A Late Fusion Approach to Cross-lingual Document Re-ranking

Dong Zhou[1], Séamus Lawless[1], Jinming Min[2], Vincent Wade[1]
Center for Next Generation Localisation
1. KDEG, School of Computer Science and Statistics, Trinity College Dublin, Dublin 2, Ireland
2. Department of Computer Science, Dublin City University, Dublin 9, Ireland
+353 1 896 1765
Dong.Zhou@scss.tcd.ie, Seamus.Lawless@scss.tcd.ie, Jinming. Min@googlemail.com, Vincent Wade@scss.tcd.ie

ABSTRACT
The field of information retrieval still strives to develop models which allow semantic information to be integrated in the ranking process to improve performance in comparison to standard bag-of-words based models. Cross-lingual information retrieval is an example of where such a model is required, as content or concepts often need to be matched across languages. To overcome this problem, a conceptual model has been adopted in ranking an entire corpus which normally exploits latent/implicit features of the text. One of the drawbacks of this model is that the computational cost is significant and often intractable in modern test collections. Therefore, approaches utilizing concept-based models for re-ranking initial retrieval results have attracted a considerable amount of study, in particular the latent concept model. However, fitting such a model to a smaller collection is less meaningful than fitting it into the whole corpus. This paper proposes a late fusion method which incorporates scores generated by using external knowledge to enhance the space produced by the latent concept method. This method is further demonstrated to be suitable for multilingual re-ranking purposes. To illustrate the effectiveness of the proposed method, experiments were conducted over test collections across three languages. The results demonstrate that the method can comfortably achieve improvements in retrieval performance over several re-ranking methods.

Categories and Subject Descriptors
H.3.3 [**Information Systems**]: Information Storage and Retrieval - *Information Search and Retrieval;* I.2.7 [**Computing Methodologies**]: Artificial Intelligence - *Natural Language Processing.*

General Terms
Algorithms, Experimentation, Languages.

Keywords
Cross-lingual Information Retrieval, Document Re-Ranking, Data Fusion, Linear Combination Model.

CIKM'10, October 26–30, 2010, Toronto, Ontario, Canada.

1. INTRODUCTION
Information retrieval (IR) often suffers from the so called "*vocabulary mismatch*" problem. A document may be semantically relevant to a query despite the fact that the specific query terms used and the terms found in the document completely or partially differ [2]. An extreme example of this is cross-lingual information retrieval (CLIR) where content or concepts need to be matched across languages. Consequently, overlap with respect to linguistic terms (or via translated words) should not be a necessary condition in query-document similarity calculation. Methods relying on the "bag-of-words" model display poor performance in many cases. In order to overcome the vocabulary mismatch problem, several solutions have been suggested which exploit semantic relations between text units. Among these methods, the latent model, the explicit model and the mixed model are commonly employed [6, 7].

However, these models have well documented drawbacks. Firstly, these methods are very computationally complex. In the latent model, complexity grows linearly with the number of dimensions and the number of documents. This has been the biggest obstacle to the widespread adoption of this kind of method. For the explicit and mixed model, the number of dimensions used to map documents onto the external knowledge space are often limited to ten thousand [3] so that it is feasible to process the large test collections used. Another problem with the explicit model is that the documents are often distributed over thousands of dimensions in which the semantic relatedness will degrade dramatically [1]. How to find these dimensions is not reported and this may significantly influence the retrieval performance.

Therefore, researchers started to consider integrating the aforementioned models into smaller, controlled document collections to address these shortcomings and assist the retrieval process. Zhou and Wade [6] proposed a Latent Dirichlet Allocation (LDA)-based method to model the latent structure of "topics" deduced from the initial retrieval results. However, due to the smaller corpus size, fitting a latent model into this corpus has less meaning than fitting the same model into a large, web-scale corpus. This means that some form of justification has to be applied to achieve better performance. A simple approach to address this problem is to directly apply the explicit or mixed model to a controlled corpus to improve ranking performance. A similar problem will arise in the latent model in this single semantic space, resulting in limited improvements.

In the setting of CLIR, this problem becomes more complicated. CLIR focuses on research into the retrieval of documents written in languages different to the language in which the query is expressed [5]. Assuming that query translation is employed, the simplest approach to addressing the cross-lingual re-ranking problem is to directly apply the monolingual methods on the results obtained using the translated query. Obviously, the drawback of this approach is that errors resulting from translation noise will be inherited by the re-ranking process, which may result in unsatisfactory performance.

To address the challenges described above, Zhou et al. further introduced a method incorporating scores generated using external knowledge to enhance the semantic space produced by the latent concept method [7]. This method is intended to produce global consistency across the semantic space: *similar entries are likely to have the same re-ranking scores with respect to the latent and manifest concepts*. This is extended in the current paper to solve the cross-lingual document re-ranking problem by automatically inducing a semantic correspondence between two languages (query language and document language) using parallel Wikipedia corpora through a late fusion approach. This correspondence is then used to project the inquiry into another language in the semantic space to accomplish the re-ranking task.

2. LATENT RE-RANKING MODEL

In this section, the problem addressed by this paper is defined. The latent document re-ranking model implemented is also described briefly.

2.1 Problem Definition

Let $\mathbb{D} = \{d_1, d_2, \dots, d_n\}$ denote the set of documents to be retrieved. Given a query q, a set of initial results $\mathbb{D}_{init} \in \mathbb{D}$ of top documents are returned by a standard IR model (initial ranker). However, typically the performance of the initial ranker can be improved upon. The purpose of our re-ranking method is to re-order a set of documents $\mathbb{D}'_{init}$ so as to improve retrieval accuracy at the most highly ranked results.

2.2 LDA-based latent model

The specific method used here is borrowed from [6], which is based on the LDA model. In this model, the topic mixture is drawn from a conjugate Dirichlet prior that remains the same for all documents. The distance between a query and a document based on this model is defined via the Kullback-Leibler divergence:

$$RS_{LDA}^{KL} = -D(MLE_q(\cdot)||LDA_d(\cdot))$$

The final score is then obtained through a linear combination model of the re-ranking scores based on the initial ranker and the latent document re-ranker, shown as follows:

$$Score^{latent-lda} \stackrel{\text{def}}{=} \lambda \cdot OS + (1-\lambda) \cdot RS_{LDA}^{KL}$$

where OS denotes original scores returned by the initial ranker and λ is a parameter that can be tuned with $\lambda = 1$ meaning no re-ranking is performed. This algorithm is named ***latent-lda***.

2.3 LSI-based latent model

Another well-known approach to the latent model is the LSI method. It is based on Singular Value Decomposition (SVD), a technique from linear algebra. It uses cosine correlation to compute the similarity between a query and a document in a SVD space to obtain RS_{LSI}^{COS} :

$$RS_{LSI}^{COS} = \cos(\overrightarrow{q_{k'}}, \overrightarrow{d_{k'}})$$

This is combined with the original score to produce the final ***latent-lsi*** algorithm score:

$$Score^{latent-lsi} \stackrel{\text{def}}{=} \lambda' \cdot OS + (1-\lambda') \cdot RS_{LSI}^{COS}$$

3. EXPLICIT RE-RANKING MODEL

In this section the concept of a cross-lingual explicit re-ranking model is presented which is based upon external knowledge resources. Please refer to [7] for details about this model applied in a monolingual setting.

A very important characteristic of Wikipedia is that articles are linked across languages. Cross-lingual links are those that link a certain article to a corresponding article in the Wikipedia database in another language. A previous analysis of this cross-lingual link structure between the German and English Wikipedia showed that 95% of these links are indeed bi-directional [4]. The existence of a language link function $lan - link_{l \to m}$ is assumed that maps an article of Wikipedia W_l to its corresponding article in Wikipedia W_m.

Given a document $d \in \mathbb{D}$ in language l, the document can be indexed with respect to another language m, by transforming the vector $\Phi_l(\vec{d})$ into a corresponding vector in the vector space that is spanned by the articles of Wikipedia in the target language:

$$\Psi_{l \to m}: \mathbb{R}^{|W_l|} \to \mathbb{R}^{|W_m|}$$

This linking function is calculated as follows:

$$\Psi_{l \to m} \langle v_l^1, \dots, v_{|W_l|}^i \rangle = \langle v_m^1, \dots, v_{|W_m|}^j \rangle$$

Where

$$v_p' = \sum_{q \in \{q* | lan-link_{l \to m}(a_{q*}) = a_p\}} v_q$$

With $1 \le p \le |W_l|$, $1 \le q \le |W_m|$. So that in order to get the representation of a document $d \in \mathbb{D}$ in language l with respect to Wikipedia W_m it is simply a case of computing the function:

$$\Psi_{m \to l}(\Phi_l(\vec{d}))$$

The score produced by this model can then be calculated as the cosine similarity[1]:

$$RS_{CL-ESA}^{COS} = \cos(\Phi_m(\overrightarrow{q_m}), \Psi_{l \to m}(\Phi_l(\overrightarrow{d_l})))$$

To apply this method to re-ranking, W_l and W_m are limited to the number of highly relevant documents for a given query. There are two alternative ways of applying this method. Suppose that the original query is written in language l and the translated query is written in language m. It is possible to use the original query to search in W_l and find the correspondent articles in W_m, or the translated query could be used to search in W_m and locate the correspondent articles in W_l:

[1] Note that RS_{ESA}^{COS} is used later for monolingual explicit model

CL-ESA1: $\Psi_{l\to m}(\Phi_l(\overrightarrow{q_l})), \Phi_m(\overrightarrow{d_m})$

CL-ESA2: $\Phi_l(\overrightarrow{q_l}), \Psi_{m\to l}(\Phi_m(\overrightarrow{d_m}))$

The final ranking score is defined as:

$$Score^{explicit-cross} = \mu' \cdot OS + (1-\mu') \cdot RS^{COS}_{CL-ESA}$$

As in the latent model, OS denotes original scores returned by the initial ranker and μ' is a parameter that can be tuned with $\mu' = 1$ meaning no re-ranking is performed. This algorithm is named as ***explicit-cross***. It has two permutations, ***explicit-cross-q and explicit-cross-d*** *which* correspond to CL-ESA1 and CL-ESA2 respectively.

4. LATE FUSION VIA MULTIPLE SCORES

Armed with the latent and explicit models defined above, the late fusion method proposed by this paper is now described. A simple assumption is taken here: the number of dimensions produced by the explicit model has to correspond to the number of dimensions induced by the latent model. Based upon this assumption, the late fusion method can be conducted so as to make a constraint: $|W_l| = k$.

This approach then takes the inputs from the original score, the latent model and explicit model and produces the final ranking score through a linear combination model, defined as:

$$Score^{fusion-lda} \stackrel{\text{def}}{=} \varsigma \cdot OS + (1-\varsigma-\tau) \cdot RS^{KL}_{LDA} + \tau \cdot RS^{COS}_{ESA}$$

and

$$Score^{fusion-lsi} \stackrel{\text{def}}{=} \varsigma' \cdot OS + (1-\varsigma'-\tau') \cdot RS^{COS}_{LSI} + \tau' \cdot RS^{COS}_{ESA}$$

These algorithms, denoted as ***fusion-lda*** and ***fusion-lsi***, can be viewed as specific to monolingual re-ranking. For cross-lingual re-ranking, the scores of the ***fusion-lda-cross*** and ***fusion-lsi-cross*** algorithms are defined as:

$$Score^{fusion-lda-cross} \stackrel{\text{def}}{=} \varsigma \cdot OS + (1-\varsigma-\tau) \cdot RS^{KL}_{LDA} + \tau \cdot RS^{COS}_{CL-ESA}$$

and

$$Score^{fusion-lsi-cross} \stackrel{\text{def}}{=} \varsigma' \cdot OS + (1-\varsigma'-\tau') \cdot RS^{COS}_{LSI} + \tau' \cdot RS^{COS}_{CL-ESA}$$

An interesting point is that the results produced by the monolingual explicit model can actually help the cross-lingual re-ranking with translated query $\Phi_m(\overrightarrow{q_m})$ and document $\Phi_m(\overrightarrow{d_m})$. This generates mixed dual-space algorithms ***fusion-lda-cross-mix*** and ***fusion-lsi-cross-mix*** for cross-lingual re-ranking:

$$Score^{fusion-lda-cross-mix} \stackrel{\text{def}}{=} \varsigma \cdot OS + (1-\varsigma-\tau_1-\tau_2) \cdot RS^{KL}_{LDA} + \tau_1 \cdot RS^{COS}_{ESA} + \tau_2 \cdot RS^{COS}_{CL-ESA}$$

and

$$Score^{fusion-lsi-cross-mix} \stackrel{\text{def}}{=} \varsigma' \cdot OS + (1-\varsigma'-\tau'_1-\tau'_2) \cdot RS^{KL}_{LSI} + \tau'_1 \cdot RS^{COS}_{ESA} + \tau'_2 \cdot RS^{COS}_{CL-ESA}$$

This concludes the description of the proposed re-ranking models and methods.

5. EVALUATION

5.1 Experimental Setup

The text corpus used in the experiment described below consisted of elements of the CLEF-2008[2] and CLEF-2009 European Library (TEL) collections[3] written in English, French and German. All of the documents in the experiment were indexed using the Terrier toolkit[4]. Prior to indexing, Porter's stemmer and a stopword list[5] were used for the English documents. A French and German analyzer[6] was used to analyze the French and German documents. The initial ranker used in this study is the classic vector space model. A Wikipedia database in English, French and German was used as an explicit concept space. Only those articles that are connected via cross-language links between all three Wikipedia databases were selected. A snapshot was obtained on the 29/11/2009, which contained an aligned collection of 220,086 articles in all three languages. Parameters are tuned so as to maximize mean average precision (MAP).

The following evaluation metrics were chosen to measure the effectiveness of the various approaches: the precision of the top 5 documents (Prec@5), the precision of the top 10 documents (Prec@10), normalized discounted cumulative gain (NDCG), MAP and Bpref. Statistically-significant differences in performance were determined using a paired t-test at a confidence level of 95%. Whenever a re-ranking method assigns different documents with the same score, the ties are broken by document ID.

5.2 Experimental Results

For the cross-lingual part of the experiments, firstly the performance of different variants of explicit models is compared to determine which one should be chosen to use in the late fusion methods. A straightforward observation is that in fact there is very little difference between the various explicit methods. In more detail, ***express-cross-d*** and ***express-cross-q***'s performance are on some occasions better than that of the ***express-mono*** algorithm. This partially confirmed that directly applying monolingual methods into the cross-lingual applications may not always produce the most beneficial results.

The performance of the fusion methods are considered with respect to the non-monolingual runs. As illustrated by Table 1, three variants exceeded the initial ranker by a statistically significant margin in many test runs. While ***fusion-lda-cross*** and ***fusion-lsi-cross*** methods sometimes delivered better performance than that of monolingual methods, more noticeable improvements were observed in the methods incorporating monolingual re-ranking scores. This cross-lingual re-ranking

[2] The test collections used in CLEF-2008 and CLEF-2009 are in fact identical.

[3] http://www.clef-campaign.org

[4] http://terrier.org

[5] ftp://ftp.cs.cornell.edu/pub/smart/

[6] http://lucene.apache.org/

LSI-based Methods								
Metrics	BL							
	French-English				German-English			
	init.	***fusion-lsi***	***fusion-lsi-cross***	***fusion-lsi-cross-mix***	***init.***	***fusion-lsi***	***fusion-lsi-cross***	***fusion-lsi-cross-mix***
Prec@5	0.48	0.5*	0.504*	0.508*	0.48	0.48	0.476	0.492*
Prec@10	0.444	0.462*	0.462	0.464*	0.416	0.442*	0.444*	0.442*
Prec@20	0.389	0.395	0.397	0.397	0.364	0.367	0.368	0.37
NDCG	0.3675	0.372	0.3717	0.373*	0.3617	0.3684*	0.3689*	0.3691*
MAP	0.2116	0.214	0.2137	0.2145	0.2097	0.2135	0.2139	0.215
bpref	0.2448	0.2501	0.25	0.2501	0.2434	0.2485	0.2485	0.251*
LDA-based Methods								
Metrics	BL							
	French-English				German-English			
	init.	***fusion-lda***	***fusion-lda-cross***	***fusion-lda-cross-mix***	***init.***	***fusion-lda***	***fusion-lda-cross***	***fusion-lda-cross-mix***
Prec@5	0.48	0.532*	0.536*	0.54*	0.48	0.488	0.488	0.492*
Prec@10	0.444	0.47*	0.472*	0.474*	0.416	0.466*	0.462*	0.47*
Prec@20	0.389	0.389	0.389	0.391	0.364	0.389*	0.39*	0.394*
NDCG	0.3675	0.3695	0.3698	0.3705	0.3617	0.3716*	0.3721*	0.3716*
MAP	0.2116	0.2164	0.2165	0.218*	0.2097	0.2234*	0.2236*	0.224*
bpref	0.2448	0.2582*	0.2579*	0.2578*	0.2434	0.2587*	0.2588*	0.2591*

Table 1. Cross-lingual Re-ranking Experimental Results. For each evaluation setting, statistically significant differences between different methods and the initial ranker are indicated by a star. Due to space restrictions, only results on the English collection are shown.

performance is a favorable indication of the stability of the late fusion technique.

6. CONCLUSIONS AND FUTURE WORK

This paper proposed and evaluated a late fusion approach for re-ordering initial retrieval results with state-of-art cross-lingual performance. This paper also proposed a way to apply the explicit model to the cross-lingual re-ranking problem, and performed a systematic comparison between different models. It would be beneficial to conduct a direct comparison between ranking and re-ranking using the proposed algorithmic variations. Future work will also include identifying different combination models rather than the linear combination model used for data fusion.

7. ACKNOWLEDGMENTS

This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (www.cngl.ie) at Trinity College Dublin and Dublin City University.

8. REFERENCES

[1] Cimiano, P., Schultz, A., Sizov, S., Sorg, P. and Staab, S. Explicit versus latent concept models for cross-language information retrieval. In *Proceedings of the 21st international jont conference on Artifical intelligence* Pasadena, California, USA, 2009. Morgan Kaufmann Publishers Inc., 1513-1518.

[2] Furnas, G. W., Landauer, T. K., Gomez, L. M. and Dumais, S. T. The vocabulary problem in human-system communication. *Commun. ACM*, 30, 11, 1987, 964-971.

[3] Potthast, M., Stein, B. and Anderka, M. A Wikipedia-Based Multilingual Retrieval Model. In *Proceedings of 30th European Conference on Information Retrieval* Glasgow, Scotland, 30th March - 3rd April 2008, 2008. Springer, 522-530.

[4] Sorg, P. and Cimiano, P. Cross-language Information Retrieval with Explicit Semantic Analysis. In *Proceedings of the Working Notes of the CLEF 2008 Workshop* Aarhus, Denmark, 17-19 September 2008, 2008.

[5] Zhou, D., Truran, M., Brailsford, T. and Ashman, H. A Hybrid Technique for English-Chinese Cross Language Information Retrieval. *ACM Transactions on Asian Language Information Processing (TALIP),7(2),* 2008. 1-35.

[6] Zhou, D. and Wade, V. Latent Document Re-Ranking. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP),* Singapore, 2009. ACL, 1571-1580.

[7] Zhou, D., Lawless, S., Min, J. M., and Wade, V. Dual-Space Re-ranking Model for Document Retrieval. In *Proceedings of The 23rd International Conference on Computational Linguistics (COLING 2010),* Beijing, China, 2010.