Semantic Representation in Text Classification Using Topic Signature Mapping

Palakorn Achananuparp, Xiaohua Zhou, Xiaohua Hu, and Xiaodan Zhang

Abstract— **Document representation is one of the crucial components that determine the effectiveness of text classification tasks. Traditional document representation approaches typically adopt a popular bag-of-word method as the underlying document representation. Although it's a simple and efficient method, the major shortcoming of bag-of-word representation is in the independent of word feature assumption. Many researchers have attempted to address this issue by incorporating semantic information into document representation. In this paper, we study the effect of semantic representation on the effectiveness of text classification systems. We employed a novel semantic smoothing technique to derive semantic information in a form of mapping probability between topic signatures and single-word features. Two classifiers, Naïve Bayes and Support Vector Machine, were selected to carry out the classification experiments. Overall, our topic-signature semantic representation approaches significantly outperformed traditional bag-of-word representation in most datasets.**

I. Introduction

One of the major criticisms of bag-of-word document representation is that it treats each individual word as an independent feature, therefore, it ignores any semantic relatedness that might exist between words. Given such assumption, document representation is likely to contain an inaccurate set of features. For example, synonymous words such as automobile and car will be considered as two different features, each with one occurrence, while polysemous words such as bank in financial bank and river bank will be treated as the same feature with two occurrences. Another related issue of bag-of-word representation is data sparseness problem. In many cases, words that appear in training documents of a given class may not appear in testing collection.

In this paper, we propose the method to address the above shortcomings by incorporating semantic information into document representation. To derive semantic information for each word vector, we employ a novel technique called context-sensitive semantic smoothing which statistically maps topic signatures to single-word features. The classification experiments on 4 well-known datasets using Naïve Bayes and Support Vector Machine (SVM) classifiers demonstrated that our method significantly outperforms a bag-of-word approach in most conditions.

Our main contributions can be summarized as follow. First, we proposed the methods to incorporate semantic information into document representation using topic-signature semantic smoothing technique. Our method is different from others in that it includes contextual information from multiword phrases, thus the mapping assignment can be more specific. For example, given the task of determining the relatedness between two single words bank and boat, it might be difficult to do so because the word bank alone has several meanings. Without any contextual information, it will be very ambiguous to judge their semantic relatedness effectively. However, by giving the phrase river bank instead of bank, it is relatively easier to relate river bank with boat. Next, we consider two additional factors in our evaluation, corpora complexity and data sparseness, and we empirically evaluate our method along those factors.

The rest of the paper is organized as followed. In section 2, we discuss related work on document representation and text classification. Next, we present the proposed methods in section 3. Section 4 and 5 describes experimental setup and results, respectively. Finally, we conclude the outcome of our work in section 6.

II. Related Work

Document representation is one of the major components that determine the performance of text classification. Therefore, it has been a major interest within text classification research community to develop techniques to improve upon traditional bag-of-word representation. These techniques can be grouped into two main approaches according to how semantic information or concepts are obtained. The first approaches rely on statistical mechanism to automatically extract conceptual words from the corpus or

Manuscript received November 30, 2007. (Write the date on which you submitted your paper for review.) This work is supported in part by NSF Career grant (NSF IIS 0448023), NSF CCF 0514679, PA Dept of Health Tobacco Settlement Formula Grant (No. 240205 and No. 240196), and PA Dept of Health Grant (No. 239667). The first author is with College of Information Science and Technology, Drexel University, Philadelphia, PA 19104 USA (corresponding author to provide phone: 267-402-0750; e-mail: pkorn@drexel.edu).

The second author is with College of Information Science and Technology, Drexel University, Philadelphia, PA 19104 USA (corresponding author to provide phone: 267-402-0750; e-mail: xiaohua.zhou@drexel.edu).

The third author is with College of Information Science and Technology, Drexel University, Philadelphia, PA 19104 USA (corresponding author to provide phone: 267-402-0750; e-mail: thu@cis.drexel.edu).

The fourth author is with College of Information Science and Technology, Drexel University, Philadelphia, PA 19104 USA (corresponding author to provide phone: 267-402-0750; e-mail: xzhang@cis.drexel.edu).

to group word into clusters. Approaches based on Probabilistic Latent Semantic Analysis (pLSA) have been proposed by [4] to build concept-based document representation. Concepts were extracted from the corpus and combined with individual terms to create a term-concept representation. Bekkerman et al [1] introduced word-cluster representation using Information Bottleneck method.

Phrase-based representation is a common approach to represent a document. Most techniques employed language models as the underlying algorithm to extract phrases. For example, Peng et al. [12] proposed statistical language models to discover n-gram phrases. Shen et al. [14] introduced an n-multigram language model to automatically extract n-multigram sequences, a frequently-occurring pattern, through Expectation-Maximization algorithm. Next, syntactic parsing was employed by [10] to extract phrases that correspond to a given syntactic relationships.

The second approaches are based on the use of background knowledge as a source of concepts and semantic information. These works can be further categorized into WordNet-based approaches and domain-ontology based approaches. Several strategies for adding and replacing terms with WordNet concepts were investigated by [6]. For instance, in add concepts strategy, each term vector is extended by new entries from WordNet concepts. Replace terms by concepts strategy works similar to the first strategy. But instead of concatenate a term vector with concepts, it removes all the terms from the vector representation if at least one corresponding concept exists. Ontological concepts from existing domain ontologies such as Medical Subject Headings (MeSH) and Unified Medical Language System (UMLS) have been employed in several studies [3][19]. In addition, Several WordNet-based semantic similarity measures have been proposed [18],[9],[13]. These techniques have been applied in various domains including text classification.

III. Proposed Method

A. Language Model and Text Classification

Document smoothing is a family of methods used in language modeling to assign a reasonable non-zero probability to any non-seen terms, indicated by their zero probability in document vector. From document representation perspective, smoothing helps mitigate data sparseness problem as it reduces a number of zero cells. A particular approach to document smoothing, called *semantic smoothing*, refers to the methods to which contextual information is incorporated into the smoothing model. The major advantage of semantic smoothing approach is the ability to recognize semantic relatedness between terms.

When semantic smoothing is applied to a text document, words that have strong semantic relationship with each other are identified by a high mapping probability. For example, *automobile* and *car* have a higher mapping probability than *automobile* and *spoon*. Several semantic smoothing techniques have been successfully applied to information retrieval [2][15][21]. It was empirically proved to be superior to traditional smoothing methods according to retrieval performance.

In our previous work, Zhou et al. [22] proposed a context-sensitive semantic smoothing (CSSS) method for language modeling information retrieval. CSSS decomposes a document into a set of weighted context-sensitive *topic signatures* and then statistically maps topic signatures to word features. Topic signatures play a significant role in word sense disambiguation as they provide a more specific context for the terms in document representation. For instance, suppose *entertainment* and *star* are topic signatures of a given document, the term *actor* in document vector is likely to have more weight than *moon* since it has more probability to occur in this context. In this study, we adopted a context-sensitive semantic smoothing method to incorporate semantic information into document representation. Semantic information was derived from the mapping probabilities between topic signatures and word features.

On the surface, our method resembles "concept vector only" strategy discussed in [6] However, while concept vector approach replaces original word features with semantic categories, CSSS keeps the original word features and employs semantic mapping probability as term weights instead of a simple term frequency. We believe this method is less susceptible to noise (unrelated words), which could be introduced into the document representation via the inclusion of certain semantic features. In subsequent sections, we describe in details about our proposed methods. First, we introduce the notion of topic signatures used in the study. Then, we discuss how topic signatures are used in the formulation of semantic information.

B. Extracting Topic Signatures

We define topic signatures as the conceptual summarization of topics in a given document. Specifically, we considered two types of topic signature representation, *multiword -phrase* and *single word*. The use of word phrases has been investigated in information retrieval community for several years. For this study, we defined multiword phrase as equivalent to a *rigid noun phrase*. Each phrase consists of two or more single adjacent words. The first word should begin with a noun or an adjective and the last words should ends with a noun. Their distance threshold is four words. We considered phrases that occur frequently in a document collection as topic signatures. A modified version of Xtract [15] was employed to extract multiword phrases from a document collection.

The use of single words as topic signatures was similar to a unigram document model smoothing by [2] where single words were statistically mapped according to their translation probability. Once both types of topic signatures

were identified, we tried to find a set of single words to represent semantic meaning of a topic signature. This way, we can smooth a document language model by statistically mapping topic signatures to single-word features. In this study, we experimented with both multiword phrase and single-word topic signatures in order to compare their effectiveness on classification performance.

C. Topic-Signature Mapping for a Set of Documents

In this section, we present the method to incorporate semantic information into document representation. Figure 1 demonstrates the overall mapping concept. First, each document was represented by a vector with single-word features (V_w). Before semantic mapping could be conducted, we identified a set of topic signatures (V_t) by indexing all documents in a collection. Once all topic signatures were extracted, the next step was to estimate the probability of mapping topic signatures to the single-word features of each document. The mathematical models to compute the mapping probability were shown in equation (3.1), (3.2), and (3.3). The final product was a feature vector having a set of indexed terms as the word features. Each cell in a vector contained the mapping probability of topic signatures to a given word feature. As illustrated above, the topic signatures provide the high-level semantic categories to each document. The semantic information is then utilized to "smooth" the document representation.

The following formulation describes the estimation of mapping probability of topic signatures to word features ($p_s(w|d_i)$). Our approach is based on a mixture model consisting two components, *a simple language model* ($p_b(w|d_i)$) and a *translation model* ($p_t(w|d_i)$), as described in equation (3.1). Due to space limitation, we advise the readers to see [17], [20], [21], and [22] for a more comprehensive discussion of related probabilistic models used in this paper.

$$p_s(w \mid d_i) = (1-\lambda)\, p_b(w \mid d_i) + \lambda p_t(w \mid d_i) \qquad (3.1)$$

Where w is a word feature in a document d_i. We employed the *translation coefficient,* λ , in equation (3.1) to control the influence of the two components in the mixture model. It has a real-number value from 0 to 1. If the translation coefficient is equal to zero, the mixture model becomes a simple language model with background smoothing; that is, $p_s(w|d_i) = p_b(w|d_i)$. In contrast, if the translation coefficient is equal to 1, the mixture model is a translation model; that is $p_s(w|d_i) = p_t(w|d_i)$.

The first component in the mixture model, a simple language model with background smoothing ($p_b(w|d_i)$), defined in (3.1), can be computed by:

$$p_b(w \mid d_i) = (1-\alpha)\, p_{ml}(w \mid d_i) + \alpha p(w \mid D) \qquad (3.2)$$

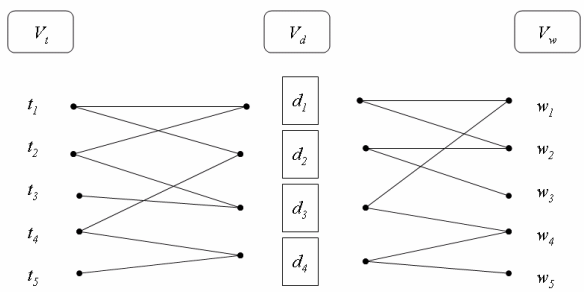

Fig. 1. Topic signatures mapping for a set of documents. V_t, V_d, and V_w are topic signature set, document set, and word set, respectively.

Where w is a word feature in a document d_i and D is a document collection. The *background coefficient* α in (3.2) is used to control the influence of the collection model $p(w|D)$. It has a real-number value from 0 to 1. $p_{ml}(w|d_i)$ is a maximum likelihood estimator.

Lastly, the translation model $p_t(w|c_i)$ that statistically maps topic signatures of a document to the single-word features is described in equation (3.3) below:

$$p_t(w \mid d_i) = \sum_k p(w \mid t_k)\, p(t_k \mid d_i) \qquad (3.3)$$

where t_k represents the k-th topic signature while $p(t_k|d_i)$ can be derived from maximum likelihood estimation. After the mixture model had been trained, we replaced term weight value in a corresponding cell with topic signature mapping probability. With this approach, a word feature in document representation remained the same while semantic information was added into each entry in the matrix.

IV. Experiments

To evaluate the impact of semantic representation on text classification performance, we carried out a series of classification experiments on 4 benchmark datasets. We used Naïve Bayes and SVMLight [7] as the underlying classifiers. The same set of experiments was conducted on each classifier to compare the representation effectiveness on different algorithms.

A. Datasets

Four benchmark datasets were selected to carry out the experiments. These are 20-newsgroup (20NG), LA Times of TREC Disk 5, Reuters, and Topic Detection and Track Corpus version 2 (TDT2).

20NG is collected from 20 Usenet newsgroups. It contains 20 classes with a total of 19,997 articles; each class comprises about 1,000 articles. LA Times of TREC Disk 5 contains a sampling of roughly 40% of news articles published by the Los Angeles Times between Jan 1, 1989 to December 31, 1990. It contains 111,084 articles in 22 sections (e.g. financial, entertainment, sports, etc.). Articles in top 15 sections were selected to be indexed. If a section

contained more than 2,001 articles, only the first 3,000 were selected. Ultimately, the remaining 21,523 articles were indexed. Reuters-21578 corpus contains 21,578 articles from Reuters newswire in 114 categories. TDT2 consists of 64,500 news articles from major news agencies published in 1998. The total of 10,212 articles with unique labels was indexed.

Since multiword phrase extraction, the estimation of translation probability, and the estimation of background model require a large amount of documents, therefore, we indexed the entire document collection. On the other hand, only a subset of documents was selected to be classified.

B. Document Preprocessing

We extracted individual words from the title and body sections of each document while ignoring content from metadata section. Stop words were removed according to a common stop word list. After that, the remaining words were stemmed. Next, we extracted multiword phrases from each corpora using Xtract. Then, we calculated the probability of translating a topic signature (multiword phrases and single-words) to a word feature. We assumed that word features with translation probability less than 0.001 are not semantically related to a given topic signature. Consequently, we treated those terms as noise and removed them. Then, the translation probabilities of the remaining words were renormalized accordingly. We excluded any documents containing less than 5 unique single-word features after indexing. After the exclusion, 20NG contained 19,660 articles left for the experiments. 17,913 articles in 10 sections were selected from LA Times. Reuters had 8,882 articles left while TDT2 contained 7,094 news articles.

C. Evaluation Setting

The main objective of the experiments was to compare the classification performance between traditional bag-of-word representation and the proposed semantic representations. In addition to different representations, we investigated the effects of document representation with respect to corpora complexity and data sparseness.

As discussed by [1], the contribution of document representation on classification performance is also affected by corpora complexity (measured in terms of vocabulary space). To evaluate this, we included both high-complexity corpus, such as 20NG and LA Times, and low-complexity corpus, such as Reuters and TDT2. Next, we evaluated the effect of data sparseness based on different training data size. With smaller training dataset, the model is likely to encounter more unseen terms than larger training dataset; hence, more zero cells in document vectors. To achieve this, we randomly partitioned training data into a given fraction. A large training dataset contains 33% of total documents while a small training dataset contains 1% of documents. For a given percentage of training data, we obtained the average performance from 10 random runs. Each random run was controlled by a random seed of 10. All experiments were conducted using Dragon Toolkit [23].

Based on our parameter tuning, we set background coefficient α in equation (3.2) to 0.5. Translation coefficient λ in equation (3.1) was set to 0.4 for datasets with high complexity (20NG and LA Times) and 0.1 for datasets with low complexity (Reuters and TDT2).

All baseline experiments used a bag-of-word representation with a simple term frequency as a term weighting scheme. We chose CHI feature selector with an alpha value of 0.02 as a feature selector for baseline Naïve Bayes conditions. In baseline SVM cases, we selected document frequency selector with minimum frequency of 5 as a feature selector. Since feature selection has no effect on semantic smoothing, we did not use it in other conditions. Furthermore, we also experimented with TFIDF as additional baseline comparison for SVM classifier.

D. Evaluation Criteria

We adopted a standard set of evaluation metrics: precision (P), recall (R), and F1-measure [16], in our evaluation of text classification performance. Precision is defined as the proportion of actual positive class members with respect to all positive class members. Recall is defined as the proportion of predicted positive class members with respect to all actual positive class members. F1 is the harmonic mean of precision and recall which is computed by the following formula:

$$F1 = \frac{2 \times P \times R}{P + R} \tag{4}$$

Furthermore, we evaluated the performance of multi-category classification using micro-average and macro-average measures. Micro-average assigns equal weight to every document while macro-average assigns equal weight to every category disregarding its frequency. If each category has the same data distribution, then micro-F1 and micro-F1 are basically the same. Statistical significant tests were performed on an improvement of F1 scores using a paired t-test at the significant level of $p<0.01$ (99% confidence interval).

V. Results

A. Naïve Bayes Classifier

The performance of Naïve Bayes classifier is displayed in table 1. Overall, topic-signature semantic representation approach significantly outperformed traditional bag-of-word representation in most datasets (except LA Times) at the significance level of $p<0.01$.

In the case where baseline approach already performed very well (above 95%), i.e. TDT2 cases, semantic representation also outperformed baseline representation, e.g. in TDT2 with 33% training data. Nevertheless, there are quite a few cases in which baseline still performed better than semantic conditions especially on many micro-F1

TABLE I
THE COMPARISON OF DOCUMENT REPRESENTATION TYPES ON NAIVE BAYES CLASSIFIER'S PERFORMANCE.

Dataset	Training Data Size	Micro-F1		
		Baseline	Topic Signature	
			Phrase	Word
20NG	33%	**0.757**	0.747	0.754
	1%	0.382	**0.456**	0.445
LA Times	33%	0.717	0.720	**0.722***
	1%	**0.528**	0.516	0.492
Reuters	33%	**0.894**	0.887	0.886
	1%	**0.764**	0.744	0.743
TDT2	33%	0.965	**0.972**	**0.972**
	1%	**0.921**	0.911	0.909
Dataset	**Training Data Size**	**Macro-F1**		
		Baseline	**Topic Signature**	
			Phrase	**Word**
20NG	33%	**0.751**	0.735	0.742
	1%	0.372	**0.454**	0.444
LA Times	33%	0.698	0.705*	**0.707***
	1%	0.489	**0.492***	0.471
Reuters	33%	**0.775**	0.741	0.743
	1%	0.307	**0.429**	0.427
TDT2	33%	0.955	**0.965**	0.964
	1%	0.847	**0.850***	0.846

The best result for each case is in bold. * denotes the change is not statistically significant at p<0.01.

TABLE II
THE COMPARISON OF DOCUMENT REPRESENTATION TYPES ON SVM CLASSIFIER'S PERFORMANCE.

Dataset	Training Data Size	Micro-F1		
		Baseline	Topic Signature	
			Phrase	Word
20NG	33%	0.799	**0.809**	0.807*
	1%	0.473	**0.506**	0.476*
LA Times	33%	0.782	0.785*	0.780
	1%	0.525	**0.553**	0.530*
Reuters	33%	0.942	**0.947**	**0.947**
	1%	**0.755**	0.752	0.750
TDT2	33%	0.978	**0.981**	0.980
	1%	0.888	**0.901***	0.900*
Dataset	**Training Data Size**	**Macro-F1**		
		Baseline	**Topic Signature**	
			Phrase	**Word**
20NG	33%	0.794	**0.806**	0.802*
	1%	0.465	**0.497**	0.468*
LA Times	33%	**0.766**	0.764*	0.759
	1%	0.492	**0.506**	0.485*
Reuters	33%	**0.881**	0.881*	0.879*
	1%	**0.318**	0.292*	0.290*
TDT2	33%	0.970	**0.974**	**0.974**
	1%	0.764	**0.791**	0.790*

The best result for each case is in bold. * denotes the change is not statistically significant at p<0.01.

results. A possible explanation is that our semantic approach estimated mapping probability at a document-level while Naïve Bayes classifier performs better with smoothing mechanism at a document-class level.

The results in most cases look promising. The highest performance gain in all conditions is Reuters with 1% training set in which semantic approach achieved the relative improvement of 39.74% (from 0.307 to 0.429) on macro-F1. Semantic representation also performed quite well on 20NG dataset. For instance, the improvement at 1% training data was 19.37% (from 0.382 to 0.456) on micro-F1 measure and 22.04% (from 0.372 to 0.454) on macro-F1 measure.

In addition, we found that semantic representation performed relatively well in high-complexity corpora, such as 20NG, than in low-complexity corpora, such as TDT2 according to the magnitude of improvement. In this category, multiword-phrase topic signature representation performed better than single-word topic signatures. This confirms our expectation that contextual information from multiword phrases is helpful for mapping assignments.

In regards to data sparseness factor, we observed that as the data became sparser (from 33% training data to 1% training data), the performance gain also increased significantly, e.g. in 20NG and Reuters's macro-F1 cases. Nonetheless, apart from those datasets, there were cases where semantic representation did not significantly affect performance in different data sparseness levels. Therefore, it is difficult to draw a general conclusion about data sparseness factor from the current results.

B. SVM Classifier

The performance of SVM classifier is displayed in table 2. Overall, semantic representation using topic signatures approach significantly outperformed traditional bag-of-word representation in most datasets at the significance level of p<0.01.

The biggest improvement in SVM is in 20NG with 1% training data condition in which semantic approach achieved the relative improvement of 6.98% (from 0.473 to 0.506) on micro-F1 measure and 6.88% (from 0.465 to 0.497) on macro-F1 measure.

We noticed an interesting issue about improvement gap between Naïve Bayes and SVM classifier. Although the performance improvement in SVM experiment is statistically significant, the magnitude of changes in many SVM cases is relatively smaller than those in Naïve Bayes cases. Although these results are not what we initially expected, they are not entirely incomprehensible. One possible explanation for the fact that semantic representation in Naïve Bayes worked much better than semantic representation in SVM is that document presentation with topic signature mapping is based on the probabilistic mechanism which is more compatible with Naïve Bayes

TABLE III
THE COMPARISON OF DOCUMENT REPRESENTATION TYPES ON SVM CLASSIFIER'S PERFORMANCE HAVING TFIDF AS BASELINE.

Dataset	Training Data Size	Micro-F1				
		TFIDF Baseline	Topic Signature			
			Phrase	Word	Phrase+ TFIDF	Word+ TFIDF
20NG	33%	0.833	0.809	0.807	0.839	**0.841**
	1%	0.555	0.506	0.476	0.582	**0.586**
LA Times	33%	0.781	0.785*	0.780*	**0.803**	**0.803**
	1%	0.531	0.553*	0.530*	**0.572**	0.571
Reuters	33%	0.947	0.947*	0.947*	**0.956**	**0.956**
	1%	0.558	**0.752**	0.750	0.689	0.688
TDT2	33%	0.979	0.981	0.980	**0.983**	**0.983**
	1%	**0.904**	0.901	0.900*	0.896	0.896
Dataset	**Training Data Size**	**Macro-F1**				
		TFIDF Baseline	**Topic Signature**			
			Phrase	**Word**	**Phrase+ TFIDF**	**Word+ TFIDF**
20NG	33%	0.829	0.806	0.802	0.837	**0.839**
	1%	0.546	0.497	0.468	0.573	**0.577**
LA Times	33%	0.762	0.764*	0.759	**0.789**	**0.789**
	1%	0.491	0.506*	0.485	**0.529**	**0.529**
Reuters	33%	0.901	0.881	0.879	**0.915**	0.914
	1%	0.228	**0.292**	0.290	0.237*	0.236*
TDT2	33%	0.973	0.974*	0.974*	**0.978**	**0.978**
	1%	**0.819**	0.791	0.790	0.787	0.787

The best result for each case is in bold. * denotes the change is not statistically significant at $p<0.01$.

classifier than with SVM mechanism.

On the issue of corpora complexity, we observed that semantic representation performed relatively better on highly complex datasets (e.g. 20NG and LA Times) than low complex datasets (Reuters and TDT2). This could be explained by the fact that it only took a few single words to describe documents in low-complexity corpora. Thus, the benefit of using multiword phrase was fairly minimal. On the other hand, a large vocabulary space was likely to benefit more from the coarser granularity features. Furthermore, we noticed that as data sparseness increased, the performance gain also increased in most datasets, except Reuters. Thus, we concluded that semantic representation is generally effective in high data sparseness corpora, with a few exceptions.

Lastly, the performance of SVM classifier with TFIDF baseline is displayed in table 3. Interestingly, semantic representation with topic signatures alone did not outperform TFIDF baseline in most conditions. To further the issue, we tested another strategy by linearly combining TFIDF score with semantic mapping probability. It turned out that the "combined" representation did improve the classification performance over TFIDF baseline. This result suggested that higher performance improvement could be achieved with other term weighting strategies.

VI. CONCLUSIONS

In this paper, we investigated the effects of semantic representation on the performance of text classification. Although many approaches to incorporate semantic information into text classification have been proposed, the effect on classification performance is still subject to the testing corpus and the best strategy is still up to debate.

We proposed a method to enhance document representation with semantic information based on a topic signature mapping. Semantic information was derived by estimating the probability of mapping topic signatures to single word-features. Two types of topic signatures were introduced: multiword phrase and single-word. We used Xtract to automatically obtain multiple word phrases from all documents in a given document collection. After all topic signatures were identified, we began the estimation of semantic mapping between topic signatures and word features of each document.

We conducted the evaluation of our method on 4 datasets using two classifiers, Naïve Bayes and Support Vector Machine. Overall, topic-signature semantic representation approaches significantly outperformed traditional bag-of-word representation in most datasets at the significance level of $p<0.01$. The performance improvement in Naïve Bayes cases was greater than those in SVM. One possible

explanation for Naïve Bayes-favored improvement is that our method is based on the probabilistic mechanism which is more compatible with Naïve Bayes classifier. Moreover, we believe that Naïve Bayes' results can be improved further by performing topic signature mapping at a class model level instead of a document level since it is more compatible with Naïve Bayes' mechanism.

The evaluation confirms several advantages of multiword phrase signatures over single-word signature. First, phrase-word mapping was more effective than word-word mapping because of contextual information available in multiword phrases. Efficiency wise, phrase-word mapping took less time than word-word mapping since a number of single words were relatively greater than a number of multiword phrases. Therefore, the calculation complexity in word-word mapping case was higher.

The results from Naïve Bayes and SVM experiments confirms that our method worked well across corpora complexity levels as shown by high improvement in many datasets. Particularly, SVM experiments showed that multiword-phrase topic signatures worked better on high complexity corpora. However, we did not find a definite conclusion about data sparseness factor although it could be generally observed that our method worked better in most high-sparseness datasets, with a few exceptions.

In the future, we have an idea to improve Naïve Bayes performance by performing topic signature mapping at a class level. With this method, we believe the performance improvement can be significantly raised because the mechanism is more compatible with the Naïve Bayes' mechanism. In addition, we plan to experiment with different strategies of combining semantic score with baseline term weighting score. Furthermore, we plan to conduct additional experiments to compare the effectiveness of our method with background knowledge approaches such as WordNet-based methods. Lastly, we would like to experiment with hybrid approach of combining semantic representation from WordNet with topic signature representation.

ACKNOWLEDGMENT

This work is supported in part by NSF Career grant (NSF IIS 0448023), NSF CCF 0514679, PA Dept of Health Tobacco Settlement Formula Grant (No. 240205 and No. 240196), and PA Dept of Health Grant (No. 239667).

REFERENCES

[1] Bekkerman, R., El-Yaniv, R., Tishby, N., and Winter, Y.: Distributional Word Clusters vs. Words for Text Categorization. Journal of Machine Learning Research, 3(2003), pp. 1183—1208.

[2] Berger, A. and Lafferty, J.: Information Retrieval as Statistical Translation. In Proceedings of the 22nd ACM SIGIR, pp. 222—229 (1999)

[3] Bloehdorn, S. and Hotho, A.: Boosting for text classification with semantic features. In the Workshop on Text-based Information Retrieval (TIR-04) at the 27th German Conference on Artificial Intelligence, September (2004)

[4] Cai, L. and Hoffman, T.: Text categorization by boosting automatically extracted concepts. In Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 182—189, Toronto, CA (2003)

[5] Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., and Harshman, R. A.: Indexing by latent semantic analysis. Journal of the American Society of Information Science, 41(6), pp. 391—407 (1990)

[6] Hotho, A., Staab, S., and Stumme, G.: Ontologies improve text document clustering. In Proceedings of the 3rd International Conference on Data Mining, November 19-22, pp. 541-544 (2003)

[7] Joachims, T.: Text categorization with support vector machines: Leaning with many relevant features. In Proceedings of European Conference on Machine Learning, pp 137-142 (1998)

[8] Jelinek, F. and Mercer, R.: Interpolated estimation of markov source parameters from sparce data. Pattern Recognition in Practice, E.S. Gelseman and L.N. Kanal, Eds., pp. 381—402 (1980)

[9] Leacock, C. and Chodorow, M.: Combining local context and WordNet similarity for word sense identification. In: Fellbaum, C. (Ed.), WordNet: An electronic lexical database, MIT Press, Cambridge, MA. pp. 265-283.

[10] Lewis, D. D.: Representation quality in text classification: An introduction and experiment. In Proceedings of a Workshop on Speech and Natural Language, Hidden Valley, Pennsylvania (1990)

[11] McCallum, A. and Nigam, K.: A comparison of event models for Naïve Bayes text classification. AAAI Workshop on Learning for Text Categorization, pp. 41—48 (1998)

[12] Peng, F., Schuurmans, D., and Wang, S.: AugmentingNaïve Bayes classifiers with statistical language models. Information Retrieval, 7(34), pp. 317-345 (2004)

[13] Resnik, P.: Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In Proceeding of International Joint Conference for Artificial Intelligence (IJCAI-95), pp.448—453 (1995)

[14] Shen, D., Sun, J. T., Yang, Q., and Chen, Z.: Text Classification Improved through Multigram Models. In Proceedings of the ACM Fifteenth Conference on Information and Knowledge Management (CIKM'06). Arlington, USA. November 6-11 (2006)

[15] Smadja, F.: Retrieving collocations from text: Xtract. Computational Linguistics, 19(1), pp. 143—177 (1993)

[16] Van, R.C.: Information Retrieval. Butterworths, London, second edition (1979)

[17] Wei X. and Croft, W. B.: LDA-based document models for ad-hoc retrieval. In Proceedings of the 29th ACM SIGIR, pp. 178-185

[18] Wu, Z. and Palmer, M.: Verb Semantics and Lexical Selection. In Proceedings of the 32nd Annual Meeting of the Associations for Computational Linguistics (ACL'94), pp. 133—138, Las Cruces, New Mexico (1994)

[19] Yetisgen-Yildiz, M. and Pratt, W.: The effect of feature representation on medline document classification. In Proceedings of the American Medical Informatics Association Fall Symposium, Washington D.C. (2005)

[20] Zhai, C. and Lafferty, J.: A Study of Smoothing Methods for Language Models Applied to Ad hoc Information Retrieval. In Proceedings of the 24th ACM SIGIR Conference on Research and Development in IR, pp. 334—342 (2001)

[21] Zhou, X., Hu, X., Zhang, X., Lin, X., and Song, I. Y.: Context-sensitive Semantic Smoothing for Language Modeling Approach to Genomic Information Retrieval. In the 29th Annual International ACM SIGIR Conference (ACM SIGIR 2006), August 6-11 Seattle, WA, pp. 170—177 (2006)

[22] Zhou, X. Zhang, X., and Hu, X.: Semantic Smoothing of Document Models for Agglomerative Clustering. In the 12th International Joint Conference on Artificial Intelligence (IJCAI 2007), January 6-12, India, pp. 2922-2927 (2007)

[23] Zhou, X., Zhang, X., and Hu, X.: Dragon Toolkit: Incorporating Auto-learned Semantic Knowledge into Large-Scale Text Retrieval and Mining. In proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI), October 29-31, Patras, Greece (2007)