Building a Statistical Machine Translation Engine for Improved Automated Translation from English to Estonian

Machine translation into morphologically rich languages such as Estonian is challenge, due to problems stemming from lexical sparsity and the sheer amount of grammatical features in the grammar of such a language. This paper explores the problems and solutions of building a statistical machine translation engine from scratch for automated translation from English to Estonian, using the open-source Moses toolkit (Koehn et al 2007) and utilizing morphogen (Schlinger et al 2013), another open-source tool for improving machine translation, specifically for morphologically rich languages.

Moses is used to automatically train translation models, using text pairs in the form of parallel corpora. morphogen is implemented in two phases. First, an inflection model is learned, that is used to target word inflection based on the context. The learned model is used to produce sentence-specific translation phrases, which can be used to augment the translation grammar, thus improving the quality of the translation in the area of inflected words and phrases.

As expected, automated translation into Estonian entails a number of difficulties mainly stemming from its complex case system and the free variability of the case forms - some of which is somewhat context-dependent, and indeed not quite free. This is a work in progress, and the conclusions and solutions proposed in this paper should be taken as such.

References

Schlinger, E., Chahuneau, V., and Dyer, C. (2013). morphogen: Translation into morphologically rich languages with synthetic phrases. *The Prague Bulletin of Mathematical Linguistics*, 100, 51–62.

Koehn, P., 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, June 2007*.