Cross-Language Pseudo-Relevance Feedback Techniques for Informal Text

Chia-Jung Lee and W. Bruce Croft

Center for Intelligent Information Retrieval
School of Computer Science, University of Massachusetts, Amherst
{cjlee,croft}@cs.umass.edu

Abstract. Previous work has shown that pseudo relevance feedback (PRF) can be effective for cross-lingual information retrieval (CLIR). This research was primarily based on corpora such as news articles that are written using relatively formal language. In this paper, we revisit the problem of CLIR with a focus on the problems that arise with informal text, such as blogs and forums. To address the problem of the two major sources of "noisy" text, namely translation and the informal nature of the documents, we propose to select between inter- and intra-language PRF, based on the properties of the language of the query and corpora being searched. Experimental results show that this approach can significantly outperform state-of-the-art results reported for monolingual and cross-lingual environments. Further analysis indicates that inter-language PRF is particularly helpful for queries with poor translation quality. Intra-language PRF is more useful for high-quality translated queries as it reduces the impact of any potential translation errors in documents.

Keywords: Informal text, discussion forum, cross-language information retrieval, pseudo-relevance feedback

1 Introduction

The task of cross-lingual information retrieval (CLIR) attempts to bridge the mismatch between source and target languages using approaches such as query and document translation. Previous techniques [1][31] addressed CLIR using different benchmark collections[1] that are written with relatively formal language (e.g., news articles). There are, however, many applications that involve relatively informal language. Social sites like discussion forums often contain slang or abbreviations that will result in frequent translation errors. Effective CLIR for this type of data may require modifications to the existing techniques.

Informal text poses several problems for machine translation (MT). It is likely that both document and query translations contain a significant proportion of mistranslated and untranslated terms. An example query from our data set is: "*What are people saying about real-name tweeting registration that was imposed*

[1] CLEF and NTCIR are two examples.

on March 16th 2012 in China" in source language English. The term *tweeting* was translated into 啁啾 (chirping) in the target language, Chinese. The cause of this mistranslation lies in the parallel collection used to train the MT engine, where the only instances of *tweeting* referred to birds *chirping*. Similarly, correct document translation is complicated by the creativity of social site users. For example, 微博 (Weibo) is often written as 围脖 (surrounding neck) in forum documents. This is an amusing pun because the pronunciation of both terms is identical in Chinese. These types of noise, in addition to the noise produced by translation errors, can easily result in significant topic drift in translated queries and collections.

In this paper, we explore techniques to improve CLIR for a collection of informal text. We propose a new technique based on feature-driven cross-lingual pseudo relevance feedback that expands translated queries with intra-language or inter-language feedback terms. Given source queries Q_s and target corpus C_t as input, both query and document translation are performed to generate the translated $Q_{T(s)}$ and $C_{T(t)}$. We propose to use inter- and intra-language PRF to reduce noise produced by poor query and document translation, respectively. Intra-language PRF extracts terms from top ranked documents in C_t retrieved by $Q_{T(s)}$. This type of feedback helps to mitigate any drop in performance due to poor document translations. For poorly translated queries, we consider the opportunity of recovering semantics from PRF performed on source language. Accordingly, inter-language PRF first retrieves documents in $C_{T(t)}$ using Q_s, based on which it then locates the aligned documents in C_t and extracts feedback terms. For example, despite *tweeting* being translated to *chirping*, retrieval on the source corpus is able to discover documents discussing *real-name registration*. We may recover the lost term, *tweeting*, by locating their parallel documents in the target language and directly extracting terms from these documents.

We evaluate existing techniques and our selective PRF model using a recently created collection of web forum posts. Test queries were manually created and associated relevant posts were judged using a pooling technique across multiple retrieval techniques. We explore the language pair English and Chinese[2]. Our aim is to select between intra- and inter-language PRF for each query. Experimental results show that this selective PRF model can significantly improve retrieval performance over several monolingual and cross-lingual baselines.

The remainder of this paper is laid out as follows. Section 2 discusses related work on CLIR and PRF. Section 3 describes the proposed approach. We show the evaluation results in Section 4 and Section 5 concludes the paper.

2 Related Work

Figure 1 provides an overview of CLIR approaches that have been studied. For clarity, we denote the languages used in original query and document collection respectively as L_1 and L_2. Thus the translated query and document would be

[2] The crawled forum data is Chinese, the queries are posed in English, and documents were judged in Chinese.

generated in languages L_2 and L_1. Query translation [5][27] is one of the most common methods of bridging the language gap. This approach first translates the original topics into L_2, performs monolingual retrieval on the language-compatible index, and produces a single ranked list of documents. Alternatively, document translation [3][15] translates the original document collection into topic language L_1 and the original queries are used for retrieval. Hybrid approaches [6] consider both query and document translation. As a retrieval system generates a final single ranked list, approaches that conduct multiple retrieval runs need to merge lists using fusion techniques [9][18][25]. Gey et al [11] proposed bypassing the merge process by building a single index that contains documents in all languages and concatenating a source query with all language translations as a new query. Chen et al [5], however, showed that the combined approach is empirically less effective than the query translation approach. Studies have also been conducted on different translation techniques [23][33].

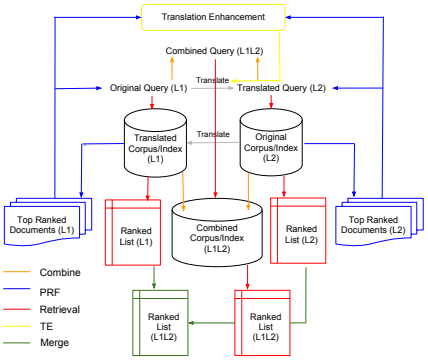

Fig. 1: An overview of related work on CLIR.

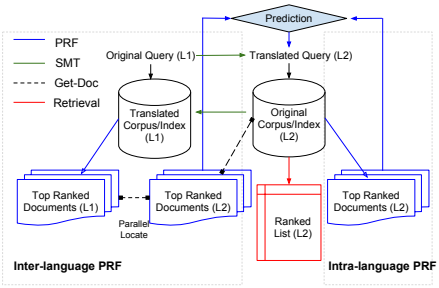

Fig. 2: An overview of the proposed approach.

The practice of PRF has been effectively applied in various monolingual and cross-lingual IR environments. For monolingual retrieval, a number of approaches [16][29][32] have been proposed to improve the performance and robustness of PRF. Metzler et al. [21] proposed a feature-based approach, latent concept expansion, to model term dependencies. For cross-lingual settings, PRF can be applied in different retrieval stages of pre-translation, post-translation [25] and the combination of both [1][19][26]. Chen and Gey [6] further demonstrated that rank fusion of returned lists obtained from L_1 and L_2 is able to improve retrieval performance in most circumstances. In addition to expanding queries, as shown at the top of Figure 1, He and Wu [13] used PRF to enhance query translation by adjusting translation probabilities and resolving out-of-vocabulary terms.

Recent research has shown that monolingual retrieval performance can be improved by the use of another assisting language. Chinnakotla et al [7] used a second language to improve the performance of PRF based on a framework "MultiPRF". Similarly, Na and Ng [22] proposed translating documents into an auxiliary language which then served as a semantically enhanced representation for supplementing the original bag of words.

3 OUR APPROACH

Query expansion using PRF provides a method for automatic local analysis [30]. A typical instantiation is to retrieve an initial set of documents with the original query, assume the top retrieved documents are relevant, and expand the original query with the significant terms extracted from those pseudo-relevant documents. The newly constructed query is used to retrieve the final result.

Although PRF has been shown to be helpful for CLIR [1], it is not clear how effective PRF can be for a collection of informal text and its associated queries. Queries of different characteristics may benefit differently from different feedback techniques. In particular, when query translation quality is poor, we hypothesize that inter-language PRF is more useful, considering the *tweeting* example in Section 1. On the other hand, intra-language PRF can be more helpful when query translation quality is good as it reduces the impact of any potential translation errors in documents. Document translation errors can happen often, particularly in a collection of informal text. For example, for the query *How should I maintain a car?*, we identify some documents losing relevance because 汽车的脚, meaning "feet" of a car, was used to refer to tires of a car in Chinese. Yet it has nothing to do with car tires after it is translated into *motor vehicles of feet*. Translation of named entities can also cause problems in translated documents. For example, 富士康 (*Foxconn*) is translated into *fuji kang* in our system.

Noting that queries can benefit differently from different types of PRF, we propose to select between inter- and intra-language PRF for each individual query, as opposed to previous work that mostly focused on applying the same technique to all query instances. Figure 2 provides a high level overview of our approach. In the following, we denote a user-issued query as Q_{L_1} in source language L_1 and the aim is to retrieve a set of documents from corpus C_{L_2} in target language L_2. Query and document translations are performed to generate Q_{L_2} and C_{L_1} resulting in a bilingual dataset.

3.1 Intra-language and Inter-language PRF

Intra-language PRF is an implementation of PRF using only the target language L_2. The translated queries Q_{L_2} are issued against the target language index I_{L_2} of the collection C_{L_2}. The set of pseudo relevance feedback terms are then extracted from the top ranked documents. Metzler and Croft [21] further proposed Latent Concept Expansion (LCE) that models term dependencies during expansion and have showed significant improvements in retrieval effectiveness. The terminology, *intra-language* PRF (intra-PRF), is used to indicate that other language, L_1, is not used in the process, as shown in the right part of Figure 2.

Inter-language PRF (inter-PRF) modifies the PRF framework to use the translated collection. We construct inter-PRF by first retrieving documents from the translated corpus C_{L_1} using source queries Q_{L_1}, as shown in the left part of Figure 2. One difference from intra-PRF is that the retrieved documents are now written in source language L_1, meaning that the feedback terms cannot be directly compared against target language index I_{L_2}. We therefore locate the

target language documents aligned with the retrieved set and extract significant terms directly from those documents in C_{L_2}. These terms are then used to expand the translated query, Q_{L_2}, for the second retrieval pass.

3.2 Selecting between Intra-PRF and Inter-PRF

Another contribution of this paper is to compare the relative utility of these different feedback techniques and properly integrate them together. We hypothesize that inter-PRF is more helpful for queries that are malformed due to poor query translation. Meanwhile, intra-PRF is better for correctly translated queries and can compensate for document translation errors. To combine relevance signals from inter- and intra-PRF, we formulate the problem as a classification task, such that either intra-PRF or inter-PRF is selected as feedback to the input query using a per-query basis.

To integrate the two sets of PRF terms, we propose to estimate the weights $\{\phi_i\}$ that assess the importance of each component, constrained by $\sum_i \phi_i = 1$ and $\phi_i \geq 0$, into the retrieval model as shown in Equation 1:

$$\begin{aligned} sc(Q, D) &= \theta \cdot g_q(Q, D) + \bar{\theta} \cdot g_p(PRF, D) \\ &= \theta \cdot g_q(Q, D) + \bar{\theta} \cdot \sum_{\substack{i \in \{intra,\\ inter\}}} \phi_i \cdot g_p(PRF_i, D) \end{aligned} \tag{1}$$

In Equation 1, $g_q(\cdot)$ and $g_p(\cdot)$ are retrieval functions taking query or feedback terms as input, based on which the functions search the index and produce relevance scores. Parameter $\theta \in [0, 1]$ controls the belief in relevance estimators and $\bar{\theta} = 1-\theta$. For brevity, Q denotes a query in target language L_2, and PRF_{intra} and PRF_{inter} respectively represent the set of corresponding PRF terms.

We further need to estimate the weights $\{\phi_i\}$ guiding the model to incorporate PRF terms either from the intra- or inter-language method. Given a retrieval performance metric $m(.)$, we can represent the effectiveness of a query with PRF by $m(Q, PRF_{intra})$ or $m(Q, PRF_{inter})$. We then define the relative effectiveness of using intra- and inter-PRF as $\hat{m} = m(q, PRF_{intra}) - m(q, PRF_{inter})$. Intuitively, a query is more suited to intra-PRF if $\hat{m}$ is positive while inter-PRF is more suitable if $\hat{m}$ is negative. A classification decision is then naturally guided by the sign function $sign(\hat{m})$ where labels $+1$ and -1 respectively stand for the choice of intra- and inter-PRF. Supposing a classifier c is established based on $sign(\hat{m})$, we use a binary belief estimator $\mathbb{1}_c$ that takes the predicted results from c as input and outputs a binary value 0 when $c = -1$ or 1 when $c = +1$. Replacing $\{\phi_i\}$ in Equation 1 with output of indicator function $\mathbb{1}_c$ we can obtain a rigidly-classified retrieval model.

In addition, we can estimate $\{\phi_i\}$ probabilistically such that the estimation would alleviate the penalty of binary mis-classification. Specifically, we transform the Support Vector Machine (SVM) prediction results into posterior probabilities [10][24], and incorporate these probabilities into the retrieval model indicating the importance of each PRF set. Alternatively, the hyper-parameter learning

framework proposed [2] could be used for learning the cross-lingual weights. We leave this as future work.

The retrieval functions $g_q(\cdot,\cdot)$ and $g_p(\cdot,\cdot)$ can be modeled using a wide variety of methods. Metzler and Croft [20] proposed an effective family of retrieval models using Markov Random Field (MRF) for term dependency modeling. This approach has consistently outperformed bag-of-words retrieval models. We use the Sequential Dependency Model (SDM) [20], an instantiation of the MRF model, to compute $g_q(Q, D)$ for a query Q and a document D. For PRF terms, we adopt unigram query likelihood to compute $g_p(\cdot,\cdot)$ as dependency between feedback terms is less likely to improve effectiveness.

3.3 Features

The feature set, consisting a total number of 42 features, is used to select between intra- and inter-PRF for each query.

We compute the number of nouns, verbs, and named entities in Q_{L_1} and Q_{L_2}. The degree of conformity between these statistical distributions can be good indicators about how well the syntactic structure and semantics have been retained after the translation process.

He and Ounis [12] proposed using pre-retrieval predictors for inferring query performance. For a query predicted to have high performance, using PRF from the same side of language could be more effective than that of another side. We adopt the simplified query clarity score (SCS) and query scope (QS) from [12] to characterize both Q_{L_1} and Q_{L_2}. More details can be found in [12].

We compute the collection frequency and inverse document frequency of each query term $t_i \in Q_{L_1}$ and $t_i \in Q_{L_2}$. We then generate the features by taking the minimum, maximum, and average frequencies among the results of each language respectively. We additionally include the standard deviation and the max min ratio of *idf* as features as in [12]. Features based on the query length of Q_{L_1} and Q_{L_2} are included to provide an estimate of the length variation after translation.

Intuitively, for a query, if the retrieved document set for each language is similar, the translation quality for both the query and the document set is likely to be high. Accordingly, we compute the intersection of the retrieved document ranked lists produced by $Q_{L_1} \rightarrow C_{L_1}$ and $Q_{L_2} \rightarrow C_{L_2}$ at ranks 10, 100, 500, and 1000. More sophisticated post-retrieval predictors such as [8][14] can be used.

We consider the degree of collection coherence between query terms and feedback terms. Denoting a feedback term for a query Q as z_i and a query term as t_j, we compute Pointwise Mutual Information (PMI) between the pairwise combinations of z_i and t_j for query instance Q. PMI provides a semantic similarity measure that sorts lists of important neighbor words from a large corpus.

4 Experiments

4.1 Experimental Setup and Dataset

Document Set: All of our experiments are conducted over a collection of Chinese (L_2) language forum posts. All posts were translated into English (L_1) using

state-of-the-art machine translation techniques, as part of the Broad Operational Language Technology project (BOLT) [3]. The original data set of 287,783 threads was collected by the Linguistic Data Consortium (LDC)[4]. Threads may contain multiple posts. By splitting the threads into posts, we obtain 2,416,869 posts (documents). The dataset spans a wide range of discussions and documents with informal language are prevalent across the entire collection.

Query and Relevance Judgments: We manually constructed 50 natural language English queries Q_{en} and aim to retrieve documents from the Chinese forum collection. With forum posts being the search target, we created queries that request suggestions or opinions on current events/topics that are common in social media such as discussion forums. We generated relevance judgments based on a three level assessment $\{0, 1, 2\}$, representing non-relevant, partially relevant and relevant. The judgments were done by two bilingual assessors and a document is assigned the maximum relevance level of the two. Relevance judgments were collected using a pooling method based on multiple well-known retrieval approaches. A total number of 2,495 judged documents were collected (49.9 posts/query), among which 1,072 were judged relevant (21.44 posts/query). The ratio of the number of relevant documents to the total number of judged is relatively high partly because quoting and reusing are popular in discussion forums.

Statistical Machine Translation of Queries: Several machine translation tools were used to provide the noisy channels that we are investigating in this paper. Our first translation engine was the open source tool Moses and Giza++ [5] that implements the statistical (or data-driven) approach using sentence-aligned English and Chinese Sougo news articles corpus. Our alternative translations for queries come from the BOLT project and the Google Translate tool. Our three translated query sets are denoted as Q_{mos}, Q_{bolt} and Q_{gt}. In total, we explore 4 types of instances of the 50 query topics, including 1 original query (English) and 3 query translation instances (Chinese).

Other Open Source Tools: Indexes are created for both original C_{L_2} and translated C_{L_1} corpora using Indri[6]. We use a support vector machine (SVM) [4] to classify when a query should use inter- or intra-PRF. Predictions are conducted using 10-fold cross validation. We extract linguistic features using the tools built by the Stanford Natural Language Processing Group[7] and Harbin Institute of Technology [8] for Chinese NER.

Retrieval Setup: We evaluate the results using the top 1000 retrieved documents, and report mean average precision (MAP), precision@ 10 (P@10) and normalized discounted cumulative gain@k (n@10). We use Dirichlet smoothing with $\mu = 2500$ and fix θ in Equation 1 as 0.8. PRF parameters are set to use

[3] http://www.darpa.mil/Our_Work/I2O/Programs/Broad_Operational_Language_Translation_(BOLT).aspx
[4] http://www.ldc.upenn.edu/
[5] http://www.statmt.org/moses/index.php?n=Main.HomePage
[6] http://www.lemurproject.org/indri/
[7] http://nlp.stanford.edu/
[8] http://ir.hit.edu.cn/ltp/

top 10 documents and top 20 feedback terms. Note that these parameters can be tuned to improve performance. In this paper, we only apply the selective cross-lingual PRF approach to the Chinese side (i.e., the original corpus). This is because relevance feedback can be less useful to the English side as the document translation quality is often poor due to the informal text.

4.2 Retrieval Experiments

In this section, we report retrieval performance of the proposed methods and baselines. We compare our approach to two strong monolingual baseline approaches including SDM [20] and LCE [21]. We also consider two strong cross-lingual retrieval methods that fuse two monolingual ranked lists produced using LCE. As shown in Equation 2, we implement CombSUM $C_{sum}(Q, D)$ and CombMNZ $C_{mnz}(Q, D)$ [17][28] with four kinds of revised score functions $S_R(\cdot, \cdot)$. These include raw similarity score, normalized raw score, logarithm of normalized raw score and rank-based score [18]. $C_{mnz}(Q, D)$ revises $C_{sum}(Q, D)$ by considering the binary existence of document D_{L_i} in top k retrieved documents, where k is a parameter set to 500 in our experiments. w is set to 0.5 assuming no prior knowledge on language side estimation.

$$\begin{aligned} C_{sum}(Q, D) &= wS_R(Q_{L_1}, D) + \bar{w}S_R(Q_{L_2}, D) \\ C_{mnz}(Q, D) &= \big(\sum_{L_i} I(D_{L_i} \in \{TopK\}\big) \cdot C_{sum}(Q, D) \end{aligned} \tag{2}$$

Recall that we have four groups of queries Q_{mos}, Q_{bolt}, Q_{gt} and Q_{en}. In addition to the proposed approach, we generate an oracle run by selecting intra- or inter-PRF according to the true label $\hat{m}$ of the query. The oracle run defines an upper bound for the approach under these experimental settings. Note that, SDM+PRF_{intra} essentially implements LCE as discussed in Section 3.1. For English queries Q_{en}, we report the performance only on SDM and SDM+PRF_{intra}. In this case, we are performing monolingual retrieval against English (L_1).

Table 1 shows the retrieval results of the proposed approaches, where the best performance for each query set is underlined [9]. For all translated query sets, the classification-based retrieval models SDM+PRF$_{svm}$ (rigid) and SDM+PRF$_{psvm}$ (probabilistic) consistently show significant improvements over the strong baselines SDM and/or SDM+PRF$_{intra}$, and approach oracle performance.

The benefit of each technique investigated varies between different query sets of different translation quality. We first quantify query translation quality by manually assigning a tag of being good or poor. A query is labelled *poor* when the assessor could not even guess the original intent of the translated query. A query is labelled *good* when the original intent is partially or fully preserved. Figure 3 shows that Google Translate Q_{gt} outperforms Q_{bolt}, and Q_{bolt} achieves better quality than Q_{mos}. Note that better performance can be tuned for Moses and Giza++ by considering different parameters. In Table 1, for

[9] Oracle runs are not considered.

Query	Metric	SDM	PRF$_{intra}$	PRF$_{inter}$	PRF$_{svm}$	PRF$_{psvm}$	PRF$_{ora}$	C_{mnz}	C_{sum}
Q_{mos}	MAP	.2155	.2330†	.2709$^{\dagger}_{\star}$	.2721$^{\dagger}_{\star}$	.2753$^{\dagger}_{\star}$	.2786$^{\dagger}_{\star}$	.2617	.2616
	P@10	.2860	.3060†	.3560$^{\dagger}_{\star}$	.3580$^{\dagger}_{\star}$	.3610$^{\dagger}_{\star}$	.3640$^{\dagger}_{\star}$	.3440	.3340
	n@10	.2780	.2977†	.3408$^{\dagger}_{\star}$	.3412$^{\dagger}_{\star}$	.3450$^{\dagger}_{\star}$	.3439$^{\dagger}_{\star}$	.3235	.3235
Q_{bolt}	MAP	.2827	.2975	.3081†	.3150$^{\dagger}_{\star}$	.3100†	.3194$^{\dagger}_{\star}$	.3136	.3111
	P@10	.3200	.3380	.3760$^{\dagger}_{\star}$	.3760$^{\dagger}_{\star}$	.3800$^{\dagger}_{\star}$	.3840$^{\dagger}_{\star}$	.3720	.3680
	n@10	.3219	.3377	.3627$^{\dagger}_{\star}$	.3778$^{\dagger}_{\star}$	.3792$^{\dagger}_{\star}$	.3789$^{\dagger}_{\star}$	.3641	.3662
Q_{gt}	MAP	.3419	.3542	.3540	.3602†	.3612$^{\dagger}_{\star}$	.3658$^{\dagger}_{\star}$	.3542	.3419
	P@10	.4100	.4380†	.4320†	.4410†	.4500$^{\dagger}_{\star}$	.4480$^{\dagger}_{\star}$	.4360	.4360
	n@10	.3960	.4274†	.4220†	.4303†	.4303†	.4362$^{\dagger}_{\star}$	.4095	.4080
Q_{en}	MAP	.2284	.2525†						
	P@10	.3200	.3340						
	n@10	.3005	.3154†						

Table 1: Retrieval performance of the proposed and baseline methods. All SDM+PRF$_i$ is abbreviated as PRF$_i$. † and ⋆ denote significant difference over SDM and SDM+PRF$_{intra}$ with $p < 0.05$.

Q_{mos}, the most poorly translated query set, the advantages of SDM+PRF_{inter} are clearly apparent. This validates our hypothesis that translations with poor quality would benefit more from inter-PRF. For high quality translations such as Q_{gt}, intra-PRF is more effective than inter-PRF. For Q_{bolt}, the performance ordering resembles that of Q_{mos} with smaller gap between inter- and intra-PRF.

The two rightmost columns of Table 1 report the retrieval performance for the best result among the four scoring methods using CombSUM and CombMNZ. While CombMNZ and CombSUM are strong benchmarks, the results show that SDM+PRF_{psvm} significantly outperforms the fusion approaches in most cases. Consistent with [18], CombMNZ slightly outperforms CombSUM and normalized functions can be more effective than unnormalized or rank-based similarities.

4.3 Per-Query Retrieval Variation

One of the main purposes of this paper is to demonstrate how queries are influenced differently by different PRF techniques. Figure 4 shows the relative performance gain of applying intra- or inter-PRF against a baseline method SDM (y-axis) using a per-query basis (x-axis). That is, for each query in Q_{bolt}, the y-axis displays either Δ_{inter} or Δ_{intra} as defined in Equation 3, where $m(\cdot)$ computes the metric MAP. Figure 4 is sorted according to non-decreasing Δ_{inter}.

$$\Delta_i = m(SDM + PRF_i) - m(SDM), i \in \{inter, intra\} \tag{3}$$

We observe, from Figure 4, that there is a small fraction of queries that receive negative feedback from inter- or intra-PRF (i.e., $\Delta_{inter} < 0$ or $\Delta_{inter} < 0$). This can be because either the query is too hard or the translation is too poor. On the other hand, queries can be improved using the two techniques at the

same time, providing an explanation to why SDM+PRF_{psvm} may outperform SDM+PRF_{oracle} in some cases. More importantly, the trends of Δ_{inter} and Δ_{intra} often intersect each other across the query set, indicating proper integration of inter- and intra-PRF can result in overall optimized performance as previously shown in Table 1.

	Good	Poor
Q_{mos}	30 (60%)	20 (40%)
Q_{bolt}	38 (76%)	12 (24%)
Q_{gt}	48 (96%)	2 (4%)
Avg	38.7 (77%)	11.3 (23%)

Fig. 3: Translation quality (number of *poor* or *good* query instances) for different query sets.

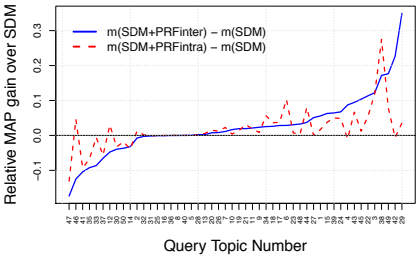

Fig. 4: Relative MAP gain of cross-lingual PRF techniques over SDM for each Q_{bolt} query.

4.4 Combining Inter- and Intra-PRF

We now explore the effects of combining inter- and intra-PRF using linear weighting. Figure 5 shows the retrieval performance based on our original weighted retrieval model (Equation 1). We vary the weight ϕ_{intra} for intra-PRF from 0 to 1 by steps of 0.05, and $\phi_{inter} = 1 - \phi_{intra}$. Note that the weight configurations are fixed across all queries, as opposed to the prediction framework where the PRF class for each query would be rigidly or probabilistically selected. For all three types of translated query sets, we observe a general trend that over-weighting ϕ_{intra} results in decreased retrieval performance. For Q_{mos}, we find the performance negatively correlates with the increase of ϕ_{intra}, indicating that inter-PRF outperforms intra-PRF consistently for queries of poor translation quality. For Q_{gt}, the increase of ϕ_{intra} initially improves retrieval effectiveness. The peak performance is reached around $\phi_{intra} \cong 0.3$ and further increase of ϕ_{intra} results in performance drop. The performance trend for Q_{bolt} is similar to Q_{mos}.

5 Conclusions

We investigate the task of CLIR on a large collection of forum posts. The translation noise is increased by informal text used in discussion forums. We consider two types of PRF for CLIR and propose a solution that selects between inter- and intra-PRF on a per-query basis. Experimental results show that the selective cross-lingual PRF approach significantly improves the performance over strong monolingual and cross-lingual baselines. We find that queries with poor translation quality benefit most from inter-PRF as document translation is more reliable in such cases. Intra-PRF is more useful for queries with good query translation accuracy as it reduces the impact of any translation error in documents.

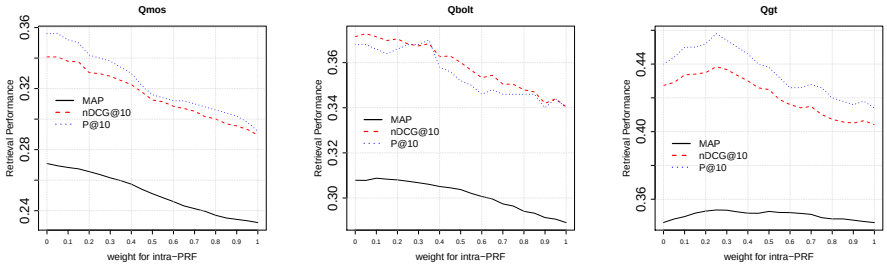

Fig. 5: Retrieval performance for weighted combination of inter- and intra-PRF.

In future work, we are interested in investigating other types of queries and other techniques for CLIR. We also consider the potential of integrating thread information in the discussion forum for better smoothing.

Acknowledgments

This work was supported in part by the Center for Intelligent Information Retrieval and in part by IBM subcontract #4913003298 under DARPA prime contract #HR001-12-C-0015. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

References

1. Lisa Ballesteros and W. Bruce Croft. Resolving ambiguity for cross-language retrieval. In *Proc. of SIGIR*, SIGIR '98, pages 64–71, 1998.
2. Michael Bendersky and W. Bruce Croft. Modeling higher-order term dependencies in information retrieval using query hypergraphs. In *Proc. of SIGIR*, pages 941–950.
3. Martin Braschler and Peter Schäuble. Experiments with the eurospider retrieval system for clef 2000. In *Proc. of CLEF 2000*, pages 140–148. Springer-Verlag, 2001.
4. Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.
5. Aitao Chen. Multilingual information retrieval using english and chinese queries. In *In*, pages 44–58. Springer-Verlag, 2002.
6. Aitao Chen and Fredric C. Gey. Multilingual information retrieval using machine translation, relevance feedback and decompounding. *Inf. Retr.*, 7(1-2):149–182, January 2004.
7. Manoj K. Chinnakotla, Karthik Raman, and Pushpak Bhattacharyya. Multilingual prf: english lends a helping hand. In *Proc. of SIGIR*, pages 659–666.
8. Kevyn Collins-Thompson and Paul N. Bennett. Predicting query performance via classification. In *Proc. of ECIR*, pages 140–152, 2010.
9. Gordon V. Cormack, Charles L A Clarke, and Stefan Buettcher. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proc. of SIGIR*, SIGIR '09, pages 758–759, 2009.
10. Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

11. Fredric C. Gey, Hailing Jiang, Vivien Petras, and Aitao Chen. Cross-language retrieval for the clef collections - comparing multiple methods of retrieval. In *Revised Papers from the Workshop of Cross-Language Evaluation Forum on Cross-Language Information Retrieval and Evaluation*, CLEF '00, pages 116–128, 2001.
12. Ben He and Iadh Ounis. Inferring query performance using pre-retrieval predictors. In *Proc. Symposium on String Processing and Information Retrieval*, pages 43–54. Springer Verlag, 2004.
13. Daqing He and Dan Wu. Translation enhancement: a new relevance feedback method for cross-language information retrieval. In *Proc. of CIKM*, pages 729–738, 2008.
14. Oren Kurland, Anna Shtok, David Carmel, and Shay Hummel. A unified framework for post-retrieval query-performance prediction. In *Proc. of ICTIR*, pages 15–26.
15. Adenike M. Lam-Adesina and Gareth J. F. Jones. Exeter at clef 2003: Experiments with machine translation for monolingual, bilingual and multilingual retrieval. In *CLEF*, pages 271–285, 2003.
16. Victor Lavrenko and W. Bruce Croft. Relevance based language models. In *Proc. of SIGIR*, SIGIR '01, pages 120–127, 2001.
17. Joon Ho Lee. Combining multiple evidence from different properties of weighting schemes. In *Proc. of SIGIR*, pages 180–188, 1995.
18. Joon Ho Lee. Analyses of multiple evidence combination. pages 267–276. ACM Press, 1997.
19. Gina-Anne Levow. Issues in pre- and post-translation document expansion: untranslatable cognates and missegmented words. In *Proc. of AsianIR '03*, pages 77–83, 2003.
20. Donald Metzler and W. Bruce Croft. A markov random field model for term dependencies. In *Proc. of SIGIR*, SIGIR '05, pages 472–479, 2005.
21. Donald Metzler and W. Bruce Croft. Latent concept expansion using markov random fields. In *Proc. of SIGIR*, SIGIR '07, pages 311–318, 2007.
22. Seung-Hoon Na and Hwee Tou Ng. Enriching document representation via translation for improved monolingual information retrieval. In *Proc. of SIGIR*, pages 853–862, 2011.
23. Douglas W. Oard. A comparative study of query and document translation for cross-language information retrieval. In *Proc. of the third conference of the association for machine translation in the Americas*, 1998.
24. John C. Platt. *Probabilities for SV Machines*, pages 61–74. 2000.
25. Yan Qu, Alla N. Eilerman, Hongming Jin, and David A. Evans. The effects of pseudo-relevance feedback on mt-based. In *CLIR," RIAO 2000, Content-based Multi-Media Information Access. CSAIS*, pages 46–60, 2000.
26. Monica Rogati and Yiming Yang. Cross-lingual pseudo-relevance feedback using a comparable corpus. In *Evaluation of Cross-Language Information Retrieval Systems*, pages 151–157. 2002.
27. Jacques Savoy. Report on clef-2001 experiments: Effective combined query-translation approach. In *Evaluation of Cross-Language Information Retrieval Systems*, pages 27–43. 2002.
28. Joseph A. Shaw, Edward A. Fox, Joseph A. Shaw, and Edward A. Fox. Combination of multiple searches. In *TREC-2*, pages 243–252, 1994.
29. Ellen M. Voorhees. Query expansion using lexical-semantic relations. In *Proc. of SIGIR*, SIGIR '94, pages 61–69, 1994.
30. Jinxi Xu and W. Bruce Croft. Query expansion using local and global document analysis. In *Proc. of SIGIR*, pages 4–11, 1996.
31. Jian yun Nie, Michel Simard, Pierre Isabelle, Richard Dur, and Université De Montréal. Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the web. In *Proc. of SIGIR*, pages 74–81, 1999.
32. Chengxiang Zhai and John Lafferty. Model-based feedback in the language modeling approach to information retrieval. In *Proc. of CIKM*, CIKM '01, pages 403–410.
33. Jiang Zhu and Haifeng Wang. The effect of translation quality in mt-based cross-language information retrieval. In *Proc. of ACL*, ACL-44, pages 593–600, 2006.