IR Project Proposal

Krzysztof Czuba

February 26, 1998

In the project I'd like to investigate the different approaches to Translingual Information Retrieval. In the first part, I will use a human translated parallel corpus (UNICEF) for retrieval and methods that require training. In the second part, the human translated corpus will be replaced by a machine translated one. The automatically created corpus will be used in methods that require training, like LSI, or require translation mates, like PRF. Since the retrieval will be done from the same set of human translated documents, it will be possible to evaluate the effect of using the imperfect, automatically obtained corpus.

The idea of using an MT system is motivated by the easy availability of general-purpose machine translation. The MT system I'm planning to use is SYSTRAN, which is available on the Web from the AltaVista site. The system translates into Spanish which is the language the UNICEF collection is also available in. The advantage of using SYSTRAN is its large vocabulary which allows it to produce translations which do not contain any untranslated words.

Although the cost of obtaining a machine translated corpus in this case is low, the quality of the resulting corpus might not be sufficient for high retrieval quality. This is due to the general-purpose character of the MT system: the resulting translations might not be precise enough. On the other hand, the MT system might be able to use some syntactic and semantic knowledge in picking the translations of words which is not available to systems using plain vocabulary lookup (like DICT).

The main disadvantage I envision at this point is the fact that current MT systems operate only at the sentence-level and the disambiguation of the lexical choices is very local. Also, the system will most probably run into the paraphrase problem since MT systems are likely to pick the same translation for closely related words. This particular problem cannot be solved by creating a large size corpus.

If time permits, I will also run experiments in which queries/documents are translated using the MT system, and the translations are used to retrieve documents.

The results for part one of the project will be compared with the results reported in [1] and [6]. Part two will be evaluated against the results of part one and the results in the papers. The recall/precision curves, f1 and 11-point averages will be used to compare the different systems with and without MT.

I would like to have access to the LSI system (or share it with someone).

The timeline:

- March 2nd: project web page;
- March 10th: investigate the availability of all the systems and required corpora/query sets;
- March 17th: preliminary results for the most computation/effort intensive systems: LSI/DICT; a plan and code for obtaining the bilingual corpora by using the MT system;
- March 24th: results for the basic part with the UNICEF corpus for all the methods;
- March 30th: created the automatically translated corpora;

- April 10th: finished the computation/effort intensive methods using the generated corpus;
- April 24th: write-up, presentation.

References

[1] Jaime Carbonell, Yiming Yang, Robert Frederking, Ralf D. Brown, Yibing Geng, and Danny Lee. "Translingual Information Retrieval: A Comparative Evaluation". In Proceedings of Fifteenth International Joint Conference on Artificial Intelligence (IJCAI-97).

[2] Dumais S., Landauer T., and Littman M. "Automatic Cross-Linguistic Information Retrieval using Latent Semantic Indexing".

[3] Hull D., Grefenstette G, "Experiments in Multilingual Information Retrieval", Rank Xerox Research Center

[4] Landauer T., Littman M. (1995) "A statistical method for language-independent representation of the topical content of text segements", Bellcore.

[5] Littman M., Keim G., (1997) "Cross-Language Text Retrieval with Three Languages", CS-1997-16, Duke University.

[6] Littman M., Jiang F. (1998) "A comparison of Two Corpus-Based Mehods for Translingual Information Retrieval".

[7] Oard D. "Alignment of Spanish and English TREC Topic Descriptions", University of Maryland.

[8] Oard D. W., "Alternative approaches for Cross-Language Text Retrieval", University of Maryland.

[9] Oard D. W., Dorr B., "A Survey of Multilingual Text Retrieval", CS-TR-3615, University of Maryland.

[10] Rehder B., Littman M., Dumais S., Landauer T. "Automatic 3-Language Cross-Language Information Retrieval with Latent Semantic Indexing".

[11] Sheridan P., Ballerini J. P., P. Schäuble, "Building a Large Multilingual Test Collection from Comparable News Documents", Proceedings of Workshop on Cross-linguistic Information Retrieval, held in conjunction with the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Zürich, Switzerland, August 1996.

[12] Sheridan P., Braschler M., Schäuble P., "Cross-Language Information Retrieval in a Multilingual Legal Domain", First European Conference on Research and Advanced Technology for Digital Libraries, Pisa, Italy, September 1997.

[13] Sheridan P., Wechsler M., Schäuble P., "Cross-Language Speech Retrieval: Establishing a Baseline Performance", Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Philadelphia, Pennsylvania, July 1997.

[14] Wechsler M., Sheridan P., Schäuble P., "Multi-Language Text Indexing for Internet Retrieval", Proceedings of the 5th RIAO Conference, Computer-Assisted Information Searching on the Internet, Montreal, Canada, June 1997.