Rapid-Response Machine Translation for Unexpected Languages

Douglas W. Oard*
College of Information Studies and UMIACS
University of Maryland, College Park, MD 20742
Franz Josef Och
University of Southern California Information Sciences Institute
4676 Admiralty Way, Marina Del Rey, CA 90292
oard@umd.edu, och@isi.edu

Abstract

Statistical techniques for machine translation offer promise for rapid development in response to unexpected requirements, but realizing that potential requires rapid acquisition of required resources as well. This paper reports the results of experiments with resources collected in ten days; about 1.3 million words of parallel text from five types of sources and a bilingual term list with about 20,000 term pairs. Systems were trained with resources individually and in combination, using an approach based on alignment templates. The use of all available resources was found to yield the best results in an automatic evaluation using the BLEU measure, but a single resource (the Bible) coupled with a small amount of in-domain manual translation (less than 6,000 words) achieved more than 85% of that upper baseline. With a concerted effort, such a system could be built in a single day.

1 Introduction

In June of 2003, the U.S. Defense Advanced Research Projects Agency (DARPA) organized a "surprise language" evaluation to determine the extent to which language technologies being developed under the Translingual Information Detection, Extraction and Summarization (TIDES) program can be rapidly deployed. In preparation for that evaluation, the Linguistic Data Consortium (LDC) organized a ten-day "data collection dry run" in March 2003 to try out procedures for obtaining and/or creating the language resources that this community will need in June. The Philippine language Cebuano was chosen for the dry run, and eleven institutions contributed to the resulting data collection and construction effort over the next ten days (Oard, 2003).

Several teams used the resulting resources as a basis for constructing systems that worked with English and Cebuano. Dictionary-based Cross-Language Information Retrieval (CLIR) proved to be a tractable task, with batch experiments demonstrating respectable retrieval effectiveness after three days (Oard et al., 2003) and two fully integrated interactive CLIR systems available by the tenth day. Machine Translation (MT) proved to be a bigger challenge, however. The interactive CLIR systems were forced to rely on term-by-term gloss translation, because suitable statistical machine translation results were not available during the dry run. Our purpose in this paper is to explore the potential for building MT systems using the resources that were constructed, with an eye towards contributing MT results for use in integrated systems during the surprise language experiment in June.

Interactive CLIR systems that are designed for users that cannot read the language in which the documents are written rely on MT for three purposes:

- As a precursor for summarization, to support selection of documents for further examination based on examination of a list of summaries (e.g., headlines) that are rendered in a language that the searcher can read.
- Directly, to support relevance assessment based on the full text of translated documents that the user chooses to examine.
- Directly, to support the ultimate use of the content of documents that the searcher feels would have utility for their intended purpose.

In the next section, we briefly review prior

* This work was performed while at USC-ISI.

work on rapid development of machine translation systems. We then describe the resources that were produced during the surprise language dry run, the statistical machine translation system that we trained using those resources, and the evaluation measures that we used to assess the accuracy and fluency of the resulting translations. After discussion of the results that we obtained, we conclude the paper by identifying several potential directions for future work on this topic.

2 Prior Work

Perhaps the simplest approach to rapid development of translation capabilities is to perform term-by-term gloss translation using a bilingual term list. For example, Resnik and Oard used gloss translations to assess the ability of English speakers to manually categorize Japanese directory entries (Oard and Resnik, 1999). The first foray that we are aware of into rapid development of statistical machine translation systems was by a team at the 1999 Johns Hopkins Summer Workshop. They built a Chinese-to-English MT system in one day, although the parallel text collection that they used had been assembled over an extended period at considerable expense by the Linguistic Data Consortium (Al-Onizan et al., 1994). Germann was the first to try similar techniques with rapidly developed resources, building a Tamil-to-English MT system by manually translating 24,000 words of Tamil into English in a six week period (Germann, 2001).

All of this work was done before automatic evaluation of MT system output using measures such as BLEU (Papineni et al., 2001) became commonplace, so a wide variety of techniques were used to assess the results of those early efforts. Resnik and Oard used a classification consistency measure, the Hopkins workshop team assessed their output by inspection, and Germann performed a suite of task-based user studies. In each case, the MT results were found to be useful.

3 Cebuano Resources

The surprise language dry run resulted in production of six types of resources that that we used in our experiments:[1]

- **B:** 912,775 words of verse-aligned parallel text from the **Bible**, provided by the University of Maryland. Both the English and the Cebuano bibles were obtained from the Web in character-coded form. The English Bible was the World English Bible (WEB) version. Verse alignment was used in lieu of sentence alignment for this data.

- **E:** 214,327 words of parallel text from **examples** of usage that were automatically extracted from a printed bilingual dictionary after scanning and Optical Character Recognition (OCR) by the Johns Hopkins University. No postediting was performed on the OCR results.

- **M:** 23,761 words of parallel text that was **manually** created at the USC Information Sciences Institute by three native speakers of Cebuano who evidenced fluency in (spoken) English. Cebuano news and editorials were translated into English. The resulting translations were prepared rapidly (about 2 hours per page) by individuals that were not trained as professional translators, so they are often somewhat disfluent. The set of available translations was divided as follows: (1) a 4,220-word training set consisting of all the news translations (from two translators), for use in training MT systems, (2) a 5,895-word development set (from the same two translators), consisting entirely of editorials, used for parameter tuning, and (3) a 13,646-word evaluation set (from all three translators), consisting entirely of editorials, for use in computing BLEU scores. The contributions to the evaluation set were balanced across translators.

- **N:** 138,408 words of parallel text from *Ang Bayan*, the **newsletter** of the Philippine Communist Party. The text of each newsletter was extracted from PDF files by Carnegie-Mellon University and sentence-aligned aligned at the USC Information Sciences Institute.

[1] Word counts are based on a count of tokens on the English side of each corpus.

- **T:** A bilingual **term** list with 20,491 translation pairs that was produced by the LDC by merging all available bilingual terms lists from Web and commercial sources and removing duplicates. Translation pairs were used for training the statistical machine translation system as if they were were (very short!) aligned sentences. The resulting alignments are thus extremely good, but no useful information about translation probabilities could be discerned from this collection alone.
- **W:** 57,914 words of parallel text from **Web** pages provided in chunk-aligned form by the University of Maryland, subsequently aligned at sentence-level by New York University. These Web pages were approximately equally divided between four categories: cultural, evangelical, folk tales, and miscellaneous (which included documents about sports and math, documents from the United Nations, and a few manually rekeyed phrase lists provided by the LDC).

4 Translation Approach

Our statistical translation model is based on the alignment template approach (Och and Ney, 2002). In this translation model, a sentence is translated by segmenting the input sentence into phrases, translating these phrases, and reordering the translations in the target language. In addition to the feature functions described in (Och and Ney, 2002), our system includes a phrase penalty (the number of phrases used), multiple language models, and special alignment features. The different feature functions $h_m, m = 1, \ldots, M$, some of which are based on alignment templates, are combined in a log-linear form:

$$\begin{aligned} Pr(e_1^I|f_1^J) &= p_{\lambda_1^M}(e_1^I|f_1^J) \\ &= \frac{\exp[\sum_{m=1}^M \lambda_m h_m(e_1^I, f_1^J)]}{\sum_{e'^I_1} \exp[\sum_{m=1}^M \lambda_m h_m(e'^I_1, f_1^J)]} \end{aligned}$$

Because each of the feature functions that we use is an information content measure derived from a probabilistic model, these feature functions are more 'informative' than the binary feature functions used in standard maximum entropy models.

For search, we use a dynamic programming beam-search algorithm to explore a subset of all possible translations (Och et al., 1999) and extract n-best candidate translations using A* search (Ueffing et al., 2002). These n-best candidate translations are the basis for discriminative training of the model parameters with respect to translation quality (Och, 2003). Three English language models are interpolated (one from a large collection of news, one from the English Bible, and one from the English side of the training data) using linear interpolation constants learned on the development test collection of translated editorials.

5 Evaluation Measures

In recent years, various methods have been proposed to automatically evaluate machine translation quality by comparing hypothesis translations with reference translations which try to approximate human assessment and often achieve an astonishing degree of correlation to human subjective evaluation of fluency and adequacy (Doddington, 2002; Papineni et al., 2001).

Here, we use the BLEU score to assess the translation quality. This criterion computes the geometric mean of the precision of word n-grams of various lengths between a hypothesis and a set of reference translations multiplied by a penalty factor BP for short sentences:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} \frac{\log p_n}{N}\right)$$

Here p_n denotes the precision of n-grams in the hypothesis translation (i.e., the fraction of the word sequences that could be found in *some* reference translation). A BLEU score of 0.0 corresponds to system output that is very different from all of the reference translations, a score of 1.0 corresponds to system output that is identical to some patchwork combination of the reference translations. In this paper, we use a single reference translation. To avoid a bias towards the specific style of one translator, reference translations from multiple sources are needed. Three translators made balanced contributions to our test corpus. We also show 95% confidence intervals for each value, computed using bootstrap resampling (Press et al., 2002).

6 Results

Figure 1 shows the results of our experiments. English is the language of instruction in Philippine schools, so Cebuano documents often include some English loan words. Untranslated Cebuano documents therefore receive a nonzero BLEU score (4.6%). Among single sources, the Web collection (9.4%) and the Bible (9.0%) did the best. Combinations of sources generally outperformed single sources, with the best pairing being the bilingual term list and the Web collection (10.0%). Adding additional sources beyond two generally yielded small further improvements, with the best overall results (10.2%) coming from use of all available sources.

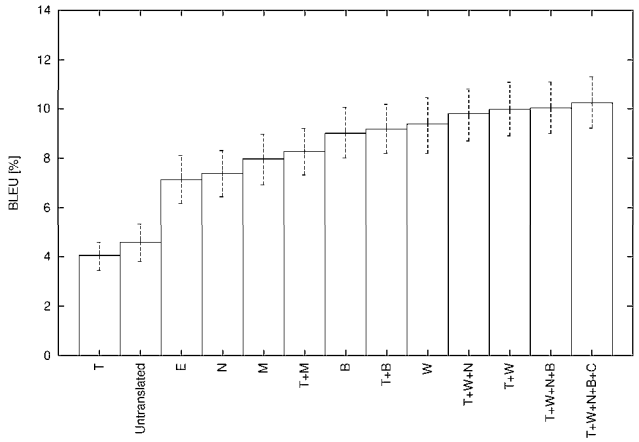

Figure 1: BLEU scores (%), by condition, with 95% confidence intervals.

The confidence intervals establish a partition on the individual sources; sources that are grouped yield results that are statistically indistinguishable (at 95% confidence), but across groups there is a clear preference order. **W** and **B** form the top-performing group, **M**, **N** and **E** the next, and **T** the worst. Indeed, the bilingual term list seems to be of no value on its own, but to cause no harm (and perhaps some small benefit) when used with any parallel text collection. This behavior is consistent with the inability of our technique to discern translation probabilities from a bilingual term list alone.

Examples of usage extracted from a printed dictionary seem to do relatively poorly, which may result from the presence of OCR errors and/or errors in the automatic extraction of those examples (effectively, a form of alignment error). The multi-source Web collection outperformed the larger newsletter collection that we had expected to be a good domain match for this task, so we might be tempted to speculate that diversity is more useful than size for this task. The Bible (our largest single collection) did relatively well, however, and our smallest collection (4,220 words of manually translated news) came out right in the middle. So our results do not point clearly to a size-diversity tradeoff. Perhaps future experiments with subsets of the Web collection might shed addition light on that question.

It is also worth mentioning that the accuracy of sentence alignment varies from one collection to another, with the best alignments most likely being for the Web collection, the Bible, and the manual translations. Interestingly, these are the three best single sources. Our assessment of alignment accuracy here is based on our understanding of the process by which the alignments were performed; we would want to actually measure alignment accuracy directly before placing too much stock in this intriguing observation.

Automatic measures of translation effectiveness are useful during system development because improvements in those measures have been shown to be correlated with human assessment of improved translation quality. There is, however, not presently any way to establish a target level for BLEU scores that would indicate utility for a specific purpose. An example of MT output drawn from the configuration with the best BLEU score is shown below:

```
question transparent is our government ?
of salem arellano , mindanao scoop , 17
november 2002 of so that day the seminar
that was held in america that from the
four big official of the seven the place
in mindanao run until is in davao . the
purpose of the seminar , added of members
orlando maglinao , is the resistance to
cause the corruption in the government is
be , ue , ue of our country .
```

For comparison, the human translation of the same passage is shown below:

```
question is our government transparent ?
by salem arellano , mindanao scoop , 17
november 2002 in the past days , there
```

```
was a seminar held in america participated
by the top four positions of the seven
places here in mindanao extended up to
davao . according to councilor orlando
maglinao , the purpose of the seminar was
fighting corruption , which has been a
great factor on what 's happening to our
government .
```

We have not conducted a task-based evaluation, but it appears to us that the MT output would often be adequate to at least judge the topical relevance of a document in an information retrieval context. The utility of these translations to support tasks that require a greater degree of understanding of nuance (e.g., report preparation) is open to serious question, however.

7 Conclusions and Future Work

To the best of our knowledge, this is the first time that anyone has tried such a diverse set of resources for a single language pair in a statistical MT framework. Although we must caveat our conclusions with the fact that they are based on only a single set of experiments, we are now in a position to offer some guidance for application of similar techniques to other language pairs. First, it appears that our approach to statistical MT is fairly robust, generally obtaining at least some benefit as additional resources are added. Second, simple resources such as the Bible and a moderately large bilingual term list proved to be sufficient to double the BLEU score over that which could be achieved based on loan words and other exact string matches alone. This is already a strong baseline, since Cebuano exhibits a large number of loan words from English, so even larger gains may be possible in other language pairs. Third, the resulting translations appear to the eye to be of some value for some tasks; user studies would be needed before a stronger claim could be made.

These results point to a couple of interesting opportunities for further work on this problem. Germann found that postediting yielded markedly better results from a manual translation effort in which translators were generating English rather than their native language (Germann, 2001). Postediting takes time, of course, so it would be useful to characterize the nature of the tradeoff between postediting and generation of additional (unedited) translations when the total time investment is held constant. Another interesting question is whether postediting the evaluation collection (thus better modeling the actual translation task) would affect our results. Another important question is the dependence of our techniques on the availability of a representative development test collection. Would drawing the development collection from Web resources (which may be somewhat less representative) adversely affect the results? The present test collection could be used to answer that question.

There are also some important questions that could be answered by extending the present test collection. Perhaps the most intriguing of them is whether we selected the right translation direction for the manually prepared translations. Answering that question would require a new manual translation effort, a complex undertaking (because of the need to find and coordinate native speakers). We chose not to do so in this case for two reasons (1) we would not have been able to read the resulting translations, adding to our logistic challenges in a time-constrained task, and (2) dialect effects might be more pronounced in Cebuano than in English. When we did this in June for Hindi, we again choose to translate into English for the same reasons. On that occasion, our colleagues at Carnegie-Mellon University chose to translate from English into Hindi. We have not yet had a chance to compare notes on this, but this experience with Hindi may ultimately yield some insight into the implications of choosing one translation direction over another.

At practical level, this experience and our experiences in June with Hindi have also revealed some simple enhancements that can improve the utility of our system for downstream applications. For the June exercise, we provided both Web-based and socket-based facilities for on-demand translation. The Web system proved to be useful when exploring new collections, and the socket service was used by New York University for as part of an interactive cross-language question answering system and by Alais-i to display cross-document co-reference results. Capitalization and punctuation reattachment were included in these systems. We also developed a separate version of our system that was tuned

for speed, reducing the size of the beam search somewhat and omitting the reordering of alignment templates. With that system, we were able to translate about 50,000 documents in a few days (using multiple processors), and the results were used to support research on summarization and information retrieval at a number of sites. Machine translation is, of course, always a means to an end rather than an end in itself. These sorts of interactions with those who want to build capabilities that we provide into their systems can therefore be a useful source of insight.

So what have we learned? Well, we haven't yet produced a useful MT system in a day using only resources that we didn't have the day before. But we have, for the first time, demonstrated that we could have. There are still many interesting questions to explore that might ultimately lead us to the ability to build even better systems that quickly. And, of course, there are some languages for which even obtaining a Bible and a moderately large bilingual term list in character-coded form could be a challenge in itself. But we have clearly established the point of departure; such systems can be built, and it is now up to us to build them well.

Acknowledgments

The authors are grateful to the translators and to the sites that provided and aligned dictionaries and parallel text for their invaluable assistance with resource generation, and to Rahul Bhagat, Uli Germann, Kevin Knight, and Anton Leuski for their help along the way. This work has been supported in part by DARPA cooperative agreement N660010028910.

References

Yaser Al-Onizan, Jan Curin, Michael Jahr, Kevin Knight, John Lafferty, Dan Melamed, Franz-Josef Och, David Purdy, Noah A. Smith, and David Yarowsky. 1994. Statistical machine translation. Technical report, Johns Hopkins University, Baltimore, MD, July. Summer Workshop Final Report, http://www.clsp.jhu.edu/ws99/projects/mt/.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proc. ARPA Workshop on Human Language Technology.*

Ulrich Germann. 2001. Building a statistical machine translation system from scratch: How much bang for the buck can we expect? In *ACL 2001 Workshop on Data-Driven Machine Translation*, Toulouse, France.

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

Douglas W. Oard, David Doermann, Bonnie Dorr, Daqing He, Philip Resnik, Amy Weinberg, William Byrne, Sanjeev Khudanpur, David Yarowsky, Anton Leuski, Philipp Koehn, and Kevin Knight. 2003. Desparately seeking Cebuano. In *Third Conference on Human Language Technologies*, Edmonton, Canada, May.

Douglas W. Oard. 2003. Surprise: It's Cebuano! *Team TIDES*, April.

Franz Josef Och and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, PA, July.

Franz J. Och, Christoph Tillmann, and Hermann Ney. 1999. Improved alignment models for statistical machine translation. In *Proc. of the Joint SIGDAT Conf. on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 20–28, University of Maryland, College Park, MD, June.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Sapporo, Japan, July.

Kishore A. Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. Bleu: a method for automatic evaluation of machine translation. Technical Report RC22176 (W0109-022), IBM Research Division, Thomas J. Watson Research Center, Yorktown Heights, NY, September.

William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. 2002. *Numerical Recipes in C++*. Cambridge University Press, Cambridge, UK.

Nicola Ueffing, Franz Josef Och, and Hermann

Ney. 2002. Generation of word graphs in statistical machine translation. In *Proc. Conference on Empirical Methods for Natural Language Processing*, pages 156–163, Philadelphia, PE, July.