Morphological Generation of German for SMT

Alexander Fraser, Marion Weller, Aoife Cahill, Fabienne Fritzinger
Institute for NLP, University of Stuttgart, Germany
contact: fraser@ims.uni-stuttgart.de

0.1 Introduction

We participated in the ACL WMT 2009 shared task for translation of German to English, and English to German. We used the Moses open source system, combined with morphological processing. For German to English, we had the only constraint system comparable with the open-data systems. One of the reasons the system performed well was strong reduction of the German vocabulary, through a simplistic corpus-driven algorithm with minimal linguistic knowledge, which performed aggressive inflection removal and compound splitting.

For the English to German task, we submitted a two-step system. In the first step we translated English to the reduced German representation (the same representation which we used as the input to our German to English system). In the second step, we built another Moses system to "translate" the reduced German representation back to normal German through the addition of inflection and merging of split compound words. This two-step system was the worst constraint system submitted (and in fact the only constraint system that differed from all other constraint systems which otherwise scored as a group).

Our presentation at the Research Workshop of the Israel Science Foundation on Machine Translation and Morphologically-rich Languages will describe work done as a direct reaction to the poor performance of our English to German system, and to the poor performance of SMT systems in general for this translation task. Our current English to German system for news translation improves by 0.84 BLEU over the baseline and uses sophisticated morphological generation based on SMOR, the University of Stuttgart morphological analyzer/generator of German, and BitPar, a state-of-the-art parser of German also developed in-house.

0.2 Overview of the translation process

The work we will describe is currently focused on generalizing to better model German NPs and PPs. We particularly want to ensure that we can generate novel German NPs, where what we mean by novel is that the inflected realization is not present in the training data (neither in the parallel data nor in the monolingual data used for language modeling).

In order to ensure coherent German NPs, we model *case*, *gender*, *number*, and *definiteness* agreement. This is a diverse group of features. *Number* of the German noun and *definiteness* of the article are often easily determined given the English source and the word alignment. *Gender* of the noun is innate and often difficult to determine given the English source word (for example: inanimate objects in German have genders that seem to be arbitrary to non-native speakers). *Case* is a function of the slot in the subcategorization frame of the verb. There is agreement in all of these features in an NP. For instance the *gender* of an article is determined by the head noun, while the *definiteness* of an adjective is determined by the choice of indefinite or definite article, etc.

In terms of translation, we can have a large number of surface forms. For instance, English "blue" can be translated as German blau, blaue, blauer, blaues, blauen. We try to predict which form is correct given the context. Our system is able to generate forms which were not seen in the training data. We follow a two-step process: Translate to "blau" (the stem), and then predict features (e.g., we might predict *case*=nominative, *gender*=feminine, *number*=singular, *definiteness*=definite) to generate the form "blaue").

We begin building an SMT system by parsing the German training data with BitPar. Following this we extract morphological features from the parse. Next, we lookup the surface forms in the SMOR morphological analyzer. We resolve conflicts any conflicts between the parse and possible analyses according to SMOR and then choose the SMOR analysis which is consistent. Finally, we output the "stems" of the German text, with the addition of markup (which will be discussed later).

We then build a standard Moses system translating from English to stems (however, we use a single additional factor to obtain the coarse POS for each stem). Once we are given a sequence of stems and POS, we predict the correct inflection using a sequence model and finally generate the final surface forms.

0.3 Stem Markup

In the presentation we will present a comparison between different markup styles for the stems (we will also compare this with the two-step system we submitted to ACL WMT 2009). The best markup so far is:
- Nouns are marked with *gender* and *number*.
- Prepositions are marked with the *case* their object takes (this moves some of the difficulty in predicting *case* from the inflection prediction step to the stem translation step).
- Articles are reduced to "definite-article" or "indefinite-article".
- Contractions consisting of a preposition and an article (such as German "beim", which can mean "at the") are split and each part is marked as above.
- For verbs, we use the surface form of the inflected verb form (rather than the stem) in the hope that this can act as a weak surrogate to the subcategorization frame for predicting *case*, see below.
- For all other words, we use their stems (except for words not covered by SMOR, where we use surface forms).

0.4 Inflection Prediction

We currently use 4 simple HMMs for solving the inflection prediction problem. Each linguistic feature is modeled independently and has a different input representation based on the previously described markup. The input consists of a sequence of coarse POS tags, and for those stems that are marked up with the relevant feature, this feature value. So, for example, for prediction of *definiteness*, article POS tags will additionally be marked as definite or indefinite in the input. The output of the HMM consists of a POS and a feature value prediction. For example, *definiteness* must be predicted for adjectives, so adjective POS tags in the output will be marked as definite or indefinite. For the prediction of *case* we use the inflected verb form in place of the POS, so that the *case* of NPs to the left and right of the verb can be (somewhat) improved. Finally, given the stem, the POS tag and the relevant features, each surface form is generated using SMOR.

When doing monolingual prediction (i.e. using the stems of clean text), we obtain an overall accuracy of about 91% on ambiguous words (the baseline of taking the most frequent inflection for a stem has about a 59% accuracy, accuracy is measured by checking whether the generated form is the same as the input form). The prediction works quite well for *gender*, *number* and *definiteness*, which are local features to the NP that normally agree with the explicit markup output by the stem translation system (for example, the *gender* of a common noun, which is marked in the stem markup, is usually successfully propagated to the rest of the NP). However, prediction of *case* does not always work well.

0.5 Conclusion

When we combine the two steps (stem translation and inflection prediction) we obtain an improvement of 0.84 BLEU over a baseline system for the ACL WMT 2009 shared task on news translation. We make two observations. In our current evaluation, we use a very similar word alignment for both the baseline and the morphological system. We will realize further improvements by obtaining an improved word alignment by eliminating morphological tags on the German side and removing markers of *number* from the English side. The first step of translating from English to German stems (with the markup we previously discussed) is substantially easier than translating directly to inflected German (we see BLEU scores on stems+markup that are over 2.0 BLEU higher than the BLEU scores on inflected forms when running MERT; and the number of 4-gram matches is substantially improved, which is partially due to simple vocabulary reduction, but also seems to be due to better word ordering through less sparsity in the ordering and language models).

In current work we are trying to improve inflection prediction by switching from the 4 simple HMM models to a single CRF over all of the linguistic features that encodes the same dependencies as each of the HMMs, but can also be strongly lexicalized (so the prediction of inflection is conditioned on both the English input and the German stems). In future work, we would like to condition the inflection prediction on the parse tree of the English input, and we would also like to experiment with target-side German syntax models for stem translation, both with a view towards improving *case* prediction. This work is funded by the German Research Foundation: DFG grant "Models of Morphosyntax for Statistical Machine Translation".