Understanding compound words: a new perspective based on compositional distributional semantics

Marco Marelli, University of Milano-Bicocca

In the present work I discuss CAOSS (Compounding as Abstract Operations 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 the 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 meaning representations for untrained word combinations (e.g., "snow building").

Model predictions are tested against psycholinguistic data obtained from the conceptual combination and word processing literature. CAOSS is shown to mirror evidence related to the processing of novel compounds, and in particular the impact of relational information: CAOSS simulations correctly predict both relational priming effects (Gagné, 2001) and relational dominance effects (Gagné & Shoben, 1997) on behavioral data. Moreover, model predictions are proved to be useful for understanding the impact of semantic transparency in the processing of familiar compounds: CAOSS-based estimates play a central role in explaining semantic effects in compound processing, when examining both lexical decision latencies (Balota et al., 2007) and fixation times during text reading (Cop et al., 2016).

The model simulations indicate that compositionality-related phenomena are reflected in language statistics. Human speakers are able to learn these aspects from language experience and automatically apply them to the processing of any word combination. The present model is flexible enough to emulate this procedure, predicting sensible relational similarities and correctly capturing the contribution to semantic transparency provided by compositional operations. Taken together, the model simulations indicate that a compositional perspective on compound-word meaning is crucial for understanding the processing of both novel and familiar combinations.