Eliciting explicit knowledge from domain experts in direct intrinsic evaluation of word embeddings for specialized domains

van Boven, G.; Bloem, J.

Publication date
2021

Document Version
Final published version

Published in
Human Evaluation of NLP Systems (HumEval)

Eliciting explicit knowledge from domain experts in direct intrinsic evaluation of word embeddings for specialized domains

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

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

Abstract

We evaluate the use of direct intrinsic word embedding evaluation tasks for specialized language. Our case study is philosophical text: human expert judgements on the relatedness of philosophical terms are elicited using a synonym detection task and a coherence task. Uniquely for our task, experts must rely on explicit knowledge and cannot use their linguistic intuition, which may differ from that of the philosopher. We find that inter-rater agreement rates are similar to those of more conventional semantic annotation tasks, suggesting that these tasks can be used to evaluate word embeddings of text types for which implicit knowledge may not suffice.

1 Introduction

Philosophical research often relies on the close reading of texts, which is a slow and precise process, allowing for the analysis of a few texts only. Supporting philosophical research with distributional semantic (DS) models (Bengio et al., 2003; Turney and Pantel, 2010; Erk, 2012; Mikolov et al., 2013) has been proposed as a way to speed up the process (van Wierst et al., 2016; Ginammi et al., in press; Herbelot et al., 2012), and could increase the number of analysed texts, decreasing reliance on a canon of popular texts (cf. addressing the great unread, Cohen, 1999). However, we cannot evaluate semantic models of philosophical text using a general English gold standard, as philosophical concepts often have a very specific meaning. For example, the term *abduction*, usually meaning a kidnapping, denotes a specific type of inference in philosophy (Douven, 2017). Therefore, models must be evaluated in a domain-specific way.

The critical difference between the general case and the philosophy case is the following. It is easy to find native speakers of e.g. British English who have a good intuition of the meaning of its terms in general use, and the relations between them. This yields e.g. the SimLex-999 word similarity dataset (Hill et al., 2015), covering frequent words and their typical senses. More difficult is finding 'native speakers' who have an intuition of the meaning of the terms used by a particular philosopher. The only candidate would be that philosopher themselves, and even then, the meaning of some of the terms used is the result of explicit analysis and definition rather than implicit language knowledge of the philosopher. Uncommon terms with highly specific meanings are explicitly defined and debated, leading them to differ between philosophers or even within the works of a single philosopher. Any accurate evaluation or annotation would require expert knowledge, and methods that can incorporate explicit knowledge, rather than judgements based on implicit knowledge of a standard language or jargon by one of its speakers.

We test two direct evaluation methods for DS models described by Schnabel et al. (2015) on our case study, the works of Willard V. O. Quine, a 20th century American philosopher. Instead of native English speaking crowdworkers, we selected expert participants who have studied this philosopher extensively. We aim to test whether these methods produce reliable results when participants need to use explicit rather than implicit knowledge, and consider the methods to be successful if inter-rater agreement matches that of other semantic evaluations. More broadly, our methodological findings apply to evaluation of DS models for specialized domains, language for specific purposes (LSP), historical language varieties or other language (varieties) for which no native annotators are available.

2 Related work

Most intrinsic evaluations compare word embedding similarities (e.g. in terms of cosine distance)

Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval), pages 107–113
Online, April 19, 2021.

to premade datasets of human similarity or relatedness judgements. Sets of words are created and evaluated on semantic relations by participants, and the similarity between the assessments and an embedding space is used as a measure of performance. In specific domains, examples of such datasets of term ratings can be found for identifier names in source code (Wainakh et al., 2019), in the medical domain (Pakhomov et al., 2010, 2011; Pedersen et al., 2007) and in geosciences both for English (Padarian and Fuentes, 2019) and Portuguese (Gomes et al., 2021). The last two studies compare domain-specific embeddings to general domain embeddings and both find that the former perform better. A problem of these indirect datasets is that only naturally occurring, often high-frequency terms without any spelling variations, are evaluated, while DS models include many more variations (Batchkarov et al., 2016).

Direct intrinsic evaluation methods, where participants respond directly to the output of models, can be categorized as *absolute* and *comparative* intrinsic evaluation (Schnabel et al., 2015). The former method evaluates embeddings individually and compares their final scores, while in the latter participants directly express their preferences between models. To our best knowledge, the only example of a domain-specific direct human evaluation is Dynomant et al. (2019) who evaluate French embeddings of health care terms by a human evaluation in which two medical doctors rate the relevance of the first five nearest neighbours of target terms from models trained on in-domain text.

In the philosophical domain some evaluations have been conducted with other methods, sometimes incorporating expert explicit knowledge, but none are direct. In each of these studies the work of Quine is utilized as data. Firstly, Bloem et al. (2019) propose a method of evaluating word embedding quality by measuring model consistency, not making use of expert knowledge. Secondly, Oortwijn et al. (2021a) construct a conceptual network which serves as a *ground truth* of expert knowledge. They compare the similarity of embeddings for target philosophical terms to their position in the manually created network. Here, the conceptual relatedness between terms is restricted to the property of sharing hypernyms, and only terms that were predefined in the ground truth can be considered for evaluation. Betti et al. (2020) introduce a more elaborate ground truth that is concept-focused, including more types of conceptual relations and including irrelevant as well as relevant terms for better evaluation of model precision. Still, evaluation remains restricted to terms in the ground truth. Only using direct evaluation methods we can attempt to evaluate all model output.

3 Task description

We perform a synonym detection task and a coherence task. In these tasks, participants are asked to judge model-generated candidate terms that semantic models deem closest to a target term. In the synonym detection task, participants select the most similar word to target term t out of a set of options: the k-nearest neighbours of the target term in each model that is being compared. In the coherence task, the participant selects a semantical outlier in a set of words, where one of the words is not close to t in the model. We refer to Schnabel et al. (2015) for details and a comparison to other tasks for general semantic evaluation. Our participant instructions are based on Schnabel et al., who use the instructions of the WordSim-353 dataset (Finkelstein et al., 2001). But as this study focuses on explicit knowledge, several adjustments are needed.

Although explicit knowledge is easier to verbalize than implicit knowledge, it involves controlled rather than automatic processing (Bowles, 2011; Ellis, 2004, 2005), so our version of the task might take longer. Yet in order to retain the required focus, the test should not take too long. We therefore conduct a pilot study in which response times are measured to estimate task durations, and we adapt the size of the main study accordingly.

The original task instructions do not define similarity, while other studies define it as co-hyponymy (Turney and Pantel, 2010) or synonymy (Hill et al., 2015). According to Batchkarov et al. (2016) defining similarity is difficult as it depends on the context and downstream application in which the terms are used. We keep a consistent context, both training and evaluating in the domain of a particular philosopher, although the concern of capturing the multidimensional concept of *similarity* in a single number is valid also in this context. Gladkova and Drozd (2016) claim participants are likely to prefer synonyms when asked to select the most similar word. In this study we are looking to find any relationship present, rather than a specific one, and expect the experts to explicitly consider this, so we ask for *relatedness*. Gladkova and Drozd further

argue that when asked for relatedness participants must choose between various relations present, a choice that can be subjective or random, and might reflect other factors such as "frequency, speed of association, and possibly the order of presentation of words". The first two factors are alleviated in this study as the participants must take Quine's definitions of words into account rather than their own. This forces participants to think their answers through, which should reduce the association effects typical of fast-paced online studies. To account for effects of order of presentation, we randomize the order of the options. In our instructions, we define relatedness as synonymy, meronymy, hyponymy, and co-hyponymy, and provide examples. Participants are also allowed to base their judgements on other types of relations.

After the experiment we present a post-test survey, querying the types of relation participants based their judgements on, and task difficulty. Furthermore, we change the option *I don't know the meaning of one (or several) of the words* in the synonym detection task to be *None of these words is even remotely related*, and also include a similar option in the coherence task, namely *No coherent group can be formed from these words*. This is done to avoid any random selection of words when there are no meaningful relations, making the responses more accurate. As we aim to gather explicit knowledge, participants are allowed to look up relevant information on presented words. For reproducibility, our instructions (and results) are included in the supplementary materials. [1]

As the tasks require participants to be experts on the work of Quine, the number of possible participants is limited. Although participants are philosophers trained to work precisely and make consistent judgements, subjectivity can be a risk as participant must choose the relation they deem most important, while lacking context. We use inter-rater agreement to evaluate this. We report joint probability of agreement (percentage of agreement) as we have added the *none* options to avoid chance agreement. As joint probabilities cannot be compared across studies, we also report Cohen's κ.

All experiments[2] are conducted on the survey platform *Qualtrics*[3]. Participants are asked to execute the experiments in a silent environment.

4 Case study for philosophy

We make use of the QUINE corpus (v0.5, Betti et al., 2020), which includes 228 philosophical books, articles and bundles Quine authored, consisting of 2.15 million tokens in total. As target terms for evaluation, we use Bloem et al.'s (2019) test set for the the book *Word & Object* (Quine, 1960), one of Quine's most influential works. It consists of 55 terms that were selected from the book's index. We used 10 of these terms in the pilot study, 25 in the synonym detection task, 14 in the coherence task and 6 in both experiments.[4]

One Quine expert participated in the pilot study. The pilot study consists of short versions of the two tasks, both testing five target terms. In the synonym detection task, each target term has six candidate related terms from the models, that the participant should choose between. Each term is tested three times with candidates of differing similarity from the model (nearest neighbour ranks $k \in \{1, 5, 50\}$). The pilot coherence (outlier) task has ten questions. The average response time for the synonym detection task was 109.5s and 42.1s for the coherence task. Because for the first task this was higher than anticipated, we reduced the number of ranks to two and divided the task across two separate surveys.

4.1 Experiment 1: Synonym detection task

Three experts on the work of Quine, including the participant of the pilot study, participated in this experiment. They all hold a Master's degree in philosophy and have studied the philosopher extensively.

This task includes 31 target words, which are all tested on two ranks k, with $k \in \{1, 10\}$, resulting in 62 questions. Of the 45 test set terms not used in the pilot study, we took the fifteen highest frequency terms ($n > 275$) and the sixteen lowest ($n < 84$). The experiment was conducted through two surveys, each consisting of 31 questions, lasting around 50 minutes, with a break halfway. [5]

The data from one of the participants was excluded, as the participant indicated that the test was too difficult and that their expertise on the work of Quine did not suffice. Moreover the response times of this participant were a lot lower than for the other participants. For this experiment, the overall

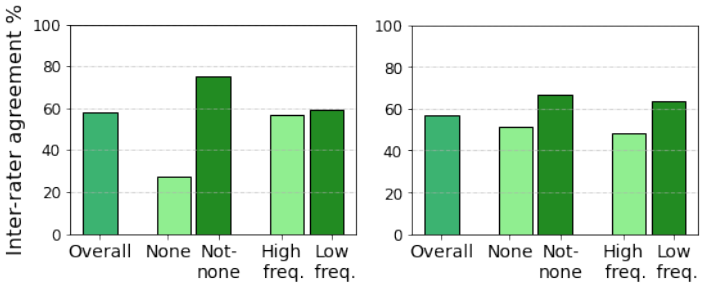
[1] To be found at https://github.com/gvanboven/direct-intrinsic-evaluation

[2] Experiments were approved by The Ethics Committee of the Faculty of Humanities of the University of Amsterdam.

[3] www.qualtrics.com

[4] Listed in the supplemental materials, with frequencies

[5] Example surveys and raw results for each participant are included in the supplementary materials.



(a) (b)

Figure 1: Inter-rater agreement in different conditions of (a) the synonym detection task and (b) the coherence task

	Response time	
	Exp. 1	**Exp. 2**
Overall	45.5 s	25.8 s
None	53.4 s	28.2 s
Not-none	43.4 s	23.4 s
High frequency	35.7 s	26.9 s
Low frequency	54.9 s	24.9 s

Table 1: Response times in different conditions of 1. synonym and 2. coherence tasks

inter-rater agreement was 58.1%, with $\kappa = 0.492$.

4.2 Experiment 2: Coherence task

Two of the participants from the previous experiments also participated in this study. 20 target words are used: the 14 test set terms not used in the pilot or Exp. 1, and the 3 highest and lowest frequency terms from Exp. 1. We divide these into eleven low frequency words ($n < 142$) and nine high frequency words ($n > 187$). Using 3 DS models this results in 60 questions, the test takes approximately 40 minutes with a break. The inter-rater agreement was 56.7%, with $\kappa = 0.345$.

5 Analysis

To assess whether the method was successful we discuss some reliability metrics and examine disagreement examples. First of all, the fact that the data from one participant had to be excluded confirms the high standard of expertise required for participating in our version of the tasks. The results might have differed had there been more or different participants. However, other studies on expert explicit knowledge also execute tasks with two (Dynomant et al., 2019) or three (Padarian and Fuentes, 2019; Gomes et al., 2021) participants.

Inter-rater agreement scores for the two tasks were 58.1% ($\kappa = 0.492$) and 56.7% ($\kappa = 0.345$), indicating moderate or fair agreement. Batchkarov et al. (2016) found the average inter-rater agreement of two raters of the WordSim-353 (Finkelstein et al., 2001) and MEN (Bruni et al., 2014) dataset to lie between $\kappa = 0.21$ and $\kappa = 0.62$. Thus, agreement scores in this study are not lower than that of commonly used similarity datasets, despite participants having to agree on another person's semantics and including a *None* option.

Both experiments yield lower inter-rater agreement for the *None* option than for the other choices, shown in Figure 1(a) and (b). Response times were also higher for the *None* option in both tasks (Table 1), suggesting this choice is more difficult. Most disagreement thus concerned the presence of a semantic relationship, but if the annotators agreed there was one, they mostly preferred the same relation. This suggests a *None* option increases annotation quality in general. In the coherence task, there was more agreement on low than high frequency words, which may be due to their lesser ambiguity.

According to the post-test survey, participants mostly based their judgements on sharing the same super term. Relationships that were used without being listed in the instructions were antonymy, forming a technical bigram term together, having the same stem and being used in the same context by Quine. We see this reflected when examining some examples of disagreement. In Table 2, we see disagreement on the related term for *adjectives* because both chosen terms have a relation to this target term, but these are two different relations. We see agreement for *information*, as *collateral information* is a meaningful bigram in Quine's thought experiment on radical translation. In Table 3 we see disagreement on the *ambiguity* outlier. While *believe* has a tenuous relation to *ambiguity*, participant 2 may have considered this relation too tenuous and went for *none*. One expert stated that unclear boundaries of the *none* option were the reason for many *none* disagreements. The *sense datum* disagreement was guessed to be over a rare non-mathematics sense of *divisibility* that one participant remembered but the other might not have.

6 Discussion

In the post-test survey, participants commented that it was sometimes difficult to select the most related

adjectives	*information*	*application*
translation	learning	**numbers**$_1$
embodying	reduction	ambiguity
modifiers$_2$	**collateral**$_{12}$	multiplicity
specious_ present	application ordered_pair	subtraction belong
verbs$_1$	*None*	abbreviative
None		***None***$_2$

Table 2: Example of disagreement and agreement in the synonym detection task. To be read vertically, with target terms in italics. Bolded/marked model terms were chosen by participants to be related to the target term.

ambiguity	*objects*	*sense_datum*
parts	object	**prediction**$_1$
phoneme	physical	construction
believe$_1$	**them**$_{12}$	**divisibility**$_2$
None$_2$	*None*	*None*

Table 3: Example of disagreement and agreement in the coherence task. Bolded/marked terms were chosen by participants to be outliers, underlined terms were model outliers with lower word embedding similarity.

word, as different relations were present and selecting the most important one is partly a matter of preference. Such ambiguity is prevalent in any semantic annotation task in which context is unspecified, and in other language annotation tasks in which no explicit choice is made in the guidelines among possible competing valid interpretations (Plank et al., 2014). As noted by Sommerauer et al. (2020), justified disagreement is possible, though detecting it requires meta-annotation and this is in itself a difficult task. However, it might yield additional insights, i.e. that certain DS models might prioritize certain relation types in their nearest neighbours, and that these are equally valid because the experts disagreed on them. Disagreement can also be caused by poorly specified tasks and insufficient conceptual alignment among annotators, especially when the goal is creating a ground truth (Oortwijn et al., 2021b) or otherwise annotating for a specific theory or interpretation.

In future experiments, more specific instructions on when to consider a relation to be relevant, or guidelines on prioritizing certain relations over others, can reduce the difficulty of the task. Our expert participants used many semantic relation types in their interpretation with no clear hierarchy among them. However, applying this to DS model evaluation may require more insight into what exactly the geometric relationships between embeddings in a DS model capture. It may also be interesting for philosophers to make use of models trained to represent particular relations, such as antonymy (Dou et al., 2018). With more specific instructions explicitly directing participants to prioritize or ignore specific relations, our evaluation approach can be adapted to evaluate such models and we expect higher agreement in this type of task. In other cases different interpretations can be desirable, e.g. where there is no hierarchy of relations and a model should capture *relatedness* in a broad sense. For this purpose, we should consider allowing multiple answers — while a forced choice helps to elicit implicit knowledge, explicit knowledge may not always support a categorical decision, though this adds the complication of deciding when an option is relevant enough, similar to the *none* option.

Our results show that absolute and comparative intrinsic evaluation tasks can be used to agree on semantic relatedness between word embeddings even when the target language variety is highly specific. By instructing domain experts to perform the evaluation task using explicit expert knowledge rather than implicit knowledge, inter-rater agreement rates similar to other semantic annotation tasks can be reached. Due to the inherent lack of context in evaluating type-based non-contextual word embeddings, participants struggled with the generality of the task. Based on our analysis and post-test survey, we expect more specific guidelines on word relatedness to increase the reliability of the annotators' judgements, while limiting their generalizability. The addition of a *None* option seemed particularly beneficial for obtaining more reliable annotations based on explicit knowledge. We expect these findings to apply in the context of other domains for which no 'native' annotators are available — for example, language for specific purposes (LSP), historical language varieties or idiolects. In future work, the absolute and comparative intrinsic evaluation tasks we have described can be used to compare the quality of the representations of different word embedding models on these specialized language varieties.

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).

References

Miroslav Batchkarov, Thomas Kober, Jeremy Reffin, Julie Weeds, and David Weir. 2016. A critique of word similarity as a method for evaluating distributional semantic models. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 7–12.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.

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.

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.

Melissa A Bowles. 2011. Measuring implicit and explicit linguistic knowledge: What can heritage language learners contribute? *Studies in second language acquisition*, pages 247–271.

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49:1–47.

Margaret Cohen. 1999. *The sentimental education of the novel*. Princeton University Press.

Zehao Dou, Wei Wei, and Xiaojun Wan. 2018. Improving word embeddings for antonym detection using thesauri and SentiWordNet. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 67–79. Springer.

Igor Douven. 2017. Abduction. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, summer 2017 edition. Metaphysics Research Lab, Stanford University.

Emeric Dynomant, Romain Lelong, Badisse Dahamna, Clément Massonnaud, Gaétan Kerdelhué, Julien Grosjean, Stéphane Canu, and Stefan J Darmoni. 2019. Word embedding for the French natural language in health care: comparative study. *JMIR medical informatics*, 7(3):e12310.

Rod Ellis. 2004. The definition and measurement of L2 explicit knowledge. *Language learning*, 54(2):227–275.

Rod Ellis. 2005. Measuring implicit and explicit knowledge of a second language: A psychometric study. *Studies in second language acquisition*, 27(2):141–172.

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

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web*, pages 406–414.

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

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.

Diogo da Silva Magalhães Gomes, Fábio Corrêa Cordeiro, Bernardo Scapini Consoli, Nikolas Lacerda Santos, Viviane Pereira Moreira, Renata Vieira, Silvia Moraes, and Alexandre Gonçalves Evsukoff. 2021. Portuguese word embeddings for the oil and gas industry: Development and evaluation. *Computers in Industry*, 124:103347.

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.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

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

Yvette Oortwijn, Jelke Bloem, Pia Sommerauer, Francois Meyer, Wei Zhou, and Antske Fokkens. 2021a. 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 (NAACL-HLT)*. In press.

Yvette Oortwijn, Thijs Ossenkoppele, and Arianna Betti. 2021b. Interrater disagreement resolution: A systematic procedure to reach consensus in annotation tasks. In *Proceedings of the First Workshop on Human Evaluation of NLP Systems (HumEval)*.

José Padarian and Ignacio Fuentes. 2019. Word embeddings for application in geosciences: development, evaluation, and examples of soil-related concepts. *Soil*, 5(2):177–187.

Serguei Pakhomov, Bridget McInnes, Terrence Adam, Ying Liu, Ted Pedersen, and Genevieve B Melton. 2010. Semantic similarity and relatedness between clinical terms: an experimental study. In *AMIA annual symposium proceedings*, page 572. American Medical Informatics Association.

Serguei VS Pakhomov, Ted Pedersen, Bridget McInnes, Genevieve B Melton, Alexander Ruggieri, and Christopher G Chute. 2011. Towards a framework for developing semantic relatedness reference standards. *Journal of biomedical informatics*, 44(2):251–265.

Ted Pedersen, Serguei VS Pakhomov, Siddharth Patwardhan, and Christopher G Chute. 2007. Measures of semantic similarity and relatedness in the biomedical domain. *Journal of biomedical informatics*, 40(3):288–299.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Linguistically debatable or just plain wrong? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 507–511.

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

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.

Pia Sommerauer, Antske Fokkens, and Piek Vossen. 2020. Would you describe a leopard as yellow? Evaluating crowd-annotations with justified and informative disagreement. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4798–4809.

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

Yaza Wainakh, Moiz Rauf, and Michael Pradel. 2019. Evaluating semantic representations of source code. *CoRR*, abs/1910.05177.

Pauline van Wierst, Sanne Vrijenhoek, Stefan Schlobach, and Arianna Betti. 2016. Phil@Scale: Computational Methods within Philosophy. In *Proceedings of the Third Conference on Digital Humanities in Luxembourg with a Special Focus on Reading Historical Sources in the Digital Age. CEUR Workshop Proceedings, CEUR-WS.org*, volume 1681, Aachen.