Ad hoc and Multilingual Information Retrieval at IBM

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

January 27, 1999

1 Introduction

IBM participated in two tracks at TREC-7: ad hoc and cross-language. In the adhoc task we contrasted the performance of two different query expansion techniques: local context analysis and probabilistic model. Two themes characterize IBM's participation in the CLIR track at TREC-7. The first is the use of statistical methods. In order to use the document translation approach, we built a *fast* (translation time within an order of magnitude of the indexing time) French⇒English translation model trained from parallel corpora. We also trained German⇒French and Italian⇒French translation models entirely from comparable corpora. The unique characteristic of the work described here is that all bilingual resources and translation models were learned automatically from corpora (parallel and comparable.) The other theme is that the widely varying quality and availability of bilingual resources means that language pairs must be treated separately. We will describe methods for using one language as a pivot language in order to decrease the number pairs, as well as methods for merging the results from several retrievals.

2 Adhoc

2.1 System Description

We used two different multi-pass strategies in TREC-7 automatic ad-hoc experiments, both of them based on improving the document scores given by the Okapi formula [1] by combining them with scores obtained with expanded queries. To construct the expanded queries we tried the local context analysis approach [2] and also the probabilistic model [3].

The data preprocessing stage was the same as the one applied in our TREC-6 system and described in [4]. We used statistical tokenizer, part-of-speech tagger [5] and morphological analyzer on both the description fields of the queries and content bearing fields of the documents. Filler query prefixes were filtered out by mechanism similar to the one described in [4]. We have collected unigrams and bigrams based on the morphed data using a 540459 word vocabulary and a list of 514 stop words.

We used Okapi formula [1] for the first-pass scoring the same way as in [4]. Unigrams and bigrams in the intersection of the query and document contributed a score of:

$$s = \frac{tf}{c_1 + c_2 \times \frac{dl}{avdl} + tf} \times w^{(1)}, \tag{1}$$

where tf and qtf are the document and query counts for a given n-ngram, dl is the document length, $avdl$ is the average length of the documents in the collection and $w^{(1)}$ is the inverse document frequency, computed as:

$$w^{(1)} = log(\frac{N - n + 0.5}{n + 0.5}),$$

where N is the total number of documents in the corpus and n is the number of documents containing a given n-gram. In the Eq.(1) we used $c_1 = 0.5, c_2 = 1.5$ for unigram scoring and $c_1 = 0.05, c_2 = 0.05$ for the bigrams. The first pass score was a linear combination of unigram and bigram scores given by Eq.(1), with the unigram scores weight set to 0.8 and bigram scores weight equal to 0.2. First pass results for query description and title fields are summarized in Table 1, line 1 and Table 2 line 1, respectively.

2.2 Query Expansion with Local Context Analysis

In this experiment we applied an approach similar to the one described in [2], with some modifications in the way the inverse document frequencies were handled and in expanded terms weighing.

We used passages of 200 non-stop words, overlapping by a half of their length. The original queries were expanded by adding 100 unigrams based on top 100 passages.

The expanded queries were used to score both the documents and the passages with Okapi formula. Passage scores were later converted into new document scores in a way where the document score was given by the score of its highest scoring passage. Final scores were obtained as weighed combination of the two, with the ratio between document and passage scores set to 40/60. The results of these experiments for query description and title fields are listed in Table 1, line 2 and Table 2 line 2, respectively.

	TREC-5		TREC-6		TREC-7	
	AveP	P20	AveP	P20	AveP	P20
pass1	0.1757	0.2650	0.1769	0.3050	0.1865	0.3760
LCA, os	0.2010	0.3050	0.2057	0.3090	0.2336	0.4010
px, ps	0.1951	0.2850	0.1901	0.3010	0.2075	0.3770
px, os	0.1974	0.3000	0.1863	0.3070	0.2047	0.3750

pass1: first pass Okapi scoring, unigram nad bigram terms
LCA: local context analysis query expansion
px: probabilistic model query expansion
ps: second pass probabilistic model scoring
os: second pass Okapi scoring

Table 1: Results of experiments on TREC-5, TREC-6 and TREC-7 sets: ad-hoc, automatic, short topics.

	TREC-5		TREC-6		TREC-7	
	AveP	P20	AveP	P20	AveP	P20
pass1	0.1234	0.1820	0.1943	0.3250	0.1749	0.3440
LCA, os	0.1714	0.2520	0.2392	0.3390	0.2502	0.3720

pass1: first pass Okapi scoring, unigram nad bigram terms
LCA: local context analysis query expansion
os: second pass Okapi scoring

Table 2: Results of experiments on TREC-5, TREC-6 and TREC-7 sets: ad-hoc, automatic, title.

2.3 Query Expansion with Probabilistic model

Our probabilistic model based query expansion technique was the same as the one described in [4]. After the first pass Okapi scoring, top 40 documents were used to determine the additional unigrams, by thresholding the probabilistic model scores. New bigrams were the ones found in at least 15 of the top 40 documents.

We tried using both probabilistic model and Okapi formula for second pass scoring with expanded queries. Both the first and second pass scores were modified using the method for scoring correlated features, described in [6]. Scores of the original and expanded terms were then combined using 80/20 weighing ratio. The results of these test runs may be found in Table 1, lines 3 and 4.

2.4 Conclusion

We have experimented with various query expansion and scoring algorithms in the context of TREC-5, TREC-6 and TREC-7 tasks. All the query expansion methods brought an improvement of average precision over the baseline first pass Okapi scoring. Among the query expansion methods, LCA technique caused the most significant benefit, with the two of probabilistic model based methods bringing roughly the same but smaller improvement.

3 Crosslanguage track

3.1 Introduction

IBM's participation in the cross-language track at TREC-7 involved building separate systems for all four document languages : English, French, German, and Italian. We focused our attention on the English queries, although the techniques we studied would also have been applicable to the other query languages. Four runs were submitted, covering both long and short queries. ("Long" queries used all three fields, <Title>, <Description>, and <Narrative>. "Short" queries used just the traditional <Description> fields.) All query processing was fully automatic. We varied our strategy somewhat between runs: this paper will focus on the techniques used in runs *ibmcl7cl* and *ibmcl7cs*. A unifying theme of these runs is the extensive use of statistical methods, reflecting the long history of statistical approaches to machine translation in our group. [7] In fact, all bilingual dictionaries and translation models used in these runs were learned automatically from corpora. [1] We treated each document language as a separate IR system. Unlike last year's task [8], this year's task involved merging the ranked lists of documents from each system.

Our overall approach to cross-language information retrieval has been to translate the documents, rather than queries, since there is more varied context in the documents. Once the documents have been translated, we use familiar IR techniques such as the Okapi formual [1], and probabilistic models [3] that have been successful used by our group in the ad-hoc tasks at previous TREC's. [9, 4] Most of our work in cross-lingual retrieval has focused on French. We developed a "Fast Document Translation" algorithm that was trained on a parallel corpus, incorporated word sense disambiguation (also learned from the parallel corpus) and by ignoring word order, was able to translate the entire French section of the SDA in a reasonable amount of time (in fact, within an order of magnitude of the amount of time spent indexing the collection!) For retrieval from Ger-

[1] The runs *ibmcl7al* and *ibmcl7as* used all of the above methods, but also incorporated query translation from English to German and Italian using Altavista (Systran, http://babelfish.altavista.digital.com). The motivation was that different translation systems would complement each other. The incorporation was through a linear combination of scores. The result was a modest improvement in overall performance.

man text, we did not have a parallel corpus available, so we used comparable corpus methods to create appropriate training data for our machine translation methods. Italian was treated identically to German. We also studied the use of French as pivot language, so that we could combine our resources for retrieving French documents with an English query with our resources for retrieving German documents with a French query to produce a system for retrieving German documents with an English query. Finally we also explored simple schemes for merging disjoint sets of documents retrieved from different IR system.

3.2 Statistical Machine Translation

The statistical approach to machine translation assumes that with any pair of English and French sentences (E, F) (of length $|E|$ and $|F|$ words, respectively), with $E = e_1 ... e_{|E|}$ and $F = f_1 ... f_{|F|}$ one can associate a probability that E is a translation of F. The most probable English translation E of a given French sentence is then given by

$$\hat{E} = \arg\max_E P(E|F). \tag{2}$$

Modeling $P(E|F)$ depends upon being able to factor it into terms representing individual pairs of words. This factorization is accomplished by introducing a word-by-word alignment between the sentences, motivated by the idea that there are many words in one language ("perfume") which are highly correlated with a word in the other language ("parfum"). We denote the alignment of a sentence pair as A. A typical representation of the alignment is to assign to each word e_i in E an integer $a_i \in \{1...|F|\}$ indicating that it is associtated with f_{a_i}.

There are many ways to factor Eqn. (2) into terms involving words. We follow [7] and introduce Model 1

$$p(E, A|F) = \frac{\epsilon}{(|E|+1)^{|F|}} \prod_i t(e_i|f_{a_i}) \tag{3}$$

As originally described, Model 1 was used in a source-channel framework. This approach is computationally expensive and therefore difficult to incorporate into a document-translation based IR system. The principle difficulty is that the search space in Eq. (2) covers variation in word order and other features that are largely irrelevant to information retrieval. However, extracting a bilingual dictionary from the trained model is easy: for each French word f tabulate the English word e that maximizes $t(e|f)$.

3.3 "Fast Translation"

We have extended Model 1 into a more versatile method that is able to translate phrases and to disambiguate the sense of words during translation. [10] In order to incorporate context into the model, we note that the existence of the alignment A allows each English word to have a context not only of surrounding English words, but also of French words surrounding the word to which it is aligned. We denote the number of values of i for which $a_i = j$ as the *fertility* n_j of the j'th word in F. We have proposed a different decomposition of the basic equation into word-by-word terms:

$$p(E, A|F) = \left[\prod_{i=1}^{|F|} p_n(n_i|n_1^{i-1}, F)\right] p_a(A|N, F) \left[\prod_{j=1}^{|E|} p_s(e_j|e_1^{j-1}, A, F)\right] \quad (4)$$

where the fertilities $n_1, ...n_{|F|}$ are collectively denoted N.

In order to generate the translation of a French sentence, one first picks the fertilities of the French words with probabilities p_f (as opposed to the total number of English words, as in Model 1), then one picks an alignment with probability p_a as constrained by the fertilities. Finally, English words are picked with probabilities p_s as translations of the French words, based on the context in the French sentence. Different "senses" or meanings of the French word, as disambiguated by its context, are reflected in the different choices of Enlgish words generated (i.e. *pomme* may be rendered as *potato* or as *apple* depending upon whether or not it is followed by *de terre*.) This model is trained under the observation that most of the probability is likely to be associated with the Viterbi alignment (or at most a few neighboring alignments) [7]. We note that an approximate alignment can be easily computed from many other models, and have had considerable success using alignments computed from Model 1, as described above. (Alignment probabilities are easily found from a rearrangement of Eq. (3).) Approximating the fertility and sense terms so that only local context matters leaves us with fertility and sense models which are simply 4-gram langauge models:

$$p_n(n_i|n_1^{i-1}F) \approx p_n(n_i|f_i, f_{i-}, f_{i+}) \quad (5)$$

$$p_s(e_j|e_1^{j-1}, A, F) \approx p_s(e_j|f_{a_j}, f_{a_j-}, f_{a_j+}). \quad (6)$$

Here we will take local context of a word f_i to be the previous and next non-stop words, denoted f_{i-} and f_{i+}, respectively. and treat the middle factor on the right-hand-side of Eq. (4) as an irrelevant constant. The translation model is completely specified by these functions. The two functions that must be modeled are simply conditional probabilities of the occurrence of a word given some information about the local context of the word, a problem familiar from speech recognition. [11]

In order to translate French text with model, for each French word, the fertility is predicted with p_n. Then, p_s is used to select which n of several possible choices of English words are likely translates. The resulting translation is incorporated into our information retrieval system by simply indexing the translated documents. Translating the 3 years of SDA newswire required an average of 28 hours for each year of newswire text on an RS-6000 Model 590. This translation rate is much faster than other published accounts of using document translation on corpora of comparable size [12] and, in fact, is within an order of magnitude of the amount time spent on other processing of the documents (part-of-speech tagging, morphological stemming, building the inverted index, etc.) The combined (MT+IR) system achieved an average precision of 0.3400 on TREC-6 long queries, the best result of which we are aware on that query set. Incorporating the fertility and sense models results in an $18\% - 19\%$ improvement in average precision over merely using the statistical constructed dictionary implicit in Model 1.

3.4 Comparable Corpora

The English-French retrieval system was trained using a parallel corpus, a particularly valuable linguistic resource. An important issue is whether a similar system can be trained using more readily available linguistic resources, such as a comparable corpus. A comparable corpus differs from a parallel corpus in two important aspects. First the similar documents in the corpus are not translations of each other, but are composed independently. Second, these comparable documents may not be aligned with each other. The SDA newswire itself is an example of a comparable corpus. The French, German, and Italian sections are not translations of each other, but since they report news events from the same time periods (including local Swiss events), we nevertheless expect that articles about the same event contain correlations useful for the training of statistical machine translation algorithms.

In order for this corpus to be useful for machine translation purposes, we select an aligned subset of comparable articles and treat it as a parallel corpus. (Another approach to comparable corpora involves comparing the context of words across the languages without aligning specific articles. [13]) Our approach is motivated by the observation that names are frequently spelled identically in French, German, and Italian. Passages that contain the same name (or better, the same names) even though they are in different languages, are more likely to be about the same event. Of course, names that are more common are less informative. Such an approach to comparable corpora alignment has already been utilized. [14] These features suggest the following algorithm:

(1) Index the French and German SDA into passages of, for example, 50 words.

(2) Formulate an initial dictionary of bilingual word-pairs (either known translates or words that are spelled identically in both languages)

(3) Compute Okapi scores of documents in one language against those in the other, counting those word-pairs from the bilingual dictionary as equivalent or matching. It is convenient to score only those documents that were published on approximately the same date. For each French passage, retrieve the best German passage. For each German passage, retrieve the best French passage.

(4) If a French and a German passage retrieve each other, regard them as aligned.

(5) Treat the subset of aligned German and French passages as training data for the machine translation.

(6) Train a machine translation system, and extract a new dictionary of bilingual word-pairs.

(7) Repeat starting at step (3).

We performed two alignments with this procedure: French-German, and French-Italian. Both were seeded with identically spelled words. After two iterations we obtained 23261 "aligned" French and Italian passages, and 90453 "aligned" French and German passages. There were approximately 35% more "aligned" articles after the second iteration than after the first. We did not check the quality of the alignments, but regarded them as a test of our ability to train a machine translation system with a noisy training data. Translating the German corpus (including the NZZ) into French with the dictionary produced by this method, and retrieving using the French TREC-6 queries (long version) produced an average precision of 0.2361, which was about 68% of our (German) monolingual performance. As a percentage of monolingual performance, this was similar to that obtained with a French-English dictionary constructed from parallel corpora, although we caution that the tasks are not strictly comparable.

3.5 Pivot Language

Having developed resources for retrieving French documents given English queries, and for retrieving German documents given French queries, it is desirable to be able to combine these resources in order to retrieve German documents given English queries. There are several methods of combining these resources.

(1) Direct translation: Combine the German⇒French translation system with the French⇒English translation system directly, by translating the German documents into French, and then translating them into English.

(2) Convolution: Convolve the German⇒French translation system with the French⇒English translation system to obtain a German⇒English translation system. This operation is suggested by the mathematical structure of the translation model.

$$t(g|e) = \sum_f t(g|f)t(f|e) \tag{7}$$

In effect, we sum over several possible translations in the intermediate language.

An alternative approach is to combine the information retrieval systems themselves, rather than the underlying translation systems, by using query expansion. An appealing feature of this method is its generality: different implementations of cross-language IR systems (document translation, query translation, LSI, etc.) can be combined. Our approach is as follows:

(1) Use the English query in the English-French CLIR system to retrieve French documents.

system	AveP	P20	%baseline
1	0.3478	0.4136	100%
2	0.2361	0.2955	68%
3	0.1577	0.1977	45%
4	0.1301	0.1455	37%
5	0.2295	0.2636	66%

Table 3: Retrieval from German documents, TREC-6 long queries: (1) = German queries (monolingual) (2) = French queries (G$\Rightarrow$F translation), (3) = English queries (G$\Rightarrow$F$\Rightarrow$E translation) (4) = English queries, convolution translation, (5) = English queries, French query expansion (see text).

(2) Formulate a French query based on the top-n French documents.

(3) Use this query in the French-German CLIR system to retrieve German documents.

Although more sophisticated query-expansion techniques could easily be incorporated, we formed the new French query by simply merging all non-stopwords in the top-n French documents. We found that $n = 3$ worked much better than $n = 2$ or $n = 1$, and that there was a relatively smaller loss in average precision for $n > 3$. We found that the query-expansion technique substantially outperformed the methods involving combining the translation models (see table 3.)

3.6 Merging

We implemented the English-French and English-German retrieval systems as described above, guided by the results of the above experiments. We implemented the English-Italian system by blindly following the structure of the English-German system. Since the goal of the CLIR track is to produce a single list of relevant documents across all languages, it is necessary to merge the results from each system. Scores produced by the Okapi formula (or similar IR formulae) are not directly comparable, because of the different languages and differing quality of the underlying translation resources. What is needed is a simple means to estimate the probability that a document D is relevant $Pr(\rho)$ based on previous performance of the IR system. Note that probabilistic models such as [3] are not comparable either, even though the mathematics suggests that they are modeling the probability of relevance. In fact, because they are trained by pseudorelevance feedback (off of first-pass Okapi scores) they are no more comparable than the scores of the first pass.

We can estimate probability of relevance from precision at rank $P(R)$ by

$$(R+1)P(R+1) - RP(R) = Pr(\rho|R) \tag{8}$$

Because of the very limited amount of training data available, it is essential that

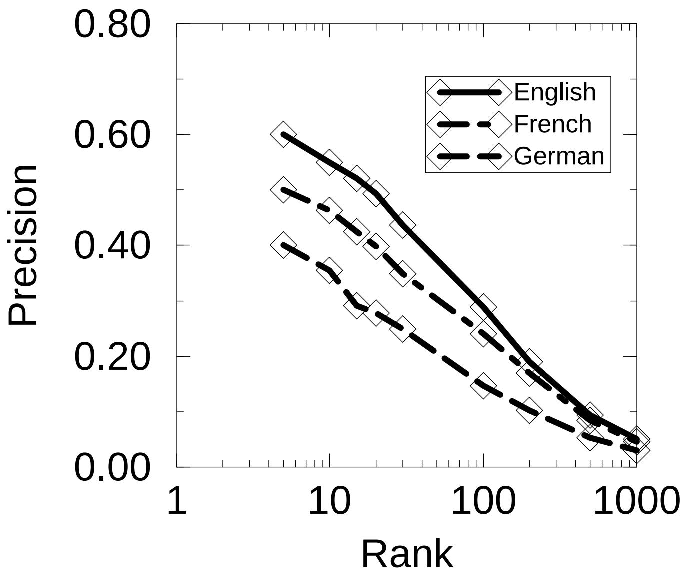

Figure 1: Precision vs. log(Rank)

	TREC-7 (short)	TREC-7 (long)
interleave	0.1912	0.2574
merged	0.2212	0.2942
% rel.gain	15%	14%

Table 4: Merging of documents: interleaving (top) and modeling a system-wide probability of relevance (bottom).

we use only a few parameters to describe each system. We note in Figure 2 that precision is approximately a linear function of the $log R$. (We do not claim that there is an underlying scaling law, or that we expect the linearity to hold over more decades, merely that this is a simple interpolation scheme that allows us to describe the precision of the system, using only two parameters.) Thus we can estimate the probability of a relevance of a document as simple function only of its rank in the original retrieval. Thus we can merge a disjoint set of documents retrieved from different systems by sorting on the estimate of $P(\rho)$. We note that this procedure does not use any information about the magnitude of the scores. Furthermore it merges documents in the same proportions and in the same order for all queries.

We calibrated our merging system on the TREC-6 queries against French and German documents, and assumed that Italian would be similar to the German. We compare this estimate from the simple interleaving of equal numbers of documents that would be obtained if $Pr(\rho|R)$ is chosen to be any arbitrary decreasing function of R identical for all systems. The result was a substantial improvement in overall performance, even though the system was calibrated on the results of TREC-6 queries.

3.7 Conclusion

We emphasize that all of the methods described here are statistical in nature and that all bilingual lexicon used were learned automatically from corpora. Although statistical machine translation has long relied on parallel corpora, we have shown how these methods can also be extended to non-parallel, comparable corpora. Since linguistic resources vary widely in both size and quality between language pairs, it is necessary to develop separate systems for each language pair. Therefore we have also developed methods to address the merging problem, and successfully used a pivot language in order to reduce the number of language pairs.

4 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] J. Xu and W. B. Croft 1996 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.

[3] Ernest P. Chan, Santiago Garcia, Salim Roukos 1998 Probabilistic Modeling for Information Retrieval with Unsupervised Training Data. In *Proceedings, The Fourth International Conference on Knowledge Discoveryand Data Mining,* AAAI Press, pg.159.

[4] M. Franz and S. Roukos, "TREC-6 Ad-hoc Retrieval", in *The 6th Text REtrieval Conference (TREC-6).*

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

[6] M. Franz, S. Roukos 1998. A Method for Scoring Correlated Features in Query Expansion. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* Melbourne, Australia, pp. 337-338.

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

[8] D.Harman and E.Voorhees, "Overview of the Sixth Text REtrieval Conference (TREC6)",in *The 6th Text REtrieval Conference (TREC-6).*

[9] E.Chan, S.Garcia, S.Roukos, "TREC-5 Ad Hoc Retrieval Using K Nearest-Neighbors Re-Scoring" in *The 5th Text REtrieval Conference (TREC-5)* ed. by E.M. Voorhees and D.K.Harman.

[10] 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)

[11] L.R. Bahl, F.Jelinek, and R.L. Mercer, "A Maximum Likelihood Approach to Continuous Speech Recognition", in *IEEE Transactions on Pattern Analysis and Machine Intelligence* 5 (2), 1983.

[12] D.W. Oard, P.Hackett, "Document Translation for Cross-Language Text Retrieval at the University of Maryland", in *The 6th Text REtrieval Conference (TREC-6)* ed. by E.M. Voorhees and D.K.Harman.

[13] P.Fung, "A Statistical View on Bilingual Lexicon Extracction: From Parallel to Non-parallel Corpora", in *Machine Translation and the Information Soup* ed. by D.Farwell, L.Gerber, and E.Hovy. (1998)

[14] B.Mateev, E.Munteanu, P.Sheridan, M.Wechsler, and P.Schäuble, "ETH TREC-6: Routing, Chinese, Cross-Language and Spoken Document Retrieval" in in *The 6th Text REtrieval Conference (TREC-6)* ed. by E.M. Voorhees and D.K.Harman. (1997)