The ISI/USC MT System

Ignacio Thayer, Emil Ettelaie, Kevin Knight, Daniel Marcu, Dragos Stefan Munteanu,
Franz Joseph Och[1]*, Quamrul Tipu*

USC Information Sciences Institute
4676 Admiralty Way, Suite 1001
Marina del Rey, CA 90292

Abstract

The ISI/USC machine translation system is a statistical system based on a phrase translation model that is trained on bilingual parallel data. This translation model is combined with several other knowledge sources in a log-linear manner. The weights of the individual components in the log-linear model are set by an automatic parameter-tuning method. The system described here has been developed for translating news text, and is a simplified version of the one we participated with in the NIST 2004 MT evaluation. We give a brief overview of the components of the system and discuss its performance at IWSLT.

1. The ISI/USC Machine Translation System

Our machine translation system uses a log-linear model to combine several different knowledge sources into a direct model of translation. The 12 different models used to score hypothesized translations are given in Table 1. We also give more in-depth descriptions of the major components.

1.1. Translation Model

At the core of the system is the alignment template translation model, which learns many-to-many mappings between word sequences from parallel bilingual data. A sentence is translated by segmenting a source-language sentence into phrases, translating these phrases with the ones observed in the training data, and reordering the target-language phrases. More details about the alignment template approach to machine translation used here are given in [1], [2].

For the IWSLT evaluation for Chinese- and Japanese-to-English, we trained the alignment template system on the 20,000 lines of bilingual basic travel expressions provided by the organizers. For the "additional" evaluation condition for Chinese, we used 6 of the allowed corpora provided by LDC. For the "unrestricted" evaluation condition for Chinese, we used 167M words of parallel news and political data obtained from LDC in addition to the provided data. When mixing the provided in-domain data with out-of-domain data, the in-domain data was weighted by a factor of 5, and was resegmented with the LDC segmenter.

1.2. Language Model

A smoothed trigram model was also used to score hypothesized translations. We used the SRI Language Modelling Toolkit to train a language model smoothed with Kneser-Ney discounting. For all of the evaluation conditions, a language model was trained on the English half of the parallel corpus used for alignment-template training. For the "additional" and "unrestricted" evaluation conditions, an additional language model was used that was trained on 800M words of monolingual news text. Each language model is considered an independent information source, and is weighted separately in the global log-linear model.

1.3. Minimum Error Rate Training

The individual model weights of the log-linear model are set using a parameter tuning procedure that minimizes the error rate of a given evaluation function (such as the BLEU score) on a held-out test corpus. Setting model weights in order to minimize the error of the function used for testing has been shown to provide better results than maximum-likelihood training [3]. For this evaluation, we optimize parameters to achieve the best performance with respect to the BLEU score. We split the provided development data into two equally sized corpora that were used separately for minimum error training and testing.

2. Results

The results achieved by our system are displayed in Table 2. We submitted 3 Chinese-to-English configurations and one Japanese-to-English configuration. The 20,000 sentences of basic travel expression data provided during the evaluation ("supplied" data) is included in the training data for all of the systems. Where allowed, we use a language model trained on 800M words of news data ("lm"). For the "additional" and "unrestricted" evaluation conditions, we use 6 of the allowed LDC corpora ("LDC"), and for the unrestricted data track, we use all of the data allowed in the NIST evaluation (a superset of the 6 corpora in "LDC"). It should be noted that because of time constraints, minimum error training was not run on the "unrestricted" Chinese-to-English system. Instead, the model weights from the "supplied+LDC+lm" sub-

[1] Now at Google, Inc.

Component	Description
Alignment Template Model	Phrase-level $p(e\|f)$
Language Model	Smoothed 3-gram model
Word Penalty	Bonus for longer sentences
Alignment Template Penalty	Bonus for longer alignment templates
Left Monotone	Penalizes left non-monotonicity
Right Monotone	Penalizes right non-monotonicity
Model 1	Full-sentence $p(e\|f)$ IBM Model 1 probability
Inverse Model 1	Alignment Template $p(f\|e)$ Model 1 probability
Lexicon-Backup Penalty	Penalty for using word-based translations
Jump Penalty	Penalty for non-monotonicity
Missing Word Penalty	Penalty for unaligned content words
Lexical Smoothing	Weighting of word-to-word translation probabilities

Table 1: Scoring components incorporated into the log-linear model

Data Condition	Chinese	Japanese
Supplied	37.42	40.08
Additional (supplied+LDC+lm)	44.05	N/S
Unrestricted (supplied+lm+NIST)	24.3	N/S

Table 2: Results on the 3 evaluation conditions. Minimum error training was not run on the Chinese-to-English "unrestricted" system because of time constraints. For Japanese-to-English, only one system was submitted.

mission were used.

The best results were achieved by the system "supplied+LDC+lm", which used the supplied data (weighted by a factor of 5), 6 of the LDC corpora allowable in the additional data track, plus the additional language model trained on 800M words of news data. Note that this is better than we reported after the evaluation, as we made an error in submission.

The worst results were achieved when using all of the out-of-domain news and political data. This experiment was run to gauge the effect of a large amount of news data (167M words) on translation performance in another domain, but was handicapped by the fact that because of insufficient time, the model weights were not optimally tuned.

3. References

[1] Koehn, P., and Och, F. J., Marcu, D., "Statistical Phrase-Based Translation", Proceedings of the Human Language Technology Conference 2003 (HLT-NAACL 2003), Edmonton, Canada, May 2003.

[2] Och, F. J. and Ney, H., "The Alignment Template Approach to Statistical Machine Translation", Accepted for publication in Computational Linguistics, 2004.

[3] Och, F. J., "Minimum Error Rate Training for Statistical Machine Translation", ACL 2003: Proc. of the 41st Annual Meeting of the Association for Computational Linguistics, Japan, Sapporo, July 2003.