Desparately Seeking Cebuano

Douglas W. Oard, David Doermann, Bonnie Dorr, Daqing He, Philip Resnik, and Amy Weinberg
UMIACS, University of Maryland, College Park, MD, 20742
(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, Philipp Koehn and Kevin Knight
USC Information Sciences Institute, 4676 Admiralty Way, Marina Del Rey, CA 90292
(leuski,koehn,knight)@isi.edu

Abstract

This paper describes an effort to rapidly develop language resources and component technology to support searching Cebuano news stories using English queries. Results from the first 60 hours of the exercise are presented.

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 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 24% of the population of the Philippines, and 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 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, and more recently with the University of Southern California. 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 posing English queries to find relevant Cebuano news articles from the period immediately following the bombing.

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:

People. People who know the language are an excellent source of insight, and universities are an excellent place to find such people. We were able to locate a speaker of Cebuano within 50 feet of one of our offices, and to schedule an interview with a second Cebuano speaker within 36 hours of the announcement of the language.

Scholarly 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 develop a rudimentary stemmer within 36 hours.

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++. An evaluation by a native Cebuano speaker of a stratified random sample of 88 translation pairs showed remarkably high precision. On a 4-point scale with 1=correct and 4=incorrect the most frequent 100 words averaged 1.3, the next 400 most frequent terms averaged 1.6, and the 500 next most frequent terms after that averaged 1.7. 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 our previously developed STRAND system to locate likely translations in the Internet Archive. Those runs were not yet complete 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 12 hours of the announcement we had zoned page images from a Cebuano-English dictionary that was available commercially in Adobe Page Description Format (PDF) 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. Again, we achieved good precision, with a sampled word error rate for OCR of 6.9% and a precision for a random sample of translation pairs of 87%. Part of speech tags were also extracted, although they are not used in our process.

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 CLIR system described in the next section was therefore constructed using only English resources that were (or could have been) pre-assembled, plus a Cebuano-English bilingual term list, a rule-based stemmer, 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 have developed 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 and Pirkola's structured query method, 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.

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 50 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 paired two-tailed t-test (with $p<0.05$) to assess the statistical significance of observed differences. Our initial configuration, without stemming, obtained a mean inverse rank of 0.14, which is a statistically significant improvement over no translation at all (mean inverse rank 0.02 from felicitous cognate and loan word matches). The addition of Cebuano stemming resulted in a reduction in mean inverse rank to 0.09. Although the reduction is not statistically significant in that case, the result suggests that our initial stemmer is not yet useful for information retrieval tasks.

The other key capability that is needed is title and document translation. We can accomplish this in one of two ways. The simplest approach is to reverse the bilingual term list, and to 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 most likely 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. A more sophisticated way is to build a statistical machine translation system using parallel text. We built our first statistical machine translation system within 40 hours of the announcement, and one sentence of the resulting translation using each technique is shown below:

```
Cebuano: 'ang rebeldeng milf, kinsa
lakip sa nangamatay, nagdala og
backpack nga dunay explosives nga
niguba sa waiting lounge sa airport,
matod sa mga defense official.'

Term-by-term translation:
'(carelessness, circumference,
conveyence) rebeldeng milf, who lakip
(at in of) nangamatay, nagdala og
backpack nga valid explosives nga
niguba (at, in of) waiting lounge
(at, in, of) airport, matod (at, in,
of) mga defense official'

Statistical translation: 'who was
accused of rank, ehud og niguba
waiting lounge defense of those dumah
milf rebeldeng explosives backpack
airport matod official.'
```

At this point, term-by-term translation is clearly the better choice. But as more parallel text becomes available, we expect the situation to reverse. The LDC is preparing a set of human reference translations that will allow us to detect that changeover point automatically using the NIST variant of the BLEU measure for machine translation effectiveness.

4 Conclusion

The results reported in this paper were accomplished by a team of 20 people with expertise in various facets of te task that invested about 250 person-hours over two and a half days. As additional Cebuano-specific evaluation resources are developed, we expect to gain additional insight into the quality of these early resources. Moreover, once we see what works best for Cebuano by the end of the process, we plan to revisit our process design with an eye towards better optimizing our initial time investments. We expect to be able to address both of those points in detail by the time of the conference.

This exercise was originally envisioned as a dry run to work out the kinks in our process, and indeed we have already learned a lot on that score. First, we learned that our basic approach seems sound; we built the key components of an interactive CLIR system in about 40 hours, and by the 60-hour point we had some basis for believing that each of those components could at least minimally fulfill their role in a fully integrated system. Some of our time was, however, spent on things that could have been done in advance. Perhaps the most important of these was the development of an information retrieval test collection using the Bible. That job, and numerous smaller ones, are now done, so we expect that we will be able to obtain similar results with about half the effort next time around.

Acknowledgments

Thanks to Clara Cabezas, Tim Hackman, Margie Hinonangan, Burcu Karagol-Ayan, Okan Kolak, Huanfeng Ma, Grazia Russo-Lassner, Michael Subotin, Jianqiang Wang and the LDC! This work has been supported in part by DARPA contract N660010028910.