Challenging Distributional Models with a Conceptual Network of Philosophical Terms

Yvette Oortwijn[♣†] **Jelke Bloem**[♣] **Pia Sommerauer**[*]
Francois Meyer[*] **Wei Zhou**[♣] **Antske Fokkens**[*†]
[♣] Institute for Logic, Language and Computation, University of Amsterdam
[*]Computational Linguistics & Text Mining Lab, Vrije Universiteit Amstersterdam
[†]Department of Mathematics and Computer Science, Eindhoven University of Technology
{y.oortwijn,j.bloem}@uva.nl, {pia.sommerauer,a.s.fokkens}@vu.nl
francoisrmeyer@gmail.com, csunset14330@gmail.com

Abstract

Computational linguistic research on language change through distributional semantic (DS) models has inspired researchers from fields such as philosophy and literary studies, who use these methods for the exploration and comparison of comparatively small datasets traditionally analyzed by close reading. Research on methods for small data is still in early stages and it is not clear which methods achieve the best results. We investigate the possibilities and limitations of using distributional semantic models for analyzing philosophical data by means of a realistic use-case. We provide a ground truth for evaluation created by philosophy experts and a blueprint for using DS models in a sound methodological setup. We compare three methods for creating specialized models from small datasets. Though the models do not perform well enough to directly support philosophers yet, we find that models designed for small data yield promising directions for future work.

1 Introduction

Philosophers apply text analysis to understand and delineate the precise meaning of concepts and the relations between them in a given text. This includes comparative research that investigates differences in how concepts are viewed in different philosophical schools or by individual philosophers. Betti and van den Berg (2014) point out that comparative research on concepts should follow a conceptual model approach. This approach states that we should not look at shifts of individual concepts in isolation, but rather address changes of a conceptual model as a whole. In such a system, relations between concepts are made explicit and comparative studies should identify how such relations change. Previous studies have shown that distributional methods can be used to support philosophical research by retrieving passages relevant to concepts in an author's work (e.g., the concept of *grounding* within the work of Bernard Bolzano, van Wierst et al., 2016; Ginammi et al., 2020), but can we also generate distributional semantic (DS) models that are precise enough to identify differences in concepts?

This paper takes a first stab at addressing this question. In particular, we address the challenges involved in dealing with highly technical domain-specific terms that are defined in small corpora. As such, our use case has properties difficult for DS modeling, but typical for disciplines working with comparatively limited data. We compare domain-specific embeddings created using Word2Vec (Mikolov et al., 2013a,b) and a count-based SVD model (Levy et al., 2015) to those created by Nonce2Vec (Herbelot and Baroni, 2017), specifically designed for dealing with tiny data. Taking into account previous work criticizing the use of DS models for detecting sense-shift, we construct a data-specific ground truth, apply multiple evaluation metrics and verify whether results are stable across various random initializations. Our results confirm that SVD representations are superior to Word2Vec for small data and show that Nonce2Vec outperforms Word2Vec and, in most cases, SVD. However, results are currently not accurate enough for providing evidence or new insights to philosophers. Nevertheless, we are hopeful that better results can be obtained in the future by optimizing Nonce2Vec to deal with small rather than tiny data and by creating a bigger, more balanced ground truth.

The main contributions of this paper are (1) a new ground truth of philosophical concepts linked to a clean philosophical corpus that is particularly challenging to model; (2) a blueprint for investigating DS models for domain specific research; (3) a comparative study of different approaches of creating embeddings for highly domain-specific terms.[1]

[1]The ground truth, details of results and code can be found on GitHub: https://github.com/YOortwijn/

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2511–2522
June 6–11, 2021.

After presenting related work, we describe the philosophical context: requirements, corpus and our ground truth. In Section 4, we outline how the DS models we use are created. We then present our evaluation and results in Section 5 which is followed by our conclusions and discussion.

2 Related Work

In this section we cover (1) other work related to distributional semantics (DS) for specific concepts and conceptual change (2) critical reflection on evaluation and the methodology involved and (3) work on small datasets and identification of domain specific meaning.

2.1 DS for Concepts and Conceptual Change

A well-known application of DS is the use of diachronic word embeddings to track and analyze changes in the meaning of words over periods of time (Kim et al., 2014; Kulkarni et al., 2015; Mitra et al., 2015; Hamilton et al., 2016b,a; Kenter et al., 2015; Tahmasebi and Risse, 2017; Montariol and Allauzen, 2019; Giulianelli et al., 2020, e.g.). Most of these approaches study what is called *semantic shift*, which is the change in (dominant) sense of a specific word by comparing the word's meaning representations in different time periods (Kutuzov et al., 2018). DS methods have also been used to study concepts related to gender and intersectionality (Herbelot et al., 2012), studying cultural stereotypes (Lewis and Lupyan, 2019) or harm-related concepts in psychological research papers (Vylomova et al., 2019).

Wevers and Koolen (2020) survey three ways in which distributional semantic representations can help trace concept change. However, none of these methods requires historians of ideas to fix initial and testable hypotheses on the meaning of concepts as Sommerauer and Fokkens (2019) recommend on the basis of Betti and van den Berg (2014). Betti and van den Berg argue that concepts are not isolated, but part of conceptual models. Sommerauer and Fokkens (2019) show that translating conceptual models to words representing them is one of the challenges involved in using DS models for studying conceptual change. They ground their conceptual model of 'Racism' in literature by sociologists, anthropologists and historians, but argue that domain experts would ideally be involved directly, as is done in the current paper. Betti et al. (2020) introduce a concept-focused ground truth designed by domain experts, QuiNE-GT, where paragraphs of philosophical text are annotated in terms of their relation to a conceptual model of the concept of *naturalized epistemology* in Quine's works. We also make use of conceptual modeling methodology to build a ground truth, but our task is to extract knowledge on target term relations rather than to perform an information retrieval task searching for paragraphs relevant to a research question. While QuiNE-GT contains exhaustive lists of words pertaining to a particular research question, we aim for broader coverage of different terms used by Quine and their relations.

2.2 Methodological Challenges

An interdisciplinary collaboration with domain experts can lead to hypotheses about shifts or nearest neighbors of specific terms, which can be tested by methods also used for detecting sense shift. These methods are not without challenges. The meaning representations are affected by random factors such as initialization and order of example (Hellrich and Hahn, 2016a) and frequency effects (Dubossarsky et al., 2017). A major obstacle in addressing these critical points is the lack of high quality evaluation sets (Tahmasebi et al., 2018; Kutuzov et al., 2018) and a tendency to use a single evaluation metric (Gladkova and Drozd, 2016) while each metric has downsides (Bakarov, 2018) .

Evaluations on small sets of hand-picked examples that exhibit strong sense-shift (e.g. Hamilton et al. (2016a)) leave it unclear whether they are also suitable for making new discoveries or exploring data. van Aggelen et al. (2019) introduce a large-scale evaluation set derived from a thesaurus and show that performance of distributional methods is much lower on this more challenging set.

These critical findings stress the need for methodologies that allow us to establish the quality of embeddings and to tell the difference between a stable, reliable finding and an artefact of the method. Dubossarsky et al. (2017) propose the use of shuffled and synchronic corpora for verification. Rosenfeld and Erk (2018) use synthetic words that consist of two real words merged together that result in a shift of these words' senses for evaluation. Sommerauer and Fokkens (2019) recommend stress-testing through control words (that should not change) and by comparing results on multiple models. We supplement these proposals for

Challenging_DMs.

diachronic models by providing methods that can be used as a strict test of synchronic model quality, independently of measuring change (so that frequency effects are not a risk). To ground these methods, we introduce a novel, high quality ground truth containing fine-grained meaning distinctions in the philosophical domain. We stress-test our findings by applying multiple evaluation metrics and control for random factors by initializing our models multiple times.

2.3 Dealing with Small Datasets

In addition to the challenges outlined above, we are faced with the issue that domain-specific corpora are typically small, i.e. up to a few million tokens rather than web scale. Learning embeddings from small corpora is not an easy task, where SVD models outperform Word2Vec (W2V) (Sahlgren and Lenci, 2016, on 1M words, Asr et al., 2016, on 8M words), and learning them for rare words presents further difficulty (Luong et al., 2013). Nonce2Vec (N2V) (Herbelot and Baroni, 2017) addresses this issue through 'high-risk' incremental learning with an initially high but decaying learning rate, allowing them to learn embeddings from single sentences (called tiny data). Faruqui et al. (2015) incorporate ontological information from lexical-semantic databases as a postprocessing step, which can be done when training data is sparse. However, when working in a specific domain, such as the texts of a particular philosopher, words may have different and specific uses, and general-purpose evaluation resources or training data do not always reflect these meanings (Betti et al., 2020). Bloem et al. (2019) confirm the domain-specific character of philosophical writings showing that two vectors for the same word, one trained on Wikipedia and one trained on the works of a specific philosopher, can have low similarity, especially for high-frequency terms. Shoemark et al. (2019) find that the top ranked words are domain-specific to the Twitter data they used. Wohlgenannt et al. (2019) evaluate DS models trained on two fantasy book series by having domain experts manually compile evaluation datasets addressing the relevant word senses, incorporating domain knowledge in both training and evaluation. Roy et al. (2019) propose incorporating text annotation of in-domain vocabulary and semantic relations into the word embeddings to improve the quality of domain-specific word embeddings learned from relatively small data sets.

In this paper, we investigate how different approaches for learning embeddings deal with the domain specific concepts we are dealing with. We compare Herbelot and Baroni's N2V to continuing training with W2V and to directly creating SVD models on our corpus.

3 Philosophical Background and Data

The goal of this section is to provide some insights into the process of interpreting philosophical texts and the use case for our experiments. We briefly describe the process and challenges of close-reading and how it could be supported by DS models. Then we present a corpus of philosophical texts and a ground truth for philosophy.

3.1 Philosophical Questions

Many philosophical research questions focus on the interpretation and comparison of philosophical views expressed in writing. These questions revolve around specific concepts and how they are defined and viewed by different philosophers. Often, different philosophers use the same terms to describe different concepts. For example, Quine sees *reference* as a relation between a singular term and a physical object, where a physical object is not part of *reality*, but of our *ontology* (Quine, 1960). This is opposed to many other philosophers, who take what we refer to and what we receive stimulation from as the same thing, i.e., physical objects in reality.

To make solid comparisons between views, it is necessary to determine which concepts are closely related to each other, or which concept pairs stand in similar relations to others. To do this, philosophical experts practice close-reading. The interpretation of only a single passage requires close-reading and expertise of not only the work the passage is in, but often also other works by the same author or even other authors. Conclusions are often drawn on a small subset of the relevant available data. It almost always requires making a selection of sources to consider and thus allows for cherrypicking data. The use of computational linguistic methods, instead, could make it possible to consider all available data as a basis for evidence, and thereby prevent biased source selection.

Differences in the interpretation of a term in different authors' work can be understood as a difference in its relations to other terms. A difference

in how terms can be clustered together would then show a difference in the conceptual relations these terms have to each other. Computational methods that can capture this aspect of meaning can be applied in various stages of philosophical research.

Exploration. In the first stages of research, a philosopher might have a single or a few passages or terms that should be interpreted. At this point, they may want a rough overview of other passages or terms relevant to the one(s) under consideration. DS models may help the researcher to find relevant passages without any of the search terms they may use in key term search. These passages can provide input for more directed searches and be a start for a traditional research path with close-reading of the identified passages. The recall of the method for this application need not be very high: as long as the researcher identifies some new relevant passages without being overloaded with irrelevant ones and the selection is not biased towards a specific interpretation, DS models enrich the philosopher's research.

Testing Hypotheses about the Text. When a researcher already has some competing hypotheses for interpretation based on close-reading of some works or passages, or based on secondary literature, DS models can help to compile evidence for both hypotheses and compare the results. If there are multiple possible interpretations of a term, a DS model could provide insight into which terms are most closely related to this term, giving evidence for the correct interpretation. If the outcome of such a comparison is to be used as direct evidence, it is essential that the DS model is highly accurate and a methodology is applied to distinguish veritable observations from noise. A researcher may however also use these results in a more surveying manner. In this case, more accuracy is needed than in the case of identifying passages, but a certain amount of error is acceptable. In this paper, we aim to investigate the level of accuracy we can obtain on philosophical text with either of these applications in mind (surveying hypotheses or providing evidence for a hypothesis).

3.2 Quine in Context Corpus

We make use of a large corpus that comprises the virtually complete *oeuvre* in English of Willard V. O. Quine, the QUINE corpus (Version 0.5, Betti et al., 2020),[2] for creating our DS models. The corpus includes texts on various topics, from formula-heavy logic works to philosophy of language. Version 0.5 of this corpus consisting of 228 books and articles by Quine, containing 2,150,356 word tokens and 38,791 word types. It is a high quality corpus where scanned page images were OCR-processed and corrected manually.

3.3 A Ground Truth in Philosophy

Establishing a *ground truth* for philosophical concepts is not trivial (see e.g. van den Berg et al. (2018), Betti et al. (2020)). We address this by building on the methods described by Betti and van den Berg (2014) for building conceptual models. Instead of trying to understand the meaning of a term in isolation, we focus on the interrelations of terms.

We base our ground truth on Quine's *Word and Object* (Quine, 1960), which encompasses many of the terms and themes that Quine discusses throughout the rest of his work.[3]. We obtain this book's most important terminology from its index. The philosophical expert on our team established a conceptual network representing the term-clusters and relations. The expert categorized each word as either belonging to one of five clusters (LANGUAGE, ONTOLOGY, REALITY, MIND, META-LINGUISTIC) or as a relational term (i.e. part of either the *reference* or *regimentation* relation that connects (parts) of clusters to each other). Any two terms in the same cluster can be seen as conceptually related (e.g. *noun* and *verb* are conceptually related since they are both linguistic items and are therefore both in the LANGUAGE cluster). The *reference* relation connects terms from the *language* and *ontology* cluster, i.e. elements of language *refer to* elements of the ontology. *Regimentation* connects parts of the *language* and *meta-linguistic* cluster. So the terms that are clustered together are semantically similar to each other, while the relational terms are related terms that are not necessarily semantically similar. Our conceptual network contains 74 clustered terms and 43 relational terms (overlapping the 74). The conceptual network was checked independently by two other philosophers specialized in

[2]The corpus was derived from copyrighted works by Betti et al. (2020). The corpus is available to researchers that can show they own the original works. Replication instructions are available here: https://github.com/YOortwijn/QuiNE-ground-truth

[3]A more detailed and accessible explanation of the conceptual network, including further motivation for the categorization of terms can be found at https://github.com/YOortwijn/Challenging_DMs

Quine. There was a 100% consensus among the experts on the clustering of the 74 terms and relations of the 43 terms. Since these terms are core terms in the work of Quine for which most experts agree on their coarse interpretation, high consensus was expected. However, differences in interpretations and disagreement between experts is more likely upon more fine-grained analysis and even though consensus was expected, a fourth consulted expert may still disagree with the interpretation.

Even high-quality DS models have certain limitations when it comes to representing words accurately due to their architecture (e.g. expressing very fine-grained differences and polysemy). We identified the following potential challenges prior to examining vector representations from our DS models: First, terms that are related by the *reference* relation might be closer to each other than to other terms in their respective clusters. For instance, a *singular term* (cluster LANGUAGE) refers to a *physical object* (cluster ONTOLOGY). Therefore, they might be closer to each other than to other terms in their clusters (*relative clause* and *class*, respectively). Second, the LANGUAGE and METALINGUISTIC clusters are relatively similar. While they can be distinguished in *Word and Object* by their relation to ontology and regimentation but this is not necessarily the case for all of Quine's works. Examples of terms that could be misplaced due to this are *article* and *noun*. Third, there are terms that are comparatively distinct from the other terms in their cluster (but nevertheless clear members of the cluster), such as *phoneme* in the REALITY cluster. Fourth, the clusters contain some polysemous terms and terms that can be used in both a technical and a non-technical way within Quine's works, e.g., *name*, *particular*, *context*, *form*. Finally, some terms, such as *prelinguistic quality space*, might have an extremely low number of occurrences.

Based on these observations, we divide our ground truth in the following subsets: (1) terms that should be assigned to the correct cluster and (2) terms that could be assigned to a wrong but also plausible cluster given the corpus and the first two potential challenges by way of the *reference* or *regimentation* relation. The focus will be on (1), but (2) will be used in the first task.

4 Training DS Models

Bloem et al. (2019) noted that reasonable embeddings for some philosophical terms can be learned from Wikipedia-data. As a baseline, we include a model trained exclusively on a 2019 Wikipedia dump using default Word2Vec (W2V), *wikipedia-W2V*. Multi-word target terms were linked by underscores to have a single vector per target term and 85 of the 99 target terms are in the vocabulary of this model. We test an SVD count-based model, using the PPMI-SVD approach from Levy et al. (2015) and two predictive approaches for creating our DS models: W2V (Mikolov et al., 2013a,b) in its Gensim implementation (Řehůřek and Sojka, 2010), as well as Nonce2Vec (Herbelot and Baroni, 2017, N2V) adapted for small, in-domain data situations (Bloem et al., 2019). To learn an embedding for a specific term, N2V uses the sentences in which this term occurs to map it into a previously learned general-domain semantic background space trained on Wikipedia data.[4] This is done by initializing the vector for the target term to the sum of the background space vectors of words in the in-domain context sentence from the Quine corpus, following Lazaridou et al. (2017). Training then takes place with an initial high learning rate and parameter decay, while the background space is frozen and only the target term is learned. Using W2V, we learn embeddings for specific terms by training only on the in-domain context sentences of our target terms. We test two initialization methods: random initialization, and using the additive model of N2V. We also modify N2V to have a random initialization condition for comparison, giving us four conditions: *W2V-random*, *N2V-random*, *W2V-additive* and *N2V-additive* (the N2V default).

4.1 Preprocessing

We carry out various preprocessing steps to ensure that we (1) find the maximum of target term mentions and (2) regularize the contexts so we can exploit the full potential of the small corpus. In part, we make use of the preprocessing steps already performed on the QUINE corpus (v0.5), which was sentence-split and tokenized using UCTO[5] and lemmatized using Spacy[6] using its core model for English.[7] The QUINE corpus features a rather high number of mathematical expressions. Rather than treating them as unique expressions, they were nor-

[4] We used the same Wikipedia dump for *wikipedia-W2V*.

[5] https://languagemachines.github.io/ucto/

[6] https://spacy.io/

[7] Spacy English core model: en_core_web_sm

malized by replacing them by the symbol *XfZ* for formulas, and *XsZ* for symbols. We assume that the specific expressions do not add to the distributional information.

For (1), we need to ensure that all instances of the terms in the evaluation set are identified in the corpus. We search for all morphological variants of the target terms and replace them by the unmarked singular form, by means of a manually created list. Furthermore, many of the target terms consist of two or more words, which should receive a single representation. As with the Wikipedia baseline, we search for all mentions of the target terms in the corpus and join all target terms from the ground truth that consist of multiple words (MWEs) by underscores to turn them into a single token. We did not handle MWEs that were not target terms, so no automatic MWE identification took place.

4.2 Hyperparameter Tuning

We propose a framework for fine-tuning models specifically designed for domain-specific experiments with small data. As the size of our ground truth is comparatively limited (for computational purposes), we do not want to 'waste' portions of it for fine-tuning. Instead, we use 'proxy' terms and a 'proxy' corpus to evaluate and compare models on an artificial task. We aim to select data representative of the target data (inspired by fine-tuning for low-resource languages, Søgaard, 2011).

Terms and Corpus. As target terms we select 20 technical terms from the legal domain. Similar to the philosophical target terms, many technical legal terms have distinct or more specific meanings in legal scholarship as opposed to generic corpora. To select a proxy corpus, we compare the contexts of the target terms to the contexts of the legal terms in four candidate corpora: the British Law Corpus (BLC), the Open Access Journal corpus, Wikipedia, and the British National Corpus (BNC). We compare the contexts in terms of easily computable metrics which characterize properties we expect to have an impact on training a DS model: average relative frequency of all the context words, their average polysemy (in terms of WordNet synsets (Fellbaum, 2010; Miller, 1995)), their entropy (based on unigram frequency), type count, token count, and type/token ratio. We rank each corpus by similarity to the Quine corpus on each metric. Out of the four corpora, Wikipedia and the BNC had an average rank of 1.8, while the BLC was the least similar with 4. Out of the two equal choices in terms of means, the Wikipedia corpus was more similar to the Quine corpus in terms of variance, so we chose this corpus for extracting contexts of the legal proxy terms.

Task. As we do not have a conceptual ground truth for the legal terms, we rely on an artificial task. We approximate embedding quality in terms of consistency. Bloem et al. (2019) define a model as 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)". We test whether a model creates consistent representations of a term when trained on only a subset of its contexts using artificial examples in the following way: Our artificial examples consist of contexts of two terms, which are merged to become one pseudo-term. Since the pseudo-term's contexts are split evenly between contexts of *term1* and *term2*, its embedding is expected to be somewhere halfway between the embeddings of the two terms.[8] We train separate vectors $\vec{t_1}$ and $\vec{t_2}$ for *term1* and *term2* on the basis of 100 occurrences of each, as well as $\vec{t_p}$ for the pseudo-term *term1_term2*, based on 50 occurrences of each component term. We then compute the vector half-way between $\vec{t_1}$ and $\vec{t_2}$. In a consistent model, the cosine similarity between this vector and $\vec{t_p}$ should be high. In tuning, we perform a grid search and take the average of this metric computed over 10 random pairs of legal terms for each hyperparameter combination.[9]

The results show that our models can learn vectors for artificial combined terms that are consistent with the middle point between the vectors of the two component terms in vectorial space. There is great variation for different hyperparameter sets. Average cosine similarities varied from 0.08 to 0.87 (N2V-additive) or 0.96 (W2V-random). We found that with the additive initialization, lower learning rates performed better, while with the random initialization, higher number of negative samples had the greatest impact on the consistency scores. For N2V, the lowest parameter decay

[8] Our assumption on the expected position of the pseudo-term embedding oversimplifies the nature of DS models. 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 in tuning, not in testing our models.

[9] The full parameter space can be found in our code repository.

rates performed best, probably because our artificial terms have more occurrences (50 and 100) than N2V was designed for (1-4). The initial high learning rate is a core feature of N2V, so we also include the best setting with a learning rate of 1 as an additional condition (*N2V-additive-a1*).

5 Results

The tuned models were evaluated against the ground truth. This section presents multiple evaluation tasks and results to (1) explore different aspects of model quality and (2) stress-test our findings. In these tasks, we use the 74 terms from the conceptual network that were clustered by the experts.

Cluster similarity Our similarity task is defined as follows: Given a target term t_t, a term from the same cluster t_{sc} and a term from a different cluster t_{dc}, we test whether the target term t_t is closer to t_{sc} than to t_{dc}. If the cosine similarity between t_t and t_{sc} is higher than the cosine similarity between t_t and t_{dc}, it is counted as correct, else as incorrect. We carry out this comparison for all possible term combinations and report the percentage of correct outcomes. We report the proportion of target-terms that are classified in the correct cluster. We also show the proportion of target terms that are clustered incorrectly but plausibly given their relation (via *reference* or *regimentation*) to other clusters. We exclude all three terms that are out of vocabulary in any of the models, as a difference in target terms distorts the comparison. This way, we ensure that all models are evaluated on the same terms.

Table 1 shows that N2V outperforms the W2V models in most cases. The best performance is by N2V with additive initialization (standard N2V), pairing 65.0% correct according to the clusters, and 72.6% when additional relations between terms are also considered correct. The count-based SVD model performs similarly well. These are the only two models that beat the Wikipedia baseline. The best W2V model (W2V-random), pairs 56.4% correct according to the clusters, and 64.3% with additional relations. To evaluate the stability of our best result, we train 25 identically parameterized models as in Hellrich and Hahn's (2016b) reliability metric and obtain similarity scores in a range of 64.04%-65.22% (mean 64.65%, cf. 64.95% in testing) indicating high stability.

Model	Sim.	Oth. Rel	Dunn
N2V-additive	**64.95%**	7.68%	**0.56**
N2V-additive-a1	56.09%	8.38%	0.24
N2V-random	55.18%	8.39%	0.19
W2V-additive	52.00%	6.44%	0.12
W2V-random	56.44%	7.88%	0.08
SVD	**65.19%**	7.18%	0.35
wikipedia-W2V	59.58%	6.17%	0.17

Table 1: Outcome Cluster Similarity & Dunn Index

Dunn index The Dunn Index (DI) is a general metric of cluster quality and can be used to measure how well embeddings from the same cluster are clustered in semantic space (Huang et al., 2016). It is the ratio of the minimum inter-cluster distance to the maximum cluster size, and higher values indicate tighter clusters and better separation.

The DI results in table 1 confirm that N2V models outperform W2V models. N2V with additive initialization achieved the highest DI value 0.56, followed by the SVD model (0.35). We can compare this to Huang et al. (2016), who used DI to evaluate word embeddings in the medical domain, using six semantic clusters taken from an expert-defined controlled vocabulary of medical terms using far larger data sources (e.g. PubMed and Wikipedia). In their experiment with 800 terms (we have 99) and six clusters (comparable to our five), their DI scores were 0.16-0.20 for a bag-of-words baseline model, and 0.43 (PubMed) to 0.25 (Wikipedia) for W2V. This is comparable to our wikipedia-W2V condition which scored 0.17 on our clusters and data, indicating that our task is more difficult. In light of this, the 0.56 DI of our N2V-additive model seems quite good, while the 0.08 of W2V-random indicates poor cluster quality. But why did N2V cluster better than SVD while the two did not differ much in the cluster similarity task? DI is determined by both inter and intra distances. We found SVD has greater intra-cluster distances (0.51 inter, 1.47 intra) than N2V (0.55, 0.99) after normalization to unit vectors. This means clusters are more compact in the N2V model, potentially making the cluster similarity task easier.

K-means clustering and Centroids We clustered terms from each model into five clusters using the K-means clustering algorithm from scikit-learn (Pedregosa et al., 2011) and evaluated using three of its performance evaluation metrics: (i) ad-

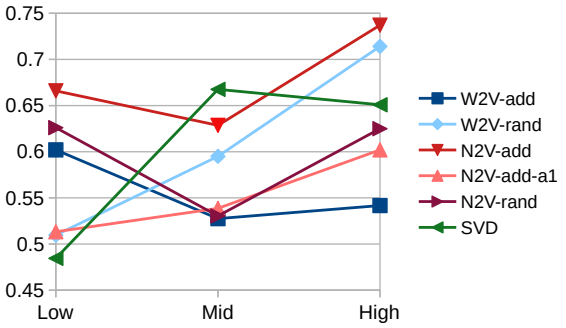

Figure 1: Similarity scores from Table 1 split by term frequency. Low $\leq$ 49 tokens, mid = 50 - 750, high $\geq$ 750.

justed Rand index, (ii) adjusted mutual information and (iii) Fowlkes-Mallows index. Results for (i) and (ii) show scores close to zero for all models within the bounded range [-1,1], indicating results close to random. The best model is the SVD model ((i) 0.12, (ii) 0.17). On (iii), with scores in range [0,1], the best N2V model (0.48) outperforms the best SVD and W2V. Manual inspection of the clusters shows that in many cases the majority of terms is put into a single cluster and the other clusters have only a few terms in them.

We also applied a centroid-based approach to evaluate clustering. We calculated the mean of the normalized vectors for each cluster to determine its centroid. We then calculated the F-score by checking for each term whether its cosine distance was closer to its cluster centroid than to another. The best performing model is W2V-random (F-score: 0.10), followed by N2V-random (0.08). All other models perform approximately equally bad (0.04).

K-nearest neighbors In our final evaluation, we classify terms into clusters using K-nearest neighbors (KNNs). We compute the macro-averaged F1 score for each term using leave-one-out cross validation. For both k=3 and k=1, the SVD model performs best with an F-score of respectively 0.45 and 0.42. For k=3, the best N2V outperforms W2V, while for k=1 the best W2V outperforms N2V, scoring almost the same as the SVD model. Manual inspection shows that for all models most of the terms from any cluster are either classified as part of the language or the meta-linguistic cluster, which are the two largest clusters.

Frequency effects To further explore our findings, we performed the cluster similarity evaluation task again, but with the target terms split by frequency. This allows us to see how the quantity of training data affects the cluster similarity. We distinguish between low-frequency terms (1-49 occurrences, n=22), medium-frequency terms (50-750 occurrences, n=55) and high-frequency terms (750-6730 occurrences, n=19, cutoffs were set to have a reasonable number of terms in the low and high frequency class). For reference, N2V was designed to train on 1-4 occurrences of a term, while for W2V, more is better. We expect additive initialization to outperform random initialization for low frequencies where an informed initialization can make up for a lack of training data.

Figure 1 shows that most models benefit from more data, but N2V clearly outperforms W2V in the low frequency condition, even with random initialization. Secondarily, models with additive initialization outperform their randomly initialized variants, possibly due to the transfer of domain-specific information for low-frequency terms noted by Bloem et al. (2019). *N2V-add-a1* forms an exception, where the high learning rate may cause massive changes to the initial vector position after only a few training occurrences, performing worse than random. SVD does not pattern with N2V here, performing very poorly on the low frequency terms. This model performs best in the 50-750 occurrence range. The SVD models cannot benefit from additive initialization and should therefore be compared to the randomly initialized models.

In the high-frequency range, N2V again performs best, probably due to the low rates of parameter decay selected in the hyperparameter tuning. As expected, W2V performs quite well with more data in its standard random initialization condition. Unexpectedly, it performs quite poorly with the additive initialization. This might be an issue with our tuning process: as all our artificial terms had a relatively low frequency of 100, the tuning task may have selected a model that relies too much on the initialization, and learns poorly for the *w2v-additive* condition. This shows the importance of the tuning data resembling the target data closely.

6 Conclusion and Discussion

The results show that, in general, N2V and SVD represent the ground truth clusters better than W2V on this type of data. Furthermore, we see that using N2V or SVD for smaller, domain-specific data outperforms a larger domain-general W2V model

trained on a large corpus. N2V is able to learn higher-quality embeddings than W2V from small texts, as it was designed to, and we confirm previous work showing that the same holds for count-based models to a limited extent.

Clustering methods (centroid and k-means) do not detect anything close to the clusters defined in the ground truth, whereas more fine-grained methods (cluster similarity and KNN) do yield results that are clearly above chance. The evaluation in terms of the Dunn index is also promising. Despite the overall low performance, we take this as an indication that the models group the terms with some systematicity. Furthermore, the rankings of the different models remain consistent across evaluations. Arriving at the same results through various methods can be seen as a fulfillment of Sommerauer and Fokkens's (2019) stress-test requirement.

Manual inspection of the clusters indicates that the imbalance in the (already very small) dataset is problematic for a K-nearest neighbors classifier, which assigned almost all words to the two biggest clusters. We expect that the same may hold for centroids and k-means. In hindsight, we could have controlled for this by extending the dataset beyond the terms in the Index of *Word and Object*. While this might have provided more accurate insights, we expect that most use-cases that work with small (or even tiny) data are most likely also working with similarly unbalanced data. Standard machine-learning techniques aiming to abstract over examples are most likely not able to pick up (potentially weak) signals based on just a few examples. We therefore consider fine-grained and example-based methods a more promising direction.

Overall, research on small data is still in an early phase. We see that models designed to work with tiny data outperform others on low-frequency terms, but yield only slightly better or comparative results when compared on mid- or high-frequency terms. It has to be considered that these models are overall very similar to the standardly used models. Future research should explore more balanced approaches which combine the strengths of both versions. For instance, by adjusting the settings of N2V based on the frequency of a target term.

From the perspective of a philosopher who may want to make use of DS models to support their work, the results we obtained in this study are not good enough yet. The minimum for exploratory work would be that the vast majority of the terms is correctly clustered and all categories are exemplified. Currently most terms are placed in the two largest categories, which might even give high accuracy but still does not represent the data well. Thereby, exploration of the data with these models could give a wrong impression about how the terms relate to each other. For hypothesis testing, the required accuracy depends on the hypothesis being tested, but in principle it is possible when the model no longer makes clear mistakes (but it may not be able to always distinguish between conceptual and relational connections and still make errors on clear borderline cases). Unfortunately, this level of accuracy was not yet reached either. More broadly, as studying language change through diachronic word embeddings adds a layer of complexity beyond the synchronic word embeddings we investigated, we expect that diachronic word embeddings trained on small data sets will not be able to reflect actual conceptual change and thus directly support philosophical research at this stage.

We do, however, think that our results have laid the groundwork for using DS models for exploratory purposes. We should keep in mind that the conceptual network based on *Word and Object* calls for very fine-grained distinctions. While this may still be too challenging, we expect that exploring differences in terms used by different authors could be more realistic.

Acknowledgements

Oortwijn and Meyer's initial contribution to this research was funded by the Vrije Universiteit's Network Institute. Sommerauer and Fokkens were funded by the Netherlands Organization of Scientific Research (NWO) PGW.17.041 awarded to Pia Sommerauer and NWO VENI grant 275-89-029 awarded to Antske Fokkens. Bloem and Oortwijn were funded by NWO VICI grant 277-20-007 awarded to Arianna Betti. Oortwijn was also funded by NWO grant 314-99-117 awarded to Bettina Speckmann and by the Human(e) AI grant *Small data, big challenges* funded by the University of Amsterdam. We would like to thank Thijs Ossenkoppele and Arianna Betti for their evaluation of the ground truth. We furthermore thank the eIdeas research group and anonymous reviewers for feedback. All remaining errors are our own.

References

Fatemeh Torabi Asr, Jon Willits, and Michael Jones. 2016. Comparing predictive and co-occurrence based models of lexical semantics trained on child-directed speech. *Cognitive Science*.

Amir Bakarov. 2018. A survey of word embeddings evaluation methods. *Computing Research Repository*, arXiv:1801.09536. Version 1.

Arianna Betti, Martin Reynaert, Thijs Ossenkoppele, Yvette Oortwijn, Andrew Salway, and Jelke Bloem. 2020. Expert concept-modeling ground truth construction for word embeddings evaluation in concept-focused domains. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6690–6702.

Arianna Betti and Hein van den Berg. 2014. Modelling the history of ideas. *British Journal for the History of Philosophy*, 22(4):812–835.

Jelke Bloem, Antske Fokkens, and Aurélie Herbelot. 2019. Evaluating the consistency of word embeddings from small data. In *Proceedings of Recent Advances in NLP*, Varna, Bulgaria. INCOMA Ltd. Shoumen, Bulgaria.

Haim Dubossarsky, Daphna Weinshall, and Eitan Grossman. 2017. Outta control: Laws of semantic change and inherent biases in word representation models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1136–1145. Association for Computational Linguistics.

Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, and Noah A Smith. 2015. Retrofitting word vectors to semantic lexicons. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1606–1615. Association for Computational Linguistics.

Christiane Fellbaum. 2010. Wordnet. In *Theory and Applications of Ontology: Computer Applications*, pages 231–243. Springer.

Annapaola Ginammi, Jelke Bloem, Rob Koopman, Shenghui Wang, and Arianna Betti. 2020. Bolzano, Kant and the traditional theory of concepts - A computational investigation [in press]. In *The Dynamics of Science: Computational Frontiers in History and Philosophy of Science*, Pittsburgh. Pittsburgh University Press.

Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. 2020. Analysing lexical semantic change with contextualised word representations. *arXiv preprint arXiv:2004.14118*.

Anna Gladkova and Aleksandr Drozd. 2016. Intrinsic evaluations of word embeddings: What can we do better? In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 36–42, Berlin, Germany. Association for Computational Linguistics.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 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.

William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016b. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501. Association for Computational Linguistics.

Johannes Hellrich and Udo Hahn. 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.

Johannes Hellrich and Udo Hahn. 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.

Aurélie Herbelot and Marco Baroni. 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.

Aurélie Herbelot, Eva von Redecker, and Johanna Müller. 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.

Jian Huang, Keyang Xu, and VG Vinod Vydiswaran. 2016. Analyzing multiple medical corpora using word embedding. In *2016 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 527–533. IEEE, IEEE.

Tom Kenter, Melvin Wevers, Pim Huijnen, and Maarten de Rijke. 2015. Ad hoc monitoring of vocabulary shifts over time. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, pages 1191–1200, New York, NY, USA. Association for Computing Machinery.

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal analysis of language through neural language models. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 61–65. Association for Computational Linguistics.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*, pages 625–635. Association for Computing Machinery.

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397. Association for Computational Linguistics.

Angeliki Lazaridou, Marco Marelli, and Marco Baroni. 2017. Multimodal word meaning induction from minimal exposure to natural text. *Cognitive Science*, 41:677–705.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Molly Lewis and Gary Lupyan. 2019. What are we learning from language? Associations between gender biases and distributional semantics in 25 languages.

Thang Luong, Richard Socher, and Christopher Manning. 2013. Better word representations with recursive neural networks for morphology. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 104–113. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 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. Neural Information Processing Systems Foundation, Inc.

George A Miller. 1995. Wordnet: a lexical database for English. *Communications of the Association for Computing Machinery*, 38(11):39–41.

Sunny Mitra, Ritwik Mitra, Suman Kalyan Maity, Martin Riedl, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2015. An automatic approach to identify word sense changes in text media across timescales. *Natural Language Engineering*, 21(5):773.

Syrielle Montariol and Alexandre Allauzen. 2019. Empirical study of diachronic word embeddings for scarce data. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 795–803. INCOMA Ltd. Shoumen, Bulgaria.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Willard Van Orman Quine. 1960. *Word and Object*. MIT press.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling With Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA.

Alex Rosenfeld and Katrin Erk. 2018. Deep neural models of semantic shift. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 474–484. Association for Computational Linguistics.

Arpita Roy, Youngja Park, and Shimei Pan. 2019. Incorporating domain knowledge in learning word embedding. In *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 1568–1573. IEEE, IEEE.

Magnus Sahlgren and Alessandro Lenci. 2016. The effects of data size and frequency range on distributional semantic models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 975–980. Association for Computational Linguistics.

Philippa Shoemark, Farhana Ferdousi Liza, Dong Nguyen, Scott Hale, and Barbara McGillivray. 2019. Room to glo: A systematic comparison of semantic change detection approaches with word embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 66–76. Association for Computational Linguistics.

Anders Søgaard. 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, Association for Computational Linguistics.

Pia Sommerauer and Antske Fokkens. 2019. Conceptual change and distributional semantic models: an

exploratory study on pitfalls and possibilities. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 223–233. Association for Computational Linguistics.

Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2018. Survey of computational approaches to diachronic conceptual change. *CoRR*, abs/1811.06278.

Nina Tahmasebi and Thomas Risse. 2017. On the uses of word sense change for research in the digital humanities. In *International Conference on Theory and Practice of Digital Libraries*, pages 246–257. Springer, Cham Springer 2017.

Astrid van Aggelen, Antske Fokkens, Laura Hollink, and Jacco van Ossenbruggen. 2019. A larger-scale evaluation resource of terms and their shift direction for diachronic lexical semantics. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 44–54. Linköping University Electronic Press.

Hein van den Berg, Arianna Betti, Thom Castermans, Rob Koopman, Bettina Speckmann, Kevin Verbeek, Titia van der Werf, Shenghui Wang, and Michel A Westenberg. 2018. A philosophical perspective on visualization for digital humanities. In *3rd Workshop on Visualization for the Digital Humanities (VIS4DH2018)*. VIS4DH.

Pauline van Wierst, Sanne Vrijenhoek, Stefan Schlobach, and Arianna Betti. 2016. Phil@ Scale: Computational methods within philosophy. In *CEUR Workshop Proceedings*, volume 1681. CEUR Workshop Proceedings.

Ekaterina Vylomova, Sean Murphy, and Nicholas Haslam. 2019. Evaluation of semantic change of harm-related concepts in psychology. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 29–34. Association for Computational Linguistics.

Melvin Wevers and Marijn Koolen. 2020. Digital begriffsgeschichte: Tracing semantic change using word embeddings. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, pages 1–18.

Gerhard Wohlgenannt, Ariadna Barinova, Dmitry Ilvovsky, and Ekaterina Chernyak. 2019. Creation and evaluation of datasets for distributional semantics tasks in the digital humanities domain. *Computing Research Repository*, arXiv:1903.02671. Version 1.