Short Term Scientific Mission
Uppsala University
Department of Linguistics and Philology

Hosts: Prof. Joakim Nivre, Prof. Jörg Tiedemann
Visitor: Laura Mascarell Espuny

May 18th - May 29th

1 Introduction

Current Statistical Machine Translation (SMT) systems translate at sentence-level ignoring inter-sentential context information [Koehn, 2009]. This discourse unawareness leads to incorrect translation of words or phrases that are ambiguous in the context of the sentence. For example, the term *face* can be translated either into German *Wand* or *Gesicht*. However, knowing that it co-refers back to the compound *north face* in the previous sentence would help the SMT system to correctly disambiguate the term and translate it into *Wand*.

This issue is addressed in [Mascarell et al., 2014]. In general, the method automatically detects compounds XY (e.g. *Nordwand* can be split into *Nord* X and *Wand* Y) that in the next sentences are co-referenced back with their last constituent Y (e.g. *Wand*). Since the compound gets more context information from X, the translation of Y in the compound is more specific in that particular context. The caching method then enforces this translation to all the elements Y that co-refer to the compound, improving the correctness of the translation.

The aim of this STSM is to improve the translation of ambiguous words that co-refer to multiword expressions, with special focus on compounds. We extend the work in [Mascarell et al., 2014], so instead of always enforcing the translation of Y, the SMT system also considers other feature scores to decide whether to enforce the same translation. For example, the system does not modify the translation of Y when it results in a worse language model score. To do so, we need to include our method as a feature model, so the resulting score can compete with the ones given by other features. This approach will help to detect false positives in the automatic detection of the XY-Y pairs, increasing the overall translation quality of the SMT system.

2 Description of the work carried out

We developed our feature model in the discourse-oriented decoder Docent [Hardmeier et al., 2013], which implements a stochastic variant of the hill climbing algorithm. At every stage of the search, the decoder produces a complete translation of the whole document. The search algorithm accepts a new state (i.e. a new translation of the document), when its document score (i.e. a combination of the scores taken from the feature models included) is higher than the last accepted.

Our feature model computes for the current document translation the proportion of the total number of XY-Y pairs that use the same translation. It then takes the logarithm to put the score on the same scale as the other features models included in Docent:

$$log\left(\frac{Count(\text{same translation } Y)}{Count(XY\text{-}Y \text{ pairs})}\right) \tag{1}$$

For our experiments, we trained a German-French translation model with Moses [Koehn et al., 2007] on 285'877 sentences from the Text+Berg corpus [Bubenhofer et al., 2013]. In our Docent configuration, the output from running Moses is the initial translation of each document. The test set is a collection of 261 small documents that contain compounds XY and their coreferences Y, randomly sampled from Text+Berg. We automatically annotated these XY-Y pairs in the test set using the MMAX2 annotation tool [Müller and Strube, 2006].

We intend to continue the experiments carried out during the STSM. Specifically, we plan to run the experiments using different weights for our feature, to see how they affect the MT quality.

References

[Bubenhofer et al., 2013] Bubenhofer, N., Volk, M., Klaper, D., Weibel, M., and Wüest, D. (2013). Text+Berg-korpus (release 147_v03). Digitale Edition des Jahrbuch des SAC 1864-1923, Echo des Alpes 1872-1924 und Die Alpen 1925-2011.

[Hardmeier et al., 2013] Hardmeier, C., Stymne, S., Tiedemann, J., and Nivre, J. (2013). Docent: A document-level decoder for phrase-based statistical machine translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 193–198, Sofia, Bulgaria. Association for Computational Linguistics.

[Koehn, 2009] Koehn, P. (2009). *Statistical Machine Translation*. Cambridge University Press, New York, NY, USA.

[Koehn et al., 2007] Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al. (2007). Moses: Open source toolkit for statistical machine translation. In

Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions, pages 177–180. Association for Computational Linguistics.

[Mascarell et al., 2014] Mascarell, L., Fishel, M., Korchagina, N., and Volk, M. (2014). Enforcing consistent translation of german compound coreferences. In *Proceedings of the 12th Konvens Conference*.

[Müller and Strube, 2006] Müller, C. and Strube, M. (2006). Multi-level annotation of linguistic data with mmax2. *Corpus technology and language pedagogy: New resources, new tools, new methods*, 3:197–214.