

# The semantics of meaning: distributional approaches for studying philosophical text

Francois Meyer  
University of Amsterdam  
francoismeyer@gmail.com

Yvette Oortwijn  
University of Amsterdam  
yvette.oortwijn@gmail.com

Pia Sommerauer  
Vrije Universiteit Amsterdam  
pia.sommerauer@vu.nl

Jelke Bloem  
University of Amsterdam  
j.bloem@uva.nl

Arianna Betti  
University of Amsterdam  
a.betti@uva.nl

Antske Fokkens  
Vrije Universiteit Amsterdam  
antske.fokkens@vu.nl

## ABSTRACT

Distributional semantic models have risen to prominence in Natural Language Processing (NLP). Subsequently they have also been used in the digital humanities for studying conceptual change. This paper investigates the use of distributional models in philosophy. We propose a methodology for testing whether the models can be used for philosophical analysis. The methodology includes constructing a ground truth for philosophical terms, tuning distributional models for small data, learning embeddings for philosophical terms, and evaluating the learned embeddings using the constructed ground truth. We present results obtained with this methodology and show that Nonce2Vec, a model designed to operate on small text corpora, outperforms the more established Word2Vec in our evaluation framework. We also discuss some of the issues and potential pitfalls of applying distributional models to philosophical analysis.

## 1 INTRODUCTION

In the analysis of philosophical texts, philosophers are concerned with understanding and delineating the precise meaning of certain concepts in a given text, as well as the relations between them. Doing this manually requires close reading of a text and an appreciation of the subtleties involved in the philosophical concepts under discussion. This is a time-consuming method of research and makes the study of large corpora difficult. For this reason, being able to support philosophical research with computational tools would be a valuable addition to the field.

In this paper we investigate the possibility of studying conceptual meaning through distributional semantic methods. Distributional semantics is based on the idea that the meaning of a word can be derived from the contexts in which the word appears [4]. There are many techniques that try to leverage this idea by constructing vectorial representations for words that are based on the contexts in which these words occur in a text corpus. So-called predictive models have been very successful in constructing representations that capture the semantic and syntactic properties of words [1]. Predictive models learn representations, which are known as word embeddings, that can be used to predict the contexts in which a word occurs. Word2Vec [18, 19] has emerged as the most popular of these models. It relies on large corpora since many example contexts are required to construct a semantically representative

vector for a word. In this project we aim to determine whether such models can be sensibly applied to philosophical texts to learn representations for technical terms based on very few occurrences of the terms in a philosophical text.

Since distributional semantic models are designed to learn from very large text corpora, applying them to shorter philosophical texts presents a number of challenges. Models like Word2Vec are usually trained on corpora of at least tens of millions of words, while technical philosophical terms occur in the order of tens or hundreds of times in philosophical texts. To deal with such challenges we will consider different distributional semantic models and compare how well they handle philosophical texts. While most models require large corpora for training, models like Nonce2Vec [11] have been designed to learn high-quality embeddings from smaller corpora. Working with small text corpora is a significant challenge in NLP. There are many use cases, such as low-resource languages and specialized domains, for which large corpora are not available. The philosophical use case presents an opportunity to analyze the performance of distributional semantic models in a small data setting.

To evaluate these models, we will perform experiments using a corpus made up of all philosophical text written by the author Willard Van Orman Quine, consisting of 228 articles, books and bundles preprocessed in such a way that they are ready for digital analysis. These experiments will be assessed by philosophical knowledge of the corpus and its concepts.

The overarching goal of this project is to determine whether distributional semantic models can be used in the analysis of philosophical texts and, more broadly, for working with small, specialized data. Specifically, we are interested in whether the precise meaning of a concept in a philosophical text can be represented by these models. Furthermore, we want to determine whether the subtle differences between the natural language meaning of a concept and the philosophically relevant meaning of a concept can be captured. This project is highly methodological in nature; the objective is to explore the use of a novel methodology (outlined in Section 3) in the history of ideas.

We apply this methodology to the Quine corpus and learn embeddings for technical philosophical terms with Word2Vec and Nonce2Vec. In Section 4 we present the results obtained at different stages of the methodology and finally compare

Word2Vec and Nonce2Vec using our novel evaluation framework. We show that Nonce2Vec outperforms Word2Vec, indicating that the embeddings learned by Nonce2Vec reflect the constructed philosophical ground truth better than those learned by Word2Vec.

## 2 RELATED WORK

Word embeddings have played a significant role in many of the recent performance gains in Natural Language Processing. Using word embeddings to represent words has been shown to improve performance in many standard NLP tasks [25].

Word embeddings have been used in the digital humanities. Diachronic word embeddings have been used to track and analyse changes in the meaning of words over periods of time [7, 8, 14–16]. [12] showed that distributional semantics can be a useful tool in the analysis of discursive trends. Although these results have been promising, [9, 10] showed that the word representations learned for such analyses display random fluctuations. This is largely because they are learned from smaller text corpora than the models are designed to learn from.

[11] proposed Nonce2Vec, a modification of Word2Vec explicitly designed to learn from smaller text corpora. Nonce2Vec is able to quickly learn a high-quality embedding for a target word by modifying Word2Vec’s embedding initialization and training procedures. It initializes the embedding of the target word as the sum of the embeddings of all the known words in the target word’s contexts. During training it only adjusts the embedding of the target word, keeping the embeddings of all other words constant, while also using training parameter values that ensure faster learning.

Aside from results on small data in general, it has also been shown that the application of computational tools to corpora in philosophy can be a valuable contribution to philosophical research [5, 24].

## 3 METHODOLOGY

An important contribution of distributional models could be to automatically discover shifts in meaning between different corpora (e.g. by different authors or by the same author but written in different time periods). As suggested by [2], this type of investigation should follow a conceptual model approach. To establish whether this could be done reliably, it is necessary to first evaluate how well distributional models capture subtle aspects of meaning. However, before we can use these computational methods in philosophical research at all, we need to determine whether the methods are reliable. In order to see how, and to what extent, computational methods can be used to understand and analyze philosophical data, we developed methods and experiments to test the effectiveness of the distributional semantic models. The overall approach consists of five parts:

- (1) Defining precisely what we want the models to learn from the philosophical data.
- (2) Constructing a conceptual network of technical terms used by Quine that represents their meanings in Quine’s work.

- (3) Tuning the hyperparameters of the distributional semantic models that will be used to learn embeddings.
- (4) Learning embeddings for the technical terms in the conceptual network by applying the tuned models to the Quine corpus.
- (5) Evaluating how well these models capture the philosophically relevant meanings of the technical terms.

### 3.1 Defining the task

In order to evaluate the reliability of distribution models, we need to test how well these models are able to capture the subtle aspects of the meaning of philosophical terms. Many terms in philosophy are used in a similar way as they are in natural language, but often the meaning of these terms in philosophy is more specific and technical than in natural language. To evaluate whether these precise meanings are captured, we need to evaluate how well the embeddings reflect the meaning of the terms based on how the closeness of embeddings in the vectorial space.

A problem for this task is that it is not clear what the relation “closeness in vectorial space” captures. Therefore, the task is to classify relations between terms in the dataset as precisely as possible and compare these relations with the output of the distributional model. In order to do this we need to establish a ground truth that not only identifies whether or not terms are related but in what way terms are related. An example of a specific way in which terms might be related is by being near synonyms, in this case the terms will be similar in meaning. If the distribution model identifies many terms as being close in vectorial space if they are similar in meaning, then the sense of relatedness that being close in vectorial space captures could be partially identified with similarity in meaning.

Therefore, the task is to create an evaluation data set that specifies different ways in which terms in the data set are related to each other and then evaluate whether the distributional model outputs terms as being close in vectorial space that are related by the same ways of being related to each other. The focus will be on a specific form of relatedness that will be discussed in section 3.2, conceptual relatedness.

### 3.2 Constructing the conceptual network

The most common way to evaluate distributional semantic models of data, is to establish and compare the output to a *ground truth*. However, constructing this is not trivial for philosophical texts (see e.g. [23]).

In our case, we choose to base the construction of a ground truth on one of the most important works in the corpus, Word and Object [20]. This work is the most natural choice for the construction of evaluation data, since it encompasses many of the terms and themes that Quine discusses throughout his work. To obtain the most important terminology from this work, we took terms from the index of Word and Object, as the most important terminology is likely to be listed there.

By use of prior knowledge of the corpus and further study of the concepts in this book, we first constructed a visual representation of how the most important terms in the work relate

to each other. While this is a very useful way to understand the relations between terms for philosophers, it is not usable for the evaluation of the models. In order to get to a more computationally manageable format, we decided to translate the visual representation to a semi-ontology, meaning that for many terms we semi-formally defined its relations to other terms (e.g. *reference* is a relation between a *singular term* in language and an *object* in our ontology).

This resulted in a conceptual network of all the index terms of *Word and Object* [1960]. In general, we determined that all terms in the corpus can either be categorized as part of one of five clusters of terms (language, ontology, reality, mind, linguistic), or as a relation between terms in these clusters. Specifically, for every term in the index (with the exception of names and some more complicated cases), we determined whether it fell in either one of these five clusters or denotes a relation between specific terms within the clusters. For instance, one of the clusters is *reality* and contains terms such as *stimulus* which are defined to be part of reality.

The results reflect the meaning of the term by showing how it related to other terms. If two terms are in the same cluster, this means that they are conceptually related to each other. For instance, for any term  $x$  in the “language” cluster, we know that  $x$  is a linguistic item. Some of the relations between terms represent analogy judgments; what “reference” is to “singular terms” is what “being true of” is to “general terms”. More generally, for a relation between a term in the “language” cluster and a term in the “ontology” cluster, we know that it is a relation between a linguistic item and an object in the ontology. Moreover, some of the terms are related to each other by being near synonyms. For instance, “physical thing” and “material object” both terms in the “ontology” cluster and are used almost interchangeably.

### 3.3 Tuning model hyperparameters

Distributional semantic models have many parameters that have to be specified, such as the dimensionality of the embeddings and the size of the window around a word that is considered its context. The training algorithm also depends on a number of parameters that control how much the embeddings are adjusted at different stages of training. [17] showed that these parameters play a considerable role in the quality of the resulting embeddings. Choosing good parameter values is especially important when models are trained on small corpora. The models have to learn from very few contexts, so how the context is defined and how much a model should learn from each context are important factors.

We are interested in finding hyperparameter values that would be useful for our task, as described in Section 3.1. We want to use the Quine corpus and the conceptual network to evaluate our models on this task, so we avoid using the Quine corpus or the technical terms in the conceptual network to tune our models. Doing so could lead to parameter values that only work well for this particular corpus and the set of technical terms in the conceptual network. Instead, we propose a framework for tuning distributional models that can be applied to any use case involving small domain-specific

corpora. The strategy is to select hyperparameters that lead to consistent embeddings for terms in an artificial corpus. The artificial corpus is designed to contain contexts that are similar to those of the target corpus (the Quine corpus in our case). We now outline the main components of the tuning strategy.

(i) *Choosing terms for tuning.* The terms for which we want to learn embeddings are technical philosophical terms: the index terms of *Word and Object*. To approximate this type of data, we select technical terms from another domain - law. Many technical legal terms also have distinct meanings in legal scholarship as opposed to natural language. We identified 20 legal terms to use in our procedure.

(ii) *Choosing a corpus for context extraction.* The types of contexts surrounding the target words play a crucial role in the resulting embedding representation of the target words. The Quine corpus has characteristics that distinguish it from other natural language corpora. The type of contexts that occur in the corpus are different to the contexts that occur in general natural language corpora. It consists of philosophical texts, so it contains detailed discussions involving technical terms. We want to tune our models with contexts that are similar to those of the Quine corpus, since this would ensure that the chosen hyperparameter values would be applicable to the Quine corpus, yet there is no other source of Quine “big data” that we can use.

This problem is comparable to the problem of training models for an endangered language, for which little data is available. For that, the technique of data point selection has been proposed: from another language, only the data that is similar to data from the endangered language is selected, so that a model of the endangered language can be trained using data from different languages [22]. Similarly, we select another corpus of non-philosophical text that matches certain quantifiable characteristics of our Quine texts, out of a list of candidate corpora. This required a method for comparing different candidate corpora in terms of their contexts. For this we develop the notion of context characterization - computing a number of metrics that summarize the characteristics of the contexts in a corpus. We use the following quantifiable features:

- Word frequency: the average relative frequency of all the words in a context.
- Polysemy: the average number of WordNet synsets that the words in a context are associated with.
- Entropy: an information theoretic measure indicating how “surprising” the words in a context are.
- Number of words per sentence.
- Number of unique words per sentence.
- Type/token ratio per sentence: how many different words are used relative to the length of the sentence.

For each of the technical terms in the conceptual network we computed metrics based on these features in the Quine corpus. We then compared four candidate corpora to the Quine corpus to decide which one we would use to tune our models. We identified four possible corpora, ranging from highly technical to general usage - the British Law Corpus, the Open Access Journal corpus, Wikipedia, and the British National Corpus.

For each of the legal terms we extracted contexts from each of the candidate corpora. We then computed the context features for each corpus and compared the metrics to those of the Quine corpus (using methods for comparing the distributions of data sets). Out of the four corpora, the Wikipedia corpus was found to be most similar to the Quine corpus. Therefore we used the contexts of the legal terms extracted from the Wikipedia corpus to generate artificial examples and tune our models.

(iii) *Evaluating model consistency with artificial examples.* We generate artificial examples that can be used to evaluate a model for consistency. The examples consist of contexts from two terms, which are merged to become one pseudo-term. This procedure is based on the hypothesis that since the pseudo-term’s contexts are split evenly between contexts of *term1* and *term2*, its embedding should be half-way between the embeddings of the two terms.

We use these artificial examples because we need some measure of model quality to be able to select the better set of parameters in tuning our model. Yet, we do not yet want to use our conceptual network to decide what model represents concepts better, as that would bias the model towards modeling that particular set of concepts. Therefore, we use consistency over artificial examples for this purpose. Computing consistency does not require any meaning ground truth, it only different compares models. Following Bloem et al. [3], we say that a model is consistent if “its output does not vary when its input should not trigger variation (i.e. because it is sampled from the same text or domain)”. In our case, we want the model to produce similar meaning representations when it is presented with different sets of sentences discussing the same (artificial) term from the same corpus.

The artificial examples are generated using the legal terms chosen in (i) and the Wikipedia corpus chosen in (ii) to ensure that they are similar to the Quine corpus. Each artificial example involves two legal terms (*term1* and *term2*) and is generated by extracting contexts for both terms from the same corpus, replacing occurrences of either term with a pseudo-term *term1\_term2*, and shuffling the contexts. Half of the resulting artificial corpus consists of *term1* contexts and half of it consists of *term2* contexts. To control for possible effects of data size, each artificial example has the same number of contexts. The number of contexts used was 100, as the number of contexts we can find for the Quine terms in the Quine corpus is in that order of magnitude. To evaluate vector space consistency for a pair of terms we do the following:

- (1) We separately train embeddings for *term1* and *term2* from contexts of these terms extracted from the Wikipedia corpus.
- (2) We train an embedding for the pseudo-term from the artificial contexts of *term1\_term2*.
- (3) We compute the vector half-way between the embeddings of *term1* and *term2*.
- (4) We compute the cosine similarity of this vector and the embedding of *term1\_term2*. A high cosine similarity is seen as a good indicator of consistency.

Our hypothesis on the expected position of the pseudo-term embedding does oversimplify the nature of the semantic spaces

of word embeddings. The structure of semantic spaces and the distances between embeddings are still poorly understood, and it is not guaranteed that the embedding of a merged term should ideally be positioned in between its two constituent terms. However, we only assume that such a middle position is a good approximation when evaluating the consistency of a distributional semantic model using artificial data. During the tuning procedure, we apply steps 1 to 4 to several pairs of legal terms for each hyperparameter setting and compute the average cosine similarity obtained in step 4. We select the parameter values resulting in the highest average cosine similarity.

### 3.4 Learning embeddings

Before we can create new embeddings based on the Quine corpus, a number of preprocessing steps need to be carried out in order to deal with the particularities of the texts. Much of Quine’s work contains formal language, which is unusual for corpora used to learn word embeddings. Since many of the symbols appear quite frequently (e.g. logical variable  $x$ ) and in different contexts, it is possible that they would adversely affect the embeddings learned from the corpus. The corpus was preprocessed to prevent this. In addition to the usual preprocessing often applied to corpora in NLP (removing digits and punctuation, converting all text to lowercase) all one-letter words were removed from the corpus. The goal of this preprocessing was to remove all logical symbols from the corpus, leaving only Quine’s technical discussions for the distributional semantic models to learn from.

The distributional semantic models were trained on the full Quine corpus to learn embeddings for the technical terms in the conceptual network. The hyperparameters of the models were set to the values obtained by the procedure of Section 3.3.

### 3.5 Evaluation

As described at the start of this section, the main goal of this project is to determine whether distributional semantic models could be useful tools for philosophical analysis. For embeddings to be useful to a philosopher, they would have to capture the philosophically relevant meanings of technical terms. We designed an experiment that determines to what extent our learned embeddings accomplish this.

We drew inspiration from established evaluation methods for word embeddings [21], many of which test whether the surrounding neighbourhood of word’s embedding contains embeddings of words that are semantically similar or related. The general idea is that related terms should be closer to each other in the resulting vector space than unrelated terms. An example of this is the use of the SimLex-999 dataset [13], which contains pairs of English words with similarity scores as judged by native speakers of English. In a good distributional semantic model, the cosine similarities between embeddings are expected to approximate these human-rated similarity scores – the scores serve as a semantic ground truth or gold standard. These evaluation datasets do not include philosophical terminology though, so it has not been possible until now to

Ranking	Word frequency	Polysemy	Entropy	Sentence length	Unique words	Type/token ratio
1	Wiki	BLC	Wiki	Wiki	Wiki	BNC
2	OAJ	BNC	BNC	BNC	BNC	Wiki
3	BNC	OAJ	OAJ	OAJ	OAJ	OAJ
4	BLC	Wiki	BLC	BLC	BLC	BLC

**Table 1: A quantitative comparison of the contexts that occur in the Quine corpus to four other corpora - Wikipedia (Wiki), the British National Corpus (BNC), the Open Access Journal corpus (OAJ), and the British Law Corpus (BLC). For each context feature the candidate corpora are ranked from most similar to least similar to the Quine corpus.**

<i>contract + felony</i>	he entered a plea agreement on january to a <i>contract_felony</i> charge johnson had been signed to a recording <i>contract_felony</i> with bna records in august he signed a new three year <i>contract_felony</i> with tottenham
<i>admissible + sentence</i>	he began serving his <i>admissible_sentence</i> in may the second treadway confession remains <i>admissible_sentence</i> american league lawyers appealed the <i>admissible_sentence</i>
<i>custody + misdemeanor</i>	zelenka has already confessed and been taken into <i>custody_misdemeanor</i> the charge can be a <i>custody_misdemeanor</i> or a felony the inquest is mandatory with a jury where the death occurs in <i>custody_misdemeanor</i>

**Table 2: Some samples from the generated contexts of three pseudo-terms. For each pseudo-word the the contexts of two different legal terms were extracted from Wikipedia and merged to create an artificial corpus.**

apply this kind of evaluation, based on a gold standard, to embeddings of philosophical terms. Previously, only evaluation metrics that measure qualities of an embedding without reference to manually created gold standard data were used in the domain of philosophy [3]. To be able to evaluate our embeddings extrinsically, based on input from philosophers, we need our own evaluation dataset.

Since we are interested specifically in the philosophical meaning of terms in the works of Quine, we used our conceptual network to specify whether or not the *Word and Object* [20] index terms are related to each other. We consider terms that are in the same cluster to be more related to each other than terms in different clusters. We expect terms in the same cluster to have embeddings that are closer to each other (having higher cosine similarity scores) than terms in other clusters.

To test this hypothesis we analysed the embeddings of the terms in the conceptual network that we learned, as described in the previous section. For each term in the conceptual network we randomly sampled one term in its cluster and one term in a different cluster. We then computed which of these sampled terms are closer to the original term in the embedding space.

This evaluation method was extended to a measure of the quality of learned embeddings. By computing the success rate of a set of embeddings (how often a term from the same cluster is closer to the target term than a term from a different cluster) we obtain a metric that summarises how well embeddings reflect the conceptual network. This allowed us to compare the embeddings learned by different models, which gave us an indication of which models are best suited for this kind of application.

## 4 RESULTS

We present the results of preliminary attempts at applying the methodology outlined in the previous section. We experiment with two distributional semantic models - Word2Vec and Nonce2Vec.

### 4.1 Context characterisation

We applied the context characterisation developed in Section 3.3 (ii) to the Quine corpus and four other corpora. The features were also computed for the legal terms in four candidate corpora. We then compared the distribution of the features in each of the four corpora to the Quine corpus features. We did this by finding the deciles of the distributions and computing the difference between the deciles of the Quine corpus and the candidate corpora. This enabled us to rank each of the candidate corpora based on their similarity to the Quine corpus for a particular feature. The resulting rankings are shown in Table 1. The rankings show that out of the four candidate corpora, the Wikipedia corpus is most similar to the Quine corpus in terms of its contexts. It was most similar to the Quine corpus for as many as 4 out of the 6 features computed.

### 4.2 Hyperparameters

We generated artificial examples of the kind described in Section 3.3 (iii) for all the possible pairs of legal terms. Some artificial contexts are presented in Table 2. We then randomly selected 10 pairs of terms to tune our model hyperparameters with.

We trained Word2Vec with different hyperparameters and computed the average cosine similarity (of the pseudo-term embedding and the expected embedding) for each parameter setting over all 10 pairs. The average cosine similarity varied widely, from 0.2362 (the least consistent model) to 0.9665

<i>noun</i>	<i>description</i>	<i>eternal sentence</i>		<i>quantifier</i>
1. designate	indeterminacy of translation	affirmative meaning	stimulus	indefinite singular term
2. elimination	designate	explication		<b>copula</b>
3. <b>demonstrative</b>	<b>verb</b>	phoneme		verb
4. referential opacity	<b>substantive</b>	<b>article</b>		<b>syntax</b>
5. homonymy	<b>noun</b>	<b>mass term</b>		relative term

**Table 3: Examples of the neighbourhoods around terms in the embedding space learned by Word2Vec. The five nearest neighbours of the terms are shown. Terms within the same cluster are indicated with bold text.**

<i>noun</i>	<i>description</i>	<i>eternal sentence</i>		<i>quantifier</i>
1. <b>verb</b>	context	<b>indefinite term</b>	<b>singular</b>	<b>predication</b>
2. <b>adjective</b>	<b>function</b>	<b>demonstrative</b>		<b>conjunction</b>
3. <b>substantive</b>	<b>conjunction</b>	<b>adjective</b>		<b>connective</b>
4. designate	<b>construction</b>	<b>pronoun</b>		open sentence
5. <b>relative term</b>	material	<b>noun</b>		description

**Table 4: Examples of the neighbourhoods around terms in the embedding space learned by Nonce2Vec. The five nearest neighbours of the terms are shown. Terms within the same cluster are indicated with bold text.**

(the most consistent model). The most influential parameters were those that control control how much the embeddings are adjusted during training.

We also tuned the hyperparameters of Nonce2Vec with this procedure. The average cosine similarity of different parameter settings varied less widely, from 0.5627 to 0.9489. The most influential parameters were once again those that control control how much the embeddings are adjusted during training. The subsampling rate, a parameter that controls how often very frequent words are ignored during training, was also highly influential on the resulting consistency.

### 4.3 Evaluation

We learned embeddings for the terms in the conceptual network by training Word2Vec and Nonce2Vec on the entire Quine corpus. We used the hyperparameters that led to the highest consistency in our artificial experiments. We then applied the evaluation strategy outlined in Section 3.5.

Word2Vec did not group the embeddings of terms within a cluster closer to each other than those of terms in different clusters. If the embedding of any term is compared to the embeddings of two other terms - one within its cluster and one in a different cluster - there is no tendency for the term’s embedding to be closer to the term within its cluster. Table 3 demonstrates this by showing the five nearest neighbours of four terms in the conceptual network. The surrounding neighbourhoods of these terms do not contain more terms from the same cluster than would be expected from random chance.

However, the embeddings produced by Nonce2Vec did reflect some of the clusters in the conceptual network. 60% of the time that a term is sampled from the same cluster as a target term it is closer in the embedding space than a term sampled from a different cluster. This is significantly better

than the Word2Vec embeddings, which shown no improvement over the 50% baseline of randomly guessing which term would be closer. The grouping together of terms from the same cluster in the embedding space can clearly be seen when the nearest neighbours of some of the terms are analysed. Table 4 demonstrates this by showing the five nearest neighbours of four terms in the conceptual network. The surrounding neighbourhoods of these terms tend to contain terms from the same cluster.

## 5 DISCUSSION & CONCLUSION

Nonce2Vec is designed to learn high-quality embeddings from small corpora. However, this does not necessarily mean that it can learn embeddings from domain-specific corpora that capture the technical meanings of terms. Testing this requires the creation of an evaluation framework that determines to what extent the domain-specific meanings of technical terms are reflected in the learned embeddings. This is what our methodology achieves for the philosophical domain. The novel evaluation framework that we have proposed evaluates to what extent the embeddings learned by a model reflects the constructed philosophical ground truth. The results show that Nonce2Vec is able to capture the meaning of philosophical terms in the Quine corpus better than Word2Vec.

The results presented in this paper have demonstrated the importance of a methodologically sound approach to evaluating distributional semantic models in the digital humanities. Our methodology evaluates how well embeddings capture the technical meanings of terms in a single corpus. In the future this approach could be extended to other applications of word embeddings in the digital humanities. It would be possible to develop a methodology that evaluates how well embeddings capture the subtle differences in the meanings of terms in different texts and across different domains (e.g. philosophical

texts compared to natural language). Another application of interest would be evaluating how well diachronic embeddings reflect changes in the meanings of technical terms over time.

In section 3, it was noted that there is an ambiguity in what the relation “closeness in vectorial space” or “cosine similarity” captures. In general, the notions similarity and relatedness are problematic due to their context sensitivity. As Goodman [6, 445] argues, when judging whether one thing is more similar to something than something else, we not only have to make a selection between relevant properties, but also determine how important these properties are to the relative similarity. This means that there is not one way in which we can judge similarity, but something can both be said to be similar and not similar to something else according to different considerations. Therefore, if the considerations that should be taken into account and the relative weighing of these considerations is not fixed beforehand, similarity judgments do not identify one specific relation.

In the present paper, we identified a specific way in which terms can be related or similar and evaluated whether closeness in vectorial space captures this forms of relatedness. It should be noted that, in general, the distributional model can never tell us that two terms are not related, due to this context-sensitive nature of relatedness and similarity. What it can show is that there is a sense of relatedness such that two terms are more closely related to each other than to another term. Before we can use the model to support the philosophical analysis of terms and how terms might change over time or across authors, the sense of relatedness that the distributional model captures needs to be more precisely identified.

## REFERENCES

- [1] Baroni, M., Dinu, G., and Kruszewski, G. (2014). Don’t count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 238–247. Association for Computational Linguistics.
- [2] Betti, A. and van den Berg, H. (2014). Modelling the history of ideas. *British Journal for the History of Philosophy*, 22(4):812–835.
- [3] Bloem, J., Fokkens, A., and Herbelot, A. (2019). Evaluating the consistency of word embeddings from small data. Unpublished manuscript.
- [4] Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*.
- [5] Ginammi, P., Koopman, R., Wang, S., Bloem, J., and Betti, A. (2013). Bolzano, kant, and the traditional theory of concepts. In *The Dynamics of Science: Computational Frontiers in History and Philosophy of Science*.
- [6] Goodman, N. (1972). Seven strictures on similarity. In *Problems and Projects*, pages 437–446. Bobs-Merril.
- [7] Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2016a). Cultural shift or linguistic drift? comparing two computational measures of semantic change. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2116–2121. Association for Computational Linguistics.
- [8] Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2016b). Diachronic word embeddings reveal statistical laws of semantic change. *CoRR*, abs/1605.09096.
- [9] Hellrich, J. and Hahn, U. (2016a). An assessment of experimental protocols for tracing changes in word semantics relative to accuracy and reliability. In *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 111–117. Association for Computational Linguistics.
- [10] Hellrich, J. and Hahn, U. (2016b). Bad company—neighborhoods in neural embedding spaces considered harmful. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2785–2796. The COLING 2016 Organizing Committee.
- [11] Herbelot, A. and Baroni, M. (2017). High-risk learning: acquiring new word vectors from tiny data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 304–309. Association for Computational Linguistics.
- [12] Herbelot, A., von Redecker, E., and Müller, J. (2012). Distributional techniques for philosophical enquiry. In *Proceedings of the 6th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 45–54. Association for Computational Linguistics.
- [13] Hill, F., Reichart, R., and Korhonen, A. (2015). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.
- [14] Kenter, T., Wevers, M., Huijnen, P., and de Rijke, M. (2015). Ad hoc monitoring of vocabulary shifts over time. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management, CIKM ’15*, pages 1191–1200, New York, NY, USA. ACM.
- [15] Kim, Y., Chiu, Y., Hanaki, K., Hegde, D., and Petrov, S. (2014). Temporal analysis of language through neural language models. *CoRR*, abs/1405.3515.
- [16] Kulkarni, V., Al-Rfou, R., Perozzi, B., and Skiena, S. (2014). Statistically significant detection of linguistic change. *CoRR*, abs/1411.3315.
- [17] Levy, O., Goldberg, Y., and Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.
- [18] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.
- [19] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013.*, pages 3111–3119.
- [20] Quine, W. V. O. (1960). *Word and object*. MIT press.
- [21] Schnabel, T., Labutov, I., Mimmo, D., and Joachims, T. (2015). Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 298–307, Lisbon, Portugal. Association for Computational Linguistics.
- [22] Søgaard, A. (2011). Data point selection for cross-language adaptation of dependency parsers. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 682–686. Association for Computational Linguistics.
- [23] Van Den Berg, H., Betti, A., Castermans, T., Koopman, R., Speckmann, B., Verbeek, K., Van der Werf, T., Wang, S., and Westenberg, M. A. (2018). A philosophical perspective on visualization for digital humanities. In *3rd Workshop on Visualization for the Digital Humanities (VIS4DH2018)*.
- [24] Van Wierst, P., Vrijenhoek, S., Schlobach, S., Betti, A., et al. (2013). Phil@ scale: Computational methods within philosophy. In *DHLU*.
- [25] Young, T., Hazarika, D., Poria, S., and Cambria, E. (2017). Recent trends in deep learning based natural language processing. *CoRR*, abs/1708.02709.