Extending memory-based machine translation to phrases[1]

Maarten van Gompel Antal van den Bosch Peter Berck

ILK Research Group, Tilburg centre for Cognition and Communication, Tilburg University, P.O. Box 90153, NL-5000 LE Tilburg, The Netherlands

Abstract

We present a phrase-based extension to memory-based machine translation. This form of example-based machine translation employs lazy-learning classifiers to translate fragments of the source sentence to fragments of the target sentence. Source-side fragments consist of variable-length phrases in a local context of neighboring words, translated by the classifier to a target-language phrase. We compare three methods of phrase extraction, and present a new decoder that reassembles the translated fragments into one final translation. Results show that one of the proposed phrase-extraction methods—the one used in Moses—leads to a translation system that outperforms context-sensitive word-based approaches. The differences, however, are small, arguably because the word-based approaches already capture phrasal context implicitly due to their source-side and target-side context sensitivity.

1 Phrase-based memory based machine translation

A memory-based machine translation (MBMT) system divides into a training subsystem, producing a translation model, and a translation subsystem [3, 4, 1]. A parallel corpus is used for phrase extraction and example generation, i.e. the generation of translations of source fragments to target fragments. These fragments, with as its main constituent an aligned pair of phrases, are compressed, rather than merely stored, in the training phase. In testing, unseen source-language sentences in a test corpus are also transformed into fragments, which the memory-based classifier maps onto a distribution of target-language fragment translations. A decoder then reassembles all translated fragments together into one sentence, searching through and choosing between alternative solutions when more than a single target sentence can be built out of the predicted fragments.

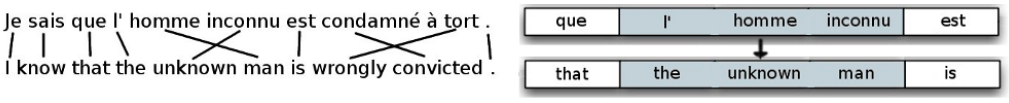

Figure 1: *Left:* A word alignment between a French and English sentence, *Right:* A phrase-based training example in context

We assume a word-alignment between all sentence pairs in the parallel corpus. Figure 1 (left) illustrates such a word-aligned sentence pair. On the basis of this, we create example fragment translations that serve as training examples. Examples are composed as follows: The feature vector consists of a phrase from the source sentence, with one context word on the left side, and one context word on the right side. The class consists of the target-language phrase that aligns to the source-language phrase.

To determine phrases in the source- and target-language sentences in the parallel corpus available for training, we make use of a phrase-translation table [2]. In addition we include two other approaches to phrase extraction for comparison. The *phrase-list approach* is a straightforward method that extracts frequent n-grams only from the source-language side of the training corpus, and stores this in a *phrase list*. When using

[1]This paper was presented at the 3rd Workshop on Example-Based Machine Translation, 12–13 November 2009, Dublin City University,Dublin, Ireland

this method, each source sentence is matched against the phrase list, and whenever a phrase is found, we follow the word-alignments from the phrase and assume that the sequence of words it points to is the aligned target phrase. The *marker-based approach* [4] segments a sentence into non-overlapping chunks, splitting whenever so-called marker words occur, i.e. closed-class function words.

The fact that we may end up with multiple examples covering the same words in the source sentence, we can speak of various possible *fragmentations* of the source sentence S that each map to different target phrases. In other words, the decoding part of the translation system needs to search heuristically in a large space of possible outcomes. The decoding procedure starts by generating an initial hypothesis: a translation hypothesis based on the segmentation of the source sentence into phrases that were translated with the highest confidence, and concatenating the respective target phrases in the source-side order. A hypothesis can be modified in two main ways: (1) the order in which the hypothesis fragments are assembled can be changed, and (2) the choice of hypothesis fragments can be changed, i.e. other hypothesis fragments with an equal or lower translation probability could be tried. Using beam search, the decoder maximizes a score function that expresses the *fidelity* and the *fluency* of the tested target hypothesis. A quantification of fluency is provided by a trigram-based statistical language model with back-off smoothing on the target language, while fidelity is expressed by the probabilities generated by the memory-based translation model. In addition, a distortion factor is added into the multiplication.

2 Results and Conclusions

The experimental results on obtained by training and testing on two benchmark translation corpora for Dutch to English translation, one of movie subtitles and another consisting of medical texts, tell us that memory-based machine translation can be extended from translating fixed-length word trigrams to translating phrases of arbitrary length. Of the three tested methods of phrase extraction, the Moses phrase-translation table approach emerges as the best solution, producing a BLEU score of 23.0 on the movie subtitles test data (versus 21.6 for the word-based version of the system that only maps trigrams of source words to target words, but uses the same decoder) and 30.7 on medical test text (versus 27.2 for the word-based approach; a previous MBMT system [1] that maps source trigrams of words to target trigrams attains a 30.1 BLEU score).

In sum, on the two test sets the phrase-based system improves over all previous memory-based approaches. Nevertheless, the impact of phrases in comparison to word-based MBMT is relatively limited. A potential explanation for this limited effect is that word-based MBMT that maps trigrams of source words to trigrams of target words can be seen as implicitly phrase-based already. The approach followed in [3, 1] implicitly capturing all phrases up to length three. Our current approach changes this only slightly by turning the source-side trigrams into variable-width examples of Moses phrases surrounded by their left and right neighboring words, and predicting variable-width target-side phrases at the output, starting from single words.

References

[1] S. Canisius and A. van den Bosch. A constraint satisfaction approach to machine translation. In *Proceedings of the 13th Annual Conference of the European Association for Machine Translation (EAMT-2009)*, pages 182–189, 2009.

[2] P. Koehn. Pharaoh: A beam search decoder for phrase-based statistical machine translation models. In R.E. Frederking and K. Taylor, editors, *Proceedings of the American Machine Translation Association*, volume 3265 of *Lecture Notes in Computer Science*, pages 115–124. Springer, 2004.

[3] A. van den Bosch and P. Berck. Memory-based machine translation and language modeling. *The Prague Bulletin of Mathematical Linguistics*, 91:17–26, 2009.

[4] A. van den Bosch, N. Stroppa, and A. Way. A memory-based classification approach to marker-based EBMT. In *Proceedings of the METIS-II Workshop on New Approaches to Machine Translation*, pages 63–72, 2007.