Linguistics, German Compounds and Statistical Machine Translation. Can they all get along?

Carla Parra Escartín
University of Bergen
Bergen, Norway
carla.parra@uib.no

Stephan Peitz
RWTH Aachen University
Aachen, Germany
peitz@cs.rwth-aachen.de

Hermann Ney
RWTH Aachen University
Aachen, Germany
ney@cs.rwth-aachen.de

Here, we report different experiments created to study the impact of using linguistics to preprocess German compounds prior to translation in Statistical Machine Translation (SMT). German compounds are a known challenge both in Machine Translation (MT) and Translation in general as well as in other Natural Language Processing (NLP) applications. In the case of SMT, this challenge is usually overcome by splitting the compounds into their constituents to decrease the number of unknown words and improve the results of evaluation measures like the BLEU score (Papineni et al., 2001). When translating from German into Romance languages such as Spanish, a further challenge is faced as German compounds are translated as phraseological units.

Examples 1 and 2 show the splittings of the German compounds *Warmwasserbereitung* and *Wärmerückgewinnungssysteme* and their translations into English and Spanish.

(1)

Warm	*Wasser*	*Bereitung*
caliente	agua	preparación
warm	water	production

[ES]: 'Preparación de agua caliente'
[EN]: 'Warm water production'

(2)

Wärme	*Rückgewinnung*	*s*	*Systeme*
calor	recuperación	Ø	sistemas
heat	recovery	Ø	Systems

[ES]: 'sistemas de recuperación de calor'
[EN]: 'heat recovery systems'

As may be observed in 1 and 2, in Spanish not only there is word reordering, but also there is usage of other word categories such as prepositions. While the examples above are quite simple, the work done by researchers such as Angele (1992); Gómez Pérez (2001) and Oster (2003) for the pair of languages German→Spanish shows that the translational equivalences in Spanish not only are very varied, but also unpredictable to a certain extent.

To assess to which extent it is necessary to deal with German compounds as a part of preprocessing in SMT systems, we have tested two different compound splitters and strategies, such as adding lists of compounds and their translations to the training set. Concretely, we have used the the state-of-the-art splitter implementation by Popović et al. (2006) and the splitter developed by Weller and Heid (2012).

Two corpora have been used: the TRIS corpus and the Europarl corpus for German→Spanish. In the case of TRIS only the subcorpus corresponding to the construction domain has been used.

The results obtained in our experiments seem to indicate that the manually compiled list of compounds added to the training corpus helped to increase the probabilities of alignment of 1:n correspondences and thus the compound translations in the phrase tables are better.

References

Angele, S. (1992). *Nominalkomposita des Deutschen und ihre Entsprechungen im Spanischen. Eine kontrastive Untersuchung anhand von Texten aus Wirtschaft und Literatur.* München: iudicium verlag GmbH.

Gómez Pérez, C. (2001). *La composición nominal alemana desde la perspectiva textual: El compuesto nominal como dificultad de traducción del alemán al español.* Ph. D. thesis, Departamento de Traducción e Interpretación, Universidad de Salamanca, Salamanca.

Oster, U. (2003). *Los términos de la cerámica en alemán y en español. Análisis semántico orientado a la traducción de los compuestos nominales alemanes.* Ph. D. thesis, Departament de Traducció i Comunicació, Universitat Jaume I, Castellón.

Papineni, K., S. Roukos, T. Ward, and W.-J. Zhu (2001, September). Bleu: a Method for Automatic Evaluation of Machine Translation. IBM Research Report RC22176 (W0109-022), IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, NY 10598.

Popović, M., D. Stein, and H. Ney (2006). Statistical machine translation of german compound words. In *Proceedings of the 5th international conference on Advances in Natural Language Processing*, FinTAL'06, Berlin, Heidelberg, pp. 616–624. Springer-Verlag.

Weller, M. and U. Heid (2012, May). Analyzing and Aligning German compound nouns. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association.