Morphology Generation for Statistical Machine Translation

Ann Clifton
Simon Fraser University
Burnaby, BC, Canada
aca69@sfu.ca

Anoop Sarkar
Simon Fraser University
Burnaby, BC, Canada
anoop@cs.sfu.ca

1 Introduction

State of the art Machine Translation (MT) systems tend to perform poorly when translating into languages with rich morphological systems. When multiple morphemes inflect a single word stem, the total lexicon of surface forms can be very large, and any given surface form may occur infrequently in a text, leading to a significant data sparsity problem.

In this paper, we examine how morphological information incorporated into MT models can be used to address this problem. Our first approach is adapted from previous work (Popović and Ney, 2004; de Gispert and Mariño, 2008), which segments the morphologically complex language into stems and suffixes and then trains the MT system on the segmentations, treating each as an individual word; the motivation is that productive stems and suffixes in isolation may reduce sparsity, as well as increase the level of lexical symmetry between source and target in the translation model.

The second approach we use was proposed in (Minkov et al., 2007; Toutanova et al., 2008), in which complex words in the target are stemmed before translation, and then the morphologically inflected forms are generated for the target stem output as a post-processing step by predicting a sequence of morphemes using a Maximum Entropy Markov Model (MEMM). We implemented this approach using Conditional Random Fields (CRFs) instead of MEMMs.

2 Approach and Experiments

In our first set of experiments, we performed both supervised and unsupervised segmentation on the data, onto which we added word-internal morphology boundary markers for recovery of the surface forms after decoding; we then used this segmented data to train a phrase-based MT system.

For the morphology-generation approach, we follow (Minkov et al., 2007; Toutanova et al., 2008), who use MEMMs to predict inflected forms from sets of inflected surface forms which have been gathered to correspond to each stem. We use CRFs instead of MEMMs and we use a smaller set of features, dropping the syntax-based features.

To isolate the generation task, we begin by stemming the reference translations and using the CRF to predict the morphological tag sequences for the stems. For the full translation task, we obtain the surface forms corresponding to the tags by using n-gram language model disambiguation (over morphemes) in surface form prediction. In the full pipeline, we train a phrase-based MT system on stemmed text and use our CRF model to predict the inflection tag sequences on the MT output stems, and from the inflection tags produce the surface word using the language model. We then combined the two approaches using the CRF model to predict suffixes for hanging stems in the segmentation-trained MT output.

English to Finnish We chose Finnish as our target language because of the richness of its morphological system and because MT systems performed most poorly on producing Finnish (Koehn, 2005). We used English-Finnish data from the Europarl data set and trained all four MT models on those sentences 40 words or less, 977335/943159 sentences for the unsupervised/supervised segmented data respectively, and 986166 sentences for the word and stem-only data. The test data was also from Europarl, the same set as in (Koehn, 2005) with 2000 sentences.

Morpheme segmentation To derive the stem and morphological information by supervised means, we ran the Omorfi FST morphological analyzer (Omorfi, 2007) on the data, from which we extracted for each word its lemma, part-of-speech (POS) tag, and morphologically decomposed inflection form. When no analysis was generated,

we retained the unstemmed word with null feature values. We used the FST analyzer in hopes that it might be more effective at yielding productive morphological patterns in fusional type languages with a great degree of morphophonemic alternation that make straightforward segmentation difficult. To derive the unsupervised segmentation for greater coverage and generality, we used Morfessor (Creutz and Lagus, 2005). Using these segmentations, we trained four phrase-based MT models (in addition to the word-trained baseline model): the supervised and unsupervised stem and suffix segmentations, and a model trained on the stems alone.

CRF-based morpheme prediction We trained the CRF model on lexical and morphological features using CRFSGD (Bottou, 2007). For the lexical features, we considered unigrams and bigrams of up to two of the previous and next lemmas and POS-tags, as well as the unigrams and bigrams of the previous two POS-tags for prediction bigrams. For the tag sequence morphological features, we represented each of the series of morphological tags making up the inflection prediction set as a vector of feature values corresponding to each morphological category, using the categories of number, case, and person. To recover the fully inflected word forms corresponding to the CRF output lemma plus inflection tag sequence, we used a word-based language model to predict which inflected word should correspond to an ambiguous lemma plus tag sequence. For the tag-sequence CRF training set, we used 315,247 Finnish sentences and the same 2000 sentence test set.

We tested the CRF morphology prediction model on a stemmed version of a Finnish text as well as on the (ostensibly noisier) output of the Moses MT system trained on stemmed input from the supervised analysis. Since the unsupervised stems-only model performed more poorly than the baseline, we did not use this output to test the morphology generation post-processing.

We then tested the CRF suffix prediction model on the output of the MT system trained on segmented stems and suffixes, applying it only to hanging stems translated without a following suffix.

3 Results and Ongoing Work

Table 1 shows that the MT models trained on both supervised and unsupervised segmentations outperformed the baseline word-based model.

Model	BLEU score
Baseline-words	14.39
USV segs	14.94
SV segs	14.58
SV stems with tag-CRF	10.09
SV segs with suffix-CRF	14.58

Table 1: Model scores. USV/SV refer to unsupervised/supervised; segs refer to models trained on segmentation with morphology in decoding; stems refers to models trained on stems alone with morphology generated in post-processing.

For the evaluation of the CRF morphology prediction model, we first compared the output of the CRF model on the stemmed reference translations to the original inflection sequences generated by the morphological analyzer before applying the language model to predict the corresponding surface forms. The best results were obtained using features that were conjunctions of prediction bigrams and observation context. For the CRF trained on tag sequences, the accuracy was 83.85% and 79.18%.

We then looked at the performance on MT system output, to see if the performance gains of translating stems would outweigh errors in surface form prediction. For the tag-sequence CRF, this was evaluated after using the language model to recover the surface forms. When we applied the CRF tag/suffix prediction model to the hanging stems in the segmentation-trained MT output, the BLEU scores remained the same.

We found that the models that use the morphology in translation was able to improve the BLEU scores over the baseline. In addition, that higher accuracy in inflection-tag sequence prediction was obtained using overlapping features would seem to confirm the utility of using a CRF model for capturing dependencies between predictions.

To improve upon the results for using the CRF to boost the segmented MT output, we are currently running experiments that include augmenting the CRF model feature set with lexical and syntactic bilingual features of the aligned corpora, and using discontinuous phrase-based and lattice-based decoding.

References

Leon Bottou. Stochastic Gradient CRFs homepage. http://leon.bottou.org/projects/sgd. 2007

Mathias Creutz and Krista Lagus. Inducing the morphological lexicon of a natural language from unannotated text. *Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning*, 2005.

Adria de Gispert and José Mariño. On the impact of morphology in English to Spanish statistical MT. *Speech Communication*, 2008.

Phillip Koehn. Europarl: a parallel corpus for statistical machine translation. *Proceedings of MT Summit*, 2005.

Phillip Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open source toolkit for statistical machine translation *Annual Meeting of ACL, demonstration session*, 2007.

John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional random fields: probabilistic models for segmenting and labeling data. *Proceedings of the 18th International Conference on Machine Learning*, 2001.

Andrew McCallum, Dayne Freitag, and Fernando Pereira. Maximum entropy markov models for information extraction and retrieval. *Proceedings of the 17th International Conference on Machine Learning*, 2000.

Einat Minkov, Kristina Toutanova, and Hisami Suzuki. Generating complex morphology for machine translation. *Association for Computational Linguistics*, 2007.

Maja Popović and Hermann Ney. Towards the use of word stems and suffixes for statistical machine translation. *4th International Conference on Language Resources and Evaluation (LREC)*, 2004.

Hisami Suzuki and Kristina Toutanova. Learning to predict case markers in Japanese. *Association for Computational Linguistics*, 2006.

Kristina Toutanova and Hisami Suzuki. Generating case markers in machine translation. *Association for Computational Linguistics*, 2007.

Kristina Toutanova, Hisami Suzuki, and Achim Ruopp. Applying morphology generation models to machine translation *Proceedings of ACL-08: HLT*, 2008.

http://kitwiki.csc.fi/twiki/bin/view/KitWiki/OMorFiHome 2007.