Improved Cross-Language Retrieval using Backoff Translation

Philip Resnik,[1,2] Douglas Oard,[2,3] and Gina Levow[2]
Department of Linguistics,[1]
Institute for Advanced Computer Studies,[2]
College of Information Studies,[3]
University of Maryland
College Park, MD 20742
{resnik,gina}@umiacs.umd.edu, oard@glue.umd.edu

ABSTRACT

The limited coverage of available translation lexicons can pose a serious challenge in some cross-language information retrieval applications. We present two techniques for combining evidence from dictionary-based and corpus-based translation lexicons, and show that backoff translation outperforms a technique based on merging lexicons.

1. INTRODUCTION

The effectiveness of a broad class of cross-language information retrieval (CLIR) techniques that are based on term-by-term translation depends on the coverage and accuracy of the available translation lexicon(s). Two types of translation lexicons are commonly used, one based on translation knowledge extracted from bilingual dictionaries [1] and the other based on translation knowledge extracted from bilingual corpora [8]. Dictionaries provide reliable evidence, but often lack translation preference information. Corpora, by contrast, are often a better source for translations of slang or newly coined terms, but the statistical analysis through which the translations are extracted sometimes produces erroneous results. In this paper we explore the question of how best to combine evidence from these two sources.

2. TRANSLATION LEXICONS

Our term-by-term translation technique (described below) requires a translation lexicon (henceforth *tralex*) in which each word f is associated with a ranked set $\{e_1, e_2, \ldots e_n\}$ of translations. We used two translation lexicons in our experiments.

2.1 WebDict Tralex

We downloaded a freely available, manually constructed English-French term list from the Web[1] and inverted it to French-English format. Since the WebDict translations appear in no particular order, we ranked the e_i based on target language unigram statistics calculated over a large comparable corpus, the English portion of the Cross-Language Evaluation Forum (CLEF) collection, smoothed with statistics from the Brown corpus, a balanced corpus covering many genres of English. All single-word translations are ordered by decreasing unigram frequency, followed by all multi-word translations, and finally by any single-word entries not found in either corpus. This ordering has the effect of minimizing the effect of infrequent words in non-standard usages or of misspellings that sometimes appear in bilingual term lists.

2.2 STRAND Tralex

Our second lexical resource is a translation lexicon obtained fully automatically via analysis of parallel French-English documents from the Web. A collection of 3,378 document pairs was obtained using STRAND, our technique for mining the Web for bilingual text [7]. These document pairs were aligned internally, using their HTML markup, to produce 63,094 aligned text "chunks" ranging in length from 2 to 30 words, ∼8 words on average per chunk, for a total of ∼500K words per side. Viterbi word-alignments for these paired chunks were obtained using the GIZA implementation of the IBM statistical translation models.[2] An ordered set of translation pairs was obtained by treating each alignment link between words as a co-occurrence and scoring each word pair according to the likelihood ratio [2]. We then rank the translation alternatives in order of decreasing likelihood ratio score.

3. CLIR EXPERIMENTS

Ranked tralexes are particularly well suited to a simple ranked term-by-term translation approach. In our experiments, we use top-2 balanced document translation, in which we produce exactly two English terms for each French term. For terms with no known translation, the untranslated French term is generated twice (often appropriate for proper names). For French terms with one translation, that translation is generated twice. For French terms with two or more translations, we generate the first two translations in the tralex. Thus balanced translation has the effect of introducing a uniform weighting over the top n translations for each term (here $n = 2$).

Benefits of the approach include simplicity and modularity — notice that a lexicon containing ranked translations is the only requirement, and in particular that there is no need for access to the internals of the IR system or to the document collection in order to

[1] http://www.freedict.com

[2] http://www.clsp.jhu.edu/ws99/projects/mt/

.

perform computations on term frequencies or weights. In addition, the approach is an effective one: in previous experiments we have found that this balanced translation strategy significantly outperforms the usual (unbalanced) technique of including all known translations [3]. We have also investigated the relationship between balanced translation and Pirkola's structured query formulation method [6].

For our experiments we used the CLEF-2000 French document collection (approximately 21 million words from articles in *Le Monde*). Differences in use of diacritics, case, and punctuation can inhibit matching between tralex entries and document terms, so we normalize the tralex and the documents by converting characters to lowercase and removing all diacritic marks and punctuation. We then translate the documents using the process described above, index the translated documents with the Inquery information retrieval system, and perform retrieval using "long" queries formulated by grouping all terms in the title, narrative, and description fields of each English topic description using Inquery's #sum operator. We report mean average precision on the 34 topics for which relevant French documents exist, based on the relevance judgments provided by CLEF. We evaluated several strategies for using the WebDict and STRAND tralexes.

3.1 WebDict Tralex

Since a tralex may contain an eclectic mix of root forms and morphological variants, we use a four-stage backoff strategy to maximize coverage while limiting spurious translations:

1. Match the **surface form** of a document term to **surface forms** of French terms in the tralex.

2. Match the **stem** of a document term to **surface forms** of French terms in the tralex.

3. Match the **surface form** of a document term to **stems** of French terms in the tralex.

4. Match the **stem** of a document term to **stems** of French terms in the tralex.

We used unsupervised induction of stemming rules based on the French collection to build the stemmer [5]. The process terminates as soon as a match is found at any stage, and the known translations for that match are generated. The process may produce an inappropriate morphological variant for a correct English translation, so we used Inquery's English kstem stemmer at indexing time to minimize the effect of that factor on retrieval effectiveness.

3.2 STRAND Tralex

One limitation of a statistically derived tralex is that any term has *some* probability of aligning with any other term. Merely sorting translation alternatives in order of decreasing likelihood ratio will thus find *some* translation alternatives for every French term that appeared at least once in the set of parallel Web pages. In order to limit the introduction of spurious translations, we included only translation pairs with at least N co-occurrences in the set used to build the tralex. We performed runs with $N = 1, 2, 3$, using the four-stage backoff strategy described above.

3.3 WebDict Merging using STRAND

When two sources of evidence with different characteristics are available, a combination-of-evidence strategy can sometimes outperform either source alone. Our initial experiments indicated that the WebDict tralex was the better of the two (see below), so we adopted a reranking strategy in which the WebDict tralex was refined according a voting strategy to which both the original WebDict and STRAND tralex rankings contributed.

Condition	MAP
STRAND ($N = 1$)	0.2320
STRAND ($N = 2$)	0.2440
STRAND ($N = 3$)	0.2499
Merging	0.2892
WebDict	0.2919
Backoff	0.3282

Table 1: Mean Average Precision (MAP), averaged over 34 topics

For each French term that appeared in both tralexes, we gave the top-ranked translation in each tralex a score of 100, the next a score of 99, and so on. We then summed the WebDict and STRAND scores for each translation, reranked the WebDict translations based on that sum, and then appended any STRAND-only translations for that French term. Thus, although both sources of evidence were weighted equally in the voting, STRAND-only evidence received lower precedence in the merged ranking. For French terms that appeared in only one tralex, we included those entries unchanged in the merged tralex. In this experiment run we used a threshold of $N = 1$, and applied the four-stage backoff strategy described above to the merged resource.

3.4 WebDict Backoff to STRAND

A possible weakness of our merging strategy is that inflected forms are more common in our STRAND tralex, while root forms are more common in our WebDict tralex. STRAND tralex entries that were copied unchanged into the merged tralex thus often matched in step **1** of the four-stage backoff strategy, preventing WebDict contributions from being used. With the WebDict tralex outperforming the STRAND tralex, this factor could hurt our results. As an alternative to merging, therefore, we also tried a simple backoff strategy in which we used the original WebDict tralex with the four-stage backoff strategy described above, to which we added a fifth stage in the event that fewer than two WebDict tralex matches were found:

5. Match the **surface form** of a document term to **surface forms** of French terms in the STRAND tralex.

We used a threshold of $N = 2$ for this experiment run.

4. RESULTS

Table 1 summarizes our results. Increasing thresholds seem to be helpful with the STRAND tralex, although the differences were not found to be statistically significant by a paired two-tailed t-test with $p < 0.05$. Merging the tralexes provided no improvement over using the WebDict tralex alone, but our backoff strategy produced a statistically significant 12% improvement in mean average precision (at $p < 0.01$) over the next best tralex (WebDict alone). As Figure 1 shows, the improvement is remarkably consistent, with only four of the 34 topics adversely affected and only one topic showing a substantial negative impact.

Breaking down the backoff results by stage (Table 2), we find that the majority of query-to-document hits are obtained in the first stage, i.e. matches of the term's surface form in the document to a translation of the surface form in the dictionary. However, the backoff process improves by-token coverage of terms in documents by 8%, and gives a 3% relative improvement in retrieval results; it also contributed additional translations to the top-2 set in approximately 30% of the cases, leading to the statistically significant 12% relative improvement in mean average precision as compared to the baseline using WebDict alone with 4-stage backoff.

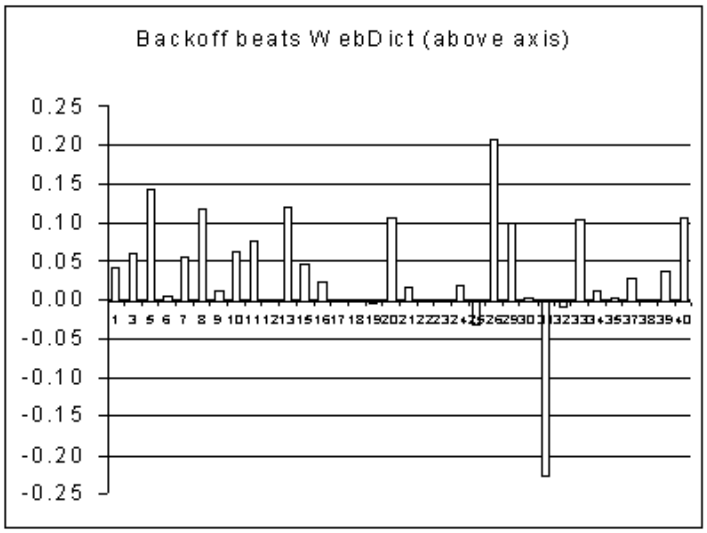

Figure 1: WebDict-to-tralex backoff vs. WebDict alone, by query

Stage (forms)	Lexicon matches
1 (surface-surface)	70.38%
2 (stem-surface)	3.18%
3 (surface-stem)	0.46%
4 (stem-stem)	0.98%
5 (STRAND)	8.34%
No match found	16.66%

Table 2: Term matches in 5-stage backoff

5. CONCLUSIONS

There are many ways of combining evidence from multiple translation lexicons. We use tralexes similar to those used by Nie et al. [4], but our work differs in our use of balanced translation and a backoff translation strategy (which produces a stronger baseline for our WebDict tralex), and in our comparison of merging and backoff translation strategies for combining resources. In future work we plan to explore other combinations of merging and backoff and other merging strategies, including post-retrieval merging of the ranked lists.

In addition, parallel corpora can be exploited for more than just the extraction of a non-contextualized translation lexicon. We are currently engaged in work on lexical selection methods that take advantage of contextual information, in the context of our research on machine translation, and we expect that CLIR results will be improved by contextually-informed scoring of term translations.

6. ACKNOWLEDGMENTS

This research was supported in part by Department of Defense contract MDA90496C1250 and TIDES DARPA/ITO Cooperative Agreement N660010028910,

7. REFERENCES

[1] L. Ballesteros and W. B. Croft. Resolving ambiguity for cross-language retrieval. In W. B. Croft, A. Moffat, and C. V. Rijsbergen, editors, *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 64–71. ACM Press, Aug. 1998.

[2] T. Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74, March 1993.

[3] G.-A. Levow and D. W. Oard. Translingual topic tracking with PRISE. In *Working Notes of the Third Topic Detection and Tracking Workshop*, Feb. 2000.

[4] J.-Y. Nie, M. Simard, P. Isabelle, and R. Durand. Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the web. In M. Hearst, F. Gey, and R. Tong, editors, *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 74–81, Aug. 1999.

[5] D. W. Oard, G.-A. Levow, and C. I. Cabezas. CLEF experiments at Maryland: Statistical stemming and backoff translation. In C. Peters, editor, *Proceedings of the First Cross-Language Evaluation Forum*. 2001. To appear. http://www.glue.umd.edu/∼oard/research.html.

[6] D. W. Oard and J. Wang. NTCIR-2 ECIR experiments at Maryland: Comparing structured queries and balanced translation. In *Second National Institute of Informatics (NII) Test Collection Information Retrieval (NTCIR) workshop*. forthcoming.

[7] P. Resnik. Mining the Web for bilingual text. In *37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, College Park, Maryland, June 1999.

[8] P. Sheridan and J. P. Ballerini. Experiments in multilingual information retrieval using the SPIDER system. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Aug. 1996.