- Original BERT paper (~28K citations): https://arxiv.org/abs/1810.04805v2

In order to derive one conceptor from a certain concept expressed in text, we first derive a set of word embeddings representing the various occurences of a concept in text. In order to obtain multiple word embeddings, rather than a single one, for each individual concept, we employ *contextual* word embeddings. This type of embeddings only captures the semantics of one particular occurence of a concept in text (i.e. in a specific phrase, sentence, paragraph, etc.). In concept theory jargon, contextual embeddings denote exemplars, rather than prototypes. In computational linguistics jargon, contextual embeddings denote tokens, rather than types. This is in stark contrast to non-contextual word embeddings, which consistently map a given word to a single prototypical embedding.

The model we use for embedding concepts is a pre-trained BERT, due to its fitting architecture and large body of related work. For completeness, BERT is an unsupervised model whose training objective includes the task of reconstructing short texts which have been previously corrupted by masking (i.e. omitting) random tokens. This is conceptually similar to denoising autoencoders, in that the task of recovery forces the model to internalize relevant input representations. However, due to the particularities of BERT's architecture (i.e. self-attention layers), the model manages to learn representations of individual tokens expressed in the input text, rather than only an aggregated representation of the complete input text as a whole. This makes BERT and related models fitting candidates as sources of contextual word embeddings for defining conceptors.

We also note that BERT and other transformer models tend to operate with sub-word tokens which are often closer in size to morphemes than to words. Assuming that discrete concepts can be more adequately represented by words compared to morphemes, we average sub-word embeddings into word embeddings.