Learning to Classify Short Text with Actively Acquired Knowledge from Universal Knowledge Base

Peng Qi
School of Software, Tsinghua University
Beijing 100084, China
qipeng.thu@gmail.com

Xiaoming Jin
School of Software, Tsinghua University
Beijing 100084, China
xmjin@tsinghua.edu.cn

Dou Shen
CityGrid Media
12729 Northup Way
Bellevue, WA 98006
doushen@gmail.com

ABSTRACT

Short text is becoming dominant in user-generated content with the popularity of microblogs and social network. Text classification, as a primary solution of understanding text is facing challenges for short text due to data sparseness. One way to solve the sparseness is to introduce external data sources or called universal knowledge base (UKB), to enrich short text. Previous study has proved the effectiveness when the UKB is relevant to the short-text data set under study. However, this situation is not always true since in most cases, the accessible UKB can be too general, with lots of irrelevant information. Therefore, a fundamental problem yet to be studied is how to best utilize UKB by leveraging the relevant information and avoiding the negative impact from irrelevant information. In this paper, we present an effective solution for this problem by actively acquiring related knowledge from the UKB, and then seamlessly enrich the original short text data. Moreover, we devised an iterative framework for the knowledge selection and short text enrichment process based on the well studied topic modeling algorithms. We thoroughly compared our proposed solution with the state-of-art approaches on public data sets. Experimental results show that our solution can reduce the short-text classification error by 5∼20% in worst case without losing efficiency.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Text analysis*; I.2.6 [**Artificial Intelligence**]: Learning

General Terms

Algorithm

Keywords

Short Text, Text Classification, Universal Knowledge Base, Actively Aquired Knowledge

1. INTRODUCTION

Text classification has been the primary approach for understanding large text corpus. Many methods such as k-nearest neighbors (k-NN) [7], Naïve Bayes [14], maximum entropy [17], and support vector machines (SVMs) [6] have been studied and applied to various benchmarks (Reuters-21578 [16], WebKB [25], etc.) and other large scale text data sets. They can usually achieve relatively satisfactory results.

Recently, the volume of short text in various forms such as Web search queries, forum & chat messages, product reviews, microblogs, and so on, is exponentially increasing. Directly applying conventional text classification methods on short text does not work in most cases due to the fundamental difference between short text and traditional documents. Short text is usually too short to provide a meaningful word distribution, and meanwhile their sparseness determines that they are noisier and less topic-focused. As a result, given two pieces of short text, it is hard to measure their semantic similarity, which fails the conventional text classification methods.

There are two sorts of popular approaches in recent studies to tackle this problem. One approach is utilizing search engines to obtain unstructured (the Internet) or structured (Wikipedia[1], etc.) data sources to acquire more related background information for further similarity measurement or input data enrichment. Often used features from this approach are searching snippets, search result titles/abstracts, and page counts [1, 4]. However, this approach faces the disadvantage of low efficiency, as searching in large-scale datasets can be rather time-consuming, which can hardly meet the need of real-time web applications. Another approach focuses on topic models and word ontology, where usually a structured universal dataset is used, and topic models are estimated from such datasets to filter noise, perform word sense disambiguation, and most importantly, provide context and word correlation information for short text data. Recent studies have proved that using latent topics estimated from such universal datasets, relatively satisfactory accuracy can be achieved in both classifying and clustering tasks [19, 21]. However, these approaches often fail to consider about topic coverage or granularity of the universal dataset, which may result in less satisfactory results if the universal dataset contains too many irrelevant topics with the short texts. For example, suppose we are to classify NIPS papers into two categories – physical neurology and computational neuroscience, with a universal dataset Wikipedia, on which a topic model that is too coarse to distinguish these two fields, is built. Such a situation will no doubt become a challenge for state-of-art works that haven't taken top-

[1] http://www.wikipedia.org/ (In many works, the English version of Wikipedia is preferred: http://en.wikipedia.org/)

ic coverage of the universal dataset into consideration. However, a selected subset of the UKB relevant to these two categories can provide more satisfactory performance.

Inspired by the idea of using external knowledge topics to enrich short texts, combining with the intuition of improving the performance of short text classification with actively acquired knowledge, we present a general framework in this paper. The framework aims to deal with peculiarly distributed short texts with very large knowledge base, by actively acquiring relevant knowledge from the knowledge base to enrich short text. The underlying idea is that, for all classification tasks, we introduce a common very large universal knowledge base (UKB), then use the UKB to enrich the content of the short texts, and finally build a classifier on the enriched short texts. The framework is mainly based on the recent successful latent topic analysis models such as pLSA [11] and LDA [2], and powerful machine learning methods such as maximum entropy and SVMs. The advantages of the proposed framework include:

- **Universality** The proposed framework is designed to fit all classification tasks where direct approach cannot provide satisfactory outcome due to data sparseness, which is not merely limited to short text classification tasks.[2] More importantly, for all classification tasks in one given field, only *one* universal knowledge base covering "world knowledge" is sufficient as our framework is able to select the most relevant document in such a UKB.
- **Reduce Sparseness** By enriching them with latent topic information of relevant UKB documents, we reduce the sparseness of short texts and thus achieve better classification accuracy.
- **Iterative Approach** Our proposed framework works in a iterative style, which can diminish the undesired noise in input training dataset of short texts to an acceptable level, and be used for different tasks considering various time limit and classification accuracy requirements.
- **Easy to Implement** The proposed framework is based on mature topic model and machine learning algorithms, thus it is easy to be implemented and run on multiple platforms (and open source resources are affluent).

This paper is organized as follows. Related work of the field is introduced in Section 2, which provides the reader with background information of our work. In Section 3, we will pose the problem that we seek to solve and give a brief analysis. Section 4 concerns mainly about how our proposed framework is structured and how it functions, and Section 5 contains carefully designed experiments of the framework and detailed evaluations. Finally, Section 6 gives a conclusion of our work and offers some discussion of possible future directions.

2. RELATED WORK

"Learning to Classify Short and Sparse Text & Web with Hidden Topics from Large-scale Data Collections" by Phan *et al* 2008 [21] is probably the study most relevant to our work, which attempts to enrich input short text with latent topic information selected from the universal dataset in order to improve short text classification outcome. The main difference is that this work is a one-pass process, while ours can be run for certain iterations so that the UKB can be made the most of, where a selected subset of the UKB is used to enrich the data instead of the whole universal dataset, as is in Phan *et al*'s work.

Another related work that makes use of structured universal dataset is the utilization of searching engines to enhance short text clustering by Banerjee *et al* [1], where Wikipedia is also used as the universal knowledge source in a direct way. Bollegala *et al* [4] also applied searching engine to improve word similarity measurement by using page counts and searching snippets as part of the measurement. These works, although proven effective, depend severely on the implementation of the searching engine, namely how the documents are indexed, searched, ranked, and even drawn snippets from. Besides, the reliance on searching engines limits the performance of such approaches, and thus makes them less useful in real-time applications that requires relatively high efficiency.

On the other hand, similarity measurement or short text classifications methods based on taxonomy or generative topic models can eliminate this undesired reliance on searching engines, meet the performance need to some extent, and link the words and short texts together in a semantic way. In these approaches, researchers often employ topic models such as LDA [2] and pLSA [11], and universal datasets (Wikipedia, Wordnet, etc.) as reliable external knowledge bases. Apart from [21], Nguyen *et al* [19] also studied the use of universal dataset in unsupervised learning, and the use of both pictorial information and short annotations to annotate future images. Similarly, Hu *et al* [12] also employed Wikipedia and Wordnet as external semantic knowledge base in their study of short text clustering. The use of universal data set is not limited to work on short texts. Hu *et al* [13], as well, found the use of such semantic knowledge base effective in clustering long text documents.

3. PROBLEM SPECIFICATION

To enhance readability, we herein state that in the rest of this paper, mathematical values are denoted in bold font if they are matrices or vectors and in plain font if they are scalars.

3.1 Problem Analysis

Short texts are usually text documents that consist of tens of or up to a hundred words, which is usually sparser and noisier than long text documents. This sparseness and noise poses a great challenge for classification tasks on such documents, and leads to unsatisfactory classification accuracy in various short text classification tasks from document domain taxonomy to content disambiguation.

In recent studies, *Universal Knowledge Base (UKB)* is more and more used in short text classification and other learning tasks, to provide intra-class relations by discovering the underlying latent topics within short texts [1,21]. In these approaches, the short texts are enhanced by the "latent topics" or "substantial definitions" discovered from the UKB.

However, in UKB-related researches, scholars often used a manually selected UKB, i.e. the UKB used in the experiments are carefully selected so that the content is close to the short texts in the classification tasks. This limits the use of UKB approaches in short text classification, because in most real-world cases, we only have access to an unselected UKB (e.g. Wikipedia) from which auto-

[2] In our work, however, for the sake of readability and convenience of explanation, we take short text classification problem as a perfect example.

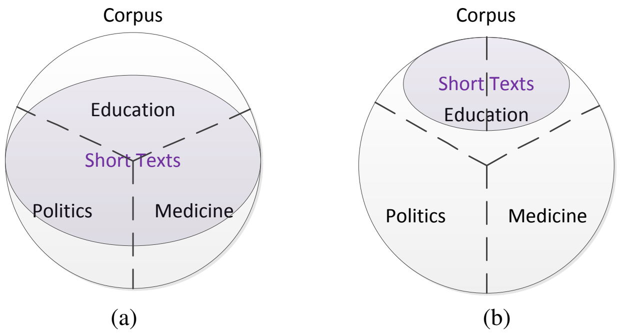


Figure 1: A typical situation where selection of UKB documents is needed

matically selecting semantically relevant information is often costly, if not intractable, given the size of the UKB. On the other hand, directly using such UKB will be less helpful to our classification tasks where the latent topics may be more irrelevant.

3.2 Instincts of the Proposed Framework

To cope with the problems of classifying short texts with UKB, we come up with the idea to select *relevant* documents from the UKB before state-of-art methods are run, in order to make the most of a unselected UKB, and refine the classification results on the selected UKB as well.

This idea can be easily be explained with the example shown in Figure 1. In this example, the UKB (or in text-related context called "Corpus") contains 3 latent topics, namely *Education, Politics,* and *Medicine*, respectively. The work of previous researchers are often under the situation of Figure 1(a), where the UKB (denoted by $\mathcal{U}$ in the literature) covers exactly the latent topics needed to distinguish short texts (denoted by $\mathcal{S}$ in the literature). However, when dealing with the situation of Figure 1(b), where the corpus and its latent topics are the same while the short texts fit only in some of its latent topics, it incontrovertibly degenerate the boosted distinguishing power of the classifier trained on the short texts, compared to that scenario depicted in Figure 1(a).

This problem can be easily tackled down by tuning the number of latent topics of the UKB ($K^{\mathcal{U}}$), or using a multiple-granularity of topics to fit the given classification task [5]. However, both involve the adjustment of the parameter $K^{\mathcal{U}}$, which is very inefficient since the topic model of the whole UKB needs to be reestimated for many times to find a practical value, and classification accuracy is very sensitive to this parameter [21]. If, instead, a small portion of the UKB documents are selected to form a *related corpus* for the given classification task, then estimate a *related latent topic model* on the selected corpus, much time tuning and examining the effect of $K^{\mathcal{U}}$ can be saved, and the classification accuracy can be better enhanced. For instance, in Figure 1(b), the current $K^{\mathcal{U}}$ is 3 and a classification task is given to distinguish short texts related to *primary school* and *high school*. Instead of tuning $K^{\mathcal{U}}$ to a greater value, we can simply use the existing topic model to give an estimation of relatedness between each document in $\mathcal{U}$ and the task $\mathcal{S}$, by making the most of estimated parameters – topic-word distribution of each UKB latent topic, the proportion of each latent topic for each UKB document, etc. Even if for Figure 1(a), we can expect a better selection of the UKB can achieve a better classification accuracy due to the increased intra-class correlation and the decreased interclass correlation.

A brute-force method of estimating the relatedness between a UKB document and the short texts is the cosine measurement, which reflects the angle between documents when represented by term frequency (TF) or term frequency inverse document frequency (TF-IDF) vectors. We can compute the cosine measurement between each UKB document and short text, and rank the documents by their maximum relatedness with certain short text. This approach is straight-forward, however, as we will show in Section 5, it not only suffers from efficiency problem, but also is affected by sparseness and noises in the short texts.

Thus, we need to use a less sparse and less noisy representation of the short texts in order to achieve better selection of UKB documents. A direct instinct is using topic models. By estimating topic models for short texts, we can smoothen the noise and as well discover important terms in the short texts. One thing to notice is, that the topics estimated from short texts may still be sparse compared to that of the UKB. Hence we introduced an iterative framework to overcome this sparseness. A detailed discussion of UKB document selection will be covered in Subsection 4.2.

4. PROPOSED FRAMEWORK

In this section, we present the proposed framework which aims at building a classifier that classifies short text with latent topics learnt from actively acquired knowledge (documents) from very large-scale universal dataset. Figure 2 gives a brief illustration of this framework.

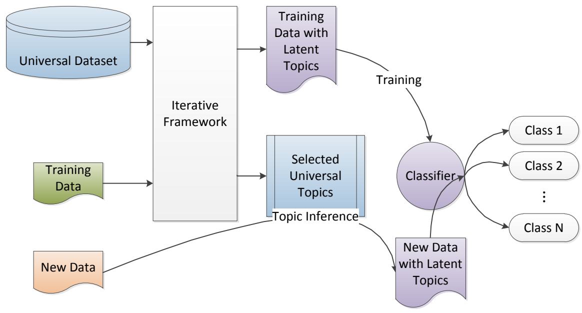
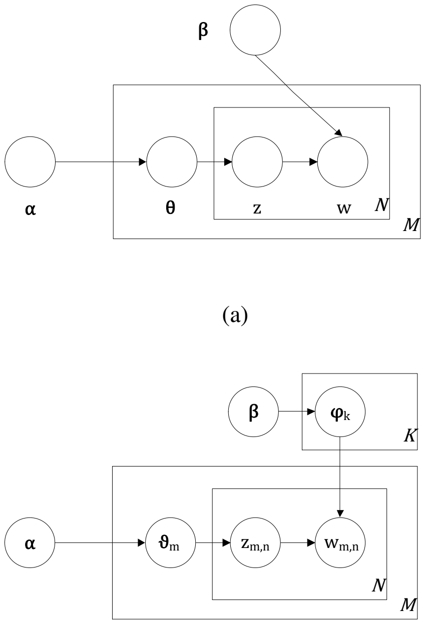


Figure 2: Brief Illustration of Proposed Framework

The framework can be roughly divided into three parts: (i) relevant document selection from the UKB, where a iterative sub-framework is used, (ii) short text data enrichment with selected documents, and (iii) building the classifier. Here, we will discuss in details the selection and enrichment parts, which are the basic components of the iterative sub-framework, while the construction of the classifier will be covered in Subsection 5.2.

4.1 Hidden Topic Analysis Models

Recently, many researchers have studied intensively the latent semantics in text documents, and among them two outstanding examples are Probabilistic Latent Semantic Analysis (pLSA) [11] and Latent Dirichlet Allocation (LDA) [2]. Both of these works provide a probabilistic insight on latent topics, where the text documents can be viewed as a generated production over a latent distribution controlled by a set of latent variables. In this work, we adopt LDA in our experiments of estimating latent topic models for text documents.



(a)

(b)

Figure 3: The structure of Latent Dirichlet Allocation

Latent Dirichlet Allocation is a generative model that models perfectly, but not limited to, text documents. In LDA, each document vector $\boldsymbol{w}$ in a corpus D is assumed generated by a three-step procedure: (i) Choose $N \sim \text{Possion}(\xi)$ (This parameter is not a critical part of the model); (ii) Choose $\boldsymbol{\theta} \sim \text{Dir}(\boldsymbol{\alpha})$; (iii) For each of the N words w_n, first choose a topic $z_n \sim \text{Multinomial}(\boldsymbol{\theta})$, then choose a word w_n from $p(w_n|z_n, \boldsymbol{\beta})$, a multinomial probability conditioned on the topic z_n. This procedure can be perfectly depicted by a three layer structure, shown in Figure 3(a).

From Figure 3(a), it is evident that LDA is a three-layer model. The outmost plate, called the "corpus plate", consists of two variables $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ that are shared among all documents in the corpus. $\boldsymbol{\alpha}$ is the parameter of the Dirichlet Distribution $\text{Dir}(\boldsymbol{\alpha})$, and $\beta_{ij} = p(w_j = 1|z_i = 1)$, i.e. the likelihood of the presence of a word w_j given a certain topic z_i. In practice, we note that without more prior knowledge of the corpus, $\boldsymbol{\alpha}$ is set the same value on all its dimensions, so we replace it with a scalar α in the context, and let the context determine its dimensionality. Similarly, $\boldsymbol{\beta}$ is often viewed as a matrix filled with a single fixed value β, while the probability distribution $p(w_j = 1|z_i = 1)$ is denoted by a variable $\boldsymbol{\varphi}_{ij}$, where $\boldsymbol{\varphi}_k \sim \text{Dir}(\beta)$ for each $k \in [1, K]$, which forms the "topic plate" and where K is the total number of latent topics and assumed known(See Figure 3(b), the Bayesian network of LDA). The second plate is the "data collection plate", where the topic portion ϑ_m are drawn from the Dirichlet distribution for each document $\boldsymbol{w}_m$, $m \in [1, M]$. The third plate is called the "document plate", where each of the N_m words in document $\boldsymbol{w}_m$ is chosen conditioned on both the word-topic marginal distribution $\boldsymbol{\varphi}$, and the chosen topic $z_{m,n}$, which is chosen from a multinomial distribution with parameter $\boldsymbol{\vartheta}_m$. From Figure 3(b) it is easy to derive the following joint distribution of all known and hidden variables given the Dirichlet parameters:

$$p(\boldsymbol{w}_m, \boldsymbol{z}_m, \boldsymbol{\vartheta}_m, \boldsymbol{\Phi}|\alpha, \beta) = p(\boldsymbol{\Phi}|\beta) \prod_{n=1}^{N_m} p(w_{m,n}|\boldsymbol{\varphi}_{z_{m,n}}) p(z_{m,n}|\boldsymbol{\vartheta}_m) p(\boldsymbol{\vartheta}_m|\alpha)$$

By by integrating over $\boldsymbol{\vartheta}_m$, $\boldsymbol{\Phi}$ and summing over $\boldsymbol{z}_m$, we can obtain the following likelihood of a document $\boldsymbol{w}_m$ as follows:

$$p(\boldsymbol{w}_m|\alpha, \beta) = \int\int p(\boldsymbol{\vartheta}_m|\alpha) p(\boldsymbol{\Phi}|\beta) \prod_{n=1}^{N_m} p(w_{m,n}|\boldsymbol{\vartheta}_m, \boldsymbol{\Phi}) d\boldsymbol{\Phi} d\boldsymbol{\vartheta}_m$$

Finally, the likelihood of the whole data collection $\boldsymbol{\mathcal{W}}$ is given by calculating the product of all the likelihoods of the documents $\{\boldsymbol{w}_m\}_{m=1}^{M}$:

$$p(\boldsymbol{\mathcal{W}}|\alpha, \beta) = \prod_{m=1}^{M} p(\boldsymbol{w}_m|\alpha, \beta) \tag{1}$$

Estimating LDA's parameters by directly and accurately maximizing the likelihood in Equation 1, unfortunately, is intractable. However, multiple alternatives exists, such as variational methods [2], expectation propagation [18], and Gibbs Sampling [10]. Gibbs Sampling is a special case of Markov chain Monte Carlo (M-CMC) [8], which, in complicated parameter estimation tasks, often yields an algorithm that is easy to implement, requires little memory, and is competitive in speed and performance with existing algorithms [9].

An intensive discussion of Gibbs Sampling can be found in the technical report [10]. Hereby, we provide only the most crucial formula to implement Gibbs Sampling for parameter estimation of LDA. Let $\boldsymbol{w}$ be the vector of all words, and $\boldsymbol{z}$ be the corresponding vector of topic assignment, of the whole collection $\boldsymbol{\mathcal{W}}$. The topic assignment for a particular word depends on the current topic assignment of all the other word positions. More specifically, the topic assignment of a particular word t is sampled from the following multinomial distribution[3]:

$$p(z_i = k|\boldsymbol{z}_{\neg i}, \boldsymbol{w}) = \frac{n_{k,\neg i}^{(t)} + \quad \beta_t}{\left[\sum_{v=1}^{V} n_k^{(v)} + \beta_v\right] - 1} \frac{n_{m,\neg i}^{(k)} + \alpha_k}{\left[\sum_{j=1}^{K} n_m^{(j)} + \alpha_j\right] - 1} \tag{2}$$

where $n_{k,\neg i}^{(t)}$ is the number of times the word t is assigned to topic k except the current assignment; $\sum_{v=1}^{V} n_k^{(v)} - 1$ is the total number of words assigned to topic k except the current assignment; $n_{m,\neg i}^{(k)}$ is the number of words in document m assigned to topic k except the current assignment; and $\sum_{j=1}^{V} n_m^{(j)} - 1$ is the total number of

[3] Here we note that for preciseness, the subscripts for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ in the original Gibbs Sampling formula are reserved. This does not contradict the claim hereinabove that LDA hyperparameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are usually set the same value for each dimension.

words in document m except the current word t. After finishing Gibbs Sampling, two matrices $\mathbf{\Phi}$ and $\mathbf{\Theta}$ are calculated as follows:

$$\varphi_{k,t} = \frac{n_k^{(t)} + \beta_t}{\sum_{v=1} V n_k^{(v)} + \beta_v} \tag{3}$$

$$\vartheta_{m,k} = \frac{n_m^{(k)} + \alpha_k}{\sum_{j=1} K n_m^{(j)} + \alpha_j} \tag{4}$$

After estimating the matrices $\mathbf{\Phi}$ and $\mathbf{\Theta}$ for the current collection $\boldsymbol{\mathcal{W}}$, we can also use the estimated model to infer the topic assignment of a new collection $\underline{\boldsymbol{\mathcal{W}}}$. Similar to Equation 2, The topic assignment of a particular word t in $\underline{\boldsymbol{w}}$ (vector of all words in the new collection) depends on the current topics of all the other words in $\underline{\boldsymbol{w}}$ and the topics of all the other words in $\boldsymbol{w}$ as follows:

$$\begin{aligned} & p(\underline{z}_i = k | \underline{\boldsymbol{z}}_{\neg i}, \underline{\boldsymbol{w}}; \boldsymbol{z}, \boldsymbol{w}) = \\ & \frac{n_k^{(t)} + \underline{n}_{k,\neg i}^{(t)} + \beta_t}{\left[\sum_{v=1}^{V} n_k^{(v)} + \underline{n}_k^{(v)} + \beta_v\right] - 1} \frac{n_{\underline{m},\neg i}^{(k)} + \alpha_k}{\left[\sum_{j=1}^{K} n_{\underline{m}}^{(j)} + \alpha_j\right] - 1} \end{aligned} \tag{5}$$

After Gibbs Sampling on the estimated model, the topic assignments of the documents in $\underline{\boldsymbol{W}}$, the matrix $\underline{\mathbf{\Phi}}$, is calculated as follows:

$$\vartheta_{\underline{m},k} = \frac{n_{\underline{m}}^{(k)} + \alpha_k}{\sum_{j=1}^{K} n_{\underline{m}}^{(j)} + \alpha_j} \tag{6}$$

4.2 Selecting Relevant Documents from UKB

Intuitively, to select relevant documents from UKB, utilizing searching engines is a straight-forward choice. However, the actual performance of this approach depends on selection of keywords, preprocessing procedures of the searching engine(stemming, stop-word removal, etc) and probably most importantly, scoring and ranking policy. Besides, as is mentioned hereinabove, this approach can be very slow that cannot meet the need of online applications. Another approach, the cosine measurement, as mentioned in Subsection 3.2, may be affected severely by sparseness and noise of the short texts, and is too time-consuming.

Accordingly, we propose a more independent and robust method based on topic models and the Kullback-Leibler divergence [15] to select relevant UKB documents. The K-L divergence is a nonsymmetric measure of the difference between two probability distributions, which is defined as

$$D_{\mathrm{KL}}(P|Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$$

K-L divergence measures the expected number of extra bits required to code samples from P when using a code based on Q, rather than using a code based on P, i.e. information gain from Q to P. In this approach, we first estimate topics for both the original short text training data and the universal knowledge base, and denote the topics $T_i^{\mathcal{S}}$ and $T_j^{\mathcal{U}}$, respectively, where $i = 1..K^{\mathcal{S}}$ and $j = 1..K^{\mathcal{U}}$. Then for each pair $(T_i^{\mathcal{S}}, T_j^{\mathcal{U}})$, we calculate the K-L divergence $D_{\mathrm{KL}}(T_i^{\mathcal{S}}|T_j^{\mathcal{U}})$, which, the smaller, indicates that $T_i^{\mathcal{S}}$ is more similar to $T_j^{\mathcal{U}}$ or a part of $T_j^{\mathcal{U}}$, or in this context, the universal topic $T_j^{\mathcal{U}}$ "covers" the short text topic $T_i^{\mathcal{S}}$. An example that may help with understanding the K-L divergence in this context is as follows. Let $T_{i_1}^{\mathcal{S}}$={study: 0.3; learn: 0.7}, $T_{i_2}^{\mathcal{S}}$={research: 0.6; study: 0.4}, $T_j^{\mathcal{U}}$={study: 0.3; learn: 0.699; research: 0.001}, then the K-L divergences will be $D_{\mathrm{KL}}(T_{i_1}^{\mathcal{S}}|T_j^{\mathcal{U}}) = 0.3 \times \log(0.3/0.3) + 0.7 \times \log(0.7/0.699) = 6.21 \times 10^{-4}$, $D_{\mathrm{KL}}(T_{i_2}^{\mathcal{S}}|T_j^{\mathcal{U}}) = 0.6 \times \log(0.6/0.001) + 0.4 \times \log(0.4/0.3) = 1.72$. This meets our intuition that $T_{i_1}^{\mathcal{S}}$ and $T_j^{\mathcal{U}}$ are similar in that they are probably related to education, while $T_{i_2}^{\mathcal{S}}$ may focus more on scientific research. Note that we hereby assume that all words that occur in the short text occur in the UKB, or else the K-L divergence can be undefined.[4]

After $K^{\mathcal{S}} \times K^{\mathcal{U}}$ K-L measurements, we obtain a matrix $(\boldsymbol{M}_{\mathrm{KL}})_{K^{\mathcal{S}} \times K^{\mathcal{U}}}$. Then for each document $\boldsymbol{w}_i$ in the UKB, we have its topic assignment $\boldsymbol{\vartheta}_i$ (where each dimension $\boldsymbol{\vartheta}_{ij} = P(\boldsymbol{z}_j|\boldsymbol{w}_i)$), and $\boldsymbol{\vartheta}_i$ is an $1 \times K^{\mathcal{U}}$ column vector. Then the document's relevance with the short text data R_i is given as

$$\hat{\boldsymbol{\vartheta}}_i = \boldsymbol{M}_{\mathrm{KL}} \cdot \boldsymbol{\vartheta}_i$$

$$R_i = \min_j \hat{\boldsymbol{\vartheta}}_{ij}$$

To illustrate this, we give an example as follows. Suppose we have two short text topics $T_1^{\mathcal{S}}$ and $T_2^{\mathcal{S}}$, two UKB topics $T_1^{\mathcal{U}}$ and $T_2^{\mathcal{U}}$, where $T_1^{\mathcal{S}}$ is identical with $T_1^{\mathcal{U}}$, and $T_2^{\mathcal{S}}$ is identical with $T_2^{\mathcal{U}}$. We assume that $T_1^{\mathcal{U}}$ and $T_2^{\mathcal{U}}$(and thus $T_1^{\mathcal{S}}$ and $T_2^{\mathcal{S}}$) are irrelevant so that $X = D_{\mathrm{KL}}(T_1^{\mathcal{U}}|T_2^{\mathcal{U}})$ and $Y = D_{\mathrm{KL}}(T_2^{\mathcal{U}}|T_1^{\mathcal{U}})$ are very large positive numbers. Thus we have

$$\boldsymbol{M}_{\mathrm{KL}} = \begin{bmatrix} 0 & X \\ Y & 0 \end{bmatrix}$$

Hence for three UKB documents whose $\boldsymbol{\vartheta}$-vectors are $\boldsymbol{\vartheta}_1 = [1\,0]^T$, $\boldsymbol{\vartheta}_2 = [0\ 1]^T$ and $\boldsymbol{\vartheta}_3 = [0.5\ 0.5]^T$, we can calculate the relevance respectively as $R_1 = \min\{0, Y\} = 0, R_2 = \min\{X, 0\} = 0, R_3 = \min\{0.5X, 0.5Y\}$. These R_i's represent the best possible single-topic relevance of the UKB document i with the short text topics. Note that in this relevance measurement we consider only the maximum dimension of the product vector as we are looking for UKB documents that can provide more discriminative context for short texts. We then sort UKB documents by this relevance value ascentantly and choose the top documents as selected UKB documents. Note that we use only the relative value of R_i's while the absolute value hardly has any physical meaning.

In classification applications, in order to provide better interclass distinction, the fore stated measurement to select UKB documents are employed to each class of short text training data to avoid interclass latent topics, which is much more common in short texts than that of long documents mainly due to their sparseness. Thus the actual total number of topics estimated for the short texts is $|\mathcal{L}| \times K^{\mathcal{S}}$, where $|\mathcal{L}|$ is the cardinality of class label set, i.e. the total number of class labels. However, in this context, we will continue to use $K^{\mathcal{S}}$ for latent topic count in short texts, but this number is assigned to each class of short texts in the training data.

4.3 Enriching Short Texts with Selected Data

In the work of Phan *et al* [21], a similar enrichment task is done by adding semantically meaningless topic terms to the original short

[4] In our experimental practice, words that only appear in short text vocabulary are simply ignored when K-L divergence is calculated. This is sound since we by no means have any other knowledge about these words except from the UKB.

text data to provide concurrence, reduce data sparseness and generate similar distribution of training and testing set of short texts. Simply stated, Phan *et al*'s approach is: first estimate a topic model of the universal dataset, then infer both training and testing set of short texts in this topic model, the obtained topic proportions are then used to add semantically meaningless topic terms(for instance, if the topic proportion of topic i in a short text is in $[0.05, 0.10)$), and finally the classifier is built on such enriched data.

In our evaluation, we simplified the enrichment of short texts by simply appending inferred $\underline{\boldsymbol{\vartheta}}$ vectors (See Equation 6) to the original short text data. Similarly to Phan *et al*'s work, we employed a Maximum Entropy classifier, for which the discretization of $\underline{\boldsymbol{\vartheta}}$ is essential. For each short text document $\underline{\boldsymbol{w}}$, the enriched short text is given as follows (written in a MatLab fashion):

$$\underline{\boldsymbol{w}}' = [\underline{\boldsymbol{w}}; \lfloor \gamma \cdot \underline{\boldsymbol{\vartheta}} \rfloor]$$

Here, γ is a magnification factor, which we will fix in all our experiments.

Until now, we have introduced the two key components, UKB document selection and short text enrichment, of the iterative sub-framework, which is depicted in Figure 4. With the input training set of short text data, we first estimate a topic model to select UKB documents, and then use the topic model of the selected documents to enrich the short text. In the next iteration, the enriched short texts are used to help select UKB documents, and thus a iterative refinement of both the training set of short texts and (the topic model of) the selected UKB documents is performed. During such iterations, the short texts tend to converge to a more similar feature space as the selected UKB documents. After enough iterations, the sub-framework outputs the refined training set of the short texts as well as the topic model of the refined selected UKB documents, and the latter can be used to enrich new unseen data for classification. The performance of this sub-framework will be examined in details in Section 5.

5. EXPERIMENTS & EVALUATION

5.1 Experiment Dataset

The dataset used in this work is collected by Xuan-Hieu Phan [21], which consists of two parts: *Short Texts* and *Corpus*. The task is to classify the *Short Texts* given *Corpus* as the universal knowledge base. The dataset is prepared based on a set of topics and topic-oriented keywords. Topics include "Arts", "Business", and "Culture", etc. As for keywords, take topic "Business" as an example, keywords vary from *advertising, e-commerce* to *finance*.

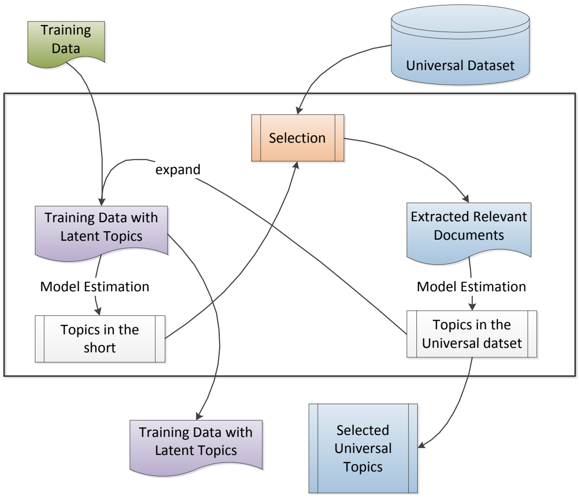

Training Data
Universal Dataset
Selection
expand
Training Data with Latent Topics
Extracted Relevant Documents
Model Estimation
Model Estimation
Topics in the short
Topics in the Universal datset
Training Data with Latent Topics
Selected Universal Topics

Figure 4: Iterative sub-framework

This sub-framework works in the following steps:

1. Initially, $K_0^{\mathcal{U}}$ topics are estimated for the UKB
2. $K^{\mathcal{S}}$ topics are estimated for the input training data $\mathcal{S}$
3. η percent[5] of the UKB documents are selected by their rank given with regard to each class
4. $K^{\mathcal{U}}$ topics are estimated for the selected corpus
5. The original short texts are assigned topics of the selected corpus, and enriched to form the input training data of the next iteration[6] (Goto Step 2)
6. After several iterations, we obtain enriched training data and a topic model for selected corpus in the last iteration, and the latter can be used to infer and enrich new short text data

Short Texts. Short texts are collected from searching snippets on the Google searching engine. Specifically, by using search engines we can typically get *searching snippets* that consist of a URL, a short title, and a short description. The short texts used in this dataset is obtained by performing search transactions with predefined topic-oriented keywords as above, and the top 20 or 30 ranked searching snippets(specifically the short descriptions) are then collected. Then the snippets are labeled based on the topic from which the keyword was drawn.

Corpus. The corpus is prepared by crawling Wikipedia pages corresponding to above mentioned keywords and relevant pages by following outgoing hyperlinks with JWikiDocs [23][7].

Some of the statistics of the dataset is given in Table 1.

[5] This percentage is dependant on the UKB used, and the specific classification task. However, as we will show in Subsection 5.3, η is often small and thus can be determined efficiently in practice.

[6] We expect this iterative sub-framework can improve classification accuracy, by boosting the accuracy of topics estimated in the short texts, and then those in the selected corpus in turn. More specifically, the semantical relation provided by the topic model of the selected corpus can provide help (for example word sense disambiguation) for short text topic estimation, by providing more concurrence among words of the same latent topic in them. The enriched topic vectors, though did not directly influenced the selection of UKB documents, their positive effect on the short text topics will lead to a more specific K-L divergence based selection. We find this idea somehow similar to that of [3]. However, we note that this iterative framework shall not show expected power when $K^{\mathcal{S}} = 1$, since under that situation, the topic estimated from the enriched short text shall remain the same as the previous iteration (if we ignore the enriched words that won't help in the K-L measurement).

[7] JWikiDocs: http://jwebpro.sourceforge.net

Table 1: Summary of the Dataset

Short Texts			
No.	Domain	# Training Data	# Test Data
I	Business	1,200	300
II	Computers	1,200	300
III	Culture-Arts-Entertainment	1,880	330
IV	Education-Science	2,360	300
V	Engineering	220	150
VI	Health	880	300
VII	Politics-Society	1,200	300
VIII	Sports	1,120	300
	Total	10,060	2,280

Corpus
Raw Data:3.5GB;$\|Docs\| = 461,177$
Preprocess: Duplicate documents, HTML tags,navigation links, stop and rare (thresold = 30) words are removed
Final Data: 240MB; $\|Docs\| = 71,986$;$\|Vocabulary\| = 60,649$

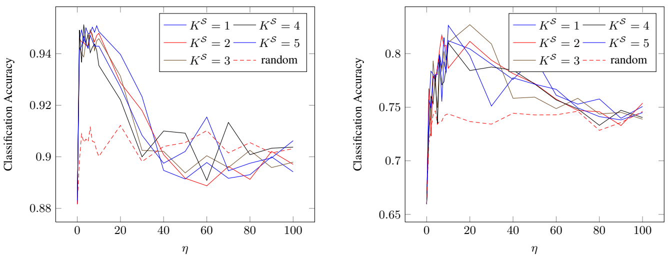
5.2 Experiment Methodology

We designed two experiments to evaluate short text classification accuracy both within one iteration and over several iterations. The first experiment is designed to examine the classification accuracy on multiple classification tasks with the same UKB, and the second experiment is for examining the effectiveness of the iterative framework.

In the first experiment, we set the percentage of UKB documents used to help enrich short text expressions, η (varies in $\{1, 2, \ldots, 10, 20, \ldots, 100\}$); the number of topics estimated in each short text class (varies from 1 to 5) and that in selected corpus (varies from $|\mathcal{L}|$ to $5|\mathcal{L}|$); and the number of short text classes $|\mathcal{L}|$ (varies from 2 to 8), in order to examine both scenarios depicted in Figure 1, i.e. use the proposed framework to deal with multiple classification tasks, which cover different scope of the UKB. In this experiment, we simplified the multiple possible combinations of classes, and used classes I-II to I-VIII for simplicity, as numbered in Table 1. The control group is designed similar to Phan *et al*'s work [21], where no selection or iterations are involved, and the "selected corpus" consists of the same η percent of randomly chosen documents from the UKB.

The second experiment uses some of the conclusions of the first experiment (See Subsection 5.3), where η is fixed to 10, and $K^{\mathcal{U}}$ is fixed to be identical to $|\mathcal{L}|$. We run this experiment on all 8 classes of short text classes, and observe whether an improvement exists in classification accuracy, when we run the experiment for up to 50 iterations.

In all the above experiments, some settings are shared. $K_0^{\mathcal{U}}$ is fixed to 50 for our proposed method, and the hyperparameters α and β in LDA are fixed to 0.5 and 0.1, respectively. Each setting of the experiments is run for 10 independent runs, and the final classification accuracy is given by averaging over the 10 runs. In LDA, while estimating topic models for the UKB, the selected corpus, and the short texts, we observed that 200 iterations is enough and fixed this value. Similarly, the number of iterations of LDA topic inference is fixed to 100. Our experiment is implemented under a Java environment, where the LDA [24][8] and the MaxEnt classifier [22][9] are both implemented by Xuan-Hieu Phan (JGibbLDA with Cam-Tu Nguyen).



(a) $|\mathcal{L}| = 2$

(b) $|\mathcal{L}| = 8$

Figure 5: Classification Accuracies of Parameter $K^{\mathcal{U}} = |\mathcal{L}|$

5.3 Evaluation

Figure 5 shows clearly that our proposed framework of actively acquiring relevant knowledge from the UKB to boost short text classification not only beats the performance of that with randomly chosen documents, but also its peak accuracy is much higher than that of using the whole UKB. This proves our instincts that some knowledge of the UKB are more useful than the others or even the whole UKB in boosting short text classification accuracy in both scenarios that the UKB covers exactly the same domain of the short text (Figure 1(a) and Figure 5(b)), and that short texts are only fitted in a subset of UKB's domains (Figure 1(b) and Figure 5(a)). It also proves that our proposed framework is able to discover *this relevant knowledge* from the UKB, and use it to improve short text classification performance. We note here that the selection of the parameter η is quite simple — η is usually quite small[10], and a rough study on the given UKB and classification task when this framework is adopted elsewhere should be easy and efficient.

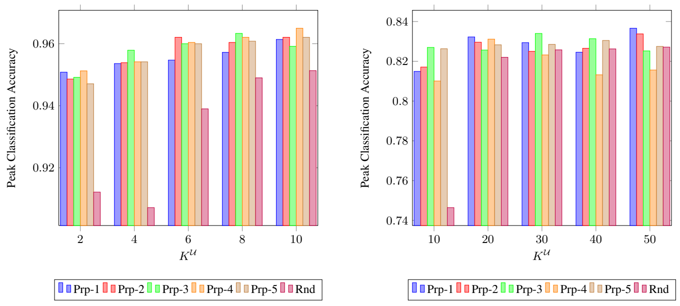
From Figure 6, we can clearly see that our proposed framework is relatively invariant to the change of $K^{\mathcal{S}}$, i.e. the variance of $K^{\mathcal{S}}$ affects little on the classification accuracy of the proposed framework, so in subsequent experiments, we fix this parameter to 1. Moreover, our proposed method is less sensitive to the parameter $K^{\mathcal{U}}$, which leads to the advantage that when this framework is employed, there is no much need to tune the factor $K^{\mathcal{U}}$ carefully, which is very time-consuming and inconvenient. For simplicity, setting $K^{\mathcal{U}}$ to $|\mathcal{L}|$ is enough for the framework to achieve a relatively high accuracy. Having observed the detailed experimental data carefully, we explain this phenomenon as follows. The essence of the use of UKB in improving short text classification performance is making use of the rich semantic context of the UKB to unite the sparse short texts of the same class, and provide semantical

[8] JGibbLDA: http://jgibblda.sourceforge.net/

[9] JMaxEnt is part of the project JTextPro. JTextPro: http://jtextpro.sourceforge.net/

[10] In this work, the peak classification accuracy occurs at around $\eta = 10$, for both the proposed method and randomly selected documents from the UKB. Furthermore, in our primitive experiments with another UKB (corpus) randomly sampled from the entire Wikipedia, the peak accuracy occurs at around $\eta = 5$ or less, which make it easier to choose this parameter when the UKB is more "universal".

[11] In this figure, "Prp-x" denotes the bar corresponding to the proposed framework, with parameter $K^{\mathcal{S}} = x$; "Rnd" denotes the random control group.



(a) $|\mathcal{L}| = 2$

(b) $|\mathcal{L}| = 8$

Figure 6: Peak Classifications Accuracies of Tuning $K^{\mathcal{U}}$ [11]

background information to distinguish short texts among the classes. The previous approaches (mainly [21]) tackles down this problem by fine-tuning on the parameter $K^{\mathcal{U}}$, which determines that whether the useful latent topics are well distinguished from each other, and also that the useful ones are well distinguished from the useless ones. When enriching short texts, the topic inference step implicitly "selects" topics that are useful in the UKB by assigning each short text document more weights on them, and "discards" the useless ones vice versa. Thus, it is important that this parameter is well tuned. In our proposed framework, however, the most useful documents, which contains the useful latent topics to help distinguish short texts amongst classes, are selected from the UKB explicitly, so directly setting the parameter to the number of total classes is enough to distinguish those "useful" topics in the selected corpus. [12]

Figure 7 is an evident illustration that our proposed framework can reach competitive performance with the previous work, which reaches its peak performance at the cost of tuning the parameter $K^{\mathcal{U}}$. Combined with the discussion of choice of η discussed above, this proves that our proposed method can reduce the requirement of time in previous works on reaching a relatively high classification accuracy by selecting a small portion of the UKB, as well as fixing the parameter $K^{\mathcal{U}}$ to an value that is more reasonable. What is more, this proposed method performs significantly better than previous works in scenarios depicted in Figure 1(b), and thus is capable of improving multiple classification tasks with *only one* universal knowledge base that covers knowledge of all.

Figure 8, it is evident that the classification accuracy is improved, though slightly, over the iterations. In fact, the major motivation of the iterative approach is the low intra-class cohesion of the short text caused by its sparsity (See Figure 4. Thus, we can expect a more significant improvement on more sparse data, using this iterative approach.

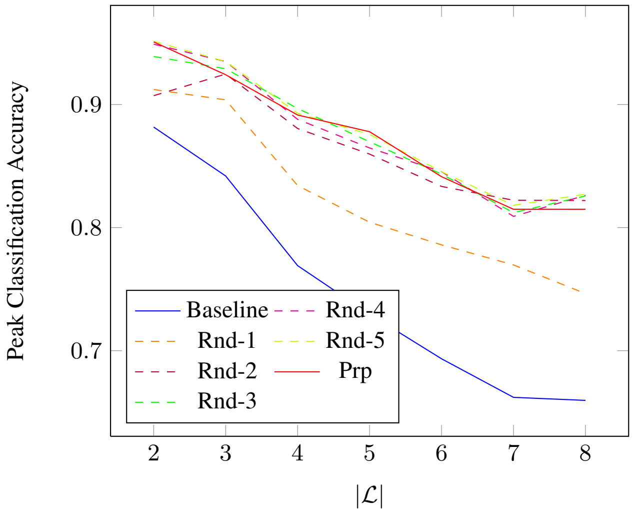


Figure 7: Peak Classification Accuracy Reached under Multiple Settings of $|\mathcal{L}|$[13]

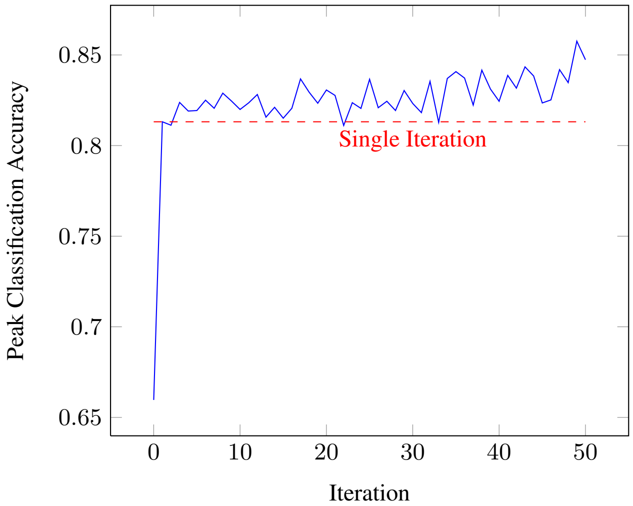


Figure 8: Classification Accuracies over Iterations
(Settings in each iteration: $|\mathcal{L}| = 8$, $K^{\mathcal{S}} = 3$, $K^{\mathcal{U}} = 3|\mathcal{L}|$, and η=10)

[12] We note here that (i) when $K^{\mathcal{U}}$ is set too large, both the performance of the proposed framework and previous work will suffer, for that one evident latent topic may be divided into multiple obscure ones, and this again gives rise to data sparsity and weakens the distinguishing assistance; and (ii) in our observation, the variance of $K^{\mathcal{U}}$ does not have a major effect on the choice of the parameter η to achieve the peak accuracy.

[13] In this figure, "Rnd-x" denotes the bar corresponding to the random control group, with parameter $K^{\mathcal{U}} = x|\mathcal{L}|$; "Prp" denotes the proposed framework with parameters $K^{\mathcal{S}} = 1$ and $K^{\mathcal{U}} = |\mathcal{L}|$; "Baseline" stands for the classification accuracies with the raw short texts.

6. CONCLUSION & DISCUSSION

In this paper, we presented a general framework to improve classification accuracy for, but not limited to, short text documents. Besides its universality for sparse data classification tasks, our proposed framework can reduce data sparsity with *only one* universal knowledge base for tasks of one type of classification tasks (e.g. short texts or images), since it can make the most of the UKB by actively acquiring the most relevant knowledge from it to enrich the sparse data. This approach can not only reduce classification error but also simplify the choice of critical parameters, which in previous works are tuned painstakingly. Furthermore, we devised an iterative framework to further refine the classification accuracy, given the semantic information of both the sparse data and the UKB.

In the development of this work, we see two major problems that could be directions of future works. One is the application of the framework to unsupervised learning tasks (e.g. short text clustering), and the other is the use of such framework on multiple UKBs with different modalities, such as text, audio, and images. The former, as we were in the progress of the experiments presented in this paper, posed great challenge on how to properly make use of the poor semantical information provided in the sparse data. This draws the problem again back to the fundamental problem of short text classification: sparsity and noise. Since short text documents are sparse and noisy, we observed in our primitive works that if the topic models are not estimated class-wise on short texts, many topics would tend to be a mixture of the real latent semantics. The latter problem, from our perspective, is more innovative and challenging in the development of machine learning algorithms, and even the whole field of artificial intelligence. Improving short text classification with UKB, the method itself, is itself a transfer learning problem [20] in our opinion. We can consider the UKB the source domain, while view short texts as the target domain, since they differ drastically on word distribution, document distribution, and even vocabulary coverage. Inspired by the excellent transfer learning algorithm, *self-taught learning* [26], which the author argued a highly probable manner of learning that human beings adopt naturally, we here pose the question: Is it possible to make use of knowledge from UKBs of different modalities to accomplish certain learning task, probably with self-taught learning method, like human beings learn vocabularies from visual and acoustic sensations? Hopefully these could be inspiration for future researchers.

7. REFERENCES

[1] S. Banerjee, K. Ramanathan, and A. Gupta. Clustering short texts using wikipedia. In *Proc. ACM SIGIR*, 2007.

[2] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3, 2003.

[3] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proc. Computational Learning Theory*, 1998.

[4] D. Bollegala, Y. Matsuo, and M. Ishizuka. Measuring semantic similarity between words using web search engines. In *WWW 2007 / Track: Semantic Web*, pages 757–766, May 2007.

[5] Mengen Chen, Xiaoming Jin, and Dou Shen. Short text classification improved by learning multi-granularity topics. In *Proc. IJCAI*, 2011.

[6] Corinna Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20, 1995.

[7] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*. John Wiley & Sons, 1973.

[8] Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE PAMI*, 6:721–741, 1984.

[9] Thomas L. Griffiths and Mark Steyvers. Finding scientific topics. In *Proc. of the National Academy of Science of USA*, 2004.

[10] Gregor Heinrich. Parameter estimation for text analysis. Technical report, 2005.

[11] T. Hofmann. Probabilistic latent semantic analysis. In *Proc. Uncertainity in Artificial Intelligence*, 1999.

[12] Xia Hu, Nan Sun, Chao Zhang, and Tat-Seng Chua. Exploiting internal and external semantics for the clustering of short texts using world knowledge. In *Proc. CIKM*, Nov. 2009.

[13] Xiaohua Hu, Xiaodan Zhang, Caimei Lu, E.K. Park, and Xiaohua Zhou. Exploiting wikipedia as external knowledge for document clustering. In *Proc. KDD*, 2009.

[14] T. Joachims. A probabilistic analysis of the rocchio algorithm with tfidf for text categorization. In *Int. Conf. Machine Learning*, 1997.

[15] S. Kullback and R. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, pages 79–86, 1951.

[16] D. D. Lewis. Reuters-21578 text categorization test collection. http://www.research.att.com/ lewis/reuters21578.html, 1997.

[17] D. Liu and J. Nocedal. On the limited memory bfgs method for large-scale optimization. *Mathematical Programming*, 6:503–528, 1989.

[18] Thomas Minka and John Lafferty. Expectation-propagation for the generative aspect model. In *Proc. Uncertainty in Artificial Intelligence*, 2002.

[19] C. Nguyen and T. Tokuyama. Bridging semantic gaps in information retrieval: Context-based approaches. *ACM VLDB 10*, September 2010.

[20] S. Pan and Q. Yang. A survey on transfer learning. *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, 2009.

[21] X. Phan, L. Nguyen, and S. Horiguchi. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In *WWW 2008 / Refereed Track: Data Mining - Learning*, pages 91–100, April 2008.

[22] Xuan-Hieu Phan. Jtextpro: A java-based text processing toolkit. http://jtextpro.sourceforge.net/, 2006.

[23] Xuan-Hieu Phan. Jwikidocs: A java-based wikipedia crawling toolkit. http://jwebpro.sourceforge.net/, 2007.

[24] Xuan-Hieu Phan and Cam-Tu Nguyen. Jgibblda: A java implementation of latent dirichlet allocation (lda) using gibbs sampling for parameter estimation and inference. http://jgibblda.sourceforge.net/, 2006.

[25] CMU World Wide Knowledge Base (Web->KB) project. Webkb datasets. http://www.cs.cmu.edu/ webkb/.

[26] R. Raina, A. Battle, H. Lee, B. Packer, and A. Ng. Self-taught learning: Transfer learning from unlabeled data. In *Proceedings of the 24th International Conference on Machine Learning*, 2007.