TectoMT – a deep-linguistic core of the combined Chimera MT system

Martin Popel, Roman Sudarikov, Ondřej Bojar, Rudolf Rosa, Jan Hajič

Charles University in Prague, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics, Malostranské nám. 25, CZ-11800 Prague 1, Czech Republic

{popel,sudarikov,bojar,rosa,hajic}@ufal.mff.cuni.cz

Abstract. Chimera is a machine translation system that combines the TectoMT deep-linguistic core with phrase-based MT system Moses. For English–Czech pair it also uses the Depfix post-correction system. All the components run on Unix/Linux platform and are open source (available from Perl repository CPAN and the LINDAT/CLARIN repository). The main website is https://ufal.mff.cuni.cz/tectomt. The development is currently supported by the QTLeap 7th FP project (http://qtleap.eu).

TectoMT and Chimera

TectoMT (the deep-linguistic core of Chimera) is an open-source MT system based on the Treex platform for general natural-language processing. TectoMT uses a combination of rule-based and statistical (trained) modules ("blocks" in Treex terminology), with a statistical transfer based on HMTM (Hidden Markov Tree Model) at the level of a deep, so-called tectogrammatical representation of sentence structure. In the Chimera combination, TectoMT is complemented by a Moses PB-SMT system (factored setup with additional language models over morphological tags) and optionally also by an automatic postprocessing (correction) component called Depfix. Chimera can be thus characterized as a hybrid system that combines statistical MT with deep linguistic analysis and automatic post-correction system, which is useful especially for translation into inflectionally rich languages. The three systems are combined serially: TectoMT runs first, then an additional Moses phrase table is extracted from TectoMT's input and output. The additional table is then used in a weighted combination with a large Moses translation table to produce pre-final output. Depfix then re-parses the output (as well as input) and generates the final output based on rules reflecting morphosyntactic properties of the target language.

Chimera was transferred from English–Czech to additional three language pairs (English to Dutch, Portuguese and Spanish) within the QTLeap 7th EU project.

References

Dušek, O., Gomes, L., Novák, M., Popel, M., Rosa, R. (2015). New Language Pairs in TectoMT. *Proceedings of the 10th Workshop on Machine Translation*, ISBN 978-1-941643-32-7, ACL, Stroudsburg, PA, USA, 98–104.

Rosa, R., Dušek, O., Novák, M., Popel. M. (2015). Translation Model Interpolation for Domain Adaptation in TectoMT. *Proceedings of the 1st Deep Machine Translation Workshop*, ISBN 978-80-904571-7-1, Charles University in Prague, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics, Prague, Czech Republic, 89–96.

Bojar, O., Tamchyna, A. (2015). CUNI in WMT15: Chimera Strikes Again. *Proceedings of the 10th Workshop on Machine Translation*, ISBN 978-1-941643-32-7, ACL, Stroudsburg, PA, USA, 79–83.