Investigating the Representation of Semantic Relations of Chinese Noun-Noun Compound in Transformer-based Language Models

He Zhou, Emmanuele Chersoni and Yu-Yin Hsu

Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University

A noun-noun compound consists of two nouns but functions as a single noun that denotes an entity. However, the semantic relation between the two nouns can be significantly varied, as the meaning of a noun-noun compound is not merely the combination of the meanings of each part. For example, in *aiqing*$_{n_1}$*gushi*$_{n_2}$, where *aiqing* (*love*) modifies *gushi* (*story*), it means *a story that is ABOUT love*, while in *minjian*$_{n_1}$*gushi*$_{n_2}$, *minjian* (*folk*) modifies *gushi* (*story*) denoting *a story that is FROM the folk*. The semantic relations between the modifier and the head noun differ, even though both compounds share the same head noun. In this study, we investigate the semantic compositionality of Chinese noun-noun compounds from a computational linguistics perspective. Specifically, we examine how the Transformer-based language models represent the latent semantic relation knowledge existing within Chinese noun-noun compounds.

To initiate our study, we constructed a dataset of Chinese noun-noun compounds with semantic relation annotations. The annotation process involved two steps: first, identifying noun-noun compounds from nouns, and second, identifying semantic relation(s) between the modifier and the head nouns. Consequently, the dataset comprises of 2,083 noun-noun compounds, each labeled with one or more possible semantic relations.

To explore how the Transformers-based language models process noun-noun compounds and encode the semantic relations, we constructed featuring groups of compounds with shared lexical or semantic features. We then examined whether representations extracted from the language models can differentiate between noun-noun compounds based on whether they share the same semantic relation. We found: (1) Transformer-based language models do encode semantic relations within compounds, but to a limited extent, with this encoding being more pronounced in the middle layers of models. (2) Regarding compound representations, despite that using the mean-pooled embeddings of the modifier and the head noun encoded relatively more semantic relation information, encoder-only models such as BERT, RoBERTa, and multilingual BERT tend to encode more such information using solely the modifier noun. On the contrary, decoder-only models such as LLaMA3, Qwen3, and Deepseek-llm tend to encode more using solely the head noun. This can be attributed to the architecture differences between the two types of models. (3) Taking a closer look at encodings for different semantic relations, the models can capture more semantic relation information that denotes agentive actions such as MAKE. They also performed modestly with the purposive relation FOR and the locative relation IN. However, their capability to encode the more general relation ABOUT and the essive relation BE is comparatively limited.