Model Fusion Experiments for the Cross Language Speech Retrieval Task at CLEF 2007

Muath Alzghool and Diana Inkpen

School of Information Technology and Engineering
University of Ottawa
{alzghool, diana}@site.uottawa.ca

Abstract This paper presents the participation of the University of Ottawa group in the Cross-Language Speech Retrieval (CL-SR) task at CLEF 2007. We present the results of the submitted runs for the English collection. We have used two Information Retrieval systems in our experiments: SMART and Terrier, with two query expansion techniques: one based on a thesaurus and the second one based on blind relevant feedback. We proposed two novel data fusion methods for merging the results of several models (retrieval schemes available in SMART and Terrier). Our experiments showed that the combination of query expansion methods and data fusion methods helps to improve the retrieval performance. We also present cross-language experiments, where the queries are automatically translated by combining the results of several online machine translation tools. Experiments on indexing the manual summaries and keywords gave the best retrieval results.

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
Data Fusion, Retrieval Models, Query Expansion

1 Introduction

This paper presents the third participation of the University of Ottawa group in the Cross-Language Speech Retrieval (CL-SR) track, at CLEF 2007. We present our systems, followed by results for the submitted runs for the English collection. We present results for many additional runs for the English collection. We experimented with many possible weighting schemes for indexing the documents and the queries, and with several query expansion techniques. Several researchers in the literature have explored the idea of combining the results of different retrieval strategies, different document representations and different query representations; the motivation is that each technique will retrieve different sets of relevant documents; therefore combining the results could produce a better result than any of the individual techniques. We propose new data fusion techniques for combining the results of different Information Retrieval (IR) schemes. We applied our data fusion techniques to monolingual settings and to cross-language settings where the queries are automatically translated from French and Spanish into English by combining the results of several online machine translation (MT) tools. At the end we present the best results, when manual summaries and manual keywords were indexed.

2 System Description

The University of Ottawa Cross-Language Information Retrieval systems were built with off-the-shelf components. For the retrieval part, the SMART [3, 11] IR system and the Terrier [2, 10] IR system were tested with many different weighting schemes for indexing the collection and the queries.

SMART was originally developed at Cornell University in the 1960s. SMART is based on the vector space model of information retrieval. We used nnn.ntn, ntn.ntn, lnn.ntn, ann.ntn, ltn.ntn, atn.ntn, ntn.nnn, nnc.ntc,

ntc.ntc, ntc.nnc, lnc.ntc, anc.ntc, ltc.ntc, atc.ntc weighting schemes [3 ,11]; lnn.ntn performs very well in CLEF-CLSR 2005 and 2006 [6,1] .

Terrier was originally developed at University of Glasgow. It is based on Divergence from Randomness models (DFR) where IR is seen as a probabilistic process [2, 10]. We experimented with the In(exp)C2 weighting model, one of Terrier's DFR-based document weighting models.

For translating the queries from French and Spanish into English, several free online machine translation tools were used. The idea behind using multiple translations is that they might provide more variety of words and phrases, therefore improving the retrieval performance. Seven online MT systems [6] were used for translating from Spanish and from French into English. 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. We used the combined topics for all the cross-language experiments reported in this paper.

We have used two query expansion methods. The first one is based on the Shoah Visual History Foundation thesaurus provided with the Mallach collection; our method adds two items and their alternatives (synonyms) from the thesaurus, based on the similarity between the thesaurus terms and the title field for each topic. More specifically, to select two items from the thesaurus, we used SMART with the title of each topic as query and the thesaurus terms as documents, using the weighting scheme lnn.ntn. After computing the similarity, the top two thesaurus terms were added to the topic; for these terms all the alternative terms was also added to the topic. For example, in topic 3005, the title is "Death marches", and the most similar terms from the thesaurus are "death marches" and "deaths during forced marches"; the alternative terms for theses terms are "death march" and "Todesmärsche". Table 1 shows two entries from the thesaurus; each entry contains six types of fields: name − contains a unique numeric code for each entry, label − a phrase or word which represents the entry, alt-label − contains the alternative phrase or the synonym for the entry, usage − contains the usage or the definition of the entry. There are two more relations in the thesaurus: is-a and of-type, which contain the numeric code of the entry involved in the relation. The second query expansion method extracts the most informative terms from the top-returned documents as the expanded query terms. In this expansion process, 12 terms from the returned documents (the top 15 documents) were added to the topic, based on Bose-Einstein 1 model (Bo1) [4,10]; we have put a restriction on the new terms: their document frequency must be less than the maximum document frequency in the title of the topic. The aim of this restriction is avoid more-general terms being added to the topic. Any term that satisfies this restriction will be a part of the new topic. We have also up weighted the title terms five times higher than the other terms in the topic.

Table 1. The top two entries from the thesaurus that are similar to the topic title "Death marches".

```
<keyword>
    <name>9125</name>
    <alt-label>death march</alt-label>
    <alt-label> Todesmärsche</alt-label>
    <broader-term>15445</broader-term>
    <label>death marches</label>
    <of-type>5289</of-type>
    <usage>Forced marches of prisoners over long distances, under heavy guard and
extremely harsh conditions. (The term was probably coined by concentration camp
prisoners.)</usage>
</keyword>
<keyword>
    <name>15460</name>
    <broader-term>15445</broader-term>
    <label>deaths during forced marches</label>
    <of-type>4109</of-type>
    <usage>The daily experience of individuals with death during forced marches
that was not the result of executions, punishments, arbitrary killings or
suicides.</usage>
</keyword>
```

For the data fusion part, we proposed two methods that use the sum of normalized weighted similarity scores of 15 different IR schemes as shown in the following formulas :

$$Fusion1 = \sum_{i \in IR\ schems} [W_r^4(i) + W_{MAP}^3(i)] * NormSim_i \qquad (1)$$

$$Fusion2 = \sum_{i \in IR\ schems} W_r^4(i) * W_{MAP}^3(i) * NormSim_i \qquad (2)$$

where $W_r(i)$ and $W_{MAP}(i)$ are experimentally determined weights based on the recall (the number of relevant documents retrieved) and precision (MAP score) values for each IR scheme computed on the training data. For example, suppose that two retrieval runs r1 and r2 give 0.3 and 0.2 (respectively) as MAP scores on training data; we normalize these scores by dividing them by the maximum MAP value: then $W_{MAP}(r1)$ is 1 and $W_{MAP}(r2)$ is 0.66 (then we compute the power 3 of these weights, so that one weight stays 1 and the other one decreases; we chose power 3 for MAP score and power 4 for recall, because the MAP is more important than the recall). We hope that when we multiply the similarity values with the weights and take the summation over all the runs, the performance of the combined run will improve. $NormSim_i$ is the normalized similarity for each IR scheme. We did the normalization by dividing the similarity by the maximum similarity in the run. The normalization is necessary because different weighting schemes will generate different range of similarity values, so a normalization method should applied to each run. Our method is differed than the work done by Fox and Shaw in 1994 [5] and Lee in 1995 [7]; they combined the results by taking the summation of the similarity scores without giving any weight to each run. In our work we weight each run according to the precision and recall on the training data.

3 Experimental Results

3.1 Submitted Runs

Table 2 shows the results of the submitted results on the test data (33 queries). The evaluation measure we report is the standard measure computed with the trec_eval script (version 8): MAP (Mean Average Precision) and Recall. The information about what fields of the topic were indexed is 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 and which document fields were indexed; [8,9] give more information about the training and test data . For the uoEnTDtManF1 and uoEnTDtQExF1 runs we used the Fusion1 formula for data fusion; and for uoEnTDtQExF2, uoFrTDtF2, and uoEsTDtF2 we used the Fuison2 formula for data fusion. We used blind relevance feedback and query expansion from the thesaurus for the uoEnTDtManF1, uoEnTDtQExF1, and uoEnTDtQExF2 runs; we didn't use any query expansion techniques for uoFrTDtF2 and uoEsTDtF2.

Our required run, English TD, obtained a MAP score of 0.0855. Comparing this result to the median and average of all runs submitted by all the teams that participated in the track (0.0673, 0.0785) [9], our result was significantly better (based on a two-tailed Wilcoxon Signed-Rank Test for paired samples at $p < 0.05$ across the 33 evaluation topics) with a relative improvement of 21% and 8%; there is a small improvement using Fusion1 (uoEnTDtQExF1) over Fusion2 (uoEnTDtQExF2), but this improvement is not significant.

Table 2. Results of the five submitted runs, for topics in English, French, and Spanish. The required run (English, title + description) is in bold.

Runs	MAP	Recall	Fields	Description
uoEnTDtManF1	0.2761	1832	TD	English: Fusion 1, query expansion methods, fields: MANUALKEYWORD + SUMMARY
uoEnTDtQExF1	**0.0855**	1333	TD	English: Fusion 1, query expansion methods, fields: ASRTEXT2004A + AUTOKEYWORD2004A1, A2
uoEnTDtQExF2	0.0841	1336	TD	English: Fusion 2, query expansion methods, fields: ASRTEXT2004A + AUTOKEYWORD2004A1, A2
uoFrTDtF2	0.0603	1098	TD	French : Fusion 2, fields: ASRTEXT2004A + AUTOKEYWORD2004A1, A2
uoEsTDtF2	0.0619	1171	TD	Spanish : Fusion 2, fields: ASRTEXT2004A + AUTOKEYWORD2004A1, A2

3.2 Comparison of Systems and Query Expansion Methods

In order to compare between different methods of query expansion and a base run without query expansion, we selected the base run with the weighting scheme lnn.ntn, topic fields title and description, and document fields ASRTEXT2004A, AUTOKEYWORD2004A1, and AUTOKEYWORD2004A2. We used the two techniques for query expansion, one based on the thesaurus and the other one on blind relevance feedback (denoted Bo1 in Table 3). We present the results (MAP scores) with and without query expansion, and with the combination of both query expansion methods, on the test and training topics. According to Table 3, we note that both methods help to improve the retrieval results, but the improvement is not significant on the training and test data; also the combination of the two methods helps to improve the MAP score on the training data (not significantly), but not on the test data.

Table 3. Results (MAP scores) for Terrier and SMART, with or without relevance feedback, for English topics (using the TD query fields).

	System	**Training**	**Test**
1	lnn.ntn	0.0906	0.0725
2	lnn.ntn +thesaurus	0.0941	0.0730
3	lnn.ntn +Bo1	0.0954	0.0811
4	lnn.ntn+ thesaurus+ Bo1	0.0969	0.0799

3.3 Experiments using Data Fusion

We applied the data fusion methods described in section 2 to 14 runs produced by SMART and one run produced by Terrier; all runs was produced using a combination of the two methods of query expansion as described in section 2. Performance results for each single run and fused runs are presented in Table 4, in which % change is given with respect to the run providing better effectiveness in each combination on the training data. The Manual English column represents the results when only the manual keywords and the manual summaries were used for indexing the documents using English topics, the Auto-English column represents the results when automatic fields are indexed from the documents (ASRTEXT2004A, and AUTOKEYWORD2004A1, A2) using English topics. For cross-languages experiments the results are represented in the columns Auto-French, and Auto-Spanish.

Data fusion helps to improve the performance (MAP score) on the test data The best improvement using data fusion (Fusion1) was on the French cross-language experiments with 21.7%, which is statistically significant while on monolingual the improvement was only 6.5% which is not significant. Also, there is an improvement in the number of relevant documents retrieved (recall) for all the experiments, except Auto-French on the test data, as shown in Table 5. We computed these improvements relative to the results of the best single-model run, as measured on the training data. This supports our claim that data fusion improves the recall by bringing some new documents that were not retrieved by all the runs. On the training data, the Fusion2 method gives better results than Fusion1 for all cases except on Manual English, but on the test data Fusion1 is better than Fusion2. In general, the data fusion seems to help, because the performance on the test data in not always good for weighting schemes that obtain good results on the training data, but combining models allows the best-performing weighting schemes to be taken into consideration.

The retrieval results for the translations from French were very close to the monolingual English results, especially on the training data, but on the test data the difference was significantly worse. For Spanish, the difference was significantly worse on the training data, but not on the test data.

Experiments on manual keywords and manual summaries showed high improvements, the MAP score jumped from 0.0855 to 0.2761 on the test data.

Table 4. Results (MAP scores) for 15 weighting schemes using Smart and Terrier (the In(exp)C2 model), and the results for the two Fusions Methods. In bold are the best scores for the 15 single runs on the training data and the corresponding results on the test data. Underlined are the results of the submitted runs.

Weighting scheme	Manual English		Auto-English		Auto-French		Auto-Spanish	
	Training	Test	Training	Test	Training	Test	Training	Test
nnc.ntc	0.2546	0.2293	0.0888	0.0819	0.0792	0.055	0.0593	0.0614
ntc.ntc	0.2592	0.2332	0.0892	0.0794	0.0841	0.0519	0.0663	0.0545
lnc.ntc	0.2710	0.2363	0.0898	0.0791	0.0858	0.0576	0.0652	0.0604
ntc.nnc	0.2344	0.2172	0.0858	0.0769	0.0745	0.0466	0.0585	0.062
anc.ntc	0.2759	0.2343	0.0723	0.0623	0.0664	0.0376	0.0518	0.0398
ltc.ntc	0.2639	0.2273	0.0794	0.0623	0.0754	0.0449	0.0596	0.0428
atc.ntc	0.2606	0.2184	0.0592	0.0477	0.0525	0.0287	0.0437	0.0304
nnn.ntn	0.2476	0.2228	0.0900	0.0852	0.0799	0.0503	0.0599	0.061
ntn.ntn	0.2738	0.2369	0.0933	0.0795	0.0843	0.0507	0.0691	0.0578
lnn.ntn	0.2858	0.245	**0.0969**	**0.0799**	0.0905	0.0566	0.0701	0.0589
ntn.nnn	0.2476	0.2228	0.0900	0.0852	0.0799	0.0503	0.0599	0.061
ann.ntn	0.2903	0.2441	0.0750	0.0670	0.0743	0.038	0.057	0.0383
ltn.ntn	0.2870	0.2435	0.0799	0.0655	0.0871	0.0522	0.0701	0.0501
atn.ntn	0.2843	0.2364	0.0620	0.0546	0.0722	0.0347	0.0586	0.0355
In(exp)C2	**0.3177**	**0.2737**	0.0885	0.0744	**0.0908**	**0.0487**	**0.0747**	**0.0614**
Fusion 1	0.3208	0.2761	0.0969	0.0855	0.0912	0.0622	0.0731	0.0682
% change	1.0%	0.9%	0.0%	6.5%	0.4%	21.7%	-2.2%	10.0%
Fusion 2	0.3182	0.2741	0.0975	0.0842	0.0942	0.0602	0.0752	0.0619
% change	0.2%	0.1%	0.6%	5.1%	3.6%	19.1%	0.7%	0.8%

4 Conclusion

We experimented with two different systems: Terrier and SMART, with combining the various weighting schemes for indexing the document and query terms. We proposed two approaches for query expansion, one based on the thesaurus and another one based on blind relevance feedback. The combination of the query expansion methods obtained a small improvement on the training and test data (not statistically significant according to a Wilcoxon signed test).

Our focus this year was on data fusion: we proposed two methods to combine different weighting scheme from different systems, based on weighted summation of normalized similarity measures; the weight for each scheme was based on the relative precision and recall on the training data. Data fusion helps to improve the retrieval significantly for some experiments (Auto-French) and for other not significantly (Manual English).

The idea of using multiple translations proved 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 could produce radically different translations.

Combining query expansion methods and data fusion helped to improve the retrieval significantly comparing to the median and average of all required runs submitted by all the teams that participated in the track.

In future work we plan to investigate more methods of data fusion, removing or correcting some of the speech recognition errors in the ASR content words, and to use speech lattices for indexing.

Table 5. Results (number of relevant documents retrieved) for 15 weighting schemes using Terrier and SMART, and the results for the Fusions Methods. In bold are the best scores for the 15 single runs on training data and the corresponding test data; underlined are the submitted run

Weighting scheme	Manual English		Auto-English		Auto- French		Auto- Spanish	
	Training	Test	Training	Test	Training	Test	Training	Test
nnc.ntc	2371	1827	1726	1306	1687	1122	1562	1178
ntc.ntc	2402	1857	1675	1278	1589	1074	1466	1155
lnc.ntc	2402	1840	1649	1301	1628	1111	1532	1196
ntc.nnc	2354	1810	1709	1287	1662	1121	1564	1182
anc.ntc	2405	1858	1567	1192	1482	1036	1360	1074
ltc.ntc	2401	1864	1571	1211	1455	1046	1384	1097
atc.ntc	2387	1858	1435	1081	1361	945	1255	1011
nnn.ntn	2370	1823	1740	1321	**1748**	**1158**	1643	1190
ntn.ntn	2432	1863	1709	1314	1627	1093	1502	1174
lnn.ntn	2414	1846	1681	1325	1652	1130	1546	1194
ntn.nnn	2370	1823	**1740**	**1321**	1748	1158	**1643**	**1190**
ann.ntn	2427	1859	1577	1198	1473	1027	1365	1060
ltn.ntn	2433	1876	1582	1215	1478	1070	1408	1134
atn.ntn	2442	1859	1455	1101	1390	975	1297	1037
In(exp)C2	**2638**	**1823**	1624	1286	1676	1061	1631	1172
Fusion 1	2645	1832	1745	1334	1759	1147	1645	1219
% change	0.3%	0.5 %	0.3%	1.0%	0.6%	-1.0%	0.1%	2.4%
Fusion 2	2647	1823	1727	1337	1736	1098	1631	1172
% change	0.3%	0.0%	0.8%	1.2%	-0.7%	-5.5%	-0.7%	-1.5%

References

1. Alzghool M. and Inkpen D. : Experiments for the Cross Language Speech Retrieval Task at CLEF 2006. In Proceedings of CLEF 2006, Lecture Notes in Computer Science, Springer-Verlag 4730, 2007, pp.778-785
2. Amati, G. and van Rijsbergen, C. J. : Probabilistic models of information retrieval based on measuring the divergence from randomness. ACM Transactions on Information Systems, Vol. 20, No. 4, October (2002) 357–389.
3. Buckley C., Salton G., and Allan J.: Automatic retrieval with locality information using SMART. In Text REtrieval Conference (TREC-1), March (1993) 59–72.
4. Carpineto C., de Mori R., Romano G., and Bigi B.: An information-theoretic approach to automatic query expansion. ACM Transactions on Information Systems (TOIS), Vol. 19, No. 1, January (2001) 1-27.
5. Fox, E.A. and Shaw, J.A. (1994). Combination of multiple searches. Proceedings of the Third Text REtrieval Conference (TREC-3). National Institute of Standards and Technology Special Publication 500-215.
6. Inkpen D., Alzghool M., and Islam A.: Using various indexing schemes and multiple translations in the CL-SR task at CLEF 2005. In Accessing Multilingual Information Repositories, 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005, Vienna, Austria, 21-23 September, (2005).
7. Lee, J.H. (1995). Combining multiple evidence from different properties of weighting schemes. Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.180-188.
8. Oard D.W., Soergel D., Doermann D., Huang X., Murray G.C., Wang J., Ramabhadran B., Franz M. and Gustman S.: Building an Information Retrieval Test Collection for Spontaneous Conversational Speech, in Proceedings of SIGIR, (2004).
9. Oard D.W., J., Jones G. J. F., Pecina P., et al: Overview of the CLEF 2007 cross-language speech retrieval track. In Working Notes of the CLEF- 2007 Evaluation, Budapest, Hungary, (2007).
10. Ounis I., Amati G., Plachouras V., He B., Macdonald C. and Johnson D.: Terrier Information Retrieval Platform. In 27th European Conference on Information Retrieval (ECIR 05), (2005). http://ir.dcs.gla.ac.uk/wiki/Terrier
11. Salton G. and Buckley C.: Term-weighting approaches in automatic retrieval. Information Processing and Management, Vol. 24, No. 5, (1988) 513-523.