What Is the Basic Semantic Unit of Chinese Language? A Computational Approach Based on Topic Models

Qi Zhao[1], Zengchang Qin[1,2], and Tao Wan[2]

[1] Intelligent Computing and Machine Learning Lab,
School of Automation Science and Electrical Engineering,
Beihang University, Beijing, China
[2] Robotics Institute, Carnegie Mellon University, USA
zhaoqi1861@gmail.com, {wantao,zcqin}@andrew.cmu.edu

Abstract. Chinese language has been generally regarded as a Subject-Verb - Object (SVO) language and the basic semantic unit is the Chinese word that is usually consisted by two or more Chinese characters. However, word-centered structure of Chinese language has been controversial in linguistics. Some recent research in computational linguistics in Chinese language suggests that the character-based models perform better than the word-based models in some applications such word segmentation. In this paper, the word-based topic models and the character-based models are tested for modeling Chinese language, respectively. By empirical studies, we demonstrated the effectiveness of using Chinese characters as the basic semantic units. These two models have close performance in text classifications while the character-based model has a better quality in language modeling and a much smaller vocabulary. By testing on a bilingual corpus, three independent topic models based on Chinese words, Chinese characters and English words are trained and compared to each other. we verify the capability of topic models in modeling semantics by experiments across Chinese and English. The classification accuracy can also be boosted up by aggregating the classification results from the three independent topic models.

1 Introduction

Most of the Natural Language Processing (NLP) research are focus on the English language and the techniques used in study English can be easily extended to other alphabetic languages. As to Chinese, though with the most users in the world, because of the tower of Babel, the study has lagged behind. Most of researchers study Chinese by focusing on segmentation, tagging and parsing [19,16]. Little work has been done to understand the semantic structure of the Chinese language [23].

In linguistic typology, subject-verb-object (SVO) is a sentence structure where the subject comes first, the verb second, and the object third. Languages may be classified according to the dominant sequence of these elements. SVO the second most common order found in the world, after SOV, and together, they account for more than 85% of the world's languages [14,25]. Unlike English, Chinese has a distinct morphology. Thus, some theories that applied in English cannot be directly used in studying Chinese. It is even controversial that the Chinese is a SVO language [22]. In the Chinese writing system, the characters are monosyllabic, each usually corresponding to a spoken syllable

M. Kanazawa et al. (Eds.): MOL 12, LNAI 6878, pp. 143–157, 2011.

with a basic meaning. However, although Chinese words may be formed by characters with basic meanings, a majority of words in Mandarin Chinese require two or more characters to write and have the meaning that is different from but somehow related to the characters they are made from. The morpheme of Chinese language is the Chinese character and Chinese text is written without word boundaries. Effectively recognizing Chinese words is like recognizing collocations in English, substituting characters for words and words for collocations. That is also one of the main reasons that Chinese segmentation becomes an uneasy problem. For example, give a Chinese sentence S_1: *jì suàn ji huì xìa guó jì xìang qí* in Fig. 1, it can be segmented into the following two possible ways that both make sense. This also inspires us to study the character-based model to avoid such difficulties.

Fig. 1. Two possible ways of Chinese word segmentation for the given sentence S_1

The computational approach of studying natural language is referred to as Natural Language Processing (NLP). One of the biggest challenges in this research area is that how machines can understand the semantic meaning of natural language. The classical NLP research heavily used knowledge-based approaches with hand-crafted rules from linguistics. In recent years, statistical methods have been widely used in NLP by considering natural language as data which can be studied using machine learning models [13]. Topic models are such a class of machine learning models using hierarchical probabilistic relations for analyzing discrete data collections (e.g, words). It assumes that documents are mixtures of some latent topics, and each topic is a probability distribution over vocabulary [8]. Most of research are focus on studying English documents. It has been validated to be an effective method in modeling English documents in semantic level. Most former research in applying topic models to explore Chinese documents also choose the Chinese word as the basic term of documents [21,23]. However, unlike English and the majority of western languages, the basic structural unit of Chinese language is not necessarily to be word based on some linguistics research [22]. In this paper, we developed a series of experiments of using topic models based on characters and words to answer the question proposed in the title of this paper.

This paper is structured as the following: topic models is introduced in Section 2. In Section 3, we discuss the Chinese language modeling based on Chinese words and characters using a well-known topic model called Latent Dirichlet Allocation. The empirical results are given and discussed in Section 4. The conclusions and future work are given in the end.

2 Topic Models

The essential idea of using topic models in language modeling is that a document can be represented by a mixture of latent topics and each topic is a distribution over the vocabulary. Early research in the study of topic modeling is probabilistic latent semantic indexing (pLSI) [7] [10] which is modified from classical LSI by introducing a latent topic variable. In this model, a document index d and a word w_n are conditionally independent given a latent topic z:

$$p(d, w_n) = p(d) \sum_z p(w_n|z)P(z) \tag{1}$$

An obvious shortcoming of this model is that it cannot assign probability to a previously unseen document owing to its use of document index as d. So the model need to be retrained when new documents are added. And this yields the linearly growth of the number of parameters to be estimated with the number of documents. Latent Dirichlet Allocation (LDA) [6] proposed by Blei *et. al* solved these problems by treating the topic mixture weights as a k-dimensional random variable θ with Dirichlet distribution. Dirichlet distribution is a family of continuous multivariate exponential probability distributions over the simplex of positive vectors that sum to one. A k-dimensional Dirichlet variable θ has the following probability density:

$$p(\theta|\alpha) = Dir(\alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \cdots \theta_k^{\alpha_k - 1} \tag{2}$$

where $\theta_i \geqslant 0$, $\sum_{i=1}^{k} \theta_i = 1$, $\alpha_i > 0$ and $\Gamma(x)$ is the Gamma function.

LDA is a generative model which specify a probabilistic procedure by which documents can be generated. The Dirichlet distribution is conjugate to Multinomial distribution and this property will facilitate the inference and estimation of the model. LDA assumes the generative process for each document $\mathbf{w}$ in a corpus $\mathcal{D}$ as follows:

1. Choose $\theta \sim \mathrm{Dir}(\alpha)$.
2. For each of the N words w_n:

a) Choose a topic $z_n \sim \mathrm{Multinomial}(\theta)$.
b) Choose a word w_n from $p(w_n|z_n, \beta)$, a multinomial probability conditioned on the topic z_n.

In the above process, β is a $k \times V$ matrix where k is the number of topics, V denote the size of vocabulary and $\beta_{ij} = p(w_j|z_i)$. This process is illustrated by the graphical model in Fig. 2. The generative process described above did not make any assumptions with regard to the order of words in a document, which is known as the bag-of-words assumption. Based on the generative relation (see Fig. 2) of these random variables, the joint probability of θ, $\mathbf{z}$ and $\mathbf{w}$ can be calculated by the following equation when given α and β as hyper-parameters:

$$p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta) \tag{3}$$

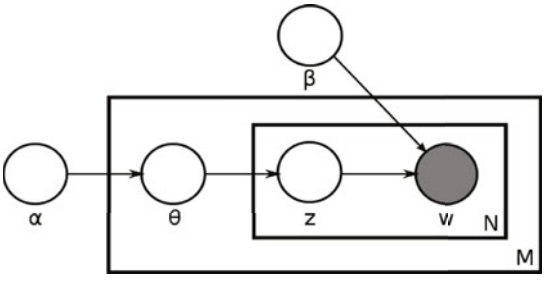

Fig. 2. Graphical representation of Latent Dirichlet Allocation

By integrating over θ and summing over z_n, the marginal distribution of document **w** is as follow:

$$p(\mathbf{w}|\alpha,\beta) = \int p(\theta|\alpha) \prod_{n=1}^{N} \sum_{z_n} p(z_n|\theta) p(w_n|z_n,\beta) d\theta \tag{4}$$

The above LDA model has been widely used in NLP such as text classification [18], information retrieval [20], and question answering system [17]. Excepting pLSI and LDA mentioned above, other topic models are generally evolved from LDA by alternating the structure of LDA or introducing new variable, such as correlated topic model and dynamic topic model. Correlated topic model [4] is able to capture the relationship between topics by replacing the Dirichlet distribution with logistic normal distribution. Dynamic topic model [5] improve LDA by relaxing the assumption that documents are exchangeable within the corpus and further provide a way to capture the evolution of topics. Several approximate inference algorithms can be applied to LDA, including variational approximation [6], Gibbs sampling [8,18], and expectation propagation [15]. In this paper, we employ the basic LDA with variational inference method in the subsequent experiments.

3 Topic Models with Chinese Characters and Words

Which entity, Chinese word or Chinese character, is the basic semantic unit in Chinese is still a controversial issue, partly because of the evolution of this language through nearly 3400 years' history [24]. In the early years of Chinese, each character has very specific semantic meaning like an English word. New characters can be created[1] to represent new meanings when necessary. In the modern Chinese language, unlike an English word, a Chinese character has a broad range of sense that is open to be combined with other characters to be more semantically specific. For example, character *rén* means *person*. By combining with the character *bìng* (illness), it becomes *bìng rén* which has more specific meaning of *patient*; combining with the character *nán* (male), it becomes

[1] The new character can be created as a pictograph or following some rules of assembling basic characters or strokes. There are some comprehensive research to study the evolution of Chinese characters, e.g. [1].

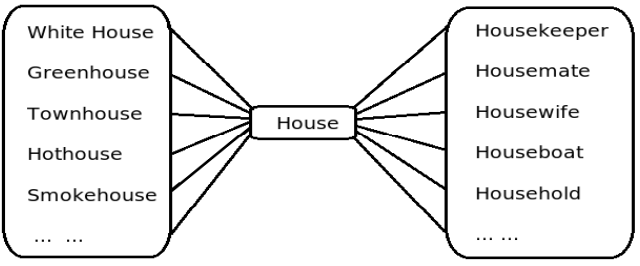

Fig. 3. An example of compound words of using "house" as the center word in English . It is similar to the composition of Chinese words using Chinese characters.

nán rén (man); combining with the character *lìe* (hunt), it becomes *lìe rén* (hunter). Another example is that the Chinese word for "Internet" is "*yīn tè wǎng*", where *yīn tè* is the transliterate of "Inter" and "*wǎng*" is the paraphrase of "net". Though modern Chinese characters serve as the building bricks of Chinese words and gradually lost their specific meanings, some linguists like Xu [22] still argues that the basic semantic unit of Chinese language should be Chinese characters but not Chinese words [2]. In this paper, we will not going deep into his linguistic theory but using a computational approach to study this phenomenon.

Most topic models treat a document as a bag-of-words [3] while it is fairly reasonable for English, the equivalent bag-of-characters assumption in modeling Chinese documents may look totally nonsense, because almost all Chinese words are made of grouping and ordering of two or more Chinese characters - these Chinese characters are heavily depending on each other. As we have seen from the above, a Chinese word can be considered as a compound of the basic semantic unit in some different ways (e.g., endocentric, exocentric, copulative and appositional). Some English compound words are also follow this way of generating new words by grouping two separate words such as the 'house' example given in the Fig. 3.

Topic models have been widely used in analyzing English text. However, unlike English and the majority of western languages, the basic structural unit of Chinese language is character instead of word (these is not space between words in Chinese). The character-based approach has recently received a great attention. It has been proved to be more effective in word segmentation by considering the word segmentation as a character tagging problem [19]. Some Chinese linguists point out that Chinese characters are quite expressive in meaning and emphasis should be paid to character in Chinese language research, while not to word. Previous research are inclined to use Chinese word as substitute of English word in Chinese language research [21,23]. Hence, in

[2] Xu [22] also argues that the Chinese should not be regarded as a SVO language and the modern Chinese syntax is heavily biased by wester linguistics and overlooking its evolution from the ancient Chinese language.

[3] The bag-of-words assumption has obvious shortcomings of ignoring the orders of words. Some research has been done to relax this assumption, e.g. [5,9].

this paper, we apply topic model based on Chinese word and character, respectively, to compare their performances.

Perplexity is a measure of the ability of a language model to generalize to unseen data. It is defined as the reciprocal geometric mean of the likelihood of corpus $\mathcal{D}$ containing M documents:

$$Perplexity(\mathcal{D}) = \exp\left\{-\frac{\sum_{d=1}^{M}\log p(\mathbf{w}_d)}{\sum_{d=1}^{M} N_d}\right\} \tag{5}$$

where $\mathbf{w}_d$ represents the dth document in the training corpus, N_d is the number of words containing in $\mathbf{w}_d$, and $p(\mathbf{w}_d)$ is the posterior probability $\mathbf{w}_d$ giving this model. Giving the language model LDA, $p(\mathbf{w}_d) = p(\mathbf{w}|\alpha,\beta)$ (Equation 4). The model with a lower perplexity means better ability in generalization. In the following experiments, we'll first compare the LDA models based on Chinese words and characters in perplexity measure.

4 Experimental Studies

In this section, we conducted a series of experiments to compare the performances of Chinese word-based LDA and Chinese character-based LDA in text modeling and classification as well as experiments of exploring topic model's semantic analyzing ability by using a bilingual corpus. In these experiments, we used a LDA software called LDA-C [4], which is an implementation in C with the variational Expectation Maximization (EM) method.

4.1 Data Descriptions

The Chinese corpus used in Section 4.2 is a news archive for classification provided by the Sogou laboratory [5]. The documents in this corpus are news articles collected from the website of the Sohu.com, which is one of the China's biggest Internet media company. These news articles are manually edited and classified into 10 classes that cover military, education, tourism and other topics (see Table 1 for details). In the original corpus, each class contains 8000 documents. But few of these documents contain nothing but some meaningless non-Chinese symbols. Therefore those documents contain less than 100 Chinese characters will be ignored in our research. We then selected 10000 documents with 1000 documents per class from the corpus as our experiment data which is named as NEWS1W and made it public available at the project page [6]. When we do the Chinese word-based topic modeling, we use a Chinese word segmentation tool ICTCLAS-09 [7] in our experiments. We also removed rare terms that appears less than 10 times across the whole corpus. For those terms that appears in over 50% of

[4] LDA-C can be downloaded from: `http://www.cs.princeton.edu/~blei/lda-c/`

[5] The corpus can be obtained at: `http://www.sogou.com/labs/dl/c.html`

[6] Dataset is available at: `http://icmll.buaa.edu.cn/projects/topicmodel/dataset/`

[7] ICTCLAS is an integrate Chinese lexical analysis system which provides a tool for Chinese word segmentation. `http://ictclas.org/`

the documents, we consider them as stop words and remove them from the corpus as well. After the above preprocessing steps, the corpus NEWS1W contains 25673 unique Chinese words and 4142 unique Chinese characters.

In Section 4.3, we used a Chinese-English bilingual corpus collected from the Yeeyan.com [8] which is a very popular online platform for volunteers to translate high quality magazine or newspaper articles from major western media (mainly English) to Chinese. From the website of Yeeyan.com, we crawled 7800 articles and each of them contains the original English article and the corresponding Chinese translation (parallel bilingual corpus). These articles are categorized into 8 classes: business, technology, culture, sports, health, nature, life and society. By the same method as described in the last paragraph we preprocessed the corpus and finally select 3200 text documents for experiments and named the corpus as BIL3200 [6]. However, we didn't stem English words in BIL3200, therefore, the words "google" and "google's" could be considered as different words. In summary, there are 21272 unique Chinese words, 3626 unique Chinese characters and 22769 unique English words in BIL3200.

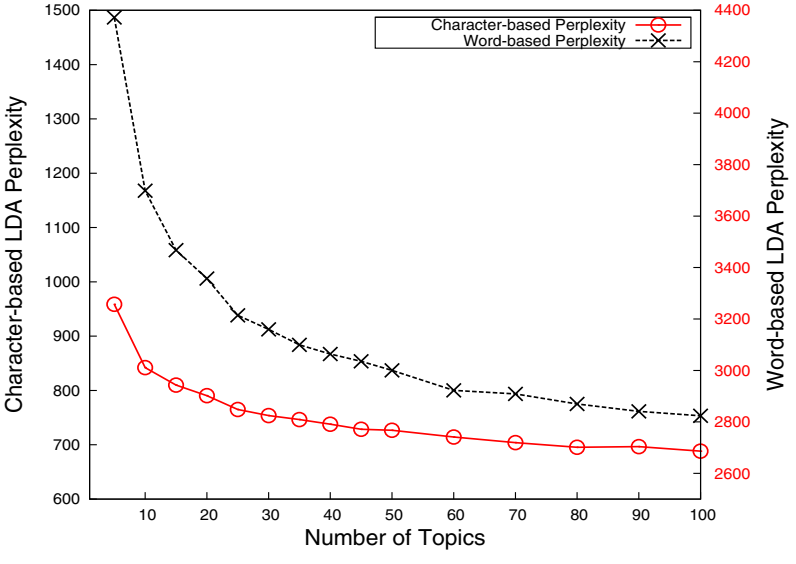

Fig. 4. Perplexity of the character-based LDA and the word-based LDA on NEWS1W

4.2 Language Modeling and Text Classification

To compare the generalization ability of the word-based LDA and the character-based LDA, we first compare their perplexity in language modeling. In this experiment, 10% of the documents in NEWS1W were held out as test set and the remaining 90% documents were used to train topic models based on characters and words, respectively. The perplexity with topic number varies from 5 to 100 is shown in the Fig. 4. As we can see

[8] Website: `http://www.yeeyan.org/`

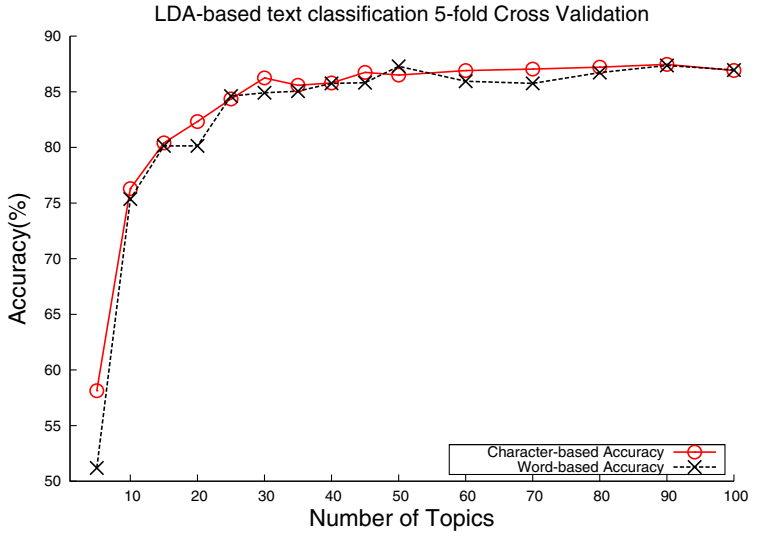

Fig. 5. 5-fold cross-validation classification accuracy using SVM on 9000 training documents of NEWS1W using LDA based on Chinese characters and words

from the results, the perplexity of both models decreased with increment of number of topics. Obviously, character-based LDA (the curve marked with circles) achieved much better performance in perplexity than word-based LDA (the curve marked with crosses), that means the character-based topic model has a better performance in language modeling. One possible reason is that the size of word vocabulary is much larger than the size of character vocabulary. A considerable part of words in the word vocabulary appear only few times in the corpus and this yields a lower log likelihood in perplexity calculation comparing to the character-based model.

In topic models, each document can be represented as a distribution over topics by parameter θ. If two documents are semantically close, it means the distance between two corresponding topic distributions is small, too. By using topic distributions, we can quantitatively measure the semantic (dis)similarities between two documents. In such a way, documents of each class can be mapped to a k-dimensional space where k is the number of topics. In such a space, we can use discriminative machine learning algorithms, such as Neural Network (NN), k-Nearest Neighbor (kNN) or Support Vector Machine (SVM), to classify these documents based on the *semantic distance* measure in terms of dissimilarity between topic distributions θ. In particular, we employ the SVM to the classification task on the NEWS1W dataset for its effectiveness in text classifications such as question classification [11]. We adopted libsvm[9] with its default RBF kernel in our experiments. The training and test accuracy with different number of topics k is shown in Fig. 5 and 6, respectively.

[9] LIBSVM is an integrated software for support vector classification, regression and distribution estimation. `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`

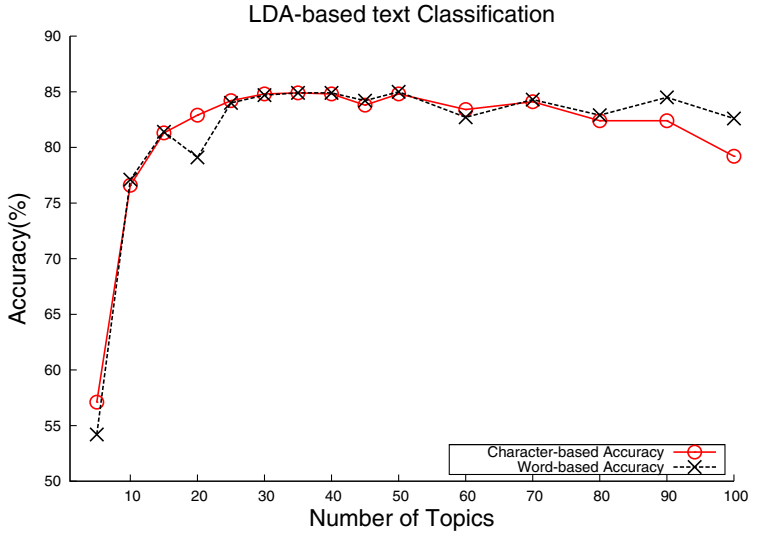

Fig. 6. Test accuracy using SVM on 1000 test documents of NEWS1W using LDA based on Chinese characters and words

As we can see from Fig. 5, the training accuracy of using both of the word-based and the character-based topic models are close to each other. Both accuracy rises with the increment of number of topics at first, and swing slightly around 85% when the number of topics reaches 30. To further compare the models' ability in classifying unseen document in LDA, we classify the test data of 1000 text documents and results shown in Figure 6 are also quite close for these two models. It is surprising because Chinese characters have always been regarded as incomplete semantic unit because it has to be combined with other characters to formulate a word which can precisely indicate a semantic concept. From these experiments, we can see that the topic models based on Chinese characters are as good as word-based topic models. In fact, the character-based LDA model even had a better accuracy at some number of topics and it has a much smaller (around 1/5) vocabulary than the word-based LDA model. Furthermore, Table 1 and 2 show the details of classification results on each class using Precision (P), Recall (R) and F_1 measure which is defined by:

$$F_1 = \frac{2R * P}{R + P} \tag{6}$$

4.3 Classification on a Bilingual Corpus

If the topic model is really able to analyze documents at the semantic level, it will be interesting to see its performance on bilingual documents that describe exactly the same contents but in different languages - they should have similar semantic measurements in terms of topic distributions. Therefore, in this section, we design a series of experiments on the parallel bilingual corpus BIL3200 by using the LDA models based on Chinese

Table 1. Classification results of the word-based LDA model with 30 topics on NEWS1W

	Military	Edu.	Tourism	Culture	IT	Job	Auto	Sports	Finance	Health
P	87%	94%	89%	64%	83%	78%	86%	100%	96%	80%
R	95%	80%	86%	85%	75%	83%	81%	97%	87%	78%
F_1	91%	86%	87%	73%	79%	80%	84%	98%	91%	79%

Table 2. Classification results of the character-based LDA model with 30 topics on NEWS1W

	Military	Edu.	Tourism	Culture	IT	Job	Auto	Sports	Finance	Health
P	95%	93%	90%	63%	72%	74%	99%	100%	95%	81%
R	94%	77%	81%	77%	84%	84%	91%	94%	87%	79%
F_1	94%	84%	85%	69%	78%	79%	95%	97%	91%	80%

characters (denoted by LDA-CC), Chinese words (LDA-CW) and English words (LDA-EW). The documents in BIL3200 are randomly divided into training set and test set by the ratio of 9:1. Then, we separately train the 3 individual models (LDA-CC, LDA-CW and LDA-EW) in order to obtain the topic distributions of each document. Figure 7 shows top 20 terms of 3 topics extracted independently from the 3 models. It is surprising to see that these 3 topics are semantically close which are talking about internet service or technology.

Subsequently we classify 320 test documents (10% of BIL3200) using the same method as described in the Section 4.2. The accuracy based on these 3 types of models are illustrated in Figure 8. The the classification accuracy of all the 3 models are getting closer with the increase of topic numbers. Even the biggest accuracy difference is no more than 6%. When the topic number is 40 and 50, the accuracy for these 3 models are almost identical (around 65%). This can be regarded as an evidence that topic models can effectively clustering semantically closed terms into topics and use them to model text documents. Given a document represented in two languages, the relative semantic relations are kept the same and the topics in different language may semantically mean the similar concepts (e.g., the topics shown in Fig. 7).

From the above two classification experiments on NEWS1W and BIL3200, we can see that LDA-CC and LDA-CW have indistinguishable performances in text classification. These empirical evidence show that the semantic content of a document can be well modeled by both LDA-CC and LDA-CW. Even Chinese characters are not considered as the basic semantic unit of Chinese language in classical linguistics, they can effectively represent the semantic content of a document as well as Chinese words in this computational approach. In other words, when we employ computational approach of modeling Chinese language by considering the semantic relation in terms of considering term occurrences. Chinese characters, which are as good as Chinese words, can be used as the basic semantic units.

Chinese character (Topic 23)	Chinese word (Topic 01)	English word (Topic 30)
社	公司	google
络	网络	search
万	网站	says
联	用户	yahoo
百	服务	news
户	客户	s
互	技术	company
交	市场	like
亿	已经	best
索	提供	people
界	新	world
智	互联网	turkeys
卡	企业	new
第	年	turkey
三	数据	art
者	美元	years
信	信息	internet
服	谷歌	market
务	家	google's
名	软件	video

Fig. 7. Top 20 words of three topics trained from topics models based on Chinese characters (left), Chinese words (middle) and English words (right). These three topics are semantically similar for containing characters or words related to the Internet technology or service.

4.4 Semantic Similarity of Documents

In topic models, each document can be represented by distribution on topics. This allows us to measure semantic distance (or dissimilarity) of documents using the Euclidean distance [10] of the topic distribution vectors θ. The semantic distance between document A and B is denoted by $D(A, B)$, so that we can say document A is semantically closer to B than C, i.e., $D(A, B) < D(A, C)$ if:

$$|\theta_A - \theta_B| < |\theta_A - \theta_C| \tag{7}$$

Given three documents A, B and C, if Eq. 7 holds in LDA-CC, it supposes that this relation should also hold in both other two models if the topic models can capture the semantic contents correctly. E.g.:

$$D_{LDA-CC}(A, B) < D_{LDA-CC}(A, C) \Longleftrightarrow D_{LDA-CW}(A, B) < D_{LDA-CW}(A, C) \tag{8}$$

[10] Kullback-Leibler (KL) distance is often used in measuring the divergence between two probability distributions, however, KL distance is asymmetric so that we use Euclidean distance is this experiment.

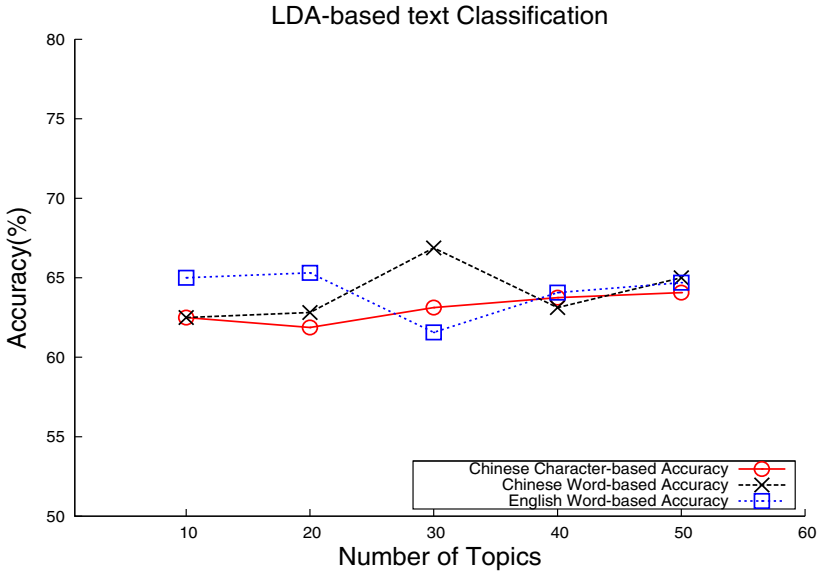

Fig. 8. Classification accuracy of the 320 test bilingual documents from BIL3200 using LDA based on Chinese character (LDA-CC), Chinese word (LDA-CW) and English word (LDA-EW)

$$D_{LDA-CW}(A,B) < D_{LDA-CW}(A,C) \Longleftrightarrow D_{LDA-EW}(A,B) < D_{LDA-EW}(A,C) \quad (9)$$

$$D_{LDA-CC}(A,B) < D_{LDA-CC}(A,C) \Longleftrightarrow D_{LDA-EW}(A,B) < D_{LDA-EW}(A,C) \quad (10)$$

where $D_{LDA-CC}(A,B)$ represents the semantic distance of document A and B using LDA-CC.

In the following experiments, we randomly pick 80 documents from the 320 BIL3200 test documents and measure the semantic distance between any two documents in LDA-CC, LDA-CW and LDA-EW, respectively. There are $C^2_{80} = 3160$ distances and $C^2_{3160} = 4991220$ relations given 3 documents. Among all the semantic distance comparisons, the percentage of the relation (Eq. 8) holds in both given models are shown in Table 3. The LDA-CC model and the LDA-CW models are the most similar pairs in semantic modeling; LDA-CW and LDA-EW are more similar comparing to

Table 3. Comparisons of document similarity relations between topic models based on Chinese characters, words, and English words when $k = 50$

	Total	Identical	Different
LDA-CC Vs LDA-CW	4991220	73.04%	16.96%
LDA-CC Vs LDA-EW	4991220	67.55%	32.45%
LDA-CW Vs LDA-EW	4991220	70.10%	29.90%
LDA-CC Vs LDA-CW Vs LDA-EW	4991220	55.34%	44.66%

LDA-CC and LDA-EW. There two facts are both intuitively true because the bridge between Chinese and English language is still based on word-to-word translation. The results in Table 3 also give a supporting evidence that the topic models are indeed modeling at the semantical level of the documents. Otherwise, we cannot found such relations holds in these 3 independently trained models.

4.5 Boosting of Classification Accuracy

To further explore the relations between these 3 models, we compare the predicted class labels using 320 test documents when the number of topics was set to 50. The detailed results are shown in Table 4: the column of "Identical" shows to the number of documents (and its percentage of all 320 test documents) which are classified identically by the models for comparisons in the first column. The column of "Different" is the number of documents (and percentage) which were classified differently. As we can see from the results, given a particular document, it is very likely to have the same classification no matter in which language or be character-based or word-based style. This result can sufficiently unveil the hidden capabilities of using topic models in modeling text semantics. The column of "Identical True" is the number of documents and percentage which are correctly classified among all the identical predictions in the column "Identical". The column of "Identical False" are the number of documents which are not correctly classified among the identical predictions from both models.

For example, the first row of Table 4 compares the models LDA-CC and LDA-CW. Among all 320 test documents, 246 (76.88% of 320) documents have the same predicted class labels by these two models. Among these 246 documents, 186 (75.61% of 246) are correctly classified while 60 are not, so that final prediction accuracy of considering both models is 75.61%. For each individual model, their accuracy is around 65% given the topic number $k = 50$. However, if we use voting from two or more models, the accuracy of classification can be significant improved (like boosting). By considering two models together, the accuracy can be improved approximately 10% from 65% to 75% (even higher if across languages, e.g. CC vs EW and CW vs EW). The best accuracy can be obtained is 83% by aggregating results from all the 3 models, that is nearly 20% higher than each individual models. This experiment provides a new way for boosting the Chinese text classification rate by training topic models independently on Chinese words and characters and combined their classification results afterwards.

Table 4. Model comparisons in text classification on BIL3200. Based on the voting of the 3 individual models, the classification accuracy can be boosted up to 83% while each individual model only has accuracy around 65% (see Fig. 8).

Model Comparisons	Identical	Different	Identical True	Identical False
LDA-CC vs LDA-CW	246 (76.88%)	74 (23.12%)	186 (75.61%)	60 (24.39%)
LDA-CC vs LDA-EW	224 (70.00%)	96 (30.00%)	175 (78.13%)	49 (21.87%)
LDA-CW vs LDA-EW	228 (71.25%)	92 (28.75%)	178 (78.07%)	50 (21.93%)
LDA-CC vs LDA-CW vs LDA-EW	200 (62.50%)	120 (37.50%)	166 (83.00%)	34 (17.00%)

5 Conclusions

In this paper, we use a computational approach to study semantic modeling of Chinese language. Latent Dirichlet Allocation (LDA), a well-known topic model, is employed to model Chinese texts based on Chinese words and Chinese characters. The empirical results showed that the character-based LDA model performs as good as the word-based LDA model in semantics modeling and classification. This research suggests that the topic model with Chinese characters can also effectively capture the semantic contents in text documents. The computational evidence presented in this paper supports Xu's [22] argument that the Chinese characters can be used as the basic semantic units in Chinese language modeling.

The main contribution of the papers is as the follows: (1) Based on comprehensive empirical studies, the results show that character-based LDA model has a better quality in language modeling than the word-based LDA model. (2) Both models have similar performance in text classification, due to its much smaller vocabulary size the character-based model is more preferable than the word-based model in real-world practice. (3) Based on the test on a bilingual corpus, results show that the topic models can indeed model text documents in a semantic level. The semantic relations between documents can be detected even using different language or different types of basic data (i.e., characters or words). (4) By aggregating the individual LDA classifiers that with only characters or words (or translations if available), we can improve the accuracy significantly in text classification.

Though topic models based on Chinese character achieved competent performance in modeling documents, its limitation is obvious: a considerable number of Chinese words have the semantic meaning irrelevant to the meanings of its consisting characters. On the other hand, topic model based on Chinese words did not make use of the connections between those words and their consisting characters. Therefore, how to build a topic model incorporating probabilistic relations between words and characters will be out future work.

Acknowledgment. This work is partially funded by the NCET Program of MOE, China and the SRF for ROCS. The second author also thanks the China Scholar Council for the visiting fellowship (No. 2010307502) to CMU. He also thanks Prof. Weidong Zhan of Peking University for useful discussions and Benny Zhang of CMU for his contributions in the early stage of this research.

References

1. Barbara, A.: The Nature of the Chinese Character. Simon, New York (1991)
2. Bishop, M.C.: Pattern Recognition and Machine Learning. Springer, Heidelberg (2006)
3. Blei, D.M., Griffiths, T., Jordan, M.I., Tenenbaum, J.: Hierarchical Topic Models and the Nested Chinese Restaurant Process. In: Thrun, S., Saul, L., Schoelkopf, B. (eds.) Advances in Neural Information Processing Systems (2004)
4. Blei, D.M., Lafferty, J.D.: Correlated Topic Models. In: Advances in Neural Information Processing Systems, vol. 18. MIT Press, Cambridge (2006)

5. Blei, D.M., Lafferty, J.D.: Dynamic Topic Model. In: Proceedings of the 23rd ICML, Pittsburgh, USA (2006)
6. Blei, D.M., Ng, A., Jordan, M.I.: Latent Dirichlet Allocation. Journal of Machine Learning Research 3, 993–1022 (2003)
7. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. Journal of the American Soc. of Inform. Sci. 41 (1990)
8. Griffiths, T.L., Steyvers, M.: Finding Scientific Topics. Proceedings of the National Academy of Science 101, 5228–5235 (2004)
9. Griffiths, T.L., Steyvers, M., Blei, D.M., Tenenbaum, J.B.: Integrating topics and syntax. In: Advances in Neural Information Processing Systems, vol. 17 (2005)
10. Hofmann, T.: Probabilistic Latent Semantic Analysis. In: Proceedings of UAI 1999, Stockholm (1999)
11. Huang, Z., Thint, M., Qin, Z.: Question Classification using Head Words and their Hypernyms. In: Proceedings of EMNLP, pp. 927–936 (2008)
12. Li, C., Sandra, T.: Mandarin Chinese: A Functional Reference Grammar. University of California Press, Los Angeles (1981) ISBN 978-0520066106
13. Manning, C., Schutze, H.: Foundations of Statistical Natural Language Processing. MIT Press, Cambridge (1999)
14. Maurits, L., Perfors, A., Navarro, D.: Why are some word orders more common than others? A uniform information density account. In: Proceedings of NIPS (2010)
15. Minka, T., Lafferty, J.: Expectation-propagation for the generative aspect model. In: Uncertainty in Artificial Intelligence (2002)
16. Ng, H.T., Low, J.K.: Chinese part-of-speech tagging: one-at-a-time or all-at- once? word-based or character-based. In: Proceedings of EMNLP, pp. 277–284 (2004)
17. Qin, Z., Thint, M., Huang, Z.: Ranking Answers by Hierarchical Topic Models. In: Chien, B.-C., Hong, T.-P., Chen, S.-M., Ali, M. (eds.) IEA/AIE 2009. LNCS, vol. 5579, pp. 103–112. Springer, Heidelberg (2009)
18. Steyvers, M., Griffiths, T.: Probabilistic topic models. In: Landauer, T., McNamara, D., Dennis, S., Kintsch, W. (eds.) Latent Semantic Analysis - A Road to Meaning (2007)
19. Wang, K., Zong, C., Su, K.-Y.: A character-based joint model for Chinese word segmentation. In: Proceedings of CoLing, pp. 1173–1181 (2010)
20. Wei, X., Croft, W.B.: LDA-based document models for ad-hoc retrieval. In: SIGIR (2006)
21. Wu, Y., Ding, Y., Wang, X., Xu, J.: A comparative study of topic models for topic clustering of Chinese web news. Computer Science and Information Technology (ICCSIT) 5, 236–240 (2010)
22. Xu, T.Q.: Fundamental structural principles of Chinese semantic syntax in terms of Chinese Characters. Applied Linguistics 1, 3–13 (2001) (In Chinese)
23. Zhang, Y., Qin, Z.: A topic model of Observing Chinese Characters. In: Proceedings of the 2nd International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC), pp. 7–10 (2010)
24. http://www.en.wikipedia.org/wiki/Chinese_language
25. http://www.en.wikipedia.org/wiki/Subject_Verb_Object