Comparing Contextual and Static Word Embeddings with Small Philosophical Data

Wei Zhou
Institute for Natural
Language Processing
University of Stuttgart
weizhou14330@gmail.com

Jelke Bloem
Institute for Logic,
Language and Computation
University of Amsterdam
j.bloem@uva.nl

Abstract

For domain-specific NLP tasks, applying word embeddings trained on general corpora is not optimal. Meanwhile, training domain-specific word representations poses challenges to dataset construction and embedding evaluation. In this paper, we present and compare ELMo and Word2Vec models trained/finetuned on philosophical data. For evaluation, a conceptual network was used. Results show that contextualized models provide better word embeddings than static models and that merging embeddings from different models boosts task performance.

1 Introduction

Statistical distributions of terms in context can be used to characterize their semantic behavior (Lenci, 2018). This is the fundamental idea that distributional models of language are built upon. When trained on large corpora, these models can provide valid word representations which can be further utilized in various downstream NLP tasks. Two common models are Word2Vec's skipgram model (W2V, Mikolov et al., 2013) and the ELMo model (Peters et al., 2018). Although word embeddings pretrained on large corpora provide good meaning representations, using them in domain-specific tasks does not achieve good results (Nooralahzadeh et al., 2018). This is because the semantic space of a certain domain can be different from that of general language. For instance, the word *substance* refers to matter in ordinary language but in philosophical contexts it is a technical term from metaphysics pertaining to entities (Robinson, 2020).

Contextualized models like ELMo address this problem to some extent, but require a lot of domain-specific data to obtain a tailored model, while such datasets are typically smaller. The main contributions of this paper are:

- We trained and finetuned ELMo models with a philosophical corpus.
- We examined and compared models with contextual embeddings and static embeddings with intrinsic evaluations.
- We experimented with combining finetuned W2V and pretrained ELMo embeddings for representations of philosophical terms.

2 Related work

Efforts have been made in the following directions with regards to creating domain-specific word embeddings. Firstly, the construction of domain-specific corpora. Roy et al. (2017) appended manual annotations (predicate-argument structure) to training data in the field of cybersecurity. The additional annotation makes the original dataset more suited to the task of training cybersecurity embeddings. Secondly, refinement can be carried out on existing embeddings. Boukkouri et al. (2019) combined W2V embeddings trained on a small domain-specific corpus with ELMo embeddings and evaluated them on a clinical entity recognition task. They found combined embeddings outperformed embeddings trained on large corpora in the medical domain. Lastly, one can also explore suitable models for training with small amount of data. Herbelot and Baroni (2017) proposed a refined W2V model called Nonce2Vec (N2V), which learns word meanings from tiny data. The N2V model takes a high-risk learning approach with heightened learning rate and larger window size to process contexts greedily. Besides N2V, simple additive models have also proven to work well on small data (Lazaridou et al., 2016; Bloem et al., 2019).

As for evaluations of domain-specific word embeddings, the usual approach is to design in-domain

tasks (Nooralahzadeh et al., 2018) or ground truths (Betti et al., 2020; Oortwijn et al., 2021). However, many domains lack such evaluation data. Bloem et al. (2019) proposed a general evaluation metric of *consistency*, based on the idea that a stable model could provide similar word embeddings given the same term across similar sources.

3 Task Description

The overall goal of this paper is to examine and compare different models by evaluating word embeddings trained on a small philosophy corpus.

3.1 Dataset

We used the dataset from Bloem et al. (2019), consisting of version 0.4 of the QUINE corpus (Betti et al., 2020) and evaluation terms. This corpus is made up of all philosophical texts written by the author Willard Van Orman Quine, consisting of 228 articles, books and bundles. The corpus consists of OCR-processed, manually corrected text and contains about 2 millions tokens after tokenization.

3.2 Model

ELMo We trained two ELMo models of different sizes and finetuned one. For training, we used the above dataset with a split of training data (17000 sentences) and testing data (5016 sentences). For finetuning, we continued training a pre-trained ELMo model[1] on the philosophical texts. Key training parameters for the three ELMo models can be found in Appendix A. The learning rate was set to default (0.2) for all models.

Word2Vec We trained Word2Vec skipgram models in the Gensim (Rehurek and Sojka, 2011) implementation with our data based on a pretrained-256 dimensional embedding model: the Nonce2Vec background model (Herbelot and Baroni, 2017) trained on Wikipedia data. We used consistency (Bloem et al., 2019) as a metric to choose the best hyper-parameters. It is measured as cosine similarity between two vectors of the same seed word. Our seed words were chosen from general philosophical terms[2] (Appendix B), excluding target terms from the Quine dataset used for evaluation (Appendix D). We abandoned terms ending with -ism and multi-token terms and selected terms whose frequency is over 50 in our corpus. We selected sentences containing seed words, divided each selected set into two parts and combined each part with the rest of the corpus. As a result, we have three corpora in total: the whole corpus and two sub-corpora. The model with the background semantic space is most consistent with a learning rate of 0.005 and led to 0.97 cosine similarity.

Nonce2Vec We trained a Nonce2Vec (N2V) model with consistency as the metric for tuning hyper-parameters. Unlike the W2V models, N2V only changes the embeddings of targeted terms, with the remaining semantic space frozen. We measured consistency with seed words, as we did with W2V. Since N2V is designed to be trained on "tiny data", we limited the contexts of each target term to up to 10 sentences both during tuning and model training. With the best selected hyperparameters, the model has a 0.97 consistency score.

3.3 Combined Embeddings

According to Boukkouri et al. (2019), combining contextualized word embeddings with their static counterparts works better on downstream tasks than merely using contextualized or static ones. The combination methods used in their paper were concatenation and addition. In our study, we further explored whether assigning weight works better than simply adding the two types of embeddings. The new embeddings are defined as $E_{mix} = \alpha * E_{elmo} + (1 - \alpha) * E_{w2v}$, where α is the weight assigned to the ELMo embeddings and $(1 - \alpha)$ the W2V. We experimented 11 values from 0 to 1 for α with an interval of 0.1.

3.4 Evaluation

We evaluated models based on word embeddings of specific terms. These terms were proposed by Oortwijn et al. (2021) as a ground truth for evaluation. They constructed a conceptual network of all relevant index terms of Quine's *Word and Object* (1960). The index terms were categorized by domain experts into one of the six clusters they defined (language, ontology, reality, mind, metalinguistic and relational terms, reproduced in Appendix D. We generated word embeddings for the 73 terms in the first five categories. 30 of these terms have a frequency less than 100 in the corpus ($n < 100$), 9 terms over 1000 ($n \geq 1000$) and 34 in between ($100 \leq n < 1000$). For ELMo, type embeddings were generated by averaging token embeddings for the same type in different contexts. For multi-token terms that were not in the model's

[1] https://github.com/allenai/bilm-tf

[2] source1 URL, source2 URL

vocabulary, we used the averaged embeddings to represent the whole. This is done for all models and evaluations and no other multiword term processing takes place (e.g. on the corpus). The embeddings were evaluated by the following metrics:

Cluster similarity Following Oortwijn et al. (2021), for each term, we sampled a term in the same category and a term in the different category and compared their similarity with the original term. We then calculated the probability that the cosine similarity between the same category terms was higher than that of a different category. We performed the sampling process 100 times for each terms and averaged the scores as our final scores.

Rank For each target term, we find the top 5 nearest (besides itself) terms by cosine similarity. Each term accounts for 0.2 score if it is in the same category as the target term. The highest score for a target term is therefore 1. We then added up the scores for all target terms as the rank score. There are 73 terms in total. However, the highest rank score is not 73, but 71.4: in the category *Mind*, there are only two terms, which means for each term in *Mind*, the highest score is 0.2 rather than 1.

Dunn index is used to measure how well embeddings of terms in the same category cluster (following e.g. Huang et al., 2016). A higher number suggests better clustering, which means a small variance between members of a cluster, and large differences between means of each cluster.

Gap is similar to cluster similarity, except that we consider pairs of all terms in this case. We calculated the cosine similarity between each two terms. We then averaged the overall similarity of the terms from the same sets and from the different sets and got their gaps.

4 Results

The main results are shown in Table 1. Our results show that, except for the pretrained ELMo model, ELMo models generally provide better embeddings than the W2V model. This might be attributed to the sequential structure of ELMo, which encodes neighbouring information based on contexts. To better understand the performance scores, we divided the results of rank score and cluster similarity into two conditions, namely the single-token terms and multi-token terms. Table 2 shows the results. The rank score and cluster similarity of the W2V model are lower than those of the ELMo small model in both single and multi-token terms' cases. However, we found that in the single-token terms case, there is a bigger difference of the rank (2.8 versus 0.8) and cluster similarity (0.13 versus 0.04) between the two models. This might be because the meaning of multi-token terms are less context dependent. Since we averaged the embeddings for each subtoken within the multi-tokens, the final representation of the multi-tokens already encodes some neighbouring information. By contrast, for single-token terms, ELMo is better in incorporating neighbouring information than the W2V model.

As for the combined models, we found that the rank score performance increased greatly from 39.4 (only ELMo) to around 44 (combined). However, there is nearly no difference between the combined models and the finetuned W2V. This suggests our finetuned W2V model already provides a reasonable semantic space for the Quine data, and adding additional information does not improve it. We also experimented with merging the embeddings from both models. The results for the rank score and cluster similarity are shown in Figure 2. Contrary to our expectation that increasing the portion of pretrained ELMo decreases both scores, there is

Model	**Sim**	**Rank**	**Dunn**	**gap**
E_s	0.69	**49.2**	0.44	0.08
E_m	0.69	46.6	0.37	0.07
E_pre	0.65	39.4	0.41	0.05
E_ft	**0.74**	48.0	0.40	0.10
W2V_ft	0.65	45.6	0.39	0.07
N2V	0.67	43.2	**0.52**	**0.14**
E_preW_ft+	0.66	44.8	0.43	0.06
E_preW_ftc	0.67	44	0.42	0.05

Table 1: Evaluation results. E = ELMo, s = small, m = medium, pre = pretrained, ft = finetuned. The last two models provide combined embeddings, where + = addition, c = concatenation. All models have dim=256 except for E_m and E_preW_ftc with dim=512.

Term	**single-token**		**multi-token**	
Model	**Rank**	**Sim**	**Rank**	**Sim**
E_s	23.2	0.69	26	0.79
W2V_ft	20.4	0.56	25.2	0.75
Δ	**2.8**	**0.13**	0.8	0.04

Table 2: Results of single and multi-token terms on rank and cluster similarity. For Sim, only the original terms were considered, instead of resampled ones.

a peak for both scores when the portion of W2V embeddings is around 0.3-0.4. It seems that combining pretrained language model embeddings with W2V needs to be examined carefully to find the sweetest point. Non-linear combination could also be explored in future work.

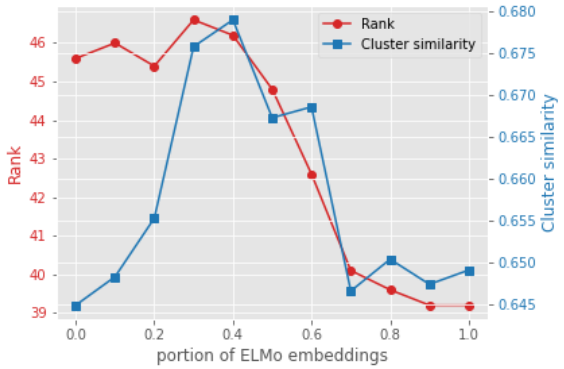


Figure 1: Rank score and cluster similarity of the merged embeddings as the portion of ELMo embedding is increased from 0 to 1.

From Table 1 and consistent with Oortwijn et al. (2021), we observed that the N2V model provides the highest Dunn index (0.52). When we scale the distance numbers to the same level for all models, we observe that the maximal intra-cluster distance in N2V is smaller than in other models. One reason could be that due to limited contexts and increased learning rate, the N2V model aggressively learns new meanings so the new meanings encode less noisy information, such as old meanings or contextual meanings. This enables outliers to be closer to their cluster centroids, resulting in a lower maximal intra-cluster distance used in Dunn index calculation. A higher intra-cluster similarity could also explain why the N2V model has a higher gap score.

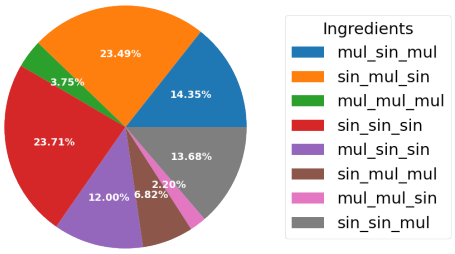


Figure 2: Distributions of eight types of errors from ELMo small model. Mul = multi-token, sin =single-token. The order of the token type corresponds to: original terms, same-cluster terms, different-cluster terms.

Semantic error analysis of terms in this dataset can only be performed by Quine domain experts. However, we can examine some superficial features. From Table 2, we observed different performance from single-token and multi-token terms. To examine the influence of single/multi-token terms on evaluation scores, we took both the correct and incorrect cases from cluster similarity and categorized them into 8 types (2*2*2) based on the token type (single or multi) of original terms, sampled-same-group term and sampled-different-group term. Figure 2 shows the results for the error case. The all-single-term case which accounts for the largest portion, nearly one fourth of all errors. The next two biggest error sources are confusion between the single and multi-token terms: in the sin_mul_sin case, instead of predicting the original term (sin) and same-cluster term (mul) to be more similar, the model predicted the original term and different-cluster terms (sin) as more similar. The same observation can be found in the mul_sin_mul case. When we look at the mul_mul_sin and the sin_sin_mul types from the correct case, we found they together account for nearly a half of all correct cases. This indicates that terms of the same type (single/multi) have the tendency to be closer, which could be the result of averaging subtoken embeddings in ELMo, comparable to the *sum effect* observed by Kabbach et al. (2019). We present the term distribution from the ELMo small model in Appendix C. We conclude that full multi-token term processing would be preferable but small datasets may not provide enough instances of each. N2V should be less affected by this due to its training on contexts of the full multi-token term even if it is low-frequent.

5 Conclusions

In this study, we pretrained/finetuned ELMo and W2V models with a small corpus of philosophical texts and compared them using intrinsic evaluation methods. We also explored combining the two kinds of embeddings. Our main conclusions are: 1) ELMo models provide better embeddings than the finetuned W2V model despite the small data size, except a pretrained model without tuning, which performs worse. 2) Concatenating and adding embeddings does not bring extra value in this study; however, when merging embeddings from different models, performance can be gained by tuning the contribution of each model.

References

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 *COLING 28*, pages 6690–6702, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jelke Bloem, Antske Fokkens, Aurélie Herbelot, and Computational Lexicology. 2019. Evaluating the consistency of word embeddings from small data. In *RANLP*.

Hicham El Boukkouri, Olivier Ferret, Thomas Lavergne, and Pierre Zweigenbaum. 2019. Embedding strategies for specialized domains: Application to clinical entity recognition. In *ACL*.

Aurélie Herbelot and Marco Baroni. 2017. High-risk learning: acquiring new word vectors from tiny data. In *EMNLP*.

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

Alexandre Kabbach, Kristina Gulordava, and Aurélie Herbelot. 2019. Towards incremental learning of word embeddings using context informativeness. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 162–168, Florence, Italy. Association for Computational Linguistics.

Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. 2016. Multi-agent cooperation and the emergence of (natural) language. *arXiv preprint arXiv:1612.07182*.

Alessandro Lenci. 2018. Distributional models of word meaning. *Annual Review of Linguistics*, 4(1):151–171.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS 2013.*, pages 3111–3119. Neural Information Processing Systems Foundation, Inc.

Farhad Nooralahzadeh, Lilja Øvrelid, and Jan Tore Lønning. 2018. Evaluation of domain-specific word embeddings using knowledge resources. In *LREC*.

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 *NAACL*, pages 2511–2522, Online. ACL.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *NAACL*, pages 2227–2237, New Orleans, Louisiana. ACL.

Radim Rehurek and Petr Sojka. 2011. Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, 3(2).

Howard Robinson. 2020. Substance. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Spring 2020 edition. Metaphysics Research Lab, Stanford University.

Arpita Roy, Youngja Park, and Shimei Pan. 2017. Learning domain-specific word embeddings from sparse cybersecurity texts. *ArXiv*, abs/1709.07470.

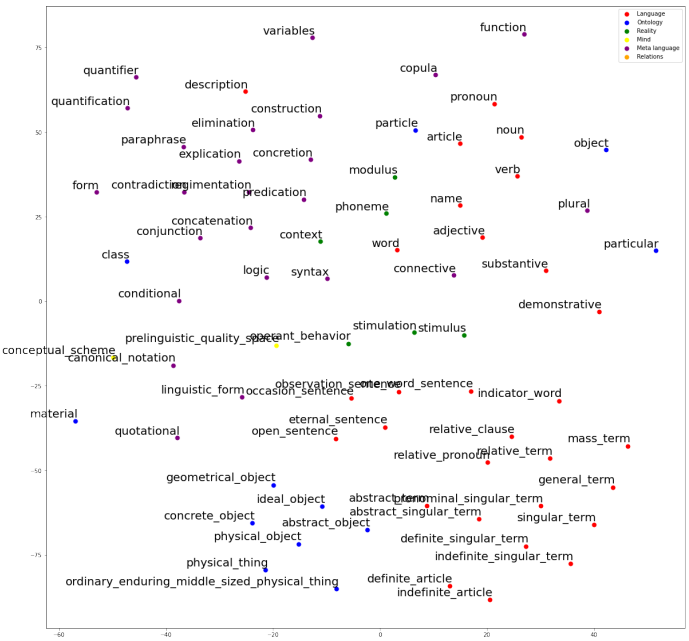
A ELMo hyper-parameters

	LSTM		Char cnn		Data	
Model	**Output dim**	**Hidden dim**	**N char**	**Embed dim**	**Type**	**Token**
ELMo s	256	1024	261	16	philosophy	1.6M
ELMo m	512	2048	261	16	philosophy	1.6M
Pretrained	256	1024	261	16	miscellaneous	800M
Finetuned	256	1024	261	16	combined	combined

B Seed words

inference	deductive	argument
analytical	antecedent	necessary
effect	cause	epistemology
extension	intension	extensional
formal	freedom	identity
argument	hypothetical	induction
categorical	infinity	intension
extension	justice	logical
moral	truth	ontology
perceptual	relativity	identity
premise	reason	theoretical
property	reasoning	extension
proposition	practical	relation
nature	analysis	disposition
subjective	analytic	critical
substance	appearance	experience
synthetic	belief	empirical
analytic	concept	formal
knowledge	practical	reason
logical	pure	standpoint
maxim	reality	subject
objective	rational	subjective
perspective	real	system
existence	perspective	spirit
fallacy	paradox	verification
meaning	science	symbol
analogy	paradox	intuition
inference	predicate	judgment
essential	sense	synthetic
extension	simplicity	theoretical
illusion	state	understanding
deductive	hypothetical	will
intensional	ideology	being
fact	imagination	use
mention	valid	

C Term distribution from ELMo small model



Term distribution after t-SNE dimension reduction for the ELMo small embeddings. Note that the Dunn index for the clusters after dimension reduction is 0.05 (down from 0.44), so there is a large information loss and this visualization does not fully represent the 256-dimensional model.

D Target terms

Language	Ontology	Reality	Mind	Metalinguistic
Pronominal singular term	Ordinary enduring middle sized physical thing	Operant behavior	Prelinguistic quality space	Canonical notation
Abstract term	Class	Modulus	Conceptual scheme	Paraphrase
Adjective	Concrete object	Stimulation		Variables
Article	Physical object	Phoneme		Concatenation
Definite article	Ideal object	Stimulus		Concretion
Indefinite article	Geometrical object	Context		Conditional
Mass term	Material			Conjunction
Demonstrative	Object			Connective
Description	Abstract object			Construction
General term	Particle			Contradiction
Singular term	Particular			Copula
Definite singular term	Physical thing			Form
Indefinite singular term				Function
Eternal sentence				Quantification
Indicator word				Quantifier
Name				Quotational
Noun				Predication
Relative term				Plural
Substantive				Regimentation
Observation sentence				Elimination
Occasion sentence				Explication
Open sentence				Linguistic form
Pronoun				Logic
Abstract singular term				Syntax
Relative clause				
Relative pronoun				
One word sentence				
Word				
Verb				