Domain Adaptation with Latent Semantic Association for Named Entity Recognition

Honglei Guo Huijia Zhu Zhili Guo Xiaoxun Zhang Xian Wu and Zhong Su
IBM China Research Laboratory
Beijing, P. R. China
{guohl, zhuhuiji, guozhili, zhangxx, wuxian, suzhong}@cn.ibm.com

Abstract

Domain adaptation is an important problem in named entity recognition (NER). NER classifiers usually lose accuracy in the domain transfer due to the different data distribution between the source and the target domains. The major reason for performance degrading is that each entity type often has lots of domain-specific term representations in the different domains. The existing approaches usually need an amount of labeled target domain data for tuning the original model. However, it is a labor-intensive and time-consuming task to build annotated training data set for every target domain. We present a domain adaptation method with latent semantic association (LaSA). This method effectively overcomes the data distribution difference without leveraging any labeled target domain data. LaSA model is constructed to capture latent semantic association among words from the unlabeled corpus. It groups words into a set of concepts according to the related context snippets. In the domain transfer, the original term spaces of both domains are projected to a concept space using LaSA model at first, then the original NER model is tuned based on the semantic association features. Experimental results on English and Chinese corpus show that LaSA-based domain adaptation significantly enhances the performance of NER.

1 Introduction

Named entities (NE) are phrases that contain names of persons, organizations, locations, etc. NER is an important task in information extraction and natural language processing (NLP) applications. Supervised learning methods can effectively solve NER problem by learning a model from manually labeled data (Borthwick, 1999; Sang and Meulder, 2003; Gao et al., 2005; Florian et al., 2003). However, empirical study shows that NE types have different distribution across domains (Guo et al., 2006). Trained NER classifiers in the source domain usually lose accuracy in a new target domain when the data distribution is different between both domains.

Domain adaptation is a challenge for NER and other NLP applications. In the domain transfer, the reason for accuracy loss is that each NE type often has various specific term representations and context clues in the different domains. For example, {“economist”, “singer”, “dancer”, “athlete”, “player”, “philosopher”, ...} are used as context clues for NER. However, the distribution of these representations are varied with domains. We expect to do better domain adaptation for NER by exploiting latent semantic association among words from different domains. Some approaches have been proposed to group words into “topics” to capture important relationships between words, such as Latent Semantic Indexing (LSI) (Deerwester et al., 1990), probabilistic Latent Semantic Indexing (pLSI) (Hofmann, 1999), Latent Dirichlet Allocation (LDA) (Blei et al., 2003). These models have been successfully employed in topic modeling, dimensionality reduction for text categorization (Blei et al., 2003), ad hoc IR (Wei and Croft., 2006), and so on.

In this paper, we present a domain adaptation method with latent semantic association. We focus

Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the ACL, pages 281–289,
Boulder, Colorado, June 2009.

on capturing the hidden semantic association among words in the domain adaptation. We introduce the LaSA model to overcome the distribution difference between the source domain and the target domain. LaSA model is constructed from the unlabeled corpus at first. It learns latent semantic association among words from their related context snippets. In the domain transfer, words in the corpus are associated with a low-dimension concept space using LaSA model, then the original NER model is tuned using these generated semantic association features. The intuition behind our method is that words in one concept set will have similar semantic features or latent semantic association, and share syntactic and semantic context in the corpus. They can be considered as behaving in the same way for discriminative learning in the source and target domains. The proposed method associates words from different domains on a semantic level rather than by lexical occurrence. It can better bridge the domain distribution gap without any labeled target domain samples. Experimental results on English and Chinese corpus show that LaSA-based adaptation significantly enhances NER performance across domains.

The rest of this paper is organized as follows. Section 2 briefly describes the related works. Section 3 presents a domain adaptation method based on latent semantic association. Section 4 illustrates how to learn LaSA model from the unlabeled corpus. Section 5 shows experimental results on large-scale English and Chinese corpus across domains, respectively. The conclusion is given in Section 6.

2 Related Works

Some domain adaptation techniques have been employed in NLP in recent years. Some of them focus on quantifying the generalizability of certain features across domains. Roark and Bacchiani (2003) use maximum a posteriori (MAP) estimation to combine training data from the source and target domains. Chelba and Acero (2004) use the parameters of the source domain maximum entropy classifier as the means of a Gaussian prior when training a new model on the target data. Daume III and Marcu (2006) use an empirical Bayes model to estimate a latent variable model grouping instances into domain-specific or common across both domains. Daume III (2007) further augments the feature space on the instances of both domains. Jiang and Zhai (2006) exploit the domain structure contained in the training examples to avoid over-fitting the training domains. Arnold et al. (2008) exploit feature hierarchy for transfer learning in NER. Instance weighting (Jiang and Zhai, 2007) and active learning (Chan and Ng, 2007) are also employed in domain adaptation. Most of these approaches need the labeled target domain samples for the model estimation in the domain transfer. Obviously, they require much efforts for labeling the target domain samples.

Some approaches exploit the common structure of related problems. Ando et al. (2005) learn predicative structures from multiple tasks and unlabeled data. Blitzer et al. (2006, 2007) employ structural corresponding learning (SCL) to infer a good feature representation from unlabeled source and target data sets in the domain transfer. We present LaSA model to overcome the data gap across domains by capturing latent semantic association among words from unlabeled source and target data.

In addition, Miller et al. (2004) and Freitag (2004) employ distributional and hierarchical clustering methods to improve the performance of NER within a single domain. Li and McCallum (2005) present a semi-supervised sequence modeling with syntactic topic models. In this paper, we focus on capturing hidden semantic association among words in the domain adaptation.

3 Domain Adaptation Based on Latent Semantic Association

The challenge in domain adaptation is how to capture latent semantic association from the source and target domain data. We present a LaSA-based domain adaptation method in this section.

NER can be considered as a classification problem. Let X be a feature space to represent the observed word instances, and let Y be the set of class labels. Let $p_s(x, y)$ and $p_t(x, y)$ be the true underlying distributions for the source and the target domains, respectively. In order to minimize the efforts required in the domain transfer, we often expect to use $p_s(x, y)$ to approximate $p_t(x, y)$.

However, data distribution are often varied with the domains. For example, in the economics-to-

entertainment domain transfer, although many NE triggers (e.g. "company" and "Mr.") are used in both domains, some are totally new, like "dancer", "singer". Moreover, many useful words (e.g. "economist") in the economics NER are useless in the entertainment domain. The above examples show that features could change behavior across domains. Some useful predictive features from one domain are not predictive or do not appear in another domain. Although some triggers (e.g. "singer", "economist") are completely distinct for each domain, they often appear in the similar syntactic and semantic context. For example, triggers of person entity often appear as the subject of "visited", "said", etc, or are modified by "excellent", "popular", "famous" etc. Such latent semantic association among words provides useful hints for overcoming the data distribution gap of both domains.

Hence, we present a LaSA model $\theta_{s,t}$ to capture latent semantic association among words in the domain adaptation. $\theta_{s,t}$ is learned from the unlabeled source and target domain data. Each instance is characterized by its co-occurred context distribution in the learning. Semantic association feature in $\theta_{s,t}$ is a hidden random variable that is inferred from data. In the domain adaptation, we transfer the problem of semantic association mapping to a posterior inference task using LaSA model. Latent semantic concept association set of a word instance x (denoted by $SA(x)$) is generated by $\theta_{s,t}$. Instances in the same concept set are considered as behaving in the same way for discriminative learning in both domains. Even though word instances do not appear in a training corpus (or appear rarely) but are in similar context, they still might have relatively high probability in the same semantic concept set. Obviously, $SA(x)$ can better bridge the gap between the two distributions $p_s(y|x)$ and $p_t(y|x)$. Hence, LaSA model can enhance the estimate of the source domain distribution $p_s(y|x;\theta_{s,t})$ to better approximate the target domain distribution $p_t(y|x;\theta_{s,t})$.

4 Learning LaSA Model from Virtual Context Documents

In the domain adaptation, LaSA model is employed to find the latent semantic association structures of "words" in a text corpus. We will illustrate how to build LaSA model from words and their context snippets in this section. LaSA model actually can be considered as a general probabilistic topic model. It can be learned on the unlabeled corpus using the popular hidden topic models such as LDA or pLSI.

4.1 Virtual Context Document

The distribution of content words (e.g. nouns, adjectives) is usually varied with domains. Hence, in the domain adaptation, we focus on capturing the latent semantic association among content words. In order to learn latent relationships among words from the unlabeled corpus, each content word is characterized by a virtual context document as follows.

Given a content word x_i, the virtual context document of x_i (denoted by vd_{x_i}) consists of all the context units around x_i in the corpus. Let n be the total number of the sentences which contain x_i in the corpus. vd_{x_i} is constructed as follows.

$$vd_{x_i} = \{F(x_i^{s_1}), ..., F(x_i^{s_k}), ..., F(x_i^{s_n})\}$$

where, $F(x_i^{s_k})$ denotes the context feature set of x_i in the sentence s_k, $1 \leq k \leq n$.

Given the context window size {-t, t} (i.e. previous t words and next t words around x_i in s_k). $F(x_i^{s_k})$ usually consists of the following features.

1. Anchor unit $A_C^{x_i}$: the current focused word unit x_i.
2. Left adjacent unit $A_L^{x_i}$: The nearest left adjacent unit x_{i-1} around x_i, denoted by $A_L(x_{i-1})$.
3. Right adjacent unit $A_R^{x_i}$: The nearest right adjacent unit x_{i+1} around x_i, denoted by $A_R(x_{i+1})$.
4. Left context set $C_L^{x_i}$: the other left adjacent units $\{x_{i-t}, ..., x_{i-j}, ..., x_{i-2}\}$ ($2 \leq j \leq t$) around x_i, denoted by $\{C_L(x_{i-t}), ..., C_L(x_{i-j}), ..., C_L(x_{i-2})\}$.
5. Right context set $C_R^{x_i}$: the other right adjacent units $\{x_{i+2}, ..., x_{i+j}, ..., x_{i+t}\}$ ($2 \leq j \leq t$) around x_i, denoted by $\{C_R(x_{i+2}), ..., C_R(x_{i+j}), ..., C_R(x_{i+t})\}$.

For example, given x_i="*singer*", s_k="*This popular new singer attended the new year party*". Let the context window size be {-3,3}. $F(singer)$ = $\{singer, A_L(new), A_R(attend(ed)), C_L(this), C_L(popular), C_R(the), C_R(new)\}$.

vd_{x_i} actually describes the semantic and syntactic feature distribution of x_i in the domains. We construct the feature vector of x_i with all the observed context features in vd_{x_i}. Given vd_{x_i} =

$\{f_1, ..., f_j, ..., f_m\}$, f_j denotes jth context feature around x_i, $1 \leq j \leq m$, m denotes the total number of features in vd_{x_i}. The value of f_j is calculated by Mutual Information (Church and Hanks, 1990) between x_i and f_j.

$$Weight(f_j, x_i) = \log_2 \frac{P(f_j, x_i)}{P(f_j)P(x_i)} \quad (1)$$

where, $P(f_j, x_i)$ is the joint probability of x_i and f_j co-occurred in the corpus, $P(f_j)$ is the probability of f_j occurred in the corpus. $P(x_i)$ is the probability of x_i occurred in the corpus.

4.2 Learning LaSA Model

Topic models are statistical models of text that posit a hidden space of topics in which the corpus is embedded (Blei et al., 2003). LDA (Blei et al., 2003) is a probabilistic model that can be used to model and discover underlying topic structures of documents. LDA assumes that there are K "topics", multinomial distributions over words, which describes a collection. Each document exhibits multiple topics, and each word in each document is associated with one of them. LDA imposes a Dirichlet distribution on the topic mixture weights corresponding to the documents in the corpus. The topics derived by LDA seem to possess semantic coherence. Those words with similar semantics are likely to occur in the same topic. Since the number of LDA model parameters depends only on the number of topic mixtures and vocabulary size, LDA is less prone to over-fitting and is capable of estimating the probability of unobserved test documents. LDA is already successfully applied to enhance document representations in text classification (Blei et al., 2003), information retrieval (Wei and Croft., 2006).

In the following, we illustrate how to construct LDA-style LaSA model $\theta_{s,t}$ on the virtual context documents. Algorithm 1 describes LaSA model training method in detail, where, Function $AddTo(data, Set)$ denotes that $data$ is added to Set. Given a large-scale unlabeled data set D_u which consists of the source and target domain data, virtual context document for each candidate content word is extracted from D_u at first, then the value of each feature in a virtual context document is calculated using its Mutual Information (see Equation 1 in Section 4.1) instead of the counts when running

Algorithm 1: LaSA Model Training

```
Inputs:
• Unlabeled data set: D_u;
Outputs:
• LaSA model: θ_{s,t};
Initialization:
• Virtual context document set: VD_{s,t} = ∅;
• Candidate content word set: X_{s,t} = ∅;
Steps:
begin
    foreach content word x_i ∈ D_u do
        if Frequency(x_i) ≥ the predefined threshold then
            AddTo(x_i, X_{s,t});
    foreach x_k ∈ X_{s,t} do
        foreach sentence S_i ∈ D_u do
            if x_k ∈ S_i then
                F(x_k^{S_i}) ⟵
                {x_k, A_L^{x_k}, A_R^{x_k}, C_L^{x_k}, C_R^{x_k}};
                AddTo(F(x_k^{S_i}), vd_{x_k});
        AddTo(vd_{x_k}, VD_{s,t});
    • Generate LaSA model θ_{s,t} with Dirichlet distribution on VD_{s,t}.
end
```

LDA. LaSA model $\theta_{s,t}$ with Dirichlet distribution is generated on the virtual context document set $VD_{s,t}$ using the algorithm presented by Blei et al (2003).

1	2	3	4	5
customer	theater	company	Beijing	music
president	showplace	government	Hongkong	film
singer	courtyard	university	China	arts
manager	center	community	Japan	concert
economist	city	team	Singapore	party
policeman	gymnasium	enterprise	New York	Ballet
reporter	airport	bank	Vienna	dance
director	square	market	America	song
consumer	park	organization	Korea	band
dancer	building	agency	international	opera

Table 1: Top 10 nouns from 5 randomly selected topics computed on the economics and entertainment domains

LaSA model learns the posterior distribution to decompose words and their corresponding virtual context documents into topics. Table 1 lists top 10 nouns from a random selection of 5 topics computed on the unlabeled economics and entertainment domain data. As shown, words in the same topic are representative nouns. They actually are grouped into broad concept sets. For example, set 1, 3 and 4 correspond to nominal person, nominal organization and location, respectively. With a large-scale unlabeled corpus, we will have enough words assigned to each topic concept to better approximate the underlying semantic association distribution.

In LDA-style LaSA model, the topic mixture is drawn from a conjugate Dirichlet prior that remains the same for all the virtual context docu-

ments. Hence, given a word x_i in the corpus, we may perform posterior inference to determine the conditional distribution of the hidden topic feature variables associated with x_i. Latent semantic association set of x_i (denoted by $SA(x_i)$) is generated using Algorithm 2. Here, Multinomial($\theta_{s,t}(vd_{x_i})$) refers to sample from the posterior distribution over topics given a virtual document vd_{x_i}. In the domain adaptation, we do semantic association inference on the source domain training data using LaSA model at first, then the original source domain NER model is tuned on the source domain training data set by incorporating these generated semantic association features.

Algorithm 2: Generate Latent Semantic Association Set of Word x_i Using K-topic LaSA Model

Inputs:
- $\theta_{s,t}$: LaSA model with multinomial distribution;
- $Dirichlet(\alpha)$: Dirichlet distribution with parameter α;
- x_i: Content word;

Outputs:
- $SA(x_i)$: Latent semantic association set of x_i;

Steps:

begin
- Extract vd_{x_i} from the corpus.
- Draw topic weights $\theta_{s,t}(vd_{x_i})$ from $Dirichlet(\alpha)$;
- **foreach** f_j *in* vd_{x_i} **do**
  - draw a topic $z_j \in \{1,...,K\}$ from Multinomial($\theta_{s,t}(vd_{x_i})$);
  - $AddTo(z_j, Topics(vd_{x_i}))$;
- Rank all the topics in $Topics(vd_{x_i})$;
- $SA(x_i) \longleftarrow$ top n topics in $Topics(vd_{x_i})$;

end

LaSA model better models latent semantic association distribution in the source and the target domains. By grouping words into concepts, we effectively overcome the data distribution difference of both domains. Thus, we may reduce the number of parameters required to model the target domain data, and improve the quality of the estimated parameters in the domain transfer. LaSA model extends the traditional bag-of-words topic models to context-dependence concept association model. It has potential use for concept grouping.

5 Experiments

We evaluate LaSA-based domain adaptation method on both English and Chinese corpus in this section. In the experiments, we focus on recognizing person (PER), location (LOC) and organization (ORG) in the given four domains, including economics (Eco), entertainment (Ent), politics (Pol) and sports (Spo).

5.1 Experimental setting

In the NER domain adaptation, nouns and adjectives make a significant impact on the performance. Thus, we focus on capturing latent semantic association for high-frequency nouns and adjectives (i.e. occurrence count $\geq$ 50) in the unlabeled corpus. LaSA models for nouns and adjectives are learned from the unlabeled corpus using Algorithm 1 (see section 4.2), respectively. Our empirical study shows that better adaptation is obtained with a 50-topic LaSA model. Therefore, we set the number of topics N as 50, and define the context view window size as {-3,3} (i.e. previous 3 words and next 3 words) in the LaSA model learning. LaSA features for other irrespective words (e.g. token unit "the") are assigned with a default topic value N+1.

All the basic NER models are trained on the domain-specific training data using RRM classifier (Guo et al., 2005). RRM is a generalization Winnow learning algorithm (Zhang et al., 2002). We set the context view window size as {-2,2} in NER. Given a word instance x, we employ local linguistic features (e.g. word unit, part of speech) of x and its context units (i.e. previous 2 words and next 2 words) in NER. All Chinese texts in the experiments are automatically segmented into words using HMM.

In LaSA-based domain adaptation, the semantic association features of each unit in the observation window {-2,2} are generated by LaSA model at first, then the basic source domain NER model is tuned on the original source domain training data set by incorporating the semantic association features. For example, given the sentence *"This popular new singer attended the new year party"*, Figure 1 illustrates various features and views at the current word w_i= "$singer$" in LaSA-based adaptation.

			→ Tagging →		
Position	w_{i-2}	w_{i-1}	w_i	w_{i+1}	w_{i+2}
Word	$popular$	new	$singer$	$attend$	the
POS	adj	adj	$noun$	$verb$	$article$
SA	$SA(popular)$	$SA(new)$	$SA(singer)$	$SA(attend)$	$SA(the)$
.....					
Tag	t_{i-2}	t_{i-1}	$\mathbf{t_i}$		

Figure 1: Feature window in LaSA-based adaptation

In the viewing window at the word "$singer$" (see Figure 1), each word unit around "$singer$" is codified with a set of primitive features (e.g. POS, SA, Tag), together with its relative position to "$singer$".

Here, "SA" denotes semantic association feature set which is generated by LaSA model. "Tag" denotes NE tags labeled in the data set.

Given the input vector constructed with the above features, 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.

In our evaluation, only NEs with correct boundaries and correct class labels are considered as the correct recognition. We use the standard Precision (P), Recall (R), and F-measure ($F = \frac{2PR}{P+R}$) to measure the performance of NER models.

5.2 Data

We built large-scale English and Chinese annotated corpus. English corpus are generated from wikipedia while Chinese corpus are selected from Chinese newspapers. Moreover, test data do not overlap with training data and unlabeled data.

5.2.1 Generate English Annotated Corpus from Wikipedia

Wikipedia provides a variety of data resources for NER and other NLP research (Richman and Schone, 2008). We generate all the annotated English corpus from wikipedia. With the limitation of efforts, only PER NEs in the corpus are automatically tagged using an English person gazetteer. We automatically extract an English Person gazetteer from wikipedia at first. Then we select the articles from wikipedia and tag them using this gazetteer.

In order to build the English Person gazetteer from wikipdedia, we manually selected several key phrases, including "births", "deaths", "surname", "given names" and "human names" at first. For each article title of interest, we extracted the categories to which that entry was assigned. The entry is considered as a person name if its related explicit category links contain any one of the key phrases, such as "Category: human names". We totally extracted 25,219 person name candidates from 204,882 wikipedia articles. And we expanded this gazetteer by adding the other available common person names. Finally, we obtained a large-scale gazetteer of 51,253 person names.

All the articles selected from wikipedia are further tagged using the above large-scale gazetteer. Since human annotated set were not available, we held out more than 100,000 words of text from the automatically tagged corpus to as a test set in each domain. Table 2 shows the data distribution of the training and test data sets.

Domains	Training Data Set		Test Data Set	
	Size	PERs	Size	PERs
Pol	0.45M	9,383	0.23M	6,067
Eco	1.06M	21,023	0.34M	6,951
Spo	0.47M	17,727	0.20M	6,075
Ent	0.36M	12,821	0.15M	5,395

Table 2: English training and test data sets

We also randomly select 17M unlabeled English data (see Table 3) from Wikipedia. These unlabeled data are used to build the English LaSA model.

	All	Domain			
		Pol	Eco	Spo	Ent
Data Size(M)	17.06	7.36	2.59	3.65	3.46

Table 3: Domain distribution in the unlabeled English data set

5.2.2 Chinese Data

We built a large-scale high-quality Chinese NE annotated corpus. All the data are news articles from several Chinese newspapers in 2001 and 2002. All the NEs (i.e. PER, LOC and ORG) in the corpus are manually tagged. Cross-validation checking is employed to ensure the quality of the annotated corpus.

Domain	Size	NEs in the training data set			
	(M)	PER	ORG	LOC	Total
Pol	0.90	11,388	6,618	14,350	32,356
Eco	1.40	6,821	18,827	14,332	39,980
Spo	0.60	11,647	8,105	7,468	27,220
Ent	0.60	12,954	2,823	4,665	20,442
Domain	Size	NEs in the test data set			
	(M)	PER	ORG	LOC	Total
Pol	0.20	2,470	1,528	2,540	6,538
Eco	0.26	1,098	2,971	2,362	6,431
Spo	0.10	1,802	1,323	1,246	4,371
Ent	0.10	2,458	526	738	3,722

Table 4: Chinese training and test data sets

All the domain-specific training and test data are selected from this annotated corpus according to the domain categories (see Table 4). 8.46M unlabeled Chinese data (see Table 5) are randomly selected from this corpus to build the Chinese LaSA model.

5.3 Experimental Results

All the experiments are conducted on the above large-scale English and Chinese corpus. The overall performance enhancement of NER by LaSA-based

	All	Domain			
		Pol	Eco	Spo	Ent
Data Size(M)	8.46	2.34	1.99	2.08	2.05

Table 5: Domain distribution in the unlabeled Chinese data set

domain adaptation is evaluated at first. Since the distribution of each NE type is different across domains, we also analyze the performance enhancement on each entity type by LaSA-based adaptation.

5.3.1 Performance Enhancement of NER by LaSA-based Domain Adaptation

Table 6 and 7 show the experimental results for all pairs of domain adaptation on both English and Chinese corpus, respectively. In the experiment, the basic source domain NER model M_s is learned from the specific domain training data set D_{dom} (see Table 2 and 4 in Section 5.2). Here, $dom \in \{Eco, Ent, Pol, Spo\}$. F^{in}_{dom} denotes the top-line F-measure of M_s in the source trained domain dom. When M_s is directly applied in a new target domain, its F-measure in this basic transfer is considered as baseline (denoted by F_{Base}). F_{LaSA} denotes F-measure of M_s achieved in the target domain with LaSA-based domain adaptation. $\delta(F) = \frac{F_{LaSA} - F_{Base}}{F_{Base}}$, which denotes the relative F-measure enhancement by LaSA-based domain adaptation.

Source → Target	Performance in the domain transfer				
	F_{Base}	F_{LaSA}	$\delta(F)$	$\delta(loss)$	F_{Top}
Eco→Ent	57.61%	59.22%	+2.79%	17.87%	F^{in}_{Ent}=66.62%
Pol→Ent	57.5 %	59.83%	+4.05%	25.55%	F^{in}_{Ent}=66.62%
Spo→Ent	58.66%	62.46%	+6.48%	47.74%	F^{in}_{Ent}=66.62%
Ent→Eco	70.56 %	72.46%	+2.69%	19.33%	F^{in}_{Eco}=80.39%
Pol→Eco	63.62%	68.1%	+7.04%	26.71%	F^{in}_{Eco}=80.39%
Spo→Eco	70.35%	72.85%	+3.55%	24.90%	F^{in}_{Eco}=80.39%
Eco→Pol	50.59%	52.7%	+4.17%	15.81%	F^{in}_{Pol}=63.94%
Ent→Pol	56.12%	59.82%	+6.59%	47.31%	F^{in}_{Pol}=63.94%
Spo→Pol	60.22%	62.6%	+3.95%	63.98%	F^{in}_{Pol}=63.94%
Eco→Spo	60.28%	61.21%	+1.54%	9.93%	F^{in}_{Spo}=69.65%
Ent→Spo	60.28%	62.68%	+3.98%	25.61%	F^{in}_{Spo}=69.65%
Pol→Spo	56.94%	60.48%	+6.22%	27.85%	F^{in}_{Spo}=69.65%

Table 6: Experimental results on English corpus

Experimental results on English and Chinese corpus indicate that the performance of M_s significantly degrades in each basic domain transfer without using LaSA model (see Table 6 and 7). For example, in the "Eco→Ent" transfer on Chinese corpus (see Table 7), F^{in}_{eco} of M_s is 82.28% while F_{Base} of M_s is 60.45% in the entertainment domain. F-measure of M_s significantly degrades by 21.83 percent points in this basic transfer. Significant performance degrading of M_s is observed in all the basic transfer. It shows that the data distribution of both domains is very different in each possible transfer.

Source → Target	Performance in the domain transfer				
	F_{Base}	F_{LaSA}	$\delta(F)$	$\delta(loss)$	F_{Top}
Eco→Ent	60.45%	66.42%	+9.88%	26.29%	F^{in}_{Ent}=83.16%
Pol→Ent	69.89%	73.07%	+4.55%	23.96%	F^{in}_{Ent}=83.16%
Spo→Ent	68.66%	70.89%	+3.25%	15.38%	F^{in}_{Ent}=83.16%
Ent→Eco	58.50%	61.35%	+ 4.87%	11.98%	F^{in}_{Eco}=82.28%
Pol→Eco	62.89%	64.93%	+3.24%	10.52%	F^{in}_{Eco}=82.28%
Spo→Eco	60.44%	63.20%	+ 4.57 %	12.64%	F^{in}_{Eco}=82.28%
Eco→Pol	67.03%	70.90 %	+5.77%	27.78%	F^{in}_{Pol}=80.96%
Ent→Pol	66.64 %	68.94 %	+3.45%	16.06%	F^{in}_{Pol}=80.96%
Spo→Pol	65.40%	67.20%	+2.75%	11.57%	F^{in}_{Pol}=80.96%
Eco→Spo	67.20%	70.77%	+5.31%	15.47%	F^{in}_{Spo}=90.24%
Ent→Spo	70.05%	72.20%	+3.07%	10.64%	F^{in}_{Spo}=90.24%
Pol→Spo	70.99%	73.86%	+4.04%	14.91%	F^{in}_{Spo}=90.24%

Table 7: Experimental results on Chinese corpus

Experimental results on English corpus show that LaSA-based adaptation effectively enhances the performance in each domain transfer (see Table 6). For example, in the "Pol→Eco" transfer, F_{Base} is 63.62% while F_{LaSA} achieves 68.10%. Compared with F_{Base}, LaSA-based method significantly enhances F-measure by 7.04%. We perform t-tests on F-measure of all the comparison experiments on English corpus. The p-value is 2.44E-06, which shows that the improvement is statistically significant.

Table 6 also gives the accuracy loss due to transfer in each domain adaptation on English corpus. The accuracy loss is defined as $loss = 1 - \frac{F}{F^{in}_{dom}}$. And the relative reduction in error is defined as $\delta(loss)$= $|1 - \frac{loss_{LaSA}}{loss_{Base}}|$. Experimental results indicate that the relative reduction in error is above 9.93% with LaSA-based transfer in each test on English corpus. LaSA model significantly decreases the accuracy loss by 29.38% in average. Especially for "Spo→Pol" transfer, $\delta(loss)$ achieves 63.98% with LaSA-based adaptation. All the above results show that LaSA-based adaptation significantly reduces the accuracy loss in the domain transfer for English NER without any labeled target domain samples.

Experimental results on Chinese corpus also show that LaSA-based adaptation effectively increases the accuracy in all the tests (see Table 7). For example, in the "Eco→Ent" transfer, compared with F_{Base}, LaSA-based adaptation significantly increases F-measure by 9.88%. We also perform t-tests on F-

measure of 12 comparison experiments on Chinese corpus. The p-value is 1.99E-06, which shows that the enhancement is statistically significant. Moreover, the relative reduction in error is above 10% with LaSA-based method in each test. LaSA model decreases the accuracy loss by 16.43% in average. Especially for the "Eco→Ent" transfer (see Table 7), $\delta(loss)$ achieves 26.29% with LaSA-based method.

All the above experimental results on English and Chinese corpus show that LaSA-based domain adaptation significantly decreases the accuracy loss in the transfer without any labeled target domain data. Although automatically tagging introduced some errors in English source training data, the relative reduction in errors in English NER adaptation seems comparable to that one in Chinese NER adaptation.

5.3.2 Accuracy Enhancement for Each NE Type Recognition

Our statistic data (Guo et al., 2006) show that the distribution of NE types varies with domains. Each NE type has different domain features. Thus, the performance stability of each NE type recognition is very important in the domain transfer.

Figure 2 gives F-measure of each NE type recognition achieved by LaSA-based adaptation on English and Chinese corpus. Experimental results show that LaSA-based adaptation effectively increases the accuracy of each NE type recognition in the most of the domain transfer tests. We perform t-tests on F-measure of the comparison experiments on each NE type, respectively. All the p-value is less than 0.01, which shows that the improvement on each NE type recognition is statistically significant. Especially, the p-value of English and Chinese PER is 2.44E-06 and 9.43E-05, respectively, which shows that the improvement on PER recognition is very significant. For example, in the "Eco→Pol" transfer on Chinese corpus, compared with F_{Base}, LaSA-based adaptation enhances F-measure of PER recognition by 9.53 percent points. Performance enhancement for ORG recognition is less than that one for PER and LOC recognition using LaSA model since ORG NEs usually contain much more domain-specific information than PER and LOC.

The major reason for error reduction is that external context and internal units are better semantically associated using LaSA model. For example, LaSA

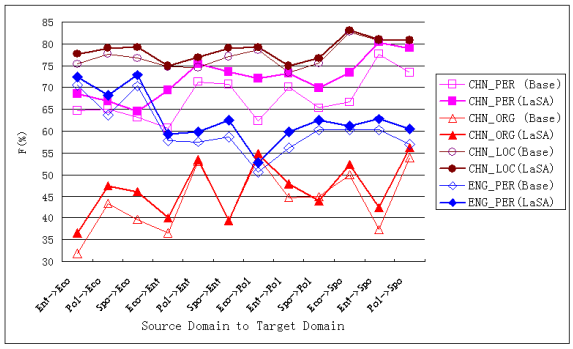

Figure 2: PER, LOC and ORG recognition in the transfer

model better groups various titles from different domains (see Table 1 in Section 4.2). Various industry terms in ORG NEs are also grouped into the semantic sets. These semantic associations provide useful hints for detecting the boundary of NEs in the new target domain. All the above results show that LaSA model better compensates for the feature distribution difference of each NE type across domains.

6 Conclusion

We present a domain adaptation method with LaSA model in this paper. LaSA model captures latent semantic association among words from the unlabeled corpus. It better groups words into a set of concepts according to the related context snippets. LaSA-based domain adaptation method projects words to a low-dimension concept feature space in the transfer. It effectively overcomes the data distribution gap across domains without using any labeled target domain data. Experimental results on English and Chinese corpus show that LaSA-based domain adaptation significantly enhances the performance of NER across domains. Especially, LaSA model effectively increases the accuracy of each NE type recognition in the domain transfer. Moreover, LaSA-based domain adaptation method works well across languages. To further reduce the accuracy loss, we will explore informative sampling to capture fine-grained data difference in the domain transfer.

References

Rie Ando and Tong Zhang. 2005. A Framework for Learning Predictive Structures from Multiple Tasks

and Unlabeled Data. In *Journal of Machine Learning Research 6 (2005)*, pages 1817–1853.

Andrew Arnold, Ramesh Nallapati, and William W. Cohen. 2008. Exploiting Feature Hierarchy for Transfer Learning in Named Entity Recognition. In *Proceedings of 46th Annual Meeting of the Association of Computational Linguistics (ACL'08)*, pages 245-253.

David Blei, Andrew Ng, and Michael Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research, 3:993–1022*.

John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain Adaptation with Structural Correspondence Learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006)*, pages 120-128.

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL'07)*, pages 440-447.

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

Yee Seng Chan and Hwee Tou Ng. 2007. Domain Adaptation with Active Learning for Word Sense Disambiguation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL'07)*.

Ciprian Chelba and Alex Acero. 2004. Adaptation of maximum entropy capitalizer: Little data can help a lot. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information and lexicography. *Computational Linguistics*, 16(1):22–29.

Hal Daume III. 2007. Frustratingly Easy Domain Adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*.

Hal Daume III and Daniel Marcu. 2006. Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research*, 26:101–126.

Scott Deerwester, Susan T. Dumais, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Radu Florian, Abe Ittycheriah, Hongyan Jing, and Tong Zhang. 2003. Named entity recogintion through classifier combination. In *Proceedings of the 2003 Conference on Computational Natural Language Learning*.

Freitag. 2004. Trained Named Entity Recognition Using Distributional Clusters. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*.

Scott Miller, Jethran Guinness, and Alex Zamanian. 2004. Name Tagging with Word Clusters and Discriminative Training. In *Proceedings of HLT-NAACL 04*.

Jianfeng Gao, Mu Li, Anndy Wu, and Changning Huang. 2005. Chinese Word Segmentation and Named Entity Recognition: A Pragmatic Approach. *Computational Linguisitc*, 31(4):531–574.

Honglei Guo, Jianmin Jiang, Gang Hu, and Tong Zhang. 2005. Chinese Named Entity Recognition Based on Multilevel Linguistic Features. In *Lecture Notes in Artificial Intelligence*, 3248:90–99.

Honglei Guo, Li Zhang, and Zhong Su. 2006. Empirical Study on the Performance Stability of Named Entity Recognition Model across Domains. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006)*, pages 509-516.

Thomas Hofmann. 1999. Probabilistic latent semantic indexing. In *Proceedings of the 22th Annual International SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99)*.

Jing Jiang and ChengXiang Zhai. 2006. Exploiting Domain Structure for Named Entity Recognition. In *Proceedings of HLT-NAACL 2006*, pages 74–81.

Jing Jiang and ChengXiang Zhai. 2007. Instance Weighting for Domain Adaptation in NLP. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL'07)*, pages 264–271.

Wei Li and Andrew McCallum. 2005. Semi-supervised sequence modeling with syntactic topic models. In *Proceedings of Twenty AAAI Conference on Artificial Intelligence (AAAI-05)*.

Alexander E. Richman and Patrick Schone. 2008. Mining Wiki Resources for Multilingual Named Entity Recognition. In *Proceedings of the 46th Annual Meeting of the Association of Computational Linguistics*.

Brian Roark and Michiel Bacchiani. 2003. Supervised and unsupervised PCFG adaptation to novel domains. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language independent named entity recognition. In *Proceedings of the 2003 Conference on Computational Natural Language Learning (CoNLL-2003)*, pages 142–147.

Xing Wei and Bruce Croft. 2006. LDA-based document models for ad-hoc retrieval. In *Proceedings of the 29th Annual International SIGIR Conference on Research and Development in Information Retrieval*.

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