Combining Query Translation and Document Translation in Cross-Language Retrieval

Aitao Chen
School of Information Management and Systems
University of California at Berkeley, CA 94720-4600, USA
aitao@sims.berkeley.edu

Fredric C. Gey
UC Data Archive & Technical Assistance (UC DATA)
University of California at Berkeley, CA 94720-5100, USA
gey@ucdata.berkeley.edu

Abstract

This paper describes monolingual, bilingual, and multilingual retrieval experiments using the CLEF 2003 test collection. The paper compares query translation-based multilingual retrieval with document translation-based multilingual retrieval where the documents are translated into the query language by translating the document words individually using machine translation systems or statistical translation lexicons derived from parallel texts. The multilingual retrieval results show that document translation-based retrieval is slightly better than the query translation-based retrieval on the CLEF 2003 test collection. Furthermore, combining query translation and document translation in multilingual retrieval achieves even better performance.

1 Introduction

One focus of this paper is on the use of parallel texts for creating statistical translation lexicons to support cross-language retrieval (bilingual or multilingual) and for creating stemmers to support both monolingual and cross-language retrieval. Another focus is on evaluating the effectiveness of translating documents by translating document words individually using translation lexicons created from machine translation (MT) systems or from parallel texts, and the effectiveness of combining query translation and document translation in cross-language retrieval. At CLEF 2003, we participated in the *monolingual*, *bilingual*, *multilingual-4*, and *multilingual-8* retrieval tasks. The retrieval system we used at CLEF 2003 is described in detail in our CLEF 2002 paper [5].

2 New Resources

This section describes the new resources that we developed and were used in our CLEF 2003 runs.

2.1 Stoplists

We gleaned 436 Swedish stopwords from a Swedish grammar book written in English [11], and created a stoplist of 567 Finnish stopwords gathered from two Finnish textbooks written in English [2, 3]. A foreign language textbook written in English usually gives the English translations of the foreign words mentioned in the textbook. We included in a stoplist those foreign words whose English translations are English stopwords. The stoplist used in indexing Finnish topics

includes additional stopwords found in the previous CLEF Finnish topics. The stoplist used in indexing Swedish topics also includes additional stopwords found in the previous CLEF Swedish topics. The CLEF 2003 topics were not used for constructing stoplists.

2.2 Base lexicons

We developed a Finnish and a Swedish base lexicons for splitting Finnish and Swedish compounds. The base lexicons were largely automatically generated. A base lexicon should include all and only the words and their variants that are not compounds. Our approach to creating a base lexicon is to start with a wordlist and then remove compounds in the wordlist. The remaining words make up the base lexicon. We first combined the Finnish words in the Finnish document collection and the Finnish words found in the Finnish version of the ispell spelling checker. We removed the words of 10 or more characters that can be decomposed into two or more component words of at least 5 characters. The decompounding procedure used was described in [5]. To decompose the long words in the initial Finnish wordlist using the decompounding procedure, we need a base lexicon. The base lexicon for splitting compounds in the initial Finnish wordlist consists of the words that are at least 7-character long, the 6-character words that occur at least 50 times in the Finnish collection, and the 5-character words that occur at least 100 times in the Finnish collection. This base lexicon was used to split compounds in the initial wordlist, and the words that can be decomposed were removed from the initial wordlist. We also removed from the initial wordlist all the words that are one to four characters long. Lastly we manually removed some of the remaining long words that look like compounds to us who do not know Finnish or Swedish. Our Finnish base lexicon consists of the remaining words on the wordlist. The Swedish base lexicon was created in the same way. Both the wordlist for the Finnish spelling checker and the wordlist for the Swedish spelling checker were downloaded from http://packages.debian.org/unstable/text/ispell.html.

2.3 Statistical translation lexicons

We downloaded the Official Journal of the European Union [1] for 1998, 1999, the first four months of 2000, and 2002 minus July and August. The documents are in PDF format and are available in eleven languages. We downloaded the PDF documents in eight languages: Dutch, English, Finnish, French, German, Italian, Spanish and Swedish. These are the same languages involved in the multilingual-8 retrieval task. The year 2002 version of the *ps2ascii* conversion program on a Linux machine was used to convert the source PDF files into text files. The converted texts are noisy in that many words were strung together. For instance, sometimes the English articles (i.e, a, an, the), pronouns, and prepositions are combined with the preceding or the following word. Sometimes two content words are concatenated, resulting in many compounds in all eight languages. In an one-page English document, in the converted texts, we found words like *culturalmix*, *andthe*, *xenophobiaand*, *inproclaiming*, *allhelped*, and more.

The original texts in PDF format are presented in two-column format. When a word is broken into two parts, a dash character is appended to the first part at the end of a line, and the second part starts on a new line. After the PDF files are converted into texts, there are many words with a dash inserted in the middle. For instance, the German word *Be-ämpfung* (the first part *Be-* appears at the end of a line, while the second part *ämpfung* starts on the following line) was converted into *Be- a"mpfung*. We did not, but should have, removed the dash and the additional space after the dash character.

Later we found another PDF to texts conversion program named *pdftotext*, also available on a Linux machine. The texts converted from PDF files using *pdftotext* look much cleaner. The same word *Be-ämpfung* was converted into *Bekämpfung* using *pdftotext*. Not only the diacritic mark was retained in its original form, but also the dash character inserted into the word was taken out by *pdftotext*.

After the diacritic marks were restored in the converted texts, the text files were aligned at the line level (a line may contain a long paragraph), then at the sentence level, after splitting a line into sentences, using a length-based alignment program [10]. Because many words in the

converted texts were joined together after conversion, we used our decompounding procedure to split compounds (including the ones created by joining two or more consecutive words in the conversion process) into their component words. About 316,000 unique compounds in the English texts were split into their component words. Our decompounding procedure does not split compounds into words that are three or fewer characters long, so we had to write special programs to split compounds that contain short words like the English article *an* or preposition *of*.

From the sentence-aligned parallel texts, we created six statistical translation lexicons using the GIZA++ toolkit [13]: 1) English to Dutch; 2) English to Finnish; 3) English to Swedish; 4) Dutch to English; 5) Finnish to English; and 6) Swedish to English. We were unable to use toolkit to create statistical translation lexicons between Italian and Spanish, German and Italian, French and Dutch, and Finish and German because of the large vocabulary sizes in these languages and limited memory on our machine (2GB). To support bilingual retrieval, we created three translation dictionaries from the sentence-aligned parallel texts based on statistical association, the maximum likelihood ratio test statistic [9]. The three dictionaries are the Italian to Spanish, German to Italian, and Finnish to German translation lexicons. We did not have adequate time to create a French to Dutch dictionary before the results were due. The procedure for creating translation lexicons from sentence-aligned parallel texts using statistical association measures is described in detail in [7].

2.4 Stemmers

In [6] we present an algorithm for automatically generating an Arabic stemmer from an Arabic to English machine translation system and an English stemmer. The words in the Arabic documents are translated individually into English, then the English translations are conflated using an English stemmer. All the Arabic words whose English translations share the same English stem are grouped together to form one cluster. In stemming, all the Arabic words in the same cluster are conflated to the same word. The words in a cluster are generally either morphologically or semantically related.

We developed one Finnish stemmer and one Swedish stemmer from the English/Finnish and English/Swedish parallel texts based on the same idea. As mentioned in section 2.3, we generated one Finnish to English and one Swedish to English translation lexicons from the parallel texts. For a Finnish word, we chose the English translation of the highest translation probability, i.e., the most likely English translation, as its translation. So every Finnish word in the parallel texts has just one English translation. The Finnish words in the parallel texts that have the same English translation were grouped together to form a cluster. The English texts were stemmed using an English morphological analyzer [8] before the Finnish/English texts were fed into GIZA++ for creating statistical translation lexicon. The English morphological stemmer maps plural nouns into the singular form, verbs into the base form, and adjectives in comparative or superlative into the positive form. The Finnish words in the same cluster were conflated into the same stem in stemming. The Swedish stemmer was generated in the same way. For example, the Swedish words *diamanten*, *diamanterna*, *diamanteroch*, *diamanthande*, *diamantrika*, *diamantsek*, *diamanter* and *diamant* were grouped into the same cluster, which also includes some other words, since the most likely English translations of these words have the same English stem *diamond* according to the statistical translation lexicon generated from the Swedish/English parallel texts. In stemming, these Swedish words were conflated into the same stem. The cluster of Finnish words whose most likely English translations share the same English stem *diamond* includes *timantit*, *timanteista*, *timanttialan*, *timanttien*, *timantteja*, *timantti*, *timanttierä*, *timantin*, *timanttialanyritysten*, *timanttiteol*, *veritimanteistä*, and more. All these words were conflated to the same Finnish stem in stemming.

2.5 English spelling normalizer

The CLEF 2003 English collection contains newspaper articles published in the U.S. and Britain. The British English spellings were changed into the American English spellings in indexing the English documents and English topics. We used a set of rules to extract the words that have both spellings in the collection and built a table that maps the British spellings to the American spellings. The table has about 2,700 entries, each entries mapping one British English word into the corresponding American English word. For example, the word *watercolour* is changed into *watercolor*, *finalised* into *finalized*, *paediatrician* into *pediatrician*, and *offences* into *offenses*.

3 Fast document translation

To translate a large collection of documents from a source language to a target language using a machine translation system can be computationally intensive and may take a long time. In this section we present an approximate but fast approach to translating source documents into a target language using bilingual lexicons derived from machine translation systems or parallel texts. We first collect all the unique words in the source documents, then translate the source words individually into the target language using a machine translation system. Once we have the translations of all the source words, we can translate a source document into the target language by replacing the words in the source document with their translations in the target language. The translation is only approximate, but very fast. It is approximate since the same source word is always translated into the same target word. When a source word has multiple meanings under different contexts in the source documents, the translations of the source word may not be the same in the target language. For example, in translating English into French, the English word *race* is translated into the French word *race*. However, the English word *race* is polysemous, it could mean *human race* or *race in sports*. When it means *race* in sports, the appropriate French translation is not *race*, but *course*. For multilingual retrieval, one can translate the document collections into the topic language using this method if one can find a MT system capable of translating documents into the topic language. When MT systems are not available but parallel texts are, one can derive a bilingual lexicon from the parallel texts, and then use the bilingual lexicon to translate the source documents into the target language by translating the documents words individually. If neither MT systems nor parallel texts are available, one can still translate documents word-by-word using bilingual dictionaries as was done in [12]. When a multilingual document collection is translated into the topic language, one can index the translated documents together and search the queries directly against the translated document collection. This approach to multilingual retrieval does not require any merging of individual ranked lists of retrieved documents as noted in [4].

4 Test collection

The document collection for the multilingual-8 IR task consists of 190,604 Dutch, 169,477 English, 55,344 Finnish, 129,806 French, 294,809 German, 157,558 Italian, 454,045 Spanish, and 142,819 Swedish newswire and newspaper articles published in 1994 and 1995. There are 60 test topics available in many languages. The multilingual-4 IR task uses the English, French, German, and Spanish documents.

5 Experimental results

All retrieval runs reported in this paper used only the *title* and *description* fields in the topics. The IDs and average precision values of the official runs are presented in bold face, other runs are unofficial ones.

5.1 Monolingual retrieval experiments

This section presents the results of our monolingual retrieval runs on eight languages. Stopwords were removed from both documents and topics, the remaining words were stemmed using either the Snowball stemmers [14] developed by Martin Porter or the automatically generated stemmers from parallel texts. For Dutch, Finnish, German and Swedish monolingual runs, the compounds were split, whenever possible, into their component words before stemming, and only their component words were retained in document and topic indexes. A compound is split into its component words only when all the component words are present in the base lexicon, otherwise it is not split and is retained as a compound in the indexes. The same decompounding procedure was applied in all four languages, using language-specific base lexicons. For automatic query expansion, 10 terms from the top-ranked 10 documents after the initial search were combined with the original query. All the monolingual runs included automatic query expansion via blind relevance feedback. Table 1 presents the monolingual retrieval results for eight languages. Column 3 gives the number of topics for each language that having at least one relevant document. The average precision values presented in the table were computed with respect to only the topics having at least one relevant document. For all monolingual runs, only the *title* and *desc* fields in the topics were used as shown in column 4. Columns 5 and 7 present the overall recall values without and with query expansion, respectively; and columns 6 and 8 present the average precision values without and with query expansion, respectively. The last column labeled *change* shows the improvement of average precision with query expansion over without it. As Table 1 shows, query expansion increased the average precision of the monolingual runs for all eight languages, the improvement ranging from 6.76% for French to 16.76% for Italian. The Finnish monolingual run **bkmonofi2**

		number	topic	without expansion		with expansion		
run id	language	topics	fields	recall	precision	recall	precision	change
bkmonoen1	English	54	TD	980/1006	0.5011	992/1006	0.5496	9.68%
bkmononl1	Dutch	56	TD	1484/1577	0.4955	1519/1577	**0.5304**	7.04%
bkmonofi1	Finnish	45	TD	462/483	0.4972	476/483	**0.5633**	13.29%
bkmonofi2	Finnish	45	TD	457/483	0.4626	472/483	**0.4962**	7.26%
bkmonofr1	French	52	TD	917/946	0.4986	923/946	**0.5323**	6.76%
bkmonode1	German	56	TD	1712/1825	0.5111	1767/1825	**0.5678**	11.09%
bkmonoit1	Italian	51	TD	770/809	0.4809	801/809	**0.5615**	16.76%
bkmonoes1	Spanish	57	TD	2214/2368	0.4556	2301/2368	**0.5091**	11.74%
bkmonosv1	Swedish	54	TD	959/1006	0.4727	987/1006	**0.5465**	15.61%
bkmonosv2	Swedish	54	TD	894/1006	0.4404	953/1006	**0.4982**	13.12%

Table 1: Monolingual IR performance.

used the Finnish statistical stemmer generated from the English-Finnish parallel texts, and the Swedish monolingual run **bkmonosv2** used the Swedish statistical stemmer generated from the English-Swedish parallel texts. All other runs presented in Table 1 used the Snowball stemmers, including the Muscat stemmers.

Table 2 presents the performances of Dutch, Finnish, German, and Swedish monolingual retrieval with different indexing features. The column labeled *stoplist* gives the average precision values of the monolingual runs when only the stopwords were removed. The last seven columns present the average precision values of the monolingual retrieval with additional indexing features. Stopwords were removed in all monolingual runs presented in Table 2. Without stemming and query expansion, decompounding alone improved the average precision from 7.10% for Finnish to 30.59% for German in comparison to the average precision when only stopwords were removed. With decompounding, stemming and query expansion, the average precision increased from 22.16% for Dutch to 52.35% for German. Note that decompounding substantially increased the German monolingual retrieval performance, so did stemming to Finnish monolingual retrieval performance. The Snowball stemmers were used, when stemming was included, in the monolingual runs pre-

sented in Table 2. Without decompounding and query expansion, the statistical Finnish stemmer increased the average precision of the Finnish monolingual retrieval from 0.3801 to 0.4304, an 13.21% increase; and the statistical Swedish stemmer increased the Swedish monolingual retrieval average precision from 0.3630 to 0.3844, an 5.90% increase. Both statistical stemmers were not as effective as the manually constructed Snowball stemmers.

			additional features						
language	topic fields	stoplist	decomp	stem	expan	decomp stem	decomp expan	stem expan	decomp stem expan
Dutch	TD	0.4342	0.4673 7.62%	0.4480 3.18%	0.4744 9.26%	0.4955 14.12%	0.5126 18.06%	0.4962 14.28%	**0.5304** 22.16%
German	TD	0.3727	0.4867 30.59%	0.4220 13.23%	0.4294 15.21%	0.5111 37.13%	0.5473 46.85%	0.4804 28.90%	**0.5678** 52.35%
Finnish	TD	0.3801	0.4071 7.10%	0.4974 30.86%	0.4204 10.60%	0.4972 30.81%	0.4469 17.57%	0.5541 45.78%	**0.5633** 48.20%
Swedish	TD	0.3630	0.4224 16.36%	0.4121 13.53%	0.4331 19.31%	0.4727 30.22%	0.4880 34.44%	0.4838 33.28%	**0.5465** 50.55%

Table 2: Evaluation of decompounding, stemming, and query expansion.

The precision of the German topic 174 with the title "Bayerischer Kruzifixstreit" (Bavarian Crucifix Quarrel) increased from 0.0937 without decompounding to 0.7553 with decompounding. The compound *kruzifixstreit* does not occur in the German document collection, while its component words, *kruzifix* and *streit*, occur 147 and 7,768 times, respectively, in the German document collection.

The Swedish topic 177 with the title "Mjölkkonsumtion i Europa" (Milk Consumption in Europe) has 9 relevant documents, but none was retrieved when compounds were not split. The precision for this topic increased from 0.0 to 0.2396 when compounds were decomposed. The compound *mjölkkonsumtion* occurs only once in the Swedish document collection. Another example is the Swedish topic 199 with the title "Ebolaepidemi i Zaire" (Ebola Epidemic in Zaire) that has 38 relevant documents in total, but the compound *ebolaepidemi* occurs only 4 times in total, once in four Swedish documents. The precision for this topic was increased from 0.2360 before decompounding to 0.6437 after decompounding. The compound *mjölkkonsumtion* was split into *mjölk* and *konsumtion*, and *ebolaepidemi* into *ebola* and *epidemi* after decompounding.

The precision for Dutch topic 171 with the title "Ijshockeyfinale in Lillehammer" (Lillehammer Ice Hockey Finals) increased from 0.0215 before decompounding to 0.3982 after decompounding. This topic has 18 relevant documents in the Dutch collection, but the compound *ijshockeyfinale* occurs only twice in total, once in two documents. After decompounding, the Dutch compound *ijshockeyfinale* was split into *ijshockey* and *finale*.

The Finnish topic 159 with the title of "Pohjanmeri, öljy ja ympäristö" (North Sea Oil Environment) has 6 relevant documents. After splitting the compound *ympristnsuojelun* into *ympristn* and *suojelun*, and the compound *Pohjanmerell* into *Pohjan* and *merell*, the precision increased from 0.0698 without decompounding to 0.4660 with decompounding. Both of the decomposed compounds occur in the *description* fields. Note that the English translations of the Dutch, Finnish, German and Swedish titles in the examples presented above are the English titles in the corresponding CLEF 2003 English topics.

5.2 Bilingual retrieval experiments

We submitted 1 Finnish to German, 1 French to Dutch, 2 German to Italian, and 2 Italian to Spanish bilingual runs. The average precision values for the six official bilingual runs (in bold face) with additional bilingual runs are presented in Table 3. For **bkbideit1**, **bkbifide1** and **bkbiites1** runs, the query words in the *title* and *desc* fields, after removing stopwords, were

translated into the document language using bilingual translation lexicons created from the Official Journal parallel texts. The bilingual translation lexicons used in these three runs were developed using the maximum likelihood ratio test statistic as the association measure. Only the top-ranked translation was retained for each query word. For the **bkbideit2** run, the German topics were translated into English, then into Italian using the L&H MT system, For the **bkbiites2** run, the Italian topics were translated into English, then into Spanish also using the L&H MT system. By the time when the results were due, we still did not have a French to Dutch translation lexicon, so we translated the French topics into English using the L&H MT system, then translated the English topic words into Dutch using the English to Dutch statistical translation lexicon built from the English-Dutch parallel texts. The English query words translated from French were individually translated into Dutch, and only the top-ranked Dutch translation for each translated English word was retained.

run id	topic fields	topic language	document language	translation resources	average precision
bkbifide1	TD	Finnish	German	parallel texts	**0.3814**
bkbifrnl1	TD	French	Dutch	L&H; parallel texts	**0.3446**
bkbideit1	TD	German	Italian	parallel texts	**0.3579**
bkbideit2	TD	German	Italian	L&H	**0.3859**
bkbiites1	TD	Italian	Spanish	parallel texts	**0.4340**
bkbiites2	TD	Italian	Spanish	L&H	**0.4003**
bkbiennl1	TD	English	Dutch	parallel texts	0.4045
bkbienfi1	TD	English	Finnish	parallel texts	0.3011
bkbienfr1	TD	English	French	L&H	0.4156
bkbiende1	TD	English	German	L&H	0.4694
bkbienit1	TD	English	Italian	L&H	0.4175
bkbienes1	TD	English	Spanish	L&H	0.4303
bkbiensv1	TD	English	Swedish	parallel texts	0.3568

Table 3: Performance of cross-language retrieval runs.

The last seven bilingual runs using English topics were used in our multilingual retrieval runs. For the English to French, German, Italian and Spanish bilingual runs, the English topics were translated into French, German, Italian and Spanish using the L&H MT system, while for the English to Dutch, Finnish and Swedish bilingual runs, the English topic words were translated into Dutch, Finnish and Swedish using the statistical translation lexicons built from the parallel texts. Again, only the top-ranked translation in a target language was retained for each English query word. The version 7.0 of L&H Power translator supports bi-directional translation between English and French, English and German, English and Italian, and English and Spanish. Our copy of the L&H Power translator does not support translation to or from Dutch, Finnish and Swedish.

All the bilingual runs applied blind relevance feedback. The top-ranked 10 terms from the top-ranked 10 documents after the initial search were combined with the initial query.

Overall, the performances of our bilingual runs are much lower than those of monolingual runs. The English to Finnish bilingual performance is only 53.45% of our best Finnish monolingual performance. French topic 192 has 19 relevant documents, all in the ATS French collection. The French title of topic 192 is "Assassinat d'un directeur de la télévision russe", its English equivalents being "Russian TV Director Murder". When the English title was translated into French using the L&H MT system, its French translation became "télé russe Directeur Murder". The word *télévision* occurs 39 times, and *télévisions* once in the 19 relevant documents in the TI, LD, TX, or ST fields, but the translated French word *télé* does not occur in the relevant documents. The word *assassinat* occurs 40 times, and *assassinats* once in the relevant documents, but the word *murder* does not occur in the relevant documents. The precision of the bilingual run for this

topic is 0.0162, while the monolingual run precision is 0.8844. Another example is the English topic 186 where the English to French bilingual performance is far below the French monolingual performance for the same topic. The English title of topic 186 is "Dutch Coalition Government" and its equivalent French title is "Gouvernement de coalition néerlandais". The English title word *Dutch* was translated into *hollandais* and the word *Netherlands* in the description into *Hollande*. Neither *Hollande* nor *hollandais* occurs in the 13 relevant French documents for this topic. The English to French bilingual performance for this topic is 0.0162, while the French monolingual performance is 0.6490.

5.3 Multilingual retrieval experiments

5.3.1 Multilingual-4 experiments

In this section, we describe our multilingual retrieval experiments using the English topics. As mentioned in section 5.2, the English topics were translated into Dutch, Finnish, French, German, Italian, Spanish and Swedish using either the L&H MT system or the statistical translation lexicons built from the parallel texts.

run id	topic language	topic fields	merging strategy	recall	precision
bkmul4en1	English	TD	raw score	4668/6145	**0.3783**
bkmul4en2	English	TD	none	4605/6145	**0.4082**
bkmul4en3	English	TD	sum of raw scores	5017/6145	**0.4260**

Table 4: Multilingual-4 retrieval performances.

Table 4 presents the results of three multilingual runs using English topics to search against the collection of documents in English, French, German and Spanish. The **bkmul4en1** run was produced by combining the English monolingual run *bkmonoen1* and three bilingual runs, *bkbienfr1*, *bkbiende1* and *bkbienes1*. The performances of the individual bilingual runs were presented in Table 3. When the results of the four individual runs were combined, the raw scores were not normalized before merging.

The **bkmul4en2** run was produced by searching the English queries against the combined collection consisting of the English documents in the English collection and the English documents translated from French, German and Spanish collections. The French, German and Spanish documents were translated into English in three steps. First, we collected all the words in the French, German and Spanish documents. Second, we translated the French, German and Spanish document words individually into English using the L&H MT system. Finally, we translated the French, German and Spanish documents into English by replacing the French, German and Spanish document words with their English translations produced in the previous step. The English documents and the translated English documents from French, German and Spanish were indexed together. This run included query expansion for which 10 terms were selected from the top-ranked 10 documents after the initial search. Note that this approach does not need to merge individual results.

The **bkmul4en3** run was produced by combining the query translation-based run **bkmul4en1** and the document translation-based run **bkmul4en2**. The relevance scores were not normalized, but summed if the same document was on both ranked lists of documents.

The run **bkmul4en2** performed better than the run **bkmul4en1** on 36 topics, but worse on 23 topics. For most of topics, the precision difference on the same topic between these two different approaches are less than 0.20. However, there are 5 topics for which the precision difference is over 0.30. The precision of topic 161 is 0.0003 in the query translation-based run **bkmul4en1**, but 0.6750 in the document translation-based run **bkmul4en2**. Topic 161 with the English title of "Diets for Celiacs" has 6 relevant documents in total in the multilingual-4 document collection, 5 being Spanish documents and 1 German document. The title word *Celiacs*, which also occurs once in the description field, was not translated into Spanish by the L&H MT system, neither into

German. The precision for topic 161 is only 0.0008 using the Spanish topic translated from the English topic to search the Spanish collection. The failure of retrieving the relevant documents in the English to Spanish bilingual run ultimately led to the poor performance of the multilingual run **bkmul4en1** on topic 161. Although the English topic word *celiacs* was left untranslated by the L&H MT system, its Spanish equivalent *celíacos* in the documents was correctly translated into the English word *celiac*. This is probably why the document translation-based run substantially outperformed the query translation-based run on this topic.

As mentioned above, the L&H MT system translated the English words *Dutch* into *hollandais* and *Netherlands* into *Hollande* in topic 186, neither actually occurring in the French documents. However, the L&H MT system translated correctly the word *néerlandais* in the French documents into the English word *Dutch*. The precision for topic 186 is 0.2213 in the **bkmul4en1** run, but 0.6167 in the **bkmul4en2** run.

5.3.2 Multilingual-8 experiments

run id	topic language	topic fields	merging strategy	recall	precision
bkmul8en1	English	TD	raw score	6342/10020	**0.3317**
bkmul8en2	English	TD	none	5864/10020	**0.3401**
bkmul8en3	English	TD	sum of raw scores	6677/10020	**0.3733**

Table 5: Multilingual-8 retrieval performances.

Table 5 presents the performances of three multilingual retrieval runs involving eight languages. The English topics were used in all three runs.

The **bkmul8en1** run was produced by combining the English monolingual run *bkmonoen1* and seven bilingual runs from English to the other seven document languages. The seven bilingual runs are *bkbiennl1*, *bkbienfi1*, *bkbienfr1*, *bkbiende1*, *bkbienit1*, *bkbienes1* and *bkbiensv1*, whose performances were presented in Table 3. The raw scores of the individual runs were not normalized before merging.

The **bkmul8en2** run was produced by searching the English queries against the combined collection consisting of the English documents in the English collection and the English documents translated from the other seven document languages. The translation of French, German and Spanish documents into English was described in section 5.3.1. The Italian documents were translated into English in the same way as were the French, German and Spanish documents. The Dutch, Finnish and Swedish documents were translated, word-by-word, into English using the statistical translation lexicons built from the parallel texts. For instance, a Finnish document was translated into English by replacing each Finnish word in the document with its most probably English translation found in the statistical Finnish to English lexicon developed from the Finnish-English parallel texts. The English documents and the translated English documents from the other seven document languages were indexed together. For query expansion, 10 terms were selected from the top-ranked 10 documents after the initial search.

The **bkmul8en3** run was the result of merging the query translation-based run **bkmul8en1** and the document translation-based run **bkmul8en2**. The relevance scores were not normalized but summed when the two runs were merged. The **bkmul8en2** run performed better than the **bkmul8en1** run on 34 topics, but worse on 25 topics.

The documents in our multilingual retrieval runs were translated out-of-context, since the words were individually translated into English from other document languages. We conjecture that using documents translated in-context would produce better results. For lack of computational resources, we did not translate the multilingual-4 document collection into English using MT systems and perform retrieval from the translated documents.

6 Conclusions

Decompounding, stemming and query expansion have been shown effective in both monolingual and cross-language retrieval. The automatically generated statistical stemmers improve retrieval performances, they are, however, not as effective as the manually created stemmers. The document translation-based multilingual retrieval is slightly better than query translation-based multilingual retrieval. Combining document translation and query translation in multilingual retrieval achieves even better performance. In the document translation-based multilingual retrieval, the documents are translated into the topic language by translating the document words individually using either MT systems or translation lexicons derived from parallel texts.

7 Acknowledgments

This research was supported in part by DARPA under contract N66001-00-1-8911 as part of the DARPA Translingual Information Detection, Extraction, and Summarization Program (TIDES).

References

[1] *http://europa.eu.int/.*

[2] M.-H. Aaltio. *Finnish for foreigners.* Otava, Helsingissa, 3rd edition, 1967.

[3] J. Atkinson. *Finnish grammar.* The Finnish Literature Society, [Helsinki], 3rd edition, 1969.

[4] M. Braschler, B. Ripplinger, and P. Schäuble. Experiments with the eurospider retrieval system for clef 2001. In *Evaluation of Cross-Language Information Retrieval Systems: Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001, Darmstadt, Germany,September 2001. Revised Papers*, pages 102–110, 2002.

[5] A. Chen. Cross-language retrieval experiments at clef 2002. In C. Peters, editor, *Working Notes for the Cross-Language Evaluation Forum (CLEF) 2002 Workshop 19-20 September, Rome, Italy*, pages 5–20, 2002.

[6] A. Chen and F. C. Gey. Building an arabic stemmer for information retrieval. In *The Eleventh Text Retrieval Conference (TREC 2002).* NIST, 2002.

[7] A. Chen, H. Jiang, and F. C. Gey. Berkeley at ntcir-2: Chinese, japanese, and english ir experiments. In N. Kando, K. Aihara, K. Eguchi, and K. Kato, editors, *Proceedings of the Second NTCIR Workshop Meeting on Evaluation of Chinese & Japanese Text Retrieval and Text Summarization*, pages 5: 32–40. National Institute of Informatics, 2001.

[8] M. Z. Daniel Karp, Yves Schabes and D. Egedi. A freely available wide coverage morphological analyzer for english. In *Proceedings of COLING*, 1992.

[9] T. Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational linguistics*, 19:61–74, March 1993.

[10] W. A. Gale and K. W. Church. A program for aligning sentences in bilingual corpora. *Computational linguistics*, 19:75–102, March 1993.

[11] P. Holmes and I. Hinchliffe. *Swedish : A Comprehensive Grammar.* Routledge, London, 1994.

[12] D. W. Oard, G.-A. Levow, and G. I. Cabezas. Clef experiments at the university of maryland: Statistical sstemming and backoff translation strategies. In C. Peters, editor, *Cross-Language Information Retrieval and Evaluation: Workshop of Cross-Language Evaluation Forum, CLEF 2000, Lisbon, Portugal, September 21-22, 2000, Revised Papers*, pages 176–187, 2001.

[13] F. J. Och and H. Ney. Improved statistical alignment models. In *ACL00*, pages 440–447, Hongkong, China, October 2000.

[14] M. Porter. Snowball: A language for stemming algorithms. *Available at http://snowball.tartarus.org/texts/introduction.html*, 2001.