Some Studies on Chinese Domain Knowledge Dictionary and Its Application to Text Classification

Jingbo Zhu
Natural Language Processing Lab
Institute of Computer Software & Theory
Northeastern University, Shenyang
zhujingbo@mail.neu.edu.cn

Wenliang Chen
Natural Language Processing Lab
Institute of Computer Software & Theory
Northeastern University, Shenyang
Chenwl@mail.neu.edu.cn

Abstract

In this paper, we study some issues on Chinese domain knowledge dictionary and its application to text classification task. First a domain knowledge hierarchy description framework and our Chinese domain knowledge dictionary named NEUKD are introduced. Second, to alleviate the cost of construction of domain knowledge dictionary by hand, we use a boostrapping-based algorithm to learn new domain associated terms from a large amount of unlabeled data. Third, we propose two models (BOTW and BOF) which use domain knowledge as textual features for text categorization. But due to limitation of size of domain knowledge dictionary, we further study machine learning technique to solve the problem, and propose a BOL model which could be considered as the extended version of BOF model. Naïve Bayes classifier based on BOW model is used as baseline system in the comparison experiments. Experimental results show that domain knowledge is very useful for text categorization, and BOL model performs better than other three models, including BOW, BOTW and BOF models.

1 Introduction

It is natural for people to know the topic of the document when they see some specific words in the document. For example, when we read a news, if title of the news includes a word "姚明(Yao Ming)", as we know, "姚明(Yao Ming)" is a famous China basketball athlete in US NBA game, so we could recognize the topic of the document is about "篮球，体育(Basketball, Sports)" with our domain knowledge. In this paper, we call the specific word "姚明(Yao Ming)" as a *Domain Associated Term* (DAT). A DAT is a word or a phrase (compound words) that enable humans to recognize intuitively a topic of text with their domain knowledge. In fact, domain knowledge is a kind of commonsense knowledge. We think that domain knowledge is very useful for text understanding tasks, such as text classification, document summarization, and information retrieval.

In previous literatures, some researchers used knowledge bases for text understanding tasks(Scott et al., 1998), such as WordNet for English and HowNet for Chinese. We know that WordNet and HowNet are lexical and semantic knowledge resources. Other researchers tried to use commonsense knowledge such as field-associated terms for text understanding tasks(M. Fuketa et al., 2000, Sangkon Lee and Masami Shishibori, 2002). But the problem of limitation of size of such knowledge base is still a key bottleneck for using domain knowledge dictionary for text understanding tasks, and how to solve it is an ongoing research focus.

In the following content, we try to give answers to four questions: 1)What is our Chinese domain knowledge dictionary NEUKD? 2)How to learn DATs from a large amount of unlabelled data? 3)How to use the Chinese domain knowledge dictionary NEUKD for text classification? 4)Due to the problem of limitation of size of domain knowledge dictionary, how to solve the

problem and improve performance of text classification using domain knowledge dictionary?

2 Domain Knowledge Dictionary

We first introduce briefly domain knowledge hierarchy description framework (DKF) which includes three levels: *Domain Level* (DL), *Domain Feature Level* (DFL) and *Domain Associated Term Level* (DATL). The DL is the top level which defines many domains, such as "体育(Sports)", "军事(Military Affairs)". The DFL is the second level which defines many domain features. A domain defined in the DL has a lot of domain features defined in the DFL. For example, domain "军事(Military Affairs)" has many domain features, such as "军队(Army Feature)", "武器(Weapon Feature)" and "战争(War Feature)". The DATL is the third level which defines many domain associated terms. Many domain associated terms could indicate a same domain feature defined in the DFL. For example, some domain associated terms, such as some domain associated terms, such as "中东战争(Mid-East War)", "伊拉克战争(Iraq War)" and "阿富汗战争(Afghanistan War)", indicate domain feature "战争(War)".

Since 1996 we employed a semi-automatic machine learning technique to acquire domain knowledge from a large amount of labeled and unlabeled corpus, and built a general-purpose domain knowledge dictionary named NEUKD according to the domain knowledge hierarchy description framework(Zhu Jingbo et al., 2002). Items defined in the NEUKD include domain associated term, domain feature and domain. Currently 40 domains, 982 domain features and more than 610,000 domain associated terms are defined in the NEUKD. Some instances of NEUKD are given in Table 1. Because the size of NEUKD is limited, so in following content we will study machine learning techniques to solve the problem of using NEUKD for text classification task.

Domain Associated Terms	Domain Features	Domain
姚明 (Yao Ming)	篮球，运动员 (Basketball, Athlete)	体育 (Sports)
三峡工程 (The Sanxia project)	水利工程 (Irrigation Project)	水利 (Irrigation Works)
赛季 (Match Season)	比赛 (Match)	体育 (Sports)
阿森纳队 (Arsenal Team)	足球 (Football)	体育 (Sports)
中国工商银行 (Industrial and commercial bank of China)	银行 (Bank)	金融 (Finance)

Table 1. Some instances defined in the NEUKD

3 Bootstrapping-based DAT Learning Algorithm

To extend domain knowledge dictionary, in this paper, we will use a *feature learning algorithm based on bootstrapping* (FLB)(Zhu Jingbo et al., 2004) to learning new DATs. In the FLB learning procedure, some seed words are given in advance. In fact, seed words are some important DATs. For example, ten seed words of domain "金融(finance)" are 股票(stock), 金融(finance), 贷款(loan), 证券(stock), 财经(finance and economics), 银行(bank), 税收(tax), 外汇(foreign exchange), 投资(investment) and 股市(stock market).

The FLB learning procedure is described as follows:

- **Initialization**: Use a small number of seed words initialize DAT set
- **Iterate Bootstrapping**:
  - **Candidate DAT Learner**: Learn some new DATs as candidate DATs from unlabeled data.
  - **Evaluation**: Score all candidate DATs, and select top-n best DATs as new seed words, and add them into DAT set.

In the beginning of algorithm, all words except stopwords in the unlabeled corpus could be

considered as candidate DATs. In fact, we can regard bootstrapping as iterative clustering. In the evaluation step of FLB algorithm, RlogF metric method(Ellen Riloff, Rosie Jones, 1999) is used as evaluation function which assigns a score to a word(candidate DAT). The score of a word is computed as:

$$M(w_i) = Log_2 F(w_i, X) \times R_i \qquad (1)$$

Where $F(w_i,X)$ is the frequency of co-occurrence of word w_i and X (set of seed words) in the same sentence, $F(w_i)$ is the frequency of w_i in the corpus, and $R_i=F(w_i,X)/F(w_i)$. The RlogF metric tries to strike a balance between reliability and frequency: R is high when the word is highly correlated with set of seed words, and F is high when the word and X highly co-occur in the same sentence.

In the experiments, we use the corpus from 1996-1998 People's Daily as unlabeled data which has about 50 million words. For domain "金融(finance)", we select ten seed words shown in above example, the bootstrapping-based DAT learning algorithm obtains 65% precision performance within top-1000 new learned DATs according to human judgment.

4 Domain Knowledge based Text Classification

In this paper, naïve Bayes(NB) model (McCallum and K.Nigam, 1998) is used to build text classifier. We want to study how to use our Chinese domain knowledge dictionary NEUKD to improve text categorization.

4.1 BOW Model

The most commonly used document representation is the so called vector space model(G.Salton and M.J.McGill, 1983). In the vector space model, documents are represented by vectors of terms (textual features, e.g. words, phases, etc.). Conventional bag-of-words model (BOW) uses common words as textual features. In the comparison experiments, we use the BOW model as baseline NB system.

4.2 BOTW Model

As above mentioned, more than 610000 domain associated terms (DATs) are defined in the NEUKD, such as "姚明(Yao Ming) ", "三峡工程(The Sanxia project)", and "中国工商银行(Industrial and commercial bank of China)" shown in table 1. We use domain associated terms and common words as textual features, called BOTW models (short for bag-of-terms and words model). For example, in the previous examples, the DAT "三峡工程(The Sanxia project, Sanxia is a LOCATION name of China)" can be used as a textual feature in BOTW model. But in BOW model(baseline system) we consider two common words "三峡(The Sanxia)" and "工程(project)" as two different textual features.

4.3 BOF Model

Similar to BOTW model, we use domain features as textual features in the NB classifier, called BOF model (short for bag-of-features model). In BOF model, we first transform all DATs into domain features according to definitions in the NEUKD, and group DATs with same domain features as a cluster, called ***Topic Cluster***. For Examples, Topic Cluster "体育(sports)" includes some DATs, such as "赛季(match season)", "阿森纳队(Arsenal)", "奥运会(Olympic Games)", "乒乓球(Table Tennis)", "姚明(Yao Ming)". In BOF model, we use topic clusters as textual features for text categorization. Also the classification computation procedure of BOF model is same as of BOW model.

4.4 BOL Model

To solve the problem of the limitation of NEUKD, in this paper, we propose a machine learning technique to improve BOF model. The basic ideas are that we wish to learn new DATs from pre-classified documents, and group them into the predefined topic clusters which are formed and used as textual features in BOF model discussed in section 4.3. Then these new topic clusters could be used as textual features for text categorization. We call the new model as BOL model(short for bag-of-learned features model) which could be consider as an extended version of BOF model.

First we group all DATs originally defined in NEUKD into a lot of topic clusters as described in BOF model, which are used as seeds in following learning procedure. Then we group other words (not be defined in NEUKD) into these topic clusters. The Learning algorithm is described as following:

- **Preprocessing**: Text segmentation, extracting candidate words, and sort the candidate words by CHI method. As above mentioned, all candidate words except stopwords which

are not defined in NEUKD will be grouped into topic clusters in this process.

- **Initialization**: These words, which are defined in NEUKD, are first added to corresponding topic clusters according to their associated domain features, respectively.
- **Iteration**: Loop until all candidate words have been put into topic clusters:
  - Measure similarity of a candidate word and each topic cluster, respectively.
  - Put the candidate word into the most similar topic cluster(Note that a word can only be grouped into one cluster).

The important issue of above procedures is how to measure the similarity between a word and a topic cluster. Chen Wenliang *et. al.*(2004) proposed a measure for word clustering algorithm used in text classification. So in this paper, we use Chen's measure to measure the similarity between a word and a topic cluster in above learning algorithm. The similarity of a word w_t and a topic cluster f_j is defined as

$$S = \lambda(w_t)(\xi(w_t, w_t \vee f_j) + \xi(f_j, w_t \vee f_j))$$

$$\lambda(w_t) = \frac{N(w_t) + N(f_j)}{\sum_{i=1}^{|L|} N(f_i) + |W|} \quad (2)$$

Where

$$\xi(w_t, w_t \vee f_j) = \frac{P(w_t)}{P(w_t) + P(f_j)} \times D(P(C \mid w_t) \,\|\, P(C \mid w_t \vee f_j))$$

$$\xi(f_j, w_t \vee f_j) = \frac{P(f_j)}{P(w_t) + P(f_j)} \times D(P(C \mid f_j) \,\|\, P(C \mid w_t \vee f_j))$$

Where we define the distribution $P(C|w_t)$ as the random variable over classes C, and its distribution given a particular word w_t. $N(f_i)$ denote the number of words in the topic cluster f_i, W is the list of candidate words.

To describe how to estimate distribution P(C|f) , we first assume that in the beginning of learning procedure, only a word w_1 is included in topic cluster f_1, we could say that $P(C|f_1) = P(C|w_1)$. When a new word w_2 is added into topic cluster f_1, we could get a new topic cluster f_2. How to estimate the new distribution $P(C|f_2)$ is key step, where $f_2=w_2 \vee f_1$. We could use the following formula (3) to estimate distribution $P(C|f_2) = P(C|w_2 \vee f_1)$. Similarly, we could know if the new word w_n is added into topic cluster f_{n-1} to form a new topic cluster f_n, we also could estimate $P(C|f_n)=P(C|w_n \vee f_{n-1})$ following this way, and so on.

$$P(C \mid f_2) = P(C \mid w_2 \vee f_1) = \frac{P(w_2)}{P(w_2) + P(f_1)} P(C \mid w_2) + \frac{P(f_1)}{P(w_2) + P(f_1)} P(C \mid f_1) \quad (3)$$

We turn back the question about how to measure the difference between two probability distributions. Kullback-Leibler divergence is used to do this. The KL divergence between two class distributions induced by w_t and w_s is written as

$$D(P(C \mid w_t) \,\|\, P(C \mid w_s)) = -\sum_{j=1}^{|C|} P(c_j \mid w_t) \log\left(\frac{P(c_j \mid w_t)}{P(c_j \mid w_s)}\right) \quad (4)$$

In preprocessing step, the CHI statistic measures the lack of independence of feature t and category c.

$$\chi 2(t,c) = \frac{N(AD - BC)^2}{(A+C)(B+D)(A+B)(C+D)}$$

Where t refers to a feature and c refers to a category, A is the number of times t and c co-occur, B is the number of times t occurs without c, C is the number of times c occurs without t, D is the number of times neither c nor t co-occur, and N is the total number of documents.

5 Experimental Results

In this paper, we use naïve Bayes for classifying documents. Here we only describe multinomial naïve Bayes briefly since full details have been presented in the paper(McCallum and K.Nigam, 1998). The basic idea in naïve Bayes approaches is to use the joint probabilities of words and categories to estimate the probabilities of categories when a document is given. Given a document *d* for classification, we calculate the probabilities of each category *c* as follows:

$$P(c \mid d) = \frac{P(c)P(d \mid c)}{P(d)} \propto P(c) \prod_{i=1}^{|T|} \frac{P(t_i \mid c)^{N(t_i \mid d)}}{N(t_i \mid d)!}$$

Where $N(t_i|d)$ is the frequency of word t_i in document *d,* T is the vocabulary and |T| is the

size of T, t_i is the i^{th} word in the vocabulary, and $P(t_i|c)$ thus represents the probability that a randomly drawn word from a randomly drawn document in category *c* will be the word t_i.

In the experiments, we use NEU_TC data set(Chen Wenliang et. al. 2004) to evaluate the performance of baseline NB classifier and our classifiers. The NEU_TC data set contains Chinese web pages collected from web sites. The pages are divided into 37 classes according to "Chinese Library Categorization"(CLCEB, 1999). It consists of 14,459 documents. We do not use tag information of pages. We use the toolkit CipSegSDK(Yao Tianshun *et. al.* 2002) for word segmentation. We removed all words that had less than two occurrences. The resulting vocabulary has about 60000 words.

In the experiments, we use 5-fold cross validation where we randomly and uniformly split each class into 5 folds and we take four folds for training and one fold for testing. In the cross-validated experiments we report on the average performance. For evaluating the effectiveness of category assignments by classifiers to documents, we use the conventional recall, precision and F1 measures. Recall is defined to be the ratio of correct assignments by the system divided by the total number of correct assignments. Precision is the ratio of correct assignments by the system divided by the total number of the system's assignments. The F1 measure combines recall (r) and precision (p) with an equal weight in the following form:

$$F_1(r,p) = \frac{2rp}{r+p}$$

In fact, these scores can be computed for the binary decisions on each individual category first and then be averaged over categories. The way is called macro-averaging method. For evaluating performance average across class, we use the former way called micro averaging method in this paper which balances recall and precision in a way that gives them equal weight. The micro-averaged F1 measure has been widely used in cross-method comparisons.

To evaluate the performance of these four models based on NB classifier, we construct four systems in the experiments, including BOW, BOTW, BOF and BOL classifier. CHI measure is used to feature selection in all text classifiers.

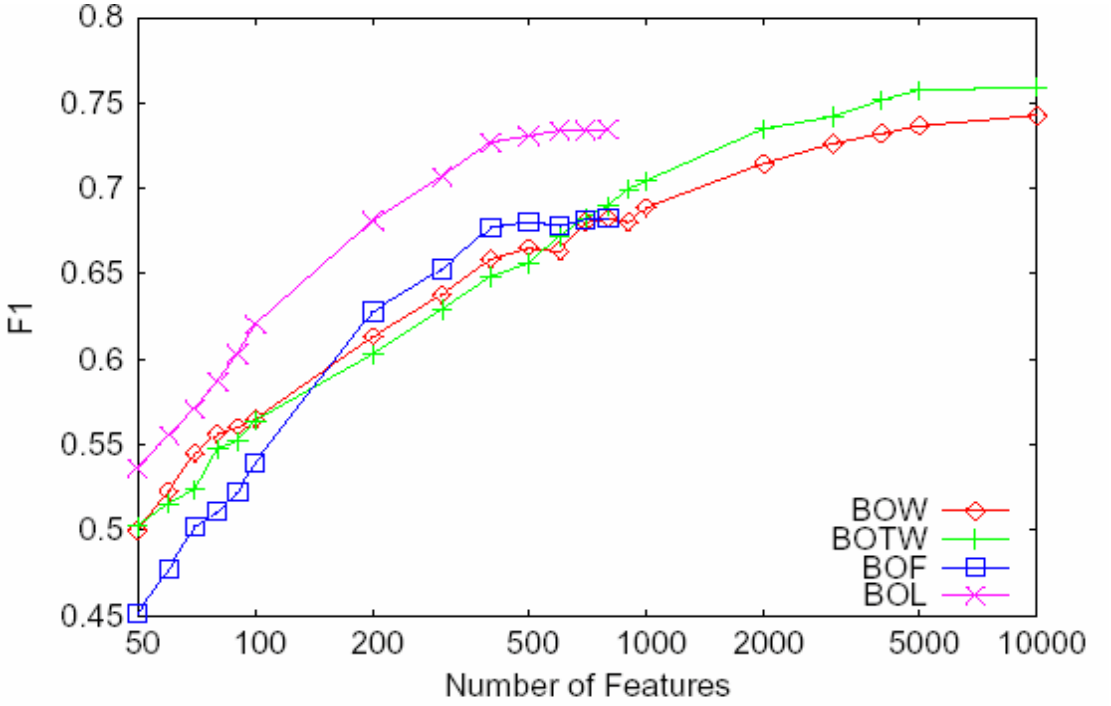

Figure 1. Experimental results of BOW, BOTW, BOF, BOL classifiers

In figure 1, we could find that BOTW classifier always performs better than BOW classifier when the number of features is larger than about 500. From comparative experimental results of BOTW and BOW classifiers, we think that domain associated items are a richer and more precise representation of meaning than common words. Because the total number of domain features in NEUKD is only 982, in figure 1 we find the maximum number of features (domain fea-

tures) for BOF and BOL classifier is less than 1000. When the number of features is between 200 and 1000, BOF classifier performs better than BOW and BOTW classifiers. It is also obvious that BOL classifier always performs better than other three classifiers when the number of features is less than 1000. As above mentioned, in BOL model, we use a machine learning technique to solve the problem of limitation of size of NEUKD, and group rest 65.01% words into predefined topic clusters as textual features in BOL model. So the classifier based on BOL model can yield better performance than BOF model.

6 Conclusions and Future Work

In this paper, we first introduce our Chinese domain knowledge dictionary NEUKD. To alleviate the cost of construction of domain knowledge dictionary by hand, we propose a boostrapping-based algorithm to learn new domain associated terms from a large amount of unlabeled data. This paper studies how to improve text categorization by using domain knowledge dictionary. To do it, we propose two models using domain knowledge as textual features. The first one is BOTW model which uses domain associated terms and common words as textual features. The other one is BOF model which uses domain features as textual features. But due to limitation of size of domain knowledge dictionary, many useful words are lost in the training procedure. We study and use a machine learning technique to solve the problem to improve knowledge-based text categorization, and propose a BOL model which could be considered as the extension version of BOF model. Comparison experimental results of those four models (BOW, BOTW, BOF and BOL) show that domain knowledge is very useful for improving text categorization. In fact, a lot of knowledge-based NLP application systems have to face the problem of limitation of size of knowledge bases. Like our work discussed in this paper, we think that using machine learning techniques is a good way to solve such problem. In the future work, we will study how to apply the domain knowledge to improve other text understanding tasks, such as information retrieval, information extraction, topic detection and tracking (TDT).

Acknowledgements

This research was supported in part by the National Natural Science Foundation of China & Microsoft Research Asia (No. 60203019), the Key Project of Chinese Ministry of Education (No. 104065), and the National Natural Science Foundation of China (No. 60473140).

References

Chen Wenliang, Chang Xingzhi, Wang Huizhen, Zhu Jingbo, and Yao Tianshun. 2004. Automatic Word Clustering for Text Categorization Using Global Information. *First Asia Information Retrieval Symposium (AIRS 2004)*, LNCS, Beijing, pp.1-6

CLCEB. 1999. China Library Categorization Editorial Board. China Library Categorization (The 4th ed.) (In Chinese), Beijing, *Beijing Library Press.*

Ellen Riloff, Rosie Jones. 1999. Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping, *Proceedings of the Sixteenth National Conference on Artificial Intelligence.*

G.Salton and M.J.McGill, 1983. An introduction to modern information retrieval, *McGraw-Hill*.

McCallum and K.Nigam. 1998. A Comparison of Event Models for naïve Bayes Text Classification, *In AAAI-98 Workshop on Learning for Text Categorization*.

M. Fuketa, S.Lee, T.Tsuji, M.Okada and J. Aoe. 2000. A document classification method by using field associated words. *International Journal of Information Sciences*. 126(1-4), p57-70

Sangkon Lee, Masami Shishibori. 2002. Passage segmentation based on topic matter, *International journal of computer processing of oriental languages*, 15(3), p305-339.

Scott, Sam, Stan Matwin. 1998. Text classification using WordNet hypernyms. *Proceedings of the COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems*, Montreal.

Yao Tianshun, Zhu Jingbo, Zhang li, and Yang Ying, 2002. Natural Language Processing- research on making computers understand human languages, *Tsinghua University Press*, (In Chinese).

Zhu Jingbo and Yao Tianshun. 2002. FIFA-based Text Classification, *Journal of Chinese Information Processing*, V16, No3.(In Chinese)

Zhu Jingbo, Chen Wenliang, and Yao Tianshun. 2004. Using Seed Words to Learn to Categorize Chinese Text. *Advances in Natural Language Processing: 4th International Conference (EsTAL 2004)*, pp.464-473