Research to Improve Cross-Language Retrieval – Position Paper for CLEF

Fredric C. Gey
e-mail: gey@ucdata.berkeley.edu

UC Data Archive & Technical Assistance,
University of California
Berkeley, CA 94720 USA

Abstract. Improvement in cross-language information retrieval results can come from a variety of sources – failure analysis, resource enrichment in terms of stemming and parallel and comparable corpora, use of pivot languages, as well as phonetic transliteration and Romanization. Application of these methodologies should contribute to a gradual increase in the ability of search software to cross the language barrier.

1 Failure Analysis

In my opinion there has been a dearth of detailed failure analysis in cross-language information retrieval, even among the best-performing methods in comparative evaluations at TREC, CLEF, and NTCIR[7]. Just as a post-mortem can determine causation in mortality, a query-by-query analysis can often shed light on why some approaches succeed and others fail. Among the sets of queries utilized in these evaluations we always find several queries which all participants perform poorly (as measured by the median precision over all runs for that query). When the best performance is significantly better than the median it would be instructive to determine why that method succeeded while others failed. If the best performance is not much better than the median, then something inherently difficult in the topic description presents a research challenge to the CLIR community. Two examples are illustrative, one from TREC and the other from CLEF.

The TREC-7 conference was the first multilingual evaluation where a particular topic language was to be run against multiple language document collections. The collection languages were the same as in CLEF (English, French, German, Italian). Topic 36, whose English title is "Art Thefts" has the French translated equivalent "Les voleurs d'art". The Altavista Babblefish translation of the French results in the phrase "The robbers of art", which grasps the significance, if not the additional precision of the original English. However, when combined with aggressive stemming, the meaning can be quite different. The Berkeley French⟶Multilingual first stemmed the word 'voleurs' to the stem 'vol', and the translation of this stem to English is 'flight' and to German 'flug,' significantly different from the original unstemmed translation. In fact our F⟶EFGI performance for this query was 0.0799 precision versus our E⟶EFGI precision of 0.3830.

For the CLEF evaluation, one query provides a significant example of the challenges facing CLIR, even with a single language such as English. Query 40 about the privatization of the German national railway was one which seems to have presented problems with all participating groups (the median precision over all CLEF multilingual runs was 0.0537 for this query). As an American group, the Berkeley group was challenged by the use of the English spelling 'privatisation' which couldn't be recognized by any machine translation softwares. The German version of the topic was not much better – in translation its' English equivalent became 'de-nationalization' a very uncommon synonym for 'privatization,' and one which yielded few relevant documents. By comparison, our German manual reformulation of this query resulted in an average precision of 0.3749 for best CLEF performance for this query.

These examples illustrate that careful post-evaluation analysis might provide the feedback which can be incorporated into design changes and improved system performance.

2 Resource Enrichment

2.1 Stemmers and Morphology

The CLEF evaluation seems to be the first one in which significant experiments in multiple language stemming and morphology was used. Some groups developed "poor man" stemmers by taking the corpus word lists and developing stem classes based upon common prefix strings. The Chicago group applied their automatic morphological analyzer to the CLEF collections to generate a custom stemmer for each language's collection[5], while the Maryland group extended the Chicago approach by developing a four-stage statistical stemming approach[14]. The availability of the Porter stemmers in French, German and Italian (from http://open.muscat.com/) also heavily influenced CLEF entries. The conclusion seems to be that stemming plays an important role in performance improvement for non-English European language, with results substantially better than for English stemming.

2.2 Parallel Corpora and Web Mining

Parallel corpora have been recognized as a major resource for CLIR. Several entries into CLEF, in particular the Johns Hopkins APL[11] used aligned parallel corpora in French and English from the Linguistic Data Consortium. More recently emphasis has been given toward mining the WWW for parallel resources. There are many sites, particularly in Europe, which have versions of the same web page in different languages. Tools have been built which extract parallel bilingual corpora from the web [13, 16]. These were applied in CLEF by the Montreal Group[12] and the Twente/TNO group[6]

2.3 Comparable Corpora Alignment

Comparable corpora are bilingual corpora which can be created through alignment of similar documents on the same topic in different languages. An example might be the foreign edition of a newspaper where stories about the same news item are written independently. Techniques for alignment require relaxation of time position (a story might appear a few days later) and the establishment of the contextual environment of topic. There has been research into the statistical alignment of comparable corpora by Picchi and Peters with Italian and English [15] and Fung with English and Chinese [2] but the techniques have not made their way into general practice. Comparable corpora will only become widely used if tools for their acquisition are created as open-source software and tools for their alignment are refined and also made available.

2.4 Geographic and Proper Names

A major need is to provide geographic and proper name recognition across languages. Proper names are often not in either machine translation programs or bilingual dictionaries, nor are geographic place names. A particular case in point was the TREC-6 cross language query CL1 about Austrian President Kurt Waldheim's connection with Nazism during WW II – one translation system translated from the German 'Waldheim' to English 'forest home'.

It has been suggested that more than thirty percent of content bearing words from news services are proper nouns, either personal and business enterprise names or geographic place name references. The availability of electronic gazetteers such as:

- National Imagery and Mapping Agency's country name files:
  http://164.214.2.59/gns/html/Cntry_Files.html

- Census Bureau's gazetteer for United States:
  http://tiger.census.gov/

- Arizona State University's list of place name servers
  http://www.asu.edu/lib/hayden/govdocs/maps/geogname.htm

- Global Gazeteer of 2880532 cities and towns around the world
  http://www.calle.com/world/

give some hope that geographic name recognition could be built into future CLIR systems.

While work has been done on extracting proper nouns in English and some other languages through the Message Understanding Conference series, it is not clear that anyone has mined parallel texts to create specialized bilingual lexicons of proper names.

3 Pivot Languages

In multilingual retrieval between queries and documents in n languages, one seems to be required to possess resources (machine translation, bilingual dictionaries, parallel corpora, etc.) between each pair of languages. Thus $O(n^2)$ resources are needed. This can be approximated with the the substitution of transitivity among $O(n)$ resources if a general purpose pivot language is used. Thus to transfer a query from German to Italian, where machine translation is available from German to English and English to Italian respectively, the query is translated into English and subsequently into Italian, and English becomes the pivot language. This method was used by the Berkeley group in TREC-7 [3] and CLEF[4]. The Twente/TNO group has utilized Dutch as a pivot language between pairs of language where direct resources were unavailable in both TREC-8 [10] and CLEF[6]. One can easily imagine that excellent transitive machine translation could provide better results than poor direct resources such as a limited bilingual dictionary. In some cases resources may not even exist for one language pair – this will be come increasingly common with the increase in the number of languages for which cross-language information search is desired. For example, a CLIR researcher may be unable to find an electronic dictionary resource between English and Malagasy (the language of Madagascar), but there are French newspapers in this former colony of France where French is still an official language. Thus, an electronic French-Malagasy dictionary may be more complete and easier to locate than an English-Malagasy one. Similarly the Russian language may provide key resources to transfer words from the Pashto (Afgan), Farsi, Tajik, and Uzbek languages (see, for example, http://members.tripod.com/G̃roznijat/b_lang/bl_sourc.html).

4 Phonetic Transliteration and Romanization

One of the most important and neglected areas in cross-language information retrieval is, in my opinion, the application of transliteration to the retrieval process. The idea of transliteration in CLIR derives from the suggestion by Buckley in the TREC-6 conference that for English-French CLIR "English query words are treated as potentially mis-spelled French words."[1] In this way English query words can be replaced by French words which are lexicographically similar and the query can be can proceed monolingually. More generally, we can often find that many words, particularly in technology areas, have been borrowed phonetically from English and are pronounced similarly, yet with phonetic customization in the borrower language. The problems of automatic recognition of phonetic transliteration has been studied by Knight and Graehl for the Japanese katakana alphabet [9] and by Stalls and Knight for Arabic[17]. Another kind of transliteration is Romanization, wherein an unfamiliar script, such as Cyrillic, is replaced by its Roman alphabet equivalent. When done by library catalogers, the transformation is one-to-one, i.e. the original script can be recovered by reverse transformation. This is not the case for phonetic transliteration where more than one sound in the source language can project to a single representation in the target language. The figure below comes from the entry for ‘economic policy’ in the GIRT special domain retrieval thesaurus of CLEF[8]. The GIRT creators have provided a translation of the thesaurus into Russian which our group

```
- <list>
  - <entry>
      <german>Wirtschaftspolitik</german>
      <russian>экономическая политика</russian>
      <translit>ekonomicheskaia politika</translit>
    </entry>
  </list>
```

Fig. 1. German-Russian GiRT Thesaurus with Transliteration

has transliterated into its Roman equivalent using the U.S. Library of Congress specification (see http://lcweb.loc.gov/rr/european/lccyr.html). It is clear that either a fuzzy string or phonetic search with English words ‘economic’, ‘policy’, or ‘politics’ would retrieve this entry from the thesaurus or from a collection of Russian documents. Generalized string searches of this type have yet to be incorporated into information retrieval systems.

5 Summary and Acknowlegments

This paper has presented a personal view of what developments are needed to improve cross-language information retrieval performance. Two of the most exciting advances in cross-language information retrieval are mining the web for parallel corpora to build bi-lingual lexicons and the application of phonetic transliteration toward search in the absence of translation resources. Comparable corpora development, which has perhaps the greatest potential to advance the field, has yet to achieve its promise in terms of impact, probably because of the lack of generally available processing tools.

I wish to thank Hailing Jiang and Aitao Chen for their support in running a number of experiments and Natalia Perelman for implementing the Russian transliteration of the GIRT thesaurus. Major funding was provided by DARPA (Department of Defense Advanced Research Projects Agency) under research grant N66001-00-1-8911, Mar 2000-Feb 2003 as part of the DARPA Translingual Information Detection, Extraction, and Summarization Program (TIDES).

References

1. C Buckley, J Walz, M Mitra, and C Cardie. Using clustering and superconcepts within smart: Trec-6. In E.M. Voorhees and D. K. Harman, editors, *The Sixth*

Text REtrieval Conference (TREC-6), NIST Special Publication 500-240, pages 107–124, August 1998.

2. Pascal Fung. A statistical view on bilingual lexicon extraction: From parallel corpora to non-parallel corpora. In D Farwell L Gerber E Hovy, editor, *Proceeding of AMTA-98 Conference, Machine Translation and the Information Soup Pennsylvania, USA, October 28-31, 1998*, pages 1–16. Springer-Verlag, 1998.
3. F. C. Gey, H. Jiang, and A. Chen. Manual queries and machine tranlation in cross-language retrieval at trec-7. In E.M. Voorhees and D. K. Harman, editors, *The Seventh Text REtrieval Conference (TREC-7), NIST Special Publication 500-242*, pages 527–540. National Institute of Standards and Technology, July 1999.
4. Fredric Gey, Hailing Jiang, Vivien Petras, and Aitao Chen. Cross-language retrieval for the clef collections - comparing multiple methods of retrieval. In *this volume*. Springer, 2000.
5. John Goldsmith, Darrick Higgins, and Svetlana Soglasnova. Automatic language-specific stemming in information retrieval. In *this volume*. Springer, 2000.
6. Djoerd Hiemstra, Wessel Kraaij, Renee Pohlmann, and Thijs Westerveld. Translation resources, merging strategies, and relevance feedback for cross-language information retrieval. In *this volume*. Springer, 2000.
7. Noriko Kando and Toshihiko Nozue, editors. *Proceedings of the First NTCIR Workshop on Japanese Text Retrieval and Term Recognition*. NACSIS (now National Informatics Institute, Tokoyo, 1999.
8. Michael Kluck and Fredric Gey. The domain-specific task of clef - structure and opportunities for specific evaluation strategies in cross-language information retrieval. In *this volume*. Springer, 2000.
9. K. Knight and J. Graehl. Machine transliteration. In *Computational Linguistics, 24(4)*, 1998.
10. Wessel Kraaij, Renee Pohlmann, and Djoerd Hiemstra. Twenty-one at trec-8: Using language technology for information retrieval. In Ellen Voorhees and D Harman, editors, *Working Notes of the Eighth Text REtrieval Conference (TREC-8)*, pages 203–217, November 1999.
11. Paul McNamee, James Mayfield, and Christine Piatko. A language-independent approach to european text retrieval. In *this volume*. Springer, 2000.
12. Jian-Yun Nie, Michel Simard, and George Foster. Multilingual information retrieval based on parallel texts from the web. In *this volume*. Springer, 2000.
13. Jian-Yun Nie, Michel Simard, Pierre Isabelle, and Richard Durand. Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the web. In *SIGIR '99: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 15-19, 1999, Berkeley, CA, USA*, pages 74–81. ACM, 1999.
14. Douglas Oard, Gina-Anne Levow, and Clara Cabezas. Clef experiments at the university of maryland: Statistical stemming and backoff translation strategies. In *this volume*. Springer, 2000.
15. Eugenio Picchi and Carol Peters. Cross language information retrieval: A system for comparable corpus querying. In Gregory Greffenstette, editor, *Cross Language Information Retrieval*, pages 81–91. Kluwer, 1998.
16. Phillip Resnick. Mining the web for bilingual text. In *Proceedings of 37th Annual Meeting of the Association for Computational Linguistics (ACL'99), College Park, Maryland, June 1999*, 1999.
17. B. Stalls and K. Knight. Translating names and technical terms in arabic text. In *Proc of the COLING/ACL Workshop on Computational Approaches to Semitic Languages, 1998*, 1998.