Adapting an Example-Based Translation System to Chinese

Ying Zhang, Ralf D. Brown, and Robert E. Frederking
Language Technologies Institute, Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213-3890 USA
{joy,ralf,ref}@cs.cmu.edu

ABSTRACT

We describe an Example-Based Machine Translation (EBMT) system and the adaptations and enhancements made to create a Chinese-English translation system from the Hong Kong legal code and various other bilingual resources available from the Linguistic Data Consortium (LDC).

1. BACKGROUND

We describe an Example-Based Machine Translation (EBMT) system and the adaptations and enhancements made to create a Chinese-English translation system from the Hong Kong legal code and various other bilingual resources available from the Linguistic Data Consortium (LDC).

The EBMT software [1, 3] used for the experiments described here is a shallow system which can function using nothing more than sentence-aligned plaintext and a bilingual dictionary; and given sufficient parallel text, the dictionary can be extracted statistically from the corpus [2]. To perform a translation, the program looks up all matching phrases in the source-language half of the parallel corpus and performs a word-level alignment on the entries containing matches to determine a (usually partial) translation. Portions of the input for which there are no matches in the corpus do not generate a translation.

Because the EBMT system does not generate translations for 100% of the text it is given as input, a bilingual dictionary and phrasal glossary are used to fill any gaps. Selection of a "best" translation is guided by a trigram model of the target language [6].

Supporting Chinese required a number of changes to the program and training procedures; those changes are discussed in the next section.

2. ENHANCEMENTS

The first change required of the translation software was support for the two-byte encoding used for the Chinese text (GB-2312, "GB" for short). Further, the EBMT (as well as dictionary and glossary) approaches are word-based, but Chinese is ordinarily written without breaks between words. Thus, Chinese input must be segmented into individual words. The initial baseline system used the segmenter made available by the LDC. This segmenter uses a word-frequency list to make segmentation decisions, but although the list provided by the LDC is large, it did not completely cover the vocabulary of the EBMT training corpus (described below). As a result, many sentences had incorrect segmentations or included long sequences which were not segmented at all or were broken into single characters. Almost every Chinese character has at least one meaning, and its meaning may be entirely different from the meaning of the word containing it. The mis-segmenting of Chinese words due to the inadequate dictionary makes it very hard to build a statistical dictionary and properly index the EBMT corpus.

To improve the performance of the Chinese segmenter, we augmented its word list by finding sequences of characters in the training corpus that belong together, based on their frequency and high mutual information. We developed a form of term extraction to find English phrases which should be treated as atomic units for translation, thus increasing the average length of "words" in both source and target languages. Finally, we also created an augmented bilingual dictionary for use in word-level alignment for EBMT by applying statistical dictionary extraction techniques to the training corpus.

As the improved segmenter and the term finder may be producing excessively long phrases or phrases which are impossible to match in the other language, we repeat the procedure of segmenting/bracketing/dictionary-building several times. On each successive iteration, the segmenter and bracketer are limited to words and phrases for which the statistical dictionary from the previous iteration contains translations. Through this iteration, we increased the size of the statistical dictionary from each step and guaranteed that all Chinese words generated by the segmenter have translations in the dictionary. This helps ensure that the EBMT engine can perform word-level alignments.

3. EXPERIMENTAL DESIGN

The primary purpose of this experiment was to determine the effect of each enhancement by operating with various subsets of the enhancements. Since it rapidly becomes impractical to test all possible combinations, we opted for the following test conditions:

1. baseline: parallel corpus segmented with the LDC segmenter and LDC dictionary/glossary
2. baseline plus improved segmenter
3. baseline plus improved segmenter and term finder
4. baseline plus improved segmenter and statistical dictionary
5. baseline plus improved segmenter, term finder, and statistical dictionary

.

For training, we had available two parallel Chinese-English corpora distributed by the LDC: the complete Hong Kong legal code (after cleaning: 47.86 megabytes, 5.5 million English words, 9 million Chinese characters) where 85% of the content (by sentence) is unique, and a collection of Hong Kong news articles (after cleaning: 24.58 megabytes, 2.67 million English words, 4.5 million Chinese characters). In addition, LDC distributes a bilingual dictionary/phrasebook, which we also used.

To determine the effects of varying amounts of training data on overall performance, we divided the bilingual training corpus into ten nearly equal slices. Each test condition was then run ten times, each time increasing the number of slices used for training the system. After each training pass, the test sentences were translated and the system's performance evaluated automatically; selected points were then manually evaluated for translation quality.

The automatic performance evaluation measured coverage of the input and average phrase length. Coverage is the percentage of the input text for which a translation is produced by a particular translation method (since the EBMT engine does not generally produce hypotheses that cover every word of input), while average phrase length is a crude indication of translation quality – the longer the phrase that is translated, the more context is incorporated and the less likely it is that the wrong sense will be used in the translation or that (for EBMT) the alignment will be incorrect. Since the dictionary and glossary remain constant for a given test condition, only the EBMT coverage will be presented.

Manual grading of the output was performed using a web-based system with which the graders could assign one of three scores ("Good", "OK", "Bad") in each of two dimensions: grammatical correctness and meaning preservation. This type of quality scoring is commonly used in assessing translation quality, and is used by other TIDES participants. Fifty-two test sentences were translated for each of four points from the automated evaluation and these sets of four alternatives presented to the graders. The four points chosen were the baseline system with 100% of the training corpus, the full system with 20% and 100% training, and the full system trained on a corpus of Hong Kong news text (cross-domain); only four points were selected due to the difficulty and expense of obtaining large numbers of manual quality judgements.

To assess the performance of the system in a different domain, as well as the effect of the trigram language model on the selection of translated fragments for the final translation, we obtained manual judgements for 44 sentences on an additional four test conditions, each trained with the entire available parallel text and tested on Hong Kong news text rather than legal sentences. These points were the cross-domain case (trained on the legal corpus) and three different language models for within-domain training: an English language model derived from the legal corpus, one derived from the news corpus, and a pre-existing model generated from two gigabytes of newswire and broadcast news transcriptions.

4. RESULTS

We discovered that there is a certain amount of synergy between some of the improvements, particularly the term finder and statistical dictionary extraction. Applying the term finder modifies the parallel corpus in such a way that it becomes more difficult for the EBMT engine to find matches which it can align, while adding dictionary entries derived from the modified corpus eliminates that effect. As a result, we will not present the performance results for Test Condition 3 (improved segmenter plus term finder); further, the data for Test Conditions 2 (improved segmenter only) and 4 (improved segmenter plus statistical dictionary) may not accurately reflect the contribution of those two components to the full system used for Test Condition 5.

Translating Legal Code

System Training	Baseline 100%	Full 20%	Full 100%	X-Dom 100%
Syntactic	42.31%	54.81%	61.06%	39.42%
Semantic	43.75%	61.54%	64.42%	34.62%

Translating Hong Kong News

Training LangModel	News Legal	News News	News Prior	Legal Legal
Syntactic	45.67%	44.71%	47.60%	34.62%
Semantic	50.00%	50.96%	51.92%	47.12%

Figure 1: Judgements – Acceptable Translations

Figure 2 shows the proportion of the words in the test sentences for which the EBMT engine was able to produce a translation, while Figure 3 shows the average number of source-language words per translated fragment. These curves do not increase monotonically because, for performance reasons, the EBMT engine does not attempt to align every occurrence of a phrase, only the N (currently 12) most-recently added ones; as a result, adding more text to the corpus can cause EBMT to ignore matches that successfully align in favor of newer occurrences which it is unable to align.

Examining Figure 3, it is clear that the fifth slice (from 40 to 50%) is much more like the test data than other slices, resulting in longer matches. In general, the closer training and test text are to each other, the longer the phrases they have in common.

Figure 1 summarizes the results of human quality assessments. The "Good" and "OK" judgements were combined into "Acceptable" and the the percentage of "Acceptable" judgements was averaged across sentences and graders. As hoped and expected, the improvements do in fact result not only in better coverage by EBMT, but also in better quality assessments by the human graders. Further, the results on Hong Kong news text show that the choice of language model does have a definite effect on quality. These results also confirm the adage that there is no such thing as too much training text for language modeling, since the model generated from the EBMT corpus was unable to match the performance of the pre-existing model generated from two orders of magnitude more text.

5. CONCLUSIONS AND FUTURE WORK

As seen in Figure 2, the enhancements described here cumulatively provide a 12% absolute improvement in coverage for EBMT translations without requiring any additional knowledge resources. Further, the enhanced coverage does, in fact, result in improved translations, as verified by human judgements. We can also conclude that when we combine words into larger chunks on both sides of the corpus, the possibility of finding larger matches between the source language and the target language increases, which leads to the improvement of the translation quality for EBMT.

We will do further research on the interaction between the improved segmenter, term finder and statistical dictionary builder, utilizing the information provided by the statistical dictionary as feedback for the segmenter and term finder to modify their results. We are also investigating the effects of splitting the EBMT training into multiple sets of topic-specific sentences, automatically separated using clustering techniques.

The relatively low slope of the coverage curve also indicates that the training corpus is sufficiently large. Our prior experience with Spanish (using the UN Multilingual Corpus [5]) and French (using

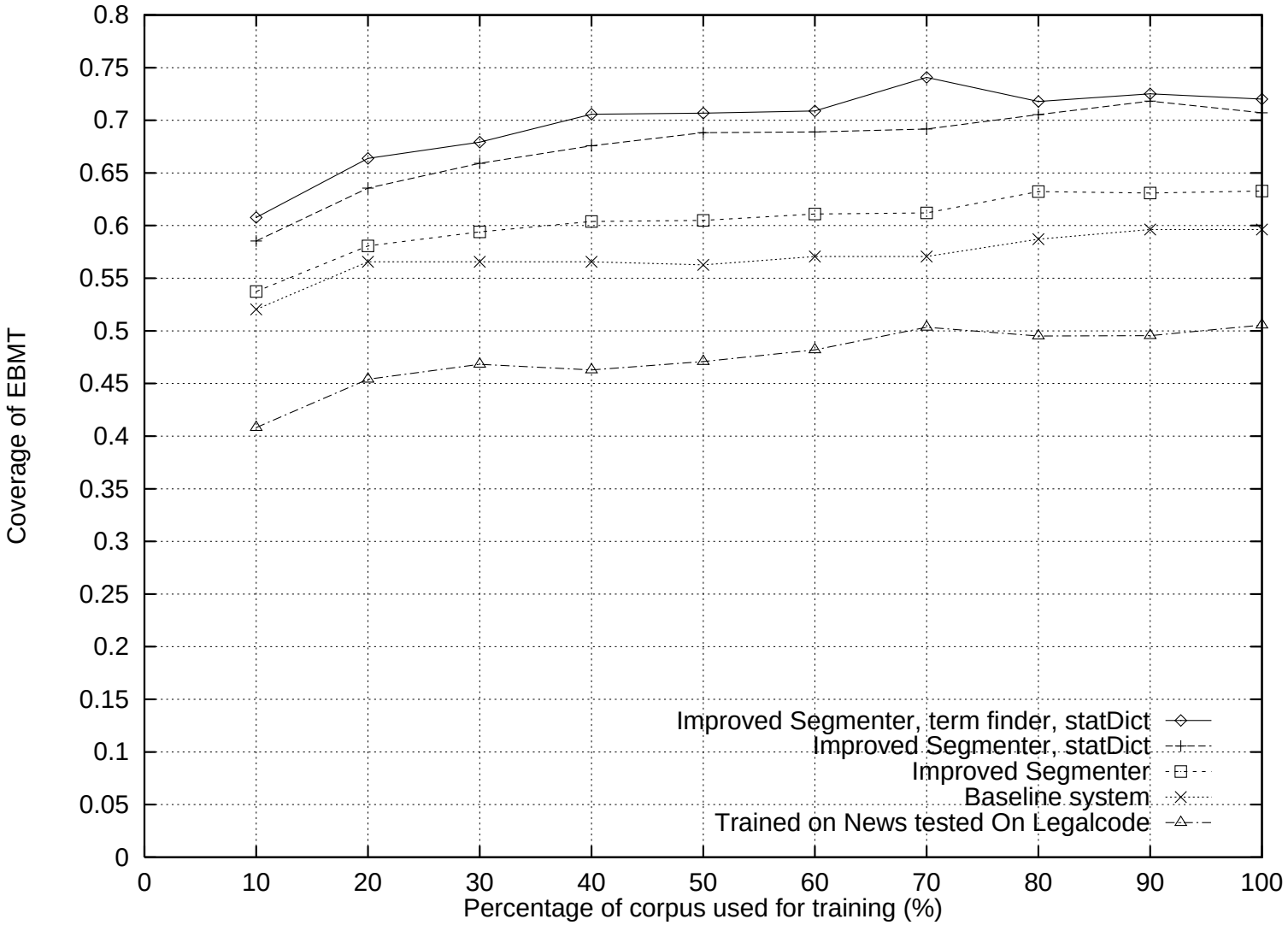

Figure 2: EBMT Coverage with Varying Training

the Hansard corpus [7]) was that the curve flattens out at between two and three million words of training text, which appears also to be the case for Chinese (each training slice contains approximately one million words of total text).

We have not yet taken full advantage of the features of the EBMT software. In particular, it supports equivalence classes that permit generalization of the training text into templates for improved coverage. We intend to test automatic creation of equivalence classes from the training corpus [4] in conjunction with the other improvements reported herein.

6. ACKNOWLEDGEMENTS

We would like to thank Alon Lavie and Lori Levin for their comments on drafts of this paper.

7. REFERENCES

[1] R. D. Brown. Example-Based Machine Translation in the PANGLOSS System. In *Proceedings of the Sixteenth International Conference on Computational Linguistics*, pages 169–174, Copenhagen, Denmark, 1996. `http://www.cs.cmu.edu/~ralf/papers.html`.

[2] R. D. Brown. Automated Dictionary Extraction for "Knowledge-Free" Example-Based Translation. In *Proceedings of the Seventh International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-97)*, pages 111–118, Santa Fe, New Mexico, July 1997. `http://www.cs.cmu.edu/~ralf/papers.html`.

[3] R. D. Brown. Adding Linguistic Knowledge to a Lexical Example-Based Translation System. In *Proceedings of the Eighth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-99)*, pages 22–32, Chester, England, August 1999. `http://www.cs.cmu.edu/~ralf/papers.html`.

[4] R. D. Brown. Automated Generalization of Translation Examples. In *Proceedings of the Eighteenth International Conference on Computational Linguistics (COLING-2000)*, pages 125–131, 2000.

[5] D. Graff and R. Finch. Multilingual Text Resources at the Linguistic Data Consortium. In *Proceedings of the 1994 ARPA Human Language Technology Workshop*. Morgan Kaufmann, 1994.

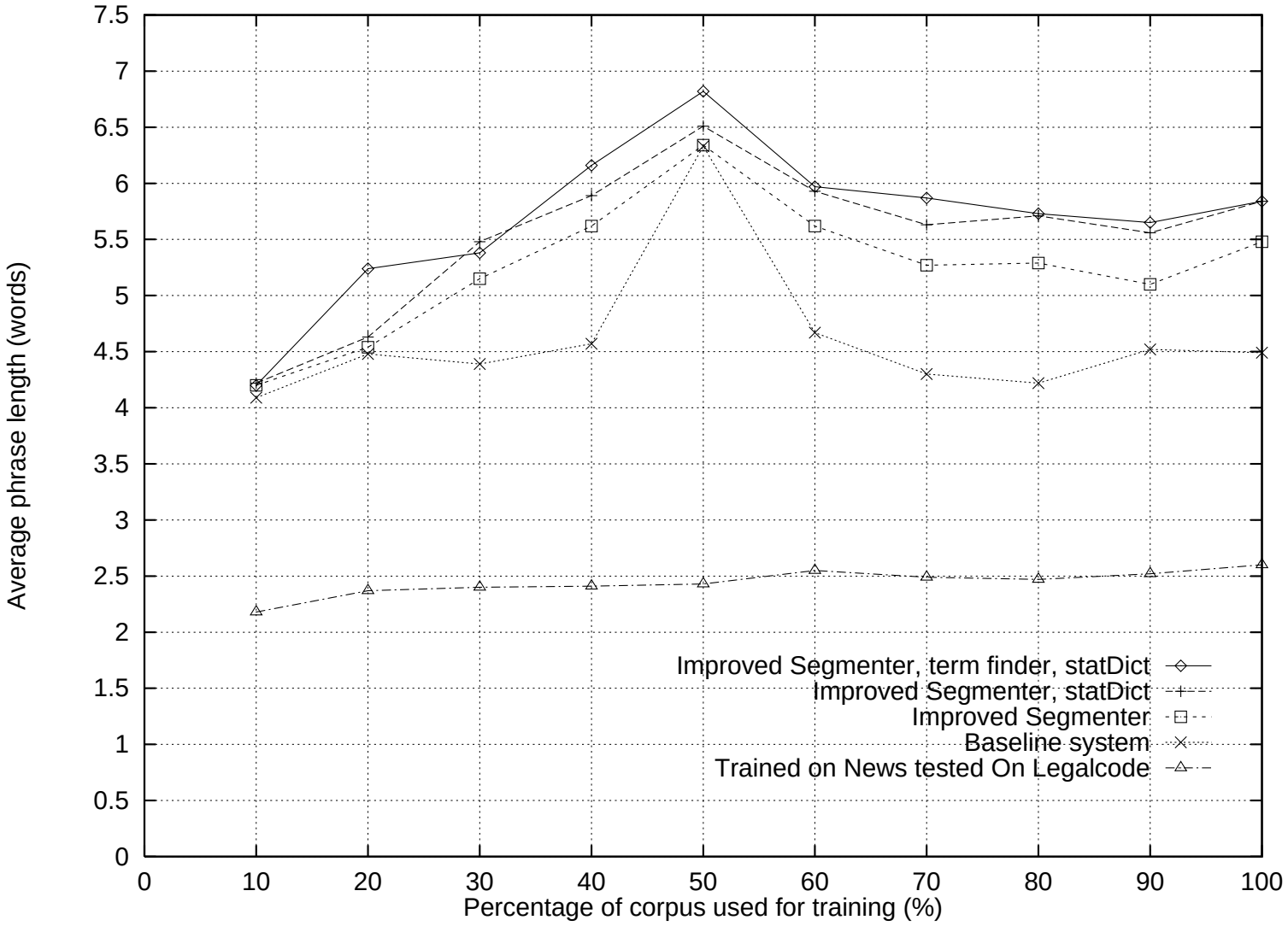

Figure 3: Average EBMT Match Lengths

[6] C. Hogan and R. E. Frederking. An Evaluation of the Multi-engine MT Architecture. In *Machine Translation and the Information Soup: Proceedings of the Third Conference of the Association for Machine Translation in the Americas (AMTA '98)*, volume 1529 of *Lecture Notes in Artificial Intelligence*, pages 113–123. Springer-Verlag, Berlin, October 1998.

[7] Linguistic Data Consortium. *Hansard Corpus of Parallel English and French*. Linguistic Data Consortium, December 1997. `http://www.ldc.upenn.edu/`.