Juri Ganitkevitch

MT Marathon 2 - Open Source Project Proposal

A phrase-based, LM-guided decoder using long *n*-Grams as gapless skip-LMs

This proposal suggests the introduction of long *n*-Grams to score the combination of two partial hypotheses. In the following we shall present a rough sketch of the proposed approach:

We assume to receive the input on a per-sentence basis. Upon receiving a new source sentence the decoder loads a filtered binary LM. This LM should be long (5-grams or higher) and will be stored in-order as well as reversed (takes up twice the space vs. enabling backwards look-up).

During the search, when given two partial hypotheses with non-overlapping coverage vectors. Assuming that on target side we have two blocks of translated text, we generate lists of *n*-Grams matching the borders of each text block.

By cross-referencing the two lists of *n*-Grams (apply some clever algorithms & heuristics here), we obtain information on how well these phrases fit next to each other, and, specifically, in what order.

Provided we have long *n*-Grams, we may also obtain information on how to glue these phrases together. Our cross-referenced *n*-Grams should have the following structure

a b c P x y z

Where abc matches the end of block 1 and xyz matches the beginning of block 2. P may be empty (in the case the phrases supposedly do well next to one another), or may contain a few words. These words typically should be the glue words we're looking for. To avoid introducing content-bearing words we can score P against the source sentence using IBM1 (actually, scoring the LM *n*-Grams upon loading against the source sentence using IBM1 might be an interesting idea in general).

Also, the glueing principle might be used to generate possible translations for words that do not appear in the phrase dictionary, provided the language model was trained on a larger monolingual corpus that just the parallel data used for the translation models.

This general idea can be extended to an incremental decoding process, where, given a hypothesis, we try to find the next phrase to translate and glue on to our already-translated block of target sentence.

All of this, along with possible further uses for bidirectional *n*-Grams constitutes our proposal for an extension to the Moses framework.