Ad hoc, Cross-language and Spoken Document Information Retrieval at IBM

Martin Franz, J. Scott McCarley, R. Todd Ward
IBM T.J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598
<franzm, jsmc, tward>@watson.ibm.com

1 Introduction

The Natural Language Systems group at IBM participated in three tracks at TREC-8: ad hoc, SDR and cross-language. Our SDR and ad hoc participation included experiments involving query expansion and clustering-induced document reranking. Our CLIR participation involved both the French and English queries and included experiments with the merging strategy.

2 Ad Hoc Track

In the TREC-8 ad hoc experiments we used a two-pass approach, in which the top documents, as ranked by the Okapi formula [1], were used to construct expanded queries, which were then used to compute the final scores. We also experimented with applying a clustering algorithm to obtain a more reliable list of passages for query expansion.

The data pre-processing agorithm was similar to the one we used in our previous TREC participations [2], [3]. It consisted of a decision tree based tokenizer, part-of-speech tagger [4] and a morphological analyzer. Filler query prefixes were removed using a database of such prefixes from previous TREC query sets. Morphed document and query unigrams and bigrams were collected using a vocabulary of 540459 words and a stop list of 514 words.

We applied the Okapi formula [1] in the first pass scoring the as described in [3]. First pass results are summarized in Table 1, line 1. Based on the first pass passage ranking, we constructed expanded queries using an LCA technique [5], modified as described in [3]. Both the documents and passages were scored with respect to the expanded queries using the Okapi formula. The final (pass 2) score of a document was computed as a combination of the document's score

	title		description		title + description	
	AveP	P20	AveP	P20	AveP	P20
pass1	0.2480	0.4010	0.2241	0.3760	0.2613	0.4270
pass2	0.2784	0.4090	0.2531	0.3950	0.3005	0.4500

Table 1: Ad hoc retrieval results, automatic.

	passage		document	
	AveP	P20	AveP	P20
baseline TREC-7	0.2032	0.3490	0.2140	0.3820
clustering TREC-7	0.2091	0.3630	0.2154	0.3920
baseline TREC-8	0.2480	0.3980	0.2481	0.3970
clustering TREC-8	0.2507	0.3910	0.2491	0.3950

Table 2: The effect of document clustering on selecting passages for query expansion, description query fields.

and the score of its highest ranking passage. Second pass results are shown in Table 1, line 2.

We also experimented with a clustering algorithm used to augment the list of passages used for query expansion, attempting to reduce the influence of the passages that rank high in the first pass scoring, but have little in common with the rest of the high ranking passages. In this experiment, we clustered the list of top 1000 passages from the first pass using a technique described in [7] and [8]. The clustering algorithm operates by reading a sequence of documents, in our case a list of passages sorted by their scores in decreasing order, and making a decision for each document either to add it to one of the existing clusters (with or without updating the cluster's profile), or to start a new cluster. After clustering the top 1000 passages, we constructed the lists of passages to be used for query expansion by selecting first the passages in the cluster created as the first and continuing by adding passages from the clusters created later, until we reached the limit of 100 pasages. Table 2 summarizes the results of the clustering experiments for both document- and passage- based second pass scoring. We used the clustering technique for our query description field based run only.

3 Spoken Document Retrieval Track

Our participation in the SDR track consisted of the reference (R1) and baseline (B1) runs. The text pre-processing and scoring techniques in our SDR experiments were based on those applied in our ad hoc entry and described in section 2. Bigram counts were collected for non-stop word pairs including pairs separated by a stop word. The number of top scoring documents used for query

	reference (R1)		baseline (B1)	
	AveP	P20	AveP	P20
pass1	0.4154	0.3940	0.3690	0.3660
pass2	0.4894	0.4530	0.4669	0.4470

Table 3: Spoken document retrieval results.

expansion was reduced to 60 to adjust for smaller size of the database.

We also tried applying a translation model to reduce the impact of speech recognition errors on the performance of the information retrieval system. In this view, there are two languages: the corpus of automatically transcribed data is considered to be one language of a parallel corpus, and a separately available corpus of manual transcriptions (of the same broadcast stories) is considered to be a separate language in a parallel corpus. Then retrieval of automatic transcriptions of broadcast news is considered to be a problem in cross-language information retrieval, since the queries (being free of speech recognition errors) more closely resemble the manual transcriptions. We then trained a statistical machine translation model of the type described in [9] to translate the *documents* from the language of automatically transcribed data into the language of hand-transcribed data. The test corpus was processed with this translation model, correcting some of the recognition errors and establishing cleaner text features to be used by the information retrieval system.

The training data was extracted from the January '98 part of the TDT2 corpus [7], which predates the SDR corpus. For the purpose of building the translation model, the output of the BBN speech recognizer served as the source language, close-captioning/manual transcripts being used as the target language. We aligned the source and target data sets at the level of sentences to form a parallel corpus. The translation model was trained on morphed representation of the corpus. We emphasize that manual transcriptions were used only in the training, not in the decoding phase.

Having trained the translation model, we applied it to translate the data produced by the BBN recognizer. Both the original and translated databases were indexed and scored separately with respect to the evaluation queries. We computed the final document scores as a linear combination of the scores of original and translated versions of the individual documents. Fig. 1 contains the average precision values for various relative weight combinations, showing a minor improvement achieved by incorporating a translation model in the system. The results of our SDR runs based on topics 74 to 123 are summarized in Table 3.

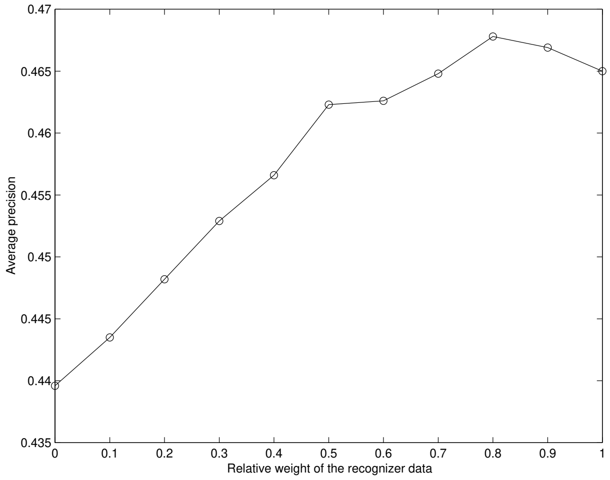

Figure 1: Combining the results based on ASR and translated data

4 Cross-language Track

4.1 Introduction

IBM's participation in the cross-language track at TREC-7 involved experiments with all four document languages : English, French, German, and Italian, and two of the query languages. Two experiments (*ibmcl8ea* and *ibmcl8ec*) were submitted based on the English queries, and two experiments (*ibmcl8fa* and *ibmcl8fc*) were based on the French queries. Our system is a composite system: we do initial retrievals for each language pair of interest, and then we merge the appropriate runs. The two experiments for each query language differed only in the merging strategy, not in the initial retrievals. The techniques studied here would also have been applicable to other query languages. All four runs used the long form of the queries. ("Long" queries used all three fields, <Title>, <Description>, and <Narrative>.) All query processing was fully automatic. We varied our strategy somewhat between the French and English query experiments. An important theme of our experiments has been that the widely varying availability and quality of bilingual resources (parallel and comparable corpora) requires that IR systems vary their strategy between language pairs accordingly. A second unifying theme of these experiments is the extensive use of statistical methods, reflecting the long history of statistical approaches to machine translation in our group. [6] In fact, all bilingual dictionaries and translation models used in these runs were learned automatically from corpora.

5 System Description

IBM's multilingual retrieval system is a composite system: a ranked list of potentially relevant documents is retrieved separately in each document language, and then these lists are merged. Because of the available bilingual resoures, the retrieval engines associated with each language pair vary somewhat between langauge pairs. The $EqFd$ is a hybrid query-translation document translation retrieval system, as described in [10], using the statistical machine translation algorithm described in [9]. Both the *English*⇒*French* query translation and the *French*⇒*English* document translation were trained from a parallel corpus (the Canadian Hansards.) The $FqEd$ retrieval system is identical, except for the interchange of *French* and *English*. The $FqGd$ system is also a hybrid query-translation document-translation IR systems, but the underlying *French*⇒*German* and *German*⇒*French* statistical translation models are trained from a comparable corpus (the SDA newswire itself), not a parallel corpus. The alignment of the comparable corpus was described in [3]. The $FqId$ system is identical. The $EqGd$ system is implemented using *French* as *pivot language*: we use the $EqFd$ system to retrieve French documents, automatically constructed a French query from these documents, and then use the $FqGd$ to retrieve German documents based on the artificial French query, as described in [3].

6 Results by Language Pairs

Because the IBM multilingual retrieval system is a composite system, it is important to observe individual aspects of our system's performance separately prior to merging. The most important aspects are the performance on the eight language pairs (systems for both French and English queries were submitted.) Results by query language and document language are shown in Tables 5. We also contrast the performance of the English query and French query systems on individual queries in the scatterplot in Fig. 2. Finally, we also contrast our systems performance on two subsets of the queries, which will have important consequences in the final merging. In analyzing the results of TREC-7, we noted that a significant fraction of the queries concern local European events, and these events are under-reported in the AP newswire. Furthermore, these queries can be automatically recognized, with reasonable reliability, by whether they specifically mention the name of a European country. This effect is shown in table 4 in which we denote the set of queries mentioning a European country E and the remainder of the queries nE. The same queries were identified as a mentioning a European country in both the English-query and French-query experiments, although this need not have been the case if the human translations of the provided queries had been looser. We suspect that this effect also correlates with the country in which the query was originally constructed, but we

document language	$\|E\|$	$\|nE\|$	$\|total\|$
English	140 (14.6%)	816	956
French	192 (33.2%)	386	578
German	327 (45.6%)	390	717
Italian	51 (30.0%)	119	170

Table 4: Number of relevant documents by document language and query subset

document language	AveP (Fr)	P20	AveP (Eng.)	P20
English	0.2952	0.3357	0.3049	0.3375
French	0.4706	0.3857	0.4186	0.3804
German	0.3142	0.3268	0.2559	0.2839
Italian	0.2788	0.1357	0.2221	0.1357

Table 5: Results by language pair

have not attempted to guess which queries were constructed in which countries.

7 Importance of Merging

Our merging strategy is to estimate the probability of relevance p of each document as a function $p(R, l_d, q)$ of the rank R that the document is retrieved by the systems for document language l_d, and also to allow this probability to depend upon features of the query q. The merging strategy as we formulate it here applies only to the merging of disjoint sets of documents. We have observed that the average precision at given rank R of information retrieval system is an approximately linear function of $log(R)$ and we can use this linearity to form a two-parameter estimate of p for that system and set of queries [3]. We have a different estimate of p for each language pair. We also have a separate estimate for the query subsets E and nE (queries mentioning a European countries, and those that do not, respectively) and we find that this results in a slight improvement in performance over the single estimate for all queries. This strategy makes only the shallowest use of information about the query and the documents and it retrieves: other information, such as the IR engine's score of the document with respect to the query has not proven beneficial. Since the average precisions for this year's queries are significantly lower than last year's, we can test the sensitivity of the overall average precision to the parameterization of the merging strategy by tuning our merging strategy to this year's queries. We

query language	submission	merging	AveP
French	ibmcl8fa	$p(R, l_d, q)$	0.2613
French	ibmcl8fc	$p(R, l_d)$	0.2600
English	ibmcl8ea	$p(R, l_d, q)$	0.2559
English	ibmcl8ec	$p(R, l_d)$	0.2515

Table 6: Results by merging strategy

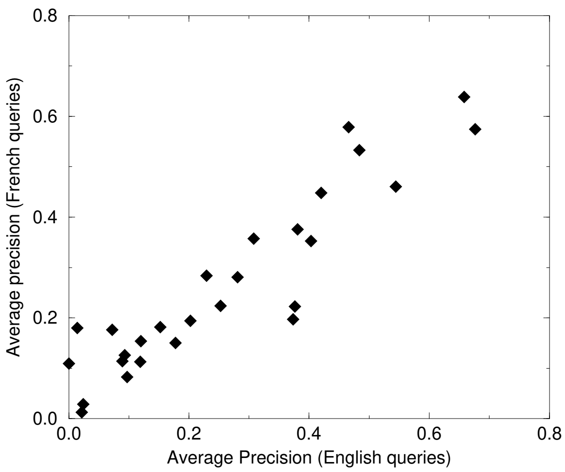

Figure 2: Scatterplot of average precision on English queries vs. French queries

find an approximate 10% improvement (average precision = 0.2803 on French queries.) These results are shown in Table 7.

8 Acknowlegments

This work is supported by NIST grant no. 70NANB5H1174.

References

[1] S.E. Robertson, S. Walker, S. Jones, M.M. Hancock-Beaulieu, M. Gatford, "Okapi at TREC-3" in *Proceedings of the Third Text REtrieval Conference (TREC-3)* ed. by D.K. Harman. NIST Special Publication 500-225, 1995.

[2] M. Franz and S. Roukos, "TREC-6 Ad-hoc Retrieval", in *Proceedings of the Sixth Text REtrieval Conference (TREC-6)* ed. by E. M. Vorhees and D.K. Harman. NIST Special Publication 500-240, 1998.

[3] M. Franz, J.S. McCarley, S. Roukos, "Ad hoc and Multilingual Information Retrieval at IBM", in *Proceedings of the Seventh Text REtrieval Conference (TREC-7)* ed. by E. M. Vorhees and D.K. Harman. NIST Special Publication 500-242, 1999.

[4] B. Merialdo, "Tagging text with a probabilistic model" in *Proceedings of the IBM Natural Language ITL,* Paris, France, pp. 161-172, 1990.

[5] J. Xu and W. B. Croft, "Query Expansion Using Local and Global Document Analysis", in *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Zurich, Switzerland, pp. 4-11, 1996.

[6] P. F. Brown et al. "The mathematics of statistical machine translation: Parameter estimation", *Computational Lingustics*, 19 (2), 263-311, June 1993.

[7] S. Dharanipragada, M. Franz, J.S. McCarley, S. Roukos, T. Ward, "Story Segmentation and Topic Detection in The Broadcast News Domain", in *Proceedings of the DARPA Broadcast News Workshop*, 1999.

[8] S. Dharanipragada, M. Franz, J.S. McCarley, S. Roukos, T. Ward, "Story Segmentation and Topic Detection for Recognized Speech", in *Proceedings of the Sixth European Conference on Speech Communication and Technology*, 1999.

[9] J.S. McCarley and S.Roukos, "Fast Document Translation for Cross-Language Information Retrieval", in *Machine Translation and the Information Soup* ed. by D.Farwell, L.Gerber, and E.Hovy., 1998.

[10] J.S. McCarley, "Should we Translate the Documents or the Queries in Cross-Language Information Retrieval?", in *37th Annual Meeting of the Association for Compuational Linguistics* College Park, MD, 1999.