FREE RESOURCES AND ADVANCED ALIGNMENT FOR CROSS-LANGUAGE TEXT RETRIEVAL

Mark W. Davis and William C. Ogden

Computing Research Laboratory
New Mexico State University
Box 30001/3CRL
Las Cruces, NM 88003
madavis@crl.nmsu.edu

ABSTRACT

For the Cross-Language Text Retrieval Track in TREC 6, NMSU experimented with a new approach to deriving translation equivalents from parallel text databases, and also investigated performing automatic, dictionary-based translation of query terms by using a dictionary that could be queried remotely via the World Wide Web. The new approach to building bilingual translation lexicons involved aligning parallel texts at the sentence level, and then performing further alignments at the sub-sentence level. The process initially chooses alignment anchors based on N-gram matches between cognate terms. Term and phrase matching is then performed between the anchor points by finding the most direct path from one anchor to the next, penalizing larger steps over runs of terms. The collected term translations are then used as equivalents for a query translation process and the translated query is then submitted to a monolingual retrieval engine. The results are compared against the performance of a monolingual French-French retrieval system, and against a translated query formed from a very high-quality bilingual dictionary accessed directly over the World Wide Web. A combined approach is also presented that uses terminology from both the dictionary and, where the dictionary lacks coverage, supplements the query translation using terms from a parallel text database.

OVERVIEW

A Cross-Language Text Retrieval (CLTR) system retrieves documents in a language that is different from the query language. Various approaches have been proposed for CLTR. An early experiment used hand-built translation thesauri (Salton, 1971). Recent work has extended the use of lexicons to make use of bilingual machine-readable dictionaries (MRDs) of a general nature to translate query terminology (Ballesteros and Croft, 1997, Davis and Ogden, 1997, Kwok, 1997 and Hull and Grefenstette 1996). A subtle variation on these methods is to use controlled-vocabulary thesauri or other specialized resources for query translation, which often provide excellent coverage of highly technical subject matter.

A dramatic alternative to the use of prepared bilingual or multilingual lexicons is to rely on the information contained in parallel texts (texts that are translations of each other) to train or derive a translation model. One approach, Latent Semantic Indexing (LSI), maps documents into a reduced-dimensionality space, based purely on term-document co-occurrence statistics in parallel texts (Dumais, Landaurer and Littman, 1995). Queries can, in turn, be mapped into the same space and the nearest documents to the query returned. In LSI, queries can be expressed in any of the shared languages of the training texts. Alternative methods include training linear, but underdetermined, translation models using an iterative least-squares minimization (Dunning and Davis, 1993), and applying stochastic optimization approaches to try to match query performance in both languages over a parallel training text set (Davis and Dunning, 1996). Approaches based on these methods have shown moderate promise, although no large-scale implementation has yet demonstrated performance that matches hybrid methods or methods that rely exclusively on MRDs for the translation task.

Despite some early experimental successes, the task of Cross-Language Text Retrieval remains dauntingly difficult, if only for the reason that resources for translation remain exceedingly expensive. Even for a system that shows startling performance in one language pair, moving to a new language pair often requires completely new resources and personnel. Parallel text is the relatively rare by-product of large-scale translation operations, while bilingual MRDs are the tightly-held intellectual property of dictionary companies, commanding impressive royalty fees for widespread application. Further, tuned lexicons for machine translation applications remain the most closely guarded inner secrets of machine translation companies. A third alternative, "comparable texts", which are matched according to topic, but not necessarily direct translations, are also plausible resources for extracting query translation terminology, but are not clearly easier to amass than true parallel text.

At NMSU's Computing Research Laboratory, we have found that the problems associated with the lack of good resources for translation are rapidly being offset by the increased availability of materials on the World Wide Web (WWW). Our approach for the TREC-6 Cross-Language Retrieval Track only uses freely available WWW resources to translate English queries into French, using a combination of new text alignment techniques for parallel text WWW resources, and bilingual MRDs. The resulting French queries are then submitted to a monolingual retrieval engine to retrieve French documents. The resulting documents could then be translated or glossed back into English using the same resources combined in an approach like that presented in Davis and Ogden (1997). Our TREC-6 submission continues to emphasize our commitment to one very practical scenario: a monolingual information retrieval user who submits a query against a collection of documents in another language, and who will then need translation aids to assess the relevance of the retrieved documents.

IS THERE A FREE LUNCH?

On-line bilingual dictionaries represent a powerful new opportunity for research and development of CLTR technology. Using a list of morphological root forms for terms in a large English text collection, custom Web robots can acquire on-line resources like bilingual resources for use in CLTR tasks. We have recently developed robots to do exactly that and have acquired reasonable "kernel" bilingual dictionaries from English to ten other languages. The number of headwords available for each language pair is small in comparison with printed dictionaries, but can be quickly expanded by a user with access to corpus analysis tools.

Languages	***Headwords***
English-Afrikaans	3,733
English-Dutch	9,853
English-Danish	3,715
English-Finnish	2,832
English-French	3,582
English-Japanese	176,528
English-Hungarian	2,479
English-Italian	2,912
English-Portuguese	2,637
English-Spanish	5,201

Table 1 Headwords for bilingual dictionaries

A comparison of the coverage of a large corpus by a kernel dictionary like the English-Spanish dictionary in the table, above, to a larger print dictionary is revealing. For a collection of 10.7 million words (TREC Spanish AFP collection) and 207,433 unique words filtered by a 30,805 word Collins bilingual dictionary headword list, case-normalized and stemmed in IR fashion (Davis, 1996), 187,103 words remain (90.2%). The 5,201-word dictionary leaves 204,227 words (98.5%).

An analysis of a randomly drawn 100 words from the unaccounted-for segment using the 30,805 word Collins dictionary indicates that 11 were abbreviations, 9 were foreign words, 49 were proper names and 31 were other words. If this pattern is representative of the collection as a whole, then abbreviations, proper names and foreign words represent a startling 69% of untranslatable words. Some abbreviations between Latin and Germanic languages go straight across (km), but some do not (NATO and OTAN), and most proper names go directly across, or require only minor accent normalization rules to account for. This pattern will not hold for translations between radically different language pairs, however, and we can expect that as CLTR is expanded to handle distally-related language pairs that the impact of these terms will grow as well.

The promise of using parallel corpora to compensate for the narrow view of dictionaries does not appear to present dramatically wider coverage of a corpus. The Spanish parallel document set from Pan American Health Organization (PAHO), consisting of 22,094 unique words drawn from 94,313 total words, leaves 201,660 words (97.2%) when filtering the same AFP document set. Making good use of parallel corpora for translation is imperfect at best, however, so the potential value of even this meager amount of coverage to CLTR applications remains suspect.

It seems that if there is a free lunch for CLTR due to free resources, then it is primarily due to the limited coverage provided by *any* translation resource, and the significant impact that direct matching of proper names and abbreviations has on retrieval performance for specialized queries. This benefit will likely disappear when, say, applying English queries to Chinese databases, without significant work for developing extensive proper name databases, or so-called *onomastica*, or for providing the ability to transliterate proper name expressions into the target language.

TREC 6 AND THE CLTR TRACK

Our experiments for the CLTR track of TREC 6 involved several freely available resources. First we remotely queried a large English-French bilingual dictionary at the University of Chicago to obtain translations of English query terms. We then supplemented the dictionary translation with additional terms derived from a parallel text database created using phrase-level alignment. All of the cross-language studies used only the French description field of the queries, a short statement of the topic.

The University of Chicago dictionary was created for use in a machine translation project and was therefore fairly clean, requiring minimal filtering to extract the key equivalent set. The dictionary contained entries for 209 out of 257 English terms (81.3%), with notable omissions including:

acupuncture
AIDS
resurgence
worldwide
franchise
pollution
Berlin
labor

Table 2 Key terms not covered by French-English dictionary

Not being able to translate AIDS in topic 7 presented perhaps the most serious deficiency for the system. Although “acupuncture” and “Berlin” translate straight across, AIDS does not. Our hypothesis was that for a high-quality bilingual dictionary like this one, the most pressing need was to improve the dictionary’s coverage for terms or phrases that were not in the original dictionary.

ALIGNMENT AND PHRASE EXTRACTION

To supplement the dictionary, we used parallel French-English parliamentary proceedings acquired automatically from the Canadian government archives. The English document set contained 51,732 words, while the French set contained 52,281 words. The documents were first aligned at the sentence and sentence-pair level using the statistical alignment procedure reported in Davis, Dunning and Ogden (1995), and which has been used to align Spanish and English document sets for past TREC experiments (Davis and Ogden, 1997). The second part of the alignment procedure involved discovering phrase and word matches between the aligned blocks at the sub-sentence level. Unlike the methods reported in Gale and Church (1991) in which the statistics are concerned only with the relative rates of co-occurrence between terms in aligned blocks, our approach emphasized the order of text within and between the French and English blocks. In this respect, our methodology perhaps most closely resembles the methods of Melamed (1996), but uses n-gram matching between cognate terms within that ordered set as the primary feature for establishing anchors. Once these anchors are established, the regions between the anchors are analyzed using a phrase-finder tuned for English and French to extract significant matching phrasal terminology to extend the bilingual dictionary. This procedure resulted in a dictionary of 7,869 pairs, including "sida" translates to "AIDS".

RESULTS

The CLTR results demonstrated extremely wide variance between the performance of the best and worst groups. A preliminary comparison between the CRL cross-language runs and all submissions shows CRL performing at or above the median on 13 out of 21 judged topics, with the cross-language system turning in the best performance of all systems on one topic, and among the worst on one other topic. Although in the preliminary results we saw cross-language results that were superior to the monolingual case, in the revised results the best of the cross-language runs were slightly poorer than the cross-language case. The results are graphed in Figure 1 on Page 6.

Run	*Average Precision-Recall (all queries)*	*% of monolingual case*
Monolingual French	0.1484	—
English-French using unmodified English-French dictionary	0.1371	92.38%
English-French using parallel-corpus derived bilingual lexicon	0.0357	24.06%
English-French using augmented dictionary	0.1428	96.23%

Table 3 Performance of CLTR approaches

The consequences of augmenting the dictionary using parallel texts are especially evident in Query 7, which contained the term AIDS. AIDS was not in the dictionary but was successfully added by the parallel text process. The average precision-recall for Query 7 went from .0095 to .1296 due to the addition of the discovered term. Indeed, all of the performance difference

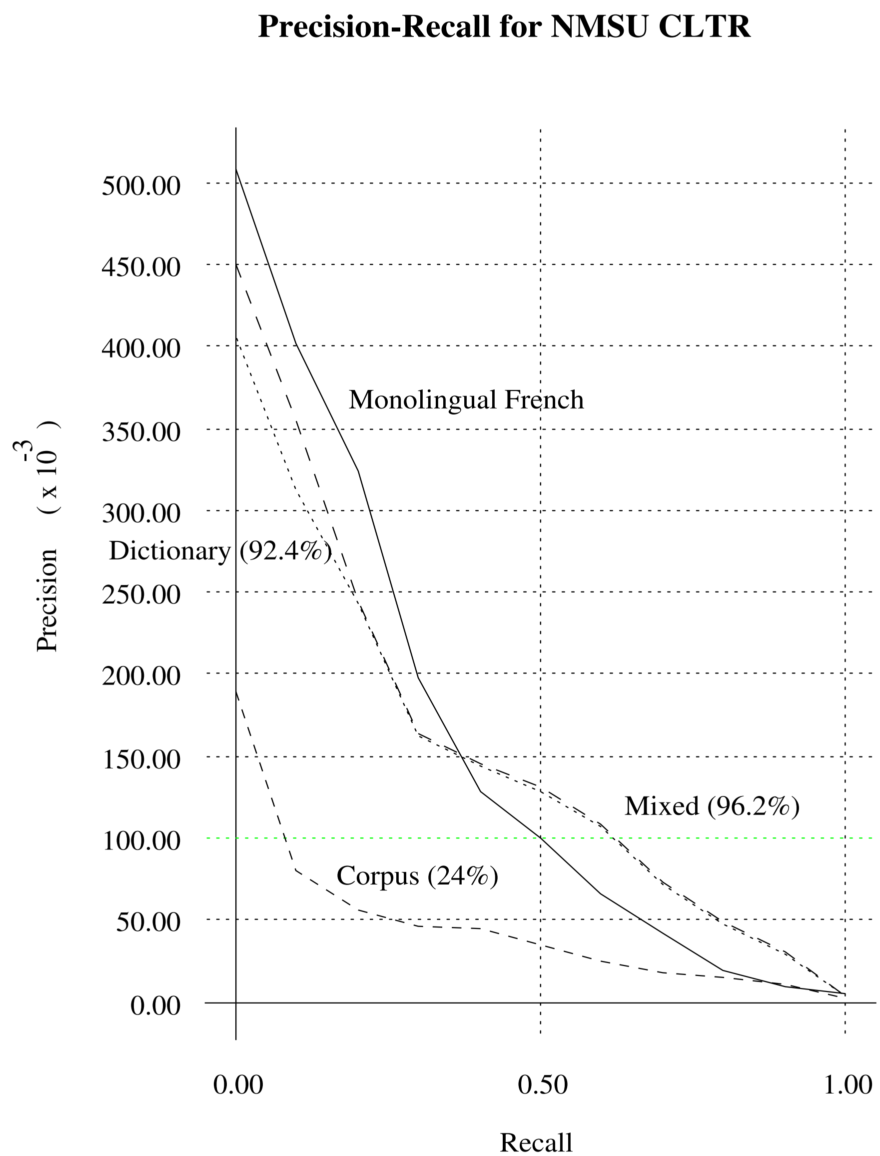

Figure 1. Average precision-recall curves for monolingual, dictionary and corpus-based cross-language retrieval techniques, as well as a method that combines the two approaches.

between the unmodified dictionary and augmented dictionary runs can be attributed to the improved performance on Query 7.

Of special interest is the comparatively poor perfomance of the corpus-based methods alone. These results bear a close similarity to those of other approaches both in TREC7 and TREC6 (Davis, 1996). The general conclusion is that extracting information and term equivalences from parallel corpora is a noisy process that results in highly noisy translation resources. The best role of such resources seems to be as a supplement to hand-prepared dictionaries, rather than as a substitute for them. Note also that extraction methods that use high-frequency term filtering to reduce error rates in turn substantially reduce the coverage of the derived dictionary to phrases and terms that already occur in hand-prepared, general bilingual resources.

CONCLUSIONS

Cross-Language Text Retrieval is a difficult problem that is compounded by the need for rare resources like parallel texts and high-quality bilingual dictionaries. Our experiments showed that a CLTR system can be successfully built using only freely-available translation resources captured from the World Wide Web.

REFERENCE

Ballesteros, L. and W. B. Croft (1997) "Phrasal Translation and Query Expansion Techniques for Cross-Language Information Retrieval," in SIGIR '97, Philadelphia, PA, July 27-31.

Davis, M. W. and Ogden, W.C. (1997) "QUILT: Implementing a Large-Scale Cross-Language Text Retrieval System," in SIGIR '97, Philadelphia, PA, July 27-31.

Davis, M. W. (1996) New Experiments in Cross-Language Text Retrieval at New Mexico State University's Computing Research Laboratory. In *Proceedings of the Fifth Text Retrieval Evaluation Conference*, Gaithersburg, MD, National Institute of Standards and Technology.

Davis, M. W. and T. Dunning (1996) Query Translation Using Evolutionary Programming for Multi-Lingual Information Retrieval II. In *Proceedings of the Fifth Annual Conference on Evolutionary Programming*, San Diego, Evolutionary Programming Society.

Davis, M. W., T. Dunning, and W. Ogden (1995) Text Alignment in the Real World: Improving Alignments of Noisy Translations Using Common Lexical Features, String Matching Strategies and N-Gram Comparisons. In *Proceedings of the Conference of the European Chapter of the Association of Computational Linguistics*. University College Dublin. March.

Dunning, T. E., and M. Davis (1993) Multi-Lingual Information Retrieval. *Memoranda in Computer and Cognitive Science*, MCCS-93-252, Computing Research Laboratory, New Mexico State University.

Dumais, S. T., T. Landaurer and M. Littman (1995) Automatic Cross-Linguistic Information Retrieval Using Latent Semantic Indexing. In *Proceedings of the Workshop on Cross-Linguistic Information Retrieval*, SIGIR'96, Zurich.

Gale, W. A. and K. W. Church (1991) A Program for Aligning Sentences in Bilingual Corpora. In *Proceedings of the 29th Annual Conference of the Association of Computational Linguistics,* 177-184, Berkeley, CA.

Hull, D. and Grefenstette, G. (1996) "Experiments in Cross-linguistic Information Retrieval" in *SIGIR96*, August, Zurich, CH.

Kwok, K. L., (1997) "Evaluation of an English-Chinese Cross-Lingual Retrieval Experiment," in Working Notes of the Cross-Language Text and Speech Retrieval Spring Symposium, AAAI-97 Spring Symposium, March 24-26, Stanford, CA.

Salton, G. (1971) "Automatic Processing of Foreign Language Documents," in *The Smart Retrieval System*, ed. Salton, G., Prentice-Hall, Englewood Cliffs, NJ.

APPENDIX 1: CLTR Questionnaire

To those of you in the CLIR track who are new to TREC, this questionnaire makes a distinction between "topics", the descriptions furnished by NIST, and "queries", the actual text submitted to your retrieval system for searching.

Queries may simply be a copy of some part or all of the topic, may be derived automatically from the topic, or may be formulated manually based on the topic description.

1. OVERALL APPROACH:

1.1 What basic approach do you take to cross-language retrieval?

[X] Query Translation
[ ] Document Translation
[ ] Other:

1.2 Were manual translations of the original NIST topics used as a starting point for any of your cross-language runs?

[ ] No
[X] Yes: The NIST-supplied English topics.

1.3 Were the automatically translated (Logos MT) documents used for any of your cross-language runs?

[X] No
[ ] Yes:

1.4 Were the automatically translated (Logos MT) topics used for any of your cross-language runs?

[X] No
[ ] Yes:

2. MANUAL QUERY FORMULATION:

2.1 If query formulation involved manual effort, how fluent was the user in the source (query) language?

[ ]

2.2 If query formulation involved manual effort, how fluent was the user in the target (document) language?

[ ]

3. USE OF MANUALLY GENERATED DATA RESOURCES:

3.1 What kind of manually generated data resources were used?

[X] Dictionaries
[ ] Thesauri
[ ] Part-of-speech Lists
[ ] Other:

3.2 Were they generated with information retrieval in mind or were they taken from related fields?

[ ] Information Retrieval
[X] Machine Translation
[ ] Linguistic Research
[ ] General Purpose Dictionaries
[ ] Other:
3.3 Were they specifically tuned for the data being searched (ie. with special terminology) or general-purpose?
[ ] Tuned for data; Please specify:
[X] General purpose
3.4 What amount of work was involved in adapting them for use in your information retrieval system.
[ ] None
[X]: robotic retrieval from WWW, filtering to eliminate duplicate headwords
3.5 Size
[X] 3582 entries
[ ] _ MBytes
3.6 Availability? - Please also provide sources/references!
[ ] Commercial
[ ] Proprietary
[X] Free: www.travlang.com and www.uchicago.humanities.edu
[ ] Other:
4. USE OF AUTOMATICALLY GENERATED DATA RESOURCES:
4.1 Form of the automatically constructed data resources?
[X] Lexicon
[ ] Thesaurus
[ ] Similarity matrix
[ ] Other:
4.2 What sort of training data was used to construct them?
[ ] Same data as used for searches:
[ ] Similar data as used for searches:
[X] Other data:
4.3 Size
[ ] _ entries
[ ] _ MBytes
4.4 Was there any manual clean-up involved in the construction process?
[ ] Yes:
[ ] No
4.5 Rough resource estimates for building the data resources (ie. an indicator of the computational complexity of the process).
[X] 0.1 hours
[ ] _ MBytes of memory used
[ ] _ temporary disk space
5. GENERAL
5.1 How dependent is the system on the data resources used? Could they easily be replaced if better sources were available?
[ ] Very dependent, _
[ ] Somewhat dependent,
[X] Easily replacable, _
[ ] Don't know
5.2 Would the approach used potentially benefit if there were better data resources (e.g. bigger dictionary or more/better aligned texts for training) available for tests?

[ ] Yes, a lot, _

[X] Yes, somewhat: see estimates of data coverage in

[ ] No, not significantly, _

[ ] Don’t know

5.3 Would the approach used potentially suffer a lot if similar data resources of lesser quality (noisier dictionary, wrong domain of terminology) were used as a replacement?

[ ] Yes a lot, _

[X] Yes, somewhat: see estimates in See “IS THERE A FREE LUNCH?” on page 2.

[ ] No, not significantly,_

[ ] Don’t know

5.4 Are similar resources available for other languages than those used?

[X] Yes, French, Italian, Portugese, Japanese, Hungarian, Aftikaans, Dutch, Danish, Finnish

[ ] No