Dublin City University at CLEF 2007: Cross-Language Speech Retrieval (CL-SR) Experiments

Ying Zhang, Gareth J. F. Jones, and 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 participated in the CLEF 2007 CL-SR English task. For CLEF 2007 we concentrated primarily on the issues of topic translation, combining this with search field combination and pseudo relevance feedback methods used for our CLEF 2006 submissions. Topics were translated into English using the Yahoo! BabelFish free online translation service combined with domain-specific translation lexicons gathered automatically from Wikipedia. We explored alternative translations methods with document retrieval based the combination of the multiple document fields using the BM25F field combination model. Our results indicate that extending machine translation tools using automatically generated domain-specific translation dictionaries can provide improved CLIR effectiveness.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.3 Information Search and Retrieval

General Terms

Measurement, Performance, Experimentation

Keywords

Speech Retrieval, Domain specific translation, Evaluation, Generalized Average Precision

1 Introduction

The Dublin City University participation in the CLEF 2007 CL-SR task focussed on extending our CLEF 2006 system to investigate combinations of general and domain-specific topic translation resources. Our 2006 participation in the CL-SR task concentrated on the combination of the multiple fields associated with the speech documents. Our study was based on using the document field combination extended version of BM25 termed BM25F introduced in [11]. In addition, we incorporate our existing information retrieval methods based on the Okapi model with summary-based pseudo-relevance feedback (PRF) [9]. Our official submissions included both English monolingual and French–English bilingual tasks using automatic only and combined automatic and manual fields. Topics were translated into English using a baseline of the online Yahoo! BabelFish machine translation system [1]. For our CLEF 2007 experiments these translations are combined with domain-specific translation lexicons gathered automatically from *Wikipedia*.

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

2 Field Combination

The English collection comprises 8104 "documents" that are manually-determined topically-coherent segments taken from 272 interviews with Holocaust survivors, witnesses and rescuers, totaling 589 hours of speech. The spoken documents are provided with a rich set of data fields, full details of these are given in [13][7]. In this work, we explored field combination based on the following fields:

- a transcription of the spoken content of the document generated using an automatic speech recognition (ASR) system, (several transcriptions are available, for experiments we use the ASR2006B field,
- two assigned sets of keywords generated automatically (AKW1,AKW2),
- one assigned set of manually generated keywords (MK),
- a short three sentence manually written summary of each segment (SUM),
- a list of the names of all individuals appearing in the segment.

Two standard methods of combining multiple document fields in retrieval are:

- to simply merge all the fields into a single document representation and apply standard single document field information retrieval methods,
- to index the fields separately, perform individual retrieval runs for each field and then merge the resulting ranked lists by summing in a process of data fusion.

The topic of field combination for this type of task with ranked information retrieval schemes is explored in [11]. That paper demonstrated the weaknesses of the simple standard combination methods and proposed 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 simple weighted summation of the multiple fields of the documents to form a single field for each document in the usual way. The importance of each document field for retrieval can be determined empirically in separate runs, the count of each term appearing in each field is multiplied by a scalar constant representing the importance of this field, and the components of all fields are then summed to form the overall single field document representation for indexing. Once the fields have been combined in a weighted sum, standard single field IR methods can be applied.

3 Okapi Retrieval System

3.1 Term Weighting

The basis of our experimental system is the City University research distribution version of the Okapi system [12]. 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 [10] 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 shown as follows,

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

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

where

$cw(i,j)$ represents the weight of term i in document j;
$n(i)$ is the total number of documents containing term i;
N is the total number of documents in the collection;
$tf(i,j)$ is the within document term frequency;
$ndl(j)$ is the normalized document length;
$dl(j)$ is the length of j;
$avgdl$ is the average document length in the collection;
k_1 and b are empirically selected tuning constants for a particular collection.

The matching score for each document is computed by summing the weights of terms appearing in the query and the document. The BM25 k_1 and b values used for our submitted runs were tuned using the 63 CLEF 2007 CL-SR English training topics. *rload* and *bigrload* take the default parameters of 4 and 5 respectively.

3.2 Pseudo-Relevance Feedback

Query expansion by pseudo relevance feedback (PRF) is a well-established procedure in both monolingual and cross-lingual IR, potentially providing some improvement in retrieval effectiveness. The method used here is based on our work originally described in [8], and modified for the CLEF 2005 CL-SR task [9]. A *summary* is made of the automatic speech recognition (ASR) transcription of each of the top ranked documents, which are assumed to be relevant to a given query. The document summary is then expanded to include all terms in the other metadata fields used in this document index. All non-stopwords in these augmented summaries are ranked using a slightly modified version of the Robertson Selection Value (RSV) [12].

In our modified version of RSV, the top t potential expansion terms are selected from the augmented summaries of the top d_1 ranked documents, but ranked using statistics from a larger number d_2 of assumed relevant ranked documents from the initial run.

The summary-based PRF method operates by selecting topical-related expansion terms from document summaries. However, since the ASR transcriptions of the conversational speech documents do not contain punctuation, we developed a method of selecting significant document segments to identify documents “summaries”. Our approach is derived from Luhn’s word cluster hypothesis. Luhns hypothesis states that significant words separated by up to five non-significant words maximum are likely to be strongly related. Clusters of these strongly related word were identified in the running document transcription by searching for word groups separated by not more than five insignificant words. Words appearing between clusters are not included in clusters, and thus can be ignored for the purposes of query expansion since they are by definition stop words. The clusters were then awarded a significance score based on the following two measures:

Luhn’s Keyword cluster method Luhns method assigns a sentence score LS for the highest scoring cluster within a sentence [8]. We adapted this method to assign a cluster score as follows:

$$LS = \frac{SW^2}{TW}$$

where SW is the number of bracketed significant words, and TW is the total number of bracketed words.

Query-biasd method This method assigns a score QS to each sentence based on the number of query terms in the sentence as follows:

$$QS = \frac{TQ^2}{NQ}$$

where TQ is the number of query terms occurring in the sentence, and NQ is the total number of terms in a query.

For each sentence (cluster), the overall sentence score SS is calculated using $SS = LS + QS$. The top s sentences (clusters) with the highest SS are then selected as the document summary.

4 MT-based Query Translation

Machine Translation (MT) based query translation uses an existing MT system to provide automatic translation. This approach has been widely used in cross-language information retrieval with good average performance when such an MT system is available for the language pair of the topic and document. In our experiments, topics were translated into English using the Yahoo! BabelFish powered by SYSTRAN [1]. 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[1] has been proved to be a valuable new source of translation information [3, 4, 5, 6]. 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 the automatic post-translation query expansion.

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

1. An English vocabulary for the domain of the test collection was constructed by performing a limited crawl of the English wikipedia[2], Category:World War II. This category is more likely to contain links to pages and subcategories concerning events, persons, places, and organizations pertaining to war crimes or crimes against humanity especially during the second world war. In total, we collected 7431 English web pages.

2. For each English page obtained, 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.

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

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

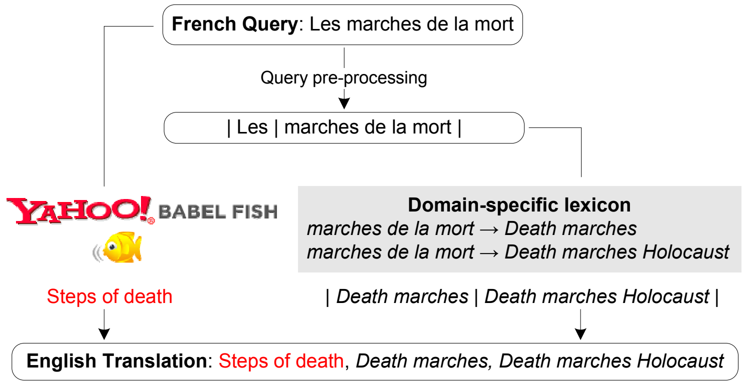
French Query: Les marches de la mort
Query pre-processing
| Les | marches de la mort |
YAHOO! BABEL FISH
Domain-specific lexicon
marches de la mort → Death marches
marches de la mort → Death marches Holocaust
Steps of death
| Death marches | Death marches Holocaust |
English Translation: Steps of death,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

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

3. We then selected the basenames of each pair of hyperlinks (German–English, French–English, and Spanish–English) as translations and added into our domain-specific lexicons. The 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.

4.2 Query translation process

As shown in Figure 1, our query translation process is performed in the following manner:

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.

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 our experimental runs for the CLEF 2007 English CL-SR task. Results are shown for combinations of manual only fields, automatic only fields and combining both manual and automatic fields. For 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. For CLIR results compare alternative topic translations resources showing MAP and precision at rank 10. Our submitted runs for the CLEF 2007 are indicated by a * 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 CLEF 2007 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. (Our submitted runs are denoted by the *.)

Term Weighting and Field Combination Based on these training runs the term weighting and field combination parameters were set as follows:

- For the *manual data field combination*, Okapi parameters $k_1 = 1.0$ and $b = 0.5$ give the best results when the document fields are weighted as MK×1, and SUM×1;
- For the *automatic data field combination*, $k_1 = 8.0$ and $b = 0.5$ perform the best when the document fields are weighted as A1K×1, AK2×1, and ASR06B×2; and
- For the *manual and automatic data field combination*, $k_1 = 3.0$ and $b = 0.6$ produce the best results when the document fields are weighted as MK×4, SUM×4, and ASR06B×1.

PRF For all our PRF runs, the top d_1 ranked documents were assumed relevant for term selection and document summaries comprised the best scoring s clusters. The RSV values to rank the potential expansion terms were estimated based on the top d_2 ranked assumed relevant documents. The top t ranked expansion terms taken from the clusters were added to the original query in each case. The original topic terms are up-weighted by a factor α relative to the expansion terms. Our PRF query expansion thus involves five parameters as follows:

t	is the number of the expansion terms selected from the summary;
s	is the number of sentences (clusters) selected as the document summary;
d_1	is the number of documents used for sentence (cluster) selection;
d_2	is the number of documents used for expansion terms ranking;
α	is the up-weighting factor.

This set of parameters were again tuned using the CLEF 2007 CL-SR English training data. We note that PRF involves selection of parameter values that are 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; $t = 40$, $s = 6$, $d_1 = 3$, $d_2 = 20$, and $\alpha = 3.0$ produce the best results for the automatic data field combination.

5.2 Field Combination and summary-based PRF

This section presents results for our field combination experiments for monolingual English retrieval. Table 1 shows results for both the baseline condition without application of PRF and with our summary-based PRF.

For the combination of the MK and SUM fields we can field than application of PRF generally produces a small improvement in performance. Note that the topics here use all three topics fields Title, Description and Narrative (TDN), and thus these results cannot be compared directly

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, $k_1 = 8.0$, $b = 0.5$. Our submitted run is denoted by the *.)

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

We then explore the combinations of the query translation and post-translation query expansion, and investigate the improvement contributed by each component in German, French, and Spanish to English CL-SR. The results of these experiments are shown in Table 2.

As shown in Table 2, in comparison to the standard BabelFish translation (BabelFish baseline), augmented translations from the domain-specific lexicons (BabelFish+LEX) led to a significant improvement (27%) in 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 should be retrieved in the German–English task. We therefore believe that the BabelFish system gives a better translation from Spanish, rather French and German, to English.

At the individual query level (shown in Table 3), we observed that retrieval effectiveness sometimes slightly degraded when the query was augmented to contain 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 of relevant documents at ranks, because they are too general in the collection. For example, “war”, “Europe”, “Poland”, “holocaust”, “country”, “Jewish”, “people”, “history”, “concentration camp”, etc. This problem may be solved if we down-weight the general-term translations during the retrieval process, so that when term frequency is used in calculating similarity, documents with many general terms will not be over-emphasized. We intend to explore this issue in further experiments.

We used the summary-based PRF to provide post-translation query expansion in all cross-lingual retrieval runs (see BabelFish+PRF and BabelFish+LEX+PRF shown in Table 2). It gave 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.

6 Conclusions

This paper has described results for our participation in the CLEF 2007 CL-SR track. In 2007 our experiments focussed on the combination of standard machine translation with domain-specific translation resources. Our results indicate that combining domain-specific translation derived from Wikipedia with the output of standard machine translation can produce substantial improvements in MAP. Further improvements can also be observed when combined with PRF. How-

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, 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 leds to a deterioration in retrieval effectiveness. (TD runs on automatic field combination, A1K×1+AK2×1+ASR06B×2, $k_1 = 8.0$, $b = 0.5$.)

ever, these trends are not observed consistently in all cases, and further investigations will focus on understanding differences in behaviour more clearly and refining our procedures for training domain-specific translation resources.

References

[1] babelfish.yahoo.com.

[2] www.wikipedia.org.

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

[4] Sisay Fissaha Adafre and Maarten de Rijke. 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, 2006.

[5] Gosse Bouma, Ismail Fahmi, Jori Mur, Gertjan van Noord, Lonneke van der Plas, and Jorg Tiedemann. 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, 2006.

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

[7] D.W.Oard, J.Wang, G.J.F.Jones, R.W.White, P.Pecina, D.Soergel, X.Huang, and I.Shafran. Overview of the CLEF-2006 cross-Language speech retrieval track. In *Proceedings of the CLEF 2006: Workshop on Cross-Language Information Retrieval and Evaluation*, Alicante, Spain, 2007. Springer.

[8] Adenike M. Lam-Adesina and Gareth J. F. Jones. Applying summarization techniques for term selection in relevance feedback. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1–9, New Orleans, Louisiana, United States, 2001. ACM Press.

[9] Adenike M. Lam-Adesina and Gareth J. F. Jones. Dublin City University at CLEF 2005: cross-language speech retrieval (CL-SR) experiments. 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 792–799. Springer, 2005.

[10] Martin F. Porter. An algorithm for su#x stripping. *Automated Library and Information Systems*, 14(3):130–137, 1980.

[11] S. E. Robertson, H. Zaragoza, and M. Taylor. Simple BM25 extension to multiple weighted fields. In *Proceedings of the 13th ACM International Conference on Information and Knowledge Management*, pages 42–49, 2004.

[12] Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. Okapi at TREC-3. In *Proceedings of the 3rd Text REtrieval Conference*, pages 109–126, 1994.

[13] R. W. White, D. W. Oard, G. J. F. Jones, D. Soergel, and X. Huang. Overview of the CLEF-2005 cross-language speech retrieval track. In *Proceedings of the CLEF 2005: Workshop on Cross-Language Information Retrieval and Evaluation*, pages 744–759, Vienna, Austria, 2006. Springer.