Empirical Study on the Performance Stability of Named Entity Recognition Model across Domains

Hong Lei Guo Li Zhang and Zhong Su
IBM China Research Laboratory
Building 19, Zhongguancun Software Park
8 Dongbeiwang WestRoad, Haidian District, Beijing, 100094, P.R.C.
{guohl, lizhang , suzhong }@cn.ibm.com

Abstract

When a machine learning-based named entity recognition system is employed in a new domain, its performance usually degrades. In this paper, we provide an empirical study on the impact of training data size and domain information on the performance stability of named entity recognition models. We present an informative sample selection method for building high quality and stable named entity recognition models across domains. Experimental results show that the performance of the named entity recognition model is enhanced significantly after being trained with these informative samples.

1 Introduction

Named entities (NE) are phrases that contain names of persons, organizations, locations, etc. Named entity recognition (NER) is an important task in many natural language processing applications, such as information extraction and machine translation. There have been a number of conferences aimed at evaluating NER systems, for example, MUC6, MUC7, CoNLL2002 and CoNLL2003, and ACE (automatic content extraction) evaluations.

Machine learning approaches are becoming more attractive for NER in recent years since they are trainable and adaptable. Recent research on English NER has focused on the machine learning approach (Sang and Meulder, 2003). The relevant algorithms include Maximum Entropy (Borthwick, 1999; Klein et al., 2003), Hidden Markov Model (HMM) (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), Robust Risk Minimization (RRM) Classification method (Florian et al., 2003), etc.

For Chinese NER, most of the existing approaches use hand-crafted rules with word (or character) frequency statistics. Some machine learning algorithms also have been investigated in Chinese NER, including HMM (Yu et al., 1998; Jing et al., 2003), class-based language model (Gao et al., 2005; Wu et al., 2005), RRM (Guo et al., 2005; Jing et al., 2003), etc.

However, when a machine learning-based NER system is directly employed in a new domain, its performance usually degrades. In order to avoid the performance degrading, the NER model is often retrained with domain-specific annotated corpus. This retraining process usually needs more efforts and costs. In order to enhance the performance stability of NER models with less efforts, some issues have to be considered in practice. For example, how much training data is enough for building a stable and applicable NER model? How does the domain information and training data size impact the NER performance?

This paper provides an empirical study on the impact of training data size and domain information on NER performance. Some useful observations are obtained from the experimental results on a large-scale annotated corpus. Experimental results show that it is difficult to significantly enhance the performance when the training data size is above a certain threshold. The threshold of the training data size varies with domains. The performance stability of each NE type recognition also varies with domains. Corpus statistical data show that NE types have different distribution across domains. Based on the empirical investigations, we present an informative sample selection method

Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006), pages 509–516,
Sydney, July 2006.

for building high quality and stable NER models. Experimental results show that the performance of the NER model is enhanced significantly across domains after being trained with these informative samples. In spite of our focus on Chinese, we believe that some of our observations can be potentially useful to other languages including English.

This paper is organized as follows. Section 2 describes a Chinese NER system using multi-level linguistic features. Section 3 discusses the impact of domain information and training data size on the NER performance. Section 4 presents an informative sample selection method to enhance the performance of the NER model across domains. Finally the conclusion is given in Section 5.

2 Chinese NER Based on Multilevel Linguistic Features

In this paper, we focus on recognizing four types of NEs: Persons (PER), Locations (LOC), Organizations (ORG) and miscellaneous named entities (MISC) which do not belong to the previous three groups (e.g. products, conferences, events, brands, etc.). All the NER models in the following experiments are trained with a Chinese NER system. In this section, we simply describe this Chinese NER system. The Robust Risk Minimization (RRM) Classification method and multi-level linguistic features are used in this system (Guo et al., 2005).

2.1 Robust Risk Minimization Classifier

We can view the NER task as a sequential classification problem. If tok_i ($i = 0, 1, ..., n$) denotes the sequence of tokenized text which is the input to the system, then every token tok_i should be assigned a class-label t_i.

The class label value t_i associated with each token tok_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 tok_i.

We assume that $P(t_i = c|x_i) = P(t_i = c|tok_i, \{t_j\}_{j\leq i})$. 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:

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

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).

2.2 Multilevel Linguistic Features

This Chinese NER system uses Chinese characters (not Chinese words) as the basic token units, and then maps word-based features that are associated with each word into corresponding features of those characters that are contained in the word. This approach can effectively incorporate both character-based features and word-based features. In general, we may regard this approach as information integration from linguistic views at different abstraction levels.

We integrate a diverse set of local linguistic features, including word segmentation information, Chinese word patterns, complex lexical linguistic features (e.g. part of speech and semantic features), aligned at the character level. In additional, we also use external NE hints and gazetteers, including surnames, location suffixes, organization suffixes, titles, high-frequency Chinese characters in Chinese names and translation names, and lists of locations and organizations. In this 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.

3 Impact of Training Data Size And Domain Information on the NER Performance

It is very important to keep the performance stability of NER models across domains in practice.

However, the performance usually becomes unstable when NER models are applied in different domains. We focus on the impact of the training data size and domain information on the NER performance in this section.

3.1 Data

We built a large-scale high-quality Chinese NE annotated corpus. The corpus size is 114.25M Chinese characters. All the data are news articles selected from several Chinese newspapers in 2001 and 2002. All the NEs in the corpus are manually tagged. Documents in the corpus are also manually classified into eight domain categories, including politics, sports, science, economics, entertainment, life, society and others. Cross-validation is employed to ensure the tagging quality.

All the training data and test data in the experiments are selected from this Chinese annotated corpus. The general training data are randomly selected from the corpus without distinguishing their domain categories. All the domain-specific training data are selected from the corpus according to their domain categories. One general test data set and seven domain-specific test data sets are used in our experiments (see Table 1). The size of the general test data set is 1.34M Chinese characters. Seven domain-specific test sets are extracted from the general test data set according to the document domain categories.

Domain	NE distribution in the domain-oriented test data set					Test set Size
	PER	ORG	LOC	MISC	Total	
General	11,991	9,820	12,353	1,820	35,984	1.34M
Politics	2,470	1,528	2,540	480	7,018	0.2M
Economics	1,098	2,971	2,362	493	6,924	0.26M
Sports	1,802	1,323	1,246	478	4,849	0.10M
Entertainment	2,458	526	738	542	4,264	0.10M
Society	916	418	823	349	2,506	0.08M
Life	2,331	1,690	3,634	763	8,418	0.39M
Science	1,802	1,323	1,246	478	4,849	0.10M

Table 1: NE distribution in the general and domain-specific test data sets

In our evaluation, only NEs with correct boundaries and correct class labels are considered as the correct recognition. We use the standard P (i.e. Precision), R (i.e. Recall), and F-measure (defined as 2PR/(P+R)) to measure the performance of NER models.

3.2 Impact of Training Data Size on the NER Performance across Domains

The amount of annotated data is always a bottleneck for supervised learning methods in practice.

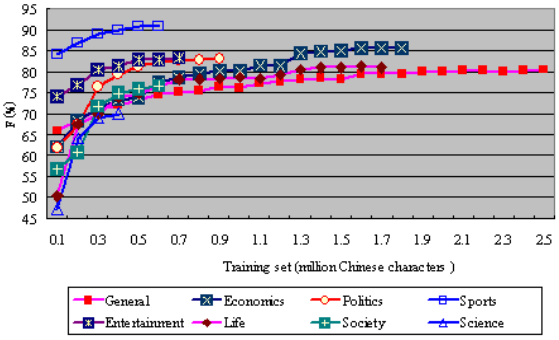

Figure 1: Performance curves of the general and specific domain NER models

Thus, we evaluate the impact of training data size on the NER performance across domains.

In this baseline experiment, an initial general NER model is trained with 0.1M general data at first. Then the NER model is incrementally retrained by adding 0.1M new general training data each time till the performance isn't enhanced significantly. The NER performance curve (labelled with the tag "General") in the whole retraining process is shown in Figure 1. Experimental results show that the performance of the general NER model is significantly enhanced in the first several retraining cycles since more training data are used. However, when the general training data set size is more than 2.4M, the performance enhancement is very slight.

In order to analyze how the training data size impacting the performance of NER models in specific domains, seven domain-specific NER models are built using the similar retraining process. Each domain-specific NER model is also trained with 0.1M domain-specific data at first. Then, each initial domain-specific NER model is incrementally retrained by adding 0.1M new domain-specific data each time.

NER Model	F(%)	Size threshold (M)	NE distribution in the training set				
			PER	ORG	LOC	MISC	Total
General	80.38	2.4	24,960	27,231	21,098	7,439	80,728
Politics	83.09	0.9	11,388	6,618	14,350	1,974	34,330
Economics	85.46	1.7	7,197	21,113	15,582	3,466	47,358
Sports	90.78	0.6	11,647	8,105	7,468	3,070	30,290
Entertainment	83.31	0.6	12,954	2,823	4,665	3,518	32,860
Society	76.55	0.6	7,099	3,279	6,946	1,909	19,233
Life	81.06	1.7	10,502	5,675	18,980	2,420	37,577
Science	70.02	0.4	1,625	3,010	2,083	902	7,620

Table 2: Performance of NER models, size threshold and NE distribution in the corresponding training data sets

The performance curves of these domain-specific NER models are also shown in Figure 1 (see the curves labelled with the domain tags). Although the initial performance of each domain-specific NER model varies with domains, the performance is also significantly enhanced in the first several retraining cycles. When the size of the domain-specific training data set is above a certain threshold, the performance enhancement is very slight as well.

The final performance of the trained NER models, and the corresponding training data sets are shown in Table 2.

From these NER performance curves, we obtain the following observations.

1. More training data are used, higher NER performance can be achieved. However, it is difficult to significantly enhance the performance when the training data size is above a certain threshold.

2. The threshold of the training data size and the final achieved performance vary with domains (see Table 2). For example, in entertainment domain, the threshold is 0.6M and the final F-measure achieves 83.31%. In economic domain, the threshold is 1.7M, and the corresponding F-measure is 85.46%.

3.3 The Performance Stability of Each NE Type Recognition across Domains

Statistic data on our large-scale annotated corpus (shown in Table 3) show that the distribution of NE types varies with domains. We define " NE density " to quantitatively measure the NE distribution in an annotated data set. NE density is defined as "the count of NE instances in one thousand Chinese characters". Higher NE density usually indicates that more NEs are contained in the data set. We may easily measure the distribution of each NE type across domains using NE density. In this annotated corpus, PER, LOC, and ORG have similar NE density while MISC has the smallest NE density. All the NE types also have different NE density in each domain. For example, the NE density of ORG and LOC is much higher than that of PER in economic domain. PER and LOC have higher NE density than ORG in politics domain. PER has the highest NE density among these NE types in both sports and entertainment domains. The unbalanced NE distribution across domains shows that news articles on different domains usually focus on different specific NE types. These NE distribution features imply that each NE type has different domain dependency feature. The performance stability of domain-focused NE type recognition becomes more important in domain-specific applications. For example, since economic news articles usually focus on ORG and LOC NEs, the high-quality LOC and ORG recognition models will be more valuable in economic domain. In addition, these distribution features also can be used to guide training and test data selection.

Domain	NE distribution in the specific domain					
	PER	LOC	ORG	MISC	ALL	Ratio (%)
Politics	167,989	180,193	105,936	30,830	484,948	16.43
Economics	117,459	200,261	352,323	76,320	746,363	25.29
Sports	129,137	73,435	98,618	33,304	334,494	11.33
Entertainment	154,193	50,408	40,444	52,460	297,505	10.08
Life	200,222	234,150	145,138	65,733	645,243	21.86
Society	63,793	53,724	43,657	21,162	182,336	6.18
Science	27,878	30,737	72,413	16,824	147,852	5.00
Others	31,723	40,730	26,666	13,926	113,045	3.83
All	892,394	863,638	885,195	310,559	2,951,786	–
Domain	NE density in the Chinese annotated corpus					Size (M)
	PER	LOC	ORG	MISC	ALL	
Politics	10.70	11.48	6.75	1.96	31.21	15.70
Economics	4.18	7.13	12.55	2.72	26.58	28.08
Sports	16.43	9.34	12.55	4.24	42.57	7.86
Entertainment	16.81	5.05	4.14	5.72	32.44	9.17
Life	5.64	6.59	4.09	1.85	18.17	35.52
Society	8.57	7.22	5.87	2.84	24.51	7.44
Science	4.30	4.74	11.17	2.60	22.82	6.48
Others	7.9	10.18	6.67	3.48	28.26	4.00
All	7.81	7.56	7.75	2.72	25.89	114.25

Table 3: NE distribution in the Chinese annotated corpus

In this experiment, the performance stability of NER models across domains is evaluated, especially the performance stability of each NE type recognition. The general NER model is trained with 2.4M general data. Seven domain-specific models are trained with the corresponding domain-specific training sets (see Table 2 in Section 3.2).

The performance stability of the general NER model is firstly evaluated on the general and domain-specific test data sets (see Table 1 in Section 3.1). The experimental results are shown in Table 4. The performance curves of the general model are shown in Figure 2, including the total F-measure curve of the NER model (labelled with the tag "All") and F-measure curves of each NE type recognition in the specific domains (labelled with the NE tags respectively).

The performance stability of the seven domain-specific NER models are also evaluated. Each domain-specific NER model is tested on the gen-

Domain	F(%) of general NER model				
	PER	LOC	ORG	MISC	ALL
General	86.69	85.55	73.59	56.00	80.38
Economic	85.11	88.22	75.91	49.53	80.50
Politics	86.26	87.00	71.31	61.50	81.90
Sports	91.87	89.03	81.67	67.41	86.10
Entertainment	84.24	85.85	68.65	60.96	79.31
Life	86.62	83.54	70.30	58.49	79.73
Society	84.53	76.16	68.89	41.14	74.50
Science	87.74	86.42	65.85	24.10	69.55

Table 4: Performance of the general NER model in specific domains

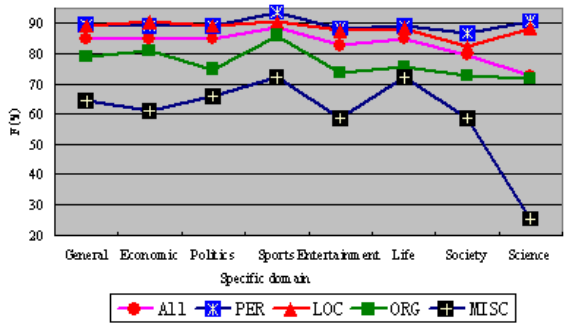

Figure 2: Performance curves of the general NER model in specific domains

NER Model	F(%) in specific domain							
	General	Economic	Politics	Sports	Entertainment	Life	Society	Science
General	80.38	80.50	81.90	86.10	79.31	79.73	74.50	69.55
Economic	75.30	85.46	74.32	72.89	68.46	76.23	65.75	68.97
Politics	73.37	66.39	83.09	76.37	71.51	74.83	67.31	53.76
Sports	71.23	62.56	68.99	90.78	73.48	71.18	64.82	53.85
Entertainment	70.82	61.52	72.04	75.34	83.31	71.80	69.10	52.50
Life	73.53	66.92	75.07	73.86	72.68	81.06	69.61	57.36
Society	70.29	62.55	72.70	70.69	72.24	74.10	76.55	53.42
Science	67.26	67.57	69.00	64.32	63.84	69.05	64.85	70.02

Table 5: Performance of NER models in specific domains

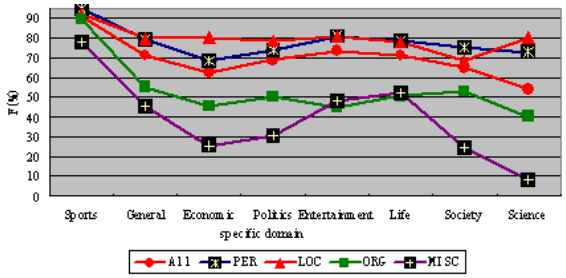

Figure 4: Performance curves of sports domain NER model in the other specific domains

eral test data and the other six different domain-specific test data sets. The experimental results are shown in Table 5. The performance curves of three domain-specific NER models are shown in Figure 3, Figure 4 and Figure 5 respectively.

From these experimental results, we have the following conclusions.

1. The performance stability of all the NER models is limited across domains. When a NER model is employed in a new domain, its performance usually decreases. Moreover, its performance is usually much lower than the performance of the corresponding domain-specific model.

2. The general NER model has better performance stability than the domain-specific NER model when they are applied in new domains (see Table 5). Domain-specific models usually could achieve a higher performance in its corresponding domain after being trained with a smaller amount of domain-specific annotated data (see Table 2 in Section 3.2). However, the performance stability of domain-specific NER model is poor across different domains. Thus, it is very popular to build a general NER model for the general applications in practice.

3. The performance of PER, LOC and ORG recognition is better than that of MISC recog-

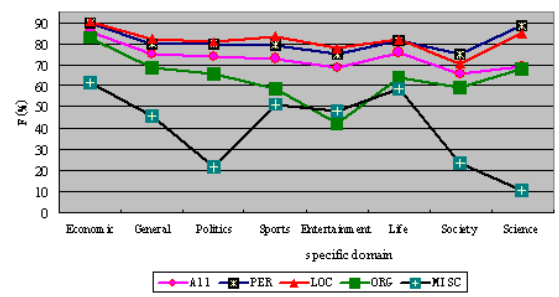

Figure 3: Performance curves of economic domain NER model in the other specific domains

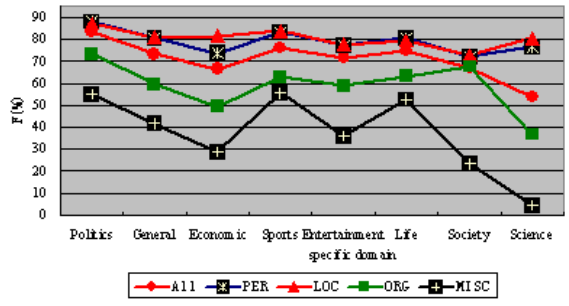

Figure 5: Performance curves of politics domain NER model in the other specific domains

nition in NER (see Figure 2 $\sim$ Figure 5). The main reason for the poor performance of MISC recognition is that there are less common indicative features among various MISC NEs which we do not distinguish. In addition, NE density of MISC is much less than that of PER, LOC, and ORG. There are a relatively small number of positive training samples for MISC recognition.

4. NE types have different domain dependency attribute. The performance stability of each NE type recognition varies with domains (see Figure 2 $\sim$ Figure 5). The performance of PER and LOC recognition are more stable across domains. Thus, few efforts are needed to adapt the existing high-quality general PER and LOC recognition models in domain-specific applications. Since ORG and MISC NEs usually contain more domain-specific semantic information, ORG and MISC are more domain-dependent than PER and LOC. Thus, more domain-specific features should be mined for ORG and MISC recognition.

4 Use Informative Training Samples to Enhance the Performance of NER Models across Domains

A higher performance system usually requires more features and a larger number of training data. This requires larger system memory and more efficient training method, which may not be available. Within the limitation of available training data and computational resources, it is necessary for us to either limit the number of features or select more informative data which can be efficiently handled by the training algorithm. Active learning method is usually employed in text classification (McCallum and Nigam et al., 1998). It is only recently employed in NER (Shen et al., 2004).

In order to enhance the performance and overcome the limitation of available training data and computational resources, we present an informative sample selection method using a variant of uncertainty-sampling (Lewis and Catlett, 1994). The main steps are described as follows.

1. Build an initial NER model (F-measure=76.24%) using an initial data set. The initial data set (about 1M Chinese characters) is randomly selected from the large-scale candidate data set (about 9M).

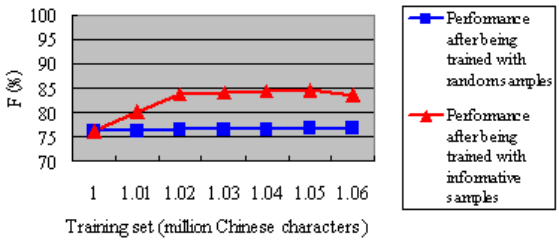

Figure 6: Performance curves of general NER models after being trained with informative samples and random samples respectively

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. The confidence score of the whole sample sentence is defined as the average of the confidence scores of all the annotations contained in the sentence. Thus, we add those sample sentences with lower confidence scores into the training set. Meanwhile, in order to keep a reasonable size of the training set, those old training sample sentences with higher confidence scores are removed from the current training set. In each retraining phase, all of the sample sentences are sorted by the confidence score. The top 1000 new sample sentences with lowest confidence scores are added into the current training set. The top 500 old training sample sentences with highest confidence scores are removed from the current training set.

3. Retrain a new Chinese NER model with the newly refined training set

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

We apply this informative sample selection method to incrementally build the general domain NER model. The size of the final informative training sample set is 1.05M Chinese characters. This informative training sample set has higher NE density than the random training data set (see Table 6).

We denote this general NER model trained with the informative sample set as "general informative model", and denote the general-domain model which is trained with 2.4M random general training data as "general random model". The performance curves of the general NER models after being trained with informative samples and random data respectively are shown in Figure 6. Experiment results (see Table 6) show that there is a significant enhancement in F-measure if using informative training samples. Compared with the random model, the informative model can increase F-measure by 4.21 percent points.

Type	Using informative sample set (1.05M)			Using random training set (2.4M)		
	F(%)	NEs	NE density	F(%)	NEs	NE density
PER	89.87	18,898	18.00	86.69	24,960	10.38
LOC	89.68	24,862	23.68	85.55	21,089	11.33
ORG	79.22	22,173	21.12	73.59	27,231	8.78
MISC	64.27	8,067	7.68	56.00	7,439	3.10
Total	84.59	74,000	70.48	80.38	80,728	33.58

Table 6: Performance of informative model and random model in the general domain

Domain	F(%) of general informative model				
	PER	LOC	ORG	MISC	ALL
Economic	89.26	90.66	81.24	61.14	84.63
Politics	89.36	89.37	74.76	65.95	84.70
Sports	93.65	90.66	86.00	72.05	88.71
Entertainment	88.38	87.54	73.88	58.32	82.74
Life	89.15	88.35	75.68	72.01	84.66
Society	86.61	82.15	72.99	58.55	79.49
Science	90.91	88.35	71.69	25.16	72.71

Table 7: Performance of the general informative model in specific domains

This informative model is also evaluated on the domain-specific test sets. Experimental results are shown in Table 7. We view the performance of the domain-specific NER model as the baseline performance in its corresponding domain (see Table 8), denoted as $F_{baseline}$. The performance of informative model in specific domains is very close to the corresponding $F_{baseline}$ (see Figure 7). We define the domain-specific average F-measure as the average of all the F-measure of the NER model in seven specific domains, denote as $\overline{F}$. The average of all the $F_{baseline}$ in specific domains is denoted as $\overline{F}_{baseline}$. The average F-measure of the informative model and the random model in specific domains is denoted as $\overline{F}_{informative}$ and $\overline{F}_{random}$ respectively. Compared with $\overline{F}_{baseline}$ ($\overline{F}$ =81.47%), the informative model increases $\overline{F}$ by 1.05 percent points. However, $\overline{F}$ decreases by 2.67 percent points if using the random model. Especially, the performance of the informative model is better than the corresponding baseline performance in politics, life, society and science domains. Moreover, the size of the informative sample set is much less than the life domain training set (1.7M).

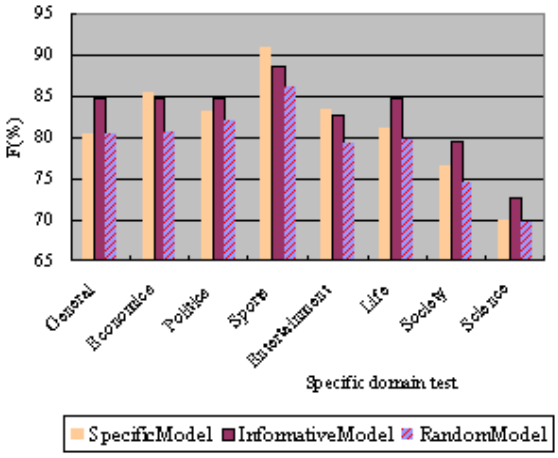

Figure 7: Performance comparison of informative model, random model, and the corresponding domain-specific models

NER model	F(%) in specific domains							$\overline{F}$
	Economic	Politics	Sports	Entertainment	Life	Society	Science	
domain-specific (baseline)	85.46	83.09	90.78	83.31	81.06	76.55	70.02	81.47
Informative	84.63	84.70	88.71	82.74	84.66	79.49	72.71	82.52
Random	80.50	81.90	86.10	79.31	79.73	74.50	69.55	78.80
NER model	$\delta(F)$ in specific domain $\delta(F) = (F - \overline{F})$							σ
	Economic	Politics	Sports	Entertainment	Life	Society	Science	
Informative	2.11	2.18	6.19	0.22	2.14	-3.03	-9.81	4.74
Random	1.7	3.1	7.3	0.51	0.93	-4.3	-9.25	4.94

Table 8: Performance comparison of informative model, random model and the corresponding domain-specific model in each specific domain

The informative model has much better performance than the random model in specific domains (see Table 8 and Figure 7). $\overline{F}_{informative}$ is 82.52% while $\overline{F}_{random}$ is 78.80%. The informative model can increase $\overline{F}$ by 3.72 percent points. The informative model is also more stable than the random model in specific domains (see Table 8). Standard deviation of F-measure for the informative model is 4.74 while that for the random model is 4.94.

Our experience with the incremental sample selection provides the following hints.

1. The performance of the NER model across domains can be significantly enhanced after being trained with informative samples. In

order to obtain a high-quality and stable NER model, it is only necessary to keep the informative samples. Informative sample selection can alleviate the problem of obtaining a large amount of annotated data. It is also an effective method for overcoming the potential limitation of computational resources.

2. In learning NER 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.

5 Conclusion

Efficient and robust NER model is very important in practice. This paper provides an empirical study on the impact of training data size and domain information on the performance stability of NER. Experimental results show that it is difficult to significantly enhance the performance when the training data size is above a certain threshold. The threshold of the training data size varies with domains. The performance stability of each NE type recognition also varies with domains. The large-scale corpus statistic data also show that NE types have different distribution across domains. These empirical investigations provide useful hints for enhancing the performance stability of NER models across domains with less efforts. In order to enhance the NER performance across domains, we present an informative training sample selection method. Experimental results show that the performance is significantly enhanced by using informative training samples.

In the future, we'd like to focus on further exploring more effective methods to adapt NER model to a new domain with much less efforts, time and performance degrading.

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.

Jian F. Gao, Mu Li, Anndy Wu, and Chang N., Huang. 2005. Chinese Word Segmentation and Named Entity Recognition: A Pragmatic Approach. *Computational Linguisitc*,31(4):531-574.

Hong L. Guo, Jian M. Jiang, Gang Hu, and Tong Zhang. 2005. Chinese Named Entity Recognition Based on Multilevel Linguistic Features. *Lecture Notes in Artificial Intelligence*,3248:90-99,Springer.

Hideki Isozaki and Hideto Kazawa. 2002. Efficient support vector classifiers for named entity recognition. In *Proceedings of Coling-2002*, pages 1-7.

Hongyan Jing, Radu Florian, Xiaoqiang Luo, Tong Zhang, and Abraham Ittycheriah. 2003. Howtogetachinesename (entity) : Segmentation and combination issues. In *EMNLP 2003*, pages 200-207.

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.

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

Andrew Kamal McCallum and K. Nigam. 1998. Employing EM in pool-based active learning for text classification. *Proceedings of 15th International Conference on Machine Learning*, pages 350-358.

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

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.

Dan Shen, Jie Zhang, Jian Su, Gou D. Zhou, and Chew L.Tan, 2004. Multi-Criteria-based Active Learning for Named Entity Recognition. *Proceedings of ACL04*, pages 589-596.

Yu Z. Wu, Jun Zhao, Bo Xu, and Hao Yu. 2005. Chinese Named Entity Recognition Based on Multiple Features. *Proceedings of EMNLP05*, pages 427-434

Shi H. Yu, Shuan H. Bai, and Paul 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.