Cross-lingual Pseudo Relevance Feedback Based on Weak Relevant Topic Alignment

WANG Xu-wen
Institute of Medical Information & Library,
Chinese Academy of Medical Sciences,
Beijing 100020
wang.xuwen@imicams.ac.cn

ZHANG Qiang
State Grid Electric Power Research Institute,
Beijing 102200
zhangqiang7@sgepri.sgcc.com.cn

WANG Xiao-jie
Beijing University of Posts and Telecommunications,
Beijing, 100876
xjwang@bupt.edu.cn

LI Jun-lian
Institute of Medical Information & Library,
Chinese Academy of Medical Sciences,
Beijing 100020
li.junlian@imicams.ac.cn

Abstract

In this paper, a cross-lingual pseudo relevance feedback (PRF) model based on weak relevant topic alignment (WRTA) is proposed for cross language query expansion on unparallel web pages. Topics in different languages are aligned on the basis of translation. Useful expansion terms are extracted from weak relevant topics according to the bilingual term similarity. Experiment results on web-derived unparalell data show the contribution of the WRTA-based PRF model to cross language information retrieval.

1 Introduction

The problem of word mismatch between queries and retrieved documents is particularly serious in cross language information retrieval (CLIR). The integration of query expansion techniques and translation knowledge is considered as an effective way to improve the CLIR performance (Ballesteros and Croft, 1998; Qu et al., 2000).

Pseudo relevance feedback (PRF) is one of the useful query optimizing technologies for monolingual retrieval tasks (Rocchio, 1971; Ruthven and Lalmas, 2003). As to the CLIR task, researchers laid more efforts on establishing an effective cross-lingual PRF mechanism on the basis of the relevance and complementary of bilingual web pages (Ballesteros and Croft, 1997; Lavrenko et al., 2002). One of the key problems is how to choose useful or relevant bilingual expansion terms.

Typical cross-lingual PRF methods assume the top retrieved documents are relevant and perform feedback calculations on the whole pseudo-relevant document level. High-frequency words are often used for expanding original queries.

In recent years, topic models were applied to more and more multilingual tasks (Wang et al., 2009; Vulic et al., 2013). Ganguly (2012) proposed an improved cross-lingual topical relevance model based on the latent topics of top ranked documents. Wang (2013) proposed a cross-lingual PRF model based on bilingual topics and showed better results on parallel or comparable corpus. However, the hypothesis of common shared bilingual topics is not always suitable for unparallel documents, since they are often poor in content relevance.

In most cases, web pages retrieved from different language fields for a specific query may lack of parallelism. There may be some common topics shared by the retrieved documents in both languages, but there are also some specific topics for source language retrieval results or target language retrieval results respectively. Only the former common shared topics would be helpful to cross-lingual PRF.

29th Pacific Asia Conference on Language, Information and Computation pages 529 - 534
Shanghai, China, October 30 - November 1, 2015
Copyright 2015 by Xuwen Wang, Qiang Zhang, Xiaojie Wang and Junlian Li

In this paper, we assume that retrieved results in different languages have independent topical distribution, but there may be some overlapping topics that describe similar or relevant content. The overlapping content is defined as weak relevant topics.

A cross-lingual PRF model based on weak relevant topic alignment (WRTA) is proposed for modeling the weak correlation between unparallel documents. Relevant topics in different languages are aligned based on translation equivalent. Then useful expansion terms are extracted from relevant topics according to their bilingual similarity.

The structure of this paper is organized as follows: section 2 introduces the structure of the WTRA-based cross-lingual PRF model; section 3 presents the comparison experiment of different PRF methods on web-derived data; the final section shows our conclusion.

2 Method

It is assumed that cross-lingual retrieval results of a specific query, although lack of parallelism or comparability, may contain some relevant content.

Firstly, we perform monolingual topic modeling for source language documents D_S and target language documents D_T respectively. A widely used topic modeling method is the Latent Dirichlet Allocation (LDA) model, which is proposed by Blei (2003). So the LDA model is employed to generate candidate topic sets. Secondly, topics in different languages are aligned based on translation equivalence. Thirdly, useful expansion terms in aligned topics are selected on the basis of translation as well as web co-occurrence features. Figure 1 shows the process of weak relevant topic alignment and expansion terms extraction.

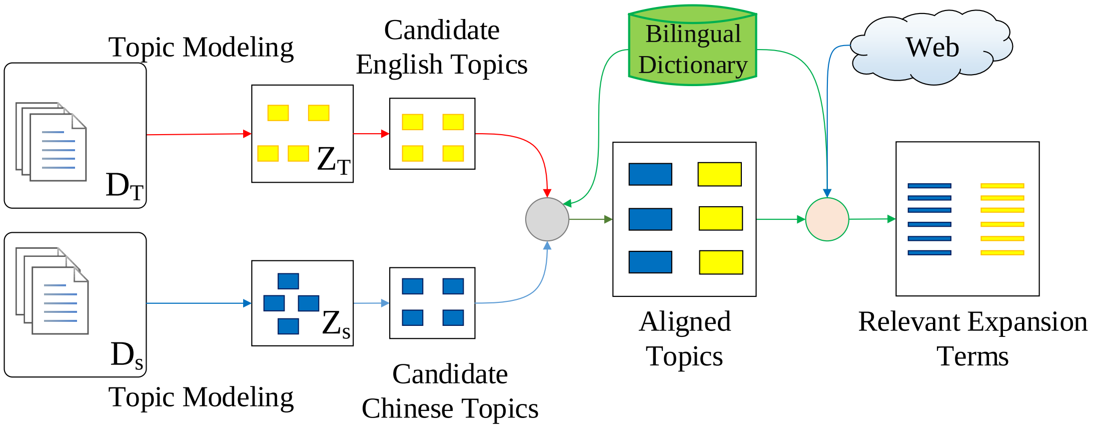

Figure. 1. Weak relevant topic alignment and extraction of relevant expansion terms

2.1 Weak Relevant Topic Alignment

For a specific query and its retrieved bilingual documents, we use the Gibbs sampling method for LDA inference (Han and Stibor, 2010) and generate two topic sets in different languages.

We need some clue for selecting candidate topics from the two topic sets. Topics that have close relation with the query or top-ranked documents are adopted as our candidate topics. Then relevant bilingual topic pairs with better translation equivalence are collected as the aligned topics.

1. Collecting candidate topics

Query related candidate topics: Topics including source language query terms Q_S or query translation terms Q_T are assumed to have directly correlation with users' query intention, namely query related topics Z_Q, see formula (1) and (2).

$$Z_Q^S = \bigcup_{z_S} \left(p\left(Q_S \mid Z_S\right) > 0 \right) = \bigcup_{z_S} \sum_{i=1}^{n} p\left(q_i^S \mid z_S\right) > 0 \quad (1)$$

$$Z_Q^T = \bigcup_{z_T} \left(p\left(Q_T \mid Z_T\right) > 0 \right) = \bigcup_{z_T} \sum_{i=1}^{n} p\left(q_i^T \mid z_T\right) > 0 \quad (2)$$

Alternative related candidate topics: The top M retrieved documents are supposed to be more relevant with users' query intention. So the top k topics with higher probability in the topic distribution $\theta(z)$ of the top M documents are adopted as the alternative related topics Z_E, see as formula (3) and (4).

Both of the query related topics Z_Q and the alternative related topics Z_E are collected as the candidate bilingual topic set Z_C, see as formula (5).

$$Z_E^S = \bigcup_{d \in D_M^S} k - \arg\max_{z} \theta_d(z) \quad (3)$$

$$Z_E^T = \bigcup_{d \in D_M^T} k - \arg\max_z \theta_d(z) \tag{4}$$

$$Z_C = Z_Q \cup Z_E \tag{5}$$

2. Topic alignment

Candidate topics in different languages are aligned according to their translation equivalence based on the machine-readable dictionary (MRD).

For a source language topic z_s and a target language topic z_t, which contain N_s terms or N_t terms respectively, the topical alignment rate is computed as formula (6). The *m* in numerator is the amount of terms in the source language topic z_s that have translation in the target language topic z_t, the *n* is the amount of terms in target language topic z_t that have translation in source language topic z_s.

$$f(z_s, z_t) = \frac{m + n}{N_s + N_t} \tag{6}$$

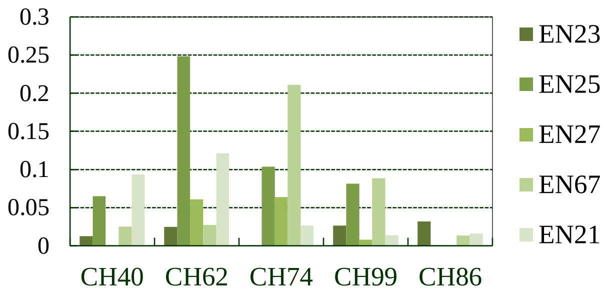

Figure. 2. The alignment rate between the candidate bilingual topics of "Information retrieval"

Figure 2 shows the alignment rate between candidate bilingual topics of the query "Information retrieval". The bi-directional translation process can be regard as a mutual multi-voting game between topics in different languages. The higher rate implies more latent relevance.

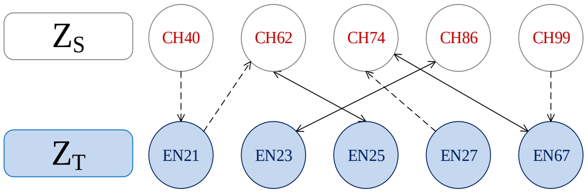

Figure. 3. The alignment of Chinese-English candidate topics of the query "Information retrieval"

Figure 3 shows the alignment relationship between candidate bilingual topics of the query. The solid arrow with two directions represents a mutual alignment between two topics, since they vote each other with the highest rate. In this case, three couples of topics are aligned successfully.

2.2 Selecting Relevant Expansion Terms

Cao et al. (2008) analyzed the potential influence of different terms to the performance of information retrieval tasks, and concluded that useful terms for query expansion in pseudo relevant documents only account for 18% in high frequency terms. Too many expansion terms may reduce the efficiency of retrieval systems (White and Marchionini, 2007).

In our work, terms from candidate topics are sorted into three categories. The first category contains semantically relevant terms that have translation or synonymy with original queries. Terms in the second category have no direct relationship with queries, but they are essential content in describing identical themes in bilingual context. The last category contains irrelevant noisy terms that should be filtered out.

To select useful expansion terms effectively, a bilingual term similarity score is computed based on web-derived data. For each pair of aligned topics, a source language term and a target language term are organized as a conjunctive query "w_s + w_t" for the real time web searching. In the real web searching, terms in different languages often co-occur in the title, snippet or URL of a retrieved multilingual webpage. So, the web co-occurrence of each pair of terms from aligned topics would be counted, see formula (7). The binary function in formula (8) represents the translation relationship between the term w_s and wt. The bilingual similarity score of the term pair is the linear combination of web co-occurrence and the translation feature, see as formula (9). The parameter λ is the weighting coefficient.

In each target language topic, terms are ranked according to the similarity score with the source language query terms, namely $\mathrm{Sim}(q_i^s, w_j^t)$. Terms with similarity score lower than the threshold μ will be filtered out.

$$f_C(w_i^S, w_j^t) = p(w_i^S, w_j^t) = \frac{\#\ \text{retrieval records including}\ (w_i^S, w_j^t)}{\#\ \text{retrieval records from IR system}} \propto \frac{N_c}{N} \tag{7}$$

$$f_T\left(w_i^S, w_j^t\right) = \mathrm{Trans}\left(w_i^S, w_j^t\right) = \begin{cases} 1, & \text{only if } \left(w_i^S, w_j^t\right) \text{are mutual translation} \\ 0, & \text{other} \end{cases} \tag{8}$$

$$\mathrm{Sim}\left(w_i^S, w_j^t\right) = \lambda\, f_T\left(w_i^S, w_j^t\right) + \left(1-\lambda\right) f_C\left(w_i^S, w_j^t\right), \qquad (0 \le \lambda \le 1) \tag{9}$$

2.3 Cross Language Pseudo Relevance Feedback Based on WRTA

Based on the above algorithm, relevant terms are obtained for cross-lingual query expansion. Figure 4 shows the CLIR process with WRTA-based PRF mechanism.

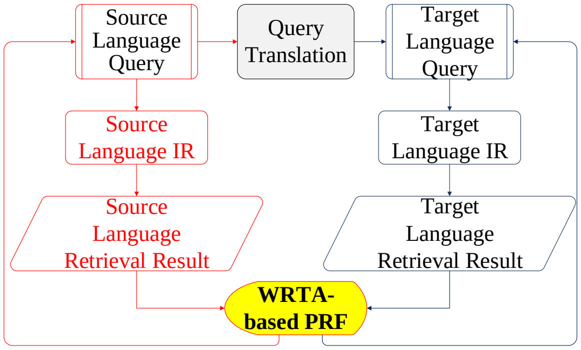

Figure. 4. CLIR process with cross language pseudo relevance feedback based on WRTA.

3 Experiments

3.1 Experimental setting and Data

We perform cross-lingual PRF experiments on a self-constructed CLIR system, namely CTP-CLIR system (Wang et al., 2013). As a prototype system, it contains a text pre-processing module, a query translation module, a retrieval model (Indri 5.2) and the pseudo relevance feedback module, which integrated various PRF mechanisms. The CTP-CLIR system could access web pages on line and retrieve local multilingual database automatically.

A Web-derived Chinese-English corpus was collected to simulate the real cross language web search task. The source language query set was selected from the Chinese science and technology concepts on CNKI. Each query contains 1 to 3 word tokens, totally 54 queries. The target language queries were the English translation of the Chinese queries, obtained from the query translation module.

The bilingual retrieval documents were collected from Google's real time retrieval results. Top 10 source language pages were crawled for each Chinese query, since most web users pay more attention to the top-ranked results in the retrieval list. The target language pages were retrieved via Google's cross-lingual retrieval. Totally 1080 web pages were collected.

Then 20 queries with poor comparable retrieval results were selected as our test set, totally 400 web pages. Other queries were saved as our training set, totally 34 queries and 680 web pages. All of the collected web pages were cleaned by the text preprocessing module and then be indexed by Indri 5.2.

Since the typical assessment criteria, such as precision or recall, shows no significant difference on the relatively small dataset, we take nDCG (Discounted Cumulative Gain) to evaluate the ranking effect of retrieval results. 27 volunteers were invited to judge the relevance of bilingual documents.

3.2 Parameters

All the parameters were tuned on the basis of our training set.

It was observed that topics from the top 1 document as well as the query related topic Z_Q contributed most to the best ranking results. So the parameters of topic alignment were configured as follows, the alternative document number M=1, the alternative topic number k=2. Each query has 1.5 pair of weak relevant topics on average. The filtering threshold of term probability in each topic σ=0.005.

The weighting coefficient of the bilingual term similarity score was set as λ=0.05, and its threshold for filtering terms μ=0.85.

The hyper parameters of the LDA model were optimized based on the training set, as follows, $\alpha = 0.1, \beta_s = 0.01, \beta_t = 0.02$. The number of training iterations was 10000.

3.3 Comparative Experiments

To examine the feedback effect of proposed method, we chose the normal CLIR results without PRF modulation as our baseline.

Various PRF methods, such as VSM-based PRF framework, LDA-based PRF model, bilingual LDA-based PRF model, etc., are also conducted before or after the query translation stage of CLIR, namely comparative experiments.

3.4 Results

Figure 5 shows the CLIR results employing different PRF methods on unparallel documents.

The first column is the result of CLIR without PRF mechanism. The second to the forth column show the results of PRF based on the Vector Space Model (VSM), namely pre-translation VSM-based PRF, post-translation VSM-based PRF and combined VSM-based PRF. The fifth to the seventh column show the results of PRF based on monolingual topic model, namely pre-translation LDA-based PRF, post-translation LDA-based PRF and combined LDA-based PRF. The eighth column is the result of bilingual LDA-based PRF, which performs integrated feedback on the basis of the bilingual LDA model. The last column shows the result of proposed WRTA-based cross-lingual PRF.

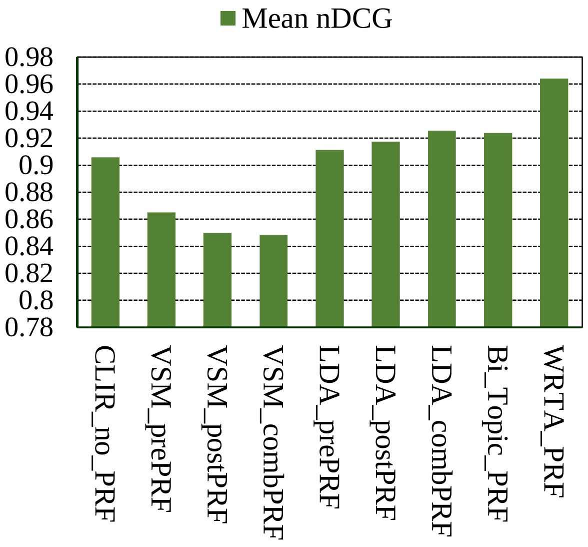

Figure. 5. Comparison of cross-lingual PRF based on WRTA and other PRF methods.

It can be observed that the VSM-based PRF methods introduced too much noise, since the feedback calculation was performed on the entire document level. The LDA-based PRF methods showed a slightly better performance than former methods, verifying the fact that a fine-grained topic may introduce more relevant terms into query expansion.

However, the PRF method based on bilingual LDA model, which used to achieve better performance than monolingual models on parallel documents, showed no advantage here, since the poor quality of the unparallel feedback documents limited the effectiveness of topical PRF methods.

In spite of the interference from the unparalleled documents, the WRTA-based PRF model achieved the highest improvement for CLIR. Expansion terms from aligned topics, which were selected based on the translation and web co-occurrence features, showed clear relevance with original queries. On one hand, noisy terms were filtered out effectively and the amount of expansion terms was reduced sharply. On the other hand, the remained expansion terms showed positive impact on the performance of CLIR on unparallel documents.

4 Conclusion

This paper describes a way to discover useful information from unparallel retrieval results for cross-lingual pseudo relevance feedback. A cross language PRF model based on weak-relevant topic alignment is proposed.

In comparison with various PRF methods, WRTA-based PRF model showed better performance and robustness in the CLIR task on less comparable documents. So it is proved to be more suitable for web oriented tasks.

It is worth noting that the effect of expansion terms for cross-lingual PRF is very complicated. The quality and quantity of expansion terms, which are influenced by the quality of translation as well as feedback documents, should be controlled carefully. Too many expansion terms may drown out valuable information, so the quantity of expansion terms is reduced sharply in our work. Noise terms are removed from candidate expansion terms effectively, so that useful terms may achieve positive feedback performance.

As to the further work, it will be necessary to introduce more multilingual knowledge resources into the cross-lingual PRF mechanism, such as Wikipedia, multilingual ontology, as well as semantic web knowledge, etc. Rich knowledge resources will be a helpful supplement for choosing relevant expansion terms, and furthermore, improving the performance of PRF model in CLIR tasks.

References

Andrzejewski D, Buttler D. Latent topic feedback for information retrieval [J]. Proceedings of the 17th Acm Sigkdd International Conference on Knowledge Discovery and Data Mining San Diego Ca Usa August 21 24 2011, 2011: 600-608.

Ballesteros L, Croft W. Statistical Methods for Cross-language Information Retrieval [J]. 1998.

Ballesteros L, Croft W. Phrasal translation and query expansion techniques for cross-language information retrieval [J]. Proceedings of the 20th Annual International Acm Sigir Conference on Research and Development in Information Retrieval, 1997, 31(SI): 84-91.

Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation [J]. The Journal of machine learning research, 2003, 3: 993-1022.

Cao G, Nie J Y, Gao J, et al. Selecting good expansion terms for pseudo-relevance feedback[C]//Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2008: 243-250.

Ganguly Debasis and Leveling Johannes and Jones Gareth J F Cross-lingual topical relevance models [C]. 24th International Conference on Computational Linguistics, 2012.

Han X, Stibor T. Efficient Collapsed Gibbs Sampling for Latent Dirichlet Allocation[J]. Jmlr, 2010.

Http://www.cnki.net/

J. J. Rocchio. Relevance feedback in information retrieval. [J]. In the SMART Retrieval System: Experiments in Automatic Document Processing, 1971:313-323

Lavrenko V, Choquette M, Croft W. Cross-lingual relevance models[J]. Proceedings of the 25th Annual International Acm Sigir Conference on Research and Development in Information Retrieval, 2002.

Orengo V, Huyck C. Relevance feedback and cross-language information retrieval[J]. Information Processing and Management an International Journal, 2006, 42(5): 1203-1217.

Qu Y, Eilerman A, Jin H. The Effect of Pseudo Relevance Feedback on MT-Based CLIR[J]. Riao 2000 Content Based Multi Media Information Access Csais, 2000.

Ruthven I, Lalmas M. A survey on the use of relevance feedback for information access systems[J]. The Knowledge Engineering Review, 2003.

Vulic I, De Smet W, Moens M. Cross-language information retrieval models based on latent topic models trained with document-aligned comparable corpora[J]. Information Retrieval, 2013.

Wang A, Li Y, Wei W. Cross language information retrieval based on LDA [J]. Intelligent Computing and Intelligent Systems. ICIS 2009.

Wang Xu-wen, Wang Xiao-jie, Sun Yue-ping, Cross-lingual pseudo relevance feedback based on bilingual topics, Journal of Beijing University of Posts and Telecommunications, Volume: 36; Issue 4; (JA) Pages: 81-84, August 2013.

Wang X, Zhang Q, Wang X, et al. LDA based PSEUDO relevance feedback for cross language information retrieval[C]// Cloud Computing and Intelligent Systems (CCIS), 2012 IEEE 2nd International Conference on-IEEE, 2012:1511-1516.

Wang X, Wang X, Zhang Q. A Web-Based CLIR System with Cross-Lingual Topical Pseudo Relevance Feedback [J]. Lecture Notes in Computer Science, Volume 8138 LNCS, 2013.

White, R.W., & Marchionini, G. (2007). Examining the effectiveness of real-time query expansion. Information Processing &Management, 43(3), 685–704.