# Multilingual PRF: English Lends a Helping Hand

Manoj K. Chinnakotla [*]
manoj@cse.iitb.ac.in

Karthik Raman
karthikr@cse.iitb.ac.in

Pushpak Bhattacharyya
pb@cse.iitb.ac.in

Department of Computer Science and Engineering
Indian Institute of Technology, Bombay
Mumbai, India

## ABSTRACT

In this paper, we present a novel approach to Pseudo-Relevance Feedback (PRF) called *Multilingual PRF (MultiPRF)*. The key idea is to harness multilinguality. Given a query in a language, we take the help of another language to ameliorate the well known problems of PRF, *viz.* (a) The expansion terms from PRF are primarily based on co-occurrence relationships with query terms, and thus other terms which are lexically and semantically related, such as morphological variants and synonyms, are not explicitly captured, and (b) PRF is quite sensitive to the quality of the initially retrieved top $k$ documents and is thus not robust. In MultiPRF, given a query in language $L_1$, it is translated into language $L_2$ and PRF is performed on a collection in language $L_2$ and the resultant feedback model is translated from $L_2$ back into $L_1$. The final feedback model is obtained by combining the translated model with the original feedback model of the query in $L_1$.

Experiments were performed on standard CLEF collections in languages with widely differing characteristics, *viz., French, German, Finnish* and *Hungarian* with *English* as the assisting language. We observe that MultiPRF outperforms PRF and is more robust with consistent and significant improvements in the above widely differing languages. A thorough analysis of the results reveal that the second language helps in obtaining both co-occurrence based conceptual terms as well as lexically and semantically related terms. Additionally, the use of the second language collection reduces the sensitivity to performance of initial retrieval, thereby making it more robust.

## Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information Search and Retrieval, Retrieval Models, Search Process

## General Terms

Algorithms, Performance, Design, Experimentation

## Keywords

Multilingual, Pseudo-Relevance Feedback, Language Models, Query Expansion

## 1. INTRODUCTION

The central problem of Information Retrieval (IR) is to satisfy the user's information need, which is typically expressed through a short (approximately 2-3 words) and often ambiguous query. The problem of matching the user's query with the documents is rendered difficult by natural language phenomena like *morphological variants*, *polysemy* and *synonymy.* Relevance Feedback (RF) tries to overcome these problems by eliciting user feedback on the relevance of documents obtained from the initial ranking and then using it to automatically refine the query. Since user input is hard to obtain, Pseudo-Relevance Feedback (PRF) [4, 30, 19] is used as an alternative, where the RF is performed by *assuming* the top $k$ documents from initial retrieval as being *relevant* to the query. Based on the above assumption, the terms in the feedback document set are analyzed to choose the most distinguishing set of terms that characterize the feedback documents and as a result the relevance of a document. The query refinement is done by adding the terms obtained through PRF, along with their weights, to the actual query.

Although PRF has been shown to improve retrieval effectiveness, it suffers from the following drawbacks: (a) due to the assumption inherent in the PRF process, *i.e.*, relevance of top $k$ documents, it is sensitive to the performance of the initial retrieval algorithm and as a result is not robust, and (b) the type of term associations obtained for query expansion is restricted to co-occurrence based relationships in the feedback documents, and thus other types of term associations such as lexical and semantic relations (morphological variants, synonymy), which are relevant in the context of the query, are not explicitly captured.

In this paper, we propose a novel approach called **Multilingual Pseudo-Relevance Feedback (MultiPRF)** to overcome both of the above limitations of PRF. We take help of a different language called herein the *assisting language.*

In MultiPRF, given a query in a source language $L_1$, the query is automatically translated into the assisting language $L_2$ and PRF performed in the assisting language. The resultant terms are translated back into $L_1$ using a probabilistic

[*] Supported by a fellowship award from Infosys Technologies Ltd., India.

*SIGIR'10,* July 19–23, 2010, Geneva, Switzerland.
Copyright 2010 ACM 978-1-60558-896-4/10/07 ...$10.00.

bi-lingual dictionary. At the same time, a feedback model is also computed in $L_1$ and finally combined with the feedback model obtained through the assisting language. The resultant model is finally used to re-rank the corpus and fetch a new ranked list of documents. Experiments on standard CLEF [3] collections in languages with widely divergent characteristics such as *French, German, Finnish* and *Hungarian* with *English* as the assisting language show that MultiPRF achieves significant performance improvement over monolingual PRF. A point about why English is used as the assisting language is in order here. English shares about 72% of the web content. Larger coverage typically ensures higher proportion of relevant documents in the top $k$ retrieval [12]. This in turn ensures better PRF. Assisting the fact is the other fact that query processing in English is a simpler proposition than in most other languages due to English's simpler morphology and wider availability of NLP tools for English.

A thorough qualitative analysis of the results reveal that MultiPRF indeed overcomes the fundamental limitations of PRF. Firstly, since it relies on the PRF in two collections of different languages, it is more robust. Secondly, the assisting language helps in obtaining both co-occurrence based conceptual terms as well as lexically and semantically related terms. The proposed approach is especially attractive in the case of languages where the original retrieval is bad due to poor coverage of the collection and/or inherent complexity of query processing (for example *term conflation*) in those languages. For example, Hungarian has only 0.2% share of web content[1] with a rich morphology. Experiments also show that MultiPRF improves over monolingual PRF even when the query translation accuracy is sub-optimal.

The organization of the paper is as follows: In section 2, we discuss the related work in the area. Section 3 explains the Language Modeling (LM) based PRF approach which is used for performing monolingual PRF and which forms our baseline. We present the MultiPRF approach: our proposed model in Section 4. Section 5 presents the experimental set up and results followed by a discussion of these results in section 6. Finally, section 7 concludes the paper by summarizing observations and outlining possible directions for future work.

## 2. RELATED WORK

PRF has been effectively applied in various IR frameworks like vector space models, probabilistic IR and language modeling [4, 15, 17, 33]. Several approaches have been proposed to improve the performance and robustness of PRF. Some of the representative techniques are (i) to refine the feedback document set [19, 24], (ii) refining the terms obtained through PRF by selecting good expansion terms [5] and (iii) using selective query expansion [1, 7] and varying the importance of documents in the feedback set [25]. Another direction of work, often reported in the TREC Robust Track, is to use a large external collection like Wikipedia or the Web as a source of expansion terms [32, 27]. The intuition behind the above approach is that if the query does not have many relevant documents in the collection then any improvements in the modeling of PRF is bound to perform poorly due to query drift.

Several approaches have been proposed for including different types of lexically and semantically related terms during query expansion. Voorhees *et al.* [28] use Wordnet for query expansion and report negative results. Recently, random walk models [16, 6] have been used to learn a rich set of term level associations by combining evidence from various kinds of information sources mentioned so far like WordNet, co-occurrence relationships, web, morphological variants *etc.,*. Metzler *et al.* [18] propose a feature based approach called *latent concept expansion* to model term dependencies.

All the above mentioned approaches use the resources available *within* the language to improve the performance of PRF. However, we make use of a *second language* (English) to improve the performance of PRF. As mentioned earlier, this is an attractive proposition for languages where the original retrieval is bad due to poor coverage and inherent complexity of query processing due to rich morphology, word compounding *etc.*

The idea of using one language to improve the accuracy of another language in a specific task has been successfully tried for the problem of Word Sense Disambiguation (WSD) [8].

A recent work by Gao *et al.* [11] uses English to improve the performance over a subset of Chinese queries whose translations in English are unambiguous. They use inter-document similarities across languages to improve the ranking performance. The computation of cross language document level similarities between English and Chinese documents is done using a bi-lingual dictionary. However, cross language document similarity measurement is in itself known to be an equally hard problem especially without using parallel or comparable corpora [10]. Moreover, the scale of their experimentation is quite small and they demonstrate their approach only on a small class of queries in a single language.

## 3. PRF IN LANGUAGE MODELING FRAMEWORK

The Language Modeling (LM) Framework for IR offers a principled approach to model PRF. In the LM approach, the document and query are modeled using multinomial distribution over words called *document language model* $P(w|D)$ and *query language model* $P(w|\Theta_Q)$ respectively. For a given query, the document language models are ranked based on their proximity to the query language model, measured using KL-Divergence.

$$\begin{aligned} Rank(D,Q) &= KL(\Theta_Q||D) \\ &= \sum_w P(w|\Theta_Q) \cdot log \frac{P(w|\Theta_Q)}{P(w|D)} \end{aligned}$$

Since the query length is short, it is difficult to estimate the query language model accurately using the query alone. In PRF, the top $k$ documents obtained through the initial ranking algorithm are assumed to be relevant and used as feedback for improving the estimation of $\Theta_Q$. The feedback documents contain a mix of both relevant and noisy terms. The actual relevant terms modeled using the feedback language model $\Theta_F$ is inferred from $D_F$ based on a Generative Mixture Model [33] formulation.

[1] `http://www.netz-tipp.de/languages.html`

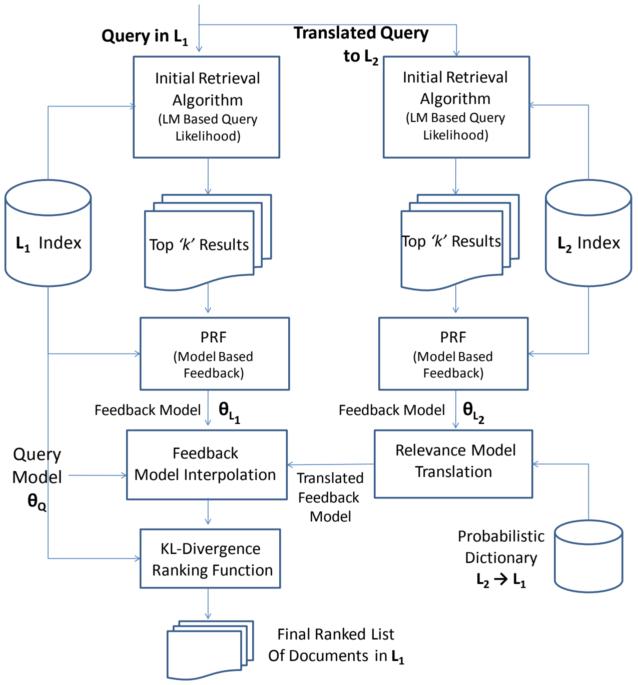

Query in $L_1$
Translated Query to $L_2$
Initial Retrieval Algorithm (LM Based Query Likelihood)
Initial Retrieval Algorithm (LM Based Query Likelihood)
$L_1$ Index
Top 'k' Results
Top 'k' Results
$L_2$ Index
PRF (Model Based Feedback)
PRF (Model Based Feedback)
Feedback Model $\theta_{L_1}$
Feedback Model $\theta_{L_2}$
Query Model $\theta_Q$
Feedback Model Interpolation
Relevance Model Translation
Translated Feedback Model
Probabilistic Dictionary $L_2 \rightarrow L_1$
KL-Divergence Ranking Function
Final Ranked List Of Documents in $L_1$

**Figure 1: Schematic of the Multilingual Pseudo-Relevance Feedback Approach**

| Symbol | Description |
|---|---|
| $\Theta_Q$ | Query Language Model |
| $\Theta^F_{L_1}$ | Feedback Language Model obtained from PRF in $L_1$ |
| $\Theta^F_{L_2}$ | Feedback Language Model obtained from PRF in $L_2$ |
| $\Theta^{Trans}_{L_1}$ | Feedback Model Translated from $L_2$ to $L_1$ |
| $P_{L_2 \rightarrow L_1}$ | Probabilistic Bi-Lingual Dictionary from $L_2$ to $L_1$ |
| $\beta, \gamma$ | Interpolation coefficients coefficients used in Multi-PRF |

**Table 1: Glossary of Mathematical Symbols used in explaining MultiPRF**

| Source Term | Top Aligned Terms in Target |
|---|---|
| **French** | **English** |
| américain | american, us, united, state, america |
| nation | nation, un, united, state, country |
| étude | study, research, assess, investigate, survey |
| **German** | **English** |
| flugzeug | aircraft, plane, aeroplane, air, flight |
| spiele | play, game, stake, role, player |
| verhältnis | relationship, relate, balance, proportion |

**Table 2: Top Translation Alternatives for some sample words in Probabilistic Bi-Lingual Dictionary**

## 3.1 Mixture Model for Estimating Feedback Model

Let $D_F = \{d_1, d_2, \ldots, d_k\}$ be the top $k$ documents retrieved using the initial ranking algorithm. Zhai and Lafferty [33] model the feedback document set $D_F$ as a mixture of two distributions: (a) the *feedback language model* and (b) the *collection model* $P(w|C)$. Assuming a fixed mixture proportion $\lambda$ in the feedback document set, the feedback language model is inferred using the EM Algorithm [9]. In the EM algorithm, the feedback model is iteratively refined by accumulating probability mass on most *distinguishing* terms which are more frequent in the feedback document set and less frequent across the entire collection. Let $\Theta_F$ be the final converged feedback model. Later, in order to keep the query focus, $\Theta_F$ is interpolated with the initial query model $\Theta_Q$ to obtain the final query model $\Theta_{Final}$.

$$\Theta_{Final} = (1 - \alpha) \cdot \Theta_Q + \alpha \cdot \Theta_F \tag{1}$$

$\Theta_{Final}$ is used to re-rank the corpus using the KL-Divergence ranking function to obtain the final ranked list of documents. Henceforth, we refer to the above PRF technique by as *Model Based Feedback (MBF)*.

# 4. MULTILINGUAL RELEVANCE FEEDBACK (MULTIPRF)

In this section, we describe our main contribution - the *Multilingual PRF* approach. The schematic of the approach is shown in Figure 1.

Given a query $Q$ in the source language $L_1$, we automatically translate the query using a query translation system into the assisting language $L_2$. We then rank the documents in the $L_2$ collection using the query likelihood ranking function [14]. Using the top $k$ documents, we estimate the feedback model using MBF described in the previous section. Similarly, we also estimate a feedback model using the original query and the top $k$ documents retrieved from the initial ranking in $L_1$. Let the resultant feedback models be $\Theta^F_{L_2}$ and $\Theta^F_{L_1}$ respectively.

The feedback model estimated in the assisting language $\Theta^F_{L_2}$ is translated back into language $L_1$ using a probabilistic bi-lingual dictionary $P_{L_2 \rightarrow L_1}(f|e)$ from $L_2 \rightarrow L_1$ as follows:

$$P(f|\Theta^{Trans}_{L_1}) = \sum_{\forall \text{ e in } L_2} P_{L_2 \rightarrow L_1}(f|e) \cdot P(e|\Theta^F_{L_2}) \tag{2}$$

The probabilistic bi-lingual dictionary $P_{L_2 \rightarrow L_1}(f|e)$ is learned from a parallel sentence-aligned corpora in $L_1 - L_2$ based on word level alignments. Tiedemann [26] has shown that the translation alternatives found using word alignments could be used to infer various morphological and semantic relations between terms. For example, in Table 2, we show the top translation alternatives for some sample words. For example, the French word *américain* (american) brings different variants of the translation like *american, america, us, united, state, america* which are lexically and semantically related. Hence, the probabilistic bi-lingual dictionary acts as a rich source of morphologically and semantically related feedback terms. During the step for translating the feedback model given in Equation 2, the translation model adds related terms in $L_1$ which have their source as the term from feedback model $\Theta^F_{L_2}$.

The final MultiPRF model is obtained by interpolating the above translated feedback model with the original query model and the feedback model of language $L_1$ as given below:

$$\Theta^{Multi}_{L_1} = (1 - \beta - \gamma) \cdot \Theta_Q + \beta \cdot \Theta^F_{L_1} + \gamma \cdot \Theta^{Trans}_{L_1} \tag{3}$$

Since we want to retain the query focus during back translation the feedback model in $L_2$ is interpolated with the translated query before translation. The parameters $\beta$ and $\gamma$ control the relative importance of the original query model, feedback model of $L_1$ and the translated feedback model

| Language | CLEF Collection Identifier | Description | Assisting Collection Used | No. of Documents | No. of Unique Terms | CLEF Topics (No. of Topics) |
|---|---|---|---|---|---|---|
| English | EN-00+01+02 | LA Times 94 | - | 113005 | 174669 | - |
| | EN-03+05+06 | LA Times 94 + Glasgow Herald 95 | - | 169477 | 234083 | - |
| French | FR-00 | Le Monde 94 | EN-00+01+02 | 44013 | 127065 | 1-40 (29) |
| | FR-01+02 | Le Monde 94, French SDA 94 | EN-00+01+02 | 87191 | 159809 | 41-140 (88) |
| | FR-03+05 | Le Monde 94, French SDA 94, 95 | EN-03+05+06 | 129806 | 182214 | 141-200 & 251-300 (99) |
| | FR-06 | Le Monde 94, 95, French SDA 94, 95 | EN-03+05+06 | 177452 | 231429 | 301-350 (48) |
| German | DE-00 | Frankfurter Rundschau 94<br>Der Spiegel 94/95 | EN-00+01+02 | 153694 | 791093 | 1-40 (33) |
| | DE-01+02 | Frankfurter Rundschau 94, Der Spiegel 94, 95, German SDA 94 | EN-00+01+02 | 225371 | 782304 | 41-140 (85) |
| | DE-03 | Frankfurter Rundschau 94, Der Spiegel 94, 95, German SDA 94, 95 | EN-03+05+06 | 294809 | 867072 | 141-200 (51) |
| Finnish | FI-02+03+04 | Aamulehti 94-95 | EN-03+05+06 | 55344 | 531160 | 91-250 (119) |
| Hungarian | HU-05 | Magyr Hirlap 2002 | EN-03+05+06 | 49530 | 256154 | 251-300 (48) |

**Table 3: Details of the CLEF Datasets used for Evaluating the MultiPRF approach. The number shown in brackets of the final column CLEF Topics indicate the actual number of topics used during evaluation.**

obtained from $L_1$ and are tuned based on the choice of collection in $L_1$ and $L_2$.

## 5. EXPERIMENTAL SETUP

We evaluate the performance of our system using the standard CLEF evaluation data [3] in four widely differing languages - French, German, Finnish and Hungarian using more than 600 topics. We use English as the assisting language. The details of the collections, their corresponding topics and the assisting collections used for MultiPRF are given in Table 3. Note that we choose the English assisting collection such that the coverage of topics is similar to that of the original corpus so as to get meaningful feedback terms. In all the topics, we only use the *title* field. We ignore the topics which have no relevant documents as the true performance on those topics cannot be evaluated.

We use the Terrier IR platform [21] for indexing the documents. We perform standard tokenization, stop word removal and stemming. We use the Porter Stemmer for English and the stemmers available through the Snowball[2] package for French, German, Finnish and Hungarian. Other than these, we do not perform any other processing on German, Finnish and Hungarian. However, in French, since some function words like *l'*, *d' etc.,* occur as prefixes to a word, we strip them off during indexing and query processing, since that caused the baseline performance to decrease. We use standard evaluation measures like *MAP*, *P@5* and *P@10* for evaluation. Additionally, for assessing robustness, we use the Geometric Mean Average Precision (GMAP) metric [23] which is also used in the TREC Robust Track [27].

The probabilistic bi-lingual dictionary used in MultiPRF was learnt automatically by running GIZA++ - a word alignment tool [20] on a parallel sentence aligned corpora. For French-English, German-English and Finnish-English language pairs, we used the *Europarl Corpus* [22] and in case of Hungarian-English, we used the Hunglish Corpus[3].

We make use of off-the-shelf translation systems available in the above language pairs. We use Google Translate[4] as the query translation system as it has been shown to perform well for query translation [29]. Later, we show that our approach is not dependent on Google Translate, and report results using a basic SMT system for query translation. For this, we evaluate the quality of the above Query Translation systems and analyze their impact on the quality of our results. In the pathological case of term not being found in English after query translation, we only perform MBF on the source language $L_1$.

We use the MBF approach explained in Section 3.1 as a baseline for all our comparisons. We use two-stage Dirichlet smoothing with the optimal parameters tuned based on the collection [34]. We tune the parameters of MBF, specifically $\lambda$ and $\alpha$, and choose the values which give the optimal performance on a given collection. We uniformly set the number of feedback documents, *i.e.,* $k$ as 10 *i.e.* top ten documents. The overall results are shown in Table 4. We observe that the optimal values of interpolation coefficients $\beta, \gamma$ in MultiPRF are almost uniform across collections and vary in the range 0.4-0.48.

## 6. RESULTS AND DISCUSSION

As in Table 4, the results show that the MultiPRF approach with English as the assisting language significantly outperforms the MBF approach across all datasets of all the chosen languages. We consistently observe significant improvements in MAP (between 4% to 8%), P@5 (between 4% to 39%) and P@10 (around 4% to 22%). The MultiPRF approach is also more robust than plain MBF as reflected in the improvements obtained in GMAP scores (between 15% to 730%). This could be attributed in part to the reduced sensitivity of our approach to the number of relevant documents in the feedback set of the source language. An analysis of the overall results reveal that MultiPRF leverages the performance in English language and adds relevant terms like morphological variants and synonyms in addition to co-occurrence based term relations. Besides this, it also

[2] `http://snowball.tartarus.org/index.php`

[3] `http://mokk.bme.hu/resources/hunglishcorpus`

[4] `http://translate.google.com`

| | MAP | | | P@5 | | | P@10 | | | GMAP | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Collection** | **MBF** | **MultiPRF** | **% Improv.** | **MBF** | **MultiPRF** | **% Improv.** | **MBF** | **MultiPRF** | **% Improv.** | **MBF** | **MultiPRF** | **% Improv.** |
| **FR-00** | 0.4220 | 0.4393 | 4.10 | 0.4690 | 0.5241 | **11.76**$^{\ddagger}$ | 0.4000 | 0.4000 | 0.00 | 0.2961 | 0.3413 | **15.27** |
| **FR-01+02** | 0.4342 | 0.4535 | **4.43**$^{\ddagger}$ | 0.4636 | 0.4818 | 3.92 | 0.4068 | 0.4386 | **7.82**$^{\ddagger}$ | 0.2395 | 0.2721 | **13.61** |
| **FR-03+05** | 0.3529 | 0.3694 | **4.67**$^{\ddagger}$ | 0.4545 | 0.4768 | **4.89**$^{\ddagger}$ | 0.4040 | 0.4202 | **4**$^{\ddagger}$ | 0.1324 | 0.1411 | 6.57 |
| **FR-06** | 0.3837 | 0.4104 | 6.97 | 0.4917 | 0.5083 | 3.39 | 0.4625 | 0.4729 | 2.25 | 0.2174 | 0.2810 | **29.25** |
| **DE-00** | 0.2158 | 0.2273 | 5.31 | 0.2303 | 0.3212 | **39.47**$^{\ddagger}$ | 0.2394 | 0.2939 | **22.78**$^{\ddagger}$ | 0.0023 | 0.0191 | **730.43** |
| **DE-01+02** | 0.4229 | 0.4576 | **8.2**$^{\ddagger}$ | 0.5341 | 0.6000 | **12.34**$^{\ddagger}$ | 0.4864 | 0.5318 | **9.35**$^{\ddagger}$ | 0.1765 | 0.2721 | 9.19 |
| **DE-03** | 0.4274 | 0.4355 | 1.91 | 0.5098 | 0.5412 | 6.15 | 0.4784 | 0.4980 | 4.10 | 0.1243 | 0.1771 | **42.48** |
| **FI-02+03+04** | 0.3966 | 0.4246 | **7.06**$^{\ddagger}$ | 0.3782 | 0.4034 | **6.67**$^{\ddagger}$ | 0.3059 | 0.3319 | **8.52**$^{\ddagger}$ | 0.1344 | 0.2272 | **69.05** |
| **HU-05** | 0.3066 | 0.3269 | **6.61**$^{\ddagger}$ | 0.3542 | 0.4167 | **17.65**$^{\ddagger}$ | 0.3083 | 0.3292 | **6.76**$^{\ddagger}$ | 0.1326 | 0.1643 | **23.91** |

**Table 4: Results comparing the performance of MultiPRF approach over the baseline MBF approach on CLEF collections. Results marked as $^{\ddagger}$ indicate that the improvement was found to be statistically significant over the baseline at 90% confidence level ($\alpha = 0.01$) when tested using a paired two-tailed t-test.**

| TOPIC NO. | ORIGINAL QUERY | TRANSLATED ENGLISH QUERY | MBF MAP | MPRF MAP | MBF - Top Representative Terms (With meaning) | MultiPRF - Top Representative Terms (With meaning) |
|---|---|---|---|---|---|---|
| FRENCH '00. TOPIC 33 | Tumeurs et génétique | Tumors and Genetics | 0.0414 | 0.2722 | malad (ill), tumeur (tumor), recherch (research), canc (cancer), yokoham, hussein | tumor (tumor), génet, canc, gen, malad, cellul (cellular), recherch |
| FRENCH '03. TOPIC 198 | Oscar honorifique pour des réalisateurs italiens | Honorary Oscar for Italian filmmakers | 0.1238 | 0.4324 | italien, président (president), oscar , gouvern (governer) , scalfaro , spadolin | film, italien, oscar, honorair (honorary) , cinem (film), cinéast (filmmaker), réalis (achieve), produit(product) |
| FRENCH '06. TOPIC 317 | Les Drogues Anti-cancer | The Anti-Cancer Drugs | 0.001 | 0.1286 | drogu (drugs), anti , trafic (trafficking), entre (between), légalis (legalise), canc , malad , cocaïn , afghanistan , iran | canc, drogu, recherch, malad, trait, taxol, glaxo, cancer |
| GERMAN '02. TOPIC 115 | Scheidungsstatistiken | Divorce Statistics | 0.2206 | 0.4463 | prozent (percent), unterstutz (supporters), frau (woman), minderjahr (underage), scheidung (divorce) | statist, scheidung , zahl (number), elt (parent), kind (child), famili, geschied (divorced), getrennt (separated), ehescheid (divorce) |
| GERMAN '03. TOPIC 147 | Ölunfälle und Vögel | Birds and Oil Spills | 0.0128 | 0.1184 | rhein (rhine), olunfall (oil spill), fluss (river), ol (oil) , heizol (fuel/oil), tank (tanker) | ol, olverschmutz (oil pollution), vogel (bird), erdol (petroleum), olp (oil slick), olunfall , gallon, vogelart (bird species) |
| FRENCH '05. TOPIC 274 | Bombes actives de la Seconde Guerre Mondiale | Active bombs of the Second World War | 0.6182 | 0.3206 | bomb, guerr(war), mondial (world), vill(city), découvert(discovery), second, explos, alemagn (germany), allemand (german) | guerr, mond(world), deuxiem (second), activ, bombard, japon(japan), hiroshim, nagasak, atom, nucléair (nuclear) |
| GERMAN '03. TOPIC 188 | Deutsche Rechtschreibreform | German spelling reform | 0.8278 | 0.6776 | deutsch, reform, spiegel (reflect), rechtschreibreform (spelling reform), sprach (language), osterreich (austria), rechtschreib (spelling), wien (vienna), schweiz (switzerland) | deutsch, reform, deutschland (Germany), clinton, deutlich (clearly), president , berlin, europa, gipfel(summit), bedeut(important) |

**Table 5: Qualitative comparison of feedback terms given by MultiPRF and MBF on representative queries where positive and negative improvements were observed in French and German collections.**

improves the performance of some queries where the PRF performance was poor to start with, by bringing in related terms through PRF in $L_2$ and back translation.

To illustrate the qualitative improvement in feedback terms, a detailed analysis of a few representative queries is presented in Table 5. Based on the above analysis, the improvements obtained by MultiPRF approach could be mainly attributed to one of the following three reasons:- (a) Retrieval Performance in $L_2$ is good and the resultant feedback model contains a lot of relevant terms, which when brought back to $L_1$ via back-translation leads to improvement. (b) During the back-translation process, important synonyms and popular morphological variants (inflectional forms) of key terms are found, which otherwise were missing from the Model-Based feedback model. and (c) A combination of both the above factors.

For example, consider the French Query *"Oscar honorifique pour des réalisateurs italiens"*, meaning "Honorary Oscar for Italian Filmmakers". Model Based Feedback on French expands the query using the top retrieved documents of the initial retrieval. However, here it introduces significant topic drift towards Oscar Scalfaro (a former Italian President) and Italian politics thus causing words such as $\{$*scalfaro, spadolin, gouvern*$\}$. However, feedback in English produces relevant terms, which on translation back into French, introduces terms such as $\{$*cinem, cinéast, réalis*$\}$. This wrenches back the focus of the query from the political domain to the intended film domain, thus leading to performance increase. Another example of this phenomenon is the query *"Les Drogues Anti-Cancer"* (Anti-Cancer Drugs). Here too MBF causes drift away from the intended meaning and instead to Drug-Trafficking, by introducing terms such as $\{$*traffic, entre, afghanistan*$\}$, which causes very poor performance on the query. MultiPRF however utilizes the good feedback performance of English on this query, to generate a set of very relevant French terms such as $\{$*recerch, taxol, glaxo*$\}$. Hence the drift from the intended meaning towards drug-trafficking is corrected, by the introduction of the above mentioned terms, which help in bringing up the performance on this query. These examples demonstrate

| Corpus | Google Translate | SMT |
|---|---|---|
| FR-01+02 | 0.93 | 0.67 |
| FR-03+05 | 0.88 | 0.77 |
| DE-01+02 | 0.93 | 0.64 |
| DE-03 | 0.81 | 0.58 |

**Table 6: Comparison of Query Translation Quality using Google Translate and SMT system trained on Europarl Corpus on a scale of 0-1.**

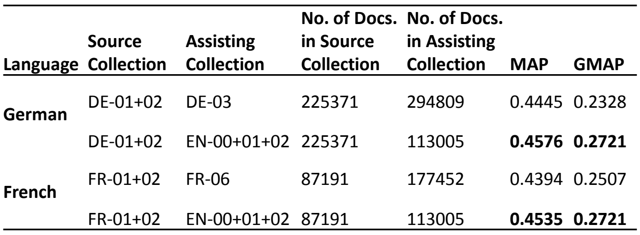
the robustness of the MultiPRF approach and the reduced sensitivity to the relevance of the top documents from the initial retrieval.

Apart from this we also see improvements on queries due to introduction of synonyms and other semantically related terms. For example, on the German query *"Ölunfälle und Vögel"* meaning "Birds and Oil Spills", MBF performs poorly with many irrelevant terms introduced in the feedback model. However English finds some relevant terms, and additionally adds many terms to the feedback model, which are synonyms/semantically related to oil spills and birds, such as {*olverschmutz, ol, olp, vogelart*}. This helps in bringing up more relevant documents while reducing drift.

## 6.1 Effect of Query Translation Quality

Accurate Query Translation is fundamental to MultiPRF. As explained earlier, we chose *Google Translate* mainly due to its ease of availability. In this section, we study the impact of varying translation quality on the performance of our approach. We train a Statistical Machine Translation (SMT) system, on the French-English and German-English language pairs, by running an off-the-shelf publicly available tools like Moses [13] on Europarl corpora. The above SMT system is quite simple because we do not perform any language-specific processing or any parameter tuning to improve the performance of the system and also it is limited by the domain of the parallel corpora which is parliamentary proceedings. To correlate the translation quality with the performance of MultiPRF, we evaluated the query translations produced by Google Translate and SMT system on a three-point scale between 0 and 1 (0 - Completely Wrong Translation, 0.5 - Translation not optimal but query intent partially conveyed and 1 - Query intent completely conveyed). The results are shown in Table 6. We compare the performance of MultiPRF using Google Translate, Basic SMT system, and ideal query translations. The ideal translations were obtained by manually fixing some of the errors in the above two systems. The performance on ideal query translations gives an idea of the upper bound on the performance of MultiPRF. The results of our evaluation are shown in Table 7.

As expected, the performance of MultiPRF on ideal translations is the best followed by Google Translate and the Basic SMT system. The results demonstrate that translation using the basic SMT system improves over monolingual MBF, especially P@5 and P@10. This shows that the performance of MultiPRF improves performance with any reasonably good query translation system.



| Language | Source Collection | Assisting Collection | No. of Docs. in Source Collection | No. of Docs. in Assisting Collection | MAP | GMAP |
|---|---|---|---|---|---|---|
| German | DE-01+02 | DE-03 | 225371 | 294809 | 0.4445 | 0.2328 |
| | DE-01+02 | EN-00+01+02 | 225371 | 113005 | **0.4576** | **0.2721** |
| French | FR-01+02 | FR-06 | 87191 | 177452 | 0.4394 | 0.2507 |
| | FR-01+02 | EN-00+01+02 | 87191 | 113005 | **0.4535** | **0.2721** |

**Table 8: Comparison of MultiPRF performance with MBF using an assisting collection in the same language. The coverage of the source and assisting collections is also given for comparison.**

## 6.2 Comparison with Assisting Collection in Same Language

One of the prime reasons for improvement in MultiPRF performance is good monolingual performance of assisting collection. The natural question which may then arise is whether the assisting collection needs to be in a different language. In this section, we study the performance of MultiPRF when the assisting collection is in the same language. Given a query, we use MBF on both source and assisting collections and interpolate the resultant feedback models. The final interpolated model is used to rerank the corpus and produce the final results. For the experiments, we use the French and German collections (FR-01+02, DE-01+02) since they have additional collections (FR-06, DE-03) with larger coverage in their own language. The results of comparison are shown in Table 8.

From the results, we notice that although the coverage of assisting collections in the source language is more than that of English, MBF still performs poorly when compared to MultiPRF. This can be attributed to the following reasons a) the MBF performance of a query, which is ambiguous or hard in the source language collection, will be bad due to the poor quality of top $k$ documents retrieved during initial retrieval. The quality of the top $k$ documents will not change if the same ambiguous query is given to assisting collection in the source language. However, if source and assisting languages differ, the ambiguity may get resolved during translation causing an improvement in MBF performance. The above intuition is confirmed by the decrease in robustness, as reflected in the GMAP scores, when the source and target languages are same. b) it still suffers from the fundamental limitation of monolingual PRF *i.e.* the expansion terms included are only based on co-occurrence relations and does not include lexically and semantically related terms.

## 6.3 Comparison with Thesaurus Based Expansion in Source Language

As discussed earlier, another major source of improvement in MultiPRF is due to the inclusion of lexically and semantically related terms. However, this alone does not justify the use of an assisting collection in a different language since the same effect could be achieved by using *thesaurus based expansion* in the source language. In this section, we show that augmenting MBF with both thesaurus based expansion and assisting collection in the same language is not effective when compared to MultiPRF.

Since there is no publicly available thesauri for the above mentioned European languages, as proposed in Xu *et al.*

| | MAP | | | | P@5 | | | | P@10 | | | | GMAP | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | MPRF | MPRF | MPRF | | MPRF | MPRF | MPRF | | MPRF | MPRF | MPRF | | MPRF | MPRF | MPRF |
| | MBF | SMT | GT | Ideal | MBF | SMT | GT | Ideal | MBF | SMT | GT | Ideal | MBF | SMT | GT | Ideal |
| **FR-01+02** | 0.4342 | 0.4494 | 0.4535 | 0.4633 | 0.4636 | 0.4818 | 0.4818 | 0.4864 | 0.4068 | 0.4239 | 0.4386 | 0.4477 | 0.2395 | 0.245 | 0.2721 | 0.2965 |
| **FR-03+05** | 0.3529 | 0.3576 | 0.3694 | 0.3762 | 0.4545 | 0.4707 | 0.4768 | 0.4889 | 0.404 | 0.4141 | 0.4202 | 0.4323 | 0.1324 | 0.1329 | 0.1411 | 0.1636 |
| **DE-01+02** | 0.4229 | 0.4275 | 0.4576 | 0.4639 | 0.5341 | 0.5523 | 0.6 | 0.6 | 0.4864 | 0.5125 | 0.5271 | 0.5386 | 0.2492 | 0.2032 | 0.2721 | 0.2816 |
| **DE-03** | 0.4274 | 0.4236 | 0.4355 | 0.4388 | 0.5098 | 0.5294 | 0.5412 | 0.5451 | 0.4784 | 0.4863 | 0.498 | 0.4922 | 0.1243 | 0.1225 | 0.1771 | 0.1981 |

**Table 7: Results comparing the performance of MultiPRF approach over the baseline MBF approach with Google Translate and another SMT system trained using Europarl corpus.**

[31], we learn a probabilistic thesaurus $P_{L \rightarrow L}$, in source language $L$, from the probabilistic bi-lingual dictionaries in L-English $P_{L \rightarrow E}$ and English-L $P_{E \rightarrow L}$. Given two words $s_1$ and $s_2$ in source language $L$ and $e$ is a word in English ($E$), $P_{L \rightarrow L}$ is given by:

$$\begin{aligned} P_{L \rightarrow L}(s_2|s_1) &= \sum_{\forall e \in E} P_{L \rightarrow L}(s_2, e|s_1) \\ &= \sum_{\forall e \in E} P_{E \rightarrow L}(s_2|e) \cdot P_{L \rightarrow E}(e|s_1) \\ & \quad \text{(Assuming } s_2,\, s_1 \text{ are independent given } e) \end{aligned}$$

Lexically and semantically related words like morphological variants and synonyms have a high probability score in $P_{L \rightarrow L}$ since they usually map to the same word in the target language. Given a query, we initially run MBF in the source language and let $\Theta_L^F$ be the resultant feedback model. Later, we use the probabilistic thesauri to expand the feedback model as follows:

$$P(f|\Theta_L^{Thesaurus}) = \sum_{\forall s \in S} P_{L \rightarrow L}(f|s) \cdot P(s|\Theta_L^F)$$

The above step includes morphological variants and synonyms for the terms in the feedback model. The final model is obtained by interpolating the $\Theta_L^{Thesaurus}$ with the MBF model $\Theta_L^F$ as shown in Equation 3.

For the above experiments, we use the FR-01+02 and DE-01+02 French and German collections. The results of comparison is shown in Figure 2. It shows that MBF with both thesaurus based expansion and assisting collection in the source language does not perform as well as MultiPRF. MultiPRF automatically combines the advantage of PRF in two different collections and thesaurus based expansion. This addresses the fundamental limitations of MBF and results in an improvement of both retrieval performance and robustness.

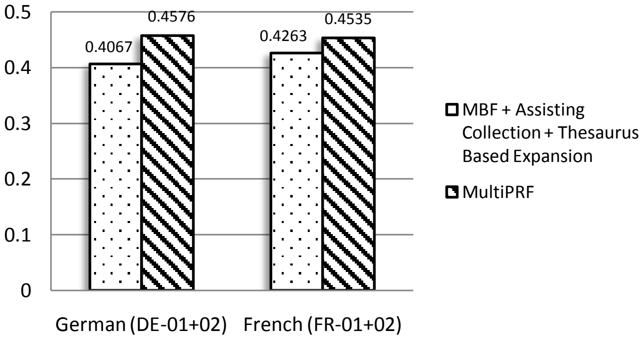


**Figure 2: MAP score comparison of MultiPRF and MBF with assisting collection in same language and Thesaurus Based Expansion. In MBF experiments, FR-06 and DE-03 were used as assisting collections for French and German respectively.**

## 7. CONCLUSION AND FUTURE WORK

We presented a novel approach to PRF called Multilingual PRF in which the performance of PRF in a language is improved by taking the help of another language collection. We also showed that MultiPRF addresses the fundamental limitations of monolingual PRF, *viz.*, (i) the inability to include term associations based on lexical and semantic relationships and (ii) sensitivity to the performance of the initial retrieval algorithm. Experiments on standard CLEF collections across a wide range of language pairs with varied degree of familial relationships show that MultiPRF consistently and significantly outperforms monolingual PRF both in terms of robustness and retrieval accuracy. Our error analysis pointed to the following contributing factors: (i) inaccuracies in query translation including the presence of out-of-vocabulary terms, (ii) poor retrieval on English query, and in a few rare cases, (iii) inaccuracy in the back translation. We feel we have taken only the first step towards a direction of work with rich potential, *viz. how a language can help another with respect to pseudo-relevance feedback.*

As part of future work, we plan to vary the assisting language and study its effect on MultiPRF performance. Also, we would like to remove the dependence of MultiPRF approach on availability of parallel corpora in the assisting language.

## 8. REFERENCES

[1] G. Amati, C. Carpineto, and G. Romano. Query Difficulty, Robustness, and Selective Application of Query Expansion. In *ECIR '04*, Sunderland, UK, pages 127–137, 2004.

[2] A. Berger and J. D. Lafferty. Information Retrieval as Statistical Translation. In *SIGIR '99*, pages 222–229, Berkeley, USA, 1999. ACM.

[3] M. Braschler and C. Peters. Cross-Language Evaluation Forum: Objectives, Results, Achievements. *Information Retrieval*, 7(1-2):7–31, 2004.

[4] C. Buckley, G. Salton, J. Allan, and A. Singhal. Automatic Query Expansion Using SMART : TREC 3. In *TREC-3*, pages 69–80, 1994.

[5] G. Cao, J.-Y. Nie, J. Gao, and S. Robertson. Selecting Good Expansion Terms for Pseudo-Relevance Feedback. In *SIGIR '08*, pages 243–250, NY, USA, 2008. ACM.

[6] K. Collins-Thompson and J. Callan. Query Expansion Using Random Walk Models. In *CIKM '05*, pages 704–711, NY, USA, 2005. ACM.

[7] S. Cronen-Townsend, Y. Zhou, and W. B. Croft. A Framework for Selective Query Expansion. In *CIKM '04*, pages 236–237, NY, USA, 2004. ACM.

[8] I. Dagan, A. Itai, and U. Schwall. Two Languages Are More Informative Than One. In *ACL '91*, pages 130–137, Morristown, NJ, USA, 1991. ACL.

[9] A. Dempster, N. Laird, and D. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, 39:1–38, 1977.

[10] T. S. Dumais, A. T. Letsche, L. M. Littman, and K. T. Landauer. Automatic Cross-Language Retrieval Using Latent Semantic Indexing. In *AAAI Technical Report SS-97-05*, pages 18–24, 1997.

[11] W. Gao, J. Blitzer, and M. Zhou. Using English Information in Non-English Web Search. In *iNEWS '08: ACM Workshop on Improving Non English Web Searching*, pages 17–24, NY, USA, 2008. ACM.

[12] D. Hawking, P. Thistlewaite, and D. Harman. Scaling Up The TREC Collection. *Information Retrieval*, 1(1-2):115–137, 1999.

[13] H. Hoang, A. Birch, C. Callison-Burch, R. Zens, R. Aachen, A. Constantin, M. Federico, N. Bertoldi, C. Dyer, B. Cowan, W. Shen, C. Moran, and O. Bojar. Moses: Open Source Toolkit for Statistical Machine Translation. In *ACL '07*, Prague, Czech Republic, pages 177–180, 2007. ACL.

[14] John Lafferty and Chengxiang Zhai. Probabilistic Relevance Models Based on Document and Query Generation. In *Language Modeling for Information Retrieval*, volume 13, pages 1–10. Kluwer International Series on IR, 2003.

[15] K. S. Jones, S. Walker, and S. E. Robertson. A Probabilistic Model of Information Retrieval: Development and Comparative Experiments. *Information Processing and Management*, 36(6):779–808, 2000.

[16] J. Lafferty and C. Zhai. Document Language Models, Query Models, and Risk Minimization for Information Retrieval. In *SIGIR '01*, pages 111–119, NY, USA, 2001. ACM.

[17] V. Lavrenko and W. B. Croft. Relevance Based Language Models. In *SIGIR '01*, pages 120–127, NY, USA, 2001. ACM.

[18] D. Metzler and W. B. Croft. Latent Concept Expansion Using Markov Random Fields. In *SIGIR '07*, pages 311–318, NY, USA, 2007. ACM.

[19] M. Mitra, A. Singhal, and C. Buckley. Improving Automatic Query Expansion. In *SIGIR '98*, pages 206–214, NY, USA, 1998. ACM.

[20] F. J. Och and H. Ney. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51, 2003.

[21] I. Ounis, G. Amati, P. V., B. He, C. Macdonald, and Johnson. Terrier Information Retrieval Platform. In *ECIR '05*, Volume 3408 of *Lecture Notes in Computer Science*, pages 517–519. Springer, 2005.

[22] K. Philipp. Europarl: A Parallel Corpus for Statistical Machine Translation. In *MT Summit*, 2005.

[23] S. Robertson. On GMAP: and Other Transformations. In *CIKM '06*, pages 78–83, NY, USA, 2006. ACM.

[24] T. Sakai, T. Manabe, and M. Koyama. Flexible Pseudo-Relevance Feedback via Selective Sampling. *ACM Transactions on Asian Language Information Processing (TALIP)*, 4(2):111–135, 2005.

[25] T. Tao and C. Zhai. Regularized Estimation of Mixture Models for Robust Pseudo-Relevance Feedback. In *SIGIR '06*, pages 162–169, NY, USA, 2006. ACM.

[26] J. Tiedemann. The Use of Parallel Corpora in Monolingual Lexicography - How Word Alignment Can Identify Morphological and Semantic Relations. In *Proceedings of the 6th Conference on Computational Lexicography and Corpus Research (COMPLEX)*, pages 143–151, Birmingham, UK, 28 June - 1 July 2001.

[27] E. Voorhees. Overview of The TREC 2005 Robust Retrieval Track. In *E. M. Voorhees and L. P. Buckland, Editors, The Fourteenth Text REtrieval Conference, TREC 2005*, Gaithersburg, MD, 2006. NIST.

[28] E. M. Voorhees. Query Expansion Using Lexical-Semantic Relations. In *SIGIR '94*, pages 61–69, NY, USA, 1994. Springer-Verlag.

[29] D. Wu, D. He, H. Ji, and R. Grishman. A Study of Using an Out-Of-Box Commercial MT System for Query Translation in CLIR. In *iNEWS '08: ACM Workshop on Improving Non English Web Searching*, pages 71–76, New York, NY, USA, 2008. ACM.

[30] J. Xu and W. B. Croft. Improving the Effectiveness of Information Retrieval with Local Context Analysis. *ACM Transactions on Information Systems*, 18(1):79–112, 2000.

[31] J. Xu, A. Fraser, and R. Weischedel. Empirical Studies in Strategies for Arabic Retrieval. In *SIGIR '02*, pages 269–274, NY, USA, 2002. ACM.

[32] Y. Xu, G. J. Jones, and B. Wang. Query Dependent Pseudo-Relevance Feedback Based on Wikipedia. In *SIGIR '09*, pages 59–66, NY, USA, 2009. ACM.

[33] C. Zhai and J. Lafferty. Model-based Feedback in the Language Modeling Approach to Information Retrieval. In *CIKM '01*, pages 403–410, NY, USA, 2001. ACM Press.

[34] C. Zhai and J. Lafferty. A Study of Smoothing Methods for Language Models applied to Information Retrieval. *ACM Transactions on Information Systems*, 22(2):179–214, 2004.