2010 Second International Conference on Intelligent Human-Machine Systems and Cybernetics

A Topic Model of Observing Chinese Characters

Yunkai Zhang
College of Software
Beihang University
Beijing, China, 100191
Email: bennyzyk@gmail.com

Zengchang Qin
Intelligent Computing and Machine Learning Lab
School of Automation Science and Electrical
Engineering, Beihang University, China, 100191
Email: zcqin@buaa.edu.cn

***Abstract*—The Topic Models are a class of hierarchical statistical models for analyzing document collections and it has become one of the most used techniques in Natural Language Processing in the recent years. It assumes that each document could be expressed as a mixture of topics and each topic could be characterized by a distribution over words. In previous research [6], like in English language, Topic Models for Chinese Language use the words as observing data. In this research, we demonstrated the effectiveness of using Chinese characters as the basic units of observing data. The comparisons with those models based on Chinese words and English words are presented.**

I. Introduction

Topic Models [1][2] are a class of generative hierarchical statistical models for analyzing collections of documents. In topic models, documents are presented as distributions over topics and topics are expressed as distributions over words.

The structure of Chinese language is different from most of western languages, such as English and French, whose basic units are words. However, Chinese language is usually considered having one more layer: Chinese words are treated as the basic unit for semantic understanding. However, Chinese words are made by combining Chinese characters. One Chinese character could be considered as one word, or it has to be combined with other Chinese characters to have some semantic meanings. In previous work, topic models for Chinese corpus used words as observing data[6]. However, currently Chinese segmentation techniques failed to provide an efficient and highly accurate method to identify words from sequences of characters.

In this paper, we only consider the problem of modeling Chinese text corpora using Chinese characters as observing data instead of Chinese words. Our work is aimed to challenge the stereotype of processing Chinese and proved that Chinese character is also a good kind of data for Topic Models. We did experiments on analyzing corpus based on Chinese characters and we also compared the experiment results with those LDA models observing Chinese words and English words.

The rest of the paper is organized as follows. Section II is short introduction to Topic Models and the Topic Model based on Chinese characters. Section III presents the results of our experiments and studies the experiments. Section IV summarizes the paper and gives the future work.

II. Topic Models

Topic models are first popularized by Thomas Hofmann's work[3] on probabilistic Latent Semantic Indexing (pLSI). In The pLSI model (shown in Figure 1), a document is represented as a probability distribution over a set of topics and each word in one document is generated from a single topic. The pLS model posits that a document label $d(d = 1, ..., D)$ and a word $w_n(n = 1, ..., W)$ are conditionally independent given an unobserved topic z:

$$p(d, w_n) = p(d) \sum_z (w_n|z)p(z|d)$$

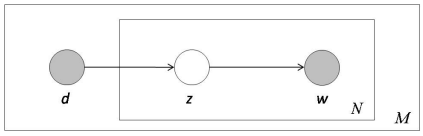

Fig. 1. Graphical model representation of pLSI model

However, the pLSI model learns the topic mixtures $p(z|d)$ only for those documents which are used for training, and thus pLSI is not well-defined for generative modelsl of documents. Besides, the number of parameters of a pLSI model grows linearly with the corpus size[1] proposed a three-level hierarchical Bayesian model called Latent Dirichilet Allocation (LDA), which has overcome the two shortness of pLSI model. The LDA (shown in Figure 2) model demonstrates that documents are represented as random mixtures over latent topics and each topic is represented as a distribution over words.

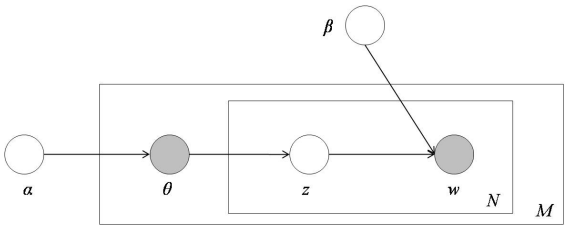

Fig. 2. Graphical model representation of LDA model

In LDA model the generative process for each document w in

978-0-7695-4151-8/10 $26.00 © 2010 IEEE
DOI 10.1109/IHMSC.2010.99

a corpus D is as following[1]:
1. Choose N from Possion (ξ)
2. Choose θ from Dir (α)
3. For each of the N words w_n:
   a) Choose a topic z_n from Multinomial(θ)
   b) Choose a word w_n from p($w_n|z_n, \beta$), a multinomial probability conditioned on the topic z_n

In LDA model, the topic mixture is expressed as a k-dimensional random variable θ with Dirichlet distribution [1]:

$$p\left(\theta|\alpha\right) = \frac{\Gamma\left(\sum_{i=1}^{k}\alpha_{\mathrm{i}}\right)}{\prod_{i=1}^{k}\Gamma(\alpha_{\mathrm{i}})}\theta_1^{\alpha_1-1}\cdots\theta_k^{\alpha_1-1}$$

Given the parameters of α and β, we can have the distribution of a document:

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

Where β is a $k * V$ matrix where $\beta_{ij} = p(w^j = 1|z^i = 1)$ and $p(w_n|z_n, \beta)$ is simply θ_i for the unique i such that $z_n^i = 1$.Finally, we can obtain the distribution of a corpus: $p(D|\alpha, \beta) =$

$$\prod_{d=1}^{M}\int p\left(\theta_{\mathrm{d}}|\alpha\right)\left(\prod_{n=1}^{N_d}\sum_{z_{dn}}p(z_{dn}|\theta_{\mathrm{d}})p(w_{dn}|z_{dn},\beta)\right)d\theta_{\mathrm{d}}$$

Blei provided a variational Expectation Maximization (EM) algorithm for the LDA model to estimate the parameters in [1], we used the same method in this research.
Specifically, we established a vocabulary of Chinese characters according to the GB-3212, which is the encoding and decoding standard of simplified Chinese characters. With the conversion of the data from characters to numbers, the input data format is as follows:

$$\begin{array}{ll} D_1 & C_1 : N_1, C_2 : N_2, \cdots C_{D_1} : N_{D_1} \\ & \vdots \\ D_i & C_1 : N_1, C_2 : N_2, \cdots C_{D_i} : N_{D_i} \\ & \vdots \\ D_m & C_1 : N_1, C_2 : N_2, \cdots C_{D_m} : N_{D_m} \end{array}$$

Where D_i is the Number of the characters in Document i, C_j is the number which indexes the j^{th} character in Document i, N_i is the number of how many times the character has occurred in the Document i and m is the size of the corpus.

III. Experiment studies

The structure of Chinese language is different from the structure of Western languages, which is usually made up by three levels that words, sentences and documents. The same thing with the Western language is that word is the basic semantic units. The difference is that Chinese has four levels: characters, words, sentences and documents. In Chinese, each word is made up by two or more characters and in most cases the meaning of each word is the combination of the meanings of the characters which form the word. In previous works[6], researchers tend to believe that words are good data for topic models since they are basic units for people to understand the meaning of documents. Because the nature of Chinese is different from English. There is no space between words. We have to do the words segmentation for using words. However the inefficiency and inaccuracy of Chinese segmentation techniques become the bottleneck of the computation based on words. Considering the fact that each character also has it basic meaning, we think that Chinese characters might be good data for language computation. Therefore we propose the topic model based on Chinese characters.
In order to evaluate the effectiveness of our model, we carried out three experiments. We used the Lancaster Corpus of Mandarin Chinese (LCMC) as our training corpus[7]. LCMC is constructed by the Department of Linguistics, Lancaster University. Besides that, we also established a bilingual training corpus that uses the source from yeeyan.org, which is a famous Chinese website for volunteers to translate magzine articles from western language (mainly English) into Chinese. We first find three categories of articles in LCMC, which are Category D: Religion, Category E: Trades, Skills and Hobbies, and Category P: Romance Fiction. We randomly selected 17 articles from each category for training. We first trained a 9-topic model and the result is shown in Figure 3.

As we can see from the Figure 3, different shapes of points are mostly located in one of the three areas in the coordinate systems. Then we trained several topic models with the pre-setting of different numbers of topics. In this training, we assume that for each document, the topic with the largest proportion value is the clustering result of this document. If the result of one document is different from the clustering results of the majority documents in the same category, we define it as a wrong result. The results of training model are shown in Table 1.

TABLE I
THE RESULT OF THE TOPIC MODEL OBSERVING CHINESE CHARACTERS

Number of Topics	9	8	7	6	5	4	3
Wrong Answer	0/51	0/51	1/51	1/51	6/51	6/51	7/51
Accuracy Rate	100%	100%	98%	98%	88%	88%	86%

As we can see from Table 1, the average accuracy of our training results is approximately 94.2%. The high accuracy of the result testified the usefulness of this model. Moreover, We found that with the number of topics decreasing, the accuracy decreases. The reason of this phenomenon is that each category in LCMC contains several themes of documents. Take Category E as an example, it contains trades, skills and hobbies, which are three totally different contents. Thus the limitation of the number of topics could prevent the LDA model from distinguishing all the documents precisely.
We did experiments to compare the performance of original topic models observing words and our topic model observing characters. Still we choose the same categories mentioned above from LCMC, which are: Category D: Religion,

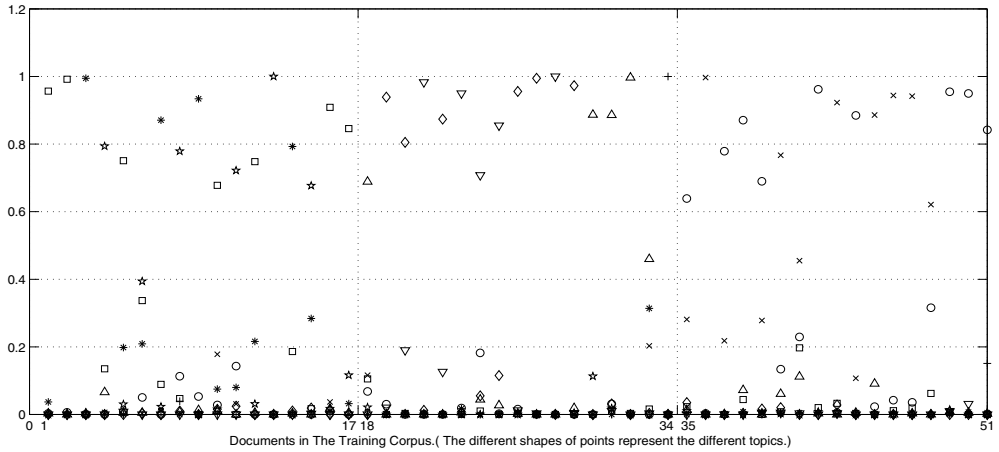

Fig. 3. The results of 9-topic training. The vertical axis represents the distribution of topics given one single document, and the horizontal axis is the list of all the training documents. T he different shapes of points represent the distribution of different topics.

Category E: Trades, Skills and Hobbies and Category P: Romance Fiction. We randomly pick 17 documents from each category for training. In LCMC every documents have already been segmented, so we can extract all the words from documents directly.We trained a 3-topic model observing both Chinese characters and Chinese words. 40 documents out of 51 documents have the same result from both topic models, which means the accordance rate of the model based on character and the model based on words is almost 80%. Furthermore, we also judge the correctness of the results, which is shown in Table 2.

TABLE II
THE RESULT OF 3-TOPIC MODEL TRAINING FOR BOTH WORDS AND CHARACTERS

Item	Error in Word model	Error in Character model	Different results between 2 models
Number	12/51	7/51	11/51

Surprisingly, we found the model observing characters outperformed the model based on words. Then we drew two three-dimensional probabilistic distribution pictures for both models, which are shown in Figure 4 and 5 respectively. We also can tell that the clusters of documents from the model based on character is better than that based on words.

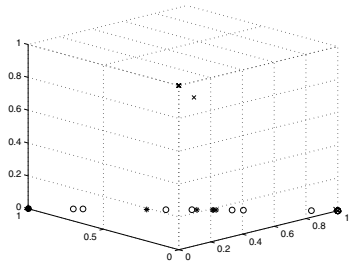

Fig. 4. Result From Model Observing Words

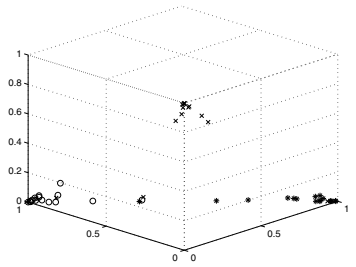

Fig. 5. Result From Model Observing Characters

The reasons why the model based on characters outshines may come from two aspects. One aspect is that we use a small training corpus so that the number of the word is too large comparing with the size of the corpus(e.g. over 8000 words in 51 documents, while only 917 characters in 51 documents). The second one is that the errors of the Chinese word segmentation in LCMC may have influence of the accuracy of word-based topic model. Though the results above cannot prove the model based on characters is better, they could at least show that topic model based on characters is indeed good for analyzing Chinese document collections. From the result of the comparison between these two topic models above, we come up with a hypothesis. Most of the characters in each cluster generated by the character-model could compose most of the words in the corresponding cluster generated by the word-model. This needs future work to prove.

In our work, we also designed an experiment to test whether

the model based on characters could really understands the semantic meaning of documents. First we choose 79 documents with three different topics, which are
Category A: finance and economics, Category B: psychology and mental science Category C: Library and Information Science from the website of Yeeyan, which is a famous Chinese websites collects documents in foreign languages and the Chinese translation versions of them. We trained 3-topic models for both English words and Chinese characters and the results are shown in Table 3.

TABLE III
THE RESULT OF 3-TOPIC MODEL TRAINING FOR BOTH ENGLISH WORDS AND CHINESE CHARACTERS

Item	Errors in English model	Errors in Character model	Different result between 2 models
Number	4/79	3/79	7/79

From the table above, we can see that the topic model based on Characters is as powerful as the topic model based on English words. Furthermore, we compared the results between these two models. Thus we drew two three-dimensional probabilistic distribution to compare the effectiveness of both models, which are shown in Figure 6 and 7 respectively. As the figures demonstrate, though the clusters generated by the Character model is not as tight as those generated by English-word model, they are still clearly separated from each other, which proves the effectiveness of the Chinese-character model.

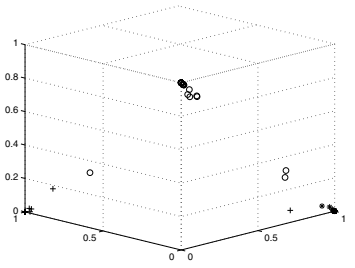

Fig. 6. Result From Model Observing English Words

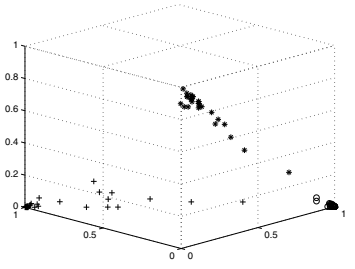

Fig. 7. Result From Model Observing Chinese Characters

In Topic Models, the k-topic distribution over each document in training corpus can be represented as a point in a k-dimensional space. Thus the semantic difference between two documents can be represented by the distance of the two corresponding points. In our work, we randomly pick three documents D_a, D_b, D_c from training corpus. Then we have three points: A, B and C from the results of the model based on English words If distance(A, B) $<$ distance(A, C), it means that from the perspective of semantic, of D_b is closer to that of D_a than D_c. Also we could have three points: A', B' and C' from the results of the model based on Chinese characters. If the relationship between AB, AC and $A'B'$, $A'C'$ is the same, we could say the model based on Chinese characters is as powerful as the model based on English words to analyze on the semantic difference between D_a, D_b, and D_a, D_c.
After calculation, we have totally 237237 comparisons from these 79 documents. The result is that, there are 34793 comparisons have the different results, and therefore the accordance rate of the two topic models is 85.33% (202444 same results out of 237237 groups of comparison). This result proves that the topic model observing Chinese characters could measure the semantic similarity between several documents with a high accuracy. In other words, this model could understand the semantic meaning of the training documents precisely as the Topic model observing English words does. Thus we believe this model is a competitive Topic Model, no matter whether it is compared with the model based on Chinese words or the model using English words.

IV. CONCLUSION

In this paper, we proposed a topic model using Chinese characters as observing data. We evaluate the performance of this model and also compared with the models based on Chinese words and English words. Our model proves to have the ability to analyze Chinese documents with high accuracy rate and the ability to understand the documents from the perspective of semantic meaning. The future work is to train more models on a larger corpus to compare different models' performance.

ACKNOWLEDGMENT

This work is partially funded by the NCET from the Ministry of Education of China.

REFERENCES

[1] Blei, D.; Ng, A.; and Jordan, M. 2003. Latent Dirichlet allocation. Journal of machine Learning Research.
[2] Steyvers, M; Griffiths, T. 2007. Probabilistic Topic Models. Latent Semantic Analysis: A Road to Meaning. Laurence Erlbaum
[3] Hoffman, T. 1999. Probabilistic latent semantic analysis In Uncertainty in Artificial Intelligence.
[4] Blei, D.; and Lafferty, J. 2007. Topic Models. Princeton University
[5] Qin, ZC; Marcus, T; Z, Huang. 2009. Ranking Answers by Hierarchical Topic Models. IEA/AIE 2009
[6] Hu, W.; Shimizu, N.; Nakagawa, H.; Huanye, S. 2008. Modeling Chinese documents with topical word-character models. Proceedings of the 22nd International Conference on Computational Linguistics
[7] McEnery T.; Xiao R. 2004. The Lancaster Corpus of Mandarin Chinese (LCMC), Department of Linguistics, Lancaster University