Cross-Lingual Document Clustering

Ke Wu and Bao-Liang Lu*

Department of Computer Science and Engineering, Shanghai Jiao Tong University
800 Dong Chuan Road, Shanghai 200240, China
{wuke, bllu}@sjtu.edu.cn

Abstract. The ever-increasing numbers of Web-accessible documents are available in languages other than English. The management of these heterogeneous document collections has posed a challenge. This paper proposes a novel model, called a domain alignment translation model, to conduct cross-lingual document clustering. While most existing cross-lingual document clustering methods make use of an expensive machine translation system to fill the gap between two languages, our model aims to effectively handle the cross-lingual document clustering by learning a cross-lingual domain alignment model and a domain-specific term translation model in a collaborative way. Experimental results show our method, i.e. C-TLS, without any resources other than a bilingual dictionary can achieve comparable performance to the direct machine translation method via a machine translation system, e.g. Google language tool. Also, our method is more efficient.

1 Introduction

The development of the World Wide Web has created the ever-increasing numbers of Web-accessible documents in languages other than English. The automated organization of these heterogeneous document collections has posed a challenge. On the other hand, the literature about cross-lingual document clustering is sparse. Typically, machine translation system is introduced to fill the gap between different languages[2,3]. In this paper, we propose a novel model, called **domain alignment translation model**, to effectively cluster the multilingual documents. Our model is inspired by the observation that its translation of a word greatly depends on the domain information of the context. In addition, our method differs widely from existing methods in that instead of the process of term translation and then clustering, the domain alignment translation model conducts term translation and clustering simultaneously by learning a bilingual domain alignment model and a domain-specific term translation model. This occurs in a collaborative way with the help of a bilingual translation dictionary

* To whome correspondence should be addressed. This work was supported in part by the National Natural Science Foundation of China under the grants NSFC 60375022 and NSFC 60473040, and the Microsoft Laboratory for Intelligent Computing and Intelligent Systems of Shanghai Jiao Tong University.

Z.-H. Zhou, H. Li, and Q. Yang (Eds.): PAKDD 2007, LNAI 4426, pp. 956–963, 2007.

after conducting monolingual document clustering on two document sets, respectively. Experimental results show that the method based on the proposed model can achieve a comparable performance with the direct machine translation method, and that in some cases, the method can even outperform the latter one greatly.

The rest of this paper is organized as follows. In Section 2, we present related work on cross-lingual document clustering. In Section 3, we describe the domain alignment translation model consisting of a cross-lingual domain alignment model and a domain-specific term translation model. A method based on the proposed model is described in detail in Section 4. Experimental results with the method on data collected from the Internet are shown in Section 5. Finally, we conclude in Section 6.

2 Related Work

The literature about cross-lingual document clustering is sparse. Evans et al. (2003, 2004) [2][3] used simple document translation for multilingual clustering in their Columbia Newsblaster system. Although they developed a simple dictionary lookup glossing system for Japanese and Russian, the system performed less well than full translation. Mathieu et al. (2004)[1] proposed a cross-lingual similarity measure for the documents, using bilingual dictionaries, employing a Shared Nearest Neighbor approach by Ertoz et al. (2001)[6] to cluster cross-lingual documents and achieving promising results. However, their method was not compared with full-fledged translation and it was not practical since it took eight hours for 3,000 documents to cluster in the cluster discovery phrase. Furthermore, Evans and Mathieu noticed a common phenomenon that found documents from the same language tending to cluster more easily than from different languages. Compared with the above two methods, Chen and Lin(2000)[4] proposed a different cluster mapping approach for cross-lingual document clustering in their multilingual news summarizer but did not conduct experiments for the clustering performance, since their system is for multilingual news summarizer. In their cross-lingual clustering, they select words with high frequency occurrence in the target language as the translations of the words in the source language.

3 Domain Alignment Translation Model

3.1 Model Description

Before describing the model, the following notations are introduced.

- S denotes a set of source words to be translated. It can be further represented as $\{w_i^S\}, i = 1 \ldots M$, where w_i^S is the ith word in S.
- T denotes a set of translated words given S. It can be further represented as $\{w_i^T\}, i = 1 \ldots M$, where w_i^T is a translation of the ith word in S. w_{ij}^T denotes the jth candidate translation of the ith word in S.

- $GEN(S)$ is a set of candidate translations given S.
- C denotes some specific domain and ζ denotes domain sets. That is, C is an element of ζ.

We use the term **domain alignment translation model** to refer to a mechanism that determine the probability $P(T, C|S)$. We need to gather the heterogeneous documents, *e.g.* Chinese documents and English documents into different groups. Compared with homogeneous documents, *e.g.* only Chinese document or only English document, there exists a wide language gap among heterogeneous documents. Meanwhile, it is our observation that a strong relationship between a translation of a word and its domain exists. For example, there are varied translations in different domains in the case of , the translation of which is export in business domain , is exit in transportation domain and is speak in politics domain etc. Accordingly, it is reasonable to search for the translation of words and the specific domain simultaneously. According to Bayes's theorem, given a set of source words S, the best T and C is the one that carry out maximization as follows:

$$\begin{aligned}\{T^*, C^*\} &= \underset{T \in GEN(S), C \in \zeta}{\arg\max} P(T, C|S) \\ &= \underset{T \in GEN(S), C \in \zeta}{\arg\max} P(C|S)P(T|C, S) \end{aligned} \tag{1}$$

where $P(C|S)$ is called cross-lingual domain alignment model and $P(T, C|S)$is called domain-specific term translation model. If we postulate that given a specific domain C and a set of source words S, its translation of each word in S is generated conditionally independently. The second term in Equation (1) can be reformulated as $P(T|C, S) = \prod_i P(w_i^T|w_i^S, C)$. Equation (1) can then be rewritten as

$$\{T^*, C^*\} = \underset{T \in GEN(S), C \in \zeta}{\arg\max} P(C|S) \prod_i P(w_i^T|w_i^S, C) \tag{2}$$

3.2 Parameter Estimation

In the section, we describes how to estimate the probabilities $P(w_{ij}^T|w_i^S, C)$ and $P(C|S)$. If we had available parallel corpus from some specific domain C, estimating $P(w_{ij}^T|w_i^S, C)$ could be the same as estimating the translation model in IBM noisy channel model. However, it is usually non-trivial to explicitly define what is the domain we need. On the other hand, it is also hard to acquire large scale parallel corpus. Therefore, we try to obtain $P(w_{ij}^T|w_i^S, C)$ from the corpus in the target language. Applying the chain rule to $P(w_{ij}^T|w_i^S, C)$, we can deduce Equation (3):

$$P(w_{ij}^T|w_i^S, C) = \frac{P(w_{ij}^T, C|w_i^S)}{P(C|w_i^S)} \tag{3}$$

If we assume that the occurrence of its translation w_{ij}^T in domain C is independent of word w_i^S, Equation (3) can be approximated through $\frac{P(w_{ij}^T, C)}{P(C|w_i^S)}$. Then we can obtain the following formula:

$$P(w_{ij}^T|w_i^S, C) = \frac{P(w_{ij}^T|C)}{P(w_i^S|C)} \cdot P(w_i^S) \tag{4}$$

Also, according to total probability formula, $P(w_i^S|C) = \sum_j P(w_{ij}^T|C)$. Therefore, Equation (4) can be written as:

$$P(w_{ij}^T|w_i^S, C) = \frac{P(w_{ij}^T|C)}{\sum_j P(w_{ij}^T|C)} \cdot P(w_i^S) \tag{5}$$

The problem of estimating $P(w_{ij}^T|w_i^S, C)$ now can be solved via estimating $P(w_{ij}^T|C)$ and $P(w_i^S)$. The probability of some translation w_{ij}^T of a source word w_i^S in a specific domain, $P(w_{ij}^T)$, can be calculated by the relative frequency of translation w_{ij}^T in the domain, that is, $P(w_{ij}^T|C) = \frac{TF(w_{ij}, C)}{TF(w, C)}$, where $TF(w_{ij}, C)$ denotes the frequency of word w_{ij} in the domain C and $TF(w, C)$ denotes the frequency of all words in the given domain. As for $P(w_i^S)$, it is actually the unigram model and thus can use the MLE estimation, smoothed by some known techniques. However, it doesn't really involve the resulting decision for optimal C and T, since it is constant in the decision-making process.

4 An Algorithm Based on the Proposed Model

In section 3, we propose a domain alignment translation model. In this section, we propose a algorithm based on the model. Simply speaking, the algorithm comprises two steps: mono-lingual document clustering; two-level search, that is, to search for term translation and the corresponding cluster that maximize $P(T, C|S)$. In the monolingual document clustering phrase, we cluster the documents in a language at an appropriate cluster number. In the search phrase, we simultaneously search the aligned clusters and term translation.

The clustering algorithm based on naïve Bayes model has been shown to be effective for high dimensional text clustering. Also, the clustering model has the similar assumption as our proposed model, which each word is generated independently in the given domain. Hence, we choose the algorithm to conduct monolingual document clustering. One can be referred to [8] for details.

On the other hand, to obtain the optimal translations and domain of a set of source words, we have to try all possible combination of their translations and the domains. However, it is computationally prohibitive. Therefore, our best option is to use a greedy algorithm toward this end. In our proposed two-level search algorithm, we just choose the set of translations with most high probability given some domain to avoid try too many candidate translations, totally ignoring the other possible translation combinations. We refer to the two-level search algorithm based on clusters as C-TLS. The algorithm is summarized

Algorithm: C-TLS(D_1, D_2,K_1,K_2,Dic)
Input: D_1: document collection in language $L1$;
D_2: document collection in language $L2$;
K_1: the number of clusters to be partitioned for D_1
K_2: the number of clusters to be partitioned for D_2
Dic: the general-purpose bilingual dictionary from $L2$ to $L1$
Steps:
1. nbEM(D_1, K_1); nbEM(D_2,K_2); %% clustering algorithm based on NB model
2. Construct the corresponding centroid v_i for each cluster c_i of D_2;
3. **For** each cluster c_i for D_2
4. **For** each cluster c_j for D_1
5. search the translation of each word with most probability for the centroid v_i in c_j ;
6. Compute and record $P(T, C|v_i)$;
End For
7. Select $< c_i, c_j^* >$ as a mapping relation if $P(T, C|v_i)$ is the highest among the recorded scores.
End For
Output: a partition of the document data given by the cluster identity vector $C = \{c_1, c_2, \ldots, c_N\}, c_i \in \{1..K\}, N = |D_1| + |D_2|$

Fig. 1. Two-level search algorithm based on clusters

in Fig. 1. In this paper, we also investigate the extreme case of the algorithm, called TLS. That is, it occurs when K_2 equals to $|D_2|$ in C-TLS.

5 Experiments

5.1 Experimental Setup

The test data is collected via RSS reader[1] . The test data comprises Chinese Web pages and English Web pages from various Web sites. They consist of news during December 2005, consisting of 6,462 English Web pages and 6,011 Chinese Web pages. We should have collected data with seven topics. Unfortunately, when we translate all Chinese Web pages into English Web pages via translation tools provided by Google language tool, there are various errors for some Web pages via Google translation tool[2], so that we have to select five topics for experimentation. They include business, education, entertainment, science and sports. The category information is obtained by RSS reader. In addition, in the experiments, we use a general-purpose Chinese-English bilingual dictionary with about 292,000 entries.

In the paper, we use average purity and average entropy for our evaluation metrics. Average entropy is used to measure mean status of how the various classes of documents are distributed within each cluster.

[1] http://www.rssreader.com/
[2] http://www.google.com/language_tools

$$AverageEntropy = \frac{1}{k}\sum_{j=1}^{k} E_j \tag{6}$$

$$E_j = -\frac{1}{\log q}\sum_{i=1}^{q}\frac{n_i^j}{n_i} * \log(\frac{n_i^j}{n_i}) \tag{7}$$

where q is the number of classes in the document collection, k is the number of partitioned clusters, n_i is the number of documents from cluster i, and n_i^j is the number of documents from cluster i assigned to category j.

The second measure is average purity that measures the average extent to which each cluster contained documents from one primary class. The purity measure is defined as follows:

$$AveragePurity = \frac{1}{k}\sum_{j=1}^{k} P_j \tag{8}$$

where P_j is the fraction of the overall cluster size that the largest class of documents assigned to that cluster represents.

5.2 Experimental Results and Discussion

In our experiments, Our main experimental results are shown in Fig. 2. All results are shown as average ± 1 standard deviation over 5 runs. The term `Google`, `Google(I2C)` and `Google(C2C)` represent our three baselines. Specifically speaking, Google refers to the method employing nbEM algorithm to all preprocessed English web pages and translated Chinese web pages, while `Google(I2C)` denotes the method making a mapping from a translated Web page to clusters of native English Web pages through nbEM and `Google(C2C)` denotes the method relating clusters of the translated web page to clusters of the native English web pages. In addition, En2Ch indicates that English is source language and Chinese is target language, whereas Ch2En indicates the reverse case.

From Fig. 2, Fig. 3 and Table 1, we can summarize the results as follows:

- `C-TLS` have better performance than `TLS` and can achieve comparable performance to `Google(C2C)` while `Google(C2C)` has to spend much time and waste much storage space on the translated documents;

Table 1. Comparison of mean time of four methods spent over different numbers of clusters

	Methods	**Time**(sec.)				
		5	10	15	20	25
1	TSL(Ch2En)	114	182	255	678	1203
2	TSL(En2Ch)	100	154	206	268	310
3	C-TSL(Ch2En)	12	18	24	36	52
4	C-TSL(En2Ch)	13	21	29	43	61

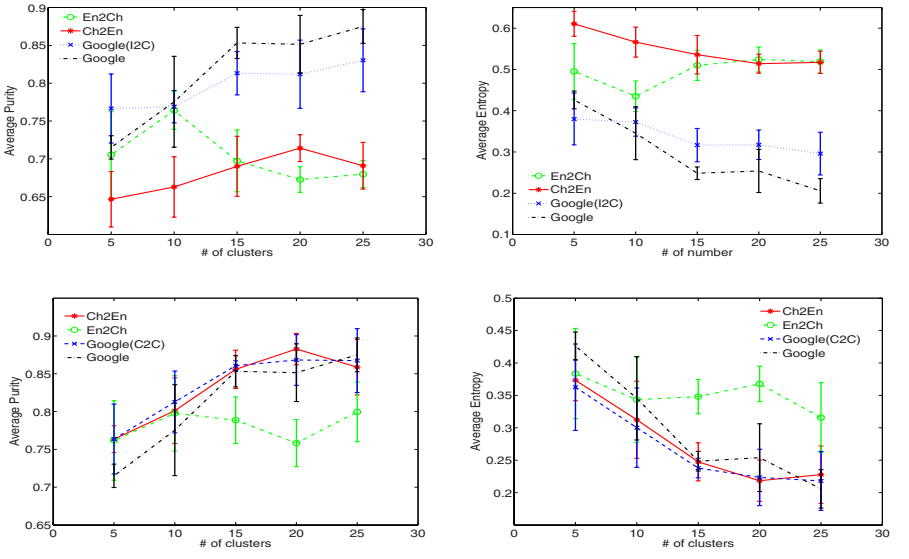

Fig. 2. Comparisons of different methods and baseline using direct machine translation. Results of `TLS`, `Google(I2C)` and `Google` are shown in the first row and results of `C-TLS`, `Google(C2C)` and `Google` are shown in the second row, where the number of clusters of English web pages is the same as one of Chinese web pages each run.

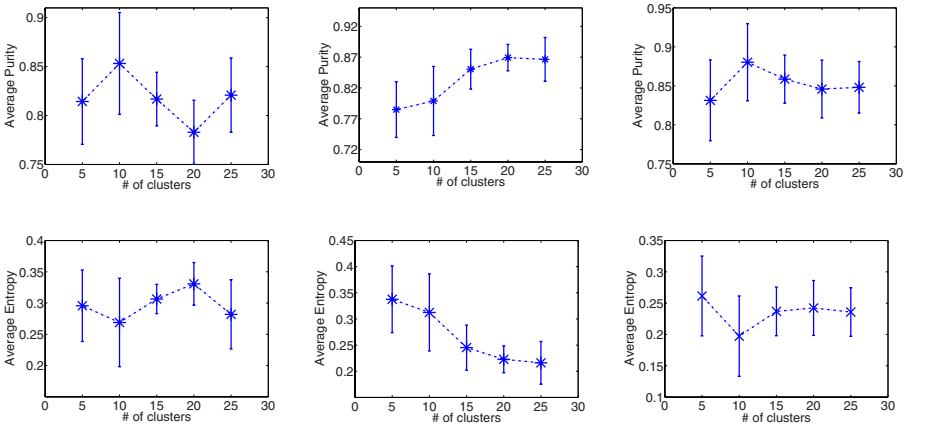

Fig. 3. Monolingual Clustering Results. Each column represents a set of results. Left-side column denotes Chinese web page clustering; middle-side column denotes English web page clustering; right-side column denotes the translated Chinese web page clustering.

- `C-TLS` achieves substantial and significant(p-value<0.05) improvements over `Google` method;
- Compared with `Google`, `Google(I2C)` and `Google(C2C)`, `TLS` and `C-TLS` is more efficient. It took about 8.3 hours for `Google`, `Google(I2C)` and

`Google(C2C)` to just translate Chinese web pages into English web pages and thus the time they spent on cross-lingual clustering is not listed in Table 1. In contrast, the longest runtime in Table 1 is about 20 minutes on Intel Pentium D 2.80GHz machine. This occurred when the number of clusters is 25 and `TLS(Ch2En)` method is used.

6 Conclusion

In this paper, we propose a novel domain alignment translation model to simultaneously conduct cross-lingual clustering and term translation. By learning a cross-lingual domain alignment model and a domain-specific term translation model in a collaborative way, we can cluster documents with a similar topic in different languages. Experimental results show our method without any resources other than a bilingual dictionary can achieve comparable performance to the direct machine translation method via Google translation tool. In our experiments, we only consider word, ignoring base phrase. We will incorporate translation of base phrase into our system in the future. On the other hand, the clustering in the source language and the clustering in the target language are related highly and thus we will explore how to reinforce their clustering quality interactively for future research.

References

1. Mathieu,B., Besançon,R., Fluhr C.: Multilingual document clusters discovery. In: RIAO'2004 proceedings, Université d'Avignon, France. (2004)
2. Evans,D., Klavans,J.L., McKeown,K.R. : Columbia Newsblaster: Multilingual News Summarization on the Web, In: Proc. HLT('04), Boston, MA. (2004)
3. Evans,D.K., Klavans,J.L.: A Platform for Multilingual News Summarization. Technical report, Columbia University Department of Computer Science. (2003)
4. Chen,H.H. and Lin,C.J. : A multilingual news summarizer. In: Proceedings of the 18th International Conference on Computational Linguistics. (2000) 159-165.
5. Hartigan,J.A. : Clustering Algorithms. John Wiley and Sons, Inc. (1975)
6. Ertoz,L., Steinbach,M., and Kumar,V. : Finding Topics in Collections of Documents: A Shared Nearest Neighbor Approach. In: Text Mine'01, Workshop on Text Mining (1st SIAM International Conference on Data Mining). (2001)
7. Carpuat,M. and Wu,D. : Word sense disambiguation vs. statistical machine translation. In: ACL 2005. (2005)
8. Meilă,M. , Heckerman,D. : An Experimental Comparison of Model-Based Clustering Methods. Machine Learning (2001) 42(1/2): 9-29