Dublin City University at CLEF 2007: Cross-Language Speech Retrieval Experiments

Ying Zhang   Gareth J. F. Jones   Ke Zhang

Centre for Digital Video Processing & School of Computing
Dublin City University, Dublin 9, Ireland

{yzhang,gjones,kzhang}@computing.dcu.ie

Abstract. The Dublin City University participation in the CLEF 2007 CL-SR English task concentrated primarily on issues of topic translation. Our retrieval system used the BM25F model and pseudo relevance feedback. Topics were translated into English using the Yahoo! BabelFish free online service combined with domain-specific translation lexicons gathered automatically from *Wikipedia*. We explored alternative topic translation methods using these resources. Our results indicate that extending machine translation tools using automatically generated domain-specific translation lexicons can provide improved CLIR effectiveness for this task.

1 Introduction

The Dublin City University participation in the CLEF 2007 CL-SR task focused on extending our CLEF 2006 system [1] to investigate combinations of general and domain-specific topic translation resources. Our 2006 system used the BM25F field combination approach [2] with summary-based pseudo relevance feedback (PRF) [3]. Our submissions to CLEF 2007 included both English monolingual and French–English bilingual tasks using automatic only and combined automatic and manual fields. The Yahoo! BabelFish machine translation system[1] was used for baseline topic translation into English; this was then combined with domain-specific translation lexicons gathered automatically from *Wikipedia*.

The remainder of this paper is structured as follows: Section 2 summarises the features of the BM25F retrieval model, Section 3 overviews our retrieval system, Section 4 describes our topic translation methods, Section 5 presents our experimental results, and Section 6 concludes the paper with a discussion of our results.

2 Field Combination

The English collection comprises 8104 "documents" taken from 589 hours of speech data. The spoken documents are provided with a rich set of data fields,

[1] babelfish.yahoo.com

full details of these are given in [4, 5]. In this work, we explored field combination based on the following fields: an automatic transcription of the spoken content (the ASR2006B field); automatically generated keywords (AKW1,AKW2); manually generated keywords (MK); manual summary of each segment (SUM); and names of all individuals appearing in the segment.

[2] demonstrates the weaknesses of standard combination methods and proposes an extended version of the standard BM25 term weighting scheme referred to as BM25F, which combines multiple fields in a more well-founded way. The BM25F combination approach uses a weighted summation of the multiple fields of the documents to form a single field for each document. The importance of each document field for retrieval is determined empirically, each field is then multiplied by a scalar constant representing the importance of this field, and the components of all fields are then summed.

3 Okapi Retrieval System

The basis of our experimental system is the City University research distribution version of the Okapi system [6]. The documents and search topics are processed to remove stopwords from a standard list of about 260 words, suffix stripped using the Okapi implementation of Porter stemming [7] and terms are indexed using a small standard set of synonyms. None of these procedures were adapted for the CLEF 2007 CL-SR test collection.

Document terms were weighted using the Okapi BM25 weighting scheme,

$$cw(i,j) = cfw(i) \times \frac{tf(i,j) \times (k_1 + 1)}{k_1 \times ((1-b) + (b \times ndl(j))) + tf(i,j)}$$

$$cfw(i) = log\left(\frac{(rload + 0.5)(N - n(i) - bigrload + rload + 0.5)}{(n(i) - rload + 0.5)(bigrload - rload + 0.5}\right)$$

where $cw(i,j)$ = the weight of term i in document j; $n(i)$ = total number of documents containing term i; N = total number of documents in the collection;, $tf(i,j)$ = within document term frequency; $dl(j)$ = length of j; $avgdl$ = average document length in the collection; $ndl(j) = dl(j)/agvdl$ is the normalized document length; and k_1 and b are empirically-tuned constants for a particular collection. $bigrload$ is an assumed number of relevant documents and $rload$ the number of these containing i. These take the standard values of 4 and 5 respectively [8]. The matching score for each document is computed by summing the weights of terms appearing in the query and the document. In this investigation we use the summary-based PRF method for query-expansion described in [3].

4 MT-based Query Translation

Machine Translation (MT) based query translation using an existing MT system has been widely used in cross-language information retrieval (CLIR) with good average performance. In our experiments, topics were translated into English using the Yahoo! BabelFish system powered by SYSTRAN. While BabelFish can provide reasonable translations for general language expressions, it is not sufficient for domain-specific terms such as personal names, organization names, place names, etc. To reduce the errors introduced by such terms during query translation, we augmented the standard BabelFish with domain-specific lexicon resources gathered from *Wikipedia*[2].

4.1 Domain-specific lexicon construction

As a multilingual hypertext medium, Wikipedia has been shown to be a valuable new source of translation information [9, 10]. Unlike the web, the hyperlinks in Wikipedia have a more consistent pattern and meaningful interpretation. A Wikipedia page written in one language can contain hyperlinks to its counterparts in other languages, where the hyperlink basenames are translation pairs. For example, the English wikipedia page `en.wikipedia.org/wiki/World_War_II` contains hyperlinks to German `de.wikipedia.org/wiki/Zweiter_Weltkrieg` , French `fr.wikipedia.org/wiki/Seconde_Guerre_mondial`, and Spanish `es.wikipedia.org/wiki/Segunda_Guerra_Mundial`. The English term “World War II” is the translation of the German term “Zweiter Weltkrieg”, the French term “Seconde Guerre mondial”, and the Spanish term “Segunda Guerra Mundial”.

Additionally, we observed that multiple English wikipedia URLs `en.wikipedia.org/wiki/World_War_II`, `en.wikipedia.org/wiki/World_War_2`, `en.wikipedia.org/wiki/WW2`, and `en.wikipedia.org/wiki/Second_world_war` are redirected to the same wikipedia page and the URL basenames “World War II”, “World War 2”, “WW2”, and “Second world war” are synonyms. Using all these English terms during query translation is a straightforward approach to automatic post-translation query expansion.

To utilize the multilingual linkage and the link redirection features, we implemented a three-stage automatic process to extract German, French, and Spanish to English translations from Wikipedia:

1. An English vocabulary list was constructed by performing a limited crawl of the English wikipedia[3], Category:World War II. This category is likely to contain links to pages and subcategories relevant to entities appearing in the document collection. In total, we collected 7431 English web pages.
2. For each English page, we extracted the hyperlinks to each of the query languages. This provided a total of 4446, 3338, and 4062 hyperlinks to German, Spanish, and French, respectively.
3. We then selected the basenames of each pair of hyperlinks (German–English, French–English, and Spanish–English) as translations and added them into our domain-specific lexicons. Non-English multi-word terms were added into the phrase dictionary for each query language. These phrase dictionaries are later used for phrase identification during query pre-processing.

[2] `www.wikipedia.org`

[3] `http://en.wikipedia.org`

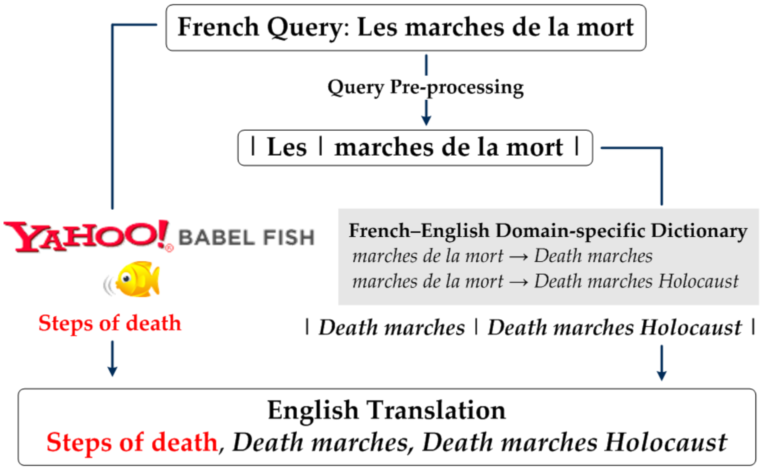
French Query: Les marches de la mort
Query Pre-processing
| Les | marches de la mort |
YAHOO! BABEL FISH
Steps of death
French–English Domain-specific Dictionary
marches de la mort → Death marches
marches de la mort → Death marches Holocaust
| Death marches | Death marches Holocaust |
English Translation
Steps of death, Death marches, Death marches Holocaust

Fig. 1. *An example of French–English query translation. (Topic numbered 3005)*

4.2 Query translation process

As shown in Figure 1, our query translation process is performed as follows:

1. Query pre-processing: We used the phrase dictionary with the maximum forward matching algorithm to segment each query Q into a list of terms $\{q_1, q_2, q_3, ..., q_n\}$.
2. Domain-specific lexicon lookup: For each query term q_i (where $i \in (1, n)$), we obtained all its English translations $\{e_{i1}, e_{i2}, e_{i3}, ..., e_{im}\}$ via a domain-specific lexicon look-up.
3. BabelFish translation: we then translated the original query Q into the English query E using the Yahoo! BabelFish system.
4. Translation results merging: For each English term e_{ij} (where $i \in (1, n)$ and $j \in (1, m)$) obtained in Step 2, we appended it to the end of the translated English query E.

5 Experimental Results

In this section we report results for combinations of manual only fields, automatic only fields and combining both manual and automatic fields. Monolingual retrieval results show precision at cutoff ranks of 5, 10 and 30, standard TREC mean average precision (MAP) and recall in terms of the total number of relevant documents retrieved for the test topic set. CLIR results compare alternative topic translations resources showing MAP and precision at rank 10. Runs formally submitted for evaluation are indicated by an asterisk in the tables.

5.1 System parameters

Our retrieval system requires a number of parameters to be set for the term weighting, field combination, and PRF components. All parameter values were set empirically using the 63 English language CLEF 2007 CL-SR training topics.

RUN Description	**Query Fields**	**Recall**	**MAP**	**P@5**	**P@10**	**P@30**
Manual field combination						
(MK×1+SUM×1, $k_1 = 1.0$, $b = 0.5$)						
Baseline	TDN	1850	0.2773	0.4909	0.4576	0.4182
*PRF	TDN	1903	0.2847	0.4970	0.4515	0.4222
Automatic field combination						
(AK1×1+AK2×1+ASR2006B×2, $k_1 = 8.0$, $b = 0.5$)						
Baseline	TD	1311	0.0735	0.1697	0.1697	0.1677
*PRF	TD	1360	0.0787	0.1697	0.1727	0.1636
Manual and automatic field combination						
(MK×4+SUM×4+ASR2006B×1, $k_1 = 3.0$, $b = 0.6$)						
*Baseline	TD	1907	0.2399	0.4364	0.3818	0.3838
*PRF	TD	1974	0.2459	0.4364	0.3818	0.3556

Table 1. Results for English monolingual retrieval. No significant difference observed.

Term weighting and field combination Based on the training runs the term weighting and field combination parameters were set as follows: *Manual data field combination*, Okapi parameters $k_1 = 1.0$ and $b = 0.5$ with document fields weighted as MK×1, and SUM×1; *Automatic data field combination*, $k_1 = 8.0$ and $b = 0.5$ document fields weighted as A1K×1, AK2×1, and ASR06B×2; and *Manual and automatic data field combination*, $k_1 = 3.0$ and $b = 0.6$ document fields weighted as MK×4, SUM×4, and ASR06B×1.

Pseudo-relevance feedback The top t ranked expansion terms taken from document summaries are added to the original query. The original topic terms are up-weighted by a factor α relative to the expansion terms. Our PRF query expansion involves five parameters as follows: $t =$ number of the expansion terms selected; $s =$ number of sentences selected as the document summary; $d_1 =$ number of documents used for sentence selection; $d_2 =$ is the number of documents used for expansion term ranking; $\alpha =$ the up-weighting factor.
PRF parameter selection is not necessarily consistent from one collection (indexed using different field combination methods) to another. Our experiments showed that $t = 60$, $s = 6$, $d_1 = 3$, $d_2 = 20$, and $\alpha = 3.0$ give the best results for the manual data field combination, and manual and automatic data field combination; while $t = 40$, $s = 6$, $d_1 = 3$, $d_2 = 20$, and $\alpha = 3.0$ produce the best results for the automatic data field combination.

5.2 Field combination and summary-based PRF

Table 1 shows monolingual English retrieval results for both the baseline condition without application of PRF and with summary-based PRF. For the combination of the MK and SUM fields it can be seen that application of PRF generally produces a small improvement in performance. Note that the topics

RUN Description	**∗French**		**Spanish**		**German**	
	MAP	P@10	MAP	P@10	MAP	P@10
BabelFish baseline	0.0476	0.1242	0.0566	0.1364	0.0563	0.1394
BabelFish+PRF	0.0501	0.1242	0.0541	0.1303	0.0655	0.1303
BabelFish+LEX	0.0606*	0.1394	0.0581	0.1394	0.0586	0.1424
∗BabelFish+LEX+PRF	0.0636*	0.1394	0.0588	0.1273	0.0617	0.1364

Table 2. Results for cross-lingual retrieval. (TD runs on automatic field combination, A1K×1+AK2×1+ASR2006B×2, $k_1 = 8.0$, $b = 0.5$. $*$ shows significance at 0.05 level.)

here use all three topics fields Title, Description and Narrative (TDN), and thus these results cannot be compared directly to any other results shown here which use only Title and Description fields (TD). Similarly for both the automatic only fields runs combining AK1, AK2 and ASR2006B, and the combination of manual and automatic fields using MK, SUM and ASR2006B, application of PRF produces a small improvement in average and high rank precision, although there appear to be some problems at lower ranks which we intend to investigate.

5.3 Yahoo! BabelFish combined with domain-specific lexicons

Table 2 shows CLIR results for standard BabelFish translation (BabelFish baseline), and augmented translations using the domain-specific lexicons (BabelFish+ LEX). The BabelFish+LEX method led to a significant improvement (27%) for the French–English retrieval task, but only 3% and 4% in Spanish–English and German–English, respectively. This can be explained by the fact that the MAP values for the baseline runs of German and Spanish are much higher than the MAP for the French baseline. We noticed that the description field of German topics sometimes contains additional explanation enclosed by square brackets. The effect of this was often that more correct documents are retrieved in the German–English task. We therefore believe that the BabelFish system gives a better translation from Spanish, rather than French and German, to English.

At the individual query level (shown in Table 3), we observe that retrieval effectiveness sometimes degrades slightly when the query is augmented to include translations from our domain-specific lexicons, despite the fact that they are correct translations of the original query terms. This occurred mainly due to the fact that additional terms result in a decrease in the rank of relevant documents because they are too general within the collection. For example, “war”, “Europe”, “Poland”, “holocaust”, “country”, “people”, “history”, etc.

We used the summary-based PRF to provide post-translation query expansion in all CLIR runs (see BabelFish+PRF and BabelFish+LEX +PRF shown in Table 2). This produced improvements of 7% for the mono-lingual run, but only provided improvements of 5%, 1%, and 5% in French–English, Spanish–English, and German–English CL-SR effectiveness. The MAP value of the French–English (BabelFish+LEX+PRF) run provides the best results among all runs submitted by participants in this task.

6 Conclusions

Our experiments for the CLEF 2007 CL-SR task focused on the combination of standard MT with domain-specific translation resources. Our results indicate that combining domain-specific translation derived from Wikipedia with the output of standard MT can produce substantial improvements in CLIR retrieval effectiveness. For example, the French term 'Hassidisme' (in Query 1166) is translated to 'Hasidic Judaism' in English, 'Varian Fry' (in Query 1133) and 'Marches de la mort' (in Query 3005) are correctly detected as phrases and thus translated as 'Varian Fry ' and 'death marches' in English, respectively. Further improvements can also be observed when combined with PRF. However, these trends are not observed consistently, and further investigations will focus on understanding differences in behaviour, and refining our procedures for training domain-specific translation resources.

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.

References

1. Jones, G.J.F., Zhang, K., Lam-Adesina, A.M.: Dublin city university at clef 2006: Cross-language speech retrieval (cl-sr) experiments. In: Proceedings of the CLEF 2006 Workshop, Alicante, Spain, Springer (2007) 794–802
2. Robertson, S.E., Zaragoza, H., Taylor, M.: Simple BM25 Extension to Multiple Weighted Fields. In: Proceedings of the 13th ACM CIKM. (2004) 42–49
3. Lam-Adesina, A.M., Jones, G.J.F.: Dublin City University at CLEF 2005: Cross-Language Speech Retrieval (CL-SR) Experiments. In: Proceedings of the CLEF 2005 Workshop, Vienna, Austria, Springer (2006) 792–799
4. White, R.W., Oard, D.W., Jones, G.J.F., Soergel, D., Huang, X.: Overview of the CLEF-2005 Cross-Language Speech Retrieval Track. In: Proceedings of the CLEF 2005 Workshop, Vienna, Austria, Springer (2006) 744–759
5. Oard, D.W., Wang, J., Jones, G.J.F., White, R.W., Pecina, P., Soergel, D., Huang, X., Shafran, I.: Overview of the CLEF-2006 Cross-Language Speech Retrieval Track. In: Proceedings of the CLEF 2006 Workshop, Alicante, Spain, Springer (2007) 744–758
6. Robertson, S.E., Walker, S., Jones, S., Hancock-Beaulieu, M., Gatford, M.: Okapi at TREC-3. In: Proceedings of the 3rd Text REtrieval Conference. (1994) 109–126
7. Porter, M.F.: An algorithm for suffix stripping. Automated Library and Information Systems **14**(3) (1980) 130–137
8. Beaulieu, M.M., Gatford, M.: Interactive Okapi at TREC-6. In: Proceedings of the 6th Text REtrieval Conference. (1997) 143–168
9. Adafre, S.F., de Rijke, M.: Discovering missing links in Wikipedia. In: Proceedings of the 3rd International Workshop on Link Discovery, Chicago, Illinois, ACM Press (2005) 90–97
10. Declerck, T., Pèrez, A.G., Vela, O., Gantner, Z., Manzano-Macho, D.: Multilingual Lexical Semantic Resources for Ontology Translation. In: Proceedings of LREC, Genoa, Italy (2006)

	Query ID	MAP		Additional Translations from Lexicons
		BabelFish	BabelFish+LEX	
French–English				
1	1345	0.0304	0.1025	Buchenwald concentration camp, Buchenwald, August 24
2	1623	0.3130	0.2960	Resistance movement, Poland
3	3005	0.0351	0.2249	Death marches, Death marches Holocaust, Schutzstaffel SS
4	3007	0.0113	0.0088	Europe, War
5	3009	0.1488	0.1247	War
6	3022	0.0568	0.0558	War, Country, Culture
7	3024	0.0010	0.0003	War
8	3025	0.0670	0.0401	War, Europe
9	3033	0.0975	0.0888	Palestine, Palestine region
German–English				
1	1133	0.1057	0.1044	Varian Fry, History, Marseille, Marseilles
2	1173	0.0461	0.0321	Art
3	3005	0.2131	0.1868	Schutzstaffel SS, Concentration camp, Concentration camps, Internment
4	3007	0.0058	0.0049	Europe
5	3009	0.1495	0.1256	War
6	3010	0.0002	0.0000	Germany, Property, Forced labor, Forced labour
7	3012	0.0003	0.0002	Germany, Jew, Jewish, Jewish People, Jews
8	3015	0.0843	0.0700	Jew, Jewish, Jewish People, Jews
9	3022	0.0658	0.0394	War, Holocaust, The Holocaust, Culture
10	3023	0.0100	0.0082	Holocaust, The Holocaust, Germany
11	3024	0.0006	0.0002	War, Holocaust, The Holocaust
12	3025	0.0857	0.0502	War, Jew, Jewish, Jewish People, Jews, Europe
13	3026	0.0021	0.0016	Concentration camp, Concentration camps, Internment
Spanish–English				
1	1173	0.0184	0.0077	Art, Literature
2	1345	0.0689	0.0596	Buchenwald concentration camp, Buchenwald, Allied powers, Allies of World War II, August 24
3	1624	0.0034	0.0003	Polonia, Poland, Holocaust, The Holocaust
4	3005	0.0685	0.0341	Posthumously, Schutzstaffel SS, Allied powers, Allies, Allies of World War II
5	3007	0.0395	0.0213	Europe, War
6	3009	0.1495	0.1256	War
7	3011	0.0413	0.0283	Holocaust, The Holocaust
8	3022	0.0661	0.0449	Holocaust, The Holocaust, Country, Culture
9	3024	0.0029	0.0016	Violence, War, Holocaust, The Holocaust
10	3025	0.0548	0.0371	War
11	3026	0.0036	0.0024	Partisans, War

Table 3. Examples of using extra translations from the domain-specific lexicons led to a deterioration in retrieval effectiveness. (TD runs on automatic field combination, A1K×1+AK2×1+ASR06B×2, $k_1 = 8.0$, $b = 0.5$.)