Mining Multilingual Topics from Wikipedia

Xiaochuan Ni, Jian-Tao Sun, Jian Hu, Zheng Chen
Microsoft Research Asia
No. 49 Zhichun Road, Beijing 100080, P.R. China
{xini, jtsun, jianh, zhengc}@microsoft.com

ABSTRACT

In this paper, we try to leverage a large-scale and multilingual knowledge base, Wikipedia, to help effectively analyze and organize Web information written in different languages. Based on the observation that one Wikipedia concept may be described by articles in different languages, we adapt existing topic modeling algorithm for mining multilingual topics from this knowledge base. The extracted "universal" topics have multiple types of representations, with each type corresponding to one language. Accordingly, new documents of different languages can be represented in a space using a group of universal topics, which makes various multilingual Web applications feasible.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing;

General Terms

Algorithms, Performance, Experimentation

Keywords

Multilingual, Wikipedia, Topic Modeling, Universal-topics

1. INTRODUCTION

In Wikipedia, each article describes one concept. Meanwhile, one concept is usually described in multiple languages, each language corresponding with one article. All documents associated with one concept (concept-unit) are similar in their topics. This motivates us to use topic modeling algorithms to mine multilingual topics from Wikipedia.

We propose a novel approach to model multilingual topics from Wikipedia data. All term by document matrices of *L* different languages are treated separately. A group of "universal" topics are used for modeling documents from different languages. The topics are inherently multilingual: each has *L* types of representations and each representation corresponds with one language. The links among documents describing the same concept are utilized to align topic representations: all these documents follow the constraint of sharing one identical topic distribution. Based on this unified modeling framework, new documents of different languages can be represented within one same vector space using the universal multilingual topics. Different from previous research work, our approach does not require additional linguistic resources like bi-lingual dictionaries or translation tools. Also, we exploit Wikipedia to extract multilingual topics applicable across multiple languages, instead of aligning documents in word or sentence level. With multilingual topics, it is very flexible to organize and utilize Web content written in different languages. Our experiments on text classification and document recommendation task indicate our topic modeling approach is effective.

Copyright is held by the author/owner(s).
WWW 2009, April 20-24, 2009, Madrid, Spain.
ACM 978-1-60558-487-4/09/04.

2. MULTILINGUAL TOPIC MODELING

We adapt Latent Dirichlet Allocation (LDA) to model multilingual topics (ML-LDA). We assume all the documents of a concept unit, although in different languages, share identical topic distribution. Figure 1 presents the graphical model of ML-LDA.

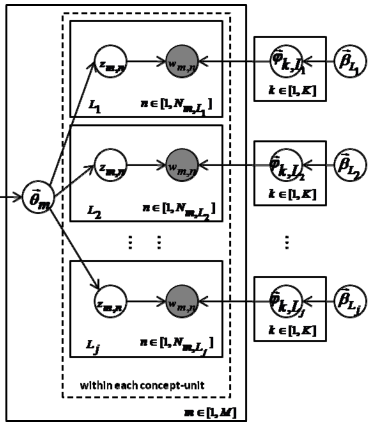

Figure 1. Graphical model representation of ML-LDA.

The notations are similar to those in LDA [1][2]. Here L_j denotes one language and $\vec{\varphi}_{k,L_j}$ denotes the word distribution for topic *k* in Language L_j. We modify Gibbs Sampling [2] method for the estimation of ML-LDA. Here in one concept-unit, documents in different languages share the same topic distribution but use different word distribution for each topic. Thus we compute $p(z_{c,i,L_j} = k \mid \vec{z}_{\neg(c,i),L_j}, \vec{w}_{L_j})$ by using:

$$\frac{n^{t}_{k,\neg(c,i),L_j} + \beta^{t}_{L_j}}{\sum_{v=1}^{V_{L_j}} (n^{v}_{k,L_j} + \beta^{v}_{L_j}) - 1} \frac{n^{k}_{m,\neg(c,i)} + \alpha_k}{\sum_{p=1}^{K} (n^{p}_{m} + \alpha_p) - 1}$$

where *t* denotes the index of the current word in Gibbs Sampling procedure. V_{L_j} is the vocabulary size of language L_j. *m* is the index of concept-unit. We compute $\vec{\varphi}_{k,L_j}$ by using:

$$\varphi_{k,t,L_j} = \frac{n^{i}_{k,L_j} + \beta^{t}_{L_j}}{\sum_{v=1}^{V_{L_j}} (n^{v}_{k,L_j} + \beta^{v}_{L_j})}$$

3. TOPIC MINING EXPERIMENTS

We have built a document-aligned comparable corpus from Wikipedia which was released on March 12, 2008. We only use 77,390 concept-units written in either English or Chinese. When ML-LDA is used, we set the hyper parameters *α* and *β* to be 0.5/*K* and 0.1 respectively, where *K* is universal-topic number ranging from 50 to 600, with 50 as the step size. For each value of *K*, the model is estimated using 200 Gibbs Sampling iterations.

Table 1 shows some example universal-topics produced by ML-LDA algorithm with *K* = 400. You can find that each universal-topic has two representations: the first line corresponds with the distribution of Chinese words and the second line is associated with English word distribution. Words on each line are ranked by probability score in decreasing order.

Table 1. Sample of Universal-topics

1st: 宇宙(universe) 理论(theory) 相对论(principle of relativity)… universe black relativity theory matter time gravitational …
2nd:足球(football) 年(year) 球(ball) 球员(football player) … football team cup national season world league scored …

We can also use another way to verify the effectiveness of our ML-LDA approach: word mapping between two languages. Given two words from different languages, we can measure their distance through their probability distributions over universal-topics. Table 2 gives two word mapping samples.

Table 2. Word mapping samples

电脑(computer): computer controller ibm plugged computers
Computer: 电脑 计算机 硬件 个人计算机 修改

4. APPLICATIONS

With universal-topics, we can map the documents of interest in different languages into the universal-topic space. In this section, we study how such representation can help cross-lingual applications. For experiment purpose, we collected a group of English and Chinese Web pages from Open Directory Project (ODP) website. 8 first level categories are used for experiments.

4.1 Cross-lingual Text Classification

Cross-lingual text classification (CLTC) addresses the problem of using texts labeled in one language to help classify texts in another language [4][5]. We have built two CLTC tasks: 1) classify Chinese pages by using the training data in English (En-to-Ch); 2) classify English pages by using the training data in Chinese (Ch-to-En). Support Vector Machine (SVM) algorithm is used as the basic classifier. Accuracy measure is adopted to evaluate the performance of the classification results.

We compare our universal-topic based approach with a translation based CLTC method ("Translation"). In "Translation" approach, the target texts are first translated into the language of source text and are then classified by the classifier trained with the source texts. Terms of Web pages are weighted by the Term Frequency (TF) scheme. The translation is based on two bilingual dictionaries (from Chinese to English and from English to Chinese respectively) downloaded from [3]. In this comparison, the ML-LDA model with K=400 is used.

Table 3. Comparison of Text Classification Results

	Universal-topic based	Translation
Ch-to-En	60.5303	48.2396
En-to-Ch	64.0129	59.7303

Table 3 shows the comparison results. We can see that, on both CLTC tasks, the universal-topic based approach outperforms the translation based method. The experimental results obtained in this section indicate that universal-topics learned from Wikipedia by ML-LDA models can indeed help CLTC tasks.

4.2 Cross-lingual Document Recommendation

In this paper, we define the cross-lingual document recommendation (CLDR) as: given a document, retrieve the related documents in a different language. Two tasks are experimented: 1) Given a Chinese Web page (source), recommend related English Web pages (Chi-to-En). 2) Given an English Web page (source), recommend related Chinese Web pages (En-to-Ch).

In this experiment, we adopt the following precision measure to empirically evaluate the recommendation results.

$$\Pr ecision = \frac{\sum_{q \in Q} \frac{|\{x \mid x \in T(n) \wedge f(x) = f(q)\}|}{n}}{|Q|}$$

where Q denotes query set, $T(n)$ denotes the set of top n most related pages and $f(\cdot)$ denotes class label. We measure the relatedness between two pages written in different languages in universal-topic space. In this experiment, the cosine similarity is utilized. We also compare our algorithm with the "Translation" approach in our experiment. In both CLDR tasks, we do recommendation for all source pages.

Table 4. Comparison of Recommendation Results

n		1	3	5	7	9
Ch-to-En	Universal-topic	0.49	0.48	0.47	0.46	0.46
	Translation	0.31	0.30	0.30	0.29	0.29
En-to-Ch	Universal-topic	0.49	0.46	0.46	0.46	0.45
	Translation	0.31	0.3	0.29	0.28	0.28

Table 4 gives the comparisons between two recommendation approaches on both CLDR tasks. The recommendation precision will decrease when the value of n grows. This is reasonable, because, when more pages are recommended, there is higher probability of recommending some irrelevant pages. We can see that the universal-topic based recommendation approach outperforms the translation based method regardless of n.

5. CONCLUSION

In this paper, we proposed a novel approach, ML-LDA, to mine multilingual topics from Wikipedia. Our experiments showed that ML-LDA is suitable for discovering multilingual topics (universal-topics). With universal-topics, documents in different languages can be represented within the same vector space. Therefore cross-lingual similarity can be measured without machine translation, which makes various cross-lingual applications feasible.

6. REFERENCES

[1] D.Blei, A.Ng and M.Jordan. Latent Dirichlet Allocation. JMLR, 3:993-1022, 2003.

[2] G.Heinrich. Parameter estimation for text analysis. Technical report, 2005.

[3] http://projects.ldc.upenn.edu/Chinese/

[4] J.Olsson, D. Oard and J.Hajic. Cross-language text classification. In Proc. of SIGIR-05, pages 645-646, 2005.

[5] Y.Wu and D.W.Oard. Bilingual topic aspect classification with a few training examples. In Proc. of SIGIR-08, pages 203-210, 2008.