Compound dictionary extraction and WordNet. A dangerous liaison?

Carla Parra Escartín
University of Bergen
Bergen, Norway

carla.parra@uib.no

Héctor Martínez Alonso
University of Copenhagen
Copenhagen, Dennmark

alonso@hum.ku.dk

In this work we present current work being done to explore ways of automatically retrieving compound dictionaries from sentence-aligned corpora using WordNet. More concretely, we focus on the pair of languages German→Spanish.

As Sag et al. (2001) argue in their seminal "pain in the neck" paper, Multiword Expressions (MWEs) are a major bottleneck for many Natural Language Processing (NLP) applications. Within the MWE community, a lot of efforts has been made to automatically extract MWEs. Ramisch (2012) offers an overview of the state of the art with respect to computational methods for MWE treatment. As Ramisch (2012) also acknowledges, in the last years there has been a shift in the MWE research, and now researchers not only focus on their identification and extraction, but also in integrating them in applications. Our research had as a starting point a real problem for human translation and machine translation (MT), and therefore is application-driven. Although we focus on compound dictionary extraction, the ultimate aim is to integrate them in Statistical Machine Translation (SMT) tasks.

As agreed by researchers in the MWE field (Moon, 1998; Cowie, 1998; Sag et al., 2001; Baldwin and Kim, 2010, etc.), one of the main issues when dealing with MWEs is that their typology has not been widely agreed upon. As regards to German compounds, they could be considered MWEs written together in one typographic word. Example 1 below exemplifies this showing the inner structure of the German compound "*Straßenlampe*" and its translation into English.

(1) *Straße en Lampe*
street Ø lamp/light
[EN]: 'street lamp // street light'

Although "*Straßenlampe*" would usually be considered a nominal compound but not a MWE, its possible translational correspondences in English are nominal compounds *and* MWEs. However, if we split the German compound into its component words as in 1, we can see that in fact it is formed by the nouns "*Straße*" and "*Lamp*" joint together with the filling element "*en*". Moreover, the translations of German nominal compounds usually correspond to phrases in the target languages, and those correspondences would in turn be MWEs. This is further illustrated in Example 2, where the inner structure of the German compound "*Warmwasserbereitungsanlagen*" is shown together with its correspondences in English and Spanish.

(2) *Warm* *Wasser* *Bereitung* *s* *Anlagen*
warm water production Ø systems
caliente agua producción Ø sistemas
[EN]: 'Warm water production systems'
[ES]: 'Sistemas de producción de agua caliente'

Based on the arguments give above, we treat German compounds and their translations as a special MWE problem.

Our working hypothesis is that the different formants of a compositional compound will share semantic features with their corresponding translational equivalents in other languages. Based on this hypothesis, we are currently running a pilot experiment to verify whether it holds true. Thus, we use German as a source language and Spanish as the target language.

Our pilot experiment consists on semantically tagging the formants of the compounds in German and their Spanish translations using WordNet, and trying to find possible overlappings across languages. To run this experiment, we have created a Gold Standard consisting of German compounds and their Spanish translations. The data has been extracted from from a 261-sentence subset of the TRIS corpus (Parra Escartín, 2012).

We expect to be able to align the split German compound with the Spanish MWE by finding a correlation between the semantic types of their formants.

If the results are positive (i.e. our hypothesis holds), we will run experiments to automatically extract compound dictionaries from aligned corpora, and these will be then used in SMT experiments. Furthermore, if our approach is successful, similar experiments could be carried out with other types of compositional MWEs.

References

Baldwin, T. and S. N. Kim (2010). Multiword Expressions. In N. Indurkhya and F. J. Damerau (Eds.), *Handbook of Natural Language Processing, Second Edition*. Boca Raton, FL: CRC Press, Taylor and Francis Group. ISBN 978-1420085921.

Cowie, A. P. (1998). *Phraseology: Theory, Analysis, and Applications: Theory, Analysis, and Applications*. Clarendon Press.

Moon, R. (1998). *Fixed Expressions and Idioms in English: A Corpus-Based Approach*. Oxford University Press.

Parra Escartín, C. (2012, May). Design and compilation of a specialized Spanish-German parallel corpus. In N. C. C. Calzolari, K. Choukri, T. Declerck, M. Uğur Doğan, B. Maegaard, J. Mariani, J. Odijk, and S. Piperidis (Eds.), *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, pp. 2199–2206. European Language Resources Association (ELRA).

Ramisch, C. (2012, September). *A generic and open framework for multiword expressions treatment: from acquisition to applications*. Ph. D. thesis, University of Grenoble (France) and Federal University of Rio Grande do Sul (Brazil), Grenoble, France.

Sag, I. A., T. Baldwin, F. Bond, A. Copestake, and D. Flickinger (2001). Multiword Expressions: A Pain in hte Neck for NLP. In *In Proc. of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*, pp. 1–15.