Desparately Seeking Cebuano

Douglas W. Oard, David Doermann, Bonnie Dorr, Daqing He, Philip Resnik, and Amy Weinberg
UMIACS, University of Maryland, College Park, MD, 20642 USA
(oard,doermann,bonnie,resnik,weinberg)@umiacs.umd.edu
William Byrne, Sanjeev Khudanpur and David Yarowsky
CLSP, Johns Hopkins University, 3400 North Charles Street, Barton Hall, Baltimore, MD 21218
(byrne,khudanpur,yarowsky)@jhu.edu
Anton Leuski
USC Information Sciences Institute, 4676 Admiralty Way, Marina Del Rey, CA 90292
(byrne,khudanpur,yarowsky)@jhu.edu

Abstract

At 4:13 A.M. Eastern Standard Time on Wednesday March 5, Cebuano was designated as the language for the TIDES surprise language dry run. This paper reports the results of the first 60 hours of our data collection and implementation effort.

1 Introduction

The Los Angeles Times reported that at about 5:20 P.M. on Tuesday March 4, 2003, a bomb concealed in a backpack exploded at the airport in Davao City, the second largest city in the Philippines. At least 23 people were reported dead, with more than 140 injured, and President Arroyo of the Philippines characterized the blast as a terrorist act (**?**). With the 13 hour time difference, it was then at 4:20 A.M on the same date in Washington, DC. Twenty-four hours later, at 4:13 A.M. on March 5, participants in the Translingual Information Detection, Extraction and Summarization (TIDES) program were notified that Cebuano had been chosen as the language of interest for a "surprise language" practice exercise that had been planned quite independently to begin on that date. The notification observed that Cebuano is spoken by 24and that it is the *lingua franca* in the south Philippines, where the event occurred.

One goal of the TIDES program is to develop the ability to rapidly deploy a broad array of language technologies for previously unforeseen languages in response to unexpected events. That capability will be formally exercised for the first time during June 2003, in a month-long "Surprise Language Experiment." To prepare for that event, the Linguistic Data Consortium (LDC) organized a "dry run" for March 5-14 in order to refine their procedures for rapidly developing language resources of the type that the TIDES community will need during the July evaluation.

Development of interactive Cross-Language Information Retrieval (CLIR) systems that can be rapidly adapted to accommodate new languages has been the focus of extensive collaboration between the University of Maryland and The Johns Hopkins University. The capability for rapid development of necessary language resources is an essential part of that process, so we had been planning to participate in the surprise language dry run to refine our procedures for sharing those resources with other members of the TIDES community. Naturally, we chose CLIR as a driving application to focus our effort. Our goal, therefore, was to build an interactive system that would allow a searcher to pose English queries to find relevant Cebuano news articles from the period immediately following the bombing.

This paper describes the first 60 hours of our data collection and implementation efforts. The next section identifies the critical language resources needed for this application and describe our process for assembling and assessing those resources. Section **??** describe the design of our CLIR system and explain the process that we used to adapt that system to Cebuano. Section **??** presents the results of an initial usability study to explore the utility of our CLIR system to searchers with and without Cebuano language skills. Finally, the paper concludes with a brief discussion of our plans for further work with Cebuano and a recounting of some of the lessons that we have already learned.

2 Obtaining Language Resources

Our basic approach to development of an agile system for interactive CLIR relies on three strategies: (1) create an infrastructure in advance for English as a query language that makes only minimal assumptions about the document language; (2) leverage the asymmetry inherent in the problem by assembling strong resources for English in advance; and (3) develop a robust suite of capabilities to exploit any language resources that can be found for

the "surprise language." We defer the first two topics to the next section, and focus here on the third.

We know of five possible sources of translation expertise:

Informants. People who know the language are an excellent source of insight, and universities are an excellent place to find people that know a wide array of languages. We were able to locate an informant within 50 feet of our office, and to schedule an interview within 36 hours of the announcement of the language.

Academic literature. Major research universities are also an excellent place to find written materials describing a broad array of languages. Within 12 hours of the announcement, reference librarians at the University of Maryland had identified a textbook on "Beginning Cebuano," and we had located a copy at the University of Southern California. Together with the excellent electronic resources located by the LDC, this allowed us to begin development of a rudimentary stemmer.

Translation lexicons. Simple bilingual term lists are available for many language pairs. Using links provided by the LDC and our own Web searches, we were able to construct an English-Cebuano term list with over 14,000 translation pairs within 12 hours of the announcement. This largely duplicated a simultaneous effort at the LDC, and we later merged our term list with theirs.

Parallel text. Translation-equivalent documents, when aligned at the word level, provide an excellent source of information about not just possible translations, but their relative predominance. Within 24 hours of the announcement we had aligned Cebuano and English versions of the Holy Bible at the word level using Giza++. The Bible's vocabulary covers only about half of the words found in typical English news text (counted by-token), so it is useful to have additional sources of parallel text. For this reason, we have extended the previously developed STRAND system to locate likely translations in the Internet Archive, the largest collection of Web documents that is presently available for research use. The first results from that process were not yet available 60 hours after the announcement when this paper was submitted.

Printed Dictionaries. People learning a new language make extensive use of bilingual dictionaries, so we have developed a system that mimics that process to some extent. Within 18 hours of the announcement we had zoned page images from a previously scanned Cebuano-English dictionary to identify each dictionary entry, performed optical character recognition, and parsed the entries to construct a bilingual term list. We were aided in this process by the fact that Cebuano is written in a Roman script, but our initial results were adversely affected by poor image quality. During the remaining 42 hours before submission of this paper, we obtained four additional printed dictionaries, broke their bindings and scanned them, and then performed entry zoning, OCR and entry parsing.

As this description illustrates, these five sources provide complementary information. Since there is some uncertainty at the outset about how long it will be before each delivers useful results, we chose a strategy based on concurrency, balancing our investment over each the five sources. This allowed us to use whatever resources became available first to get an initial system running, with refinements subsequently being made as additional resources became available. Because Cebuano and English are written in the same script, we did not need character set conversion or phonetic cognate matching in this case. The interactive CLIR system described in the next section was therefore constructed using only English resources that were (or could have been) pre-assembled, a Cebuano-English bilingual term list, the rule-based stemmer that we constructed based on the academic literature and our discussion with our informant, and the Cebuano Bible.

3 Building a Cross-Language Retrieval System

Ideally, we would like to build a system that would find whatever documents the searcher would wish to read in a fully automatic mode. In practice, fully automatic search systems are imperfect even in monolingual applications. We therefore designed an interactive approach that functions something like a typical Web search engine: (1) the searcher poses their query in English, (2) the system ranks the Cebuano documents in decreasing order of likely relevance to the query, (3) the searcher examines a list of document titles in something approximating English, and (4) the searcher may optionally examine the full text of any document in something approximating English. The intent is to support an iterative process in which searchers learn to better express their query through experience. We are only able to provide very rough translations, so we expect that such a system would be used in an environment where searchers could send documents that appear promising off for professional translation when necessary.

At the core of our system is the capability to automatically rank Cebuano documents based on an English

query. We chose a query translation architecture using backoff translation (Resnik et al., 2001) and Pirkola's structured query method (Pirkola, 1998), implemented using Inquery version 3.1p1. The key idea in backoff translation is to first try to find consecutive sequences of query words on the English side of the bilingual term list, where that fails to try to find the surface form of each remaining English term, to fall back to stem matching when necessary, and ultimately to fall back to retaining the English term unchanged in the hope that it might be a proper name or some other form of cognate with Cebuano. Accents are stripped from the documents and all language resources to facilitate matching at that final step. The key idea behind Pirkola's structured query method is to compute term weights in the query language (rather than in the document language) by separately estimating the term frequency and document frequency statistics for each query term based on that query term's set of known translation alternatives from the bilingual term list. Our present system does not employ blind relevance feedback, which is known to significantly improve cross-language search performance, but potentially at the cost of less explainability, and hence less controllability, in interactive applications. Modern Web search engines typically omit this feature for a similar reason.

Although we have chosen techniques that are relatively robust and therefore require relatively little domain-specific tuning, stemmer design is an area of uncertainty that could adversely affect retrieval effectiveness. We therefore needed a test collection on which we could try out variants of the Cebuano stemmer. We built this test collection using 34,000 Cebuano Bible verses and 600 English questions that we found on the Web for which appropriate Bible verses were known. Each question was posed as a query using the batch mode of Inquery, and the rank of the known relevant verse was taken as a measure of effectiveness. We took the mean reciprocal rank (the inverse of the harmonic mean) as a figure of merit for each configuration, and used a Wilcoxon paired signed ranked test (with p¡0.05) to assess the statistical significance of observed differences. Mean reciprocal rank is often used as a measure of effectiveness when modeling known-item retrieval tasks, and it has been found to be useful for detecting poor system configurations. The measure does not typically distinguish well among fairly similar systems, however.

The other key capability that is needed is title and document translation. We accomplish this in the simplest way possible: we reverse the bilingual term list, and we reverse the role of Cebuano and English in the process described above for query translation. Our user interface is capable of displaying multiple translations for a single term (arranged horizontally for compact depiction or vertically for clearer depiction), but searchers can choose to display only the single best translation. When reliable translation probability statistics (from parallel text) are not available, we use the relative word unigram frequency of each translation of a Cebuano term in a representative English collection as a substitute for that probability.

Our query translation process can operate in a fully automatic mode, but in order to provide greater explainability (and thus improved controllability), we have also implemented an optional user-assisted query translation mode. When that mode is selected, the system displays each Cebuano translation for a query term and allows the searcher to deselect inappropriate translations. The meaning of each known translation is indicated in English by displaying either reverse translations (English words that share the same Cebuano translation) or an example of the usage of that translation (found in the term-aligned Bible) (**?**).

4 Usability Assessment

We would like to do a small usability study with one person that knows Cebuano but is not a search professional (our informant) and one professional searcher that does not (a reference librarian). But this paper is already too long, so you will have to come to the conference to hear the results!

5 Looking Ahead

Complete details will be available by the time of the conference.

Acknowledgment

The authors are grateful to Tim Hackman, Burcu Karagol-Ayan, Okan Kolak, Anton Leuski, Huanfeng Ma, Dan Melamed, Karen Patterson, Michael Subotin, Jianqiang Wang and the Linguistic Data Consortium for their assistance with this effort. This work has been supported in part by DARPA cooperative agreement N660010028910.

References

Ari Pirkola. 1998. The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 55–63, August.

Philip Resnik, Douglas Oard, and Gina Levow. 2001. Improved cross-language retrieval using backoff translation. In *First International Conference on Human Language Technologies*. http://www.glue.umd.edu/∼oard/research.html.