Inf Retrieval
DOI 10.1007/s10791-007-9036-6

An effective and efficient results merging strategy for multilingual information retrieval in federated search environments

Luo Si · Jamie Callan · Suleyman Cetintas · Hao Yuan

Received: 11 December 2006 / Accepted: 16 October 2007

Abstract Multilingual information retrieval is generally understood to mean the retrieval of relevant information in multiple target languages in response to a user query in a single source language. In a multilingual federated search environment, different information sources contain documents in different languages. A general search strategy in multilingual federated search environments is to translate the user query to each language of the information sources and run a monolingual search in each information source. It is then necessary to obtain a single ranked document list by merging the individual ranked lists from the information sources that are in different languages. This is known as the results merging problem for multilingual information retrieval. Previous research has shown that the simple approach of normalizing source-specific document scores is not effective. On the other side, a more effective merging method was proposed to download and translate all retrieved documents into the source language and generate the final ranked list by running a monolingual search in the search client. The latter method is more effective but is associated with a large amount of online communication and computation costs. This paper proposes an effective and efficient approach for the results merging task of multilingual ranked lists. Particularly, it downloads only a small number of documents from the individual ranked lists of each user query to calculate comparable document scores by utilizing both the query-based translation method and the document-based translation method. Then, query-specific and source-specific transformation models can be trained for

L. Si (✉) · S. Cetintas · H. Yuan
Department of Computer Science, Purdue University, West Lafayette, IN 47906, USA
e-mail: lsi@cs.purdue.edu

S. Cetintas
e-mail: scetinta@cs.purdue.edu

H. Yuan
e-mail: yuan3@cs.purdue.edu

J. Callan
Language Technology Inst, School of Computer Science, Carnegie Mellon University,
Pittsburgh, PA 15213, USA
e-mail: callan@cs.cmu.edu

individual ranked lists by using the information of these downloaded documents. These transformation models are used to estimate comparable document scores for all retrieved documents and thus the documents can be sorted into a final ranked list. This merging approach is efficient as only a subset of the retrieved documents are downloaded and translated online. Furthermore, an extensive set of experiments on the Cross-Language Evaluation Forum (CLEF) (http://www.clef-campaign.org/) data has demonstrated the effectiveness of the query-specific and source-specific results merging algorithm against other alternatives. The new research in this paper proposes different variants of the query-specific and source-specific results merging algorithm with different transformation models. This paper also provides thorough experimental results as well as detailed analysis. All of the work substantially extends the preliminary research in (Si and Callan, in: Peters (ed.) Results of the cross-language evaluation forum-CLEF 2005, 2005).

Keywords Results merging · Federated search · Multilingual information retrieval

1 Introduction

Multilingual (aka "cross-lingual" or "translingual") information retrieval consists of providing a query in a source language and searching document collections in one or many other languages (Hull and Grefenstette 1996; Ballesteros and Croft 1997; Oard and Diekema 1998; Xu et al. 2001; Levow et al. 2004; Jones et al. 2005). The multilingual information retrieval task has received a lot of attention with the increasing accessibility of diverse on-line international text collections, including the World Wide Web.

Information in multiple languages is often scattered among many multilingual information sources on local area networks or the Internet. The information within these multilingual information sources is generally created and maintained by independent information providers. It is often impossible or impractical to crawl all the contents of the independent information sources and download them into a single centralized database due to the information property protection or the frequent information updates. On the other side, many information sources provide their own source-specific text search engines to enable the access of the contents. The task of searching the information in multiple languages behind the text search engines of multilingual information sources is called multilingual information retrieval in federated search environments. Particularly, this paper focuses on solutions in uncooperative federated search environments, where source-specific search engines are allowed to choose different retrieval algorithms.

There are three main problems for multilingual information retrieval in federated search environments. First, information about the contents of each information source must be learned (i.e., resource description). Second, when it is necessary, a set of multilingual information sources should be selected for a particular user query (i.e., resource selection). Third, after the ranked lists have been returned from selected sources, they must be merged into a single multilingual ranked list (i.e., results merging). Particularly, this paper addresses the results merging problem of multilingual information retrieval in federated search environments.

Most of the previous research on merging multilingual ranked lists has been done within the CLEF community (e.g., Martínez-Santiago et al. 2002; Savoy 2002, 2003; Chen and Gey 2003; Rogati and Yang 2003). Many of these algorithms can be applied in federated search environments. One simple approach is to normalize and standardize the retrieval scores in each ranked list. However, this simple approach was shown to be ineffective

(Savoy 2002). Another approach was proposed to build a source-specific logistic transformation model for an information source to estimate the probabilities of relevance for documents in the ranked lists from this source by utilizing the rank and document score information of the documents (Savoy 2002). These methods are source-specific because different model parameters are generated for different languages to estimate the probabilities of relevance. However, for different queries, the same model is applied for documents from each source. This may be problematic because a document from one source may contribute different values for different queries (e.g., there may be a lot of relevant documents from one source for query A but very few for query B). To improve the merging accuracy, a more effective method (Martínez-Santiago et al. 2002; Rogati and Yang 2003) was proposed in previous research suggesting to download all the retrieved documents, translate them into the source language and generate the final ranked lists by running a monolingual search on the translated documents. This method provides a query-specific and source-specific merging strategy. It is more effective, but it is associated with a large amount of online communication and computation costs for downloading and translating all the retrieved documents.

This paper proposes an effective and efficient algorithm for merging multilingual ranked lists. Particularly, for each user query, this algorithm only downloads a subset of the retrieved documents from the individual ranked lists. These downloaded documents are indexed and translated into the source language. Then a multilingual centralized information retrieval algorithm is applied to calculate comparable scores for all the downloaded documents by utilizing both the query-based translation method and the document-based translation method. These documents serve as the training data to estimate the transformation models, which map source-specific document scores to comparable document scores. Finally, the query-specific and source-specific transformation models are used to estimate comparable scores for all retrieved documents and thus the documents can be sorted into a single final ranked list.

The proposed merging algorithm is efficient because only a subset of the retrieved documents are downloaded and translated online. This merging algorithm is also effective. It utilizes an effective multilingual centralized information retrieval algorithm to calculate comparable document scores for the downloaded documents, which are used to train the query-specific and source-specific transformation models. Therefore, the estimated comparable document scores from the transformation models tend to be effective to sort all the retrieved documents. Furthermore, this paper shows that it is often helpful to utilize both the estimated comparable document scores and the accurate comparable document scores for results merging whenever the accurate scores are available.

The proposed merging algorithm of multilingual information retrieval in federated search environments is similar to the results merging algorithms in monolingual environments. However, the results merging problem in multilingual federated search environments is more complicated than the merging problem in monolingual environments since the language barrier has to be crossed in order to rank the documents in different languages.

Three test environments have been created using the CLEF data. Experiments have been conducted on these test environments to demonstrate the effectiveness of the query-specific and source-specific algorithm against several other alternatives. Furthermore, an extensive set of experiments is designed to investigate the effectiveness of different variants of the proposed merging algorithm. First, two variants of the proposed merging algorithm are studied by utilizing the logistic transformation model and the linear transformation model. Second, the accuracy of the proposed query-specific and source-specific algorithm is investigated when it is allowed to download different number of documents. Third, experiments are conducted to show the advantage of utilizing both the estimated document scores and the accurate document scores for results merging when the accurate document scores are available. Finally, different selection strategies for choosing to-be-downloaded documents were investigated for the merging accuracy.

The rest of the paper is arranged as follows: Section 2 introduces an effective multilingual centralized information retrieval algorithm using both query-based translation method and document-based translation method. Section 3 first discusses the previous research of a query-independent and source-specific merging method and proposes a more effective variant of it; Section 3 then proposes a new query-specific and source-specific results merging method. Section 4 discusses the experimental methodology. Section 5 briefly shows the performance of the proposed multilingual centralized information retrieval algorithm. Section 6 presents a set of experiments to demonstrate the effectiveness of the proposed results merging algorithm and investigates the behaviors of different variants of this merging algorithm. Section 7 concludes this work.

2 Multilingual centralized information retrieval

Multilingual centralized information retrieval is not the focus of this paper. However, an effective multilingual centralized information retrieval solution is very important in order to obtain high merging accuracy of multilingual results merging algorithms. This section briefly presents our multilingual centralized information retrieval algorithm.

Considerable evidence from previous research (Kamps et al. 2003; Savoy 2003; Chen and Gey 2003) has suggested utilizing multiple retrieval methods for multilingual centralized information retrieval and applying each retrieval method with multiple translation methods. One particular approach first generates results across languages by using the same retrieval algorithm with the same translation method and then merges all the disjoint ranked lists from the particular method into a simple multilingual ranked list (Chen and Gey 2003). Similarly, several simple multilingual results can be obtained by applying different retrieval algorithms. These multilingual ranked lists contain the same set of documents but have different rankings. Finally, those simple multilingual ranked lists can be combined into a more accurate multilingual ranked list. Figure 1 shows a more detailed representation of this multilingual centralized information retrieval method. This method has been shown to be more effective than several other multilingual centralized information retrieval algorithms (Chen and Gey 2003). In this paper we adopt this method for multilingual centralized information retrieval.

Section 2.1 briefly discusses the translation tool. Section 2.2 presents two multilingual information retrieval algorithms based on query-based translation and document-based translation respectively. Section 2.3 proposes a method to combine the results from multiple multilingual information retrieval algorithms.

2.1 Translation matrices by learning from parallel corpus

Translation tools have been widely used in multilingual information retrieval algorithms to cross the language barriers. The translation process in this work is mainly accomplished in a word-by-word approach by using the translation matrices generated from the European

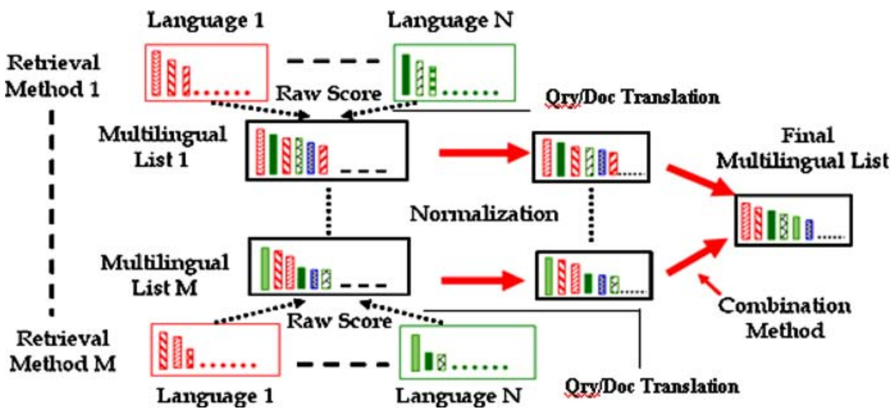

Fig. 1 The multilingual centralized information retrieval method that combines all results from a specific method into a multilingual result, and then combines the results from all methods into a final list

Parliament Proceedings Parallel Corpus,[1] which contains aligned sentences in multiple languages. Specifically, the GIZA++ (Och and Ney 2000) tool is utilized to construct the translation matrices (Brown et al. 1993) from the words in the source language (i.e., English in this work) to the words in the target language or from the words in the target language to the words in the source language. A probability value is assigned to each translation pair, indicating how probable the translation is.

2.2 Multilingual information retrieval via query translation or document translation

One straightforward multilingual information retrieval method is to translate the user query into different languages, search available information sources using the translated queries to acquire language-specific results, and merge the results into a single final multilingual ranked list.

Specifically, each query word is translated into the top 3 candidate target words in the translation matrices. Each of the three translation words is associated with a normalized weight (i.e., the sum of the three weights is 1) proportional to the corresponding weight in translation matrices. Since the vocabulary of the parallel corpus is limited, we also use word-by-word translation results from the online machine translation software Systran[2] as a complement. A weight of 0.2 is assigned to the translation by Systran and a weight of 0.8 is assigned to the translation using parallel corpus. The translated queries are used to search the index built for each language. The weight of each query term is its weight in the translation representation. The Okapi retrieval algorithm (Roberson and Walker 1994) is applied to do the retrieval. As the same retrieval algorithm is applied on the corpus of different languages using either the original query (i.e., for the collection of documents in the source language) or the translated queries of the same lengths, the raw scores in the ranked lists are somewhat comparable. Therefore, these ranked lists can be merged together by their language-specific scores into a final ranked list, which has been used in

[1] http://people.csail.mit.edu/koehn/publications/europarl

[2] http://www.systransoft.com/index.html

previous research (Chen and Gey 2003). This retrieval method is formally denoted as Qry_Tran.

An alternative multilingual information retrieval method is to translate all documents in the target languages into the source language (i.e., English in this work) and apply a single query in the source language. The retrieval method based on document translation may have an advantage compared to the retrieval method based on query translation because the translation of long documents may be more accurate in preserving the semantic meaning than the translation of short queries. In the previous research (Chen and Gey 2003), an example was shown that although one English query term is not correctly translated into the corresponding Spanish word by query translation, this Spanish word may still be correctly translated into the English query term by document translation. Translation matrices built from parallel corpus are utilized for translating documents. For each word in the target language, its top 3 English translations are considered. Five word slots are allocated to the three candidates with weights proportional to their normalized translation probabilities. The choice of using five word slots instead of only one is made in order to utilize more possible translation words rather than utilizing only the single most possible translation word. Then all of the translated documents as well as the documents in the source language (expand each term to five identical terms in order to be consistent with the translated documents) are collected into a single monolingual database and indexed. Furthermore, the queries in the source language are used to search the monolingual database with the Okapi retrieval algorithm. This method is formally denoted as Doc_Tran.

In summary, two types of multilingual information retrieval methods based on query translation and document translation are utilized in this work. All the results merging algorithms in this paper have utilized retrieval methods based on both the query translation and the document translation.

2.3 Combining multilingual ranked lists

One simple combination algorithm is to favor the documents returned by several retrieval methods as well as the high-ranking documents returned by single types of retrieval methods. Let $\mathrm{S}_{rm}(\mathrm{d}_{k\_j})$ denote the raw document score (r stands for raw score) for the jth document retrieved from the mth multilingual ranked list for the kth query, $\mathrm{S}_{rm}(\mathrm{d}_{k\_\max})$ and $\mathrm{S}_{rm}(\mathrm{d}_{k\_\min})$ represent the maximum and minimum document scores in this ranked list respectively. The normalized score of the jth document is calculated as:

$$\mathrm{S}_m(\mathrm{d}_{k\_j}) = \frac{(\mathrm{S}_{rm}(\mathrm{d}_{k\_j}) - \mathrm{S}_{rm}(\mathrm{d}_{k\_\min}))}{(\mathrm{S}_{rm}(\mathrm{d}_{k\_\max}) - \mathrm{S}_{rm}(\mathrm{d}_{k\_\min}))} \quad (1)$$

where $\mathrm{S}_m(\mathrm{d}_{k\_j})$ is the normalized document score. Then the normalized document scores in different ranked lists are summed up for each individual document and then all the documents can be ranked accordingly. This method is called the *equal weight combination method* in this paper, which can be seen as a variant of the well-known CombSum (Lee 1997; Aslam 2001) algorithm for meta search.

It is possible to design more sophisticated combination methods than the equal weight method. One simple idea is to associate different weights with different ranked lists. The desirable weights can be estimated by maximizing the combination accuracy with some training queries. However, this is not the focus of this paper. In this work, the equal weight combination method is used to calculate *comparable document scores* for documents in different languages.

3 Results merging for multilingual information retrieval in federated search environments

Results merging is an important task for multilingual information retrieval in federated search environments. It is the primary focus of this paper. In this section, several results merging algorithms are proposed to work in uncooperative multilingual federated search environments. The documents within the information sources can only be accessed through their source-specific searching services while different sources may use different retrieval algorithms. It is assumed in this work that each source is monolingual in one target language. Thus we will use the phrase "language-specific" instead of "source-specific" in the rest of this paper as these two terms are equivalent. However, the merging methods can be naturally extended to the environments where there are multiple sources per language.

Previous research (Savoy 2002, 2003) proposed the method of learning transformation models from the human relevance judgments of queries in the past to map language-specific document scores to the probabilities of relevance and thus the retrieved documents across different languages can be ranked by their estimated probabilities of relevance. However, the same model is applied for different queries within a single language in this method. This may be problematic because a document's score may vary across different queries. An alternative approach (Martínez-Santiago et al. 2002; Rogati and Yang 2003) is to translate and index all returned documents across different languages for each query and apply a centralized retrieval algorithm to compute comparable document scores. Although this method is accurate, it is often associated with a large amount of computation costs and communication costs in federated search environments.

This section proposes a new approach to learn query-specific and language-specific models of transforming language-specific document scores into comparable document scores. In particular, a small subset of returned documents from each language is downloaded, translated and indexed at retrieval time to compute comparable document scores, and then the query-specific and language-specific models are trained by both the comparable document scores and the language-specific document scores for these small subsets of documents. The trained models are applied to the ranked lists of all languages to obtain comparable document scores and finally all the returned documents are merged into a single list and ranked by their comparable scores.

This section is organized as follows: Section 3.1 describes the approach of learning query-independent and language-specific transformation models and proposes an extended method of learning the model by maximizing the mean average precision (MAP) criterion; Section 3.2 proposes the new approach of learning query-specific and language-specific results merging algorithm.

3.1 Learning query-independent and language-specific merging models with training data

To make the retrieved results from different ranked lists comparable, one natural idea is to map all of the document scores to the probabilities of relevance and rank the documents accordingly. A logistic transformation model has been used in previous studies to achieve this goal (Savoy 2002, 2003). This method has been shown to be more effective than several other alternatives such as the round robin results merging method and several simple score normalization methods (Savoy 2002, 2003). Let us assume that there are altogether I ranked lists from different languages to merge, each of them provides J documents for each query and there are altogether K training queries with human relevance judgments. Particularly, d_{k_ij} represents the jth document within the ranked list in the ith language of training query k. The pair (r_{k_ij} , $\mathrm{S}_i(\mathrm{d}_{k\_j})$) represents the rank of document j and its normalized document score (normalized by Eq. 1). Then the probability of relevance of this document can be estimated using the logistic transformation model as:

$$\mathrm{P}(\mathrm{rel}|\mathrm{d}_{k\_ij}) = \frac{1}{1 + \exp(\mathrm{a}_i \mathrm{r}_{k\_ij} + \mathrm{b}_i \mathrm{S}_i(\mathrm{d}_{k\_j}) + \mathrm{c}_i)} \tag{2}$$

where a_i, b_i and c_i are the parameters of the language-specific model that transforms all document scores from the ith language to the corresponding probabilities of relevance. Note that the same model is applied for all documents retrieved from different queries, which indicates that the model is query-independent. The optimal model parameter values are acquired generally by maximizing the log-likelihood estimation (MLE) of training data (Savoy 2002) as follows:

$$\sum_{k,i,j} \mathrm{P}^*(\mathrm{rel}|\mathrm{d}_{k\_ij}) \log(\mathrm{P}(\mathrm{rel}|\mathrm{d}_{k\_ij})) \tag{3}$$

where $\mathrm{P}^*(\mathrm{rel}|\mathrm{d}_{k\_ij})$ is the empirical probability of a particular document. This is derived from human relevance judgments data of training queries, which is 1 when a document is relevant and 0 otherwise. This objective function is convex, which guarantees the existence of a single global optimal solution.

This method treats each relevant document equally. However, this may not be reasonable in real world applications. For example, query A has two relevant documents and query B has 100 relevant documents. A relevant document for query A is considered more important to users than a relevant document for query B. Therefore, if all the queries are of equal importance to the users, a more reasonable criterion for information retrieval evaluation is to treat all the queries equally instead of the individual relevant documents. The MAP criterion reflects this idea. Assuming that there are K training queries, the value of MAP is calculated as follows:

$$\frac{1}{\mathrm{K}} \sum_{k} \left[\frac{1}{\mathrm{N}_k} \sum_{j \in \mathrm{D}_k^+} \frac{\mathrm{rank}_k^+(\mathrm{j})}{\mathrm{j}} \right] \tag{4}$$

where D_k^+ is the set of ranks of the relevant documents in the final ranked list for the kth training query, $\mathrm{rank}_k^+(\mathrm{j})$ is the corresponding rank only among relevant documents, and N_k is the number of relevant documents for the kth training query.

Based on the above observation of prior research, this paper proposes a natural extension which trains the logistic transformation model by the MAP criterion. Different sets of model parameters $\{\mathrm{a}_i, \mathrm{b}_i \text{ and } \mathrm{c}_i, 1 \le i \le \mathrm{I}\}$ generate different sets of relevant documents for all K training queries as $\{\mathrm{D}_k^+, 1 \le k \le \mathrm{K}\}$ and thus achieve different MAP values. The goal of training is to find a set of model parameters that generates the highest MAP value. However, this problem is not a convex optimization problem and there exist multiple local maximas. In this work, a common solution is used to search with multiple initial points. This new algorithm of training logistic model for maximum MAP is called the logistic model with the MAP goal, while the previous algorithm (Savoy 2002, 2003) trained for maximum likelihood is called the logistic model with the MLE goal.

3.2 Learning query-specific and language-specific merging models

Savoy's query-independent and language-specific logistic transformation model (Sect. 3.1) applies the same model on the results of different queries for each language. This is problematic when the ranked lists of different queries have similar distributions of scores but different distributions of probabilities of relevance. This suggests that it is necessary to design a query-specific model to improve the results merging accuracy.

One query-specific solution is to download and translate all returned documents (i.e., often more than 100 top ranked documents) from different languages at the retrieval time and compute comparable document scores (Sect. 2) to merge them together (Martínez-Santiago et al. 2002; Rogati and Yang 2003). Since this results merging method downloads (also indexes and translates) all the documents to be merged, it is called the *complete downloading method.*

More specifically, the complete downloading method downloads all the documents to be merged. The user query is translated into different languages and the Okapi retrieval algorithm is applied to obtain language-specific document scores. All the documents are merged by the raw scores into the first multilingual ranked list. Then, all the downloaded documents are translated into the query language. The user query is applied to the translated documents by the Okapi retrieval algorithm to obtain document scores based on the document translation method. Furthermore, all the downloaded documents are merged by the raw scores into the second multilingual ranked list. Finally, these two multilingual ranked lists are combined by the equal weight combination method into a final ranked list.

The complete downloading method is effective. However, this method is associated with a large amount of communication costs of downloading the documents and computation costs of translating and indexing many documents.

In this section, a more efficient results merging algorithm is proposed to work in the multilingual federated search environments. It only downloads and calculates comparable document scores for a small subset of returned documents and trains query-specific and language-specific models, which transform language-specific document scores to comparable scores for all returned documents.

Particularly, only the top ranked documents in the ranked list of each information source are selected to be downloaded and used to calculate comparable document scores. Let us assume that the top L documents in the ranked list of each information source are downloaded. Let the pair $(\mathrm{S}_c(\mathrm{d}_{k'\_l}), \mathrm{S}_i(\mathrm{d}_{k'\_l}))$ denote the normalized comparable document score and the normalized language-specific score for the lth downloaded document of the ith information source for the query k'. The information of the comparable document scores and the language-specific document scores of the downloaded documents serve as the training data to estimate a transformation model. For the lth downloaded document of the ith information source for the query k', a transformation function $\mathrm{f}_\theta(\mathrm{S}_i(\mathrm{d}_{k'\_l}))$ with model parameter θ maps the language-specific document score to the comparable document score.

The desired model parameters of the transformation model can be found by minimizing the sum of the squared error between actual comparable document scores and the estimated comparable document scores as follows:

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \sum_{d_{k'\_l} \in \mathrm{D_L} \in \mathrm{D_{NL}}} \left(\mathrm{S}_c(\mathrm{d}_{k'\_l}) - \mathrm{f}_\theta\left(\mathrm{S}_i(\mathrm{d}_{k'\_l})\right)\right)^2 \tag{5}$$

where θ^* is the desired model parameter, $\mathrm{D_L}$ is the set of L downloaded documents from the source. $\mathrm{D_{NL}}$ is a pseudo set of L documents that have no document representations but have pseudo normalized comparable scores of zero and pseudo normalized language-specific scores of zero. This set of pseudo documents is introduced in order to make sure that the learned model ranks documents in an appropriate way (i.e., documents with large language-specific scores are assigned with large comparable document scores and thus rank higher in the final ranked list).

There are multiple choices of the transformation models. In this work, we investigate several transformation models as follows:

$$f_\theta\left(S_i(d_{k'\_l})\right) = a_{k\_i}*S_i(d_{k'\_l}) + b_{k\_i} \tag{6}$$

which is a linear transformation model with query-specific and language-specific parameters as a_{k_i} and b_{k_i}. This is a simple transformation model that has been shown to work with a limited amount of training data for results merging of monolingual federated search (Si and Callan 2003).

One problem with the linear transformation is that the output of the linear function is not bounded, which may give excessive advantages to the returned documents from specific information sources. Since the multilingual centralized retrieval algorithm that we use generates bounded comparable document scores, it is more appropriate to design a transformation model with bounded output as:

$$f_\theta\left(S_i(d_{k'\_l})\right) = \frac{1}{1+\exp\left(c_{k\_i}*S_i(d_{k'\_l}) + d_{k\_i}\right)} \tag{7}$$

which is a logistic transformation model with query-specific and language-specific parameters as c_{k_i} and $d_{k\_i.}$

The transformation models can be learned for all information sources by downloading a small amount of documents. These models are applied to all returned documents from all sources and thus the returned documents can be ranked according to their estimated comparable scores. Note that only the language-specific document scores are used in the above transformation model in Eqs. 6 and 7 while the document ranks in the language-specific ranked lists are not considered. This choice is different from the query-independent and language-specific results merging algorithm (Eq. 2). It is used here to avoid the overfitting problem caused by the limited amount of training data as the training data of the above query-specific transformation models is less than that of the query-independent model in Eq. 2.

The transformation models can estimate comparable document scores for all retrieved documents. At the same time, exact comparable document scores are available for all of the downloaded documents. One method to take advantage of this evidence is to combine them with the estimated comparable scores. In this work, they are combined together with equal weights (i.e., 0.5). Empirical studies will be shown later to demonstrate the effectiveness of this combination strategy.

4 Evaluation methodology

The CLEF provides a good opportunity to evaluate both the multilingual centralized information retrieval algorithm and the results merging algorithm for multilingual federated search. Particularly, the algorithms proposed in this section were evaluated by two multilingual information retrieval tasks of CLEF 2005: Multi-8 two-years-on retrieval and Multi-8 results merging.

Multi-8 2-years-on retrieval task is a multilingual centralized information retrieval task, which is to search documents in eight European languages (i.e., Dutch, English, Finnish, French, German, Italian, Spanish and Swedish) with queries in a single source language (i.e., English) in a centralized environment where we have full access to all the documents.

Multi-8 results merging task is to merge the ranked lists of eight different languages into a single final list. This is viewed in this work as a results merging task within uncooperative multilingual federated search environments. Altogether, there are eight information sources that contain documents in eight different languages. The documents can only be accessed through source-specific search engines.

Twenty training queries with relevance judgments (i.e., queries 141–160) are provided to tune the behavior of results merging algorithms and the algorithms are evaluated on the other 40 test queries (i.e., queries 161–200). In this work, the query-independent and language-specific results merging methods (Sect. 3.1) utilize the 20 training queries with relevance judgments to estimate the model parameters. However, the query-specific and language-specific results merging methods do not need the relevance judgments of the 20 training queries. On the other side, these query-specific and language-specific results merging methods create pseudo training data by obtaining the actual comparable document scores online as described in Sect. 3.2.

Two sets of language-specific ranked lists (i.e., ranked lists of eight languages) from the UniNE system (Savoy 2003) and the HummingBird system[3] were provided for each query of the results merging task of CLEF 2005. The accuracies of these two sets of ranked lists on all the 60 queries can be found in Tables 1 and 2. It can be seen that the ranked lists from the UniNE system are generally more accurate than the ranked lists from the HummingBird System.

Three test environments were created based on these two sets of data: UniNE test environment (ranked lists from the UniNE system), Hum test environment (ranked lists from the HummingBird system), and the Mix environment which includes four ranked lists from the UniNE system (i.e., English, French, German and Italian) and four ranked lists from the HummingBird System (i.e., Dutch, Finnish, Spanish and Swedish). The three test environments represent three types of testing scenarios (i.e., more-accurate language-specific ranked lists, less-accurate language-specific ranked lists and language-specific ranked lists with mixed accuracy).

The information of different types of corpus statistics is utilized by different variants of the Okapi algorithm for the query-based translation method and the document-based translation method as described in Sect. 2. More specifically, query-based sampling (Callan and Connel 1999) is used to acquire 3,000 sampled documents from each information source. These documents are indexed in a centralized environment. The language-specific corpus statistics (e.g., inverse document frequency) are obtained from the sampled documents in different languages. Furthermore, the sampled documents in the seven target languages are also translated into English by the translation method as described in Sect. 2. The translated documents and the original English documents are merged into a single English collection and the corresponding corpus statistics can be derived.

Some basic text preprocessing techniques have been utilized to process the documents in multiple languages in both of the two tasks are:

[3] http://www.hummingbird.com/products/searchserver/

Table 1 Language-specific retrieval accuracy in mean average precision of results from UniNE system

Language	Dutch	English	Finnish	French	German	Italian	Spanish	Swedish
q141–q200	0.431	0.536	0.192	0.491	0.513	0.486	0.483	0.435

Table 2 Language-specific retrieval accuracy in mean average precision of results from HummingBird system

Language	Dutch	English	Finnish	French	German	Italian	Spanish	Swedish
q141–q200	0.236	0.514	0.163	0.350	0.263	0.325	0.298	0.269

- Stopword lists: The INQUERY (Turtle 1990; Callan et al. 1992) stopword list was used for English documents. Stopword lists of Finnish, French, German, Italian, Spanish and Swedish were acquired from,[4] while the snowball stopword[5] list was used for Dutch;
- Stemming: Porter stemmer was used for English words. Dutch stemming algorithm was acquired from[4] and the stemming algorithms from[6] were used for the other six languages;
- Decompounding: Dutch, Finnish, German and Swedish are compound rich languages. The same set of decompounding procedures as described in previous research (Kamps et al. 2003) was utilized.

5 Experimental results: Multilingual centralized information retrieval

Multilingual centralized information retrieval is not the focus of this work. However, the proposed federated multilingual results merging algorithms need an effective multilingual centralized information retrieval algorithm to generate training data automatically. Therefore, the accuracy of the proposed multilingual centralized retrieval algorithm is briefly presented here.

The proposed multilingual centralized retrieval algorithm in this paper combines the results from the query-based translation method and the document-based translation method. Table 3 shows the accuracies (in mean average precision) of the query-based translation (i.e., Qry_Tran) and the document-based translation (i.e., Doc_Tran) multilingual information retrieval algorithms on the training queries (i.e., queries 141–160), the test queries (i.e., queries 161–200) and the overall accuracies. Furthermore, the results from these two multilingual information retrieval algorithms are combined together by the equal weight combination method described in Sect. 2.3. The accuracy of the combination method is also shown in Table 3. Note that none of the three multilingual centralized retrieval algorithms utilize any training information from the training queries. It can be seen that the combination method substantially improves the accuracies of both the query-based translation method and the document-based translation method. The effective combination method provides strong multilingual centralized information retrieval results

[4] http://www.unine.ch/info/clef/

[5] http://www.snowball.tartarus.org/

[6] http://people.csail.mit.edu/koehn/publications/europarl

Table 3 Mean average precision of different multilingual centralized information retrieval methods

Methods	q141–q160	q161–q200	q141–q200
Qry_Tran	0.312	0.335	0.327
Doc_Tran	0.346	0.360	0.356
M2_C05	0.372	0.411	0.402

Qry_Tran represents the query-based translation method. Doc_Tran represents the document-based translation method. M2_C05 represents the equal weight combination method by combining both the results from query-based translation method and the results from document-based translations method

and serves as an important component for the proposed multilingual results merging algorithm in federated search environments.

6 Experimental results: Results merging for multilingual federated search

This section presents the experimental results of multilingual results merging algorithms in federated search environments. All the results merging algorithms were evaluated on the three test environments as described in Sect. 4.

The empirical studies in this section were designed to answer the following four questions:

(1) *How effective is the query-specific and language-specific results merging algorithm compared with other merging algorithms?* Experimental results are shown to compare the accuracy of the *query-specific* and language-specific results merging algorithm with the *query-independent* and language-specific merging algorithm.

(2) *How effective is the query-specific and language-specific results merging algorithm when it is allowed to download different number of documents?* The accuracy of the query-specific and language-specific results merging algorithm is investigated when it is allowed to *download different number of documents* to create the transformation models. Especially we investigate whether the query-specific and language-specific results merging algorithm can obtain reasonably accurate results by downloading a very small number of documents.

(3) *How effective is it to utilize both the estimated document scores and accurate comparable document scores for results merging, whenever the accurate comparable document scores are available?* Empirical studies are conducted to compare the effectiveness of two variants of the query-specific and language-specific results merging algorithm. One variant *combines both the estimated comparable document scores and the accurate comparable document scores* while the other only uses the estimated comparable document scores.

(4) *How to select the to-be-downloaded documents for estimating the transformation models?* The effect of different *selection strategies for choosing to-be-downloaded documents* on the merging accuracy is investigated particularly when the results merging algorithm is only allowed to download a small number of documents.

6.1 The effectiveness of the query-specific and language-specific results merging algorithm versus the query-independent and language-specific merging algorithm

The first set of experiments were conducted to evaluate the effectiveness of the query-specific and language-specific multilingual results merging algorithm in federated search environments. Particularly, two variants of the query-specific and language-specific models are investigated by utilizing the linear transformation model and the logistic transformation model respectively. These two merging methods do not need relevance judgments as training data. On the other side, the two query-independent and language-specific algorithms download some documents on the fly to calculate comparable document scores in order to estimate transformation models for each user query. More detailed information of the algorithms is described in Sect. 3.

More specifically, the query-specific and language-specific merging methods are compared with two variants of the query-independent and language-specific results merging methods by optimizing the maximum likelihood estimation (MLE) and the mean average precision (MAP) criterion respectively. The two query-independent and language-specific methods are trained using 20 training queries (i.e., queries 141–160) with relevance judgments. The descriptions of the two query-independent and language-specific models can be found in Sect. 3.1. The two methods serve as the baseline since the query-independent and language-specific results merging algorithm by optimizing the MLE has been shown to outperform several other merging methods in previous research (Savoy 2003).

The accuracies of all the results merging algorithms on the three test environments are shown in Table 4. It can be seen that the accuracies of the merged results in the UniNE test environment are higher than the results in the Mix test environment, while the latter are higher than the results in the Hum test environment. This is consistent with our expectation that the language-specific ranked lists of the UniNE system are more accurate than those of the HummingBird system.

Furthermore, it can be seen from the first two rows of Table 4 that the query-independent and language-specific learning algorithm optimized for the MAP criterion is consistently more accurate than the algorithm optimized for the maximum likelihood criterion. This demonstrates the advantage of directly optimizing the MAP accuracy by treating different queries equally against the strategy of optimizing the maximum likelihood that does not directly consider MAP.

Our key idea of improving the merging accuracy is to introduce the query-specific and language-specific results merging algorithm. Table 4 shows the results of two query-specific and language-specific results merging methods with the linear transformation model and the logistic transformation model respectively. Both of these two methods download and process the top 10 or top 15 documents on the fly from each ranked list to estimate the transformation models. Note that neither of these two methods requires human relevance judgments for training data. Therefore, the results from the training query set (i.e., queries 141–160) and the results from the test query set (i.e., queries 161–200) were obtained independently. It can be seen that both of the two query-specific and language-specific results merging methods substantially outperform the query-independent and language-specific methods. For example, the query-specific and language-specific merging method with the logistic transformation model is better than any query-independent method by at least 10% in any test environment. This explicitly demonstrates the advantage of the query-specific and language-specific results merging algorithm against the query-independent merging algorithm.

Table 4 Mean average precision of the merged multilingual lists of different methods in three test environments (i.e., Hum, Mix and UniNE)

Methods	Hum			Mix			UniNE		
	q141–q160	q161–q200	q141–q200	q141–q160	q161–q200	q141–q200	q141–q160	q161–q200	q141–q200
TrainLog_MLE (QI)_	0.186	0.171	0.176	0.219	0.216	0.217	0.301	0.301	0.301
TrainLog_MAP (QI)	0.210	0.192	0.198	0.232	0.268	0.256	0.322	0.322	0.327
Top_10_C05_Linear (QS)	0.223	0.224	0.224	0.270	0.293	0.285	0.330	0.356	0.347
Top_15_C05_Linear (QS)	0.227	0.232	0.230	0.275	0.302	0.293	0.337	0.366	0.356
Top_10_C05_Log (QS)	0.229	0.253	0.245	0.280	0.329	0.313	0.332	0.395	0.374
Top_15_C05_Log (QS)	0.239	0.258	0.252	0.288	0.337	0.321	0.348	0.404	0.385

TrainLog_MLE represents the query-independent and language-specific model with logistic transformation model trained with the MLE goal. TrainLog_MAP represents the query-independent and language-specific model with logistic transformation model trained with the MAP goal. Top_X_C05_Linear represents the query-specific and language-specific merging method which generates the linear transformation model from the top X documents downloaded (i.e., 5 or 10 in this case). Top_X_C05_Log represents similar results merging algorithm but with the logistic transformation model. QI indicates a query-independent model while QS indicates a query-specific model

6.2 The effectiveness of the query-specific and language-specific results merging algorithm when it is allowed to download different number of documents

The experiments in this section are conducted to investigate the accuracy of the query-specific and language-specific results merging methods when they are allowed to download different number of documents to create the transformation models. Especially the experiments show that the query-specific and language-specific results merging methods can obtain reasonably accurate results from a small number of downloaded documents.

To provide a better analysis of the proposed query-specific and language-specific multilingual results merging algorithms, we also compare them with the complete downloading algorithm that downloads all or most of the documents from the ranked lists of different languages and merges the documents by their actual comparable document scores (i.e., C_X).

The results of the query-specific and language-specific results merging algorithm with the logistic transformation model are shown in Table 5 while the results of the query-specific and language-specific results merging algorithm with the linear transformation model are shown in Table 6. It can be seen from Tables 5 and 6 that both of the two query-specific and language-specific results merging methods become more accurate when they are allowed to download more documents on the fly to estimate the transformation models. This is consistent with our expectation that more training data from more downloaded documents makes the estimated transformation models more accurate. Furthermore, it is encouraging to see that with a very limited number of downloaded documents, the Top_3_C05 method provides merging results at least as accurate as the query-independent and language-specific methods as shown in Table 4. When the methods are allowed to download 10 documents (i.e., Top_10_C05) from each ranked list, the query-specific methods generate consistently more accurate merging results than the query-independent methods. Particularly, the Top_10_C05_Log algorithm has a more than 10% advantage over the query-independent methods in all the three test environments.

The results in Tables 5 and 6 show that there is a substantial gap in the accuracies between the query-specific and language-specific results merging algorithm and the complete downloading algorithm in the HummingBird test environment, especially when the query-specific and language-specific results merging algorithm is only allowed to download a small number of documents (i.e., 3, 5 or 10). This is mainly because most of the language-specific ranked lists generated from the HummingBird system are not accurate as shown in Table 2. Therefore, the transformation models trained using the document scores of a small number of documents from the language-specific ranked lists are not accurate enough to generate estimated comparable document scores. However, the accuracy gap between the query-specific and language-specific results merging algorithm and the complete downloading method is much smaller in the mixed and the UniNE test environments, where the language-specific ranked lists are more accurate.

Another observation from Tables 5 and 6 is that the query-specific and language-specific results merging algorithm with the logistic transformation model is generally more accurate than the query-specific and language-specific results merging algorithm with the linear transformation model. This confirms the advantage of the logistic transformation model by generating bounded results of estimated comparable document scores, while the linear transformation model generates unbounded results and thus may give excessive advantages to the returned documents from specific information sources.

Table 5 Mean average precision of the merged multilingual lists of the query-specific and language-specific results merging algorithm with the logistic transformation model and the complete downloading method in three test environments (i.e., Hum, Mix and UniNE)

Methods	Hum			Mix			UniNE		
	q141–q160	q161–q200	q141–q200	q141–q160	q161–q200	q141–q200	q141–q160	q161–q200	q141–q200
Top_3_C05_Log	0.189	0.231	0.217	0.239	0.288	0.271	0.278	0.348	0.325
Top_5_C05_Log	0.212	0.241	0.232	0.255	0.306	0.289	0.298	0.375	0.349
Top_10_C05_Log	0.229	0.253	0.245	0.280	0.329	0.313	0.332	0.395	0.374
Top_15_C05_Log	0.239	0.258	0.252	0.288	0.337	0.321	0.348	0.404	0.385
Top_30_C05_Log	0.260	0.270	0.267	0.304	0.341	0.329	0.356	0.401	0.386
Top_150_C05_Log	0.276	0.296	0.290	0.313	0.358	0.343	0.360	0.411	0.394
C_150	0.290	0.302	0.298	0.312	0.348	0.336	0.356	0.382	0.373
C_500	0.315	0.333	0.326	0.323	0.359	0.347	0.356	0.384	0.374
C_1000	0.324	0.343	0.337	0.328	0.366	0.354	0.352	0.391	0.378

Top_X_C05_Log represents the query-specific and language-specific merging method which generates the logistic transformation model from the top X documents downloaded. C_X represents that the top X documents from each list are downloaded and merged by their actual comparable document scores

Table 6 Mean average precision of the merged multilingual lists of the query-specific and language-specific results merging algorithm with the linear transformation model and the complete downloading method in three test environments (i.e., Hum, Mix and UniNE)

Methods	Hum			Mix			UniNE		
	q141–q160	q161–q200	q141–q200	q141–q160	q161–q200	q141–q200	q141–q160	q161–q200	q141–q200
Top_3_C05_Linear	0.207	0.208	0.208	0.252	0.264	0.260	0.307	0.320	0.316
Top_5_C05_Linear	0.213	0.214	0.214	0.260	0.278	0.272	0.313	0.339	0.330
Top_10_C05_Linear	0.223	0.224	0.224	0.270	0.293	0.285	0.330	0.356	0.347
Top_15_C05_Linear	0.227	0.232	0.230	0.275	0.302	0.293	0.337	0.366	0.356
Top_30_C05_Linear	0.245	0.247	0.247	0.292	0.315	0.307	0.344	0.374	0.364
Top_150_C05_Linear	0.278	0.291	0.287	0.309	0.352	0.338	0.356	0.403	0.387
C_150	0.290	0.302	0.298	0.312	0.348	0.336	0.356	0.382	0.373
C_500	0.315	0.333	0.326	0.323	0.359	0.347	0.356	0.384	0.374
C_1000	0.324	0.343	0.337	0.328	0.366	0.354	0.352	0.391	0.378

Top_X_C05_Linear represents the query-specific and language-specific merging method which generates the linear transformation model from the top X documents downloaded. C_X represents that the top X documents from each list are downloaded and merged by their actual comparable document scores

6.3 Combining the estimated comparable document scores and accurate comparable document scores for results merging

The transformation models of the query-specific and language-specific results merging algorithm can estimate comparable document scores for all the retrieved documents. At the same time, actual comparable document scores are available for all the downloaded documents. The results described in Sects. 6.1 and 6.2 are obtained by combining the estimated comparable document scores and the actual comparable document scores with equal weight (i.e., 0.5 and 0.5). The experiments in this section investigate the advantage of this combination approach against the results merging method that only uses the estimated comparable document scores.

The results in Tables 7 and 8 show that the performance of the two variants of the query-specific and language-specific results merging algorithm as one variant combines estimated comparable document scores and actual comparable document scores (i.e., Top_X_C05) and the other variant only utilizes estimated comparable document scores (i.e., Top_X). It can be seen from Tables 7 and 8 that the combination method generates as accurate results as the method that only utilizes estimated document scores (i.e., when the methods are only allowed to download the top 3 documents). Furthermore, the combination method consistently generates more accurate results when more documents (i.e., top 15 and top 150) can be downloaded. This set of experiments demonstrate that it is generally helpful to utilize both the estimated document scores and the accurate comparable document scores for results merging whenever the accurate comparable document scores are available.

6.4 The strategy to select the documents to be downloaded

The query-specific and language-specific results merging algorithm needs to download some documents on the fly to calculate comparable document scores for estimating transformation models. One particular problem is which set of documents the algorithm should download in order to estimate accurate transformation models. This is an important question especially when only a small number of documents are allowed to be downloaded.

This section studies two strategies of selecting the documents to be downloaded. The first strategy selects the top 3 documents from each language-specific ranked list, which is consistent with the methods that generate the experimental results in Sects. 6.1–6.3. The second selection strategy selects the top 1, 10 and 20 documents from each language-specific ranked list.

The results of the two selection strategies are shown in Tables 9 and 10. It can be seen that the Top_[1,10,20] selection strategy is at least as effective as the Top_3 selection strategy for the linear transformation model (i.e., results in Table 10). Furthermore, the Top_[1,10,20] is substantially better than the Top_3 selection strategy for the logistic transformation model. One possible explanation for the advantage of the Top_[1,10,20] selection strategy is that the training data from the top 1, 10 and 20 documents is more comprehensive than the training data from the top 1, 2 and 3 documents. Therefore, the transformation models generated by the Top_[1,10,20] selection strategy tend to be more accurate than the transformation models generated by the Top_3 selection strategy.

Table 7 Mean average precision of the merged multilingual lists of the query-specific and language-specific results merging algorithm with the logistic transformation model in the three test environments (i.e., Hum, Mix and UniNE)

Methods	Hum			Mix			UniNE		
	q141–q160	q161–q200	q141–q200	q141–q160	q161–q200	q141–q200	q141–q160	q161–q200	q141–q200
Top_3_Log	0.187	0.229	0.215	0.244	0.285	0.272	0.282	0.347	0.325
Top_3_C05_Log	0.189	0.231	0.217	0.239	0.288	0.271	0.278	0.348	0.325
Top_15_Log	0.223	0.243	0.237	0.282	0.324	0.310	0.341	0.392	0.375
Top_15_C05_Log	0.239	0.258	0.252	0.288	0.337	0.321	0.348	0.404	0.385
Top_150_Log	0.228	0.248	0.241	0.284	0.322	0.310	0.339	0.390	0.373
Top_150_C05_Log	0.276	0.296	0.290	0.313	0.358	0.343	0.360	0.411	0.394

Top_X_C05_Log represents the query-specific and language-specific merging method which generates the logistic transformation model from the top X documents downloaded and the equal weight combination of the estimated comparable scores and the actual comparable scores. Top_X_Log is the same as Top_X_C05_Log but it only uses the estimated comparable document scores

Table 8 Mean average precision of the merged multilingual lists of the query-specific and language-specific results merging algorithm with the linear transformation model and the complete downloading method in the three test environments (i.e., Hum, Mix and UniNE)

Methods	Hum			Mix			UniNE		
	q141–q160	q161–q200	q141–q200	q141–q160	q161–q200	q141–q200	q141–q160	q161–q200	q141–q200
Top_3_Linear	0.205	0.207	0.206	0.256	0.261	0.259	0.309	0.318	0.315
Top_3_C05_Linear	0.207	0.208	0.208	0.252	0.264	0.260	0.307	0.320	0.316
Top_15_Linear	0.213	0.207	0.209	0.266	0.279	0.274	0.325	0.344	0.337
Top_15_C05_Linear	0.227	0.232	0.230	0.275	0.302	0.293	0.337	0.366	0.356
Top_150_Linear	0.211	0.210	0.210	0.264	0.288	0.280	0.319	0.354	0.341
Top_150_C05_Linear	0.278	0.291	0.287	0.309	0.352	0.338	0.356	0.403	0.387

Top_X_C05_Linear represents the query-specific and language-specific merging method which generates the linear transformation model from the top X documents downloaded and the equal weight combination of the estimated comparable scores and the actual comparable scores. Top_X_Linear is the same as Top_X_C05_Linear but it only uses the estimated comparable document scores

Table 9 Mean average precision of the merged multilingual lists of the query-specific and language-specific results merging algorithm with the logistic transformation model in the three test environments (i.e., Hum, Mix and UniNE)

Methods	Hum			Mix			UniNE		
	q141–q160	q161–q200	q141–q200	q141–q160	q161–q200	q141–q200	q141–q160	q161–q200	q141–q200
Top_3_C05_Log	0.189	0.231	0.217	0.239	0.288	0.271	0.278	0.348	0.325
Top_[1,10,20]_Log	0.216	0.232	0.227	0.271	0.311	0.298	0.319	0.372	0.355

Top_3_C05_Log means the query-specific and language-specific merging method which generates the logistic transformation model by the top 3 documents downloaded and the equal weight combination of the estimated comparable scores and the actual comparable scores. Top_[1,10,20]_C05_Log is the same as Top_3_C05_Log but it downloads the top 1, 10 and 20 documents

Table 10 Mean average precision of the merged multilingual lists of the query-specific and language-specific results merging algorithm with the linear transformation model in the three test environments (i.e., Hum, Mix and UniNE)

Methods	Hum			Mix			UniNE		
	q141–q160	q161–q200	q141–q200	q141–q160	q161–q200	q141–q200	q141–q160	q161–q200	q141–q200
Top_3_C05_Linear	0.207	0.208	0.208	0.252	0.264	0.260	0.307	0.320	0.316
Top_[1,10,20]_Linear	0.209	0.205	0.206	0.262	0.275	0.271	0.315	0.332	0.326

Top_3_C05_Linear means the query-specific and language-specific merging method which generates the linear transformation model by the top 3 documents downloaded and the equal weight combination of the estimated comparable scores and the actual comparable scores. Top_[1,10,20]_C05_Linear is the same as Top_3_C05_Linear but it downloads the top 1, 10 and 20 documents

7 Conclusion

Multilingual information retrieval is the task to search document collections in many target languages given a user query in a specific source language. Multilingual text information is often scattered in federated search environments among many independent multilingual information sources on local area networks or the Internet. The common search solution in multilingual federated search environments is to translate the user query to the target language of each information source and run a monolingual search in each information source. An important research problem is how to merge the returned ranked lists into a single multilingual ranked list, which is referred to as the results merging problem of multilingual information retrieval in federated search environments. The focus of this paper is the results merging problem of multilingual information retrieval in federated search environments.

Many algorithms have been proposed for merging multilingual ranked lists. One simple approach is to normalize the scores in each ranked list and merge all the documents by the normalized scores. However, this simple approach is not effective. Another approach utilizes relevance judgments of some training queries to build query-independent and language-specific models and thus to estimate the probabilities of relevance for all returned documents. This query-independent and language-specific results merging algorithm generates more accurate results than the simple normalization approach but it is still not very effective. A more effective method downloads and translates all retrieved documents into the query language and generates the final ranked lists by running a monolingual search on the translated documents. This approach outperforms the query-independent and language-specific results merging algorithm. However, it is associated with a large amount of online communication and computation costs for downloading and translating all the retrieved documents.

This paper presents an effective and efficient method for merging multilingual ranked lists in federated search environments. Our approach only downloads a small subset of the retrieved documents from the individual ranked lists for each user query. These downloaded documents are indexed and translated into the query language. An effective multilingual centralized retrieval algorithm is proposed to calculate comparable document scores for the downloaded documents by utilizing both the query-based translation method and the document-based translation method. Query-specific and language-specific transformation models can be built using the training information from the downloaded documents. These transformation models are applied to estimate comparable document scores for all the returned documents. Finally, all the returned documents are sorted into a single ranked list by considering the estimated comparable document scores as well as the accurate comparable document scores when such information is available.

An extensive set of experiments have been conducted to demonstrate the effectiveness of the new query-specific and language-specific multilingual results merging algorithm. Three test environments are created from the data of the multilingual results merging task in the CLEF 2005. The results indicate that the proposed query-specific and language-specific result merging algorithm is consistently more effective than the query-independent merging algorithm even with a limited number of downloaded documents (e.g., top 10). The query-specific and language-specific merging algorithm with the logistic transformation model is often more accurate than the merging algorithm with the linear transformation model. Furthermore, extensive experiments have been conducted to investigate the effectiveness of the different variants of the proposed multilingual results merging algorithm. One set of results indicate that it is helpful to utilize both the estimated comparable document scores and the accurate comparable document scores for results merging when the accurate scores are available. Another set of results have shown that the to-be-downloaded documents should be selected from a wide range of the language-specific ranked lists particularly in the case when the merging algorithm is only allowed to download a small number of documents.

There are several research problems regarding to how to further improve the effectiveness and efficiency of the multilingual results merging algorithm proposed in this paper. First, the proposed merging algorithm uses equal weights to combine the estimated comparable scores and the actual comparable scores. However, this may not be the optimal solution for different multilingual text collections. It is useful to design a learning algorithm to automatically learn the desired combination weights. Second, the experiments have shown that the training data from a larger number of downloaded documents can generate more effective transformation models and thus we can get more accurate merging results. However, it is not efficient to download many documents on the fly. A unified utility maximization model is promising to provide a good trade-off between effectiveness and efficiency by automatically adjusting the number of documents to download. It is also a very interesting direction to design merging algorithms that do not need to download and process documents online but can still achieve reasonably good merging results. This is important for search environments where downloading and processing documents online is not practical (e.g., high query load systems).

References

Aslam, J. A., & Montague, M. (2001). Models for metasearch. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ‘01, New Orleans, Louisiana, United States* (pp. 276–284). New York, NY: ACM.

Ballesteros, L., & Croft, W. B. (1997). Phrasal translation and query expansion techniques for cross-language information retrieval. In N. J. Belkin, A. D. Narasimhalu, P. Willett, & W. Hersh (Eds.), *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ‘97, Philadelphia, Pennsylvania, United States, July 27–31, 1997* (pp. 84–91). New York, NY: ACM.

Brown, P. F, Pietra, D., Pietra, D, & Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics, 19*, 263–312.

Callan, J., & Connell, M. (2001). Query-based sampling of text databases. *ACM Transactions on Information Systems, 19*(2), 97–130.

CalIan, J. P., Croft, W. B., & Harding, S. M. (1992). The INQUERY retrieval system. In *Proceedings of the Third International Conference on Database and Expert Systems Applications, Valencia, Spain* (pp. 78–83). Springer-Verlag.

Chen A., & Gey, F. C. (2003). Combining query translation and document translation in cross-language retrieval. In C. Peters, J. Gonzalo, M. Braschler, et al. (Eds.), *4th Workshop of the Cross-Language Evaluation Forum, CLEF 2003, Lecture notes in Computer Science, Trondheim, Norway* (pp. 108–121). Springer-Verlag.

Hull, D. A., & Grefenstette, G. (1996). Query across languages: A dictionary-based approach to multilingual information retrieval. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ‘96, Zurich, Switzerland, August 18–22, 1996* (pp. 49–57). New York, NY: ACM.

Jones, G. J. F., Burke, M., Judge, J., Khasin, A., Lam-Adesina, A. M., & Wagner, J. (2005). Dublin City University at CLEF 2004: Experiments in monolingual, bilingual and multilingual retrieval. In *CLEF* (pp. 207–220).

Kamps, J., Monz, C., de Rijke, M., & Sigurbjörnsson, B. (2003). The University of Amsterdam at CLEF-2003. In *Results of the CLEF 2003 Cross-Language System Evaluation Campaign, Trondheim, Norway* (pp. 71–78).

Lee, J. H. (1997). Analyses of multiple evidence combination. In N. J. Belkin, A. D. Narasimhalu, P. Willett, & W. Hersh (Eds.), *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ‘97, Philadelphia, Pennsylvania, United States, July 27–31, 1997* (pp. 267–276). New York, NY: ACM.

Levow, G. A., Oard, D. W., & Resnik, P. (2004). Dictionary-based cross-language retrieval. *Information Processing and Management, 41*, 523–547.

Martínez-Santiago, F., Martin, M., & Ureña, A. (2002). SINAI on CLEF 2002: Experiments with merging strategies. In C. Peters (Ed.), *Results of the cross-language evaluation forum—CLEF 2002* (pp. 187–196).

Oard, D., & Diekema, A. (1998). Cross-language information retrieval. In M. Williams (Ed.), *Annual review of information science* (pp. 223–256).

Och, F. J., & Ney, H. (2000). Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics, Annual Meeting of the ACL, Hong Kong, October 03–06, 2000* (pp. 440–447). Morristown, NJ: Association for Computational Linguistics.

Robertson, S. E., & Walker, S. (1994). Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In W. B. Croft & C. J. van Rijsbergen (Eds.), *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Dublin, Ireland, July 03–06, 1994* (pp. 232–241). New York, NY: Springer-Verlag New York.

Rogati, M., & Yang, Y. M. (2003). CONTROL: CLEF-2003 with open, transparent resources off-line. Experiments with merging strategies. In C. Peters (Ed.), *Results of the cross-language evaluation forum-CLEF*.

Savoy, J. (2002). Report on CLEF 2002 experiments: Combining multiple sources of evidence. In C. Peters et al. (Eds.), *Advances in cross-language information retrieval, LNCS* (Vol. 2785, pp. 66–90). Berlin: Springer-Verlag.

Savoy, J. (2003). Report on CLEF-2003 multilingual tracks. In: *Procedings of CLEF 2003, Trondheim, Norway* (pp. 7–12).

Si, L., & Callan, J. (2003). A semi-supervised learning method to merge search engine results. *ACM Transactions on Information Systems, 24*(4), 457–491.

Si, L., & Callan, J. (2005). CLEF2005: Multilingual retrieval by combining multiple multilingual ranked lists. In C. Peters (Ed.), *Results of the cross-language evaluation forum-CLEF 2005*.

Turtle, H. (1990). *Inference networks for document retrieval*. Technical Report COINS Report 90-7, Computer and Information Science Department, University of Massachusetts, Amherst.

Xu, J., Weischedel, R., & Nguyen, C. (2001). Evaluating a probabilistic model for cross-lingual information retrieval. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '01, New Orleans, Louisiana, United States* (pp. 105–110). New York, NY: ACM.