Integration of automatically-acquired multiword expressions in a hybrid machine translation system

Will Roberts and Markus Egg
{will.roberts,markus.egg}@anglistik.hu-berlin.de
Humboldt-Universität zu Berlin

WG 3: Statistical, Hybrid and Multilingual Processing of MWEs

We present a pilot experiment to integrate information about non-compositional multiword expressions (MWEs) into a hybrid machine translation (MT) system.

A list of 2 million MWE candidates is automatically extracted from the English Wikipedia using standard association measure techniques; MWEs are modelled as n-grams of inflected text, with $n \in \{2, 3\}$ (i.e., bigrams and trigrams).

The list of MWE candidates is then re-ranked in order of increasing compositionality score; the compositionality computation broadly follows the method of Salehi et al. (2015), comparing a word embedding vector representing a MWE against the word embedding vectors of its constituent words. This re-ranking makes use of word embedding models trained on the English Wikipedia, where all instances of a given MWE have been greedily replaced with a single token representing the MWE as a word-with-spaces. MWEs which have low similarity with their constituent words are judged to be non-compositional.

The least compositional MWE candidates (ca. 600K) are then integrated into a hybrid MT system. This system, TectoMT (Žabokrtský et al., 2008), uses a pipeline of statistical and rule-based tools to analyse the source language (English) up to a high-level (tectogrammatical) dependency representation, whereby semantically empty words (e.g., function words) are removed from the tree and represented as morphosyntactic properties on the remaining (semantically relevant) word nodes. The tree structure is then duplicated in the target language (Spanish), with the node lemmas and features translated by a maximum entropy model. Another pipeline of tools then turns the target tectogrammatical representation back into surface text.

MWE information is integrated into this system by identifying instances of non-compositional MWE candidates in the parallel training corpus (ca. 25M words in 1.2M sentences) using string matching on the inflected word forms. Filtering ensures that matches are made only on "treelets" (sets of words which are fully connected by

dependency relations as output by the English statistical parser), modelling MWEs as sentence constituents which may take arguments or include modification. Matched MWE instances are then collapsed into a single words-with-spaces tree node to better reflect their semantic non-compositionality. The maximum entropy model trained on the English source text with MWEs analysed in this way tries to learn how to translate English multiwords into Spanish; because MWE analysis is performed only on the source language, this transfer will be successful where an English MWE can be translated by a single Spanish lexeme, and unsuccessful where the English MWE must be translated by a Spanish phrase or MWE. In testing, the source text (ca. 18K words in 1K sentences) is analysed in the same way as during training, with the caveat that only MWEs observed during training can be used (as only these expressions are known to the maxent transfer model).

In experiments, we observe a statistically significant improvement in translation quality (0.46 BLEU points) using the MWE analysis paradigm outlined above. We find the method is highly sensitive to the compositionality score of the MWE candidates used; only the least compositional MWEs help the MT system, and including more compositional MWEs lowers MT quality below the baseline. The method is also sensitive to domain effects: Performance improvements are only observed where the parallel training data are a good thematic match for the test set.

In future work, we will replicate our results in other European languages. We will also extend analysis to the target language, allowing the MT system to learn to translate single English lexemes and English MWEs with MWEs in the target language.

References

Bahar Salehi, Paul Cook, and Timothy Baldwin. A word embedding approach to predicting the compositionality of multiword expressions. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 977–983, 2015.

Zdeněk Žabokrtský, Jan Ptáček, and Petr Pajas. TectoMT: Highly modular MT system with tectogrammatics used as transfer layer. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 167–170. Association for Computational Linguistics, 2008.