Rapidly Retargetable Interactive Translingual Retrieval

Gina-Anne Levow
Institute for Advanced
Computer Studies
University of Maryland,
College Park, MD 20742
gina@umiacs.umd.edu

Douglas W. Oard
College of Information Studies
Institute for Advanced
Computer Studies
University of Maryland,
College Park, MD 20742
oard@glue.umd.edu

Philip Resnik
Department of Linguistics
Institute for Advanced
Computer Studies
University of Maryland,
College Park, MD 20742
resnik@umiacs.umd.edu

ABSTRACT

This paper describes a system for rapidly retargetable interactive translingual retrieval. Basic functionality can be achieved for a new document language in a single day, and further improvements require only a relatively modest additional investment. We applied the techniques first to search Chinese collections using English queries, and have successfully added French, German, and Italian document collections. We achieve this capability through separation of language-dependent and language-independent components and through the application of asymmetric techniques that leverage an extensive English retrieval infrastructure.

Keywords

Cross-language information retrieval

1. INTRODUCTION

Our goal is to produce systems that allow interactive users to present English queries and retrieve documents in languages that they cannot read. In this paper we focus on what we call "rapid retargetability": extending interactive translingual retrieval functionality for a new document language rapidly with few language-specific resources. Our current system can be retargeted to a new language in one day with only one language-dependent resource: a bilingual term list.[1] Our language-independent architecture consists of two main components:

1. Document translation and indexing
2. Interactive retrieval

We describe each of these components, demonstrate their effectiveness for information retrieval tasks, and then conclude by describing our experience with adding French, German and Italian document collections to a system that was originally developed for Chinese.

2. DOCUMENT TRANSLATION AND INDEXING

We have adopted a document translation architecture for two reasons. First, we support a single query language (English) but multiple document languages, so indexing English terms simplifies query processing (where interactive response time can be a concern). Second, a document translation architecture simplifies the display of translated documents by decoupling the translation and display processes. Gigabyte collections require machine translation that is orders of magnitude faster than present commercial systems. We accomplish this using term-by-term translation, in which the basic data structure is a simple hash table lookup. Any translation requires some source of translation knowledge—we use a bilingual term list containing English translation(s) for each foreign language term. We typically construct these term lists by harvesting Internet-available translation resources, so the foreign language terms for which translations are known are typically an eclectic mix of root and inflected forms. We accommodate this limitation using a four-stage backoff statistical stemming approach to enhance translation coverage.

2.1 Preprocessing.

Differences in use of diacritic-s, case, and punctuation can inhibit matching between term list entries and document terms, so normalization is important. In order to maximize the probability of matching document words with term list entries, we normalize the bilingual term list and the documents by:

- converting characters in Western languages to lowercase,
- removing all accents and diacritics, and
- segmentation, which for Western languages merely involves separating punctuation from other text by the addition of white space.

Our preprocessing also includes conversion of the bilingual term list and the document collection into standard formats. The preprocessing typically requires about half a day of programmer time.

2.2 Four-Stage Backoff Translation.

Bilingual term lists found on the Web often contain an eclectic mix of root forms and morphological variants. We thus developed 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 source language terms in the bilingual term list.
2. Match the **stem** of a document term to **surface forms** of source language terms in the bilingual term list.

[1] For Asian languages we also use a language-specific segmentation system.

.

3. Match the **surface form** of a document term to **stems** of source language terms in the bilingual term list.
4. Match the **stem** of a document term to **stems** of source language terms in the bilingual term list.

The process terminates as soon as a match is found at any stage, and the known translations for that match are generated. Although this may produce an inappropriate morphological variant for a correct English translation, use of English stemming at indexing time minimizes the effect of that factor on retrieval effectiveness. Because we are ultimately interested in processing documents in any language, we may not have a hand-crafted stemmer available for the document language. We have thus explored the application of rule induction to learn stemming rules in an unsupervised fashion from the collection that is being indexed [2].

2.3 Balanced Top-2 Translation.

We produce exactly two English terms for each foreign-language term. For terms with no known translation, the untranslated term is generated twice (often appropriate for proper names in the Latin-1 character set). For terms with one translation, that translation is generated twice. For terms with two or more known translations, we generate the "best" two translations. In prior experiments we have found that this balanced translation strategy significantly outperforms the usual (unbalanced) technique of including all known translations [1]. We establish the "best" translations by sorting the bilingual term list in advance using only English resources. All single-word translations are ordered by decreasing unigram frequency in the Brown corpus, followed by all multi-word translations, and finally by any single word entries not found in the Brown 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. This translation strategy allows balancing of translations in a modular fashion, even when one does not have access to the internal parameters of the information retrieval system. We translate ~ 100 MB per hour using Perl on a SPARC Ultra 5.

2.4 Post-translation Document Expansion.

We implement post-translation document expansion for the foreign language stories after translation into English in order to enrich the indexing vocabulary beyond that which was available after term-by-term translation. This is analogous to the process that Singhal et al. applied to monolingual speech retrieval [4].

Term-by-term translation produces a set of English terms that serve as a noisy representation of the original source language document. These terms are then treated as a query to a comparable English collection, typically contemporaneous newswire text, from which we retrieve the five highest ranked documents. From those five documents, we extract the most selective terms and use them to enrich the original translations of the documents. For this expansion process we select one instance of every term with an IDF value above an *ad hoc* threshold that was tuned to yield approximately 50 new terms. This optional step is the slowest processing stage, with a throughput of about 20 MB per hour.

2.5 Indexing

The resulting collection is then indexed using Inquery (version 3.1p1), with the kstem stemmer and default English stopword list. Indexing is the fastest stage in the process, with throughput exceeding one gigabyte per hour.

3. INTERACTIVE RETRIEVAL

Interactive searches are performed using a Web interface. Summary information for the top-ranked documents is displayed in groups of ten per page. Document summaries consist of the date and a gloss translation of the document title. Users can inspect a gloss translation of the full text of any document if the title is not sufficiently informative. For both title and full text, the gloss translations are generated in advance using the same process as translation for indexing, with the following differences in detail:

- Terms added as a result of document expansion are not displayed.
- The number of retained translations is separately selectable for the title and for full text indexing.
- Translations are not duplicated when fewer than the maximum allowable number of translations are known.

Our goal is to support the process of *finding* documents, with the realization that the process of *using* documents may need to be supported in some other way (e.g., by forwarding relevant documents to someone who is able to read that language). We have therefore designed our interface to highlight the query terms in translated documents and to facilitate skimming by emphasizing the most common translation when multiple translations are displayed. We have found that such displays can support a classification task, even when the translation is not easy to read [3]. Documents must be classified by the user as relevant or not relevant, so our classification results suggest that this can be an effective user interface design.

4. RESULTS

We present results both for component-level performance of our language-independent retargeting modules and an assessment of the overall retargeting process.

4.1 Component-level Evaluation

We applied our retargeting approach and retrieval enhancement techniques described above in the context of the first Cross-Language Evaluation Forum's (CLEF) multilingual task. We used the English language forms of the queries to retrieve English, French, German, and Italian documents. Below we present comparative performance measures for two of the main processing components described above - statistical stemming backoff translation - applied to the English-French cross-language segment of the CLEF task. The post-translation document expansion component was applied to the smaller Topic Detection and Tracking (TDT-3) collection to improve retrieval of Mandarin documents using English.

4.1.1 Baseline CLEF System Configuration

Our baseline run was conducted as follows. We translated the $\sim 44,000$ documents from the 1994 issues of *Le Monde*. We used the English-French bilingual term list downloaded from the Web at `http://www.freedict.com`. We then inverted the term list to form a 35,000 term French-English translation resource. We performed the necessary document and term list normalization; in this case, removing accents from document surface forms to enable matching with the un-accented term list entries, converting case, and splitting clitic contractions, such as *l'horlage*, on punctuation. We trained the statistical stemming rules on a sample of the bilingual term list and document collection and applied these rules in stemming backoff. Our default condition was run with top-2 balanced translation using the Brown corpus as a source of target language unigram frequency information. Translated documents were then indexed with

	Stage 1	Stage 2	Stage 3	Stage 4
Match	70%	3%	0.5%	1%

Table 1: Percentage of document terms translated at each stage of 4-stage backoff translation with statistical stemming.

the InQuery (version 3.1p1) system, using the kstem stemmer for English stemming and InQuery's default English stopword list. Long queries were formed by concatenating the title, description, and narrative fields of the original query specification. The resulting word sequence was enclosed in an InQuery $\#sum$ operator, indicating unweighted sum.

Our figure of merit for the evaluations below is mean (uninterpolated) average precision computed using trec_eval [2] across the 34 topics in the CLEF evaluation for which relevant French documents are known.

4.1.2 Backoff Translation with Statistical Stemming

We first contrast the above baseline system with the effectiveness of an otherwise identical run *without* the stemming backoff component. Terms in the documents are thus only translated if there is an exact match between the surface form in the document and a surface form in the bilingual term list. We find that mean average precision for unstemmed translation is 0.19 as compared with 0.2919 for our baseline system including stemming backoff based on trained rules. This difference is significant at $p < 0.05$, by paired t-test, two-tailed. The per-query effectiveness is illustrated in Figure 1. Backoff translation improves translation coverage while retaining relatively high precision of matching in contrast to unstemmed effectiveness.

Backoff translation improves cross-language information retrieval effectiveness by improving translation coverage of the terms in the document collection. Using the statistical stemmer, by-token coverage of document terms increased by 7coverage. The different stages of the four-stage backoff process contributed as illustrated in 1. The majority of terms match in the Stage 1 exact match, accounting for 70% of the term instances in the documents. The remaining stages each account for between 0.5% and 3% of the document terms, while 20% of document term instances remain untranslatable. However, this relatively small increase in coverage results in the highly significant improvement in retrieval effectiveness above.

4.1.3 Top-2 Balanced Translation

Here we contrast top-2 balanced translation with top-1 translation. We retain statistical stemming backoff for the top-1 translation. We replace each French document term with the highest ranked English translation by target language unigram frequency in the Brown Corpus as detailed above, retaining the original French term when no translation is found in the bilingual term list. We achieve a mean average precision of 0.2532 in contrast with the baseline condition. This difference is significant at $p < 0.01$ by paired t-test, two-tailed. We can effectively incorporate additional translations using top-2 balanced translation without degrading performance by introducing significant additional noise. A query-by-query contrast is presented in Figure 2.

4.1.4 Document Expansion

We evaluated post-translation document expansion using the Topic Detection and Tracking (TDT-3) collection. For this evaluation, we used the TDT-1999 topic detection task evaluation framework, but because out focus in this paper is on ranked retrieval effectiveness we report mean uninterpolated average precision rather than the topic-weighted detection cost measure typically reported in TDT. In the topic detection task, the system is presented with one or more exemplar stories from the training epoch—a form of query-by-example—and must determine whether each story in the evaluation epoch addresses either the same seminal event or activity or some directly related event or activity. This is generally thought to be a somewhat narrower formulation than the more widely used notion of topical relevance, but it seems to be well suited to query-by-example evaluations. The TDT-1999 tracking task was multilingual, searching stories in both English and Mandarin Chinese, and multi-modal, involving both newswire text and broadcast news audio. We focus on the cross-language spoken document retrieval component of the tracking task, using English exemplars to identify on-topic stories in Mandarin Chinese broadcast news audio. We compare top-1 translation of the Mandarin Chinese stories with and without post-translation document expansion.[3] We used the earlier TDT-2 English newswire text collection as our side collection for expansion. We perform topic tracking on 60 topics with 4 exemplars each. Here, we report the mean average precision on the 55 topics for which there are on-topic Mandarin audio stories. The mean uninterpolated average precision for retrieval of unexpanded documents is 0.36 while post-translation document expansion raises this figure to 0.41. This difference is significant at $p < 0.01$ by paired t-test, two-tailed. The contrast is illustrated in Figure 3. Interestingly, when we tried this with French, we noted that expansion tended to select terms from the few foreign-language documents that happened to be present in our expansion collection. We have not yet explored that effect in detail, but this observation suggests that the document expansion may be sensitive to the characteristics of the expansion collection that are not immediately apparent.

4.2 The Learning Curve

We have found that retargeting can be accomplished quite quickly (a day without document expansion, three days for TREC-sized collections with document expansion), but only if the required infrastructure is in place. Adapting a system that was developed initially for Chinese to handle French documents required several weeks, with most of that effort invested in development of four-stage backoff translation and statistical stemming. Further adapting the system to handle German documents revealed the importance of compound splitting, a problem that we will ultimately need to address by incorporating a more general segmentation strategy than we used initially for Chinese. In extending the system to Italian we have found that although our statistical stemmer presently performs poorly in that language, we can achieve quite credible results even with a fairly small (17,313 term) bilingual term list using a freely available Muscat stemmer (which exist for ten languages). So although it is possible in concept to retarget to a new language in just a few days, extending the system typically takes us between one and three weeks because we are still climbing the learning curve.

5. CONCLUSION

By building on the lessons learned using the TREC, CLEF, NTCIR, and TDT collections, we have sought to build an infrastructure that can be applied to a broad array of languages. Arabic and Korean collections are expected to become available in the next year, and we are now evolving our interface to support user studies. Our approach is distinguished by support for interactive retrieval even

[2] Available at ftp://ftp.cs.cornell.edu/pub/smart/.

[3] Since Mandarin Chinese has little surface morphology, we omit backoff translation in this case.

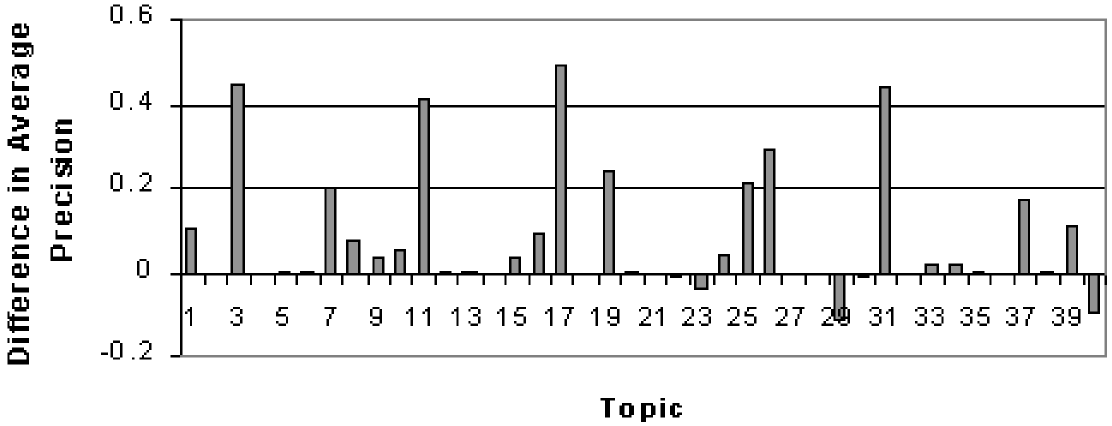

Figure 1: Comparison of effectiveness of backoff versus unstemmed translation of French documents: Bars above x-axis indicate backoff transltion outperforms unstemmed translation.

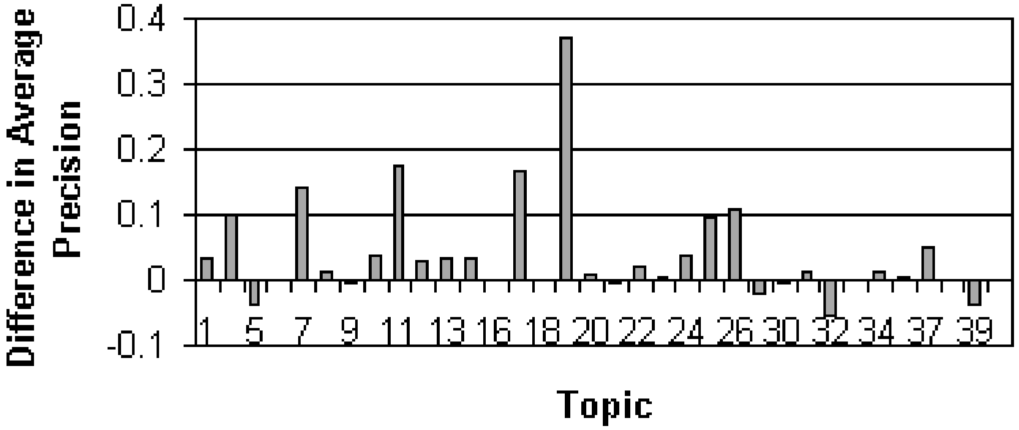

Figure 2: Comparison of effectiveness of top-2 balanced versus top-1 translation of French documents: Bars above x-axis indicate "Top-2" outperforms "Top-1"

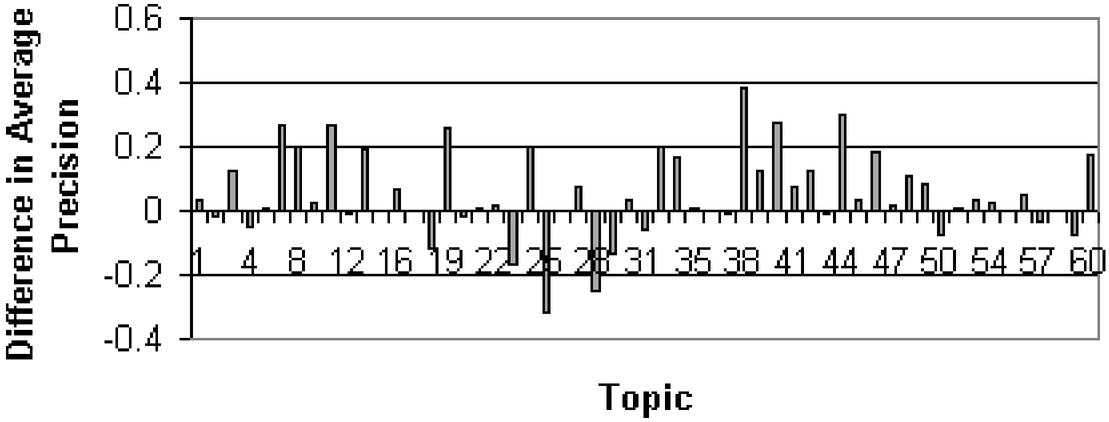

Figure 3: Comparison of effectiveness of top-1 post-translation document expansion versus bare top-1 translation of Chinese documents: Bars above x-axis indicate document expansion outperforms bare translation

in languages for which machine translation is presently unavailable, and our ultimate goal is to characterize how closely we can approximate the retrieval effectiveness users would obtain if they had the best available machine translations for the retrieved documents.

Acknowledgements

This work was supported in part by DARPA contract N6600197C8540 and DARPA cooperative agreement N660010028910.

6. ADDITIONAL AUTHORS

Clara I. Cabezas (Department of Linguistics, top University of Maryland, College Park, email: `clarac@umiacs.umd.edu`)

7. REFERENCES

[1] G.-A. Levow and D. W. Oard. Translingual topic tracking with PRISE. In *Working Notes of the Third Topic Detection and Tracking Worksho p*, Feb. 2000. http://www.glue.umd.edu/∼oard/research.html.

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

[3] D. W. Oard and P. Resnik. Support for interactive document selection in cross-language information retrieval. *Information Processing and Management*, 35(3):363–379, July 1999.

[4] A. Singhal, J. Choi, D. Hindle, J. Hirschberg, F. Pereira, and S. Whittaker. AT&T at TREC-7 SDR Track. In *Proceedings of the DARPA Broadcast News Workshop*, 1999.