University of Ottawa's Contribution to CLEF 2005, the CL-SR Track

Diana Inkpen, Muath Alzghool, and Aminul Islam
University of Ottawa
{diana,alzghool,mdislam}@site.uottawa.ca

Abstract

We present the participation of the University of Ottawa in the Cross-Language Spoken Document Retrieval task at CLEF 2005. In order to translate the queries, we combined the results of several online Machine Translation tools. For the Information Retrieval component we used the SMART system, with several weighting schemes for indexing the documents and the queries. One scheme in particular lead to better results than other combinations. We present the results of the submitted runs and of many un-official runs. We compare the effect of several translations from each language. We present results on phonetic transcripts of the collection and queries and on the combination of text and phonetic transcripts. We also include the results when the manual summaries and keywords are indexed.

Categories and Subject Descriptors

H.3. [**Information Storage and Retrieval**] H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval

General terms: Measurement, Performance, Experimentation

Keywords: Cross-Language Information Retrieval, Spoken Document Retrieval, Machine Translation

1. Introduction

This paper presents the first participation of the University of Ottawa group in CLEF, the Cross-Language Spoken Retrieval (CL-SR) track. We briefly describe the task. Then, we present our system, followed by results for the submitted runs and for many unofficial runs. We experiment with many possible weighting schemes for indexing the documents and the queries. We compare the effect of several translations of the queries and of combining the translations. We look at using phonetic transcriptions of the queries and documents instead of the original ASR-produced text, and at combining the phonetic transcripts with the text. At the end we present the best results when all available information in the collection is used.

The CLEF-2005 CL-SR test collection includes 8104 segments, 75 topics (queries), and 12359 Relevance Judgments to facilitate information retrieval experiments. Segments are the unit of retrieval in the CLEF CL-SR evaluation. Interviews with survivors of the Holocaust were manually segmented to form topically coherent segments by subject matter experts at the Survivors of the Shoah Visual History Foundation. See (Oard *et. al.*, 2004) for more details. Only the ASRTEXT2004A field (and optionally the keywords automatically extracted from it) were allowed to be indexed for the competition. This field contains ASR transcripts of the audio segments, with 38% word error rate. For contrastive studies, metadata for each segment (manual summaries, thesaurus terms, and person names) were included as additional fields that can optionally be indexed.

As a baseline, indexing the ASRTEXT2004A field using Okapi BM 25 term weights and searching with Title queries yielded an uninterpolated mean average precision of 0.0551 when run at the University of Maryland with the freely available PSE vector space search engine. This was evaluated on the 38 training topics, not on the 25 test topics for which we report results in this paper.

The topics were created in English from actual user requests and then translated into Czech, German, French, and Spanish by native speakers. An example topic in English is the following:

<top>
<num>1159
<title>Child survivors in Sweden

<desc>Describe survival mechanisms of children born in 1930-1933 who spend the war in concentration camps or in hiding and who presently live in Sweden.
<narr>The relevant material should describe the circumstances and inner resources of the surviving children. The relevant material also describes how the wartime experience affected their post-war adult life.
</top>

The Spanish, German, and Czech topics provided by the CLEF organizers contained translations of all the fields (title, description, and narrative). For French the narrative field was not translated, due to lack of time. The French topic equivalent to the above English example is the following:

<top>
<num>1159
<title>Les enfants survivants en Suède
<desc>Descriptions des mécanismes de survie des enfants nés entre 1930 et 1933 qui ont passé la guerre en camps de concentration ou cachés et qui vivent actuellement en Suède.
</top>

2. System Overview

The University of Ottawa Cross-Language IR system was built with off-the-shelf components. For translating the queries from French, Spanish, and German into English, several free online machine translation tools were used. Their output was merged in order to allow for variety in lexical choices. All the translations of a title made the title of the translated query; the same was done for the fields description and narrative. For the retrieval part, the SMART IR system (Buckley et al., 2000) was tested with many different weighting schemes for indexing the collection and the queries. The weighting schemes are combinations of term frequency, collection frequency, and length normalization components. One scheme in particular was used in the submissions because it proved to have much better performance than other combinations. For weighting document terms we used term frequency normalized by the maximum value and probabilistic collection frequency weighting with cosine normalization. For queries we used non-normalized term frequency and inverse document frequency weighting. For all languages involved in the task, this combination worked very well when all the fields of the query were used (title, description, and narrative); it worked well with title plus description, and not as well with title only.

3. Translation

For translating the topics into English we use several online MT tools. The idea behind using multiple translations is that they might provide more variety of words and phrases, therefore improving the retrieval performance. The seven online MT systems that we used for translating from Spanish, French, and German were:

1. http://www.google.com/language_tools?hl=en
2. http://www.babelfish.altavista.com
3. http://freetranslation.com
4. http://www.wordlingo.com/en/products_services/wordlingo_translator.html
5. http://www.systranet.com/systran/net
6. http://www.online-translator.com/srvurl.asp?lang=en
7. http://www.freetranslation.paralink.com

For the Czech language topics we were able to find only one online MT system:
http://intertran.tranexp.com/Translate/result.shtml

We combined their outputs by simply concatenating all the translations. All seven translations of a title made the title of the translated query; the same was done for the description and narrative fields. An example of combined output, for the French example used above, is:

<top>
<num> 1159
<title> surviving children in Sweden

surviving children in Sweden
The children survivors in Sweden
surviving children in Sweden
surviving children in Sweden
The surviving children in Sweden
surviving children in Sweden
<desc> Descriptions of the mechanisms of survival of the children born between
1930 and 1933 who passed the war in concentration camps or hidden and who currently live in Sweden.
Descriptions of the mechanisms of survival of the children born between 1930 and 1933 who passed the war in concentration camps or hidden and who currently live in Sweden.
Descriptions of the survival mechanisms of the born children between 1930 and 1933 that passed the war in co ncentration camps or hidden and that live currently in Sweden.
Descriptions of the mechanisms of survival of the children born between 1930 and 1933 who passed the war in concentration camps or hidden and who currently live in Sweden.
Descriptions of the mechanisms of survival of the children born between 1930 and 1933 who passed the war in concentration camps or hidden and who currently live in Sweden.
Descriptions of the mechanisms of survival of the children been born between 1930 and 1933 which crossed war in concentration camps or hidden and that live in Sweden nowadays.
Descriptions of the mechanisms of survival of the children born between 1930 and 1933 who passed the war in concentration camps or hidden and who currently live in Sweden.
<narr>
</top>

4. Retrieval

We used the SMART Information Retrieval (IR) system, originally developed at Cornell University in the 1960s. SMART is based on the vector space model of information retrieval (Salton, 1989). It generates weighted term vectors for the document collection. SMART preprocesses the documents by tokenizing the text into words, removing common words that appear on its stop-list, and performing stemming on the remaining words to derive a set of terms. When the IR server executes a user query, the query terms are also converted into weighted term vectors. Vector inner-product similarity computation is then used to rank documents in decreasing order of their similarity to the user query.

The newest version of SMART (version 11) offers many state-of-the-art options for weighting the terms in the vectors. Each term-weighting scheme is described as a combination of term frequency, collection frequency, and length normalization components (Salton and Buckley, 1988). The description of each component is:

- **term frequency component**

Let *tf* denote the term frequency of a term *t*; then *new_tf* weights the terms according to the following schemes:

none (n) : $new\_tf = tf$

max-norm (m) : $new\_tf = \frac{tf}{\max\_tf}$

augmented normalized (a): $new\_tf = 0.5 + 0.5 * (\frac{tf}{\max\_tf})$

where *max_ tf* is the largest *tf* value in the vector.

log (l): $new\_tf = \ln(tf) + 1.0$

square (s): $new\_tf = tf^2$

- **Merging of collection frequency component**

Let *num_docs*, *coll_freq_of_term*, and *coll _freq* denote the number of documents in the collection, the number of documents in which term t occurs, and the total number of occurrences of the term t in the collection, respectively; then *new_wt* is defined as follows:

none (n): $new\_wt = new\_tf$

inverse document frequency weight (t): $new\_wt = new\_tf * \log(\frac{num\_docs}{coll\_freq\_of\_term})$

probabilistic **(p)**: $new\_wt = new\_tf * \log(\frac{num\_docs\_coll\_freq}{coll\_freq})$

squared (s): $new\_wt = new\_tf * (\log(\frac{num\_docs}{coll\_freq\_of\_term}))^2$

- **Merging of vector normalization**

Let *m* denote the number of entries in the vector, then *norm_wt* is defined as follows:

none (n): $norm\_wt = new\_wt$

sum (s): $norm\_wt = \frac{tf}{\sum_m new\_wt}$

cosine (c): $norm\_wt = \frac{tf}{\sqrt{\sum_m new\_wt^2}}$

In this paper we employ the notation used in SMART to describe the combined schemes: xxx / xxx. The first three characters refer to the weighting scheme used to index the document collection and the last three characters refer to the weighting scheme used to index the query fields. For example, lpc/atc means that lpc was used for documents and atc for queries. lpc would apply log term frequency weighting (l) and probabilistic collection frequency weighting (p) with cosine normalization to the document collection (n). atc would apply augmented normalized term frequency (a) , inverse document frequency weight (t) with cosine normalization (c).

5. Results

Table 1 shows the results of the submitted results on test data. The evaluation measures we report are standard measures computed with the trec_eval script: map (Mean Average Precision) and bpref (Binary Preference, top R judged nonrelevant). The information about what fields of the topics that were indexed in given in the column named Fields: T for title only, TD for title + description, TDN for title + description + narrative. For each run we include an additional description of the experimental settings. For all the required runs we used the indexing scheme mpc/ntn, since it performed best. This weighting scheme worked better when all fields of the topics are indexed. The results for TDN are better than for TD or T. Table 1 does not present baseline results, but we can say that our submitted results was substantially better than the ones submitted by the other six team that participated in the task.

Language	Run	map	bpref	Fields	Description
English	uoEnTDN	0.2176	0.2005	TDN	Weighting scheme: mpc/ntn
Spanish	uoSpTDN	0.1863	0.1750	TDN	Weighting scheme: mpc/ntn
French	uoFrTD	0.1685	0.1599	TD	Weighting scheme: mpc/ntn
English	**uoEnTD**	**0.1653**	**0.1705**	**TD**	**Weighting scheme: mpc/ntn**
German	uoGrTDN	0.1281	0.1331	TDN	Weighting scheme: mpc/ntn

Table 1.Results of the five submitted runs, for topics in English, Spanish, French, and German. The required run (English, title + description) is in bold.

5.1. Comparisons of indexing schemes

Table 2 presents results for various weighting schemes document/topics. There are 3600 possible combinations of weighting schemes: 60 schemes (5 x 4 x 3) for documents and 60 for queries. We tried 240 combinations and we present in the table the results for 15 combinations (the ones, plus some other ones to show the diversity of

the results). mpc/ntn is still the best, but there are a few other weighting schemes that achieve similar performance. Some of the weighting schemes perform best when indexing all the fields of the queries (TDN), some on TD, and some on title only (T). npn/ntn was best for TD and lsn/ntn and lsn/atn are best for T.

In all the presented experiments we use stemming when indexing the collection and translated topics (except Section 5.3). Pseudo-relevance feedback was enables in the SMART system. We don't present the results here, but when we tried using an English lemmatizer (to produce base forms of inflected words) instead of a stemmer, the results were slightly worse for all settings; when using no-stemming during indexing the performance was much worse.

	Weighting scheme	**TDN**		**TD**		**T**	
		map	**bpref**	**map**	**bpref**	**map**	**bpref**
1	mpc/mts	0.2175	0.2004	0.1651	0.1707	0.1175	0.1374
2	mpc/nts	0.2175	0.2004	0.1651	0.1707	0.1175	0.1374
3	mpc/ntn	**0.2176**	0.2005	0.1653	0.1705	0.1174	0.1371
4	npc/ntn	**0.2176**	0.2005	0.1653	0.1705	0.1174	0.1371
5	mpc/mtc	**0.2176**	0.2005	0.1653	0.1705	0.1174	0.1371
6	mpc/ntc	**0.2176**	0.2005	0.1653	0.1705	0.1174	0.1371
7	mpc/mtn	**0.2176**	0.2005	0.1653	0.1705	0.1174	0.1371
8	npn/ntn	0.2116	0.1916	**0.1681**	0.1693	0.1181	0.1350
9	lsn/ntn	0.1195	0.1487	0.1233	0.1433	**0.1227**	0.1395
10	lsn/atn	0.0919	0.1456	0.1115	0.1355	**0.1227**	0.1395
11	asn/ntn	0.0912	0.1295	0.0923	0.1208	0.1062	0.1290
12	snn/ntn	0.0693	0.1327	0.0592	0.1305	0.0729	0.1113
13	sps/ntn	0.0349	0.0979	0.0377	0.1036	0.0383	0.0783
14	nps/ntn	0.0517	0.0940	0.0416	0.0791	0.0474	0.0761
15	mtc/atc	0.1138	0.1514	0.1151	0.1449	0.1108	0.1345

Table 2.Results of the various weighting schemes, for English topics. In bold are the best scores for TDN, TD, and T.

5.2. Comparison of various translations

Table 3 presents results for each translation produced by the seven online MT tools, from Spanish, French, and German into English. The last column is for the combination of all translations, as explained in Section 3. All the results in the table are for mpc/ntn, TDN (except for French where only TD was available).

The translations from German and the one from Czech had many words that were not translated, they were kept unchanged into the English output of the MT tools. These would explain the lower performance for German and Czech. The MT tool number 6 for German seems to obtain better results on the test data than the combination, but this was not the training data. In general, the combination of all translations performs better than the individual translations.

Measure	**Translation**							
	Sp1	**Sp2**	**Sp3**	**Sp4**	**Sp5**	**Sp6**	**Sp7**	**Spanish All**
map	0.1711	0.1756	0.1758	0.1563	0.1756	0.1784	0.1756	0.1863
bpref	0.1708	0.1733	0.1637	0.1563	0.1733	0.1739	0.1733	0.1750
	Fr1	**Fr2**	**Fr3**	**Fr4**	**Fr5**	**Fr6**	**Fr7**	**French All**
map	0.1547	0.1551	0.1526	0.1562	0.1551	0.1575	0.1551	0.1685
bpref	0.1554	0.1559	0.1551	0.1572	0.1559	0.1668	0.1559	0.1599
	Gr1	**Gr2**	**Gr3**	**Gr4**	**Gr5**	**Gr6**	**Gr7**	**German All**
map	0.1244	0.1238	0.1189	0.1232	0.1239	0.1491	0.1238	0.1281
bpref	0.1281	0.1286	0.1344	0.1279	0.1287	0.1633	0.1287	0.1331
	Czech							
map	0.1166							
bpref	0.1310							

Table 3.Results on the output of each Machine Translation system. Spanish, French, and German. Czech.

5.3. Results on phonetic transcriptions

In Table 4 we present results for an experiment where the text of the collection and the queries were transcribed into phonetic form and split into n-grams (groups of n sounds, n = 4 in our case) that we used for indexing (without stemming). The phonetic n-grams were produced by University of Waterloo's group. See their CLEF 2005 paper for more details.

The phonetic form might help compensate for the speech recognition errors made when the collection was produced. When the fields TD were indexed, the results are interesting. The map scores are higher than the previous results on the text form of the documents and queries (up to 28% for the translations from French), while the bpref scores are lower. When combining phonetic and text forms (by simply indexing both phonetic n-grams and text), the result are only slightly improved.

Language	map	bpref	Fields	Description
English	0.1276	0.1117	T	Phonetic, mpc/ntn
English	0.2550	0.1492	TD	Phonetic, mpc/ntn
English	0.1245	0.1198	T	Phonetic+Text, mpc/ntn
English	0.2590	0.1585	TD	Phonetic+Text, mpc/ntn
Spanish	0.1395	0.1050	T	Phonetic, mpc/ntn
Spanish	0.2653	0.1549	TD	Phonetic, mpc/ntn
Spanish	0.1443	0.1108	T	Phonetic+Text, mpc/ntn
Spanish	0.2669	0.1576	TD	Phonetic+Text, mpc/ntn
French	0.1251	0.1005	T	Phonetic, mpc/ntn
French	0.2726	0.1747	TD	Phonetic, mpc/ntn
French	0.1254	0.1023	T	Phonetic+Text, mpc/ntn
French	0.2833	0.1841	TD	Phonetic+Text, mpc/ntn
German	0.1163	0.1150	T	Phonetic, mpc/ntn
German	0.2356	0.1568	TD	Phonetic, mpc/ntn
German	0.1187	0.1159	T	Phonetic+Text, mpc/ntn
German	0.2324	0.1601	TD	Phonetic+Text, mpc/ntn
Czech	0.0776	0.0897	T	Phonetic, mpc/ntn
Czech	0.1647	0.1499	TD	Phonetic, mpc/ntn
Czech	0.0805	0.0951	T	Phonetic+Text, mpc/ntn
Czech	0.1695	0.1491	TD	Phonetic+Text, mpc/ntn

Table 4.Results on phonetic n-grams, and combination text plus phonetic transcripts for topics in English, and the translations from Spanish, French, German, and Czech. All the runs in this table use mpc/ntn.

5.4. Manual summaries and keywords

Table 5 presents the results when all the fields of the document collection were used: the manual keywords and manual summaries in addition to the ASR transcripts and the automatic keywords.

Language	map	bpref	Fields	Description
English	0.4647	0.3660	TDN	Weighting scheme: mpc/ntn, Manual fields
Spanish	0.3811	0.2988	TDN	Weighting scheme: mpc/ntn, Manual fields
French	0.3496	0.2864	TD	Weighting scheme: mpc/ntn, Manual fields
German	0.2513	0.2656	TDN	Weighting scheme: mpc/ntn, Manual fields
Czech	0.2338	0.2251	TDN	Weighting scheme: mpc/ntn, Manual fields

Table 5.Results of indexing all the fields of the collections: the manual keywords and summaries, in addition to the ASR transcripts. Again we report results of mpc/ntn scheme because they are the best.

6. Discussion

We obtained the best retrieval results among the seven teams that participated in this track. We believe that the improved performance is due to the choice of the weighting scheme used for indexing the document and query

terms. Table 2 shows that performance varies a lot with the weighting scheme; it can be much lower for the some of the classic indexing schemes.

The idea of using multiple translations proves to be good. More variety in the translations would be beneficial. The online MT systems that we used are rule-based system. Adding translations by statistical MT tools might help, since they produce radically different translations.

On the manual data, the best map score we obtained is 46%, for English topics. On automatic data the best result is 21% map score. This difference shows that the poor quality of the ASR transcripts severely hurts the performance of IR systems on this collection. In future work we plan to investigate methods of removing or correcting some of the speech recognition errors in the ASR transcripts using the method of Inkpen and Désilets (2005).

References

Chris Buckley, Gerard Salton, and James Allan. Automatic retrieval with locality information using SMART. In *Proceedings of the First Text REtrieval Conference (TREC-1)*, pages 59–72. NIST Special Publication 500-207, March 1993.

Douglas W. Oard, Dagobert Soergel, David Doermann, Xiaoli Huang, G. Craig Murray, Jianqiang Wang, Bhuvana Ramabhadran, Martin Franz, Samuel Gustman. 2004. Building an Information Retrieval Test Collection for Spontaneous Conversational Speech, in *Proceedings of SIGIR 2004*.

Diana Inkpen and Alain Désilets. 2005. Semantic Similarity for Detecting Recognition Errors in Automatic Speech Transcripts, in *Proceedings of EMNLP 2005*, Vancouver, Canada, October 2005.

Gerard Salton and Chris Buckley. 1988. Term-weighting approaches in automatic retrieval. *Information Processing and Management*, 24(5):513-523.

Gerard Salton. 1989. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley Publishing Company.