Combining Signals for Cross-Lingual Relevance Feedback

Kristen Parton[1] and Jianfeng Gao[2]

[1] Columbia University, New York, NY, USA
kristen@cs.columbia.edu
[2] Microsoft Research, Redmond, WA, USA
jfgao@microsoft.com

Abstract. We present a new cross-lingual relevance feedback model that improves a machine-learned ranker for a language with few training resources, using feedback from a better ranker for a language that has more training resources. The model focuses on linguistically non-local queries, such as [world cup] and [copa mundial], that have similar user intent in different languages, thus allowing the low-resource ranker to get direct relevance feedback from the high-resource ranker. Our model extends prior work by combining both query- and document-level relevance signals using a machine-learned ranker. On an evaluation with web data sampled from a real-world search engine, the proposed cross-lingual feedback model outperforms two state-of-the-art models across two different low-resource languages.

1 Introduction

Modern web search engines reply heavily on data-driven approaches that go beyond traditional information retrieval (IR) models by incorporating additional features into machine-learned rankers. Typical ranking features include static link analysis features like PageRank, click-through data and document classifiers [2, 6, 10]. The quality of a learned ranker depends to a large degree upon the amount of training data such as human relevance judgments, user feedback and the size of the index or web-graph.

The web is a global resource, serving users in hundreds of regions who speak hundreds of different languages. Optimizing a web search ranker for each of these language/region settings, or markets, is an expensive process, requiring a great deal of annotated data. Even after collecting annotations, ranking features derived from click-through data may not be available for markets with small numbers of users, while link analysis features such as PageRank may not be as helpful for nascent markets with fewer documents and links. Rather than collecting expensive annotated data for each new low-resource market, several strategies have been applied to exploit existing data or models. One approach is to exploit a market with more training data, such as English/US, via model adaptation (e.g. [1, 7]). Another approach, which we explore in this paper, is to use cross-lingual feedback from a high-resource market.

In this study we focus on linguistically non-local (LNL) queries, defined by [8] as concepts that are searched for by users in different markets. For instance, the concept [world cup] [copa mundial] and [coupe du monde] are LNL since they are all about the world cup. In contrast, [brooklyn beaches] is a local query. In practice, a query in

Y. Hou et al. (Eds.): AIRS 2012, LNCS 7675, pp. 356–365, 2012.
© Springer-Verlag Berlin Heidelberg 2012

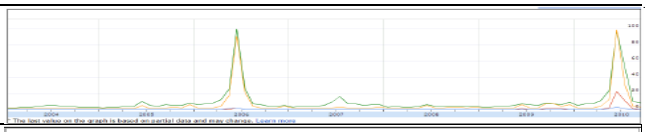
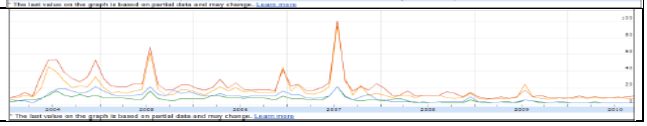
[world cup] 2004 – 2010	
[world cup] 2010 only	
[jk rowling] France, UK, US, Spain	

Fig. 1. Query volume over time for three LNL queries

language L1 is considered LNL if it has a high-confidence translation into a language L2 and the query translation is found in L2 query logs. By this metric, [8] found that at least 11.5% of Chinese queries from their dataset were LNL. Figure 1 shows query volume over time for several LNL queries. Queries in Korean, Russian, Arabic and Spanish for the concept [world cup] follow similar distributions over time, both on a large scale (2004 - 2010) and a much shorter time frame (summer 2010). Although this is not surprising, since the world cup is a time-constrained event that occurs every 4 years, it does suggest that these queries share similar user intent, even though they are in different languages. The query [jk rowling] also shows similar query volume over time in four European markets (France, UK, US, Spain), again indicating related user intent, possibly related to the publication of Harry Potter books or news headlines.

Cross-lingual relevance feedback works by retrieving results for a LNL query in the original language, L1, as well as an assisting language, L2, with a better ranker. Results from L2 are assumed to be better than L1 results, and can be used to improve L1 results, with the help of a translation dictionary. Note that this is not an instance of cross-lingual information retrieval, since the goal is still to return results in L1 only.

Our model, called the unified model, generalizes the existing cross-lingual relevance feedback models by incorporating both query expansion and document re-ranking to further amplify the signal from the high-resource ranker. We use a learning to rank approach, which requires labeled training data. Unfortunately, there are no publicly available corpora with relevance judgments for non-English web search. The datasets used in our experiments are sampled from real-world datasets indexed by a commercial search engine. We present experiments on datasets from two markets, Korean/South Korea and Russian/Russia, using English/US as the assisting market. Our evaluation shows that the proposed unified model outperforms two previous cross-lingual relevance feedback models across two different domains.

2 Cross-Lingual Relevance Feedback

Our unified model extends two previous models, MultiPRF, by Chinnakotla et al. [5], and the model proposed by Gao et al. [8], which we will refer to as DocSim. In this section, we review these models before presenting the unified model.

The baseline IR system used in our experiments is based on the language modeling (LM) framework. In this approach, documents are ranked by the similarity of their LMs θ_d to the query LM θ_q, using Kullback-Leibler divergence. Document language models are smoothed using the collection LM via Dirichlet smoothing [13]. Since search queries are often very short (2 or 3 words), the query LM θ_q is very limited. Pseudo relevance feedback (PRF) attempts to overcome this problem by assuming that the top *n* documents are relevant and extracting additional *k* query terms from them. The terms are weighted according to how often they appear in the feedback documents and how relevant the feedback documents are to the original query. The feedback relevance model θ_f is then combined with the original query model using a mixture model [11], which we will refer to as monolingual PRF (MonoPRF).

Chinnakotla et al. [5] extended the monolingual PRF model to include cross-lingual documents. Given an LNL query and search results in L1 and L2, PRF is performed in both languages. PRF terms from L2 are translated back into L1 using a probabilistic translation dictionary $P(f|e)$: $P(f|\theta_{trans}) = \sum_e P(f|e)P(e|\theta_f)$. The final model is called multilingual PRF (MultiPRF) because it does PRF in the query language (L1, θ_f) as well as in the assisting language (L2, θ_{trans}), combining them with a mixture model of Equation (1)

$$P(w|\theta_q'') = (1-\lambda-\gamma)P(w|\theta_q) + \lambda P(w|\theta_f) + \gamma P(w|\theta_{trans}) \quad (1)$$

The intuition behind the MultiPRF model is that the L2 corpus is larger than the original L1 corpus, so there are likely to be more relevant documents in L2. Doing query translation (to retrieve the assisting language feedback documents) and then back translation (to translate back the PRF model) also yields a query expansion effect, i.e., synonyms and related terms are added to the query.

Gao et al. [8] introduced the concept of LNL queries, and presented a document retrieval model for cross-lingual relevance feedback, which we will refer to as DocSim. Given an LNL query and search results in L1 and L2, a weighted bipartite graph is created over the documents, connecting L1 documents with L2 documents. The weight of each edge is the cross-lingual document similarity, which is calculated via cross-lingual cosine similarity. Finally, a relational ranking support vector machine is applied so that the ranks of L1 documents move closer to the ranks of similar L2 documents. For example, the official world cup webpage in Arabic is very similar to the official world cup webpage in English, so if the English page is ranked highly, the DocSim model will re-rank the Arabic page to also have a high rank.

3 The Unified Model

Both the MultiPRF model and the DocSim model are motivated by the same observation that there is a high-resource ranker in L2 that has better monolingual accuracy than the L1 ranker. But they are developed based on two different, yet complementary, assumptions. MultiPRF exploits the fact that L1 and L2 queries have shared *query intent*, and works via cross-lingual query expansion. In contrast, the DocSim model assumes that the documents that are relevant to an LNL query contain *related document content*, albeit in different languages.

Our unified model is intended to build on both complementary assumptions, and is a significant extension of the previous research in two aspects. First, we extend both the MultiPRF model and the DocSim model to handle web document structure. Second, the unified model takes the learning to rank framework to which a wide variety of features based on cross-lingual relevance feedback, e.g., those derived from both MultiPRF models and DocSim models, rather than being just limited to query expansion (MultiPRF) or document similarity (DocSim), are incorporated. In our implementation, we used a neural net ranker, called LambdaRank [4], which has been shown empirically to optimize NDCG (Normalized Discounted Cumulative Gain [9]). In the next section we describe the web document structural aspect of the features used by the model, and in the following section we explain the features used in the ranker in detail.

3.1 Web Document Structure

Web documents consists of multiple streams, or fields, which can be divided into content streams, such as url, title and body, and popularity streams, such as anchor text and queries used to access the page. [6] analyzed cross-stream perplexity and found that different language styles are used for composing the document body, title, anchor text, and queries. For instance, the anchor language model is more similar to the query language model than the body language model is. Therefore, each stream should be modeled separately and combined, rather than modeling the document as a single bag of words extracted from different streams. Similarly, BM25F combines weighted term frequencies from different fields, recognizing that some fields are more salient than others [12].

For cross-lingual relevance, document structure is important because popularity fields are the most useful for estimating relevance, but are also more likely to be missing for low-resource languages. Cross-lingual relevance feedback can project popularity fields from the richer market back onto the low-resource market. One potential pitfall could be translation, since popularity fields (such as anchor texts and user queries) are short and have very little context, so they are harder to translate accurately with machine translation systems than body text (or even title text), which usually consists of full sentences.

Another major advantage of incorporating web document structure into the model is speed. If relevance can be approximated by shorter document fields (such as anchor text or title), then doing cross-lingual document similarity is much cheaper. Each document similarity calculation involves word-by-word translation and then cosine similarity, and for example the model of Gao et al. [8] does $n \times m$ similarity calculations.

3.2 Ranking Features

The features used in the ranker can be grouped into three categories, monolingual features, MultiPRF features, and DocSim features.

Monolingual features include baseline ranking features that are used in almost all web search ranking models, such as PageRank (which is query-independent) and

BM25F (which is query-dependent). A baseline retrieval model and monolingual PRF model were built for each document stream. In our experimental dataset, documents have four streams: body, title, url and anchor text, as well as a bag-of-words stream "allfields". Ranking scores for each stream were defined as monolingual features.

MultiPRF features are derived as follows. A MultiPRF model was built for each document stream, and ranker scores for each stream were used as a feature. Overall, there are 5 MultiPRF features.

DocSim features are derived as follows. A single L1 document is compared to each L2 document, and then the cross-lingual similarity score, defined as a DocSim feature, is normalized and combined using a weighted average. Intuitively, this score represents the rank of similar L2 documents. This DocSim feature is computed for each of the 4 document streams. Two standard similarity functions were applied to each document-feedback document pair: Jaccard similarity and cosine similarity. As in [8], cross-lingual similarity was calculated using a translation dictionary. In addition to the cross-lingual similarity functions, similarity functions without translation were also used as features. The goal was to capture transliterations, translations and Latin spellings, which were particularly important for the url field, since the URLs were all in Latin. Certain words may also appear in Latin, even when the document is in another language (e.g., "windows"). Overall, there were two monolingual and two cross-lingual similarity functions for each stream, and 5 streams, for a total of 20 DocSim features.

4 Experimental Setup

4.1 Data

The unified model targets monolingual search in languages with few training resources. We used a re-ranking experimental paradigm where we try to improve the web search results by re-ranking documents retrieved from the entire web using a commercial search engine. We used data from two language/region settings that are linguistically and culturally different from the English/US setting, to see how well feedback from a better ranker from an unrelated domain can improve results. The domains we selected were Korean/South Korea and Russian/Russia. They are quite different from each other in order to see how well the model generalizes across domains.

Since we are only interested in linguistically non-local (LNL) queries, as defined above, we further filtered the data by selecting queries with high confidence translations and queries whose translations were present in English/US query logs. All queries were translated into English with the Bing translator public API[1]. Translations were considered high-confidence if back-translation produced a fuzzy match to the original query. The high-confidence English query translations that occurred in a large set of English/US queries were selected as LNL queries.

Given a LNL query, the English query translation was passed to the public Bing API[2] and the top 50 results were retrieved. Each result consists of a URL, title and

[1] http://www.microsofttranslator.com
[2] http://www.bing.com/developers

snippet. For each query, all URLs that were annotated for relevance were crawled, and their anchor texts were also retrieved. Many documents could not be crawled, due to dead pages or errors. Only queries with 10 or more judged documents and non-empty feedback results were kept. Table 1 shows the statistics of the final evaluation dataset used in our experiments.

Table 1. Evaluation datasets

Domain	Queries	Documents	Avg. docs/query
Korean/South Korea	134	1,986	14.8
Russian/Russia	102	1,257	12.3

4.2 IR Setup

Each crawled document was parsed and split into different streams (fields): url, title, body, anchor text and everything, which included all the other streams. Each feedback result was also split into these streams[4], and the body field was replaced with the snippet. Since the snippet contains a small amount of text highly relevant to the query, using the snippet instead of the full document retains the signal from the feedback documents, while greatly speeding up the document comparison calculation.

A unigram index was built for each document stream. Each stream was tokenized according to the document language. We used a Viterbi decoder based on a unigram model to break a URL string into tokens. As in the World Wide Web, documents from different languages exist in the same global corpus, although real search engines have more sophisticated techniques for region and language matching.

5 Results

5.1 PRF Baselines

The baseline IR model is defined in Section 2 and has only one tunable parameter, the Dirichlet parameter. The monolingual PRF model (MonoPRF) does PRF based on documents returned by the baseline IR model, and has three additional parameters: the number of feedback documents, the number of feedback terms and the mixture model parameter λ, as in Equation (1) where $\gamma = 0$. The MultiPRF model is a mixture model over the MonoPRF model and the cross-lingual PRF model, as defined in Equation (1). It has four additional parameters: the number of cross-lingual feedback documents and terms, the number of translations per feedback term, and the mixture model parameter γ.

In our experiments, the model parameters were tuned using leave-one-out cross-validation and grid search. Each model's parameters were tuned separately, so for instance, the Dirichlet parameter could end up being different for baseline, MonoPRF and MultiPRF.

Results comparing the baseline IR model with monolingual and multilingual PRF on the two LNL web datasets are shown in Figure 2. MultiPRF outperforms the baseline and the monolingual PRF model in most cases (except for NDCG at 1 for Korean). The improvements for the Russian domain are all statistically significant.

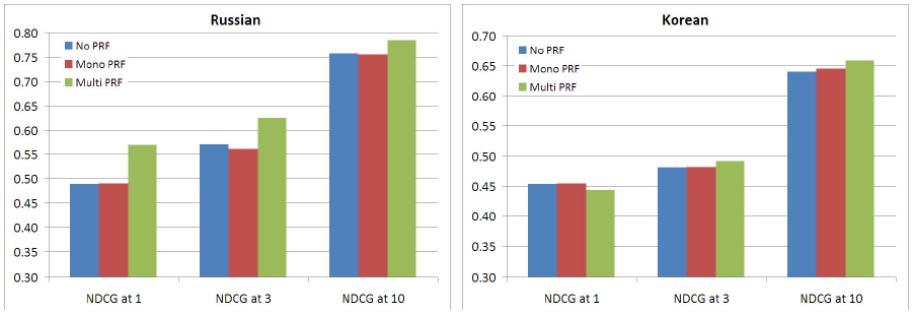


Fig. 2. Results of PRF baselines

5.2 DocSim Baseline

The DocSim model uses LambdaRank to learn a ranking model, based on the 4 document similarity features computed only on the "allfields" stream of each document, as described in Section 4.2. The LambdaRank model was a single layer neural network with 200 iterations. All reported results are from 5-fold cross-validation.

In contrast with the MultiPRF baseline, which uses query expansion for feedback, the DocSim model learns to rank L1 documents similar to the rank of similar L2 documents, based on the assumption that the L2 ranker is better. If too few feedback documents are used, there may be no similar documents to learn from. However, if too many are used, there may be too much noise in the features for the model to learn a coherent ranker. Results of applying the DocSim model with different numbers of feedback documents are shown in Figure 2. With a small number of feedback documents, the results are often worse than the baseline. However, as the number of feedback documents increases, the NDCGs improve. For the Korean domain, 25 feedback documents performed best, while for the Russian domain, 50 feedback documents were best.

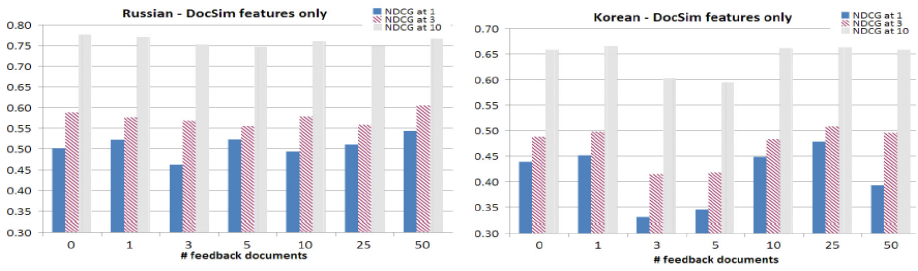


Fig. 3. Results of DocSim baseline

Table 2. Machine-learned rankers with differen features

Models	Russian NDCG at 1	NDCG at 3	NDCG at 10	Korean NDCG at 1	NDCG at 3	NDCG at 10
Baseline	50.13	58.80	77.56	43.99	46.94	65.47
MultiPRF	54.67	61.88	79.41	48.61	51.31	67.89
DocSim	55.59	59.79	77.14	47.88	50.83	67.67
Unified	56.27	62.80	79.35	47.30	51.74	68.46

5.3 Unified Model

For the final comparison, machine-learned rankers were built using baseline features (PageRank and monolingual BM25F) plus features from each model. The MultiPRF ranker had as features ranker scores from the MultiPRF rankers, while the DocSim ranker had all the DocSim features. The unified ranker had all features. Results in Table 2 shows that cross-lingual relevance feedback almost always outperforms the monolingual baseline. For the Russian domain, the unified model outperforms both the MultiPRF model and the DocSim model, except at NDCG at 10, where MultiPRF does slightly better. For the Korean domain, the unified model outperforms both DocSim and MultiPRF at NDCG at 3 and 10, but does worse at NDCG at 1. Overall, the unified model is the best performer in our experiments across both datasets.

6 Discussion and Examples

The motivating hypotheses behind the cross-lingual relevance feedback model was that linguistically non-local (LNL) queries in different languages have similar query intent and relevant documents have related content, so a poor ranker should be able to get direct feedback from a better ranker in another language. The original experiments with MultiPRF used CLEF collections, where the queries were compared over the different sites to ensure that a high percentage of them will find some relevant documents in all [language/domain] collections [3]. In contrast, extracted LNL queries were simply those that had been searched for in both languages, so there may not be as much relevant content in L2.

Surprisingly, queries that skewed heavily towards L1 were not always harmed by MultiPRF. For instance, the queries [환율] (naver) and [в контакте] (in touch) are both navigational queries to popular local websites. In the first case, the L2 ranker already knows that [naver] is a navigational query to the Korean website, so the feedback only helps. In the second case, the highest weighted English terms are still relevant, and the irrelevant terms have much lower weight.

As expected, MultiPRF did harm queries when the query translation was bad. Although query translation used a state-of-the-art MT system and the translations were filtered for "high confidence" translations via back translation, some queries were still translated poorly: [한국 일보 미국] was "hankook ilbo usa" (a partial transliteration) instead of "korea times usa"; [гадание] was "divination", instead of the more colloquial "fortune telling" or "palm reading".

Table 3. LNL queries and their retrieval results

LNL query (translation)	MonoPRF (translation)	L2 PRF	
живая природа (wildlife)	это (this), автор (author), 2010, раздел (section), alexey, далее (more), читать (read), природы (nature)	wildlife, animals, fish, service, www, utah, society, colorado, us, texas	Topic drift towards irrelevant L2 domain
работа во франции (work in France)	францию (France), туры (tours), франции (France), франция (France), работа (work), ru, отдых (relax), www, au, pair	france, work, french, employment, working, visa, travel, living, abroad, visas	L2 focuses more on work than tourism

However, even queries that are truly LNL and correctly translated can be harmed by MultiPRF. For instance, in the first example in Table 3, searching for [wildlife] in the English/US domain brings up many US-specific wildlife associations, which harm the Russian results. In the second example, the English results help re-focus the query towards working and living in France (and getting visas) instead of visiting and touring France.

7 Conclusions and Future Work

We presented a cross-lingual feedback model that aims to improve a ranker from a market with few training resources using feedback from a better ranker with richer training resources. Focusing on linguistically non-local queries allows the model to use direct feedback from the better ranker, rather than just using domain adaptation. Our model extends and generalizes prior work by incorporating both query- and document-level features. Query expansion using multilingual pseudo-relevance feedback exploits the similar *intent* of the original query and the translated query, while the document similarity features leverage related content in both languages, using translation dictionaries to bridge the cross-lingual gap. The model incorporates web document structure to further amplify the noisy signal from the better ranker. The cross-lingual unified relevance feedback model outperformed the monolingual baseline across two different domains.

While the results of this pilot study are promising, the biggest hurdle we faced was data size, and in future work we would like to apply our model to a much larger dataset. Unfortunately, we are unaware of any publicly available web search relevance judgments for languages other than English.

Another promising direction is to exploit more cross-lingual web features. For example, there are many cross-lingual anchor texts (e.g., English links pointing to Chinese pages) and user clicks (e.g., Russian queries that lead to English pages). These types of features would give stronger evidence of shared content across cross-lingual documents, or shared cross-lingual query intent. The English/US domain is also richer in popularity fields (such as anchor text and clickstream features) than some other domains. Exploiting this structural asymmetry should improve the feedback model even more.

References

1. Bai, J., Zhou, K., Xue, G., Zha, H., Sun, G., Tseng, B., Zheng, Z., Chang, Y.: Multitask learning for learning to rank in web search. In: CIKM (2009)
2. Bennett, P.N., Svore, K.M., Dumais, S.T.: Classification-enhanced ranking. In: WWW, pp. 111–120 (2010)
3. Braschler, M., Peters, C.: Cross-language evaluation forum: Objectives, results, achievements. Information Retrieval 7, 7–31 (2004)
4. Burges, C.J.C., Ragno, R., Le, Q.V.: Learning to rank with nonsmooth cost functions. In: NIPS, pp. 193–200 (2006)
5. Chinnakotla, M.K., Raman, K., Bhattacharyya, P.: Multilingual prf: english lends a helping hand. In: SIGIR, pp. 659–666 (2010)
6. Gao, J., He, X., Nie, J.-Y.: Click through-based translation models for web search: from word models to phrase models. In: CIKM, pp. 1139–1148 (2010)
7. Gao, J., Wu, Q., Burges, C., Svore, K.M., Su, Y., Khan, N., Shah, S., Zhou, H.: Model adaptation via model interpolation and boosting for web search ranking. In: EMNLP, pp. 505–513 (2009)
8. Gao, W., Blitzer, J., Zhou, M.: Using english information in non-english web search. In: Proceeding of the 2nd ACM workshop on Improving Non English Web Searching, iNEWS, pp. 17–24 (2008)
9. Jarvelin, K., Kekalainen, J.: Ir evaluation methods for retrieving highly rele-vant documents. In: SIGIR, pp. 41–48 (2000)
10. Joachims, T.: Optimizing search engines using clickthrough data. In: KDD, pp. 133–142 (2002)
11. Lavrenko, V., Croft, W.B.: Relevance-based language models. In: SIGIR, pp. 120–127 (2001)
12. Robertson, S.E., Zaragoza, H., Taylor, M.J.: Simple bm25 extension to mul-tiple weighted fields. In: CIKM, pp. 42–49 (2004)
13. Zhai, C., Lafferty, J.D.: A study of smoothing methods for language models applied to ad hoc information retrieval. In: SIGIR, pp. 334–342 (2001)