Rich morpho-syntactic descriptors for factored machine translation with highly inflected languages as target

Alexandru Ceausu
Centre for Next Generation Localisation, Dublin City University
aceausu (at) computing (dot) dcu (dot) ie

The baseline phrase-based translation approach has limited success on translating between languages with very different syntax and morphology, especially when the translation direction is from a language with fixed word structure to a highly inflected language. There are two main points to improve on: morphological translation equivalence and long range reordering. Translating the correct surface form realization of a word is dependent not only on the source word-form, but it also depends on additional morpho-syntactic information. In addition, the rich morphology of a highly inflected language permits a flexible word order, thus making difficult to model long range word order differences between languages.

Factored translation models (Koehn and Hoang, 2007) allow the integration of the linguistic information into a phrase-based translation model. We present an experiment that uses morpho-syntactic description (MSD) codes (Erjavec, 2004) as a factor in translation. Our approach is similar to several other factored machine translation experiments, such as adding the morphological features as factors (Avramidis and Koehn, 2008), adding supertags on source language (Haque et al, 2009), and modelling syntax to morphology (Yeniterzi and Oflazer, 2010), etc.

Based on the larger JRC-Acquis corpus (Steinberger et al, 2006), the SEE-ERA.net corpus (Tufis et al, 2008) contains 1200 aligned documents in several South Slavic and Balkan languages plus Czech, English, French, and German languages. The documents have morpho-syntactic description codes for Bulgarian, Greek, English, Slovene and Romanian.

Based on this corpus, we tested in a factored framework (Koehn et al, 2007) several configurations of translation, generation and reordering steps. We found that translating lemmas and morpho-syntactic descriptors prior to generating the word-forms achieved better results than the baseline phrase-based translation model.

The language pairs tested were English-Greek, English-Bulgarian, English-Slovene and English-Romanian using the SEE-ERA.net corpus. After cleaning, we split the corpus into training, development and test sets resulting in 60000 sentence pairs for training, 500 for the development test and 1000 for testing. The 4-gram word-form language models and the 5-gram MSD language models were built only using the training data sets.

We built baseline phrase-based translation models for each language pair and tuned them on the development set using MERT (Och, 2003). We followed the same steps to build the factored translation models using our proposed configuration: (i) translating lemmas, (ii) generating the possible morpho-syntactic descriptions for a given lemma, (iii) translating the associated morpho-syntactic descriptions and (iv) generating the target surface forms given the lemmas and the morpho-syntactic descriptions. In this configuration, the decoder uses two language models: one for the word-forms and another for the morpho-syntactic description codes.

We tested the systems using the BLEU score (Papineni et al, 2002). We observed improvements in accuracy ranging between 1% for Romanian and 3% for Slovenian in absolute BLEU points. Better handling of long-distance dependencies based on the MSD language model, a robust lemma translation equivalents table and a more precise selection of morphological variants are all possible explanations for the improvement in translation accuracy.

To check how the results scale to a larger corpus, we used an English-Romanian parallel corpus of one million sentence pairs. The corpus was tokenized, lemmatized and automatically annotated with morpho-syntactic description codes. For the bigger corpus, the English to Romanian factorized system achieves a BLEU score of 43% but the differences between the baseline and the factorized system are negligible.

Although the factorized model only has a marginally increase in BLEU score accuracy and detrimental effects on translation speed, we observed that it produces translations of better word order and more accurate morphological variant selection over the baseline model. Currently, we are in the process of manually evaluating the results.

References

Eleftherios Avramidis and Philipp Koehn. 2008. Enriching morphologically poor languages for statistical machine translation. In Proceedings of ACL-08/HLT, pages 763–770, Columbus, Ohio, June

Tomaz Erjavec. 2004. MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In Proc. of the Fourth Intl. Conf. on Language Resources and Evaluation, LREC'04, pp. 1535 - 1538, ELRA, Paris

Rejwanul Haque, Sudip Kumar Naskar, Yanjun Ma & Andy Way. 2009. Using Supertags as Source Language Context in SMT. In Proceedings of the 13th Annual Meeting of the European Association for Machine Translation (EAMT-09), May 14-15, 2009, Barcelona, Spain

Philipp Koehn, and Hieu Hoang. 2007. Factored Translation Models. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 868–876, Prague, June 2007

Philipp 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, Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, Prague, Czech Republic

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002, pp. 311-318

Ralph Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiş. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In Proceedings of the 5th LREC Conference, Genoa, Italy, 22-28 May, 2006, pp.2142-2147

Dan Tufiş, Svetla Koeva, Tomaž Erjavec, Maria Gavrilidou, and Cvetana Krstev. 2008. Building Language Resources and Translation Models for Machine Translation focused on South Slavic and Balkan Languages. In Marko Tadić, Mila Dimitrova-Vulchanova and Svetla Koeva (eds.) Proceedings of the Sixth International Conference Formal Approaches to South Slavic and Balkan Languages (FASSBL 2008), pp. 145-152, Dubrovnik, Croatia, September 25-28

Reyyan Yeniterzi and Kemal Oflazer. 2010. Syntax-to-Morphology Mapping in Factored Phrase-Based Statistical Machine Translation from English to Turkish, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 454–464, Uppsala, Sweden, 11-16 July 2010