Combining Multiple Models for Speech Information Retrieval

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

Abstract

In this article we present a method for combining different information retrieval models in order to increase the retrieval performance in a Speech Information Retrieval task. The formulas for combining the models are tuned on training data. Then the system is evaluated on test data. The task is particularly difficult because the text collection is automatically transcribed spontaneous speech, with many recognition errors. Also, the topics are real information needs, difficult to satisfy. Information Retrieval systems are not able to obtain good results on this data set, except for the case when manual summaries are included.

1. Introduction

Conversational speech such as recordings of interviews or teleconferences is difficult to search through. The transcripts produced with Automatic Speech Recognition (ASR) systems tend to contain many recognition errors, leading to low Information Retrieval (IR) performance (Oard et al., 2007).

Previous research has explored the idea of combining the results of different retrieval strategies; 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 IR models. We applied our data fusion techniques to the Mallach collection (Oard et al., 2007) used in the Cross-Language Speech Retrieval (CLSR) task at Cross-Language Evaluation Forum (CLEF) 2007. The Mallach collection comprises 8104 "documents" which are manually-determined topically-coherent segments taken from 272 interviews with Holocaust survivors, witnesses and rescuers, totalling 589 hours of speech. Figure 1 shows the document structure in CLSR test collection, two ASR transcripts are available for this data, in this work we use the ASRTEXT2004A field provided by IBM research with a word error rate of 38%. Additionally, metadata fields for each document include: two sets of 20 automatically assigned keywords determined using two different kNN classifiers (AK1 and AK2), a set of a varying number of manually-assigned keywords (MK), and a manual 3-sentence summary written by an expert in the field. A set of 63 training topics and 33 test topics were generated for this task. The topics provided with the collection were created in English from actual user requests. Topics were structured using the standard TREC format of Title, Description and Narrative fields. To enable CL-SR experiments the topics were translated into Czech, German, French, and Spanish by native speakers; Figure 2 and 3 show two examples for English and its translation in French respectively. Relevance judgments were generated using a search-guided procedure and standard pooling methods. See (Oard et al., 2004) for full details of the collection design.

We present results on the automatic transcripts for English queries and translated queries (cross-language) for two combination methods; we also present results when manual summaries and manual keywords are indexed.

```
<DOC>
<DOCNO>VHF[IntCode]-[SegId].[SequenceNum]</DOCNO\>
<INTERVIEWDATA>Interviewee name(s) and
birthdate</INTERVIEWDATA>
<NAME>Full name of every person mentioned</NAME>
<MANUALKEYWORD>Thesaurus keywords assigned to the
segment</MANUALKEYWORD>
<SUMMARY>3-sentence segment summary</SUMMARY>
<ASRTEXT2004A>ASR transcript produced in
2004</ASRTEXT2004A>
<ASRTEXT2006A>ASR transcript produced in
2006</ASRTEXT2006A>
<AUTOKEYWORD2004A1>Thesaurus keywords from a kNN
classifier</AUTOKEYWORD2004A1>
<AUTOKEYWORD2004A2>Thesaurus keywords from a second
kNN classifier</AUTOKEYWORD2004A2>
</DOC>
```

Figure 1. Document structure in CL-SR test collection.

```
<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>
```

Figure 2. Example for English topic in CL-SR test collection.

```
<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.
<narr>…
</top>
```

Figure 3. Example for French topic in CL-SR test collection.

2. System Description

Our Cross-Language Information Retrieval systems were built with off-the-shelf components. For the retrieval part, the SMART (Buckley, Salton, &Allan, 1992; Salton &Buckley, 1988) IR system and the Terrier (Amati &Van Rijsbergen, 2002; Ounis et al., 2005) 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 use the standard notation: weighting scheme for the documents, followed by dot, followed by the weighting scheme for the queries, each term-weighting scheme is described as a combination of term frequency, collection frequency, and length normalization components where the schemes are abbreviated according to its components variations (n no normalization, c cosine, t idf, l log, etc.) 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 (Buckley, Salton, &Allan, 1992; Salton &Buckley, 1988); lnn.ntn performs very well in CLEF-CLSR 2005 and 2006 (Alzghool &Inkpen, 2007; Inkpen, Alzghool, &Islam, 2006); lnn.ntn means that lnn was used for documents and ntn for queries according to the following formulas:

$$weight_{lnn} = \ln(tf) + 1.0 \qquad (1)$$

$$weight_{ntn} = tf \times \log \frac{N}{n_t} \qquad (2)$$

where tf denotes the term frequency of a term t in the document or query, N denotes the number of documents in the collection, and n_t denotes the number of documents in which the term t occurs.

Terrier was originally developed at the University of Glasgow. It is based on Divergence from Randomness models (DFR) where IR is seen as a probabilistic process (Amati &Van Rijsbergen, 2002; Ounis et al., 2005). We experimented with the In_expC2 (Inverse Expected Document Frequency model with Bernoulli after-effect and normalization) weighting model, one of Terrier's DFR-based document weighting models.

Using the In_expC2 model, the relevance score of a document d for a query q is given by the formula:

$$sim(d,q) = \sum_{t \in q} qtf \,.\, w(t,d) \qquad (3)$$

where *qtf* is the frequency of term *t* in the query *q*, and *w(t,d)* is the relevance score of a document *d* for the query term *t*, given by:

$$w(t,d) = \left(\frac{F+1}{n_t \times (tfn_e + 1)}\right) \times \left(tfn_e \times \log_2 \frac{N+1}{n_e + 0.5}\right) \qquad (4)$$

where

-*F* is the term frequency of t in the whole collection.

-*N* is the number of document in the whole collection.

-n_t is the document frequency of t.

-n_e is given by $n_e = N \times \left(1 - \left(\frac{1-n_t}{N}\right)^F\right)$ (5)

- ***tfn**$_e$* is the normalized within-document frequency of the term *t* in the document *d.* It is given by the normalization 2 (Amati &Van Rijsbergen, 2002; Ounis et al., 2005):

$$tfn_e = tf \times \log_e \left(1 + c \times \frac{avg\_l}{l}\right) \qquad (6)$$

where c is a parameter, tf is the within-document frequency of the term t in the document d, l is the document length, and avg_l is the average document length in the whole collection.

We estimated the parameter c of the Terrier's normalization 2 formula by running some experiments on the training data, to get the best values for c depending on the topic fields used. We obtained the following values: c=0.75 for queries using the Title only, c=1 for queries using the Title and Description fields, and c=1 for queries using the Title, Description, and Narrative fields. We select the c value that has a best MAP score according to the training data.

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 (Inkpen, Alzghool, &Islam, 2006) 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 propose two methods for combining IR models. We use the sum of normalized weighted similarity scores of 15 different IR schemes as shown in the following formulas:

$$Fusion1 = \sum_{i \in IR\ schems} [W_r^4(i) + W_{MAP}^3(i)] * NormSim_i \qquad (7)$$

$$Fusion2 = \sum_{i \in IR\ schems} W_r^4(i) * W_{MAP}^3(i) * NormSim_i \qquad (8)$$

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), and Lee in (1995); 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

We applied the data fusion methods described in section 2 to 14 runs produced by SMART and one run produced by Terrier. Performance results for each single run and fused runs are presented in Table 1, 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 transcripts are indexed from the documents, for English topics. For cross-languages experiments the results are represented in the columns Auto-French, and Auto-Spanish, when using the combined translations produced by the seven online MT tools, from French and Spanish into English. Since the result of combined translation for each language was better than when using individual translations from each MT tool on the training data (Inkpen, Alzghool, &Islam, 2006), we used only the combined translations in our experiments.

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. 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 available in the test collection showed high improvements, the MAP score jumped from 0.0855 to 0.2761 on the test data.

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 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). Our result on automatic transcripts for English queries (the required run for the CLSR task at CLEF 2007), obtained a MAP score of 0.0855. This result was significantly better than the other 4 systems that participated in the CLSR task at CLEF 2007(Pecina et al., 2007).

In future work we plan to investigate more methods of data fusion (to apply a normalization scheme scalable to unseen data), removing or correcting some of the speech recognition errors in the ASR content words, and to use speech lattices for indexing.

5. References

Alzghool, M. & Inkpen, D. (2007). Experiments for the cross language speech retrieval task at CLEF 2006. In C. Peters, (Ed.), *Evaluation of multilingual and multi-modal information retrieval* (Vol. 4730/2007, pp. 778-785). Springer.

Amati, G. & Van Rijsbergen, C. J. (2002). *Probabilistic models of information retrieval based on measuring the divergence from randomness* (Vol. 20). ACM, New York.

Buckley, C., Salton, G., & Allan, J. (1992). Automatic retrieval with locality information using smart. In *Text retrieval conferenc (TREC-1)* (pp. 59-72).

Inkpen, D., Alzghool, M., & Islam, A. (2006). Using various indexing schemes and multiple translations in the CL-SR task at CLEF 2005. In C. Peters, (Ed.), *Accessing multilingual information repositories* (Vol. 4022/2006, pp. 760-768). Springer, London.

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*. ACM, Seattle, Washington, United States.

Oard, D. W., Soergel, D., Doermann, D., Huang, X., Murray, G. C., Wang, J., Ramabhadran, B., Franz, M., & Gustman, S. (2004). Building an information retrieval test collection for spontaneous conversational speech, *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, Sheffield, United Kingdom.

Oard, D. W., Wang, J., Jones, G. J. F., White, R. W., Pecina, P., Soergel, D., Huang, X., & Shafran, I. (2007). Overview of the CLEF-2006 cross-language speech retrieval track. In C. Peters, (Ed.), *Evaluation of multilingual and multi-modal information retrieval* (Vol. 4730/2007, pp. 744-758). Springer, Heidelberg.

Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., & Johnson, D. (2005). Terrier information retrieval platform In *Advances in information retrieval* (Vol. 3408/2005, pp. 517-519). Springer, Heidelberg.

Pecina, P., Hoffmannov´a, P., Jones, G. J. F., Zhang, Y., & Oard, D. W. (2007). Overview of the CLEF-2007 cross language speech retrieval track, *Working Notes of the CLEF- 2007 Evaluation,* . CLEF2007, Budapest-Hungary.

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

Shaw, J. A. & Fox, E. A. (1994). Combination of multiple searches. In *Third text retrieval conference (trec-3)* (pp. 105-108). National Institute of Standards and Technology Special Publication.

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_expC2	**0.3177**	**0.2737**	0.0885	0.0744	**0.0908**	**0.0487**	**0.0747**	**0.0614**
Fusion 1 % change	0.3208 1.0%	0.2761 0.9%	0.0969 0.0%	0.0855 6.5%	0.0912 0.4%	0.0622 21.7%	0.0731 -2.2%	0.0682 10.0%
Fusion 2 % change	0.3182 0.2%	0.2741 0.1%	0.0975 0.6%	0.0842 5.1%	0.0942 3.6%	0.0602 19.1%	0.0752 0.7%	0.0619 0.8%

Table 1. Results (MAP scores) for 15 weighting schemes using Smart and Terrier (the In_expC2 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.

Weighting scheme	Manual English		Auto-English		Auto- French		Auto- Spanish	
	Train.	Test	Train.	Test	Train.	Test	Train.	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_expC2	**2638**	**1823**	1624	1286	1676	1061	1631	1172
Fusion 1 % change	2645 0.3%	1832 0.5 %	1745 0.3%	1334 1.0%	1759 0.6%	1147 -1.0%	1645 0.1%	1219 2.4%
Fusion 2 % change	2647 0.3%	1823 0.0%	1727 0.8%	1337 1.2%	1736 -0.7%	1098 -5.5%	1631 -0.7%	1172 -1.5%

Table 2. 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.