RC22996 (C0311-005) November 14, 2003
Computer Science

IBM Research Report

Chinese Named Entity Recognition Based on the Robust Risk Minimization Classifier

Honglei Guo, Jianmin Jiang, Gang Hu
IBM Research Division
China Research Laboratory
HaoHai Building, No. 7, 5th Street
ShangDi, Beijing 100085
China

Tong Zhang
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

Chinese Named Entity Recognition Based on the Robust Risk Minimization Classifier

Honglei Guo Jianmin Jiang Gang Hu
IBM China Research Laboratory,
HaoHai Building, No.7, 5th Street,
ShangDi, Beijing 100085, PRC
email:{guohl,jiangjm,hugang}
@cn.ibm.com

Tong Zhang
IBM T.J. Watson Research Center,
Yorktown Heights, NY, 10598, USA
email:tzhang@watson.ibm.com

Abstract

This paper presents a Chinese named entity recognition system that employs the Robust Risk Minimization (RRM) Classification method and incorporates the advantages of character-based and word-based models. From experiments on a large-scale corpus, we show that significant performance enhancements can be obtained by integrating various linguistic information (such as Chinese word segmentation, semantic types, part of speech, and named entity triggers) into a basic Chinese character based model. A novel feature weighting mechanism is also employed to obtain more useful cues from most important linguistic features. Moreover, to overcome the limitation of computational resources in building a high-quality named entity recognition system from a large-scale corpus, informative samples are selected by an active learning approach.

1 Introduction

Named entities are phrases that contain names of persons, organizations, locations etc. Named entity (NE) recognition is an important task in many natural language processing applications, such as information extraction, machine translation, etc. This task has its origin from the Message Understanding Conference (MUC) in the 1990s, and has received more and more attention in recent years. There have been a number of conferences aimed at evaluating named entity recognition systems, for example, MUC6, MUC7, CONLL2002 and CONLL2003, and the on-going ACE (automatic content extraction) evaluations (see http://www.itl.nist.gov/iad/894.01/tests/ace/).

Recent research on English named entity recognition has focused on the machine learning approach (Sang and Meulder, 2003). Algorithms which has been applied to this task include Maximum Entropy (Borthwick, 1999; Klein et al., 2003), Hidden Markov Model (Bikel et al., 1999; Klein et al., 2003), AdaBoost (Carreras et al., 2003), Memory-based learning (Meulder and Daelemans, 2003), Support Vector Machine (Isozaki and Kazawa, 2002), etc. A successful named entity recognition system may also involve sophisticated components that are hand-crafted. An example is the LTG (Mikheev et al., 1998) system which achieved the best performance in MUC-7.

What we have learned from the extensive evaluations on named entity recognition systems in recent years (such as in CoNLL and ACE) is that best statistical systems are typically achieved by using a linear classification algorithm (such as maximum-entropy or RRM employed in this paper), together with a vast amount of carefully designed linguistic features. Here, using a relatively "simple" linear classification method is important since it is difficult for more complicated learning models to effectively utilize sophisticated linguistic features. We know that successful integration of useful linguistic features is essential for a good statistical named entity system. In fact, this claim seems to be true for many other linguistic processing problems as well. From this point of view, a crucial aspect of building a high-quality statistical NLP system is how to encode useful linguistic information so that the underlying learning algorithm (say, a linear classifier) can effectively utilize. This modeling aspect is task dependent, and essential for achieving good performance. The current paper focuses on various issues we have encountered and investigated in the process of developing a high-quality Chinese named entity system.

Most existing approaches for Chinese named entity recognition use hand-crafted rules with word (or character) frequency statistics. It is only recently that machine learning based Chinese named entity recognition systems have appeared. Similar to English named entity recognition, a number of algorithms have been investigated, including Hidden Markov Model (Yu et al., 1998; Jing

et al., 2003), Maximum Entropy, class-based language model (Sun et al., 2002; Jing et al., 2003), RRM type methods (Jiang et al., 2003; Jing et al., 2003). In (Jing et al., 2003), a few methods were compared and the effect of their combination was examined.

We can argue that Chinese named entity recognition is much more difficult than English named entity recognition since Chinese is a language with more flexible linguistic structure. First, in Chinese, there is no space to mark word boundaries and no standard definition of words. No external features such as capitalization are available to help recognize Chinese named entities. It is therefore difficult to determine the boundary of Chinese named entity. Secondly, Chinese characters used in named entities are also used in constructing common words, and some often serve as single-character words in sentences. For example, the Chinese character "张", normally refers to the surname "zhang", can also refer to concepts such as "open" or "sheet". It is thus difficult to recognize named entities merely by looking at Chinese characters used.

In this paper, we present a RRM based Chinese named entity recognition system that integrates the advantages of character-based model and word-based model. A similar system has been applied to the related text-chunking problem with state of the art performance (Zhang et al., 2002). The algorithm has also been successfully used in named entity recognition. For example, in CoNLL-2003 shared task, the top system, both for the English task and for the German task, utilized RRM (Florian et al., 2003). The method was also used by top-performing systems in the recent ACE named entity evaluations.

Although we employ an algorithm that have been successfully applied to related problems, there are still many issues and challenges in creating a high-quality Chinese named entity system. In spite of our focus on Chinese, we believe that some of our observations can be potentially useful to other languages including English.

In order to capture the internal and external features of Chinese named entities, we use Chinese characters as the basic token units, and integrate Chinese word segmentation information, semantic feature, part of speech, named entity triggers etc. In order to overcome the limitation of computational resources (e.g. memory consummation, annotated corpus) in building a high-quality named entity recognition system from a large-scale corpus, we select more informative samples based on uncertainty sampling. Meanwhile, a feature weighting mechanism is used to guide the learning algorithm to focus on important linguistic features.

This paper is organized as follows. Section 2 gives a brief overview of the underlying algorithm. Section 3 describes the training and test data. Section 4 presents a Chinese named entity model which integrates the advantages of character-based and word-based model, and presents experimental results on a large-scale corpus. Section 5 describes a feature weighting mechanism for Chinese named entity recognition model. Section 6 discusses a procedure of selecting informative training samples to overcome computational limitations. Finally the conclusion is given in Section 7.

2 Robust Risk Minimization Classifier

We can view the named entity recognition task as a sequential classification problem. If we use w_i ($i = 0, 1, ..., n$) to denote the sequence of tokenized text, which is the input to the system, then every token w_i should be assigned a class-label t_i. For named entity recognition, the class-label sequence $\{t_i\}$ encodes the entity information. There are various encoding schemes. In this paper, we adopt the so-called B-I-O encoding, which we shall explain later using an example.

Our Chinese named entity recognition system employs the Robust Risk Minimization (RRM) classification method. The class label value t_i associated with each token w_i is predicted by estimating the conditional probability $P(t_i = c|x_i)$ for every possible class-label value c, where x_i is a feature vector associated with token w_i.

We assume that $P(t_i = c|x_i) = P(t_i = c|w_i, \{t_j\}_{j\leq i})$. The feature vector x_i can depend on previously predicted class labels $\{t_j\}_{j\leq i}$, but the dependency is typically assumed to be local. In the RRM method, the above conditional probability model has the following parametric form:

$$P(t_i = c|x_i, t_{i-l}, ..., t_{i-1}) = T(w_c^T x_i + b_c),$$

where $T(y) = \min(1, \max(0, y))$ is the truncation of y into the interval [0, 1]. w_c is a linear weight vector and b_c is a constant. Parameters w_c and b_c can be estimated from the training data. Given training data (x_i, t_i) for $i = 1, ..., n$, the model is estimated by solving the following optimization problem for each c (Zhang et al., 2002):

$$\inf_{w,b} \frac{1}{n} \sum_{i=1}^{n} f(w_c^T x_i + b_c, y_c^i),$$

where $y_c^i = 1$ when $t_i = c$ and $y_c^i = -1$ otherwise. The function f is defined as:

$$f(p, y) = \begin{cases} -2py & py < 1 \\ \frac{1}{2}(py - 1)^2 & py \in [-1, 1] \\ 0 & py > 1 \end{cases}$$

The generalized Winnow method in (Zhang et al., 2002) describes such a method. However in this paper, we use a different procedure that is simpler and more efficient. Our experience suggests that the difference is not important as far as accuracy is concerned.

Given the above conditional probability model, the best possible sequence of t_i's can be estimated by dynamic programming in the decoding stage (Zhang et al., 2002). As we have mentioned earlier, the Robust Risk Minimization classification method has already been successfully applied in English text chunking and named entity recognition tasks.

3 Data

IBM China Research Laboratory has created a large-scale annotated Chinese corpus (about 100M Chinese characters). All of the data are news articles selected from several Chinese newspapers (e.g. Beijing Youth Daily, Xinmin Evening News, Yangcheng Evening News etc.) in 2001 and 2002. They cover a variety of domains, such as economics, sports, entertainment, etc. All of the named entities in the corpus are annotated manually (shown in Table 1).

Type	Number of unique NEs	Occurrences of NEs
PER	126,045	874,954
LOC	58,499	1,007,369
ORG	149,379	1,010,620
MISC	47,672	157,310
Abbr.	8,633	478,216
Total	390,230	3,528,469

Table 1: Named entities (NEs) in the annotated Chinese corpus

All training data used in our experiments are selected from the annotated Chinese corpus. The size of the training data set is 9,347,578 Chinese characters. The size of the test set is 1,339,292 Chinese characters (shown in Table 2).

Type	NEs in the training data set	NEs in the test data set
PER	77,890	11,991
LOC	90,587	12,353
ORG	80,446	9,820
MISC	22,571	1,820
Total	271,494	35,192

Table 2: NEs in the training set and test set

4 Selection of Linguistic Features in Chinese Named Entity Recognition

It is difficult to achieve language independence in a very high quality named entity recognition system because different languages usually require different features. Feature design and integration is very important in the overall system design. In this section, we present a Chinese named entity recognition model by integrating the advantages of character-based and word-based models.

In this paper, we focus on recognizing four types of named entities: person (PER), location (LOC), organization (ORG), and miscellaneous named entity (MISC) that do not belong to the previous three groups (e.g. products, brands, conferences etc.). The following sentence shows an example of Chinese named entities recognized.

〈LOC〉中国〈/LOC〉外长〈PER〉唐家璇〈/PER〉与
〈LOC〉美国〈/LOC〉国务卿〈PER〉鲍威尔〈/PER〉
在〈LOC〉钓鱼台国宾馆〈/LOC〉会谈。

As mentioned earlier, we adopt the B-I-O scheme to encode the entity-information into a sequence of class-labels, so that a single class label is assigned to every token. Each token unit is tagged as either the first token unit in a named entity of type X (B-X tag), or a non-initial token unit in a named entity of type X (I-X tag), or a token unit outside of any named entities (O tag). For example, the location "钓鱼台国宾馆" is encoded as "钓鱼台(B-LOC) 国宾馆(I-LOC)".

4.1 Baseline Performance for Character-based Model and Word-based Model

In our baseline experiments of a character-based Chinese named entity recognition model and a word-based model, only some basic linguistic features are employed. These features include the basic token units (Chinese characters in the character-based model or Chinese words in the word-based model), named entity triggers and dictionaries. The baseline performances for the character-based model and the word-based model are shown in Table 3.

Type	F(%) (Word-based model)	F(%) (Character-based model)
PER	84.80	83.71
LOC	85.87	80.09
ORG	75.50	66.81
MISC	60.14	53.14
Total	80.66	75.53

Table 3: Baseline performance for character-based model and word-based model

4.2 Integrating the Advantages of Character-based Model and Word-based Model

Since both Chinese characters and Chinese words can be used as basic token units in Chinese named entity recognition, some researchers have already started to investigate the behavior of the two approaches (Jing et

al., 2003). However, previous studies didn't incorporate them into one system.

We argue that both have their own advantages. Therefore by integrating them into a single system, we can benefit from both character-based and word-based models. For example, Chinese word segmentation information, which are often useful, are not present in character-based models. On the other hand, although Chinese word information are used in word-based models, in order to achieve high performance, the boundary of the named entities should be aligned with Chinese word units. This is not always achievable.

Therefore for Chinese named entity recognition, both character-based models and word-based models have their advantages and disadvantages. Character-based models can avoid problems caused by Chinese word segmentation errors, but cannot effectively capture rich information contained in Chinese words, which are generally regarded as basic linguistic units in the Chinese language. Although word-based models implicitly use larger views to capture useful cues on named entity, they are heavily affected by Chinese word segmentation errors. In addition, since there are more words than characters, word-based models are more severely affected by the data sparseness problem. That is, we don't have sufficient data to obtain accurate statistical estimates.

Our method of integrating the character-based and word-based approaches is by creating linguistic features that incorporate both character-based features and word-based features. In order to achieve this, we use Chinese characters (not Chinese words) as the basic token units, and then map word-based features that are associated with each word into corresponding features of those characters that are contained in the word.

In general, we may regard this approach as information integration from linguistic views at different abstraction levels. Although we have so far only studied this idea on the Chinese named entity recognition task, we believe that the general concept can be potentially useful for other NLP problems (including English named entity recognition) as well. For example, if we take English words as the counterpart of Chinese characters, then we may consider English chunking segments as the counterpart of Chinese words. Our study essentially suggests the possibility of improving English named entity recognition by using a deeper syntactic structure such as chunking information. Further alone this line, we may speculate that it is potentially beneficial to incorporate even deeper syntactic parse-tree information, which gives even higher level views of a sentence. In fact, it is not difficult to find errors from the current state of the art named entity recognition systems that can only be resolved using deeper sentence level information. However, to our knowledge, the effectiveness of such potentially useful information has not been very carefully investigated. Although the current paper is only a small step in this direction, we hope that the general concept presented here can stimulate future research.

The performance difference between character-based and word-based models in Table 3 also imply that linguistic information beyond the basic linguistic features can significantly affect the quality of Chinese named entity recognition. Therefore we integrate a diverse set of local linguistic features, including word segmentation information, Chinese word patterns, complex lexical linguistic features, external named entity hints, aligned at the character level. In our system, local linguistic features of a token unit are derived from the sentence containing this token unit. All special linguistic patterns (i.e. date, time, numeral expression) are encoded into pattern-specific class labels aligned with the tokens. The features are listed below.

1. The basic token view features: Chinese characters in a $-2 \sim 2$ window surrounding the current focused basic token unit.

2. Chinese word segmentation units view window:

   (a) Chinese words in a larger 6-Chinese-word-window (i.e. $i = -3, ..., 3$) surrounding the current focused basic token unit;

   (b) Chinese words where the 5-Chinese-character-window (i.e. $i = -2, ..., 2$) is anchored.

3. Lexical linguistic features: Part of speech features and semantic features.

4. Chinese pattern features where the 5-Chinese-character-window (i.e. $i = -2, ..., 2$) is anchored, including types such as date, time, numeral expression, English word, etc.

5. The relative position of the current Chinese character in the current Chinese word, which indicates whether the current character is the initial character or not in the current Chinese word.

6. The previous two predicated tags.

7. The conjunction of the previous tag and the current token unit.

In addition, we also apply external named entity hints, including

1. 221 surnames, 517 location suffixes, 8,065 organization suffixes, 5,321 titles, 649 Chinese characters which are frequently used in Chinese names, 461 Chinese characters which are frequently used in translation names. These external hints are used to

determine whether a token unit may trigger or terminate a particular named entity class. For example, "教授"(i.e. "professor") may trigger the named entity class person.

2. Gazetteer features: gazetteer information for token units, in the form of a list of 6,160 locations and 1,656 organizations. These external gazetteers are used to determine potential classes for each token.

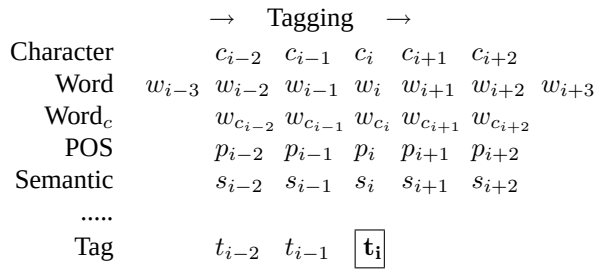

Figure 1: Linguistic feature window in Chinese named entity recognition

Figure 1 illustrates various features and views at the current character c_i, where w_{c_i} denotes the word where the Chinese character c_i is anchored. In the viewing window (we choose a window size of 2, shown in Figure 1) at the current character, each token unit around c_i is codified with a set of primitive features, together with its relative position to c_i.

Given the input vector consisted of features constructed as above, the RRM method is then applied to train linear weight vectors, one for each possible class-label. In the decoding stage, the class with the maximum confidence is then selected for each token unit. Dynamic programming can also be applied to find a sequence of class-labels with high confidence.

4.3 Experimental Results

In our evaluation, only named entities with correct boundaries and correct class labels are considered as the correct recognition. We use the standard F_β measure defined as follows.

$$F_\beta = \frac{(\beta^2+1) \times \text{Precision} \times \text{Recall}}{\beta^2 \times \text{Precision} + \text{Recall}},$$

where,

$$\text{Precision} = \frac{\text{Number of correct recognized named entities}}{\text{Number of recognized named entities}},$$

and

$$\text{Recall} = \frac{\text{Number of correct recognized named entities}}{\text{Number of correct named entities}}.$$

In this paper, we use $F = F_{\beta=1}$.

The experimental results (see Table 4) show that our integration model gives a very significant performance enhancement on Chinese named entity recognition. Chinese word segmentation information and its surrounding context information have a significant impact on performance enhancement, which can give a 5% improvement in F-measure points (shown in Table 5). Part of speech, semantic information, named entity hints also give a slightly performance enhancement (shown in Table 6).

Type	Precision(%)	Recall(%)	F(%)
PER	91.36	88.42	89.87
LOC	89.48	88.88	89.18
ORG	81.47	77.08	79.22
MISC	74.08	56.76	64.27
Total	86.79	82.49	84.59

Table 4: Integration model with complex linguistic features (learning from right to left)

Type	Precision(%)	Recall(%)	F(%)
PER	89.11	84.73	86.86
LOC	85.93	84.64	85.28
ORG	73.66	66.76	70.04
MISC	72.39	45.18	55.64
Total	82.92	76.12	79.38

Table 5: Integration model without using word segmentation information (learning from right to left)

Type	F(%) (with all features)	F(%) (without POS)	F(%) (without SEM)	F(%) (without NE hints)
PER	89.87	89.12	89.65	89.55
LOC	89.18	88.64	89.16	88.99
ORG	79.22	78.72	78.92	78.82
MISC	64.27	63.80	64.57	63.87
Total	84.59	84.00	84.46	84.28

Table 6: Performance without using POS, semantic feature (SEM) and NE hints

Since Chinese word is the basic linguistic representation unit in Chinese, we have also built a word-based model with complex linguistic features. Its performance (shown in Table 7) is lower than the full integration model. Obviously, the disadvantages of the word-based model described before have made a negative impact on the performance.

Since a sentence can be processed from left to right or from right to left during training and decoding, we also

Type	Precision(%)	Recall(%)	F(%)
PER	89.51	86.86	88.17
LOC	88.89	87.92	88.40
ORG	80.48	75.07	77.68
MISC	74.65	54.02	62.68
Total	85.85	80.89	83.30

Table 7: Word-based model with complex linguistic features (learning from right to left)

compared the difference in system performance when processing-order is taken in these two directions. The overall performance when processing is from left to right (shown in Table 8) is lower than that of from right to left (shown in Table 4). For recognizing people, the performance of processing from left to right is in fact better. However, for recognizing organizations, locations and miscellaneous named entities, the performance of processing from right to left is better. One reason is that most of the triggers for person names in Chinese are the initial characters in person names, while most of the triggers for locations, organizations, miscellaneous named entities are ending characters.

Type	Precision(%)	Recall(%)	F(%)
PER	92.16	88.10	90.08
LOC	90.33	87.10	88.68
ORG	82.80	74.16	78.24
MISC	69.96	56.20	62.23
Total	87.29	81.00	84.03

Table 8: Integration model with complex linguistic features (learning from left to right)

We have the following observations from the above experimental results:

1. By integrating character-based and word-based models, we can achieve better performance in Chinese named entity recognition than either character-based or word-based models.

2. Chinese word segmentation information and its surrounding context have a significant impact on performance.

3. Better Chinese named entity recognition model can be obtained by using more complex linguistic features.

4. The performances for person names, locations and organizations are better than that of miscellaneous named entities. We believe that the main reason for the poor performance of miscellaneous named entity is that there are less common indicative features among various miscellaneous named entities which we do not distinguish. In fact, this entity type can be further divided into thirteen categories, including products, conferences, events, brands, etc. In addition, there are a relatively small number of positive training samples for miscellaneous named entities.

5 Feature Weighting in Chinese Named Entity Recognition

Since good features can significantly enhance system performance, we may repeat features that are important two or more times in constructing the input vector to the classifier.

An equivalent method is to employ a feature weighting mechanism that assigns each feature a weight which indicates its importance in named entity recognition. For example, if the weight of Chinese word segmentation information is 3 (i.e. weight=3), it means that the value of word segmentation feature in the input vector will be added three times. This allows us to bias the learning algorithm so that the more relevant feature components have higher influence while less relevant feature components have less influence on the system decision. If the features are not weighted (i.e. weight=1), then the effect is to let the algorithm find the relevant ones by itself. By weighting those more important features according to Chinese linguistic knowledge (such as Chinese word segmentation information), performance can be further enhanced (shown in Table 9). The experiments also indicate that the performance may deteriorate when the weight for a feature is more than a certain threshold.

Type	Weight=0 F(%)	weight=3 F(%)	Weight=4 F(%)
PER	89.22	89.87	88.75
LOC	88.61	89.18	88.04
ORG	77.91	79.22	76.17
MISC	62.53	64.27	63.00
Total	83.69	84.59	82.91

Table 9: F-measure for weighting word segmentation information

6 Select Informative Training Samples by the Active Learning Approach

Since only limited external resources are available for recognizing Chinese named entities, we need to use a large number of training data, and let the learning algorithm find relevant classification patterns. In practice, the amount of annotated data is always a bottleneck for supervised learning methods. Typically a higher performance system requires more features and a larger num-

ber of training data. However, this requires larger system memory and a more efficient training method, which may not be available. Within the limitation of available computational resources, it is thus necessary for us to either limit the number of features or to select more informative data which can be efficiently handled by the training algorithm.

In order to overcome the existing computational limitation, while still being able to capture rich linguistic patterns, we build our Chinese named entity recognition model incrementally using a variant of uncertainty-sampling (Lewis and Catlett, 1994). The main steps are described as follows.

1. Build an initial recognition model (shown in Table 10) by training with an initial data set (about 1M Chinese characters) which is randomly selected from a large-scale candidate data set (about 9M Chinese characters).

2. Refine the training set by adding more informative samples and removing those redundant samples. In this refinement phase, all of the data are annotated by the current recognition model (e.g. the initial model built in Step 1). Each annotation has a confidence score associated with the prediction. In general, an annotation with lower confidence score usually indicates a wrong prediction. Therefore, we add those samples with lower confidence scores into the training set. Meanwhile, in order to keep a reasonable size of the training set, those old training samples with higher confidence scores are removed from the current training set. In each retraining phase, all of the samples are sorted by the confidence score. The top 1000 new samples with lowest confidence scores are added into the current training set. The top 500 old training samples with highest confidence scores are removed from the current training set.

3. Retrain a new Chinese named entity recognition model with the newly refined training set.

4. Repeat Step 2 and Step 3, until the performance doesn't improve any more.

Type	Precision(%)	Recall(%)	F(%)
PER	86.55	84.44	85.48
LOC	86.70	83.70	85.18
ORG	78.97	67.74	72.92
MISC	60.39	51.08	55.34
Total	82.42	76.56	79.39

Table 10: Initial model by training with an initial train data set

Using this incremental training method, we created a refined high-quality training set (about 1,051,925 Chinese characters) from a larger candidate data set (about 9M Chinese characters). Table 11 shows the number of entities contained in each data set. Since the refined training data are consisted of highly informative samples, we are able to obtain a better model (shown in Table 4).

Type	the original candidate data set	the refined training data set	the initial training data set
PER	77,890	18,898	12,575
LOC	90,587	24,862	12,869
ORG	80,446	22,173	12,197
MISC	22,571	8,067	4,691
Total	271,494	74,000	42,332

Table 11: NEs in the original candidate data set, the refined training data set and the initial training data set

Our experience with the incremental sample selection strategy suggests the following:

1. In learning named entity recognition models, annotated results with lower confidence scores are more useful than those samples with higher confidence scores. This is consistent with other studies on active learning. It also requires the underlying classifier to produce confidence estimates, which the RRM method can (by estimating the conditional probability).

2. In order to obtain a high-quality Chinese named entity recognition model, it is only necessary to keep the informative samples.

3. Informative sample-selection is an effective method for overcoming the potential limitation of computational resources, and can alleviate the problem of obtaining a large amount of annotated data.

7 Conclusion

We presented a Chinese named entity recognition system which incorporates both character-based and word-based models. From experiments performed on a large-scale corpus, our integrated Chinese named entity recognition model achieves appreciably better performance than either character-based models or word-based models alone. More generally, our approach can be regarded as a special case of utilizing information from different levels of linguistic abstractions. We believe that this idea can benefit other NLP tasks, although its effectiveness remains to be carefully investigated.

In our system, the following features have significant impacts on its performance: local Chinese characters,

Chinese word segmentation information and its surrounding context, part of speech. A feature weighting mechanism is also employed to guide the learning algorithm to focus on important linguistic feature components.

In practice, the limitation of available computational resources may become an obstacle. In order to build a high quality named entity recognition model, it is necessary to select the most informative training data. We described an incremental data selection mechanism and showed that this method can significantly improve performance.

References

Daniel M. Bikel, Richard L. Schwartz, and Ralph M. Weischedel. 1999. An algorithm that learns what's in a name. *Machine Learning*, 34(1-3):211–231.

Andrew Borthwick. 1999. *A Maximum Entropy Approach to Named Entity Recognition*. Ph.D. thesis, New York University.

Xavier Carreras, Lluís Màrquez, and Lluís Padró. 2003. A simple named entity extractor using adaboost. In *Proceedings of CoNLL-2003*, pages 152–155.

Radu Florian, Abe Ittycheriah, Hongyan Jing, and Tong Zhang. 2003. Named entity recogintion through classifier combination. In *Proceedings CoNLL-2003*, pages 168–171.

Hideki Isozaki and Hideto Kazawa. 2002. Efficient support vector classifiers for named entity recognition. In *Proceedings of Coling-2002*.

Jianmin Jiang, Honglei Guo, Gang Hu, and Tong Zhang. 2003. Chinese named entity recognition by regularized winnow algorithm. In *Proceedings of 20th International Conference on Computer Processing of Oriental Languages*.

Hongyan Jing, Radu Florian, Xiaoqiang Luo, Tong Zhang, and Abraham Ittycheriah. 2003. Howtogetachinesename (entity) : Segmentation and combination issue s. In *EMNLP 2003*.

Dan Klein, Joseph Smarr, Huy Nguyen, and Christopher D. Manning. 2003. Named entity recognition with character-level models. In *Proceedings of CoNLL-2003*, pages 180–183.

D. Lewis and J. Catlett. 1994. Heterogeneous uncertainty sampling for supervised learning. In *Proceedings of the Eleventh International Conference on Machine Learning*, pages 148–156.

Fien De Meulder and Walter Daelemans. 2003. Memory-based named entity recognition using unannotated data. In *Proceedings of CoNLL-2003*, pages 208–211.

A. Mikheev, C. Grover, and M. Moens. 1998. Description of the LTG system used for MUC-7. In *Proceedings of the Seventh Message Understanding Conference (MUC-7)*.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language independent named entity recognition. In Walter Daelemans and Miles Osborne, editors, *Proceedings of CoNLL-2003*, pages 142–147.

Jian Sun, Jianfeng Gao, Lei Zhang, Ming Zhou, and Changning Huang. 2002. Chinese named entity identification using class-based language model. In *Proceedings of Coling-2002*.

S.H. Yu, S.H. Bai, and P. Wu. 1998. Description of the kent ridge digital labs system used for muc-7. In *Proceedings of the Seventh Message Understanding Conference (MUC-7)*.

Tong Zhang, Fred Damerau, and David E. Johnson. 2002. Text chunking based on a generalization of Winnow. *Journal of Machine Learning Research*, 2:615–637.