Machine Translation and Monolingual Information Retrieval

Martin Franz
IBM T.J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598
franzm@watson.ibm.com

J. Scott McCarley
IBM T.J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598
jsmc@watson.ibm.com

Abstract

Earlier investigation into cross-language information retrieval systems that incorporate both document and query translation has shown that incorporating document translation improves retrieval performance even for human-quality query translation. Thus we view monolingual retrieval as cross-language retrieval in which the queries have already been translated and propose the incorporation of document translation. Experiments on the TREC 6 and 7 ad hoc tasks [1, 2] yield modest improvements in performance.

1 Introduction

Although many approaches to cross-language information retrieval have been tried, include non-translation based approaches, the most common approaches have involved the coupling of machine translation (MT) and traditional (monolingual) information retrieval (IR). Typical approaches include machine translating the documents into the query's language and translating the queries into the document's languages. However, recent experiments with comparable translation systems have shown that hybrid approaches outperform either extreme approach. [3] Furthermore hybrid systems mixing document translation with the monolingual baseline system outperformed the monolingual baseline. In [3] the monolingual baseline system was viewed as the human-quality limit of MT. Here we take the opposite point of view: we regard monolingual information retrieval as a form of cross-language information retrieval in which query translations have already been provided: then we incorporate document translation in order to improve the performance of monolingual IR.

2 System Description

Our algorithm for fast document translation has been previously described in some detail [4]. It is a statistical model that is a descendant of IBM Model 1 [5], which incorporates features of more advanced models such as fertility n (the number of French words associated with a given English word) and context dependence while retaining the speed and ease of training of Model 1. Very fast decoding is achieved by implementing it as a direct-channel model rather than as a source-channel model. The basic structure of the model is that an underlying Model 1 is built and then extended by using it as a basis on which to build a fertility model $p(n|e, context)$ and a sense model $p(f|e, context)$ which choose the appropriate French word in a manner sensitive to the context of the English word. Note that this a form of sensing, in which the sense of the English word is labeled by the French word into which it is translated. These models were all trained on the Hansard (Canadian parliamentary proceedings) and UN corpora, which have previously been aligned on sentence-by-sentence basis. [6]

Our IR engine has previously been described [7]. Documents preprocessing consists of part-of-speech tagging and morphological analysis. First-pass scoring is based on the Okapi formula [8] applied to passages with a length of 200 non-stop words. Query expansion involved methods similar to Local Context Analysis [9], with separate rescorings of the passages and of the entire documents, and with the final score a linear combination of the two query expansion rescorings. When query expansion was performed on the translated documents, the expansion was performed in French, not in the original English of the original documents.

3 Experiment

We have translated both the corpus and the queries of the TREC-6 and TREC-7 ad hoc tasks from English into French using fast document translation. This is the largest document translation experiment of which we are presently aware. IR experiments focused on the traditional <Description > field of the TREC-6 and TREC-7 queries. Two IR experiments were performed: the original queries were used to score (denoted S_E) the untranslated documents, and the machine translation of the queries was used to score (denoted S_{EF}) the translation of the documents. Performance in the translated domain ranged from 80% - 90% of the monolingual baseline. The scores of the documents from both systems were combined linearly for the final (hybrid system) results. A mixing of $0.7S_E + 0.3S_{EF}$ was observed to maximize the average precision of the hybrid system on the TREC-6 task, resulting in a 7% relative gain in performance. Precision vs. recall is plotted in Fig. 1, and suggests that the gain is roughly constant, except at the deepest levels of recall where it is less. Results for both TREC-6 and TREC-7, across both Okapi and LCA scores are given in Table 1. Gains were much smaller for TREC-7. Identical results would have obtained if we had trained on TREC-7

SIGIR '99 8/99 Berkley, CA, USA
© 1999 ACM 1-58113-096-1/99/0007...$5.00

and tested on TREC-6. As seen in Fig. 2, this gain is relatively robust across a range of mixing proportions. Fig. 2 also suggests that the difference between results on TREC-6 and TREC-7 is inherent in the queries, rather than a result of tuning to TREC-6. Gains for LCA scores reflected similar gains from Okapi scores. These results are consistent with previous observations. [3]

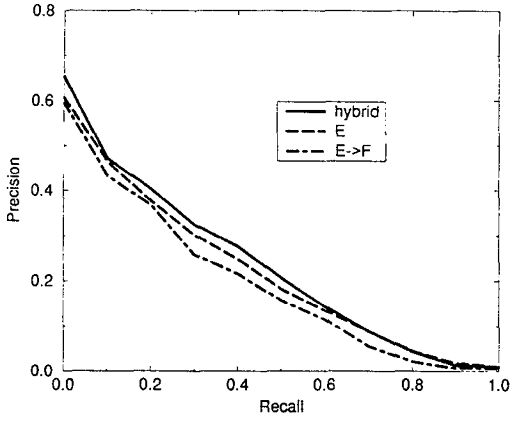

Figure 1: Recall vs. precision

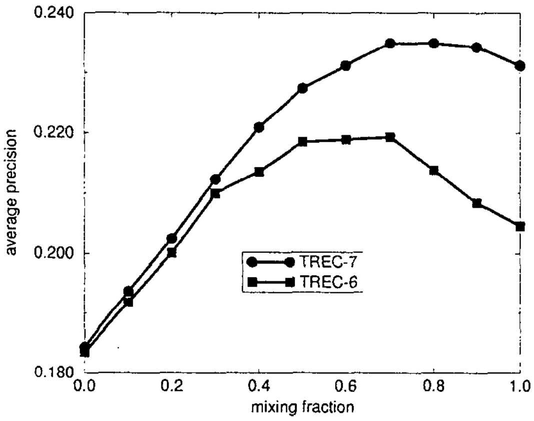

Figure 2: mixing

system	E	E⇒F	hybrid
TREC-6 (Okapi)	0.1820	0.1461	0.1947
TREC-6 (LCA)	0.2046	0.1834	0.2193
TREC-7 (Okapi)	0.1861	0.1491	0.1873
TREC-7 (LCA)	0.2312	0.1843	0.2349

Table 1: Results: averages precision of Okapi and LCA scores

4 Conclusion

We have incorporated the machine translation of both the documents and the queries into an IR engine, resulting in a modest improvement in performance on the TREC-6 ad hoc task, and a small improvement on the TREC-7 ad hoc task. They are at least two possible explanations for the gain. One is a direct form of sense disambiguation : words with ambiguous senses in English are translated into distinct French words. The translation model is able to take the context of words into account in a different manner than the IR engine. Unfortunately, inspection of the results has not yielded convincing evidence for this phenomenon in the TREC queries. Thus we suspect that the differences are due to an *indirect* sensing effect: a word in the MT output may have a different document frequency (and hence weight) than the corresponding word in the input because of translation ambiguity with respect to words that are present in *neither* the query nor the document. Further investigation is needed to understand the importances of these two effects. Furthermore, our method of incorporating this information is extremely simple and couples the machine translation to the IR in a weak manner. Future research will focus on coupling the machine translation to the IR more tightly, for example, by forming pseudowords consisting of ordered pairs of English and French words and studying retrieval in the domain of the pseudowords.

5 Acknowledgements

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

References

[1] D.Harman and E.Voorhees, "Overview of the Sixth Text REtrieval Conference (TREC6)",in The 6th Text REtrieval Conference (TREC-6), 1998.

[2] E.Voorhees and D.Harman, "Overview of the Seventh Text REtrieval Conference (TREC7)",in The 7th Text REtrieval Conference (TREC-7), 1999.

[3] J.S. McCarley, "Should we Translate the Documents or the Queries in Cross-language Information Retrieval?", in *Proceedings of the 37th Annual Meeting of the Association of Computational Linguistics*, 1999.

[4] J. S. McCarley and S. Roukos, "Fast Document Translation for Cross- langauge Information Retrieval", in *Machine Translation and the Information Soup* ed. by D. Farwell, E. Hovy, and L. Gerber, p.150.

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

[6] P.F. Brown, J.C. Lai, and R.L. Mercer, "Aligning Sentences in Parallel Corpora", in *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics.*, 1991.

[7] M.Franz, J.S.McCarley, and S.Roukos, "Ad hoc and multilingual information retrieval at IBM", in *The 7th Text REtrieval Conference (TREC-7)* ed. by E.M. Voorhees and D.K.Harman, 1999.

[8] 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.

[9] Jinxi Xu and W.B. Croft, "Query Expansion using Local and Global Document Analysis", in *19th Annual ACM SIGIR Conference on Information Retrieval*, 1996.