the "PFR" corpus). This corpus consists of about 20k sentences, annotated with word segmentation, part-of-speech tags and three named-entity tags including person (PER), location (LOC) and organization (ORG) . We use the first 6k sentences to train our NER system. Our system is then evaluated on 2k sentences from People's Daily and 1k sentences from the BBN ASR output. The results are shown in Tables 1 and 3.

To compare our system to the IBM baseline described in (Jing *et al.* 2003), we need to evaluate our system on the same corpus as they used. Among the data they used, the only publicly available corpus is a human-generated transcription of broadcast news, provided by NIST for the Information Extraction – Entity Recognition evaluation (the "IEER" corpus). This corpus consists of 10 hours of training data and 1 hour of test data. Ten categories of NEs were annotated, including person names, location, organization, date, duration, and measure. A comparison of results is shown in Table 2.

2.4 Results and discussion

From text to speech
Table 1 compares the NER performances of the same MaxEnt model on the Chinese textual PFR test data and the one-best BBN ASR hypotheses. We can see a significant drop in performance in the latter. These results support the claim that transferring NER approaches from text to spoken language is a significantly more difficult task for Chinese than for English. We argue that this is due to the combination of different factors specific to spoken Chinese. First, Chinese has a large number of homonyms that leads to a degradation in speech recognition accuracy which in turn leads to low NER accuracy. Second, the vocabulary used in Chinese person names is an open set so many characters/words are unseen in the training data.

Comparison to IBM baseline
Table 2 compares results on IEER data from our baseline word-based MaxEnt model compared with that of IBM's HMM word-based model. These two models achieved almost the same results, which show that our NER system based on MaxEnt is state-of-the-art.

Re-segmentation effect
Table 3 shows that by discarding word boundaries from the ASR hypothesis, and then re-segmenting using our MaxEnt segmenter, we obtained a better performance in most cases. We believe that some reduction in

segmentation errors due to recognition errors is obtained this way; for example, in the ASR output, two words "号 令" in "签署 了 第 四十二 号 令" are misrecognized as one word "号令", which can be corrected by re-segmentation.

Post-classification effect

Table 3 also shows that the one-pass identification/classification method yields better results than the two-pass method. However, there are still errors in the one-pass output where the bracketing is correct, but the NE classification is wrong. In particular, the type ORG is easily confusable with LOC in Chinese. Both types of NEs tend to be rather long. We propose a hybrid approach by first using the one-pass method to extract NEs, and then removing all type information, combining words of one NE to a whole NE-word and post-classifying all the NE-words again. Our results in Figure 1 show that the post-classification combined with the one-pass approach performs much better on all types.

	PER			LOC			ORG		
	P	R	F	P	R	F	P	R	F
Newspaper text	.86	.76	.81	.87	.75	.81	.83	.83	.83
1-best ASR hypothesis	.22	.18	.20	.75	.79	.77	.43	.35	.39

Table 1. NER results on Chinese speech data are worse than on Chinese text data.

Model	Precision	Recall	F-measure
IBM HMM	77.51%	65.22%	70.83%
MaxEnt	77.3%	65.4%	70.9%

Table 2. Our NER baseline is comparable to the IBM baseline.

	PER			LOC			ORG		
	P	R	F	P	R	F	P	R	F
2-pass, word	.23	.18	.20	.75	.79	.77	.43	.35	.39
1-pass, word	.25	.20	.21	.76	.84	.80	.70	.25	.36
2-pass, character	.53	.43	.48	.67	.70	.68	.75	.59	.66
1-pass, character	.60	.45	.52	.56	.69	.62	.55	.35	.43

Table 3. The character model is better than the word model, and one-pass NER is better than two-pass.

3. Using N-Best Lists to Improve NER

Miller *et al.* (1999) performed NER on the one-best hypothesis of English Broadcast News data. Palmer & Ostendorf (2001) and Horlock & King (2003) carried out English NER on word lattices. We are interested in investigating how to best utilize the *n*-best hypothesis from the ASR system to improve NER performances. From Figure 1, we can see that recall increases as the number of hypotheses increases. Thus it would appear possible to find a way to make use of the *n*-best ASR output, in order to improve the NER performance. However, we can expect it to be difficult to get significant improvement since the same figure (Figure 1) shows that precision drops much more quickly than recall. This is because the *n*th hypothesis tends to have more character errors than the (*n*-1)th hypothesis, which may lead to more NER errors. Therefore the question is, given *n* NE-tagged hypotheses, what is the best way to use them to obtain a better NER overall performance than by using the one-best hypothesis alone?

One simple approach is to allow all the hypotheses to vote on a possible NE output. In simple voting, a recognized named-entity is considered correct only when it appears in more than 30 percent of the total number of all the hypotheses for one utterance. The result of this simple voting is shown in Table 4. Next, we propose a mechanism of weighted voting using confidence measure for each hypothesis. In one experiment, we use the MaxEnt NER score as confidence measure. In another experiment, we use all the six scores (acoustic, language model, number of words, number of phones, number of silence, or NER score) provided by the BBN ASR system as confidence measure. During implementation, an optimizer based on Powell's algorithm is used to find the 6 weights (ω_k) for each score (S_k). For any given hypothesis, confidence measure is given by:

$$W = \sum_{k=1}^{6} S_k \cdot \omega_k$$

The above confidence measure is then normalized into a final confidence measure for each hypothesis:

$$\hat{W}_i = \frac{\exp(W_i)}{\sum_{l=1}^{N} \exp(W_l)}$$

Finally, an NE output is considered valid if

$$\sum_{i=1}^{N} \hat{W}_i * \delta_i(NE) > 0.3$$

$$\delta_i(NE) = \begin{cases} 1, & NE \ occurs \ in \ the \ i-th \ hypothesis \\ 0, & Otherwise \end{cases}$$

3.1 Experimental setup

We use the *n*-best hypothesis of 1,046 Broadcast News Chinese utterances from the BBN LVCSR system. *n* ranges from one to 300, averaging at 68. Each utterance has a reference transcription with no recognition error.

3.2 Results and discussion

Table 4 presents the NER results for the reference sentence, one best hypothesis, and different *n*-best voting methods. Results for the reference sentences show the upper bound performance (68% F-measure) of applying a MaxEnt NER system trained from the

Chinese text corpus (e.g., PFR) to Chinese speech output (e.g., Broadcast News). From Table 4, we can conclude that it is possible to improve NER precision by using *n*-best hypothesis by finding the optimized combination of different acoustic, language model, NER, and other scores. In particular, since most errors in Chinese ASR seem to be for person names, using NER score on the *n*-best hypotheses can improve recognition results by a relative 6.7% in precision and 1.7% in F-measure.

Results	PER		LOC		ORG	
	F	P	F	P	F	P
Reference sentence	0.71	0.75	0.78	0.77	0.56	0.72
One best	0.46	0.50	0.75	0.74	0.54	0.69
n-best simple vote	0.45	0.59	0.76	0.75	0.56	0.71
n-best weighted vote (NE score)	0.46	0.59	0.77	0.76	0.55	0.71
n-best weighted vote (all scores)	0.48	0.53	0.75	0.73	0.55	0.69

Table 4. *n*-best weighted voting with NE score gives the best performance.

4. Conclusion

We present the first known result for named entity recognition (NER) in realistic large-vocabulary spoken Chinese. We apply a maximum entropy (MaxEnt) based system to the *n*-best output of the BBN LVCSR system on Chinese Broadcast News utterances. Our results support the claim that transferring NER approaches from text to spoken language is a significantly more difficult task for Chinese than for English. We show that re-segmenting the ASR hypotheses improves the NER performance by 24%. We also show that applying post-classification improves the NER performance by 13%. Finally, we introduce a method of using *n*-best hypotheses that yields a useful 6.7% relative improvement in NER precision, and 1.7% relative improvement in F-measure, by weighted voting.

Acknowledgements. We would like to thank the Hong Kong Research Grants Council (RGC) for supporting this research in part via grants HKUST6206/03E, HKUST6256/00E, HKUST6083/99E, DAG03/04.EG30, and DAG03/04.EG09.

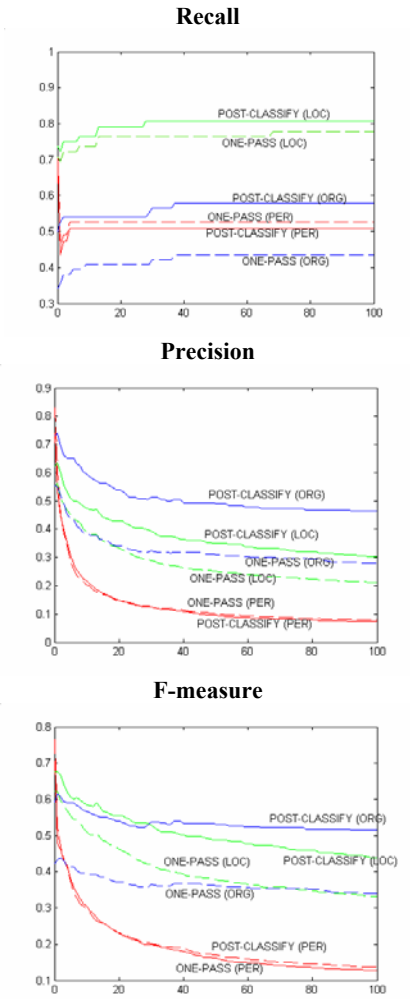

Figure 1. Post-classification improves NER performance.

References

Silviu CUCERZAN and David YAROWSKY. 1999. Language independent named entity recognition combining morphological and contextual evidence. *Proceedings of the 1999 Joint SIGDAT Conference on EMNLP and VLC*. University of Maryland, MD.

James HORLOCK and Simon KING. 2003. Discriminative Methods for Improving Named Entity Extraction on Speech Data. *Proceedings of Eurospeech 2003*. Geneva.

Institute of Computational Linguistics, Beijing University. 2001. The PFR corpus. http://icl.pku.edu.cn/research/corpus/shengming.htm.

Hongyan JING, Radu FLORIAN, Xiaoqiang LUO, Tong ZHANG and Abraham ITTYCHERIAH. 2003. HowtogetaChineseName(Entity): Segmentation and combination issues. *Proceedings of EMNLP*. Sapporo, Japan: July 2003.

David MILLER, Richard SCHWARTZ, Ralph WEISCHEDEL and Rebecca STONE. 1999. Named entity extraction from broadcast news. *Proceedings of the DARPA Broadcast News Workshop*. Herndon, Virginia: 1999. 37-40.

David D. PALMER, Mari OSTENDORF and John D. BURGER. 1999. Robust information extraction from spoken language data. *Proceedings of Eurospeech 1999*. Sep 1999.

Jian SUN, Ming ZHOU and Jianfeng GAO. 2003. A class-based language model approach to Chinese named entity identification. *Computational Linguistics and Chinese Language Processing*. 2003.

Erik F. TJONG KIM SANG. 2002. Introduction to the CoNLL-2002 Shared Task: Language-independent named entity recognition. *Proceedings of CoNLL-2002*. Taipei, Taiwan: 2002. 155-158.

Erik F. TJONG KIM SANG and Fien DE MEULDER. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. *Proceedings of CoNLL-2003*. Edmonton, Canada. 142-147.

Dekai WU, Grace NGAI and Marine CARPUAT. 2003. A Stacked, Voted, Stacked Model for Named Entity Recognition. *Proceedings of CoNLL-2003*. Edmonton, Canada: 2003. 200-203.