View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Publications at Bielefeld University

Cross-lingual Information Retrieval with Explicit Semantic Analysis

Philipp Sorg, Philipp Cimiano
Institute AIFB, University of Karlsruhe
{sorg,cimiano}@aifb.uni-karlsruhe.de

Abstract

We have participated on the monolingual and bilingual CLEF Ad-Hoc Retrieval Tasks, using a novel extension of the by now well-known Explicit Semantic Analysis (ESA) approach. We call this extension Cross-Language Explicit Semantic Analysis (CL-ESA) as it allows to apply ESA in a cross-lingual information retrieval setting. In essence, ESA represents documents as vectors in the space of Wikipedia articles, using the tfidf measure to capture how "important" a Wikipedia article is for a specific word. The interesting property of ESA is that arbitrary documents can be represented as a vector with respect to the Wikipedia article space. ESA thus replaces the standard BOW model for retrieval. In our cross-lingual extension of ESA, the cross-language links of Wikipedia are used in order to map the ESA vectors between different languages, thus allowing retrieval across languages. Our results are far behind the ones of other systems on the monolingual and ad-hoc retrieval tasks, but our motivation was to find out the potential of the CL-ESA approach using a first and unoptimized implementation thereof.

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, Wikipedia

1 Introduction

Cross-language Information Retrieval (CLIR) can be described at an abstract level as the task of retrieving documents across languages. In some sense, the CLIR task represents one extreme case of the so called *vocabulary mismatch problem*, i.e. the problem that the vocabulary of a user query and the vocabulary of relevant documents can differ substantially. The bag-of-words (BOW) model notoriously suffers from the vocabulary mismatch problem as the different dimensions are inherently orthogonal, thus neglecting relations between different words in the same language as well as across languages. Therefore, the challenging task of retrieving documents to queries in other languages requires models going beyond the traditional bag-of-words model.

When tackling the task of retrieving documents across languages, there seem to be essentially two main paradigms:

1. **Translation-based** approaches which rely either on a translation of documents or queries. For the translation of queries, one typically relies on bilingual dictionaries (compare [10], [5]).

2. Mapping of queries and documents into a **multilingual space** in which similarity between queries and documents can be computed uniformly across languages.

The first type of approaches is obviously highly dependent on the quality of the translation system used or the bilingual dictionary in question. Demner-Fusham et al. [5] have in particular shown that the coverage of the bilingual dictionary has a crucial impact on the retrieval task. As mentioned by Demner-Fushman et al., for a successful dictionary-based CLIR model, the following three steps need to be accomplished: (1) selection of the terms to be translated, (2) generation of a set of candidate translations, and (3) use of that set of candidate translations in the retrieval process.

Concerning the second type of approaches in which queries and documents are mapped into a multilingual space, there are two crucially different models:

- **latent model**: Instead of representing documents (and queries) with respect to the bag-of-word dimensions, some approaches compute "latent" concepts from the data and index documents with respect to these latent concepts. Latent concepts correspond to certain topics emerging bottom-up from the document collection. The most prominent technique here is latent semantic analysis (LSA) [4]. In fact, LSA has also been applied in cross-lingual IR settings (compare [17]). For this purposes, parallel texts are needed across languages in order to construct a matrix where the dimensions correspond to words in all languages considered. Dimensionality reduction is then applied to discover correlated words across languages. Queries and documents can then be represented in this "latent space" and retrieval can be performed in a standard fashion by calculating the cosine in this space.

- **external category model**: In contrast to retrieval models which build on latent topics or concepts, one can also choose a set of external categories, topics or concepts to define the dimensions of the vectors. These can be categories from existing thesauri, ontologies etc. The advantage is that the vectors then remain constant across different document collections, in particular also across languages. Such models presuppose that we are indeed able to index texts in various languages with respect to the multilingual space spanned by the external categories.

The latter approach based on indexing with respect to external categories is interesting in the sense that i) no parallel texts are required (e.g. in order to compute latent topics grouping words from different languages), and ii) no bilingual dictionaries are needed. Obviously, this is true only to some extent as the mapping into the external categories (across languages) might well require cross-lingual dictionaries. Gabrilovich and Markovitch [7] have for example recently presented an interesting approach in which Wikipedia articles are used as dimensions of the vectors, i.e. documents are indexed with respect to the Wikipedia article space. While Gabrilovich and Markovitch have applied this model to calculate semantic relatedness between words, this model extends straightforwardly to an IR setting, in which query and documents are mapped to a vector representing the Wikipedia article space (see for instance [9] and [6]).

An interesting characteristic of Wikipedia is that articles are linked across languages by bidirectional language links. Thus, we can in principle translate a document or query vector indexed with respect to the Wikipedia of language L_i to language L_j, thus extending straightforwardly into a cross-lingual retrieval task.

In this paper we investigate this idea closer and present an approach for cross-language IR based on Explicit Semantic Analysis. In particular, we present our system as it has been used on the CLEF monolingual and multilingual Ad-Hoc retrieval tasks. Further, we also present additional experiments on the Multext dataset conducted after the submission to the CLEF campaign in order to verify some of the parameter settings of our approach on another dataset. In order to be able to quantify the influence of the parameters, we have in particular conducted standard mating experiments on the Multext dataset.

The article is structured as follows: in the next section 2 we describe in more detail the ESA model and show how it can be used in a retrieval setting. In Section 3 we discuss how this model can be extended to a cross-lingual setting relying on the Wikipedia cross-language links. In section 4 we discuss some implementation details which are nevertheless important to understand how the overall system works on the task of cross-language IR. Finally, in Section 5 we present our results on the CLEF datasets as well as on the Multext corpus.

2 Explicit Semantic Analysis (ESA)

Explicit Semantic Analysis (ESA) [7] attempts to index or classify a given text t with respect to a set of explicitly given external categories. It is in this sense that ESA is explicit compared to approaches which aim at representing texts with respect to latent topics or concepts, as done in Latent Semantic Analysis (LSA) (see [4], [11]). Gabrilovich and Markovitch 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 as described in [7] which we briefly summarize in the following.

In essence, Explicit Semantic Analysis takes as input a text t and maps it to a high-dimensional real-valued vector space. This vector space is spanned by a Wikipedia database $W_k = \{a_1, \ldots, a_n\}$ in language L_k such that each dimension corresponds to an article a_i. This mapping is given by the following function:

$$\Phi_k : T \rightarrow \mathbb{R}^{|W_k|}$$

$$\Phi_k(t) := \langle v_1, \ldots, v_{|W_k|} \rangle$$

where $|W_k|$ is the number of articles in Wikipedia W_k corresponding to language L_k. The value v_i in the ESA vector of t expresses the *strength of association* between t and the Wikipedia article a_i. Based on a function as that defines the strength of association between words and Wikipedia articles, the values v_i can be computed as the sum of the association strength of all words of $t = \langle w_1, \ldots, w_s \rangle$ to the article a_i:

$$v_i := \sum_{w_j \in t} as(w_j, a_i)$$

One approach to define such a association strength function as is to use a tf.idf function based on the Bag-of-Words (BOW) model of the Wikipedia articles. The association strength of word w_j to article a_i is then equal to the tf.idf value of w_j in a_i:

$$as(w_j, a_i) = tf.idf_{a_i}(w_j)$$

In the literature, many different definitions of tf.idf functions based on the BOW model have been proposed (see [1]). The particular function that was used in our experiments is described in Section 4.

Essentially, for each article a_i in Wikipedia, ESA sums up all the association strengths of each word w_j appearing in the document. In this sense, the Semantic Interpreter applying ESA described in [7] essentially computes the function Φ. As output we thus get a vector representing the strength of association of our text t with respect to the articles in Wikipedia W_k. Actually, this vector thus corresponds to a ranking of the Wikipedia articles according to importance or relevance for a text t.

Given the ESA framework, we can assess the similarity between two texts $t_i, t_j \in T$, between a query q and a text t_i etc. For example, the standard cosine measure can be used to compare the vectors. In the remainder of this paper we will simply assume that the cosine is used to compare different vectors.

In fact, this framework is flexible to be applied to a variety of tasks, computing the similarity between:

- single words, which can be seen as singleton texts consisting of only one word. This can then be used to compute semantic relatedness between words as in [7]. Gabrilovich and Markovitch actually showed that their method performs better than LSI on the task of computing semantic relatedness between words.
- two documents (e.g. in a clustering task)

- a query and a document (e.g. in a retrieval. task)

In this paper we are concerned with a retrieval task, in which we are given a query q and need to rank the documents according to relevance. It should be clear from the above discussions that ESA straightforwardly extends to a retrieval scenario.

As a running example in this paper, we will use query `10.2452/460-AH` ("Scary Movies") from the 2008 CLEF Ad-hoc retrieval dataset where our system performed remarkably well. In the following table we indicate the 10 top-ranked Wikipedia articles for the query in the three languages German, English and French:

Language	English	German	French
Query	Scary Movies	Horrorfilme	Les films d'épouvante
Top 10 Wikipedia articles			
1	Scary Movie	Horror	La Plus Longue Nuit du diable
2	Horror	Audition	Barbara Steele
3	Scary Movie 3	Dark Water	Danger planetaire
4	Kazuo Umezu	Candyman	James Wan
5	James L. Venable	Prophezeiung (1979)	Dracula, mort et heureux de l'être
6	Horror and terror	Wolfen (Horrorfilm)	Seizure
7	Regina Hall	Alienkiller	Danvers (Massachusetts)
8	Little Shop of Horrors	Brotherhood of Blood	Fog (film,1980)
9	The Amityville Horror (1979 film)	Lionel Atwill	The Grudge
10	Dimension Films	Doctor X	La Revanche de Freddy

The top-10 ranked articles clearly differ between the languages. It is in particular interesting to observe that many results are actually named entities which clearly differ between languages due to a different cultural background. Consequently, the ESA vectors for the same query in different languages varies substantially, which is less optimal in a cross-language retrieval setting.

In the following section, we present our own extension to ESA called CL-ESA (Cross-language Explicit Semantic Analysis)[1], which represents a relatively straightforward extension of ESA to a cross-lingual setting. Our main aim in this paper is to discover if CL-ESA performs well in a cross-lingual retrieval setting.

3 Cross-lingual ESA (CL-ESA)

A very interesting characteristic of Wikipedia, besides the overwhelming amount of information created dynamically and in a collaborative way, is the fact that articles are linked across languages. Cross-language links are those that link a certain article to a corresponding article in the Wikipedia database in another language. A previous analysis of this cross-lingual link structure between the German and English Wikipedia showed that 95% of these links are indeed bi-directional (see [16]). The analysis of French-English and French-German links showed similar results. In the following we therefore assume the existence of a mapping function $m_{i \to j}$ that maps an article of Wikipedia W_i to its corresponding article in Wikipedia W_j.

In fact, given a text $t \in T$ in language L_i, it turns out that we can simply index this document with respect to any of the other languages $L_1, .., L_n$ we consider by transforming the vector $\Phi_i(t)$ into a corresponding vector in the vector space that is spanned by the articles of Wikipedia in the target language. Thus, given that we consider n languages, we have n^2 mapping functions of the type:

$$\Psi_{i \to j} : \mathbb{R}^{|W_i|} \to \mathbb{R}^{|W_j|}$$

This mapping is calculated as follows:

$$\Psi_{i \to j} \langle v_1, ..., v_{|W_i|} \rangle = \langle v'_1, ..., v'_{|W_j|} \rangle$$

[1] We would like to point out that we have developed and called our model CL-ESA independently of the CL-ESA approach described by Potthast et al. [13]. We discovered this work just after finishing our paper, so that CL-ESA is introduced here as a novel paradigm while it clearly has the CL-ESA approach of Potthast et al. as precedent. We thank the Web Technology & Information Systems Group of Weimar University (in particular Martin Potthast) for bearing with us in spite of missing their work in the first place and for the exchange with respect to technical details related to the implementation of the ESA approach.

where

$$v'_p = \sum_{q \in \{q^* | m_{i \to j}(a_{q^*}) = a_p\}} v_q \quad (1)$$

with $1 \leq p \leq |W_i|$, $1 \leq q \leq |W_j|$. In case that $i = j$ we thus have the identity function.

In order to get the ESA representation of a document $t \in T$ in language L_i with respect to Wikipedia W_j we simply have to compute the function $\Psi_{i \to j}(\Phi_i(t))$.

In the following table, we give the top-ranked Wikipedia articles for our running example query together with the result of mapping the German and French vectors into the English Wikipedia space:

Language	English	German → English	French → English
Query	Scary Movies	Horrorfilme	Les films d'épouvante
Top 10 Wikipedia articles			
1	Scary Movie	Horror	The Grudge
2	Horror	Audition (disambiguation)	The Devils Nightmare
3	Scary Movie 3	Dark Water	Barbara Steele
4	Kazuo Umezu	Candyman	The Blob
5	James L. Venable	Splatter film	James Wan
6	Horror and terror	Prophecy (film)	Dead and Loving It
7	Regina Hall	Wolfen (film)	Seizure (film)
8	Little Shop of Horrors	The Borrower	Danvers, Massachusetts
9	The Amityville Horror (1979 film)	Brotherhood of Blood	The Fog
10	Dimension Films	Lionel Atwill	A Nightmare on Elm Street 2: Freddy's Revenge

To illustrate the actual overlap of the ESA vectors, the next table contains the positions of the first 10 matches of the i) English ESA vector using the query of the running example and ii) the German ESA vector mapped to the English ESA space. In this case, matches are common non-zero dimensions in the ESA vector.

	Position in ranked ESA vector	
Article	*English*	*German → English*
Scary Movie	1	555
Horror	2	1
Scary Movie 3	3	288
Scary Movie 2	4	619
The Amityville Horror (1979 film)	10	262
Scary Movie 4	12	332
Horror film	15	15
Horrorpunk	16	353
Jon Abrahams	23	235
Poltergeist (film series)	29	542

The positions of these matches show that the English vector and the mapped German vector have common non-zero dimensions, but the rank of these dimensions differs a lot. In an ideal setting these ranks should be equal in both vectors.

Given the above settings, it should be straightforward to see how the actual retrieval works. The cosine between a query q_i in language L_i and a document d_j in language L_j is calculated as:

$$cos(q_i, d_j) := cos(\Phi_i(q_i), \Psi_{j \to i}(\Phi_j(d_j)))$$

This thus gives us an elegant retrieval model which is uniform across languages. A prerequisite for this model is certainly that we know the language of the query and of the different documents in order to know which mapping Ψ should be applied. We describe in the implementation section how we actually implemented a straightforward component for language detection.

4 Implementation

In this section we describe the implementation details we used for our experiments. In particular, we describe i) the document preprocessing (Section 4.1), ii) the actual ESA implementation that consists of

article preprocessing, ESA vector computation and multi-lingual mapping (Section 4.2), iii) the identification method to identify the language of a document (Section 4.3), and iv) the overall retrieval process (Section 4.4).

4.1 Preprocessing of Documents

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.

Stop-Word Filtering We used standard stop word lists in the languages English, German and French to filter out stop words.

Stemmer All terms in the documents were stemmed using Snowball Stemmers [2] available for many different languages including English, German and French.

4.2 ESA Implementation

The implementation of Cross-Lingual ESA can be divided into three steps. The first step is the preprocessing of the Wikipedia articles. This includes preprocessing of the article texts as well as the selection of articles that will be used for ESA indexing. The next step is the computation of the ESA vector, which depends on the choice of the *association strenght (as)* function that assigns the strength of association between words of the documents and Wikipedia articles. The last step is the multi-lingual mapping of the ESA vector.

In the following, the implementation of all of these steps including different variations and parameters will be explained in detail.

4.2.1 Wikipedia Article Preprocessing

The processing of the Wikipedia articles was done by using the Wikipedia tokenizer that is included in the Lucene[3] software package and then using the same methods for stop word removal and stemming as in the preprocessing of the documents. The Wikipedia tokenizer removes all Wiki markup from the text, e.g. syntax for links, headings and font styles.

The selection of articles that were used as dimensions of the ESA vector was based on different criteria. First we filtered out all redirect articles and all category articles. Then all articles with less than 100 words or less than 5 incoming pagelinks were discarded. In our first experiments, we did not perform any further selection. The results of the CLEF ad hoc retrieval are based on these settings. In the subsequent experiments on the Multext dataset, we restrict the Wikipedia articles used for ESA indexing to those that have at least a language link to one of the two other languages we consider. For example, we only consider an article of the English Wikipedia if it has a cross-language link to the German or the French Wikipedia. In absolute numbers, we used $536,896$ English, $390,027$ German and $362,972$ French articles for the ESA indexing (Wikipedia snapshot of March 12, 2008).

In the original ESA approach, Gabrilovich and Markovitch included more preprocessing and selection steps [8]. They added to the text for example the anchor text of incoming pagelinks and titles of redirects to an article. Some articles such as articles about years and similar were discarded. We have not made use of any additional similar heuristics in our implementation of the ESA/CL-ESA approach. Nevertheless, it would be interesting to study the influence of such additional heuristics in the future.

4.2.2 ESA Vector Computation

The computation of the ESA vector is based on an inverted index of the preprocessed selected Wikipedia articles. Each document of the dataset can then be treated as a query to this index. The retrieved articles with their weight can then be used to build the ESA vector.

[2] `http://snowball.tartarus.org`
[3] `http://lucene.apache.org`

The implementation of the index was done by using Lucene. As the function for computing the association strength between documents and articles, we used a customized implementation of the Lucene similarity function which computes the following function for a text $t = \langle w_1, \ldots, w_l \rangle$ and a Wikipedia article a_i of Wikipedia database W:

$$as_R(t, a_i) = (C_t)\sqrt{|a_i|}^{-1} \sum_{w_j \in t} tf_{a_i}(w_j) idf(w_j)$$

with

$$C_t = \frac{1}{\sqrt{\sum_{w_j \in t} idf(w_j)}}$$

$$tf_{a_i}(w_i) = \sqrt{\text{number of occurrences of } w_i \text{ in } a_i}$$

$$idf(w_j) = 1 + \log \frac{\text{number of articles containing } w_j}{|W| + 1}$$

The choice of as_R is motivated by the good performance on IR tasks. We therefore assume that this association strength can be used for the computation of the values of the ESA vector. The factor $\sqrt{|a_i|}^{-1}$ constitutes a normalization by length of the article. The factor $C(t)$ is only dependant on the query and does therefore not affect the relevance ranking of articles to the text t or the cosine computation.

In the experiments on the Multext dataset, we also used a different function that computes a bit valued ESA vector. This function as_{BIT} is defined as follows:

$$as_{BIT}(t, a_i) = \begin{cases} 1 & a_i \text{ contains any } w_j \in t \\ 0 & else \end{cases}$$

For both functions, the number k of articles (dimensions) considered in order to compute the ESA vector is used as a parameter. In fact, it seems that for the computation of the ESA vector *"less is more"* as conveyed by the experiments described in [6]. However, this is only the case provided that we have a reasonable way of determining which articles are most suitable. In our approach we only set those values in the ESA vector corresponding to the k articles with the highest association strength to a document t. Thus, the vectors we consider are relatively sparse with $|W| - k$ dimensions having zero values.

When using as_{BIT} to compute the ESA vector, the ranking of relevant articles for a text is still based on as_R. As this ranking is used to select k articles, as_{BIT} is not independent from as_R. The objective of using as_{BIT} however is to flatten the differences between the associated Wikipedia articles in the ESA vector.

Gabrilovich and Markovitch weighted the association strength by exploiting the pagelink structure of Wikipedia. It remains future work to adapt this method to our implementation.

4.2.3 Multi-lingual Mapping

As described above the multi-lingual mapping was done by using the cross-language links of Wikipedia. To use these links in an efficient way, some preprocessing is necessary. First we did a normalization of the target page titles of all cross-language links, as this is not done automatically in the Wikipedia database. Then we identified all cross-language links pointing to redirect pages and replaced them with language links to the article to which the redirect was leading.

In order to map the vectors from language L_i to language L_j we only use the cross-language links of Wikipedia W_i pointing to W_j. As our statistics showed that most of these links are bi-directional (95%) we did not include the links from W_j to W_i.

In some cases, two or more articles in W_i contain a cross-language link to the same article in $a \in W_j$. In this case, the new value of the ESA dimension corresponding to a was set to the sum of the values of all dimensions that correspond to the source articles in the original ESA vector (see Equation 1).

```
ESA-RETRIEVAL(Topics T, Language k, Documents D)
 1  for t ∈ T
 2  do
 3      t⃗ = Φ_k(t);
 4
 5  for d ∈ D
 6  do
 7      l := lang(d);
 8      d⃗ = Ψ_{l→k}(Φ_l(d))
 9      for t ∈ T
10      do score[t, d] = cos(t⃗, d⃗);
11
```

Figure 1: Pseudocode describing the retrieval algorithm

4.3 Language identification

In order to be able to compute the ESA vector for a document, the language of this document must be known as the computation is based on an index of a Wikipedia database in the document's language. Many document collections only contain documents in one language and thus no language identification is needed. In other cases, such as in the CLEF ad hoc retrieval task, the dataset contains documents in different languages.

In our implementation we first try to determine the language by using properties of the documents such as language annotations. If these are not available, we apply a simple heuristic to determine the language of document t as follows:

$$lang(t) := max_{L_k \in \{L_1,\ldots,L_n\}} \frac{minDim(\Phi_k(t))}{maxDim(\Phi_k(t))}$$

where $minDim(\vec{v})$ returns the value of the lowest dimension in vector $\vec{v}$ and $maxDim(\vec{v})$ returns the highest correspondingly. The intuition behind this heuristic is that a small difference between the values of the lowest and highest dimension, which is computed by the share of these values, means that the document matches good to many Wikipedia articles and it can therefore be assumed that the document is of the same language as the used Wikipedia articles. Comparing a document to Wikipedia articles in another language, there will be some mathes but the value of lowest dimension will most probably be very small.

While we have not done an extensive evaluation of this heuristic, a check showed that the quality of this heuristic is reasonable and sufficient for our purposes.

4.4 Retrieval

The implementation of the multi-lingual retrieval task is described in Figure 1 using pseudo code. In summary we first compute the ESA vector of all topics and then iterate over all documents in the dataset. The described workflow reduces the number of ESA vector computations substantially.

For the CLEF ad hoc retrieval task we were able to process the ONB dataset using all English, German and French topics in about 40 hours. The same task on the BL dataset had a runtime of approximately 60 hours.

5 Evaluation

In this section, we describe the datasets used for the evaluation. Then we present the experiments together with the different parameter settings applied. Finally, we also analyze the results of our approach with respect to different parameters using alternative measures such as the overlap of retrieved documents for the same query in different languages.

5.1 Datasets

The first dataset we used was the TEL dataset that was provided by The European Library in the context of the CLEF 2008 ad-hoc track. This dataset consists of library catalog records mainly in English, German and French but also some records in other languages. In our experiments, we used two parts of this dataset: The TEL English data provided by the British Library with mainly English records and the TEL German data provided by the Austrian National Library with mainly German records. All of these records consist of content information together with meta information about the publication. The title of the record is the only content information that is available for all records. Some records additionally contain some annotation terms. In our experiments we only used the available content information.

This dataset is challenging for IR tasks in different ways. First the text of the records is very short, only a few words for most records. Second, the dataset consists of records in different languages and retrieval methods need to consider relevant documents in all of these languages. The following examples show the complete content information of some records of the TEL English dataset:

Title or Subject	*Annotation Terms*
Strength, fracture and complexity : an international journal.	Fracture mechanics, Strength of materials
Studies in the anthropology of North American indians series.	-
Lehrbuch des Schachspiels und Einfuehrung in die Problemkunst.	Chess

The TEL English dataset contains 1,000,100 records, the TEL German dataset 869,353.

As second dataset we used the Multext JOC corpus[4]. The original data of this corpus is composed of written questions asked by members of the European Parliament on a wide variety of topics and corresponding answers from the European Commission in 9 parallel versions, published as one section of the C Series of the Official Journal of the European Community of the year 1993. The parts corresponding to the languages of the Multext project (English, French, German, Italian and Spanish) were collected and prepared in collaboration with the MLCC project. For our experiments we used the English, German and French parts. This dataset contains 3126 question/answer pairs in each language which are aligned across the languages.

5.2 CLEF Ad-hoc Experiments

The CLEF ad-hoc TEL task was divided into mono-lingual and bi-lingual tasks. 50 topics in the main languages English, German and French were provided. The topics consist 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 list ranked with respect to decreasing relevance. In line with these objectives we submitted results of six different runs to CLEF 2008. These are the results of querying English, German and French topics to the TEL English dataset and English, German and French topics to the TEL German dataset.

The following parameter settings as described in the implementation section were used for these experiments:

ESA vector length We used different lengths of the ESA vector to represent topics and records. For the topics we used $k = 10,000$, that means that 10,000 Wikipedia articles with the strongest association to a specific topic were used to build the ESA vector for this topic. For the records, we used $k = 1000$. The difference between the lengths is mainly due to performance issues. We were only able to process the huge amount of records by limiting the length of the ESA vectors for records to 1000 non-zero entries. As only 50 topics were provided, we were able to use more entries for the ESA vectors for topics. Our intention thereby was to improve recall of the retrieval by using more ESA dimensions.

Article selection In the results of the experiments submitted to CLEF, we only used the default article selection as described in the implementation section. One problem of this setting is the loss of many dimensions in the mapping process, as not all of the articles corresponding to a non-zero ESA vector

[4] http://aune.lpl.univ-aix.fr/projects/multext/

entry have a corresponding cross-language link to the Wikipedia in the target language. In this case, the information about this dimension is lost in the mapping process.

The following table contains the CLEF 2008 results of our submitted experiments measured by the Mean Average Precision (MAP) quality measure:

Dataset	*Topic language*	*MAP*
TEL English (BL)	English	17.7%
	German	7.6%
	French	3.6%
TEL German (ONB)	English	6.7%
	German	9.6%
	French	5.1%

In addition to the submitted experiments we also conducted additional experiments on the TEL dataset to better quantify and understand the impact of certain parameters on the result quality. As we were not able to evaluate the results apart from the submitted ones, we decided to examine the result overlap for queries in different languages on the same dataset. This measure can be seen as a quality measure for the capability of retrieving relevant documents across languages. Ideally, queries in different languages should result in the same set of retrieved records. We computed the result overlap for two different settings. First we used the same settings as used in the submitted results. For the second set of experiments we further restricted the Wikipedia articles that were used for ESA indexing to articles with at least one language link to one of the two other languages considered. The following table contains the result overlaps for topic pairs in different languages on the TEL English dataset:

Article restriction	*Topic language pair*	*Average result overlap*
No restriction	English - German	21%
	English - French	19%
	German - French	28%
Articles with exiting cross-language link	English - German	39%
	English - French	51%
	German - French	39%

The results show that we were able to substantially improve the retrieval methods according to the results overlap measure by restricting the Wikipedia articles. Our assumption is that the results on the retrieval task would also improve, but we did not manage to submit an additional run on time for CLEF.

5.3 Mate Retrieval on Multext JOC Corpus

As described above, the part of the Multext JOC Corpus we used consists of 3126 question/answer pairs in English, German and French. All of these documents are aligned across languages in the sense that for all documents there exist a corresponding article in the other languages. This dataset can therefore be used for mate-retrieval experiments, which allow a direct assessment of different parameters. Mate retrieval is the task of using a document as query with the objective to identify its translated counterpart in a set of documents in another language. In this case the counterpart is known in advance enabling an automatic evaluation of the mate retrieval results.

Our main goal of the mate retrieval experiments was to optimize the parameters settings for CL-ESA. We ran the experiments for various parameter settings:

ESA vector length We used different k for the maximal number of non-zero dimensions of the ESA vector, namely $k \in \{1000; 10,000; 100,000\}$.

Article selection We only used articles with existing cross-language links for the ESA vector computation as described in the implementation section.

Text selection We used different text parts of the question/answer pairs in our experiments, namely subject, question and all text consisting of subject, question and response. We always compared identic parts of queries and documents, e.g. if we used the subject as query we only matched it to the subjects of the documents in the retrieval process.

Real vs. Bit vectors In the experiments we examined the effect of using real valued ESA vectors versus bit valued ESA vectors.

As evaluation measure we used TOP-1 and TOP-10 Precision, that is the share of input document for which the mate was retrieved on position 1 or among the 10 best ranked results. The results for different text selection, ESA vector model and ESA vector lengths are presented in the following table:

The results presented in the following table are retrieval results using German queries on English documents:

			Precision	
Text	*Vector model*	*k*	*TOP-1*	*TOP-10*
Subject	real values	1000	37%	**70%**
		10,000	38%	69%
		100,000	**39%**	66%
	bit values	1000	30%	63%
		10,000	25%	54%
		100,000	15%	36%
Question	real values	1000	33%	52%
		10,000	**44%**	**69%**
		100,000	41%	65%
	bit values	1000	30%	40%
		10,000	36%	63%
		100,000	14%	37%
All text	real values	1000	29%	50%
		10,000	**46%**	**71%**
		100,000	45%	68%
	bit values	1000	27%	49%
		10,000	38%	65%
		100,000	17%	40%

The results show that using the bit valued ESA vectors yields a big loss in performance at the mate retrieval task, independently of the text parts that were used. It seems therefore to be important to use the relevance of articles to the queries that is encoded in the real values of the ESA vector representation of queries.

Looking at the number of dimensions of the ESA vector that were used, 10,000 seems to be a good value for this parameter. Using more dimensions does not yield better precision. For queries consisting of question part of the documents and all of the text, the results are even worse.

Comparing the results using different text parts as queries the differences are not significantly different. As e.g. subjects only consist of a few words but the whole documents contain several sentences, this is an unexpected result. It seems that this method works good for short queries, but with longer queries more noise is added as well and the retrieval performance therefore is not getting much better.

6 Related Work

The first approaches to Cross-lingual Information Retrieval (CLIR) were based on the translation of the query into the language of the target documents. Hull and Grefenstette presented a system that uses the term vector translation model [10]. All terms of the query are translated by looking them up in a bilingual dictionary. A problem of this approach is that many terms have multiple translations which are all added to the translated query. This leads to a loss of precision in the retrieval process. Demner-Fushman and Oard studied the effect of the size of the bilingual term list in dictionary based CLIR [5]. One of there results is that term lists with above 30,000 entries optimize the coverage of general vocabulary in their experiments. Additionally they showed that the translation of named entities is very important and substantially influences the retrieval quality. Because of that they suggest that supplemental techniques for named entity translation are useful even with large lexicons.

Another approach to CLIR is based on on Latent Semantic Indexing (LSI). LSI applied to text documents is a technique to reduce the vector representation [3]. Based on a training corpus Principal Component Analysis (PCA) on the co-occurrence matrix of words can be used to identify relevant dimensions and to construct a mapping of the original Bag-of-Words vector space to these new dimensions. For CLIR

LSI can be applied by using a parallel corpus with documents in two languages for training. Parallel documents are therefore merged co-occurrences are computed across languages. The learned model can then be used for CLIR [2] [17]. If a training corpus in multiple languages is available, containing versions of all documents in all languages, LSI can also be used for CLIR in many languages [12].

Recently emerging approaches to CLIR use the Wikipedia database as background knowledge. Schoenhofen et al. [15] presented a system that translates queries based on a small dictionary and cross-language links in Wikipedia. Afterwards the terms of the translated query are mapped to Wikipedia articles. Different features of these articles are then used to filter the query terms that are used for retrieval. This approach is different to the presented approach as they use cross-language links to translate single query terms. In our approach these links are used to define a mapping of high dimensional vector spaces, that is used to map the ESA vector representation of the whole query.

Egozi et al. presented a system for monolingual IR using Wikipedia as background knowledge [6]. This work is highly relevant for this paper as they apply Explicit Semantic Analysis [7] to IR. Additionally they propose a method to improve the ESA mapping in regards to IR tasks based on Pseudo Relevance Feedback (PSF). This is done first performing standard Bag-of-Words retrieval with a query and then using these results to select relevant dimensions of the ESA vector representation of the same query. A future challenge will be to apply these techniques as well to multi-lingual IR based on the cross-lingual ESA approach we presented in this paper.

Another approach to use PRF in multi-lingual retrieval is described in by Qu et al [14]. They examined the effects of pre-translation feedback versus post-translation feedback and identified different errors that were induced through the query expansion.

After developing our approach and submitting this paper, our literature search discovered the paper by Potthast et al. [13], who independently of us developed and presented the CL-ESA model before. In their paper, they perform extensive evaluations on two datasets: Wikipedia and the JRC Acquis dataset[5]. We also intend to use this dataset in future experimental evaluation. The approaches also differ in the way the association between a text and a Wikipedia article is computed. While Potthast et al. use the cosine similarity between a document and a Wikipedia article as weight, we have simply used the tf.idf values for this purpose.

7 Conclusion

In this paper, we have presented our CL-ESA approach and the corresponding implementation with which we have participated in this year's CLEF campaign on the monolingual and bilingual Ad-Hoc retrieval tasks. In particular, we have presented a cross-lingual extension to the Explicit Semantic Analysis (ESA) approach of Gabrilovich and Markovitch. While the results are far from satisfactory, we think that there is still a lot of potential to improve the approach in future research. Questions which seem very important to us are in how far various measures for calculating the association strength between a word (or text) and a Wikipedia article as well as the selection of Wikipedia articles influence the overall results. The interesting experiments presented in [6] show that "less is more" in the sense that considering a small number of articles can be enough provided that they are selected appropriately. In direct future work, we plan to compare our method with LSI-based cross-lingual retrieval methods to find out more in detail about the performance of our approach, being able to better quantify the weaknesses of the current implementation.

Acknowledgments

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

[5] http://langtech.jrc.it/JRC-Acquis.html

References

[1] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley, 1999.

[2] M.W. Berry and P.G. Young. Using latent semantic indexing for multilanguage information retrieval. *Computers and Humanities*, 29(6):413–429, 1995.

[3] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[4] Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990.

[5] D. Demner-Fushman and D.W. Oard. The effect of bilingual term list size on dictionary-based cross-language information retrieval. In *Proceedings of the 36th Annual Hawaii International Conference on System Sciences (HICSS'03)*, 2003.

[6] Ofer Egozi, Evgeniy Gabrilovich, and S. Markovitch. Concept-based feature generation and selection for information retrieval. In *Proceedings of the Twenty-Third Conference on Artificial Intelligence (AAAI)*, 2008.

[7] 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.

[8] Evgeniy Gabrilovich. *Feature Generation for Textual Information Retrieval using World Knowledge*. PhD thesis, Israel Institute of Technology, Kislev, 5767 Haifa, Istrael, 2006.

[9] R. Gupta and L. Ratinov. Text categorization with knowledge transfer from heterogeneous data sources. In *Proceedings of the Twenty-Third Conference on Artificial Intelligence (AAAI)*, 2008.

[10] D.A. Hull and G. Grefenstette. Querying 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*, pages 49–57, 1996.

[11] T. Landauer and S. Dumais. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychol. Rev.*, 104:211–240, 1997.

[12] M.L. Littman and Greg A. Keim. Cross-language text retrieval with three languages. Technical report, Department of Computer Science, Duke University, Durham, North Carolina, 1997.

[13] Martin Potthast, Benno Stein, and Maik Anderka. A wikipedia-based multilingual retrieval model. In Craig Macdonald, Iadh Ounis, Vassilis Plachouras, Ian Ruthven, and Ryen W. White, editors, *Proceedings of the 30th European Conference on IR Research (ECIR)*, pages 522–530, 2008.

[14] Y. Qu, A.N. Eilerman, H. Jin, and D.A. Evans. The effect of pseudo relevance feedback on mt-based clir. In *Proceedings of the RIAO (Recherche d'Information Assiste par Ordinateur) Conference*, 2000.

[15] P. Schoenhofen, A. Benzcur, I. Biro, and K. Csalogany. Performing cross-language retrieval with wikipedia. In *Proceedings of CLEF 2007*, 2007.

[16] Philipp Sorg and Philipp Cimiano. Enriching the crosslingual link structure of wikipedia - a classification-based approach. In *Proceedings of the AAAI 2008 Workshop on Wikipedia and Artifical Intelligence*, 2008.

[17] M.L. Littman S.T. Dumain, Todd A. Letsche and T.K. Landauer. Automatic cross-language retrieval using latent semantic indexing. In *Proceedings of the AAAI Symposium on CrossLanguage Text and Speech Retrieval*, 1997.