SMT Approaches for Commercial Translation of Subtitles

In this presentation, we report experiments on developing statistical machine translation (SMT) systems of practical use for the professional translation of subtitles. We present results of several methods that were tested for this task, describing both positive and negative outcomes. We believe these results to be of interest for companies considering the integration of SMT in multilingual commercial systems, and researchers interested in the use of current methods for large-scale SMT systems development in a specific domain.

The work we describe is part of the SUMAT project, funded through the EU ICT Policy Support Programme (2011-2014), whose goal is to produce machine translation systems for film and TV subtitles for seven language pairs. Nine partners are involved in the project: four subtitle companies (Deluxe Digital Studios, InVision, Titelbild, Voice & Script International) and five technical partners (Athens Technology Center, CapitaTI, TextShuttle, University of Maribor and Vicomtech).

In order to integrate SMT systems into a commercially viable translation workflow, it is vital for such systems to meet quality levels that do not hinder on the post-editing experience. Previous experiments (Bywood et al. 2012) have shown that, even in cases of increased productivity for professional translators post-editing machine-translated output, the perception and use of the systems is negatively affected overall by output of poor quality. To overcome this issue and raise SMT quality, we explored several approaches, taking into account issues of training and decoding efficiency, as well as issues regarding the integration of data from different sources and domains.

The baseline SMT phrase-based systems were trained on large numbers of translated subtitles provided by the subtitling companies (between 200,000 and 2 million subtitles per language pair), using the Moses framework (Koehn et al. 2007). To improve the baselines, two sets of experiments were performed: incorporating linguistic information (including factored models in various configurations (Koehn and Hoang 2007), syntax-based statistical translation and decompounding), and development of larger models by combining in-domain and out-of-domain data via mixture-modeling and perplexity minimization techniques (Sennrich 2012). Overall, the first approach provided little to no improvement over the baselines, whereas the second one proved successful at a comparatively lower cost.

In this talk, we will describe the main experiments and their results, offering insight on the optimal balance between development costs and the requirement for better systems accuracy in professional applications.

References

Bywood L., Georgakopoulou P., Volk M. and Fishel M. (2012). What is the Productivity Gain in Machine Translation of Subtitles? Presentation at Languages & The Media conference, Berlin, Germany.

Koehn Ph. and Hoang H. (2007). Factored Translation Models. *EMNLP-CONLL 2007*, 868-876.

Koehn Ph., Hoang H., Birch A., Callison-Burch Ch., Federico M., Bertoldi N., Cowan B., Shen W., Moran Ch., Zens R., Dyer Ch., Bojar O., Constantin A., and Herbst E. (2007). Moses: open source toolkit for statistical machine translation. ACL 2007, demonstration session.

Sennrich, Rico (2012). *Perplexity minimization for translation model domain adaptation in statistical machine translation*. EACL 2012, 539-549.