The University of Washington Machine Translation System for ACL WMT 2008

Amittai Axelrod, Mei Yang, Kevin Duh, Katrin Kirchhoff
Department of Electrical Engineering
University of Washington
Seattle, WA 98195
{amittai,yangmei,kevinduh,katrin} @ee.washington.edu

Abstract

This paper present the University of Washington's submission to the 2008 ACL SMT shared machine translation task. Two systems, for English-to-Spanish and German-to-Spanish translation are described. Our main focus was on testing a novel boosting framework for N-best list reranking and on handling German morphology in the German-to-Spanish system. While boosted N-best list reranking did not yield any improvements for this task, simplifying German morphology as part of the preprocessing step did result in significant gains.

1 Introduction

The University of Washington submitted systems to two data tracks in the WMT 2008 shared task competition, English-to-Spanish and German-to-Spanish. In both cases, we focused on the in-domain test set only. Our main interest this year was on investigating an improved weight training scheme for N-best list reranking that had previously shown improvements on a smaller machine translation task. For German-to-Spanish translation we additionally investigated simplifications of German morphology, which is known to be fairly complex due to a large number of compounds and inflections. In the following sections we first describe the data, baseline system and postprocessing steps before describing boosted N-best list reranking and morphology-based preprocessing for German.

2 Data and Basic Preprocessing

We used the Europarl data as provided (version 3b, 1.25 million sentence pairs) for training the translation model for use in the shared task. The data was lowercased and tokenized with the auxiliary scripts provided, and filtered according to the ratio of the sentence lengths in order to eliminate mismatched sentence pairs. This resulted in about 965k parallel sentences for English-Spanish and 950k sentence pairs for German-Spanish. Additional preprocessing was applied to the German corpus, as described in Section 5. For language modeling, we additionally used about 82M words of Spanish newswire text from the Linguistic Data Consortium (LDC), dating from 1995 to 1998.

3 System Overview

3.1 Translation model

The system developed for this year's shared task is a state-of-the-art, two-pass phrase-based statistical machine translation system based on a log-linear translation model (Koehn et al, 2003). The translation models and training method follow the standard Moses (Koehn et al, 2007) setup distributed as part of the shared task. We used the training method suggested in the Moses documentation, with lexicalized reordering (the `msd-bidirectional-fe` option) enabled. The system was tuned via Minimum Error Rate Training (MERT) on the first 500 sentences of the `devtest2006` dataset.

Proceedings of the Third Workshop on Statistical Machine Translation, pages 123–126,
Columbus, June 2008.

3.2 Decoding

Our system used the Moses decoder to generate 2000 output hypotheses per input sentence during the first translation pass. For the second pass, the N-best lists were rescored with the additional language models described below. We re-optimized the model combination weights with a parallelized implementation of MERT over 16 model scores on the `test2007` dataset. Two of these model scores for each hypothesis were from the two language models used in our second-pass system, and the rest correspond to the 14 Moses model weights (for reordering, language model, translation model, and word penalty).

3.3 Language models

We built all of our language models using the SRILM toolkit (Stolcke, 2002) with modified Kneser-Ney discounting and interpolating all n-gram estimates of order > 1. For first-pass decoding we used a 4-gram language model trained on the Spanish side of the Europarl v3b data. The optimal n-gram order was determined by testing language models with varying orders (3 to 5) on `devtest2006`; BLEU scores obtained using the various language models are shown in Table 1. The 4-gram model performed best.

Table 1: LM ngram size vs. output BLEU on the dev sets.

order	devtest2006	test2007
3-gram	30.54	30.69
4-gram	31.03	30.94
5-gram	30.85	30.84

Two additional language models were used for second pass rescoring. First, we trained a large out-of-domain language model on Spanish newswire text obtained from the LDC, dating from 1995 to 1998.

We used a perplexity-filtering method to filter out the least relevant half of the out-of-domain text, in order to significantly reduce the training time of the large language model and accelerate the rescoring process. This was done by computing the perplexity of an in-domain language model on each newswire sentence, and then discarding all sentences with greater than average perplexity. This reduced the size of the training set from 5.8M sentences and 166M tokens to 2.8M sentences and 82M tokens. We then further restricted the vocabulary to the union of the vocabulary lists of the Spanish sides of the de-es and en-es parallel training corpora. The remaining text was used to train the language model.

The second language model used for rescoring was a 5-gram model over part-of-speech (POS) tags. This model was built using the Spanish side of the English-Spanish parallel training corpus. The POS tags were obtained from the corpus using Freeling v2.0 (Atserias et al, 2006).

We selected the language models for our translation system were selected based on performance on the English-to-Spanish task, and reused them for the German-to-Spanish task.

4 Boosted Reranking

We submitted an alternative system, based on a different re-ranking method, called BoostedMERT (Duh and Kirchhoff, 2008), for each task. BoostedMERT is a novel boosting algorithm that uses Minimum Error Rate Training (MERT) as a weak learner to build a re-ranker that is richer than the standard log-linear models. This is motivated by the observation that log-linear models, as trained by MERT, often do not attain the oracle BLEU scores of the N-best lists in the development set. While this may be due to a local optimum in MERT, we hypothesize that log-linear models based on our K re-ranking features are also not sufficiently expressive.

BoostedMERT is inspired by the idea of Boosting (for classification), which has been shown to achieve low training (and generalization) error due to classifier combination. In BoostedMERT, we maintain a weight for each N-best list in the development set. In each iteration, MERT is performed to find the best ranker on weighted data. Then, the weights are updated based on whether the current ranker achieves oracle BLEU. For N-best lists that achieve BLEU scores far lower than the oracle, the weights are increased so that they become the emphasis of next iteration's MERT. We currently use the factor e^{-r} to update the N-best list distribution, where r is the ratio of the oracle hypothesis' BLEU to the BLEU of the selected hypothesis. The final ranker is a weighted combination of many such rankers.

More precisely, let w_i be the weights trained by MERT at iteration i. Given any w_i, we can generate a ranking y_i over an N-best list where y_i is an N-dimensional vector of predicted ranks. The final ranking vector is a weighted sum: $y = \sum_{i=1}^{T} \alpha_i y_i$, where α_i are parameters estimated during the boosting process. These parameters are optimized for maximum BLEU score on the development set. The only user-specified parameter is T, the number of boosting iterations. Here, we choose T by dividing the dev set in half: dev1 and dev2. First, we train BoostedMERT on dev1 for 50 iterations, then pick the T with the best BLEU score on dev2. Second, we train BoostedMERT on dev2 and choose the optimal T from dev1. Following the philosophy of classifier combination, we sum the final rank vectors y from each of the dev1- and dev2-trained BoostedMERT to obtain our final ranking result.

5 German → Spanish Preprocessing

German is a morphologically complex language, characterized by a high number of noun compounds and rich inflectional paradigms. Simplification of morphology can produce better word alignment, and thus better phrasal translations, and can also significantly reduce the out-of-vocabulary rate. We therefore applied two operations: (a) splitting of compound words and (b) stemming.

After basic preprocessing, the German half of the training corpus was first tagged by the German version of TreeTagger (Schmid, 1994), to identify part-of-speech tags. All nouns were then collected into a noun list, which was used by a simple compound splitter, as described in (Yang and Kirchhoff, 2006). This splitter scans the compound word, hypothesizing segmentations, and selects the first segmentation that produces two nouns that occur individually in the corpus. After splitting the compound nouns in the filtered corpus, we used the TreeTagger again, only this time to lemmatize the (filtered) training corpus.

The stemmed version of the German text was used to train the translation system's word alignments (through the end of step 3 in the Moses training script). After training the alignments, they were projected back onto the unstemmed corpus. The parallel phrases were then extracted using the standard procedure. Stemming is only used during the training stage, in order to simplify word alignment. During the evaluation phase, only the compound-splitter is applied to the German input.

6 Results

6.1 English → Spanish

The unofficial results of our 2nd-pass system for the 2008 test set are shown in Table 2, for recased, untokenized output. We note that the basic second-pass model was better than the first-pass system on the 2008 task, but not on the 2007 task, whereas BoostedMERT provided a minor improvement in the 2007 task but not the 2008 task. This is contrary to previous results in the Arabic-English IWSLT 2007 task, where boosted MERT gave an appreciable improvement. This result is perhaps due to the difference in magnitude between the IWSLT and WMT translation tasks.

Table 2: En→Es system on the test2007 and test2008 sets.

System	test2007	test2008
First-Pass	30.95	31.83
Second-Pass	30.94	32.72
BoostedMERT	31.05	32.62

6.2 German → Spanish

As previously described, we trained two German-Spanish translation systems: one via the default method provided in the Moses scripts, and another using word stems to train the word alignments and then projecting these alignments onto the unstemmed corpus and finishing the training process in the standard manner. Table 3 demonstrates that the word alignments generated with word-stems markedly improved first-pass translation performance on the `dev2006` dataset. However, during the evaluation period, the worse of the two systems was accidentally used, resulting in a larger number of out-of-vocabulary words in the system output and hence a poorer score. Rerunning our German-Spanish translation system correctly yielded significantly better system results, also shown in Table 3.

Table 3: De→Es first-pass system on the development and 2008 test set.

System	dev2006	test2008
Baseline	23.9	21.2
Stemmed Alignments	26.3	24.4

6.3 Boosted MERT

BoostedMERT is still in an early stage of experimentation, and we were interested to see whether it improved over traditional MERT in re-ranking. As it turns out, the BLEU scores on test2008 and test2007 data for the En-Es track are very similar for both re-rankers. In our post-evaluation analysis, we attempt to understand the reasons for similar BLEU scores, since the weights w_i for both re-rankers are qualitatively different. We found that out of 2000 En-Es N-best lists, BoostedMERT and MERT differed on 1478 lists in terms of the final hypothesis that was chosen. However, although the rankers are choosing different hypotheses, the chosen strings appear very similar. The PER of BoostedMERT vs. MERT results is only 0.077, and manual observation indicates that the differences between the two are often single phrase differences in a sentence.

We also computed the sentence-level BLEU for each ranker with respect to the true reference. This is meant to check whether BoostedMERT improved over MERT in some sentences but not others: if the improvements and degradations occur in the same proportions, a similar corpus-level BLEU may be observed. However, this is not the case. For a majority of the 2000 sentences, the sentence-level BLEU for both systems are the same. Only 10% of sentences have absolute BLEU difference greater than 0.1, and the proportion of improvement/degradation is similar (each 5%). For BLEU differences greater than 0.2, the percentage drops to 4%.

Thus we conclude that although BoostedMERT and MERT choose different hypotheses quite often, the string differences between their hypotheses are negligible, leading to similar final BLEU scores. BoostedMERT has found yet another local optimum during training, but has not improved upon MERT in this dataset. We hypothesize that dividing up the original development set into halves may have hurt BoostedMERT.

7 Conclusion

We have presented the University of Washington systems for English-to-Spanish and German-to-Spanish for the 2008 WMT shared translation task. A novel method for reranking N-best lists based on boosted MERT training was tested, as was morphological simplification in the preprocessing component for the German-to-Spanish system. Our conclusions are that boosted MERT, though successful on other translation tasks, did not yield any improvement here. Morphological simplification, however, did result in significant improvements in translation quality.

Acknowledgements

This work was funded by NSF grants IIS-0308297 and IIS-0326276.

References

Atserias, J. et al. 2006. FreeLing 1.3: Syntactic and semantic services in an open-source NLP library. *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006).* Genoa, Italy.

Duh, K., and Kirchhoff, K. 2008. Beyond Log-Linear Models: Boosted Minimum Error Rate Training for MT Re-ranking. To appear, *Proceedings of the Association for Computational Linguistics (ACL).* Columbus, Ohio.

Koehn, P. and Och, F.J. and Marcu, D. 2003. Statistical phrase-based translation. *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, (HLT/NAACL).* Edmonton, Canada.

Koehn, P. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation *Proceedings of MT Summit.*

Koehn, P. et al. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. *Annual Meeting of the Association for Computational Linguistics (ACL)*, demonstration session. Prague, Czech Republic.

Schmid, H. 1994. Probabilistic part-of-speech tagging using decision trees. *International Conference on New Methods in Language Processing*, Manchester, UK.

Stolcke, A. 2002. SRILM - An extensible language modeling toolkit. *Proceedings of ICSLP.*

Yang, M. and K. Kirchhoff. 2006. Phrase-based backoff models for machine translation of highly inflected languages. *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006).* Trento, Italy.