A Wikipedia-Based Multilingual Retrieval Model

Martin Potthast, Benno Stein, and Maik Anderka

Bauhaus University Weimar, Faculty of Media, 99421 Weimar, Germany
<first name>.<last name>@medien.uni-weimar.de

Abstract. This paper introduces CL-ESA, a new multilingual retrieval model for the analysis of cross-language similarity. The retrieval model exploits the multilingual alignment of Wikipedia: given a document d written in language L we construct a concept vector $\mathbf{d}$ for d, where each dimension i in $\mathbf{d}$ quantifies the similarity of d with respect to a document d_i^* chosen from the "L-subset" of Wikipedia. Likewise, for a second document d' written in language L', $L \neq L'$, we construct a concept vector $\mathbf{d}'$, using from the L'-subset of the Wikipedia the topic-aligned counterparts $d_i'^*$ of our previously chosen documents.

Since the two concept vectors $\mathbf{d}$ and $\mathbf{d}'$ are *collection-relative representations* of d and d' they are language-independent. I. e., their similarity can directly be computed with the cosine similarity measure, for instance.

We present results of an extensive analysis that demonstrates the power of this new retrieval model: for a query document d the topically most similar documents from a corpus in another language are properly ranked. Salient property of the new retrieval model is its robustness with respect to both the size and the quality of the index document collection.

1 Introduction

Retrieval models are used to assess the similarity between documents. For this purpose a retrieval model provides (*i*) the rationale for the construction of a particular document representation $\mathbf{d}$ given a real-world document d, and, (*ii*) a similarity measure φ to quantify the similarity between two representations $\mathbf{d}$ and $\mathbf{d}'$.

This paper deals with retrieval models that can be applied in a cross-language retrieval situation, i.e. when a document d is written in language L and we would like to assess its similarity to a document d' written in language L'. Our contributions are the following:

- *Multilingual Retrieval Model.* Section 2 introduces the new multilingual retrieval model CL-ESA, which overcomes many of the restrictions of previous approaches when quantifying cross-language similarity.
- *Evaluation.* Section 3 reports on experiments related to retrieval performance, dimensionality dependence, and runtime. In particular, with the so-called bilingual rank correlation a new measure for cross-lingual retrieval performance is proposed.
- *Comparable Corpus Wikipedia.* We use Wikipedia as a comparable corpus and demonstrate its usability for cross-lingual retrieval.

C. Macdonald et al. (Eds.): ECIR 2008, LNCS 4956, pp. 522–530, 2008.

Table 1. Assessment results for cross-language retrieval models with respect to six criteria: the number of languages, the computational complexity to compute the retrieval model, the number of documents which can be represented with reasonable effort, the availability of resources to construct the retrieval model, the achievable retrieval quality, and the specificity of the retrieval model to a (topic) domain.

Multilingual retrieval model		Multilinguality (# languages)	Computational complexity	Scalability (# documents)	Resource availability	Retrieval quality	Domain specificity
CL-VSM	[1,13,6]	2	low	Web	medium	medium	none
Eurovoc-based	[8,9]	21	medium	Web	poor	medium	medium
CL-LSI	[2,10]	3	high	10^4	–	very good	total
CL-KCCA	[15]	2	high	10^4	poor	very good	total
CL-RM	[5]	2	medium	Web	good	good	medium
CL-ESA		**14**	**low**	**Web**	**good**	**good**	**low**

1.1 Comparison of Related Work

Cross-language retrieval models are generalizations of monolingual models such as the vector space model (VSM), latent semantic indexing (LSI), principal component analysis (PCA), language or relevance models (RM), or—as in the case of our approach—explicit semantic analysis (ESA). However, deductions that come at no expense in a monolingual retrieval situation are difficult to be achieved between two languages L and L': terms, named entities, time or currency expressions, etc. have to be identified and mapped from L to L', which entails the problem of translation ambiguity. Basically, there are two possibilities to bridge the language barrier: (*i*) dictionaries, gazetteers, rules, or thesauri versus (*ii*) parallel corpora or comparable corpora.[1] The former provide a means to translate words and concepts such as locations, dates, and number expressions directly from L to L', whereas the latter provide "aligned" documents from L and L' that are translations of each other or that cover the same topic, and which are utilized to translate arbitrary texts.

We have analyzed the existing approaches to overview their strong and weak points; the results are comprised in Table 1: the first group shows dictionary-based approaches and the second group corpus-based approaches. For large-scale retrieval tasks the computational complexity and the resource availability disqualify the Eurovoc-based approach, CL-LSI, and CL-KCCA. Among the remaining approaches CL-RM and CL-ESA have the advantage that no direct translation effort is necessary.

2 Explicit Semantic Analysis

This section introduces the principle of cross-language explicit semantic analysis, CL-ESA, a new multilingual retrieval model which does without automatic translation capabilities. Our starting point is a recently proposed monolingual retrieval model, the explicit semantic analysis, ESA [3,4].

[1] There has been much confusion concerning corpora termed "parallel" and "comparable"; the authors of [7] provide a consistent definition.

Let D^* denote a document collection of so-called index documents, and let φ denote the cosine similarity measure. Under ESA a document d is represented as an n-dimensional concept vector $\mathbf{d}$:

$$\mathbf{d} = (\varphi(\mathbf{v}, \mathbf{v}_1^*), \ldots, \varphi(\mathbf{v}, \mathbf{v}_n^*))^T,$$

where $\mathbf{v}$ is the vector space model representation of d, $\mathbf{v}_i^*$ is the vector space model representation of the ith index document in D^*, and n is the size of D^*. If $\varphi(\mathbf{v}, \mathbf{v}_i^*)$ is smaller than a noise threshold ε the respective entry is set to zero. Let $\mathbf{d}'$ be the concept representation of another document d'. Then the similarity between d and d' under ESA is defined as $\varphi(\mathbf{d}, \mathbf{d}')$.

Rationale of the ESA retrieval model is to encode the specific knowledge of d relative to the collection D^*. In this sense each document in D^* is used as a single concept to which the document d is compared, say, $\mathbf{d}$ can be understood as a projection of d into the concept space spanned by D^*. The authors in [4] achieved with ESA an average retrieval improvement of 20% compared to the vector space model.

To function as a generic retrieval model the index document collection D^* must be of a low domain specificity: D^* should contain documents from a broad range of domains, and each index document should be of "reasonable" length. A larger subset of the documents in Wikipedia fulfills both properties.

2.1 CL-ESA

Let $\mathcal{L} = \{L_1, \ldots, L_m\}$ denote a set of languages, and let $\mathcal{D}^* = \{D_1^*, \ldots, D_m^*\}$ denote a set of index document collections where each D_i^* contains index documents of language L_i. Moreover, let $C = \{c_1, \ldots, c_n\}$ denote a set of concept descriptors. $\mathcal{D}^*$ is called a concept-aligned comparable corpus if it has the property that the ith index document, d_i^*, of each index document collection $D^* \in \mathcal{D}^*$ describes c_i in its respective language.

A document d written in language $L \in \mathcal{L}$ is represented as ESA vector $\mathbf{d}$ by using that index document collection $D^* \in \mathcal{D}^*$ that corresponds to L. Likewise, a document d' from another language $L' \in \mathcal{L}$ is represented as $\mathbf{d}'$. The similarity between d and d' is quantified in the concept space, by computing the cosine similarity between $\mathbf{d}$ and $\mathbf{d}'$.

CL-ESA exploits the following understanding of a comparable corpus alignment: if all concepts in C are described "sufficiently exhaustive" for all languages in $\mathcal{L}$, the documents d and d' are represented in comparable concept spaces under ESA, using the associated index document collections D^* and D'^* in $\mathcal{D}^*$.

CL-ESA requires a comparable corpus $\mathcal{D}^*$, and each index document collection $D^* \in \mathcal{D}^*$ should meet the requirements for the monolingual explicit semantic analysis. Again, a larger subset of the documents in Wikipedia fulfills these properties.

3 Evaluation

To analyze the power of the CL-ESA retrieval model we implemented various experiments on a multilingual parallel and a multilingual comparable corpus. The results can be summarized as follows. (*i*) Given a document d in language L, CL-ESA ranks the aligned document d' in language L' with 91% probability on the first rank. (*ii*) Given a rank ordering in language L, CL-ESA is able to reproduce this ordering in language L' at a high fidelity. (*iii*) CL-ESA is insensitive with respect to the quality of the underlying index document collection. (*iv*) CL-ESA behaves robust with respect to a wide range of the concept space dimensionality.

Altogether CL-ESA is a viable retrieval model to assess the cross-language similarity of text documents. The remainder of this section describes the experiments in greater detail.

Multilingual Corpora. In our experiments we have employed the parallel corpus JRC-Acquis [14], and the comparable corpus Wikipedia. As one of the largest corpora of its kind the JRC-Acquis corpus contains 26 000 aligned law documents per language from the European Union in 22 languages. Wikipedia has not been considered as a comparable corpus by now. This fact is surprising since up to 100 000 aligned documents are available from diverse domains and languages, and the corpus is constantly extended by Wikipedia's editors. On the downside the aligned documents may be of less quality than those of custom-made comparable corpora.

Test Collections. Two test document collections comprising 3 000 documents each were selected from the German (D, L) and the English (D', L') parts of the multilingual corpora. Both collections contain 1 000 randomly selected translation-aligned documents from JRC-Acquis, 1 000 concept-aligned documents from Wikipedia, and 1 000 not aligned documents from Wikipedia. The latter have no language link from L to L' or vice versa. In particular, we assured that the distribution of monolingual similarities among the documents in D and D' corresponds to normal orders of magnitude.

The aligned index document collections D^* and D'^* were constructed from Wikipedia so that $D^* \cap D = \emptyset$ and $D'^* \cap D' = \emptyset$: no document is index document and test document at the same time. The size $n = |D^*| = |D'^*|$ of these collections corresponds to the dimensionality of the resulting document representations in the concept space.

3.1 Experiments

This subsection describes six selected experiments from our evaluation.

Experiment 1: Cross-Language Ranking. Given an aligned document $d \in D$, all documents in D' are ranked according to their cross-language similarity to d. Let $d' \in D'$ be the aligned document of $d \in D$, then the retrieval rank of d' is recorded. Ideally, d' should be on the first or at least on one of the top ranks. The experiment was repeated for all of the aligned documents in D. The first column of Table 2 shows the recall at ranks ranging from 1 to 50. The probability of finding a document's translation- or concept-aligned counterpart on the first rank is 91%, and the probability of finding it among the top ten ranks is > 99%.

Experiment 2: Bilingual Rank Correlation. To quantify the retrieval quality related *to a set* of retrieved documents we propose a new evaluation statistic. Starting point is a pair of aligned documents $d \in D$ and $d' \in D'$, whereas the documents from D'

are ranked twice: (*i*) with respect to their cross-language similarity to d using a cross-language retrieval model, and, (*ii*) with respect to their monolingual similarity to d' using the vector space model. The top 100 ranks of the two rankings are compared using a rank correlation coefficient, e. g. Spearman's ρ, which measures their disagreement or agreement as a value between -1 and 1 respectively.

The idea of this statistic relates to "diagonalization": a reference ranking under the vector space model is compared to a test ranking computed under the CL-ESA concept space representation. The experiment is conducted for each pair of aligned documents d and d' in the test collections, averaging the rank correlations. The second column of Table 2 shows a high correlation, provided a high dimensionality of the concept space. Note that this experiment is a generalization of Experiment 1 and that it has much more explanatory power.

Experiment 3: Cross-Language Similarity Distribution. This experiment contrasts the distribution of pairwise similarities of translation-aligned documents and concept-aligned documents. The results show that, on average, for both kinds of aligned documents high similarities are computed (cf. Table 2, third column), which demonstrates that CL-ESA is robust with respect to the quality of the aligned documents in the index document collections D^* and D'^*.

Experiment 4: Dimensionality. Both retrieval quality and runtime depend on the concept space dimension of CL-ESA, which in turn corresponds the size of a language's index document collections D^* and D'^*. The dimensionality of a retrieval model affects the runtime of all subsequently employed retrieval algorithms. Under CL-ESA, documents can be represented with a reasonable number of 1 000 to 10 000 dimensions while both retrieval quality and runtime are maintained (cf. Table 2 and Figure 1).

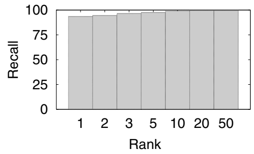
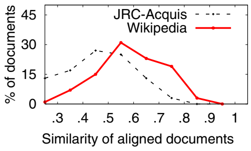
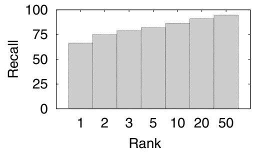
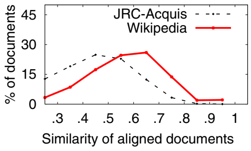
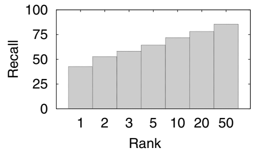
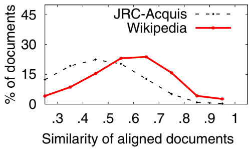
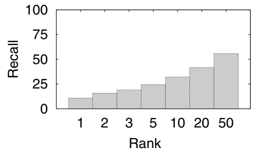
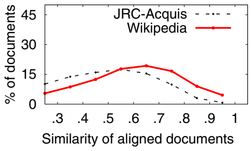
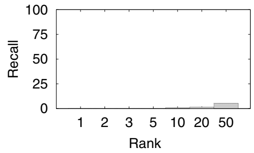
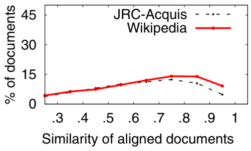
Experiment 5: Multilinguality. Starting with the two most prominent languages in Wikipedia, English and German, we study how many concepts are described in both languages, and how many are in the intersection set if more languages are considered. Currently, the Wikipedia corpus allows that documents from up to 14 languages are represented with CL-ESA (cf. Figure 1, left plot).

Experiment 6: Indexing Time. The time to index a document is between 10 to 100 milliseconds, which is comparable to the time to compute a vector space representation (cf. Figure 1, right plot). Employed hardware: Intel Core 2 Duo processor at 2 GHz and with 1 GB RAM.

3.2 Discussion

The evaluation of this section provides a framework for the adjustment of CL-ESA to the needs of a cross-language retrieval task. If, for example, a high retrieval quality is desired, documents should be represented as 10^5-dimensional concept vectors: ranking with respect to a particular query document will provide similar documents on the top ranks with high accuracy (cf. Table 2, first row). High retrieval quality comes at the price that with the current Wikipedia corpus only 2 languages can be represented at the same time, and that the time to index a document will be high (cf. Figure 1). If high retrieval

Table 2. Landscape of cross-language explicit semantic analysis: each row shows the results of three experiments, depending of the dimenionality n of the concept space

Experiment 1	**Experiment 2**		**Experiment 3**	**Dimension**
Cross-Language Ranking	**Bilingual Rank Correlation**		**CL Similarity Distribution**	n
	JRC-Acquis	Wikipedia		
	0.81	0.72		10^5
	0.46	0.61		10^4
	0.20	0.44		10^3
	0.09	0.22		10^2
	0.04	0.07		10

speed or a high multilinguality is desired, documents should be represented as 1000-dimensional concept vectors. At a lower dimension the retrieval quality deteriorates significantly. A reasonable trade-off between retrieval quality and runtime is achieved for a concept space dimensionality between 1 000 and 10 000.

Concerning the multilinguality of CL-ESA the left plot in Figure 1 may not show the true picture: if the languages in Wikipedia are not considered by their document number but by geographical-, cultural-, or linguistic relations, there may be more intersecting concepts in the respective groups. And, if only two languages are considered, the number of shared concepts between a non-English Wikipedia and the English Wikipedia will be high in most cases.

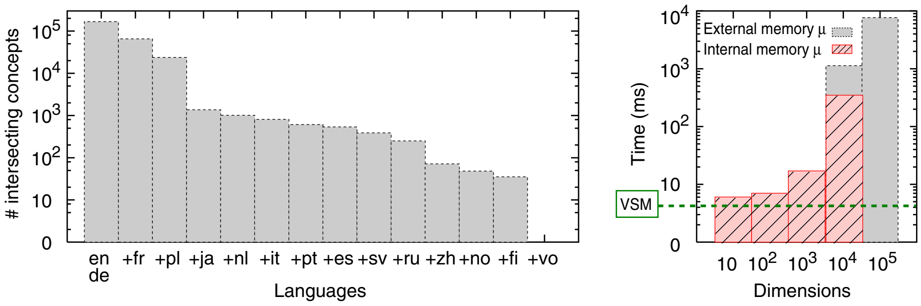


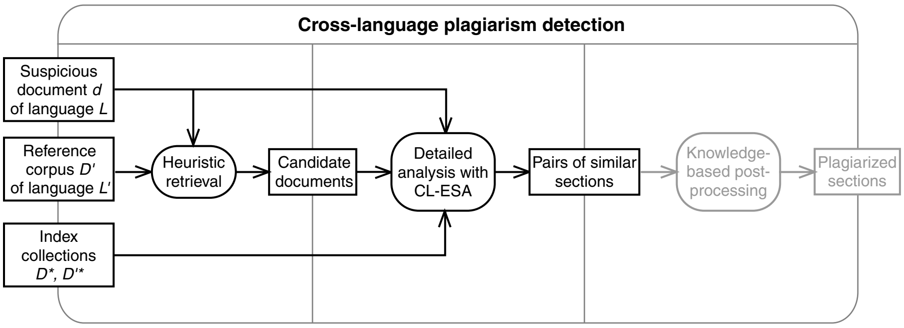
Fig. 1. Left: number of intersecting concepts among groups of languages in Wikipedia. The languages are organized in descending order wrt. the number of available documents. Right: average time to index a document under ESA, depending on the number of dimensions. We distinguish between indexes that fit in internal memory (μ), and external indexes.

4 Current Work

Our current work focuses on cross-language plagiarism detection. Plagiarism is the act of copying the work of another author and claiming it as own work. Though automatic plagiarism detection is an active field of research the particular case of cross-language plagiarism has not been addressed in detail so far.

The authors of [11] propose a three-step retrieval process to detect plagiarism, which can also be applied to detect cross-language plagiarism. Figure 2 illustrates the process. A suspicious document d of language L, which may contain a plagiarized section from a document d' in a reference corpus D' of language L', is analyzed as follows:

1. *Heuristic Retrieval.* A subset of D' is retrieved which contains candidate documents that are likely to be sources for plagiarism with respect to the content of d.
2. *Detailed Analysis.* The candidate documents are compared section-wise to d using CL-ESA for each pair of sections.



Cross-language plagiarism detection
Suspicious document d of language L
Reference corpus D′ of language L′
Index collections D*, D′*
Heuristic retrieval
Candidate documents
Detailed analysis with CL-ESA
Pairs of similar sections
Knowledge-based post-processing
Plagiarized sections

Fig. 2. A three-step process for cross-language plagiarism detection

3. *Knowledge-based Post-Processing.* Sections from the candidate documents that are similar to a section in d are processed in detail, for instance to filter cases of proper citation.

Cross-language explicit semantic analysis can be directly applied for Step 2, the detailed analysis, since it allows for a reliable assessment of cross-language similarity. However, for the preceding heuristic retrieval CL-ESA is not the best choice since a pairwise comparison of d to all documents from D' is required in order to cope with the high dimensionality of CL-ESA representations. To speed up this retrieval step we are investigating the following alternatives:

- Construction of a keyword index for D' which is queried with keywords extracted from d that are translated to L', and implementation of a focused keyword search.
- Construction of a keyword index for D' which is queried with keywords extracted from d', the machine translation of d to L', and, again, implementation of a focused search.
- Construction of a hash-based fingerprint index for D' which is queried with the fingerprint of d' [12].

The first two alternatives are based on keyword extraction as well as on cross-language keyword retrieval or machine translation technologies. The last alternative, which has the potential to outperform the retrieval recall of the first approaches, employs machine translation and similarity hashing technologies.

References

1. Ballesteros, L.: Resolving Ambiguity for Cross-Language Information Retrieval: A Dictionary Approach. PhD thesis, Director-W. Bruce Croft (2001)
2. Dumais, S., Letsche, T., Littman, M., Landauer, T.: Automatic cross-language retrieval using latent semantic indexing. In: AAAI 1997, Cross-Language, Text, and, Speech, Retrieval (1997)
3. Gabrilovich, E.: Feature Generation for Textual Information Retrieval Using World Knowledge. Phd thesis, Israel Institute of Technology (2006)
4. Gabrilovich, E., Markovitch, S.: Computing semantic relatedness using wikipedia-based explicit semantic analysis. In: IJCAI 2007, Hyderabad, India (2007)
5. Lavrenko, V., Choquette, M., Croft, W.: Cross-Lingual Relevance Models. In: SIGIR 2002, pp. 175–182. ACM Press, New York (2002)
6. Levow, G.-A., Oard, D., Resnik, P.: Dictionary-based techniques for cross-language information retrieval. Inf. Process. Manage. 41(3), 523–547 (2005)
7. McEnery, A., Xiao, R.: Parallel and comparable corpora: What are they up to? Incorporating Corpora: The Linguist and the Translator (2007)
8. Pouliquen, B., Steinberger, R., Ignat, C.: Automatic annotation of multilingual text collections with a conceptual thesaurus. In: OntoIE 2003 at EUROLAN 2003, pp. 9–28 (2003)
9. Pouliquen, B., Steinberger, R., Ignat, C.: Automatic identification of document translations in large multilingual document collections. In: RANLP 2003, pp. 401–408 (2003)
10. Rehder, B., Littman, M., Dumais, S., Landauer, T.: Automatic 3-language cross-language information retrieval with latent semantic indexing. In: TREC, pp. 233–239 (1997)
11. Stein, B., zu Eissen, S.M., Potthast, M.: Strategies for retrieving plagiarized documents. In: SIGIR 2007, pp. 825–826 (2007)

12. Stein, B.: Principles of hash-based text retrieval. In: SIGIR 2007, pp. 527–534 (2007)
13. Steinberger, R., Pouliquen, B., Ignat, C.: Exploiting multilingual nomenclatures and language-independent text features as an interlingua for cross-lingual text analysis applications. In: 4th Language Technology Conference at Information Society, Slovenia (2004)
14. Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufis, D., Varga, D.: The JRC-Acquis:A multilingual aligned parallel corpus with 20+languages. In: LREC 2006 (2006)
15. Vinokourov, A., Shawe-Taylor, J., Cristianini, N.: Inferring a semantic representation of text via cross-language correlation analysis. In: NIPS 2002, pp. 1473–1480. MIT Press, Cambridge (2003)