How Is a "Kitchen Chair" like a "Farm Horse"? Exploring the Representation of Noun-Noun Compound Semantics in Transformer-based Language Models

Mark Ormerod
Queen's University Belfast
mormerod01@qub.ac.uk

Jesús Martínez del Rincón
Queen's University Belfast

Barry Devereux
Queen's University Belfast

Despite the success of Transformer-based language models in a wide variety of natural language processing tasks, our understanding of how these models process a given input in order to represent task-relevant information remains incomplete. In this work, we focus on semantic composition and examine how Transformer-based language models represent semantic information related to the meaning of English noun-noun compounds. We probe Transformer-based language models for their knowledge of the thematic relations that link the head nouns and modifier words of compounds (e.g., KITCHEN CHAIR*: a chair* located in *a kitchen). Firstly, using a dataset featuring groups of compounds with shared lexical or semantic features, we find that token representations of six Transformer-based language models distinguish between pairs of compounds based on whether they use the same thematic relation. Secondly, we utilize fine-grained vector representations of compound semantics derived from human annotations, and find that token vectors from several models elicit a strong signal of the semantic relations used in the compounds. In a novel "compositional probe" setting, where we compare the semantic relation signal in mean-pooled token vectors of compounds to mean-pooled token vectors when the two constituent words appear in separate sentences, we find that the Transformer-based language models that best represent the semantics of noun-noun compounds also do so substantially better than in the control condition where the two constituent works are processed separately. Overall, our results shed light on the ability of Transformer-based language models to support compositional semantic processes in representing the meaning of noun-noun compounds.*

Action Editor: Kevin Duh. Submission received: 21 September 2022; revised version received: 29 April 2023; accepted for publication: 17 June 2023.

https://doi.org/10.1162/coli_a_00495

1. Introduction

One rapidly growing strand of Natural Language Processing (NLP) research is that of determining whether neural language models encode certain linguistic properties, and, if so, understanding how these properties are represented. This goal has brought together a variety of researchers and analysis techniques from fields such as machine learning, linguistics, psychology, and neuroscience (Linzen 2019; Abnar et al. 2019; Gauthier and Levy 2019; Anderson et al. 2021). With the advent of Transformer-based language models (Vaswani et al. 2017) such as BERT (Devlin et al. 2018), there has been a surge of interpretability research on this type of architecture (in particular, the study of BERT and related models has gained much popularity in a wave of research sometimes referred to as "BERTology" [Rogers, Kovaleva, and Rumshisky 2021]). To date, however, most of the research into the interpretation of Transformer-based language models has focused on their syntactic knowledge, and while there have been investigations into their semantic capabilities (e.g., Ettinger 2020; Tenney, Das, and Pavlick 2019), our understanding of how Transformers process semantic information remains largely incomplete. In contrast to syntax, where explicit representations of the grammatical structures of interest are available, a challenge faced in probing Transformer-based language models for semantics is finding suitable experimental frameworks for investigating processes relating to semantic representation and semantic composition.

In this work, we examine the extent to which Transformer-based language models have implicit knowledge of the thematic relations used in noun-noun compounds and explore how this information is encoded in the intermediary vector representations of these models. To this end, we perform layer-wise representational analysis on six different types of Transformer-based language models, covering a range of training objectives, training data, and total number of parameters.

1.1 Noun-noun Compounds and Semantic Composition

Noun-noun compounds are simple two-word phrases made up of a head noun that is modified by a modifier word. For example:

1. PUBLIC HOUSE
2. BRICK HOUSE
3. COUNTRY HOUSE

Despite their simple and consistent syntax, the meaning of the two words in a compound can combine to form a meaning for the phrase as a whole in semantically diverse ways. An interesting feature of noun-noun compounds is that, despite the semantic relation between the head noun and modifier word not being explicitly present in the phrase, their meaning is usually completely transparent to humans, even when the compound is a novel construction (van Jaarsveld and Rattink 1988). Following linguistic analysis of such compounds, we can describe their meaning with a taxonomy of *thematic relations*; that is, PUBLIC HOUSE is a house *for* the public, a BRICK HOUSE is a house *made of* brick, and a COUNTRY HOUSE is a house *located in* the country (Levi 1978; Gagné and Shoben 1997). Other approaches to compound taxonomies include Lees Robert (1960)

(an early proponent of the idea that there are a fixed number of relations for a particular head-modifier word combination), Downing (1977) (who in contrast emphasizes that the relation that describes a compound can take on any interpretation and is determined pragmatically), and more recent work such as Tratz and Hovy (2010), who create a novel taxonomic inventory by integrating several previous schemes. An alternative approach to using a taxonomy of thematic relations to interpret compounds is the dimension-based approach (Murphy 1988), which views the head noun as a schema defining a set of dimensions, each with a set of possible values. In this view, the modifier word will then fill one of these dimensions during the process of conceptual combination (Gagné and Shoben 1997). The ability of a computational model to relate the features of the two constituent words of the larger expression such that the properties of the resulting compound representation correlate with human judgment values would constitute a demonstration of semantic compositionality (Mitchell and Lapata 2010), although we would not expect the operations that would enable such a process in neural networks to be encoded in a set of systematic rules (Baroni 2020). In any case, we consider noun-noun compounds to be well-suited for investigating semantic representation and conceptual composition in Transformer-based language models—indeed, in the psycholinguistics literature, the interpretation of noun-noun compounds has proven to be a lively research area for both theories of concept representation and conceptual composition (Gagné and Shoben 1997; Murphy 2002; Estes and Hasson 2004; Devereux and Costello 2012; Lynott and Connell 2010; Maguire et al. 2007; Westerlund and Pylkkänen 2017).

In the NLP context, work on the computational interpretation of noun-noun compounds has involved classifying noun-noun compound semantic relations using a variety of features, such as semantic class information and various syntactic features (Girju et al. 2004), and lexical similarity and co-occurrence information (Ó Séaghdha and Copestake 2007; Devereux and Costello 2005). Subsequent work by Tratz and Hovy (2010) used surface features of word forms for automatically interpreting noun-noun compounds. Work by Reddy, McCarthy, and Manandhar (2011) found evidence that distributional word-space models can predict human compositionality judgments of noun-noun compounds. Other innovations in noun-noun compound interpretation include utilizing paraphrase models (Shwartz and Dagan 2018), using transfer learning and multi-task learning (Fares, Oepen, and Velldal 2018), or framing this problem as a verb paraphrasing task (Nakov 2019). More recently, Shwartz and Dagan (2019) demonstrated the power of using contextualized word embeddings (including Transformer representations) for noun-noun compound relation classification. While previous authors have generally aimed at using state-of-the-art NLP models and machine learning techniques to explore the limits of noun-noun compound relation classification, we use noun-noun compounds as a means of interpreting how Transformer-based language models build representations of the semantic relationships that exist between the constituent words.

1.2 Representational Similarity Analysis

Our analyses make use of Representational Similarity Analysis (RSA), a multivariate statistical methodology first developed in imaging neuroscience (Kriegeskorte, Mur, and Bandettini 2008). RSA allows for the comparison of different kinds of multivariate data, enabling us to compare representational vectors with both different dimensionalities (e.g., comparing two models with different hidden vector sizes) and wholly