Cross-Lingual Search over 22 European Languages

Blaž Fortuna, Jan Rupnik, Boštjan Pajntar, Marko Grobelnik, Dunja Mladenič
Institute Jozef Stefan
Jamova cesta 39, 1000 Ljubljana, Slovenia
+386 1 477 3934
blaz.fortuna@ijs.si, jan.rupnik@ijs.si, marko.grobelnik@ijs.si, dunja.mladenic@ijs.si

ABSTRACT
In this paper we present a system for cross-lingual information retrieval, which can handle tens of languages and millions of documents. Functioning of the system is demonstrated on corpus of European Legislation (22 languages, more than 400,000 documents per language). The system uses an interactive web-interface, which can take advantage of a predefined thesaurus allowing the user to dynamically re-rank the retrieval results based on the mapping onto a predefined thesaurus.

Categories and Subject Descriptors
H.3.3 [**Information Search and Retrieval**]

General Terms
Performance, Experimentation, Human Factors.

Keywords
Canonical Correlation Analysis, Cross-Lingual Information Retrieval, Dynamic Search Ranking, Search Result Visualization

1. INTRODUCTION
In this paper we present a system for cross-lingual information retrieval (CLIR) working over the multilingual corpora of European Legislation Acquis Communautaire [1]. The unique part of the corpora is an alignment over 22 official languages of European Union and annotation of documents using EuroVoc thesaurus. The system uses Canonical Correlation Analysis for correlating words from different languages and extracting language-independent latent dimensions, which serve as interlingua used to index the documents and match them against given queries [2]. Finally, during the retrieval time, EuroVoc thesaurus is used to let the user visually extend the query and re-rank the results in real-time.

2. MAIN COMPONENTS

2.1 Corpus
The system is performing search over Acquis Communautaire (AC) corpus [1]. The corpus includes more than 400,000 texts written in 22 languages, namely: Bulgarian, Czech, Danish, German, Greek, English, Spanish, Estonian, Finnish, French, Hungarian, Italian, Lithuanian, Latvian, Maltese, Dutch, Polish, Portuguese, Romanian, Slovak, Slovene, and Swedish. An alignment is provided for each language pair, resulting in 231 alignments. Note that not all of the texts appear in all the languages or are aligned with all the languages.

Most of the texts have been manually classified into the EuroVoc thesaurus. EuroVoc is a multilingual thesaurus used specially for annotating European Legislation. It consists of more than 6,000 terms which are organized hierarchically up to 8 levels deep.

2.2 Canonical Correlation Analysis
The core component of the system is Canonical Correlation Analysis [2], a method for automatically extracting latent language-independent concepts from aligned corpora of texts. The documents are indexed by the discovered latent concepts and a standard inverted index is used for the retrieval.

The presented system uses a new enhanced version of the method, which is able to linearly scale with regards to the number of documents and number of languages. This makes it especially appropriate for the task of CLIR over millions of documents from 22 different languages as in the AC corpora. Due to efficiency of the applied method the whole search engine, including the preprocessing, correlation and indexing stage, can run on a single high-end desktop computer.

2.3 Dynamic Ranking
The main interface of the search engine is a web site, like in standard search engines: the user specifies a query and gets back a list of relevant documents. The user can also specify language of the query and can filter the relevant documents by language. In case when language of the query is not specified, the system automatically classifies the query into one of the 22 languages.

The user can navigate the result list by using the SearchPoint gadget [3], which is located on the right side of the results list. The gadget visualizes the most relevant EuroVoc terms for the documents from the list and the user can choose, by a simple mouse move, to re-rank in real-time the result list by putting emphasis on the selected EuroVoc terms.

3. REFERENCES
[1] Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D., Varga, D. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006). Genoa, Italy, 24-26 May 2006.

[2] Fortuna, B., Cristianini, N., Shawe-Taylor, J. 2006. A Kernel Canonical Correlation Analysis For Learning The Semantics Of Text. Kernel methods in bioengineering, communications and image processing, edited by G. Camps-Valls, J. L. Rojo-Álvarez & M. Martínez-Ramón.

[3] Pajntar, B., Grobelnik, M., http://searchpoint.ijs.si/

SIGIR'08, July 20-24, 2008, Singapore.
ACM 978-1-60558-164-4/08/07.