Understanding compound words: a new perspective from compositional systems in distributional semantics

Marco Marelli
Ghent University
marco.marelli@ugent.be

In the present work I discuss CAOSS (Compounding as Abstract Processes in Semantic Space), a model that aims at capturing the semantic dynamics of compound processing in a data-driven framework.

In CAOSS, word meanings are represented as vectors encoding lexical co-occurrences in a reference corpus (e.g., the meaning of *snow* will be based on how often *snow* appears with other words), according to the tenets of distributional semantics (e.g., Landauer & Dumais, 1997). A combinatorial procedure is induced following Guevara (2010): given two vectors (constituent words) u and v, their composed representation (the compound) can be computed as $c = M * u + H * v$, where M and H are weight matrices estimated from corpus examples. The matrices are trained using least squares regression, having the vectors of the constituents as independent words (*car* and *wash*, *rail* and *way*) as inputs and the vectors of example compounds (*carwash*, *railway*) as outputs, so that the similarity between $M*u+H*v$ and c is maximized. In other words, the matrices are defined in order to recreate the compound examples as accurately as possible. Once the two weight matrices are estimated, they can be applied to any word pair in order to obtain a meaning representation for their combination.

CAOSS is shown to correctly predict effects related to the processing of novel compounds, and in particular the impact of relational information. Moreover, model predictions are useful for the comprehension of the role of semantic transparency in the processing of familiar compounds. Taken together, the model simulations indicate that a compositional perspective on compound-word meaning is crucial for understading the processing of both novel and familiar combinations.

References: • Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104(2)*, 211. • Guevara, E. (2010). A regression model of adjective-noun compositionality in distributional semantics. In *Proceedings of the 2010 Workshop on GEometrical Models of Natural Language Semantics* (pp. 33-37). Association for Computational Linguistics.