Cross-lingual Information Retrieval based on Multiple Indexes

Philipp Sorg, Marlon Braun, David Nicolay
Institut AIFB, Universität Karlsruhe
sorg@aifb.uni-karlsruhe.de
marlon.braun@t-online.de
davidnicolay85@yahoo.de

Philipp Cimiano
Universität Bielefeld
cimiano@techfak.uni-bielefeld.de

Abstract

In this paper we present the technical details of the retrieval system with which we participated at the CLEF09 Ad-hoc TEL task. We present a retrieval approach based on multiple indexes for different languages which is combined with a concept-based retrieval approach based on Explicit Semantic Analysis. In order to create the language-specific indices for each language, a language detection approach is applied as preprocessing step. We combine the different indices through rank aggregation and present our experimental results with different rank aggregation strategies. Our results show that the use of multiple indices (one for each language) does not improve upon a baseline index containing documents in all languages. The combination with concept based retrieval, however, results in better retrieval performance in some of the cases considered. For the bi-lingual tasks the final retrieval results of our system were the 5th best results on the BL dataset and the second best on the BNF dataset.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval

General Terms

Measurement, Performance, Experimentation

Keywords

Cross-language Information Retrieval, Explicit Semantic Analysis, Rank Aggregation, Machine Translation

1 Introduction

There are two important paradigms that can be applied to the problem of cross-language retrieval: concept-based retrieval approaches as well as approaches exploiting machine translation (MT). Concept-based methods map documents and queries into a language-independent concept space [5]. MT-based methods translate the queries or documents into the target language or into all target languages [3].

Most machine translation based approaches work for specific language pairs. The topic is given in a specific source language and all documents in the corpus are given in a defined target language.

In this paper we extend this model to be able to handle corpora containing documents in multiple languages and moreover documents containing fields in different languages. Our approach is directly motivated by the CLEF Ad-hoc TEL task. Here the target collection contains documents in different languages and the task is to find relevant documents in all languages for given topics. Our hypothesis is that retrieval can be improved by translating topics to all languages of the corpus, performing a language specific search for each translation and aggregating all the results for the single indices into one final ranking.

Another important question we address in this paper is whether concept- and MT-based techniques can be successfully combined to increase the performance of CLIR compared to concept-based and MT-based techniques alone.

For both problems, i.e. retrieval using multiple indices and combination of MT-based and concept based retrieval, relevance measures computed by different models have to be combined to an aggregated relevance score. A common approach to this problem that we also use in this paper is rank aggregation. This means that the final scores of each model are used as input values for the aggregation function. In the following we will describe the main techniques used in related work to combine different retrieval approaches.

In order to combine concept-based retrieval and term-based retrieval, Müller and Gurevych [4] use Wikipedia and Wiktionary as background knowledge to improve the retrieval performance on a mono-lingual search task. They were able to improve the performance measured by mean average precision by 34% compared to the bag-of-words baseline. Similar to our approach they use Explicit Semantic Analysis [2] for concept-based retrieval. In this paper we extend this approach to CLIR and investigate different strategies to combine evidence from different retrieval approaches.

Croft [1] describes different strategies to combine IR techniques. He shows that the task of combining the output of different retrieval systems can be modeled as the task of combining the output of multiple classifiers. He also presents different frameworks to combine multiple retrieval systems at different levels, e.g. at the representation level or at the output level. In our approach we use some of the score normalization algorithms presented by Croft. Our combination approaches are also inspired by this work, but we extend it by using machine learning to find optimal parameters of the combination.

The results of these different combination approaches show that evidence coming from different sources can be aggregated to achieve better performance of the overall retrieval system. In the context of our participation on CLEF this year, we investigate whether these techniques can also be used for the Ad-hoc task on the TEL datasets. Overall, we build on the system we presented at CLEF2008, which achieved a reasonable performance using concept based retrieval based on Explicit Semantic Analysis.

Our main contributions in this paper are the following ones:

- We extend both MT-based and concept-based retrieval into truly multi-lingual settings where not only the document collection can contain multiple languages but a document itself can contain fields in different languages. The main innovation is here that we maintain separate indices for each language and apply our combination strategies on the retrieval engines for each of these language-specific indices. Our results show that for the CLEF Ad-hoc TEL task we get a similar performance compared to a baseline system based on a single index, but no significant improvement over it.

- We also present an approach by which MT-based and concept-based retrieval (by ESA) can be combined through rank aggregation. This combination effectively increases the performance of the retrieval system for the bi-lingual task on the BL dataset using French topics and the ONB dataset using English and German topics.

The paper is structured as follows: In the Section 2 we describe our retrieval system and define MT-based retrieval, concept based retrieval as well as different aggregation approaches. In Section 3 we describe the used datasets and the preprocessing of the data. In Section 4 we present the experiments on the Ad-hoc TEL task using topics from CLEF2008, in Section 5 using topics from CLEF2009. We conclude in Section 6.

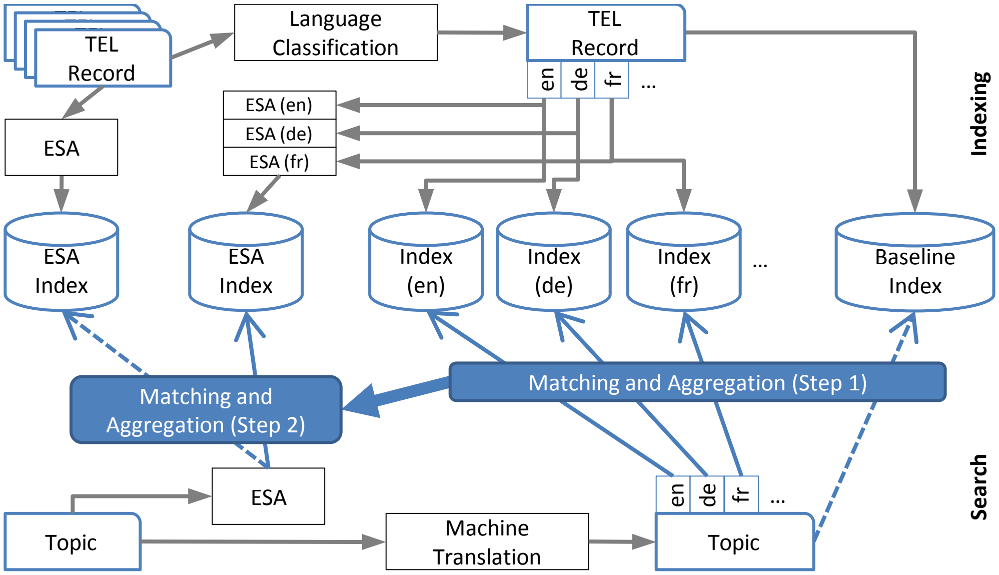

Figure 1: Figure of all used indices in our retrieval framework.

2 Approach

The main idea behind our approach is to use multiple indices (one for each language under consideration, which are all the common European languages). These are indices of fields of documents in different languages as well as concept indices of documents. The basic idea is to combine retrieval results based on the different indices. Figure 1 illustrates the different indices and processing steps which will be described in more detail in the following sections. But first we introduce some notation.

2.1 Notations

In the remaining article we use the following notations:

- $L = \{\alpha, \beta, \gamma, \ldots\}$: A set of languages.
- $D = \{d_1, \ldots, d_n\}$: A text corpus consisting of multi-lingual documents. The function $f_\alpha(d)$ selects all the document fragments of d in language α. $D_\alpha = \{f_\alpha(d_1), \ldots, f_\alpha(d_n)\}$ defines a restriction of corpus D where all document consist of their fragments in language α.
- $C = \{c_1, \ldots, c_m\}$: A set of concepts that define a concept space. Each concept has a textual description. We use c_i both to refer to concept c_i as well as to the description of c_i. The intended meaning will be clear from the context.
- $T_\alpha = \{t_{\alpha,1}, t_{\alpha,2}, \ldots\}$: A set of topics in language α that will be used to construct queries to the retrieval system. Each topic represents a certain information need. for the translation of a topic t_α to language β we will use the notation $t_{\alpha \rightarrow \beta}$.
- Statistics of a term w in document d of corpus D:
  - $\mathrm{TF}_d(w)$: Term frequency of w in document d.
  - $|d|$: Document length of d.
  - $\mathrm{DF}(w)$: Document frequency of w in corpus D.
  - $\mathrm{TF}(w)$: Term frequency of w in corpus D.
  - $n = |D|$: Number of documents.

- $|\hat{d}|$: Average document length in corpus D.

2.2 Language Detection

In our settings, the document corpus consists of multi-lingual documents which contain content in multiple languages. In our approach we assume that the parts of a document which are in different languages are identified and labeled appropriately. This is essentially the way how the f_α function described above is realized. This makes the application of language detection approaches necessary before indexing the documents (we will rely on different indices per language). In our settings the parts correspond to the fields of the documents in the TEL dataset which can be in different languages. In order to identify the language for each field, we exploit a language detection approach based on character n-grams models. The probability distributions for character sequences of the size n are used to classify text into a set of languages. We used a classifier provided by the Ling Pipe Identification Tool[1] which was trained on corpora in different languages as described in Section 3.

2.3 Machine Translation based CLIR

In the most simple case, the CLIR problem can be formulated as bilingual retrieval: given a topic t_β in language β and a set of multi-lingual documents D, find relevant documents in D_α. If all document fragments in D are of language α then $D = D_\alpha$, which is the most common scenario. In this case a MT system translating text from language β to α can be used to reduce the problem to mono-lingual retrieval by translating topic t_β to a topic $t_{\beta\to\alpha}$ in language α. Mono-lingual retrieval models can then be used to define the relevance of documents in D_α to the translated topic $t_{\beta\to\alpha}$.

In our approach we extend the bi-lingual setting to multiple languages. As shown in Figure 1 the first step is building indices for each language α that contain all terms of documents in D_α. This means that index I_α only contains information about text in language α. In the retrieval step, each topic is simultaneously translated into all languages and each translation of the topic is matched to the corresponding index. This results in a different ranking for each language. An overall ranking is computed through different aggregation approaches of these rankings which will be described in more detail in Section 2.5.

The matching of the translated topic to the language specific index is based on a mono-lingual retrieval model. In this paper we use models that have been implemented in the Terrier[2] framework.

For mono-lingual IR, we use the following retrieval models:

- DLH13

$$\mathit{Score}(t,d) := \sum_{w \in t} \mathrm{TF}_t(w) \frac{\mathrm{TF}_d(w) \log \frac{\mathrm{TF}_{d(w)}|\hat{d}||D|}{|d|\mathrm{TF}(w)} + .5 \log \left(2\pi \mathrm{TF}_d(w)(1 - \frac{\mathrm{TF}_{d(w)}}{|d|})\right)}{\mathrm{TF}_d(w) + .5}$$

- BB2

$$\begin{aligned}\mathit{Score}(t,d) \quad := \quad & \sum_{w \in t} \mathrm{TF}_t(w) \frac{\mathrm{TF}(w)+1}{\mathrm{DF}(w)(\mathrm{NTF}_d(w)+1)} \mathrm{TF}_t(w) \\ & (-\log(|D|-1) + \Phi(|D| + \mathrm{TF}(w) - 1, |D| + \mathrm{TF}(w) - \mathrm{NTF}_d(w) - 2) \\ & -\Phi(\mathrm{TF}(w), \mathrm{TF}(w) - \mathrm{NTF}_d(w)))\end{aligned}$$

with $\mathrm{NTF}_d(w) = \mathrm{TF}_d(w) \log \left(1 + \frac{|\hat{d}|}{|d|}\right)$ and $\Phi(n,m) := m + .5 \log \frac{n}{m} + (n-m) \log n$.

[1] http://alias-i.com/lingpipe/

[2] http://ir.dcs.gla.ac.uk/terrier/

- LemurTF_IDF

$$Score(t,d) := \sum_{w \in t} \mathrm{TF}_t(w) \frac{1.2\mathrm{TF}_d(w)}{\mathrm{TF}_d(w) + 1.2(.25 + .75\frac{|d|}{|\hat{d}|})} \left(\log \frac{|D|}{\mathrm{DF}(w)}\right)^2$$

2.4 Concept-based CLIR

As an instance of concept-based CLIR we build on the CL-ESA approach previously presented in [6]. For the sake of completeness we first discuss Explicit Semantic Analysis and then the cross-language extension CL-ESA.

In our retrieval system each document is mapped by ESA into a conceptual representation (the Wikipedia article space) which can be understood as an interlingua-based representation abstracting from languages which is inherently able to represent documents with fields in different languages. As shown in Figure 1 we follow two different approaches to build the index. One approach maps whole documents to the Wikipedia article space using ESA without considering that documents can contain different languages. The second approach classifies each field of a document into a corresponding language and then maps each field field into a concept vector using the a language-specific ESA instantiation.

We compare the performance of these approaches in our experiments. In both cases we rely on a single index for concept based retrieval, as the multiple languages are already considered in the concept mapping.

2.4.1 Explicit Semantic Analysis (ESA)

ESA classifies given document d with respect to a set of explicitly given external categories C. Gabrilovich and Markovitch [2] have outlined the general theory behind ESA and in particular described its instantiation to the case of using Wikipedia articles as external categories. We will basically build on this instantiation which we briefly summarize in the following.

In essence, Explicit Semantic Analysis takes as input a document d and maps it to a high-dimensional real-valued vector space. This vector space is spanned by a concept space $C_\alpha = \{c_1, \ldots, c_m\}$ in language α such that each dimension corresponds to concept c_i. This mapping is given by the following function: $\Phi_\alpha : D \rightarrow \mathbb{R}^{|C_\alpha|}$ with

$$\Phi_\alpha(d) := \langle \mathrm{AS}(d, c_1), \ldots, \mathrm{AS}(d, c_m) \rangle$$

The function AS expresses the *association strength* between d and the concept c_i. In the original ESA model AS is defined by sum of TF.IDF$_{c_i}$ values of all words of $w_j \in d$ based on the textual description of concept c_i. In previous work we examined the performance of different association strength functions for CLIR tasks [7]. Based on these result we use the following modified function:

$$\mathrm{AS}(d, c_i) := \sum_{w \in d} \frac{\mathrm{TF}_{c_i}(w)}{|c_i|} \log\left(\frac{|C|}{\mathrm{DF}(w)}\right)$$

2.4.2 Cross-lingual ESA (CL-ESA)

In this section we present the extension to ESA called CL-ESA (Cross-language Explicit Semantic Analysis). This is a relatively straightforward extension of ESA to a cross-lingual setting which we presented before in [6]. We will also describe how CL-ESA can be used for the semantic analysis for multi-lingual documents.

CL-ESA relies on the principle that concept vectors computed with respect to the Wikipedia database in one language can be translated into concept vectors with respect to another Wikipedia database relying on Wikipedia's language links[3]. This is done by mapping each dimension corresponding to article a in Wikipedia W_α to the dimension corresponding to article b in Wikipedia

[3] Cross-language links are those that link a certain article to a corresponding article in the Wikipedia database in another language.

W_β so that there exists a language link from a to b. This means that article a and b are textual descriptions of the same concept. Given this mapping it is for example possible to compare documents in language α and β based on the mapped concept vector.

In general the concept space that is used for CL-ESA needs textual descriptions of all concepts in all supported languages. We will refer to the description of concept c_i in language α by $c_{i,\alpha}$. For a multi-lingual document d CL-ESA is defined as follows:

$$\mathrm{AS}(d, c_i) := \sum_{\alpha \in L} \mathrm{AS}(d_\alpha, c_{i,\alpha}) \tag{1}$$

When CL-ESA is instantiated using the Wikipedia database, the articles have to be restricted to the articles having cross-language links to articles in all languages in L. Then all concepts represented by an article in any language have descriptions in all other languages given by the linked articles, which is needed for our model. In the following $m_{\alpha\to\beta} : W_\alpha \to W_\beta$ defines the function mapping articles from W_α according to language links to W_β.

Given a target language α for the concept representation of a multi-lingual document d with respect to Wikipedia $W_\alpha = \{a_1, a_2, \ldots\}$, the association strength defined in Equation 1 can be instantiated to Wikipedia by:

$$\mathrm{AS}_{W_\alpha}(d, a_i) := \sum_{\beta \in L} \mathrm{AS}(f_\beta(d), m_{\alpha\to\beta}(a_i))$$

Intuitively this is the association strength of a multi-lingual document d to a concept c represented by the Wikipedia article a_i in language α. This value is defined by the sum of the association strength of all fragments $f_\beta(d)$ in languages β to the concept description of c in language β. This description is given by the article in W_β to which a_i links to.

2.4.3 Retrieval using CL-ESA

Using the above defined association strength function, a mapping Φ of documents or topics to concept vectors can be defined as follows:

$$\Phi(d) := \vec{d} = \langle \mathrm{AS}(d, c_1), \ldots, \mathrm{AS}(d, c_m) \rangle$$

Given the vector representations of topics and documents, similarity measures in vector space can be used to determine the relevance of documents to topics. In our previous work we defined the following relevance function [7]:

$$rel(t, d) := \Gamma(\Pi(\Phi(q)), \Pi(\Phi(d))),$$

where Π is a projection function which reduces the dimensionality of the vector. This is done for performance issues as efficient indexing is not possible without the reduction. In our framework we use $\Pi^m_{abs}(\vec{d})$ which selects the m dimensions with highest values in $\vec{d}$, as this reduction function was shown to achieve good performance in CLIR tasks [7].

Γ defines the vector space similarity. We used the cosine similarity that is defined as

$$\Gamma_{\text{cosine}} = \frac{< \vec{t}, \vec{d} >}{\|\vec{t}\| \|\vec{d}\|}$$

2.5 Rank Aggregation

In our framework we aggregate two different kinds of rankings for a topic t. First, as we deal with multi-lingual documents and due to our separate language-specific indexing approach, for each language $\alpha \in L$ there is a ranking that expresses the relevance based on the text parts in language α. Second we compute a ranking based on the concept representation of topics and documents. In our framework we chose a two step rank aggregation approach. We first combine all text-based rankings and finally combine the resulting ranking with the concept-based ranking. In the following we describe different rank aggregation methods which we used for either the first or the second step of rank aggregation. More details will be presented in Section 4.

2.5.1 Linear Aggregation

As the first approach to aggregate different ranking scores we chose linear aggregation. This means that the final relevance score of a document is computed by the sum of all scores in the different rankings:

$$score(t, d) := \sum_{r \in R} \delta(r)\ score_r(t, d)$$

where R is a set of rankings and $\delta(r)$ a weighting function. In our experiments we use the following variations of this weighting function:

- **Normalization** using max score: $\delta(r) := 1/maxscore(r)$
  Before the aggregation, each ranking is normalized to values in $[0, 1]$. This is done by dividing each ranking score by the maximum score.

- **Normalization** using the number of retrieved documents: $\delta(r) := |r| / \sum_{r' \in R} |r'|$
  where $|r|$ is the number of retrieved documents of ranking r. This weight corresponds to the share of the number of retrieved documents for one ranking to the total number of retrieved documents for all rankings.

- **A priori weights** based on language: $\delta(r_\alpha) := P(\alpha)$
  This weighting function can applied to our first step of rank aggregation. In this case each ranking r_α is weighted by the apriori probability for a document to be in a certain language α. We use the share of text parts in language α in relation to all text parts in the corpus a apriori probability $P(\alpha)$.

2.5.2 Support Vector Machine Aggregation

As alternative approach to linear aggregation we considered rank aggregation based on Support Vector Machines (SVMs). For a given topic or document, a feature vector can be built by using the relevance score returned by each index. This is then used as input for a SVM classifier that predicts the relevance of the document on the basis of a combination of the ranking scores. This means that the results of each retrieval step on the different indices are used as feature values. The classification model is trained by using the relevance assessment available for the corpus. Each relevant document for a topic defines a positive training example, each non-relevant a negative one.

Using a linear kernel the model of the classifier corresponds to linear aggregation. By using non-linear kernels this can be extended to non-linear rank aggregation. In Section 4 we describe experiments with linear kernels and radial basis function kernels.

3 Evaluation

In this section we first introduce all datasets we used for our experiments. Then we describe the evaluation methodology and the evaluation measures. Finally we briefly present some details about our implementation.

3.1 Datasets

3.1.1 TEL Dataset

The TEL dataset was provided by the European Library in the context of the CLEF 2008/2009 ad-hoc track. This dataset consists of library catalog records of three libraries: the British Library (BL) with 1,000,100 records, the Austrian National Library (ONB) with 869,353 records and the Bibliotheque Nationale de France (BNF) with 1,000,100 records. While the BL contains a majority of English records, the ONB dataset of German records and the BNF dataset of French records, all collections also contain records in multiple languages. Each record consists of fields which again

Field	*Description*	*BL*	*ONB*	*BNF*
title	The title of the document	1	.95	1.05
subject	Keyword list of contained subjects	2.22	3.06	0.71
alternative	Alternative title	.11	.50	0
abstract	Abstract oft the document	.002	.004	0

Table 1: Average frequency of content fields of the TEL library catalog records.

BL			ONB			BNF		
Lang	*Tag*	*Det*	*Lang*	*Tag*	*Det*	*Lang*	*Tag*	*Det*
English	61.8%	76.7%	German	69.6%	80.9%	French	56.4%	77.6%
French	5.3%	4.0%	English	11.9%	8.0%	English	12.9%	8.2%
German	4.1%	2.9%	French	2.8%	2.1%	German	4.1%	3.8%
Spanish	3.1%	2.0%	Italian	1.8%	1.5%	Italian	2.3%	1.4%
Russian	2.7%	1.7%	Esperanto	1.5%	1.5%	Spanish	2.0%	1.4%

Table 2: Distribution of the 5 most frequent languages in each dataset, based on the language tags (Tag) and on the language detection model (Det).

may be of different languages. Not all of these fields describe the content of the record but contain also meta data such as the publisher name or year of publication.

As the CLEF topics are only targeted at the content fields, we first identified all content fields. Table 1 contains a list of the selected fields and the average count of each field for a record. Further we reduced additional noise by removing non-content terms like constant prefix or suffix terms from fields, e.g. the prefix term *Summary* in abstract fields.

In order to be able to use the library catalog records as multi-lingual documents as defined in Section 2 we also had to determine the language of each field. Our language detection approach is based on the language tags provided for 100.0% (BL), 89.916% (ONB), 81.64% (BNF) of all records as well as on the text-based language detection approach described in Section 2. Our analysis of the datasets showed that relying merely on the language tags introduces many errors in language assignment. First there are records tagged with the wrong language. Second, as there is only one tag per record, language detection based on tags is not adequate for records containing fields in different languages. Our language detection model determines the language for each field based on evidence from tags and from text based classification. Table 2 contains the language distribution in the TEL datasets based on the tags (Tag) as well as on our detection model (Det). A manual evaluation using a random selection of records showed that performance of the language detection approach on fields is reasonable.

3.1.2 Wikipedia Database

For concept-based retrieval we used the Wikipedia database in English, German and French as concept space. As we rely on bijective mappings between articles across languages for CL-ESA, we selected only those articles that are connected via cross-language links between all three Wikipedia databases. In this case every article is a concept having textual descriptions in English, German and French, namely the article text. Using the snapshot by 03/12/2008 for English, 06/25/2008 for French, and 06/29/2008 for German, we obtained the aligned collection of 166,484 articles in all three languages.

3.1.3 Training Corpora for Language Detection

The language detection framework requires sufficiently large corpora in all languages the classifier is trained for. We rely on the Leipzig Corpora Collection[4], which contains texts collected

[4] http://corpora.uni-leipzig.de

Test Size (characters)	1	2	4	8	16	32
Accuracy	22.59 %	34.82 %	58.55 %	81.17 %	92.45 %	97.33 %
Test Size (characters)	64	128	256	512	1024	2048
Accuracy	98.99 %	99.67 %	99.86 %	99.97 %	99.99 %	100 %

Table 3: Results of language detection using test data of different character sizes measured by classification accuracy.

from the web and newspapers, and the JRC-Acquis Multilingual Parallel Corpus[5], which contains documents published by the European Union translated in various languages.

3.2 Preprocessing

3.2.1 Language Detection

For language detection we used the n-gram language classifier included in the Ling Pipe software collection[6]. The classifier was trained using the Leipzig and JRC-Acquis corpora. When a certain language was available in both corpora we preferred the data of the Leipzig Corpus, as this showed better results in a cross validation on the training data.

We conducted multiple tests for verifying the effectiveness of the language detection model. The results showed that using a 5-gram model and a 100,000 character training is optimal in our case. Table 3 contains the classification results using different data sizes measured by the character size. The results show that the classifier achieves high performance of more than 97% accuracy for text containing more than 32 characters. As this is given for most fields in the TEL dataset this classifier is applicable for the language detection task in our framework.

3.2.2 Document Preprocessing

We used the following methods for the preprocessing of documents:

Tokenizer As tokenizer we used a standard white space tokenizer. All non-character tokens were deleted. For Wikipedia articles we also deleted all wiki markup.

Stop-Word Filtering We used standard stop word lists in the languages English, German, Finnish, French, Italian, Portugese, Swedish, which were taken from the University of Neuchatel[7], and Danish, Spanish, Dutch and Norwegian, which were taken form Ranks.nl[8].

Stemmer We used the Snowball Stemmers[9] to stem terms in English, German, French, Danish, Dutch, Finnish, Italian, Norwegian, Portugese and Swedish.

Fields in other languages than those mentioned above were not preprocessed using stemmers or stop word lists.

3.3 Evaluation Measures

The relevance assessments for the search task are provided by CLEF, resulting from a pooled manual evaluation. As evaluation measure we report mean average precision (MAP), precision at a cutoff level of 10 (P@10) and recall at a cutoff level of 100 (R@100).

[5] http://wt.jrc.it/lt/Acquis/

[6] http://alias-i.com/lingpipe/

[7] http://members.unine.ch/jacques.savoy/clef/

[8] http://www.ranks.nl/resources

[9] http://snowball.tartarus.org

3.4 Implementation

In our implementation we used different third party software tools as well as own implementations. For text based retrieval including inverted indexes and scoring models we used the Terrier IR framework. For translating the topics to various languages we used the machine translation service provided by Google[10]. We used our own implementation of CL-ESA for concept-based retrieval[11]. We also implemented an inverted concept index that allows efficient retrieval based on the concept representations of topics and documents. For example, for the ONB dataset the inverted concept index has the size of approx. 26 GB and the average processing time of a topic is approx. 135 seconds.

4 Experiments on CLEF08 Ad-hoc Topics

In this section we present the results of experiments using the CLEF08 Ad-hoc topics. As relevance assessments are available for these topics we used this task to optimize our system in respect to the retrieval model and the aggregation functions.

In all experiments we relied on the mono-lingual task, i.e. English topics for BL dataset, German topics for ONB and French topics for BNF. As all of these datasets contain documents in different languages, cross-lingual retrieval can be applied to find relevant documents in other languagesl. The mono-lingual task can therefore also be used to optimize the multi-lingual setting we propose in our framework.

4.1 Mono-lingual Retrieval Model

First we conducted experiments to optimize the retrieval models for MT based IR. As this is based on mono-lingual retrieval we compared the performance of different State-of-the-Art retrieval models. The hypothesis here was that good performance in mono-lingual retrieval should also result in good performance in cross-lingual retrieval.

We rely on the retrieval models provided by the Terrier framework in our work. We selected the best retrieval model for each dataset according to MAP and got the following best retrieval results on the different TEL datasets: MAP of .34 on the BL dataset using model DLH13, MAP of .22 on the ONB dataset using model LemurTF_IDF and MAP op .30 on the BNF dataset using model BB2. In the remainder of this paper we will report results relying on the best retrieval model for each dataset.

4.2 Rank Aggregation

As described above we defined two aggregation steps in our model. First the results of multiple text-based indexes are aggregated and afterwards the aggregated score is combined with concept-based retrieval score. In the following experiments we used again the CLEF2008 topics for the Ad-hoc mono-lingual task. The first aggregation step was evaluated on all three TEL datasets. For the evaluation of the second step we only performed experiments on the BL dataset.

4.2.1 Linear Aggregation for Multiple Indexes

The baseline for the proposed retrieval using multiple indexes is given by retrieval on a single index of all text in the documents without language classification. The performance of this baseline is shown in the first row of Table 4.

As described in Section 2 we used different normalization and weighting models for linear aggregation of the multiple indexes. Table 4 contains all results of aggregation without normalization, using max score and using the number of retrieved documents of each index and aggregation using a priori weights.

[10] http://translate.google.com

[11] http://code.google.com/p/research-esa

	BL			*ONB*			*BNF*		
Retrieval Method	*MAP*	*P@10*	*R@100*	*MAP*	*P@10*	*R@100*	*MAP*	*P@10*	*R@100*
Baseline (single index)	.34	.51	.50	.23	.36	.45	.30	.38	.56
Multiple Indexes (no norm.)	.25	.36	.45	.18	.26	.42	.22	.26	.48
Multiple Indexes (max score norm.)	.07	.08	.14	.08	.14	.22	.12	.16	.26
Multiple Indexes (num ret norm.)	.34	.51	.50	.22	.35	.42	.29	.35	.54
Multiple Indexes (a priori)	.34	.51	.50	.23	.36	.43	.30	.38	.54

Table 4: Results for MT-based retrieval on the CLEF08 mono-lingual task using a single index and using different rank aggregation methods for multiple indexes.

The results clearly show that our approaches to aggregate the results of the multiple indexes are not able to beat the baseline using a single index. Normalization based on the number of retrieved documents as well as a priori weights can both be used to achieve comparable performance in respect to MAP, P@10 and R@100. The results indicate that linear aggregation based on the multiple indexes seems not be able to improve the overall performance in this task.

As alternative approach to linear combination we experimented with Support Vector Machine based aggregation. To balance the ratio between the training data, we used all relevant documents for all topics as positive samples and randomly selected non-relevant documents as negative samples to achieve a ratio of positive/negative samples of 1/2.

As SVM implementation we used LIBSVM[12]. Using the SVM type C-SVC (c=1) with a radial basis function kernel, the training data could be classified using a 5-fold cross validation with precision of .61 and recall of .42. However when using the trained model for the actual retrieval the MAP was very low with .01. When using a linear kernel, which would lead to a classifier that is comparable to linear aggregation, we were not able to learn the model as the learning algorithm did not terminate. Our assumption is that using these kernel functions it is not possible to separate the positive and negative samples in the feature space. This would also explain the bad performance of the resulting retrieval system. It might be possible to use SVMs for rank aggregation by using other kernels, but in the scope of this paper we did not investigate that idea any further.

4.2.2 Linear Aggregation with Concept-Based Retrieval

In the technical report of last year, we presented results only based on concept-based retrieval using ESA [6]. In the current system we also investigate a modified version of the ESA-based mapping to the Wikipedia article space. The language classification step represents the TEL records as multi-lingual documents. This is used to map the documents fragments for each language to the concept space based on the Wikipedia databases in the corresponding languages. The concept vector representations of the different fragments are then combined to a single concept vector for each document as described in Section 2. Experiments on the CLEF08 mono-lingual task on the BL dataset showed an improvement of the new concept mapping model with respect to the model used in the last year experiments of 1% MAP, 7% P@10 and 5% R@100. For our experiments on the CLEF09 tasks we therefore used the new model.

In our final experiments using the CLEF08 topics we investigated the combination of MT based retrieval and concept-based retrieval. As for example suggested in [4] we also chose a linear aggregation function. The problem thereby is to find an optimal weight for each retrieval model. We approximated the optimal weight by a brute-force and systematic exploration of the parameter space. The results of this exploration for the BL dataset are presented in Figure 2. The left most bar represents MAP value giving full weight to the concept-based retrieval, while the right most bar represents the MAP giving full weight to the concept-based retrieval. The bars in between result from experiments using the combined approach with different weights. For the experiments using the CLEF2009 topics we used the best weightings derived from these experiments.

[12] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

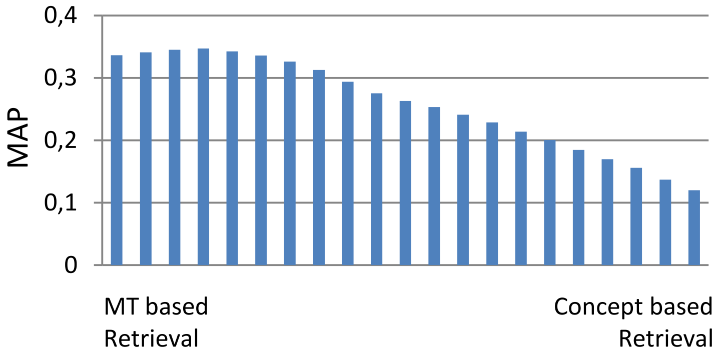

Figure 2: Results for the CLEF2008 mono-lingual ad-hoc task on the BL dataset using different weightings of MT-based retrieval and concept-based retrieval combined by linear aggregation. The left most result corresponds to MT-based retrieval, the right most to concept-based retrieval.

Topic		*BL*			*ONB*			*BNF*		
Lang.	*Retrieval Method*	*MAP*	*P@10*	*R@100*	*MAP*	*P@10*	*R@100*	*MAP*	*P@10*	*R@100*
en	Baseline (single index)	.35	.51	.55	.16	.26	.36	.25	.39	.45
	Multiple Indexes	.33	.50	.52	.15	.24	.35	.22	.34	.45
	Concept + Baseline	.35	.52	.54	.17*	.27	.37	.25	.39	.45
de	Baseline (single index)	.33	.49	.53	.23	.35	.47	.24	.35	.45
	Multiple Indexes	.31	.48	.51	.23	.34	.49	.22	.32	.43
	Concept + Baseline	.33	.49	.53	.24*	.35	.47	.24	.36	.45
fr	Baseline (single index)	.31	.48	.50	.15	.22	.31	.27	.38	.51
	Multiple Indexes	.29	.45	.47	.14	.20	.32	.25	.35	.50
	Concept + Baseline	.32	.51*	.50	.15	.22	.31	.27	.37	.50

Table 5: Results on the CLEF 2009 Ad-Hoc Task. Statistical relevant improvements according to a paired t-test with confidence level .05 are marked with *.

5 Experiments on CLEF09 Ad-hoc Topics

The CLEF09 Ad-hoc topics are similar to the topics from CLEF08. The 50 topics have the same format consisting of two fields, a short title containing 2-4 keywords and a description of the information item of interest in terms of 1-2 sentences. The objective is to query the selected target collection using topics in the same language (mono-lingual run) or topics in a different language (bi-lingual run) and to submit the results in a ranked list ordered with respect to decreasing relevance. In line with these objectives we submitted results of six different runs to CLEF08. These are the results of querying English, German and French topics to the BL, ONB and BNF datasets.

The results of our experiments are presented in Table 5. The results using multiple indexes show that this approach was not able to beat the baseline. Using a single index for the TEL records without language classification and topics only translated into the main language of each dataset achieved better performance compared to our approach based on indexes for each language and multiple translations of the topic to the matching languages. Another result is that the combination of concept-based retrieval to the MT-based retrieval was able to improve the retrieval in some cases. The improvement was significant according to a paired t-test with confidence level .05 for French topics on the BL dataset and English and German topics on the ONB dataset. However in many cases the performance was similar to the baseline without statistical significance of the difference. We could therefore not reproduce the strong improvements e.g. presented in [4].

6 Conclusion

In this paper we have presented a cross-language information retrieval approach based on multiple indexes for different languages and rank aggregation to combine the different partial results. The approach was developed in the light of the fact that the CLEF TEL dataset consists of records in different languages which also may contain fragments of more than one language. For this approach a language detection of all documents fragments of the dataset as well as translation of topics to all supported languages is necessary. Our results showed that for the CLEF08 and CLEF09 Ad-hoc task we were not able to improve retrieval result with this new model. The baseline consisting of a single index without language classification and a topic translated only to the index language achieved similar or even better results.

We also combined Machine Translation based retrieval with concept-based retrieval. The results showed that we were able to improve the baseline through the combination in some cases. However the improvement on the CLEF Ad-hoc task were not as strong as reported on other experiments in related work.

Acknowledgments

This work was funded by the Multipla project sponsored by the German Research Foundation (DFG) under grant number 38457858.

References

[1] W. Bruce Croft. Combining approaches to information retrieval. In *Advances in Information Retrieval*, pages 1–36. 2000.

[2] E. Gabrilovich and S. Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1606–1611, 2007.

[3] J. Kürsten, T. Wilhelm, and M. Eibl. CLEF 2008 Ad-Hoc Track: On-line Processing Experiments with Xtrieval. In *Working Notes of the Annual CLEF Meeting*, 2008.

[4] C. Müller and I. Gurevych. Using Wikipedia and Wiktionary in Domain-Specific Information Retrieval. In *Working Notes of the Annual CLEF Meeting*, 2008.

[5] Gerard Salton and Michael J. McGill. *Introduction to Modern Information Retrieval.* McGraw-Hill, Inc., 1986.

[6] P. Sorg and P. Cimiano. Cross-lingual Information Retrieval with Explicit Semantic Analysis. In *Working Notes of the Annual CLEF Meeting*, 2008.

[7] Philipp Sorg and Philipp Cimiano. An experimental comparison of explicit semantic analysis implementations for cross-language retrieval. In *Proceedings of the International Conference on Applications of Natural Language to Information Systems (NLDB)*, Saarbrücken, 2009.