

Towards Cross-Language and Cross-Domain Exploration of Research Platforms

Andreas Hense¹, Florian Quadt²

¹Bonn-Rhein-Sieg University of Applied Sciences, Sankt Augustin, Germany, andreas.hense@h-brs.de

²Bonn-Rhein-Sieg University of Applied Sciences, Sankt Augustin, Germany, florian.quadt@h-brs.de

INTRODUCTION

From the very beginning, the Internet has been used increasingly by researchers of different mother tongues and research areas. This linguistic diversity is reflected in multilingual web content and scientific publications. While machine translation is the most desirable grade of multilingual collaboration, fully automated and technically mature translation systems still do not exist. Nevertheless, it is possible to use the knowledge of how to translate that has been gathered so far for an application that is not as complex as machine translation: cross-language information retrieval (CLIR). CLIR relies on the fact that users of search engines can read and understand more than one single language. This is why search engines supporting CLIR present multilingual hits to the user.

We will discuss the advantages of using a concept-based interlingua and present a prototype of an interlingua-based CLIR system that has been tested with English, German, and Japanese documents. Additionally, we will focus on selected challenges in information retrieval as well as in machine translation.

USING A CONCEPT-BASED INTERLINGUA

There are several ways to design a cross-language retrieval system. The most intuitive approach is to develop a system that translates either the search query terms or the document terms into all supported languages. With respect to the support of numerous languages, this approach requires a lot of dictionaries – to be precise: the number of dictionaries grows quadratically with each newly added language. This can be avoided if there is just one single (possibly non-natural) language that all other languages have to be translated into and in which all the retrieval work is performed. Such a language is called *interlingua*. Support for new languages is added by providing a single dictionary that translates from this language into the interlingua, therefore the number of dictionaries grows linearly. This loss of complexity is illustrated in Figure 1.

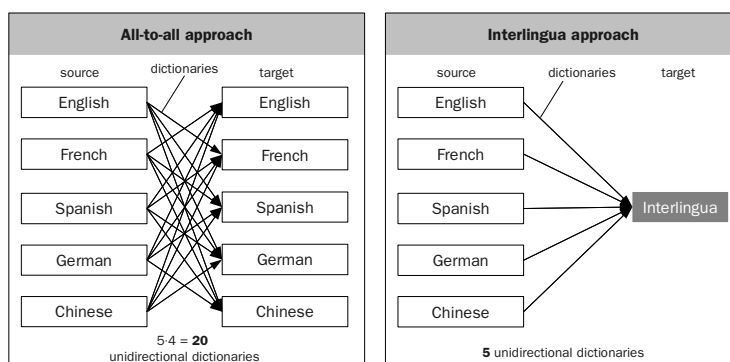


Figure 1: Comparison of the all-to-all and the interlingua approach

It is obvious that from a dictionary perspective an interlingua based search engine is less complex to maintain and to extend. If we are about to create a new language that serves as interlingua, we do not want to take over the weaknesses of natural languages such as ambiguity and synonymy. That is why we base this language on concepts to make it unambiguous.

It is important to design the interlingua carefully: Is a *vehicle* a concept? Or should we distinguish between *car*, *truck* and *motorcycle*? Or is it even necessary to differentiate a *BMW motorcycle* and a *Ferrari sports car*? This problem is still being researched eagerly in the context of semantic web, taxonomies and ontologies [1]. The problem does not stop at the question of how fine-grained a concept structure should be. There are a lot of concepts that could be structured in different ways (e. g. countries can either be classified by geographical position or by spoken languages). The hierarchies chosen will have to cover all the concepts that professional users of a specific research area may need.

CHARACTERISTICS OF THE PROTOTYPE

To describe the characteristics of the developed prototype we will distinguish the process of indexing the document base and the process of querying and presenting the results. The former is fully automated while the latter is user-interactive. The chosen interlingua is not structured (like an ontology) and describes concepts with unambiguous German words.

In order to add a text document to an index it typically passes several processing steps. These preparatory operations are heavily dependent on language and search aims (e. g. term search vs. full text search). Basic operations [2] are tokenizing, stop words removal and stemming [3]. Since the system supports cross-language retrieval and utilizes a

concept-based interlingua it introduces two additional operations: language detection and disambiguation. Figure 2 shows the implemented workflow.

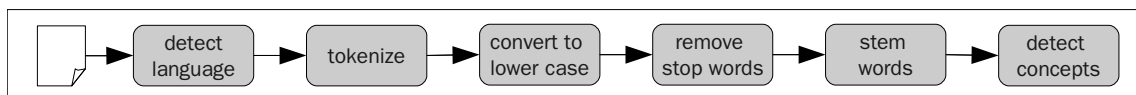


Figure 2: Possible workflow for indexing

For language detection the system generates n -grams [4] and compares them to existing language profiles. In the end the most similar language is chosen. This was successfully tested with English, German and Japanese documents.

The disambiguation feature simulates human text processing and considers surrounding words to detect the concept of ambiguous words. This requires dictionaries that cover more than word-concept pairs and additionally list signal words to differentiate one concept from another. In the following examples the signal words determining the correct concept of the word *plant* are underlined:

1. The park features a large botanic collection of native plants.

2. There is a new plant to manufacture automobiles.

The English-to-Interlingua-dictionary must contain the following entries to detect the correct concept:

word	concept	signal words
plant →	PLANT →	botanic, native [...]
plant →	BUILDING →	manufacture, automobile [...]

The disambiguation algorithm internally builds multidimensional vectors for each concept and compares the documents to them (cosine similarity). The most similar concept is chosen [5].

Since queries are usually quite short we do not apply our language detection and disambiguation algorithm to the query string. The user has to declare the query language and if the query string contains ambiguous words, the user is presented with a drop-down list in which he can select the correct concept. After this the engine looks up all documents that contain the given concepts and ranks them. For every hit the result page shows the document path and a text snippet in which the concept words are highlighted.

TECHNICAL SPECIFICATIONS

The prototype is implemented as an open-source-based web application with Java Server Pages (JSP) and JavaBeans. It runs in a Tomcat container. The index is based on the search engine Apache Lucene. Besides the web interface the search engine is additionally available as web-service and therefore can easily be integrated in other environments.

FUTURE RESEARCH

Based on the existing prototype there are some interesting aspects that could be further researched. First, it could be implemented that document domains are detected (utilizing vector comparison and domain profiles) and saved as tags in the index. If suitable domain-dictionaries are provided, the search engine could be additionally used for inter-domain retrieval. Moreover, it is quite sumptuous to create the language-to-Interlingua dictionaries manually from scratch. It could be a promising approach to use existing corpora (e.g. WordNet [6]) to initially fill the dictionaries and add user feedback controls to the result page. This way the system could learn concept words during daily use by evaluating irrelevant documents and user selected (anti-)signal words. Language is consequently a subject of change. This way the search engine core would be permanently updated and could become a useful Web 2.0 application.

There are numerous areas where cross-language retrieval is eagerly researched. The European Cross Language Evaluation Forum (CLEF) [7], for example, deals with cross-language retrieval of images and videos and grid experiments in its 2009 track.

REFERENCES

1. Soergel, D., *Multilingual thesauri and ontologies in Cross-language retrieval*. In *AAAI Symposium on Cross-language Text and Speech Retrieval*. College of Library and Information Services.
2. Baeza-Yates, R. and B. Ribeiro-Neto, *Modern Information Retrieval*. Addison Wesley, New York, 1999.
3. Porter, M., *Snowball: A language for stemming algorithms*. Available from: <http://www.snowball.tartarus.org/texts/introduction.html>, accessed 8 Jun 2009.
4. Cavnar, W. B. and J. M. Trenkle, *N-Gram-Based Text Categorization*, in *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, p. 161-175, Las Vegas, US, 1994.
5. Quadt, F., *Implementierung einer erweiterbaren Anwendung zur sprachübergreifenden Metadaten- und Volltextsuche in einer serviceorientierten Architektur*. Available from: http://www.bis.inf.fh-brs.de/pdf/master_thesis_quadt_florian.pdf, accessed 21 May 2009.
6. *WordNet*. Available from <http://wordnet.princeton.edu>, accessed 9 Jun 2009.
7. *Cross Language Evaluation Forum (CLEF)*. Available from: <http://www.clef-campaign.org>, accessed 9 Jun 2009.