Corpus-Based Query Translation for Translingual Information Retrieval

Ralf D. Brown
Language Technologies Institute, Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA 15213 USA

1 Introduction

Translingual IR (TIR) – also called Cross-Language IR – has recently begun to receive considerable attention, primarily driven by the increased availability of international text collections, including the World Wide Web. The most popular approaches in TIR use a machine-readable dictionary (MRD) for query translation[5]; their performance relative to bilingual-corpus based approaches has not previously been evaluated. In a recent set of experiments[3], both machine-translation and IR approaches to TIR were compared, including the corpus-based query translation and the general-purpose bilingual dictionary-based query translation described here.

Since most real-life queries consist of isolated words or very short phrases, full-fledged machine translation is not applicable to query translation. Instead, various word-based approaches were investigated, including:

- *Dictionary-based Term Translation* (DICT) – Look up each query term in a general-purpose bilingual dictionary, and use all its possible translations.
- *Example-based Term Translation* (EBT) – Use a sentence-aligned bilingual training corpus to find the terms that co-occur in context across languages, thus creating a corpus-based term-equivalence matrix.

Both of the above also provide some query expansion during translation.

1.1 PanEBMT Translations

The Pangloss EBMT engine (PanEBMT)[1] uses a combination of Example-Based Machine Translation (EBMT) and simple bilingual dictionary lookup to translate each sentence in its input.

The EBMT parallel corpus was derived via a sentence-aligning process primarily from the Spanish and English portions of the UN Multilingual Corpus [4], with an admixture of texts from the Pan-American Health Organization and ARPA MT evaluations. The entire corpus comprises some 685,000 sentence pairs totalling more than 250 megabytes of text.

Once phrasal matches in the corpus have been found, subsentential alignment is performed on the sentence pairs containing them. The alignment uses a number of heuristic scoring functions in addition to a bilingual dictionary for determining possible correspondences between words in the source and target sentences of a sentence pair; the highest-scoring phrase in the target sentence is returned as the translation of the matched phrase in the source sentence.

Dictionary lookups are performed using the same bilingual dictionary that the PanEBMT engine uses as part of its subsentential alignment algorithm. This dictionary is the component that differs between the DICT and EBT methods. PanEBMT also uses word-root lists containing some 88,000 Spanish forms and 26,500 English forms to improve the possible-correspondence determination during subsentential alignment.

For the queries used here, which were not full sentences, but primarily single words and short phrases, PanEBMT was only able to translate occasional phrases, so the query translations consisted nearly exclusively of terms from the dictionary. Thus the effectiveness of the two approaches depends almost entirely on the quality of the dictionaries used.

1.2 Dictionary-based Term Translation

For the dictionary-based query translation, we use a general-purpose bilingual dictionary as the PanEBMT dictionary. This dictionary was originally created for Spanish-to-English translations from a machine-readable copy of the Collins Spanish-English dictionary, with manual and statistically-generated corrections and additions.

For English-to-Spanish, we use an inverted copy of the corrected but unaugmented Spanish-English dictionary. Its size and quality are substantially lower than for Spanish-to-English (e.g. 27,200 English words vs. 51,500 Spanish words), due both to the original dictionary being intended for Spanish-to-English translation, and to the lack of augmentation.

1.3 Example-based Term Translation

Since a large sentence-aligned bilingual corpus was available, it was natural to attempt to generate a dictionary from the corpus. Such automatic generation exhibits a surprising degree of accuracy given the simplistic algorithm used to generate the dictionary – nothing more than a large co-occurrence table combined with a thresholding scheme[2]. This was used as the dictionary for the example-based term substitution.

A very nice property of the count-and-filter method is that adjusting the filtering thresholds allows us to tune a trade-off between translation accuracy and yield. Interestingly, a significant number of the erroneous translations are due to monolingual co-occurrences, e.g. "Burkina" generates translations of both "Burkina" and "Faso" because the country name "Burkina Faso" always appears as a unit in both sides of the corpus. This sort of "error" in fact provides a degree of query expansion. In addition, the number of co-occurrences of each candidate translation with the word is stored, and is used to weight the translations (by outputting each translation of a word a variable number of times according to the relative frequency of co-occurrence).

2 Empirical Evaluation

We evaluated the query-translation methods on a corpus of UN documents referring repeatedly to UNICEF, using English queries to retrieve Spanish documents. Human relevance judgments were used as the primary evaluation standard. We computed the standard 11-point average precision for each method. As a secondary performance measure, we compare the retrieval degradation when moving from monolingual to translingual IR. The idea is that perfect translingual retrieval would retrieve exactly the corresponding set of documents as monolingual retrieval, but in the target language.

For the query-translation methods, the monolingual baseline is the SMART system without relevance feedback ("SMART.basic"). The SMART.basic retrieval was optimized on the UNICEF collection, and then translingual retrievals were performed using the same parameters. For the EBT method, a variety of corpus-derived dictionaries were tested in order to optimize the dictionary-generation thresholds; the best-performing dictionary is shown here.

Table 1 shows the translingual retrieval results with different term weighting schemes. For comparison, translingual results reported by other researchers are shown in Table 2. Because the methods have been run on different corpora with different queries, direct comparisons on absolute 11-point precision-recall figures are not meaningful. However, the ratio of translingual over monolingual IR results may be more indicative of the relative power of the TIR methods – our DICT results show roughly the same ratio as found by others.

Despite the surface similarity between the DICT and EBT methods, the latter is trained to the corpus and exploits domain as well as frequency information, and therefore shows much superior results. Frequency information, in particular, provides a significant boost to the performance of the EBT method, by allowing SMART to give greater weight to the more likely translations of a word.

Method	Weighting	MIR	TIR	TIR/MIR
DICT	ntc.ntc	.4721	.2898	61%
DICT	ltc.ltc	.4306	.2340	54%
DICT	atc.atc	.3492	.1749	50%
EB-Term	ntc.ntc	.4721	.4318	91%
EB-Term	ltc.ltc	.4306	.3723	86%
EB-Term	atc.atc	.3492	.2803	80%

Table 1: Results of TIR methods

Site	Method	MIR	TIR	TIR/MIR
UMASS	CorpusPhrase	.20	.1358	68%
ETH	sim.thesaurus	.527	.278	40-53%
XEROX	Dict	.393	.235	60%
NMSU	Dict	-	-	73.5%

Table 2: Results of other TIR experiments

References

[1] R.D. Brown. Example-Based Machine Translation in the Pangloss System. In *Proceedings of the Sixteenth International Conference on Computation Linguistics*, pages 169–174, 1996.

[2] R.D. Brown. Automated Dictionary Extraction for "Knowledge-Free" Example-Based Translation. In *Proceedings of the Seventh International Conference on Theoretical and Methodological Issues in Machine Translation*, 1997.

[3] 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*, 1997.

[4] David Graff and Rebecca Finch. Multilingual Text Resources at the Linguistic Data Consortium. In *Proceedings of the 1994 ARPA Human Language Technology Workshop*. Morgan Kaufmann, 1994.

[5] D.A. Hull and G. Grefenstette. Querying Across Languages: a Dictionary-based Approach to Multilingual Information Retrieval. In *19th Ann Int ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'96)*, pages 49–57, 1996.