Compounds, Coreferences and Multiword Translations

Martin Volk, Laura Mascarell, Mark Fishel
University of Zurich
Institute of Computational Linguistics
volk@cl.uzh.ch

Current Machine Translation (MT) systems translate each sentence in isolation. However, many ambiguities can only be resolved by accessing the surrounding context. This is most obvious for the translation of pronouns. For example, when translating from English into German, the English pronoun *it* must be translated as either *er, sie* or *es* depending on the grammatical gender of the antecedent.

A similar issue arises with compounds and their co-refering elements in subsequent sentences. For example, if a sentence in a mountaineering text contains the compound *north face*, and the following sentence contains only the noun *face* as back reference, then knowing the antecedent will help to correctly translate *face* into German *Wand* rather than *Gesicht*.

In general terms we are considering situations where a sentence contains a compound XY, and the following sentence only contains the last element Y of this compound as a co-reference. Then if the element Y is ambiguous, it is obviously helpful to know the antecedent (which has more context given by X) in order to correctly translate Y.

We are dealing here with compounds as special cases of multiword expressions. Compounds may be written as joint words as in German or Swedish, or they may be written as separate words as in English or the Romance languages.

We have investigated this phenomenon in detail for the translation direction German to French. This means that we have to split the German compounds in order to find their co-refering elements. We have applied the morphological analyzer Gertwol for compound splitting, but it could be substituted by any system that does dynamic de-compounding.

In [Mascarell et al., 2014] we showed that our method improves the correctness in the automatic translation of ambiguous Y, by enforcing a more specific term in that particular context. For example, whilst the most likely translation of *Fahrt* is *course*, our method enforces *ascension* when it co-references back to *Bergfahrten*. In a corpus of 5 million tokens [Bubenhofer et al., 2014] we automatically detected 24,317 instances of a compound co-referenced by its last element. This work is based on the one-translation-per-discourse hypothesis [Carpuat, 2009], which claims that terms do not vary their meaning across a document, and therefore they should be translated consistently.

In order to find the corresponding element for Y in the target language, we need to identify the translation of the segment Y in a compound XY. To do so, we check which word from the compound translation appears as a translation candidate of Y in the phrase table of our statistical MT system. We then apply a caching technique, which remembers the translation of

the last element of a compound and propagates it to its subsequent co-references. For example, since the translation of the German word *Amt* in the compound *Bundesamt* is French *office*, its co-references are translated into *office* instead of the more frequent option *poste*.

To evaluate the correctness of the results obtained from the experiments, we performed a manual analysis of the translations enforced by our method. In a test set containing 318 pairs of a compound XY and a co-reference Y randomly sampled, our method improves 17 translations of Y, rising the correctness from 80.1% to 86.7% and the consistency by 24.1% points. In those cases where Y is not ambiguous, both the non-enforced and the enforced translation are correct. For example, in the mountaineering domain the German word *Wand* can be correctly translated into French *face* or *paroi* regardless of the antecedent translation.

Our efforts are in line with other approaches to include discourse features to increase the quality of SMT systems (cf. [Hardmeier et al., 2012], [Meyer, 2014]), but we focus on compounds and their co-refering elements as a special case of multiword translation.

We have argued that compounds and their referring elements pose a special challenge for MT systems. We have shown that crossing the sentence boundary helps the systems to improve the correct and consistent translation of co-refering elements.

References

[Bubenhofer et al., 2014] Bubenhofer, N., Volk, M., Leuenberger, F., Weibel, M., and Wüest, D., editors (2014). *Text+Berg-Korpus (Release 149 v01). Digitale Edition des Jahrbuch des SAC 1864-1923, Echo des Alpes 1872-1924 und Die Alpen, Les Alpes, Le Alpi 1925-2013*. Institut für Computerlinguistik, Universität Zürich.

[Carpuat, 2009] Carpuat, M. (2009). One translation per discourse. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, pages 19–27, Boulder, Colorado.

[Hardmeier et al., 2012] Hardmeier, C., Nivre, J., and Tiedemann, J. (2012). Document-wide decoding for phrase-based statistical machine translation. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1179–1190, Jeju Island, Korea.

[Mascarell et al., 2014] Mascarell, L., Fishel, M., Korchagina, N., and Volk, M. (2014). Enforcing consistent translation of German compound coreferences. In *Proceedings of KONVENS*, Hildesheim.

[Meyer, 2014] Meyer, T. (2014). *Discourse-level features for statistical machine translation*. Phd thesis, Idiap and EPFL, Lausanne.