Using various indexing schemes and multiple translations in the CL-SR task at CLEF 2005

Diana Inkpen, Muath Alzghool, and Aminul Islam

School of Information Technology and Engineering
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 [1], with several weighting schemes for indexing the documents and the queries. One scheme in particular led 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.

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. See [3] and [7] for more details. For the documents (segments), we indexed only the ASRTEXT2004A field and the keywords automatically extracted from it. This field contains ASR transcripts of the audio segments, with 38% word error rate. In Section 5.4 we also index the metadata for each segment (manual summaries, thesaurus terms, and person names). The topics provided with the collection were created in English from actual user requests and then translated into Czech, German, French, and Spanish by native speakers.

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 description and narrative fields. For the retrieval part, the SMART IR system [1] 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. For all languages involved in the task, the best results were obtained when all the fields of the queries were used (title, description, and narrative); it still worked well with title plus description, and not as well with title only.

3 Translation

For translating the topics into English we used 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

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 by the CLEF organizers, due to lack of time. An example of French query 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>
```

We combined the outputs of the MT systems 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 above French query, 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 concentration 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>
```

We used the combined topics for all experiments except those described in section 5.2 which investigate the effectiveness of the individual translations.

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 [5]. 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 [6]. The description of each component is:

Term Frequency Component

Let *tf* denote the term frequency of a term *t* in the document; 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): $new\_tf = 0.5 + 0.5 \cdot \frac{tf}{\max\_tf}$

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 *N* and *df* denote the number of documents in the collection and the number of documents in which term t occurs, respectively; then *new_wt* is defined as follows:

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

inverse document frequency weight (t): $new\_wt = new\_tf \cdot \log \frac{N}{df}$

probabilistic (p): $new\_wt = new\_tf \cdot \log \frac{N - df}{df}$

squared (s): $new\_wt = new\_tf \cdot (\log \frac{N}{df})^2$

Merging of Vector Normalization

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

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

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

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

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 (c). 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 the test data. The evaluation measure we report is standard measures computed with the trec_eval script: MAP (Mean Average Precision). The information about what fields of the topics 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 lnn.ntn, since it performed best on the training data. This weighting scheme worked better when all fields of the topics are indexed. The results for TDN are slightly better than for TD and better than for T. Table 1 does not present baseline results, but we can say that our submitted results were better than the ones submitted by the other six teams that participated in the task, on the required run.

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

Language	Run	MAP	Fields	Description
English	uoEnTDN	0.1366	TDN	Weighting scheme: lnn.ntn
English	**uoEnTD**	**0.1313**	**TD**	**Weighting scheme: lnn.ntn**
French	uoFrTD	0.1275	TD	Weighting scheme: lnn.ntn
Spanish	uoSpTDN	0.1156	TDN	Weighting scheme: lnn.ntn
German	uoGrTDN	0.0936	TDN	Weighting scheme: lnn.ntn

5.1 Comparison 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 best ones, plus some other ones to show the diversity of the results). lnn.ntn seems to be the best, and there might be 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). lnn.ntn is best for TDN and TD and lsn.ntn and lsn.atn are best for T. (Note that for mpc.ntn and other schemes that contain the probabilistic term “p”, due to a minor bug in Smart, some documents were returned as answer to the same query more than once. In this case, we preprocessed the results to eliminate the duplicates and kept the first 1000 distinct results for each query, to retrieve the same number of documents per query as in the other experiments).

In all the presented experiments we use stemming when indexing the collection and the translated topics (except Section 5.3). 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. Relevance feedback was not enabled in the SMART system.

5.2 Comparison of various translations

Table 3 presents results for each translation produced by the seven online MT tools, from French, Spanish, 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 lnn.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. This would explain the lower performance for German and Czech. The MT tool number 6 for French and German seems to obtain better results on the test data than the combination, but this was not the case on the training data. In general, the combination of all translations performs better than the individual translations.

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

	Weighting scheme	**TDN**	**TD**	**T**
1	lnn.ntn	**0.1366**	**0.1313**	0.1207
2	lnc.ntn	0.1362	0.1214	0.1094
3	mpc.ntn	0.1283	0.1219	0.1107
4	npc.ntn	0.1283	0.1219	0.1107
5	mpc.mtc	0.1283	0.1219	0.1107
6	mpc.mts	0.1282	0.1218	0.1108
7	mpc.nts	0.1282	0.1218	0.1108
8	npn.ntn	0.1258	0.1247	0.1118
9	lsn.ntn	0.1195	0.1233	**0.1227**
10	lsn.atn	0.0919	0.1115	**0.1227**
11	asn.ntn	0.0912	0.0923	0.1062
12	snn.ntn	0.0693	0.0592	0.0729
13	sps.ntn	0.0349	0.0377	0.0383
14	nps.ntn	0.0517	0.0416	0.0474
15	mtc.atc	0.1138	0.1151	0.1108

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

Measure	**Translation**							
	Fr1	**Fr2**	**Fr3**	**Fr4**	**Fr5**	**Fr6**	**Fr7**	**French**
MAP	0.1209	0.1196	0.1169	0.1200	0.1196	0.1288	0.1196	0.1275
	Sp1	**Sp2**	**Sp3**	**Sp4**	**Sp5**	**Sp6**	**Sp7**	**Spanish**
MAP	0.1130	0.1142	0.1016	0.0991	0.1140	0.1116	0.1142	0.1156
	Gr1	**Gr2**	**Gr3**	**Gr4**	**Gr5**	**Gr6**	**Gr7**	**German**
MAP	0.0908	0.0906	0.0853	0.0900	0.0907	0.0994	0.0906	0.0936
	Czech							
MAP	0.0822							

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 the University of Waterloo's group. See [2] for more details.

We wanted to test the hypothesis that 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 better than when only T is indexed. When combining phonetic and text forms (by simply indexing both phonetic n-grams and text), the result improved compared to using only the phonetic forms. But the MAP scores are lower than the results on the text form of the documents and queries.

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

Language	MAP	Fields	Description
English	0.0986	T	Phonetic
English	0.1019	TD	Phonetic
English	0.0981	T	Phonetic+Text
English	0.1066	TD	Phonetic+Text
French	0.0931	T	Phonetic
French	0.1052	TD	Phonetic
French	0.0929	T	Phonetic+Text
French	0.1072	TD	Phonetic+Text
Spanish	0.0898	T	Phonetic
Spanish	0.0972	TD	Phonetic
Spanish	0.0948	T	Phonetic+Text
Spanish	0.1009	TD	Phonetic+Text
German	0.0744	T	Phonetic
German	0.0782	TD	Phonetic
German	0.0746	T	Phonetic+Text
German	0.0789	TD	Phonetic+Text
Czech	0.0479	T	Phonetic
Czech	0.0583	TD	Phonetic
Czech	0.0510	T	Phonetic+Text
Czech	0.0614	TD	Phonetic+Text

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. The retrieval performance improved a lot, for all the languages. The MAP score jumped from 0.1366 to 0.277 for English, TDN, with the lnn.ntn weighting scheme. The score doubles for English queries, and for the queries translated from the other languages.

Table 5.Results of indexing all the fields of the collections: the manual keywords and summaries, in addition to the ASR transcripts (lnn.ntn).

Language	**MAP**	**Fields**	**Description**
English	0.2771	TDN	Manual fields included
French	0.2473	TD	Manual fields included
Spanish	0.2267	TDN	Manual fields included
German	0.1852	TDN	Manual fields included
Czech	0.1562	TDN	Manual fields included

6 Discussion

We obtained the best retrieval results on the required run among the seven teams that participated in this track. We tried various weighting scheme for indexing the document and query terms. Table 2 shows that performance varies with the weighting scheme; it can be lower for the some of the classic indexing schemes.

In this paper we presented the results on the test queries, but our conclusions also applied on the training queries.

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 systems. 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 around 27%, for English topics. On automatic data the best result is around 13% 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 semantic coherence measures [4].

References

1. C. Buckley, G. Salton, and J. 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).

2. C. L. A. Clarke : Waterloo Experiments for the CLEF05 SDR Track, in Working Notes for the CLEF 2005 Workshop, September, Vienna, Austria (2005)
3. D. W. Oard, D. Soergel, D. Doermann, X. Huang, G. C. Murray, J. Wang, B. Ramabhadran, M. Franz and S. Gustman : Building an Information Retrieval Test Collection for Spontaneous Conversational Speech, in Proceedings of SIGIR (2004)
4. D. Inkpen and A. Désilets : Semantic Similarity for Detecting Recognition Errors in Automatic Speech Transcripts, in Proceedings of EMNLP 2005, Vancouver, Canada, October (2005)
5. G. Salton : Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. Addison-Wesley Publishing Company (1989)
6. G. Salton and C. Buckley : Term-weighting approaches in automatic retrieval. Information Processing and Management, 24(5):513-523 (1988)
7. R. W. White, D. W. Oard, G. J. F. Jones, D. Soergel, X. Huang : Overview of the CLEF-2005 Cross-Language Speech Retrieval Track, in Working Notes for the CLEF 2005 Workshop, September, Vienna, Austria (2005)