Domain-Specific Evaluation of Word Embeddings for Philosophical Text using Direct Intrinsic Evaluation

Goya van Boven
Utrecht University
j.g.vanboven@students.uu.nl

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

Abstract

We perform a direct intrinsic evaluation of word embeddings trained on the works of a single philosopher. Six models are compared to human judgements elicited using two tasks: a synonym detection task and a coherence task. We apply a method that elicits judgements based on explicit knowledge from experts, as the linguistic intuition of non-expert participants might differ from that of the philosopher. We find that an in-domain SVD model has the best 1-nearest neighbours for target terms, while transfer learning-based Nonce2Vec performs better for low frequency target terms.

1 Introduction

Many applications of Artificial Intelligence methods to textual data rely on language models pretrained on large amounts of web text. However, this does not necessarily yield models suited to the analysis of texts in the humanities, as such texts may deviate in style, vocabulary, register and other regards. On the other hand, models trained on texts from a specific humanities domain have far less data available to learn from. As a trade-off, transfer learning can be applied, where a model is first trained on a large, general-domain dataset and then tuned on a smaller, domain-specific dataset. We compare several tuning approaches based on Word2Vec (Mikolov et al., 2013a) to a baseline of a SVD model trained only on domain-specific data, for the purpose of learning meaning representations of the terminology of a specific philosopher.

In philosophy, there is interest in supporting the close reading of texts with the use of information retrieval methods (Ginammi et al., 2021) based on distributional semantic (DS) models (Turney and Pantel, 2010; Erk, 2012; Mikolov et al., 2013a) to provide a different perspective on texts (Herbelot et al., 2012). Philosophical terms have domain-specific meanings: for example an *accident* is a non-essential property of an entity, rather than an unfortunate incident. Thus, domain-specific models and methods of evaluation are necessary.

DS models are often evaluated by comparing their performance to a *gold standard*, such as the SimLex-999 dataset (Hill et al., 2015). However, these similarity rankings concern general language terms and their typical senses, rather than domain-specific philosophical terms. Meanings of terms may differ between philosophers or even within the works of a single philosopher. Rather than modeling a standard jargon that a group of people uses, we aim to model the semantics of some particular philosopher, with no ‘native speaker’ besides that philosopher. Any evaluation of the quality of the semantic representations in this kind of model would require expert knowledge. For this reason we apply the direct intrinsic evaluation methods proposed by van Boven and Bloem (2021) with expert participants. We evaluate six models: we use Wikipedia data as a general-domain text corpus for training, and we use a domain-specific corpus of the works of Willard V. O. Quine for tuning. Quine was a 20th century American philosopher, whose works are still of great interest to philosophers, logicians and linguists. This evaluation will show us which tuning approach, if any, performs best for creating meaning representations of philosophical terms, and for digital humanities applications more broadly.

2 Related work

Suissa et al. (2022) present an overview of AI-based text analysis in digital humanities, arguing that lack of data availability characterizes the field and that domain adaptation is essential. Sommerauer and Fokkens (2019) discuss the difficulties of applying distributional semantic models to study conceptual change in digital humanities, drawing attention to frequency effects and effects of random initialization, and the importance of studying domain-relevant exemplar terms.

The 2nd International Workshop on Natural Language Processing for Digital Humanities (NLP4DH), pages 101–107
November 20, 2022. ©2022 Association for Computational Linguistics

Various digital humanities studies using word embeddings have been published, but they rarely include in-domain evaluations of those embeddings. Bjerva and Praet (2015) apply word embeddings to study relationships between persons in 6th century Latin text, but do not evaluate their model. Nelson (2021) train domain-specific word embeddings on an 18M word corpus of narratives on slavery in the American South, but do not evaluate them. Meinecke et al. (2019) use domain-specific vectors for aligning medieval text versions, extrinsically evaluating by having an expert manually inspect the resulting alignments, but without intrinsically evaluating the embeddings.

Kenter et al. (2015) use word embeddings to study vocabulary shifts and have human annotators associate words to topics and time periods for evaluation. They do not use any pre-trained models. Wohlgenannt et al. (2019) do perform in-domain evaluation, evaluating word embedding models trained on two fantasy novel book series of about 1M tokens each, manually constructing test datasets with domain experts. They compare domain-specific models including Word2Vec to a transfer learning setup where a pretrained Word2Vec model is tuned on the fantasy novel corpus. They find that an in-domain Word2Vec model outperforms the other approaches in an analogy task and a word intrusion task. Todorov and Colavizza (2020) find that fine-tuning pre-trained BERT embeddings does not help for named entity recognition in historical corpora, though this is in addition to the use of in-domain FastText embeddings.

In the philosophical domain, several domain-specific evaluation methods have been proposed, but none have directly evaluated model output. Evaluated models include the widely used $Word2Vec$ (W2V; Mikolov et al., 2013a,b) predictive model, $Nonce2Vec$ (N2V; Herbelot and Baroni, 2017) which is an adaptation of the skipgram W2V model designed for learning from few training examples in *tiny* text corpora, and count-based SVD models in the Levy et al. (2015) implementation.

Avoiding the issue of obtaining expert knowledge, Bloem et al. (2019) evaluate these models using a metric of model consistency, which rewards models that yield similar vectors when trained on different samples of the same target term within the same domain. They found that N2V outperforms a SVD baseline by this metric.

What word is most related to 'Information' ?	
a) *Learning*	b) *Reductions*
c) *Collateral*	d) *Application*
e) *Ordered Pair*	f) *None of these words is even remotely related*

Table 1: Synonym detection task example question. Here, the options are the k-nearest neighbours of target word 'information' of the various evaluated models.

What word does not belong to the group?	
a) *Numbers*	b) *Pronouns*
c) *Subtraction*	d) *Actually*

Table 2: Coherence task example question, with target word a), nearest neighbours b) and c), and outlier d).

Oortwijn et al. (2021) evaluated these models based on a conceptual network they constructed, comparing the similarity of learned embeddings for specific philosophical target terms to their position in this network. They found that domain-specific N2V and the count-based baseline models outperformed a domain-general W2V model. As the network was pre-defined, only a limited set of terms could be considered. Betti et al. (2020) propose the use of a more elaborate ground truth in evaluation that includes many relevant as well as irrelevant terms, centered around a specific concept as defined by a specific philosopher. This still would not account for creative model output.

Lastly, van Boven and Bloem (2021) proposed the use of direct intrinsic evaluation to provide more comprehensive coverage of anything the models might output. Methods from Schnabel et al. (2015) are adapted to the scenario of eliciting expert knowledge where experts respond to nearest neighbour words that the model generates. Van Boven and Bloem (2021) report competitive interrater agreement scores between experts for this method. We adopt this approach for philosophical domain model evaluation.

3 Methods and data

The direct intrinsic evaluation method consists of a *synonym detection task* and a *coherence task*, adapted from Schnabel et al. (2015). Synonym detection entails selecting the most related word to target word *t* from a set of words, which are the k-nearest neighbours of *t* in each included model. In this task, participants thus indicate their preference between the outputs of all evaluated models. Table 1 illustrates the set-up of this task. In the coherence task, which is illustrated in Table 2, expert participants are asked to identify a semantic outlier in a set of words, which is less close to t in the model than the other options. The aim of this task is to assess whether groups of words are mutually related in a small neighbourhood in the embedding, evaluating model coherence. Here, each model is evaluated individually. The combination of the two tasks provides insight into the absolute as well as the relative performance of the models. We refer to van Boven and Bloem (2021) for further details on the tasks, and an evaluation of the method. The evaluation tasks were conducted through online surveys on the platform *Qualtrics*[1].

Following Schnabel et al. (2015), we analyse the results of both tasks through a random permutation test, with the number of permutations $n = 100,000$. We use the difference in mean scores (i.e. the percentage of votes) between models as our test statistic for the synonym detection task, and for the coherence task we use the precision scores (i.e. the proportion of correctly identified outliers per model). As each model has its own qualities and the tasks evaluate different aspects, it is possible for the two tasks to yield different 'winners'.

3.1 Data

As training data we use a 140M token domain-general Wikipedia corpus and the 2.15M token QUINE corpus (v0.5, Betti et al., 2020), which includes 228 philosophical articles, books and bundles written by Quine. Following van Boven and Bloem (2021) we use the test set for the influential book *Word & Object* (Quine, 1960) by Bloem et al. (2019) as target terms for evaluation. This test set contains 55 terms selected from the index of the book, of which we use 25 in Experiment 1, 14 in Experiment 2 and 6 in both of the experiments.[2] For models that processed both datasets, the target terms were marked in the QUINE corpus so that embeddings for target terms in both corpora were learned independently of each other.

3.2 Models

We compare a $W2V$ model, two instances of a count-based SVD and three instances of the $N2V$ model, which we chose for comparability to Oortwijn et al.'s (2021) evaluation on this data. Nonce2Vec works by training target terms and their in-domain context sentences into a general-domain background model (trained on Wikipedia). We apply Word2Vec in a similar setup where we continue training the Wikipedia model only on a target term's context sentences from the QUINE corpus. This is done for comparability with N2V. Therefore, our $W2V$ and $N2V$ models are all trained on the Quine dataset with the Wikipedia corpus as a background model in a transfer learning setup, and we test different forms of transfer learning. We did not include a GloVe model as it is trained similarly to SVD and typically performs similarly to Word2Vec, and additional models would make the annotation task too lengthy for the domain experts. Rather than including BERT, we use only type-based embeddings because we consider them more transparent to the domain experts; each type only has one representation which is somewhat interpretable. We refer to Ehrmanntraut et al. (2021) for further arguments in support of their use in digital humanities applications, such as better performance on small datasets.

The first variant of the $N2V$ model, $N2V_{Add}$ is N2V's additive baseline model used for its initialization, which simply sums Wikipedia background vectors of a target term's context words (Lazaridou et al., 2017). $N2V_{Def}$ uses the default hyperparameters of Herbelot and Baroni (2017), tuned for very small datasets. $N2V_{Con}$ is tuned on Bloem et al.'s (2019) consistency metric. Learning rates and decays are lower in $N2V_{Con}$ than in the $N2V_{Def}$ model, so the tuning is less strong. The $W2V$ model is trained over 5 epochs, with start α = 1.0 and end α = 0.1.

For the count-based models, the *Hyperword* SVD-PPMI implementation of Levy et al. (2015) is used with a window size of 5. We have a transfer setup trained on both the in-domain and general domain corpus (SVD_{Q+W}), and lastly we have a baseline without transfer learning trained only on the in-domain corpus (SVD_Q).

Based on previous philosophy domain evaluations and other work indicating relatively poor W2V performance on smaller datasets (Asr et al., 2016) and rare words (Luong et al., 2013; Herbelot and Baroni, 2017), we expect N2V and SVD to outperform W2V. Furthermore, we expect N2V to outperform SVD, as this was the outcome of previous

[1] www.qualtrics.com

[2] The terms and their frequencies are those available in van Boven and Bloem's (2021) supplementary materials at https://github.com/gvanboven/direct-intrinsic-evaluation

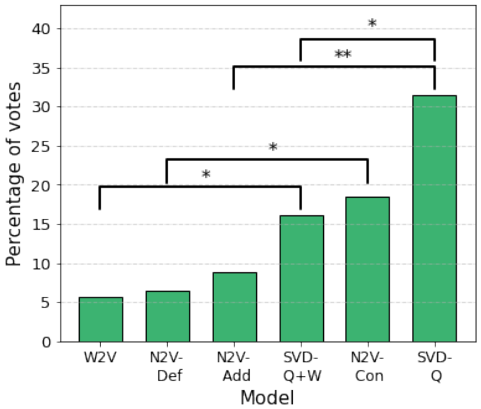

Figure 1: Results of the synonym detection task. $*$ indicates $p < 0.05$, $**$ indicates a significant p-value after Bonferroni-correction (Bonferroni, 1936).

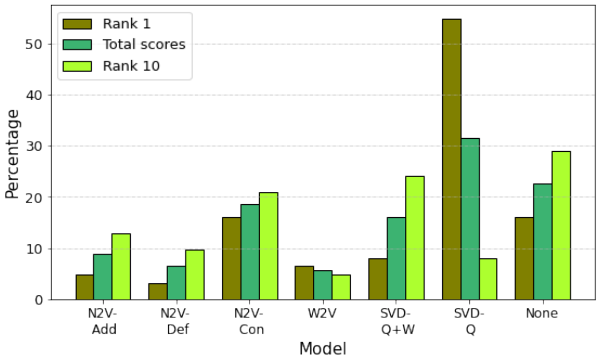

Figure 2: Synonym detection task: scores by nearest neighbour rank of the test items.

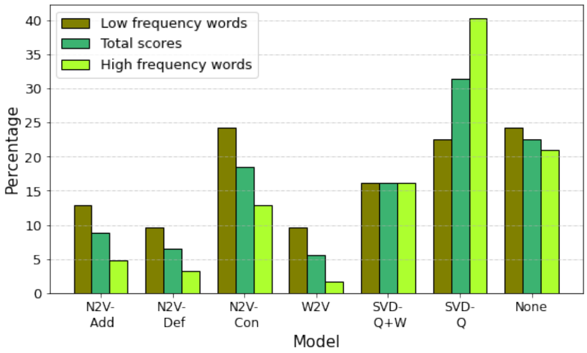

Figure 3: Synonym detection task: scores split by term frequency n. For high frequency terms, $n > 275$ and for low frequency terms $n < 84$.

evaluations by other metrics. Nevertheless, N2V is designed for tiny rather than small data. For this data size, count-based models have been shown to perform well (Sahlgren and Lenci, 2016).

4 Results

4.1 Experiment 1: Synonym detection task

Three experts on the work of Quine, who all hold a Master's degree in philosophy and have studied his work extensively, participated in this experiment. They evaluated 31 target words on two nearest neighbour ranks k, with $k \in \{1, 10\}$. We compare the mean scores of all model combinations, resulting in 15 comparisons. The total number of cast votes for best synonym is 136.[3] The data from one of the participants was excluded, as the participant indicated that the task was too difficult.

The overall scores, including the comparisons that were found to be significant, are shown in Figure 1. Figure 2 displays the scores by rank. SVD_Q receives most (31.5%) of the votes, but performs by far best on the rank 1 nearest neighbours. SVD_{Q+W} performs best on rank 10 nearest neighbours. Significant differences are found between SVD_Q and $N2V_{Add}$ ($p = 0.00079$), $N2V_{Def}$ ($p = 0.00014$) and $W2V$ ($p = 0.00009$). Inter-rater agreement is $\kappa = 0.492$. Figure 3 displays the scores split by term frequency, where $N2V_{Con}$ scores best for low frequency words and SVD_Q on high frequency words.

Surprisingly, the $N2V_{Def}$ model performs poorly even compared to $N2V_{Add}$, the Nonce2Vec baseline. This could be because $N2V_{Def}$ is designed for learning from only a few occurrences. Conversely, as the learning rate and its decay are lower in the $N2V_{Con}$ model, it may be better for representing small but not tiny datasets.

4.2 Experiment 2: Coherence task

In this task, we include the two best performing models (SVD_Q and $N2V_{Con}$) and the model that obtains the lowest score ($W2V$) in Experiment 1. Only three models were selected because they have to be evaluated one-by-one in this task. All models are tested on 20 target words and evaluated by the two expert participants. The total amount of ratings is 40 for each model. Figure 4 shows that $W2V$ performs significantly worse than both $N2V_{Con}$ and SVD_Q, while the difference between the two latter models is not significant. Inter-rater agreement is $\kappa = 0.345$.[4] Figure 5 and 6 display the scores split by term frequency, where we find

[3] 124 ratings + 12 votes that counted double as the selected option word is returned by multiple models

[4] As van Boven and Bloem (2021) discuss, the observed inter-rater agreement rates are similar to those of traditional semantic annotation tasks involving implicit knowledge.

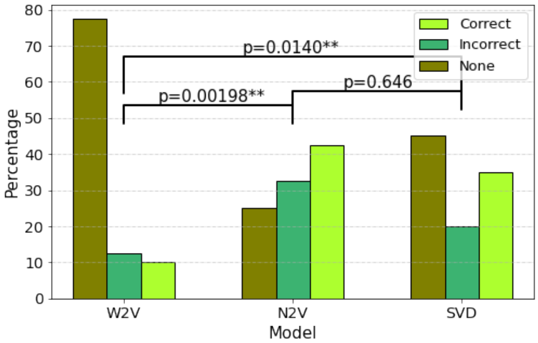

Figure 4: Results of the coherence task. *Correct* indicates the outlier was correctly identified, *Incorrect* means another words was chosen as the outlier, and *None* indicates that the option *"No coherent group can be formed from these words"* was selected.

that the best scoring $N2V_{Con}$ also performs best for high frequency words, while for low frequency words $N2V_{Con}$ and SVD_Q obtain the same score.

5 Discussion

The models that perform best are SVD_Q (the non-transfer learning baseline) and $N2V_{Con}$. SVD_Q received the most votes in the synonym detection task, and performed especially well at producing rank 1 nearest neighbours related to the target term. $N2V_{Con}$ scored higher in the coherence task, producing better clusters of related top neighbours, and producing better rank 10 neighbours in Experiment 1. This suggests SVD_Q would be more suitable for applications where top 1 precision is important, while $N2V_{Con}$ would do better in exploratory applications where a larger range of related terms is examined. The poor performance of the popular $W2V$ model for this domain and corpus size is in line with the findings of Oortwijn et al. (2021).

Both the count-based approach and the predictive approach produced a well-performing model. The hyperparameters and the transfer learning setup seemed to affect the scores more than the chosen approach. This matches Levy et al.'s (2015) claim that design choices and parameter settings influence embedding quality more than the model.

For philosophical inquiry it is important that representations of low frequency words are good, as few resources are available and low frequency terms can be crucial to understanding a concept. In the synonym detection task, $N2V_{Con}$ does better on low frequency words while SVD_Q does better on high frequency words. Conversely, in the coherence task, $N2V_{Con}$ does better on high frequency words, though its low frequency performance is still equal to SVD_Q. As $N2V_{Con}$ scored well overall on our low frequency target terms and in Bloem et al.'s (2019) consistency evaluation, it appears the most promising for modeling philosophical terms. However, it is remarkable that a simple SVD baseline performs so well compared to W2V-based transfer learning approaches. Together with results from other work (Wohlgenannt et al., 2019), this suggests that in digital humanities applications, in-domain data should be favoured over transfer learning approaches at least when 1M tokens of training data (or more)[5] are available.

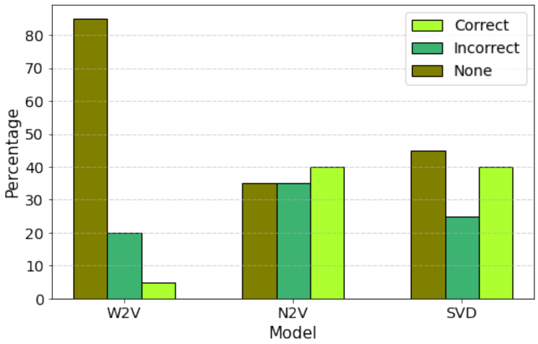

Figure 5: Scores for all models in the coherence task for low frequency terms ($n < 142$) only.

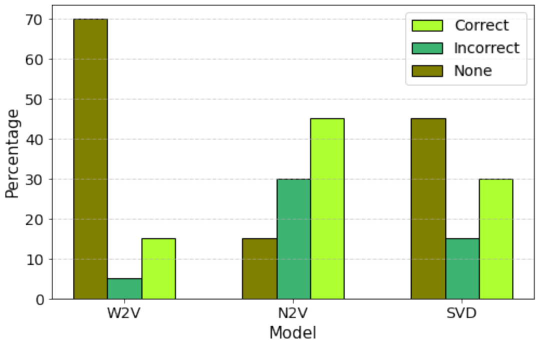

Figure 6: Scores for all models in the coherence task for high frequency terms ($n > 187$) only.

Acknowledgements

We are grateful to Yvette Oortwijn, Thijs Ossenkoppele and Arianna Betti for their input as Quine domain experts. This research was supported by VICI grant *e-Ideas* (277-20-007), financed by the Dutch Research Council (NWO).

[5] Value based on our and Wohlgenannt et al.'s (2019) in-domain corpus size, as well as the dataset sizes used in Sahlgren and Lenci's (2016) ablation study.

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. In *Annual Conference of the Cognitive Science Society*.

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, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Johannes Bjerva and Raf Praet. 2015. Word Embeddings Pointing the Way for Late Antiquity. In *Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, pages 53–57, Beijing, China. Association for Computational Linguistics.

Jelke Bloem, Antske Fokkens, and Aurélie Herbelot. 2019. Evaluating the Consistency of Word Embeddings from Small Data. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 132–141, Varna, Bulgaria. INCOMA Ltd.

Carlo Bonferroni. 1936. Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze*, 8:3–62.

Anton Ehrmanntraut, Thora Hagen, Leonard Konle, and Fotis Jannidis. 2021. Type-and Token-based Word Embeddings in the Digital Humanities. In *Proceedings of the Conference on Computational Humanities Research*, pages 16–38, Amsterdam, The Netherlands.

Katrin Erk. 2012. Vector space models of word meaning and phrase meaning: A survey. *Language and Linguistics Compass*, 6(10):635–653.

Annapaola Ginammi, Jelke Bloem, Rob Koopman, Shenghui Wang, and Arianna Betti. 2021. Bolzano, Kant and the Traditional Theory of Concepts - A Computational Investigation [final author version after R&R submitted 12 Sep, 2020]. In Andreas de Block and Grant Ramsey, editors, *The Dynamics of Science: Computational Frontiers in History and Philosophy of Science*. Pittsburgh University Press, Pittsburgh.

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, Copenhagen, Denmark. 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, Avignon, France. Association for Computational Linguistics.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. SimLex-999: Evaluating Semantic Models With (Genuine) Similarity Estimation. *Computational Linguistics*, 41(4):665–695.

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*, pages 1191–1200.

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.

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, Sofia, Bulgaria. Association for Computational Linguistics.

Christofer Meinecke, David Joseph Wrisley, and Stefan Jänicke. 2019. Automated Alignment of Medieval Text Versions based on Word Embeddings. In *LEVIA'19: Leipzig Symposium on Visualization in Applications 2019*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient Estimation of Word Representations in Vector Space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed Representations of Words and Phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Laura K Nelson. 2021. Leveraging the alignment between machine learning and intersectionality: Using word embeddings to measure intersectional experiences of the nineteenth century US South. *Poetics*, 88:101539.

Yvette Oortwijn, Jelke Bloem, Pia Sommerauer, Francois Meyer, Wei Zhou, and Antske Fokkens. 2021. Challenging distributional models with a conceptual network of philosophical terms. In *Proceedings of* *the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2511–2522, Online. Association for Computational Linguistics.

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

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, Austin, Texas. Association for Computational Linguistics.

Tobias Schnabel, Igor Labutov, David Mimno, and Thorsten Joachims. 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.

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, Florence, Italy. Association for Computational Linguistics.

Omri Suissa, Avshalom Elmalech, and Maayan Zhitomirsky-Geffet. 2022. Text analysis using deep neural networks in digital humanities and information science. *Journal of the Association for Information Science and Technology*, 73(2):268–287.

Konstantin Todorov and Giovanni Colavizza. 2020. Transfer learning for named entity recognition in historical corpora. In *CLEF (Working Notes)*.

Peter D Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37:141–188.

Goya van Boven and Jelke Bloem. 2021. Eliciting Explicit Knowledge From Domain Experts in Direct Intrinsic Evaluation of Word Embeddings for Specialized Domains. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 107–113, Online. Association for Computational Linguistics.

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