Multilingual Search for Cultural Heritage Archives via Combining Multiple Translation Resources

Gareth J. F. Jones, Ying Zhang, Eamonn Newman, Fabio Fantino
Centre for Digital Video Processing
Dublin City University
Dublin 9, Ireland
{gjones,yzhang,enewman,ffantino}
@computing.dcu.ie

Franca Debole
ISTI-CNR
Pisa
Italy
franca.debole
@isti.cnr.it

Abstract

The linguistic features of material in Cultural Heritage (CH) archives may be in various languages requiring a facility for effective multilingual search. The specialised language often associated with CH content introduces problems for automatic translation to support search applications. The MultiMatch project is focused on enabling users to interact with CH content across different media types and languages. We present results from a MultiMatch study exploring various translation techniques for the CH domain. Our experiments examine translation techniques for the English language CLEF 2006 Cross-Language Speech Retrieval (CL-SR) task using Spanish, French and German queries. Results compare effectiveness of our query translation against a monolingual baseline and show improvement when combining a domain-specific translation lexicon with a standard machine translation system.

1 Introduction

Online Cultural Heritage (CH) content is being produced in many countries by organisations such as national libraries, museums, galleries and audiovisual archives. Additionally, there are increasing amounts of CH relevant content available more generally on the World Wide Web. While some of this material concerns national or regional content only of local interest, much material relates to items involving multiple nations and languages, for example concerning events in Europe or Asia. In order to gain a full understanding of such events, including details contained in different collections and exploring different cultural perspectives requires effective multilingual search technologies. Facilitating search of this type requires translation tools to cross the language barrier between users and the available information sources.

CH content encompasses various different media, including of course text documents, images, videos, and audio recordings. Search of text documents between languages forms the focus of cross-language information retrieval (CLIR) research, while search for images is the concern of content-based image retrieval. However, whatever the media of the items they are accompanied by metadata. Such metadata may include simple factual details such as date of creation, but also descriptive details relating to the contents of the item. Multilingual searching using metadata content requires that either the metadata be translated into a language with which the user is able to search or that the search query be translated into the language of the metadata. This alternative of document or query translation is a well rehearsed argument in CLIR, which has generally concerned itself with full text document searching. However, the features of metadata require a more careful analysis. Metadata is typically dense in search terms, while lacking the linguistic structure and information redundancy of full text documents. The absence of linguistic structure makes precise translation of content problematic, while the lack of redundancy means that accurate translation of individual words

Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007), pages 81–88,
Prague, 28 June 2007.

and phrases is vital to minimise mismatch between query and document terms. Furthermore, CH content is typically in specialised domains requiring domain specific resources for accurate translation. Developing reliable and robust approaches to translation for metadata search is thus an important component of search for many CH archives.

The EU FP6 MultiMatch[1] project is concerned with information access for multimedia and multilingual content for a range of European languages. In the investigation reported in this paper we introduce the first stage multilingual search functionality of the MultiMatch system, and describe its use in an investigation for multilingual metadata search. Since at present we do not have a search test collection specifically developed for MultiMatch we use data from the CLEF 2006 Cross-Language Speech Retrieval (CL-SR) task for our experiments (Oard et al., 2006).

The remainder of this paper is organised as follows: Section 2 gives an overview of the MultiMatch search architecture, Section 3 outlines the experimental search task, Section 4 describes the translation resources used for this study, Section 5 and 6 concern our experimental setup and results, and finally Section 7 summarises our conclusions and gives details of our ongoing work.

2 MultiMatch Search System

The MultiMatch search system is centered on the MILOS Multimedia Repository system (Amato et al., 2004) which incorporates free-text search using Lucene (Hatcher and Gospodnetic, 2004) and image search using an open source image retrieval system GIFT (Müller et al., 2001). In order to support multilingual searching a number of translation tools are being developed based on standard online machine translation tools and dictionaries augmented with domain-specific resources gathered from the WWW and elsewhere. In this section we briefly introduce the relevant details of MILOS and Lucene. Since this paper focuses on text search within MultiMatch, we do not describe the multimedia features of the MultiMatch system.

2.1 MILOS: Multimedia Repository

MILOS (Multimedia dIgital Library for On-line Search) is a repository system conceived to support the distributed storage and retrieval of multimedia objects. This Multimedia Content Management System (MCMS) is able to manage not only structured data, as in databases, but also textual data (using information retrieval technologies), semi-structured data (typically in XML), mixed-mode data, and multimedia data. In MultiMatch, we use MILOS as a metadata repository to enable querying on the structure of the data stored.

MILOS has a three-tier architecture composed of three main components:

1. the XML Search Engine (XMLSE) component which manages the metadata;
2. the MultiMedia Server (MMS) component which manages the documents; and
3. the MultiMedia Digital Library service (MMDLS) component MMDLS which provides application developers with a uniform and integrated way of accessing MMS and XMLSE.

Each of these components is implemented using solutions providing flexibility, scalability, and efficiency.

2.1.1 XMLSE

XMLSE is an enhanced native XML database/repository system with special features for digital library applications. This is especially justified by the well known and accepted advantages of representing metadata as XML documents. Metadata represented with XML can have arbitrary complex structures, which allows it to handle with complex metadata schemas, and can easily be exported and imported. Our XML database can store and retrieve any valid XML document. No metadata schema or XML schema definition is needed before inserting an XML document, except optional index definitions for performance boosting. Once an arbitrary XML document has been inserted in the database it can be immediately retrieved using XQuery. This allows digital library applications to use arbitrary (XML encoded) metadata schemas

[1] www.multimatch.org

and to deal with heterogeneous metadata, without any constraint on schema design and/or overhead due to metadata translation. Thus, the native XML database/repository system is simpler than a general purpose XML database system, but offers significant improvements in specific areas: it supports standard XML query languages such as XPath and XQuery, and offers advanced search and indexing functionality on arbitrary XML documents. It supports high performance search and retrieval on heavily structured XML documents, relying on specific index structures.

Moreover XMLSE provides the possibility of using particular indexes. For example, using the configuration file of XMLSE the system administrator can associate the `<abstract>` elements of a document with a full-text index and to the MPEG-7 `<VisualDescriptor>` elements can be associated with a similarity search index. XMLSE uses Apache Lucene[2] to provide partial (or approximate) text string matching, effectively providing information retrieval functionality within MILOS. This allows XMLSE to use the ranked searching and wildcard queries of Lucene to solve queries like "find all the articles whose title contains the word XML" and so on. This application allows users to interrogate the dataset combining full text, and exact or partial match search. For example the user can look for documents whose `<metadata>` element contains the word "Switzerland". MILOS generates and submits to XMLSE the following XQuery query:

```
for $a in /document where
    $a//metadata ~ 'Switzerland'
return
    <result>
        {$a//title}, {$a//author}
    </result>
```

The query will return a list of results which consist of the title and author of all documents whose metadata contains the term "Switzerland".

2.2 Lucene

Full text search in MILOS is provided by using Lucene as a plugin. Ranked retrieval uses the standard $tf \times idf$ vector-space method provided in Lucene (Hatcher and Gospodnetic, 2004). Lucene also provides additional functionality to improve retrieval effectiveness by providing various query expansion services using techniques such as relevance feedback, although these are not used in the current investigation. Documents and search requests are preprocessed to remove stop words and stemming is applied using the standard resources supplied with Lucene.

3 Evaluation Task

The MultiMatch system will enable search from a number of CH repository sources including formally published documents, images and video, as well as material gathered from relevant WWW sources. However, in order to explore metadata search issues and evaluate our approaches to addressing related translation problems, a test collection including sample user search topics and relevance judgements is required. Since MultiMatch does not yet have such a collection available, for our current experiments we made use of the data provided for the CLEF 2006 CL-SR track (Oard et al., 2006).

The document collection comprises 8104 English documents that are manually-determined topically-coherent segments taken from 272 interviews with Holocaust survivors, witnesses and rescuers, totaling 589 hours of speech. Several automatic speech recognition transcripts are available for these interviews. However, for this study we focus on the metadata fields provided for each document: two sets of 20 automatically assigned keywords (`<AUTOKEYWORD2004A1>` and `<AUTOKEYWORD2004A2>`) determined using two different kNN classifiers, denoted by AKW1 and AKW2 respectively; a set of a varying number of manually-assigned keywords (`<MANUALKEYWORD>`), denoted by MKW; and a manual three-sentence summary written by an expert in the field (`<SUMMARY>`), denoted by SUMMARY.

The CLEF collection includes a set of 33 search topics in standard TREC format created in English, and translated into Czech, German, French, and Spanish by native speakers. Since we wish to investigate topics with minimal redundancy, for our experiments we used only the topic Title fields as our search request. Relevance judgments were generated using a search guided procedure and standard pooling methods were also provided with the collec-

[2] http://lucene.apache.org

tion. Full details of the this collection can be found in (Oard et al., 2006; White et al., 2005).

To explore metadata field search, we used various methods, described in the next section, to automatically translate the French, German, and Spanish topics into English[3].

4 Translation Techniques

The MultiMatch translation resources are based on the WorldLingo machine translation system augmented with domain-specific dictionary resources gathered automatically from the WWW. This section briefly reviews WorldLingo[4], and then describes construction of our augmentation translation lexicons and their application for query translation in multilingual metadata search.

4.1 Machine translation system

There are a number of commercial machine translation systems currently available. After evaluation of several candidate systems, WorldLingo was selected for the MultiMatch project because it generally gives good translation well between the English, Spanish, Italian, and Dutch, languages relevant to the Multimatch project[5]. In addition, it provides a useful API that can be used to translate queries on the fly via HTTP transfer protocol. The usefulness of such a system is that it can be integrated into any application and present translations in real-time. It allows users to select the source/target languages and specify the text format (e.g. plain text file or html file) of their input files. The WorldLingo translation system also provides various domain-specific dictionaries that can be integrated with translation system. A particularly useful feature of WorldLingo with respect to for MultiMatch, and potentially applications within CH in general, is that to improve the quality of translations, additional locally developed customized dictionaries can be uploaded. This enables the WorldLingo dictionaries to be extended to contain special terms for a specific domain.

4.2 Translation lexicon construction

To extend the standard dictionaries provided with WorldLingo we used the current online *wikipedia*. Wikipedia[6] is the largest multilingual free-content encyclopedia on the Internet. As of March 21 2007, there are approximately 6.8 million articles written in 250 languages available on the web, according to *Wiki Stats*[7]. Wikipedia is structured as an interconnected network of articles. Each wikipedia page can hyperlink to several other wikipedia pages. Wikipedia page titles in one language are also linked to a multilingual database of corresponding terms. Unlike the web, most hyperlinks in wikipedia have a more consistent and semantically meaningful interpretation and purpose. The comprehensive literature review presented by Adafre and Rijke (2005) describes the link structure of wikipedia. As a multilingual hypertext medium, wikipedia presents a valuable new source of translation information. Recently, researchers have proposed techniques to exploit this opportunity. Adafre and Rijke (2006) developed a technique to identify similar text across multiple languages in wikipedia using page content-based features. Boumaet et al. (2006) utilized wikipedia for term recognition and translation in order to enhance multilingual question answering systems. Declerck et al. (2006) showed how the wikipedia resource can be used to support the supervised translation of ontology labels.

In order to improve the effectiveness of multilingual metadata search, we mine wikipedia pages as a translation source and construct translation lexicons that can be used to reduce the errors introduced by unknown terms (single words and multiword phrases) during query translation. The major difference in our proposal is that the translations are extracted on the basis of hyperlinks, meta keywords, and emphasized concepts — e.g. anchor text, boldface text, italics text, and text within special punctuation marks — appearing in the first paragraph of wikipedia articles.

Meta keywords Wikipedia pages typically contain meta keywords assigned by page editors. This meta keywords can be used to assist in the iden-

[3] Due to a lack of translation resources, we did not use the Czech translations in these experiments

[4] `http://www.worldlingo.com/`

[5] Additionally, it translates well between French and English, as used in this paper

[6] `http://www.wikipedia.org/`

[7] `http://s23.org/wikistats/wikipedias_html.php?sort=good_desc`

tification of the associated terms on the same topic.

Emphasized concepts In common with standard summarization studies, we observed that the first paragraph of a wikipedia document is usually a concise introduction to the article. Thus, concepts emphasized in the introductory section are likely to be semantically related to the title of the page.

In our study we seek to use these features from multilingual wikipedia pages to compile a domain-specific word and phrase translation lexicon. Our method in using this data is to augment the queries with topically related terms in the document language through a process of *post-translation query expansion*. This procedure was performed as follows:

1. An English vocabulary for the domain of the test collection was constructed by performing a limited crawl of the English wikipedia[8], Category:World War II. This category contains links to pages and subcategories concerning events, persons, places, and organizations pertaining to war crimes or crimes against humanity especially during WWII. It should be noted that this process was neither an exhaustive crawl nor a focused crawl. The purpose of our current study is to explore the effect of translation expansion on metadata retrieval effectiveness. In total, we collected 7431 English web pages.

2. For each English wikipedia page, we extracted its hyperlinks to German, Spanish, and French. The basename of each hyperlink is considered as a term (single word or multi-word phrase that should be translated as a unit). This provided a total of 4446 German terms, 3338 Spanish terms, and 4062 French terms. As an alternative way of collecting terms in German, Spanish, and French, we are able to crawl the wikipedia in a specific language. However, a page with no link pointing to its English counterpart will not provide enough translation information.

RUN ID	**Augmented lexicon using all terms appearing in the following fields**		
	Title terms	Meta keywords	Emphasized concepts
RUN$_{mt+t}$	$\surd$	$\times$	$\times$
RUN$_{mt+m}$	$\times$	$\surd$	$\times$
RUN$_{mt+c}$	$\times$	$\times$	$\surd$
RUN$_{mt+m+c}$	$\times$	$\surd$	$\surd$

Table 1: Run descriptions.

3. For each of the German, Spanish, and French terms obtained, we used the title term, the meta keywords, and the emphasized concepts obtained from the same English wikipedia page as its potential translations.

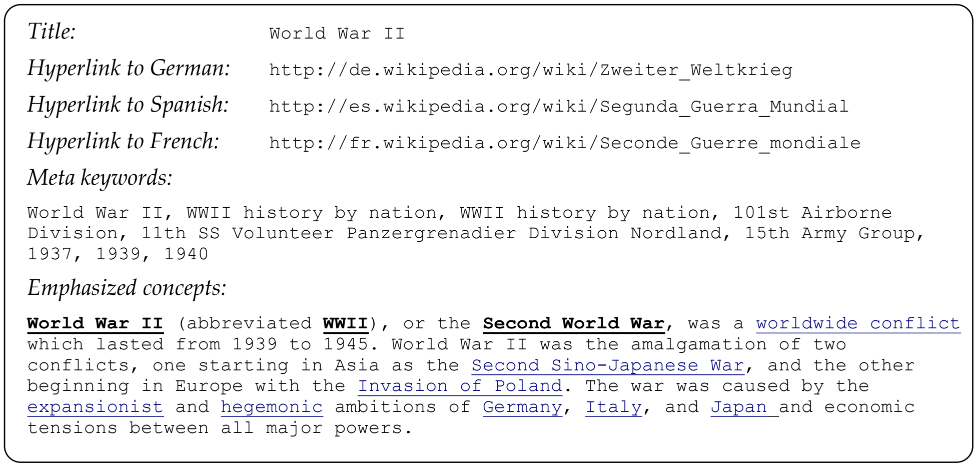
For example, consider an English page titled as “World War II”[9]. The title term, the meta keywords, the emphasized concepts in English, and the hyperlinks (to German, Spanish, and French) associated are shown in Figure 1. We first extract the basenames “Zweiter Weltkrieg” (in German), “Segunda Guerra Mundial” (in Spanish), and “Seconde Guerre mondiale” (in French) using the hyperlink feature. To translate these terms into English, we replace them using the English title term, all the English meta keywords and/or all the English emphasized concepts occurring in the same English wikipedia page. This is a straightforward approach to automatic post-translation query expansion by using meta keywords and/or emphasized concepts as expanded terms. The effects of the features described above are investigated in this work, both separately and in combination, as shown in Table 1,

5 Experimental Setup

In this section we outline the design of our experiments. We established a monolingual reference (RUN$_{mono}$) against which we can measure multilingual retrieval effectiveness. To provide a baseline for our multilingual results, we used the standard WorldLingo to translate the queries (RUN$_{mt}$). We then tested the MT integrated with different lexicons compiled using wikipedia. Results of these experiments, shown in Table 1, enable us gauge the effect of each of our additional translation resources generated using wikipedia.

[8] en.wikipedia.org

[9] http://en.wikipedia.org/wiki/World_War_II



Title: World War II

Hyperlink to German: http://de.wikipedia.org/wiki/Zweiter_Weltkrieg

Hyperlink to Spanish: http://es.wikipedia.org/wiki/Segunda_Guerra_Mundial

Hyperlink to French: http://fr.wikipedia.org/wiki/Seconde_Guerre_mondiale

Meta keywords:

World War II, WWII history by nation, WWII history by nation, 101st Airborne Division, 11th SS Volunteer Panzergrenadier Division Nordland, 15th Army Group, 1937, 1939, 1940

Emphasized concepts:

World War II (abbreviated **WWII**), or the **Second World War**, was a worldwide conflict which lasted from 1939 to 1945. World War II was the amalgamation of two conflicts, one starting in Asia as the Second Sino-Japanese War, and the other beginning in Europe with the Invasion of Poland. The war was caused by the expansionist and hegemonic ambitions of Germany, Italy, and Japan and economic tensions between all major powers.

Figure 1: Title, hyperlinks, meta keywords, and emphasized concepts (underlined terms) extracted from the English wikipedia page `http://en.wikipedia.org/wiki/World_War_II`.

The focus of this paper is not on optimising absolute retrieval performance, but rather to explore the usefulness of our translation resources. Thus we do not apply retrieval enhancement techniques such as relevance feedback which would make it more difficult to observe the impact of differences in behaviour of the translation resources. The experiments use the SUMMARY field, as an example of concise natural language descriptions of CH objects; and the AKW1 and AKW2 fields as examples of automatically assigned keyword labels without linguistic structure, with the MKW field providing similar manually assigned for keyword labels. Retrieval effectiveness is evaluated using standard TREC mean average precision (MAP) and the precision at rank 10 (P@10).

6 Results and Discussion

The results of our query translation experiments are shown in Table 2, 3, 4, and 5. For search using SUMMARY and MKW fields, the lexicon compiled using title terms provided an improvement of $7 \sim 9\%$, $7 \sim 19\%$, and $20 \sim 30\%$, in German–English, Spanish–English, and French–English retrieval task, respectively. These improvements are statistically significant at the 95% confidence level, and emphasize the importance of a good domain-specific translation lexicon.

The addition of meta keywords or emphasized concepts also improves results in most cases relative to the RUN$_{mt}$ results. However, we can see that retrieval performance degrades when the query is expanded to contain terms from both meta keywords and emphasized concepts. This occurs despite the fact that the additional terms are often closely related to the original query terms. While the addition of all these terms generally produces an increase in the number of retrieved documents, there is little or no increase in the number of relevant documents retrieved, and the combination of the two sets of terms in the queries leads on average to a slight reduce in the rank of relevant documents.

The results show that RUN$_{mt+t}$ runs provide the best results when averaged across a query set. However, when analysed at the level of individual queries different combined translation resources are more effective for different queries, examples of this effect are shown in Table 6. This suggests that it may be possible to develop a more sophisticated translation expansion methods to select the best terms from different lexicons. At the very least, it should be possible to use “context-sensitive filtering” and “combination of evidence” (Smets, 1990) approaches to improve the overall translation quality. We plan to explore this method in further investigations.

7 Conclusion and Future Work

This paper reports experiments with techniques developed for domain-specific lexicon construction to facilitate multilingual metadata search for a CH re-

RUN ID	German–English		Spanish–English		French–English	
	MAP	P@10	MAP	P@10	MAP	P@10
RUN$_{mt}$	0.0750	0.1233	0.0756	0.1250	0.0652	0.1152
RUN$_{mt+t}$	**0.0815**	**0.1516**	**0.0899**	**0.1545**	**0.0783**	**0.1333**
RUN$_{mt+m}$	0.0775	0.1266	0.0797	0.1364	0.0690	0.1030
RUN$_{mt+c}$	0.0669	0.1000	0.0793	0.1303	0.0770	0.1152
RUN$_{mt+m+c}$	0.0668	0.0968	0.0737	0.1212	0.0646	0.0970
RUN$_{mono}$	MAP = 0.1049			P@10 = 0.1818		

Table 2: Results for SUMMARY field search. (RUN$_{mt+t}$ run provides the best results in all cases.)

RUN ID	German–English		French–English		Spanish–English	
	MAP	P@10	MAP	P@10	MAP	P@10
RUN$_{mt}$	0.1158	0.1750	0.1000	0.1677	0.0903	0.1677
RUN$_{mt+t}$	**0.1235**	**0.2100**	**0.1071**	**0.2031**	**0.1171**	**0.2194**
RUN$_{mt+m}$	0.1171	0.1393	0.1023	0.2000	0.0983	0.1903
RUN$_{mt+c}$	0.1084	0.1500	0.0958	0.1636	0.1089	0.1667
RUN$_{mt+m+c}$	0.1069	0.1600	0.0947	0.1727	0.0940	0.1742
RUN$_{mono}$	MAP = 0.1596			P@10 = 0.2812		

Table 3: Results for MKW field search. (RUN$_{mt+t}$ run provides the best results in all cases.)

RUN ID	German–English		French–English		Spanish–English	
	MAP	P@10	MAP	P@10	MAP	P@10
RUN$_{mt}$	0.0264	0.0731	0.0247	0.0548	0.0316	0.0767
RUN$_{mt+t}$	**0.0273**	**0.0828**	**0.0274**	**0.0656**	**0.0406**	**0.0867**
RUN$_{mt+m}$	0.0268	0.0633	0.0258	0.0606	0.0357	0.0613
RUN$_{mt+c}$	0.0266	0.0667	0.0266	0.0636	0.0383	0.0839
RUN$_{mt+m+c}$	0.0259	0.0633	0.0260	0.0606	0.0328	0.0677
RUN$_{mono}$	MAP = 0.0388			P@10 = 0.1000		

Table 4: Results for AKW1 field search. (RUN$_{mt+t}$ run provides the best results in all cases.)

RUN ID	German–English		French–English		Spanish–English	
	MAP	P@10	MAP	P@10	MAP	P@10
RUN$_{mt}$	0.0279	0.0375	0.0347	0.0625	0.0205	0.0483
RUN$_{mt+t}$	0.0279	**0.0481**	0.0351	**0.0680**	0.0238	0.0433
RUN$_{mt+m}$	**0.0302**	0.0448	**0.0361**	0.0556	0.0223	0.0484
RUN$_{mt+c}$	0.0275	0.0414	0.0332	0.0593	0.0268	0.0548
RUN$_{mt+m+c}$	0.0299	0.0448	0.0351	0.0536	**0.0273**	**0.0581**
RUN$_{mono}$	MAP = 0.0420			P@10 = 0.0821		

Table 5: Results for AKW2 field search. (The best results are in bold.)

trieval tasks. The results show that our techniques can provide a statistically significant improvement in the retrieval effectiveness. Using a tailored translation lexicon enables us to achieve $(77\%, 78\%)$, $(86\%, 67\%)$ and $(75\%, 63\%)$ of the monolingual effectiveness in German–English, Spanish–English, and French–English multilingual metadata SUMMARY, MKW field search tasks. In addition, the multilingual wikipedia proved to be a rich resource of translations for domain-specific terms.

Intuitively, document translation is superior to query translation. Documents provide more context for resolving ambiguities (Oard, 1998) and the translation of source documents into all the languages supported by the retrieval system effectively reduces CLIR to a monolingual IR task. Furthermore, it has the added advantage that document content is accessible to users in their native languages. In our future work, we will compare the effectiveness of these two approaches to metadata search in a multilingual environment.

		MT	**Augmented lexicon using all terms appearing in the following fields**			
	Query ID	WorldLingo	Title terms	Meta keyword	Emphasized concepts	Meta keyword + Emphasized concepts
German–English	*1133*	0.6000	0.6000	0.6195	0.6092	**0.6400**
	1325	0.0000	0.0003	**0.0020**	**0.0020**	0.0018
	1623	0.2210	0.2210	**0.3203**	0.0450	0.0763
	3007	0.0000	0.0003	0.0025	0.0047	**0.0054**
	3012	0.0087	0.0087	0.0073	0.0073	**0.0097**
	3025	0.0052	0.0052	**0.0060**	0.0052	**0.0060**
Spanish–English	*1623*	0.0063	0.0063	**0.1014**	0.0084	0.0334
	3007	0.0000	0.0004	0.0028	0.0048	**0.0057**
French–English	*1133*	0.6000	0.6000	0.6195	0.6092	**0.6400**
	1345	0.0600	0.0667	**0.0809**	0.0495	0.0420
	1623	0.0750	0.0798	**0.1810**	0.0228	0.0528
	3005	0.0200	0.0232	0.0226	**0.2709**	0.1063
	3007	0.0003	0.0003	0.0024	0.0025	**0.0037**
	3025	0.0173	0.0173	**0.0178**	0.0173	**0.0178**

Table 6: Examples of MAP values obtained using different translation combinations for SUMMARY field search. (The best results are in bold.)

Acknowledgement

Work partially supported by European Community under the Information Society Technologies (IST) programme of the 6th FP for RTD - project MultiMATCH contract IST- 033104. The authors are solely responsible for the content of this paper. It does not represent the opinion of the European Community, and the European Community is not responsible for any use that might be made of data appearing therein.

References

Sisay Fissaha Adafre and Maarten de Rijke. 2005. Discovering missing links in wikipedia. In *Proceedings of the 3rd international workshop on Link discovery*, pages 90–97, Chicago, Illinois. ACM Press.

Sisay Fissaha Adafre and Maarten de Rijke. 2006. Finding similar sentences across multiple languages in wikipedia. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 62–69, Trento, Italy.

Giuseppe Amato, Claudio Gennaro, Fausto Rabitti, and Pasquale Savino. 2004. Milos: A multimedia content management system for digital library applications. In *Proceedings of the 8th European Conference on Research and Advanced Technology for Digital Libraries*, Lecture Notes in Computer Science, pages 14–25. Springer-Verlag.

Gosse Bouma, Ismail Fahmi, Jori Mur, Gertjan van Noord, Lonneke van der Plas, and Jorg Tiedemann. 2006. The university of groningen at QA@CLEF 2006 using syntactic knowledge for QA. In *Working Notes for the Cross Language Evaluation Forum 2006 Workshop*, Alicante, Spain.

Thierry Declerck, Asunciòn Gòmez Pèrez, Ovidiu Vela, Zeno Gantner, and David Manzano-Macho. 2006. Multilingual lexical semantic resources for ontology translation. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, Genoa, Italy.

Erik Hatcher and Otis Gospodnetic. 2004. *Lucene in Action (In Action series)*. Manning Publications Co., Greenwich, CT, USA.

Henning Müller, Wolfgang Müller, and David McG. Squire. 2001. Automated benchmarking in content-based image retrieval. In *Proceedings of the 2001 IEEE International Conference on Multimedia and Expo*, Tokyo, Japan. IEEE Computer Society.

Douglas W. Oard, Jianqiang Wang, Gareth J. F. Jones, Ryen W. White, Pavel Pecina, Dagobert Soergel, Xiaoli Huang, and Izhak Shafran. 2006. Overview of the CLEF-2006 cross-language speech retrieval track. In *Working Notes for the Cross Language Evaluation Forum 2006 Workshop*, Alicante, Spain.

Douglas W. Oard. 1998. A comparative study of query and document translation for cross-language information retrieval. In *Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas on Machine Translation and the Information Soup*, pages 472–483, London, UK. Springer-Verlag.

Philippe Smets. 1990. The combination of evidence in the transferable belief model. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 12(5):447–458.

Ryen W. White, Douglas W. Oard, Gareth J. F. Jones, Dagobert Soergel, and Xiaoli Huang. 2005. Overview of the CLEF-2005 cross-language speech retrievaltrack. In Carol Peters, Fredric C. Gey, Julio Gonzalo, Henning Müller, Gareth J. F. Jones, Michael Kluck, Bernardo Magnini, and Maarten de Rijke, editors, *CLEF*, volume 4022 of *Lecture Notes in Computer Science*, pages 744–759. Springer.