Expert Concept-Modeling Ground Truth Construction for Word Embeddings Evaluation in Concept-Focused Domains

Arianna Betti[♠] **Martin Reynaert**[♠♢] **Thijs Ossenkoppele**[♠] **Yvette Oortwijn**[♠♣] **Andrew Salway**[♠♡] **Jelke Bloem**[♠]

♠ Institute for Logic, Language and Computation. University of Amsterdam
♢ Tilburg School of Humanities and Digital Sciences. Tilburg University
♣ Algorithms, Geometry & Applications. Eindhoven University of Technology
♡ School of Media, Arts and Humanities. University of Sussex
{ariannabetti, thijs.ossenkoppele, yvette.oortwijn}@gmail.com,
{m.w.c.reynaert, a.j.salway, j.bloem}@uva.nl

Abstract

We present a novel, domain expert-controlled, replicable procedure for the construction of concept-modeling ground truths with the aim of evaluating the application of word embeddings. In particular, our method is designed to evaluate the application of word and paragraph embeddings in concept-focused textual domains, where a generic ontology does not provide enough information. We illustrate the procedure, and validate it by describing the construction of an expert ground truth, QuiNE-GT. QuiNE-GT is built to answer research questions concerning the concept of *naturalized epistemology* in QUINE, a 2-million-token, single-author, 20th-century English philosophy corpus of outstanding quality, cleaned up and enriched for the purpose. To the best of our ken, expert concept-modeling ground truths are extremely rare in current literature, nor has the theoretical methodology behind their construction ever been explicitly conceptualised and properly systematised. Expert-controlled concept-modeling ground truths are however essential to allow proper evaluation of word embeddings techniques, and increase their trustworthiness in specialised domains in which the detection of concepts through their expression in texts is important. We highlight challenges, requirements, and prospects for future work.

1 Introduction

A distinctive aspect of language is its capacity to express ideas and encode concepts, enabling us to treat certain natural language features as proxies for concepts and conceptual structures. Mining concepts in text corpora is a valuable enterprise in many domains, but it is no trivial feat, as the link between concepts and the words expressing them cannot be (fully) automatised.

In generic domains, concept mining can sometimes rely on resources that fix word-concept links, such as WordNet (Fellbaum, 1998) and thesauri (Kipfer, 2019). In expert domains, however, harder-to-come-by, specialised resources are needed. The issue is exacerbated in domains focusing on concepts and small corpora with high lexical variation and fine-grained meaning distinctions, like the history of the sciences and the humanities, and more broadly the history of ideas.

Importantly, the word-concept link problem is not limited to the humanities, or fields studying shifting meanings or changing concepts. It fundamentally applies to any corpora containing specialised terminology requiring substantial background knowledge for its decoding. Retrieval of relevant information in text depends on correctly identifying which terms express the concepts of interest. Any researcher or investigator struggling to sift through appreciably vast, lexically varied text collections to obtain specific, highly relevant information faces the problem, as do librarians and publishers wishing to give recommendations based on the ideas expressed in full-text documents, rather than based on metadata or user profile (Lops et al., 2011; Aggarwal, 2016). The word-concept link problem is just more *obvious* for corpora that are too big to be processed manually by otherwise very skilled humans, and *harder* the more specialised the information is.

Proceedings of the 28th International Conference on Computational Linguistics, pages 6690–6702
Barcelona, Spain (Online), December 8-13, 2020

A widely used family of techniques in concept mining are *word embeddings*. Word embeddings are mathematical representations of words exploited in distributional semantics models (DSMs) (Turney and Pantel, 2010; Erk, 2012; Clark, 2015). DSMs are statistical language models based on the so-called Distributional Hypothesis (Harris, 1954; Sahlgren, 2008), affirming that words with similar meanings tend to occur in similar contexts. DSMs can help to cluster terms into (functional) synsets, and can be integrated in downstream applications such as information retrieval systems (Clinchant and Perronnin, 2013; Zamani and Croft, 2017; Wang and Koopman, 2017), which rank units of texts, such as paragraphs and sentences, by relevance.

In certain specialised textual domains in which fine-grained study of concepts is traditionally performed exclusively by humans, the use of word embeddings on corpora of a few million tokens – too big a size for manual processing in a reasonable time – is, potentially, a game-changer (Betti et al., 2019). Technology needs validation before it can be game-changing, however (Hellrich, 2019; Sommerauer and Fokkens, 2019): word embedding techniques cannot gain a foothold in traditional domains such as history or philosophy if they cannot reasonably approximate or reliably support expert work, or if they introduce unknown biases (Fokkens et al., 2014). Unfortunately, the performance of DSMs is notoriously difficult to evaluate; it is even controversial how one should carry out evaluations in generic, big corpora domains (Gladkova and Drozd, 2016; Bakarov, 2018). Data sparsity makes for an even bigger difficulty: a few million tokens is very little data to create high quality DSMs, due to their statistical nature. What would be a proper way to evaluate the quality of distributional models for concept mining in a specialised, concept-focused domain employing small corpora? Answering this question would mean important progress for the prospect of concept mining in these domains, as we would know which models are best to choose for reliable downstream applications.

We propose a novel method to construct replicable, expert-controlled, concept-modeling ground truths (henceforth: 'ground truths$_{xc}$') for the evaluation of word and paragraph vectors built from specialised corpora in concept-focused domains. Ground truths$_{xc}$ should include or specify (a) a corpus; (b) a conceptual model and multiple research questions, distilled from the model by experts, that can be answered on the basis of the corpus; (c) lists of terms (preferably hierarchically ordered) that are manually selected and linked by experts to the concepts appearing in the model; (d) segmented units of text, ideally paragraphs, in which terms linked to the concepts appear and that are judged by experts relevant to the research questions, preferably with a varying degree of relevance. We also release a tool, HitPaRank, which speeds up the manual process of linking paragraphs (d) to concepts (b) via term lists (c). Ground truths$_{xc}$ can be used to evaluate the performance of DSMs trained on their source corpus as well as other corpora, as long as those corpora have characteristics similar to to the source corpus. Particularly valuable will be ground truths$_{xc}$ that are extensive and varied enough for studying commonly used DSMs.

The method is generally applicable, language-independent, and extensible to multilingual data. It has its theoretical basis in a specific methodology in the history of ideas known as the 'model approach' (Betti and van den Berg, 2014), and is an operationalisation of that approach. According to the model approach, concepts cannot be properly studied unless they are represented as relational networks of terms, where the latter are the result of making textual interpretation explicit.[1] Increase of replicability, objectivity and the explicitation of biases in humanities research are advantages of the method. Although the model approach has been developed to the aim of studying concept change, it is in fact applicable to any concept, irrespective of their variability.

To validate our proposal, we constructed a ground truth$_{xc}$, QuiNE-GT, concerning the concept of *naturalized epistemology* as represented in QUINE, a large corpus (c. 2 million tokens) in English by an eminent 20th century philosopher. In keeping with (a)-(d) above, QuiNE-GT consists of clusters of terms and paragraphs annotated by experts as relevant, mildly relevant or non-relevant to answer three related research questions on ideas contained in the corpus.

[1]Views of concepts differ vastly among and across philosophy, cognitive psychology and linguistics cf. e.g. Margolis and Laurence (2019); Machery (2010). The idea that we talk of concepts can be substituted by talks of terms in use forming clusters of similarity can be found in Quine (1960); see also Decock and Douven (2011).

2 Related work

Extensive work is done to apply natural language processing and machine learning techniques to problems of phrase extraction, e.g. Shang et al. (2018), and concept mining, e.g. Li et al. (2018). However, whilst such fully-automated approaches are impressive in their generality and scale of application, concept-focused domains relying on fine-grained analysis of concepts require ground truths that represent 100% accurate expert knowledge. This means that there will always be humans in the loop, as the interest in specialised fields goes to computational techniques and tools that can assist researchers in mining concepts for specialised corpora.

Ground truths$_{xc}$ are extremely rare. The theoretical methodology behind their actual and concrete construction has, to the best of our ken, never been explicitly conceptualised and properly systematised. It is on the contrary rather customary to state that the notion of ground truth is inapplicable to the humanities and the social sciences (Nguyen et al. (2020)), which are the domains in which theoretical reflection on textual representation of ideas and interpretation mostly tends to take place. Most work on historical concept change does not involve ground truths$_{xc}$ in evaluation (cf. e.g. Kenter et al. (2015), Tahmasebi and Risse (2017)).

Humanities and Social Sciences Betti and van den Berg (2016) and van den Berg et al. (2018) propose to base ground truths for concept-focused domains such as the history of ideas and intellectual history on Betti and van den Berg (2014)'s 'model approach'. These works focus however on turning conceptual models into computer science ontologies to detect concept change (as further developed in Masolo et al. (2019)), and lack details as how to actually build ground truths$_{xc}$ for use in language technology. Mencarini (2018) signals the need for the model approach's operationalisation for language technology in the social sciences.

Ginammi et al. (2020) combine the model approach with the use of DS modeling on a German 19th-century philosophy corpus. Though they themselves provide no evaluation based on ground truths$_{xc}$, and explicitly cherry-pick results of the system's output, Ginammi et al. (2020) do signal the need for proper evaluation of DSMs relying on ground truths.

Building upon Betti et al. (2019)'s mixed method, which couples the model approach with manual annotations, Ossenkoppele (2019) offers an operationalisation of the model approach. They set up a model articulating the concepts of *science*, *pseudo-science* and *epistemology* as found in a lower quality, unlemmatised and unsegmented subset of our QUINE corpus, and associate their conceptual model with explicitly documented lists of terms. The terms are then queried via simple string searches afforded by Voyant Tools.[2] By these means, a number of textual fragments are retrieved and retained if relevant to certain research questions, and then annotated. Our main theoretical innovations with respect to Ossenkoppele (2019) are two. First, the conceptualisation of a *ground truth*$_{xc}$ as the (a)-(d) package mentioned above. Second, an increase in automatisation, generalisation and replicability of the method for ground truth$_{xc}$ construction. As for practical improvements: we improve the corpus' quality and usability, and, in releasing QuiNE-GT and HitPaRank, we enlarge and refine the conceptual model, reformulate the research questions and the annotation model, expand the term lists, provide new and much more extensive lists of paragraphs, and perform more fine-grained annotations by multiple experts.

Natural Language Processing Bloem et al. (2019) apply Nonce2Vec (Herbelot and Baroni, 2017) – a modification of Word2Vec for tiny data – to provide consistent sentence embeddings for a very small corpus of two books by Quine. They select terms from the books' subject index for comparing vectors, and document the difficulties of evaluation within that domain. In expanding the Nonce2Vec line of work to a corpus of Neo-Latin philosophical text, Bloem et al. (2020) discuss DSMs evaluation based on ground truths$_{xc}$ as a plan for the future.

In stressing the danger posed by using little understood DSMs methodologies to detect concept change, Sommerauer and Fokkens (2019) fix an initial articulation of the concept of *racism* and a number of target terms as proxies for the subconcepts to be detected in the COHA and the

[2]https://voyant-tools.org/

English Google n-gram corpus; they also offer guidelines for sound methodology. Our work can be seen as thoroughly expanding on the first guideline, according to which a "wide range of verifiable hypotheses" need to be offered "to study the overall question before diving into actual changes". Our work can also be seen as extending the concerns about replicability raised in Fokkens et al. (2013) to ground truth$_{xc}$ construction.

Meyer et al. (2020) have domain experts construct a network of conceptually related terms as a ground truth for a lower quality unlemmatised subset of our Quine corpus. The ground truth consists of 74 terms clustered into 5 broad categories (*language*, *ontology*, *reality*, *metalinguistic*, *mind*), reflecting expert interpretive knowledge of Quine (1960). They propose to evaluate DSMs by comparing the mathematical closeness of the vectors for the terms created in the language modelling against the similarity/relatedness of the terms themselves as fixed in the ground truth. In the ground truth they construct, the conceptual relatedness/similarity between terms is restricted to the property of having the same hypernym; the terms are not ranked for relevance, and are not meant to capture specific concepts. No ground truth in terms of paragraphs is offered.

Domain expertise can also be incorporated into a ground truth for evaluation by means of domain-specific glossaries containing synonym, antonym and alternative form information, as demonstrated by Nooralahzadeh et al. (2018). We recognise the value in starting with domain-specific glossaries. Exactly because the latter are usually not available, we need a systematic method to identify word-concept links. This is a core challenge addressed in the current paper.

3 The QUINE Corpus and HitPaRank

We propose a general method applicable in the construction of any expert ground truth for any concept in concept-focused domains ('ground truths$_{xc}$'). We take a ground truth$_{xc}$ to have (or have to specify) the following elements: (a) a corpus; (b) a conceptual model for concepts in that corpus from which research questions and annotation model are distilled; (c) term lists associated with (b); (d) selected text units annotated with degrees of relevance to research questions/concepts. Ground truths$_{xc}$ are thus specific to the concept(s) modelled to answer specific questions, and intrinsically linked to specific corpora. Although the corpora can be expanded, this can only happen under certain strictly controlled conditions. In the next section, we illustrate our general method by referring to an example by way of validation, a ground truth$_{xc}$ (QuiNE-GT) that we built for a specific concept in a corpus (QUINE) comprising the virtually complete *oeuvre* in English of Willard V. O. Quine, a 20th century American philosopher.

We start by describing QUINE, the corpus on which QuiNE-GT is based, followed by the description of HitPaRank, a script developed by Reynaert that we found to be helpful in the execution of the general method. In the following section we expound the general method using QuiNE-GT as illustration. Note that QuiNE-GT is only one of the many possible ground truths$_{xc}$ that can be based on the same corpus constructed following the same general method.

The QUINE corpus (version 0.5). The QUINE corpus consists of virtually all of Quine's 228 books and articles, containing in total 819 documents (books are split into parts), 2,150,356 word tokens, 38,791 word types and 27,837 lemmatized word types. It includes texts in various genres and from different phases of Quine's thought on various topics, including technical, and formula-heavy writings on logic and the foundations of mathematics. The corpus exhibits a high degree of lexical variation and many instances of fine-grained meaning distinctions.[3]

The files originate from printed works, so substantial effort was required to produce a corpus of sufficiently high quality, as is often the case in computational humanities endeavors. In particular, it is key that the corpus accurately transcribes the text of the printed page, e.g. so that terms can be found by simple string matching. Also, paragraphs must be properly segmented and distinguished from headings and footnotes, and given unique identifiers for easy reference

[3]The QUINE corpus is derived from copyrighted works and is put together and processed for internal research use and for text-mining purposes only. Whilst the corpus cannot be shared at this point, we provide replication instructions here: `https://github.com/YOortwijn/QuiNE-ground-truth`

by researchers and tools alike. The use of paragraphs for ground truth$_{xc}$ construction is not mandatory, but paragraphs are especially interesting segments in philosophical works as they tend to correspond to minimal units of thought.[4] What is mandatory is the use of units that can be segmented in a replicable and automatic way (such as sections, paragraphs, or sentences). Information about word lemmas is also vital for our approach to concept mining both for query expansion (from lemmas to forms) and for mitigating sparsity when developing DSMs of small corpora. Here we briefly describe the preparation of the corpus through OCR and OCR correction, paragraph segmentation and the marking up of paragraphs and lemma information.

The books and articles were converted from images in PDF files to plaintext format using ABBYY FineReader and OCR errors were corrected manually. Books were split into separate files per chapter or section. Data irrelevant to the content of the chapters, sections, and articles was removed, including e.g. repeated headers and page numbers, and the metadata was stored separately. Some of the texts are rich in formulae and symbols which were replaced with codes as place holders, so as not to interfere unduly with linguistic analysis.

We carried out extensive semi-automated work to restore the paragraphs in all the texts which involved dealing with issues including paragraphs split by page boundaries, footnotes and headings. Some kinds of errors could be corrected via normalization scripts geared towards batches of texts that displayed similar issues, e.g. to deal with ad-hoc newlines that were erroneously present. This was complemented with manual editing and inspection of the original book images. At the same time issues relating to dashes and various forms of quotation marks were also resolved. We included metadata information (year, segment number for split books, etc) in the filenames of the texts in the corpus, for easy sorting and viewing in chronological order.

The texts were converted to FoLiA XML (van Gompel and Reynaert, 2013) using the English language module of UCTO[5], which also provided sentence segmentation and tokenization. FoLiA markup can represent text as, and maintain unique identifiers for, sections, paragraphs, sentences and words. FoLiA can also represent the lemma of each word. A dedicated FoLiA XML module[6] called on Spacy[7] using its core model for English[8] to provide lemmatization.

HitPaRank - HitPaRank[9] is a Perl script we developed that extracts the paragraphs possibly relevant to the research question from the corpus. The script has three modes of working: (i) Extraction of word ngrams (currently n $\leq$ 4) from the FoLiA XML files (in this study, from the lemmatized text layer).[10] (ii) Checking the wildcard versions of the query terms selected by researchers against the actual ngrams extracted from the corpus by (i), and returning an expanded list of query terms and their actual corpus frequencies. This mode also shrinks the wildcard version of the lists by not returning non-occurring terms. (iii) Extraction of the paragraphs that give hits on any of the query term lists and gathering statistics on all the extracted paragraphs, on the individual paragraph level and over the entire corpus.

4 A validated method for ground truth$_{xc}$ construction

Our general method for ground truth$_{xc}$ construction comprises six steps.

Step 1. Set an initial rough question and create a conceptual model An initial rough research question is set, from which concepts of interest are distilled and modelled as a network of subconcepts following Betti and van den Berg (2014). At this step, (sub)concepts usually still lack systematised association with terms in specific corpora.

[4] See note 1 in the Appendix to Ginammi et al. (2020), to be found at `https://conceptsinmotion.org/projects/ginammi-et-al-2020-appendix/`

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

[6] `https://github.com/proycon/spacy2folia`

[7] `https://spacy.io/`

[8] Spacy English core model: en_core_web_sm

[9] https://github.com/martinreynaert/HitPaRank

[10] For ngram frequency list building from larger FoLiA XML corpora the tool FoLiA-stats is recommended. It is part of the FoLiAutils toolkit. Installation of all FoLiA toolkits is greatly facilitated by the meta-installer LaMachine: `https://proycon.github.io/LaMachine/`

Step 2. Make research question(s) term-based and create initial lists of terms In order to be used for constructing ground truths$_{xc}$ for DSMs evaluation, any concept-based research questions must be turned into research questions that can be answered by successfully querying for terms or paragraphs. This is in contrast with Step 1 and typical research questions within philosophy, which are usually concept-based (van Wierst et al., 2016; Ginammi et al., 2020). To turn concept-based questions into term-based ones, all (sub)concepts at Step 1 need to be manually associated by experts to clusters of terms in a certain corpus. Through this association, a conceptual model can become an annotation model for textual corpora. At step 2 initial term lists are created, each relating to different aspects of the research questions. For example, if the question is about how a philosopher describes the relation between two concepts, then there will be at least one term list relating to each concept. The lists should ultimately consist of all terms that would indicate that the concept is present in a paragraph, e.g, for a question about plants, the term list would include different words for plants and types of plants. The initial term lists contain both full terms and wildcarded terms to capture multiple terms that share a stem, either by expanding the term forwards, backwards or both. Each term is given a ranking, which usually indicates its weight in determining relevant paragraphs (say 1 to 3, 1 being the least relevant and 3 the most). To help with compiling term lists, researchers might read text passages identified as relevant to the research questions in previous studies. Since frequently occurring ambiguous terms will result in many irrelevant paragraphs, they should be given a lower ranking or excluded. During this step and the next, researchers should expect to make extensive use of corpus analysis toolkits in order to check word and lemma usage in the corpus and their frequencies, e.g. the Autosearch system[11] developed by the Institute for the Dutch Language, based on their corpus indexing system BlackLab[12] and WhiteLab corpus interface (Reynaert et al., 2014).

Step 3. Term expansion HitPaRank is run in mode (ii) to identify all matches for the wildcarded terms found in the corpus, e.g. matching 'know*' to 'knowledge' and 'knowing' and 'acknowledge' if desired. Wildcarded terms are manually filtered and revised (if they caught too much or too little), for inclusion in expanded term lists. Lists may include artefacts of OCR errors such as e.g. 'knowedge', which should be retained in order to ensure retrieval of relevant paragraphs. The expansion is run again followed by the filtering of new wildcard matches and finalising the rankings of all terms, with some terms moving between lists if appropriate. The wildcard matches are marked as undesirable/desirable expansion, and their rankings may be changed depending on their relevance: this information is used in the next step.

Step 4. Paragraph retrieval HitPaRank is run in mode (iii) to retrieve all paragraphs matching any term in the lists (subject to markings of undesirable expansions and exact matches only), and generate statistics (what terms from what lists occur in each paragraph, corpus-level statistics about term occurrence). The output of HitPaRank allows researchers to filter the set of paragraphs returned in various ways, so that they can get a sense of the distribution of terms from different lists across paragraphs, and to select a manageable amount of paragraphs for annotation, e.g. by including only paragraphs that contain combinations of specified terms and/or terms from specified lists with specified rankings.

Step 5. Paragraph annotation The selected paragraphs are annotated for relevance to the research question, i.e. relevant, mildly relevant or not relevant. *Relevance* here is judged by experts according to the extent to which a paragraph provides clear, direct and explicit evidence for answering the research question. A paragraph that can help to answer a research question directly, without much further interpretation, is relevant. A paragraph requiring extensive expert interpretation, or only relating to a small part of the question, is marked as mildly relevant. A decision should be made as to paragraphs that explicitly endorse or reject a relevant quote from another author; they can e.g. be marked as mildly relevant, but one might want to remove

[11] `https://portal.clarin.nl/node/4222`
[12] `https://inl.github.io/BlackLab/`

paragraphs written entirely by other authors (e.g. interview questions or prefaces), and footnotes. Annotation should be done by multiple experts in multiple rounds, ideally striving for consensus.

Step 6. Corpus-dependent concept-modeling ground truth At this point, a conceptual ground truth is created that is dependent on the corpus and controlled by experts. It indicates parts of the corpus that are relevant for the research questions and that therefore should be retrieved when one is after the concept at issue. In order to use a ground truth thus obtained for evaluation of DSMs, one needs to identify the right terms and paragraphs for tuning and testing. When tuning or testing for terms, terms should be chosen that appear in the term lists, and show up in the initial reformulating of research questions into term-based questions. Terms to be chosen for tuning or testing should not be merely heuristic, that is, no term should be picked for this aim that has merely a strong correlation with interesting units of texts; the term should be conceptually relevant to the research question. When tuning or testing for paragraphs, there are a few more considerations that should be taken into account: (i) the paragraph should be high scoring, so highly relevant to the question, (ii) the paragraph should contain terms that are or could be good query terms in the lists, (iii) ideally the paragraph does not contain proper names or any other specific non-conceptual information such as names of events or places, and (iv) the paragraph should not have (many) ambiguous or rare non-technical terms.

5 QuiNE-GT

QuiNE-GT[13] is a ground truth$_{xc}$ built following the procedure described in the previous section. It comprises five lists of terms and three clusters of paragraphs relevant to three different research questions, each paragraph annotated as conveying *relevant*, *mildly relevant* or *not relevant* information for the research question. Here are the three research questions (RQ):

RQ1 is *epistemology* an autonomous enterprise with respect to *science*?

RQ2 is the term *science* only meant to include *natural science*?

RQ3 what is the relation between *epistemology* and *pseudo-science*?

It is important to note that these questions should not be read as philosophical questions about science or epistemology in an absolute sense, but as questions regarding the ideas expressed by Quine using the words *epistemology* and *science* in this corpus. QuiNE-GT is not a ground truth for e.g. epistemology but for epistemology *in Quine*. This is important to note also because the terms in italics in RQ1-RQ3 mark technical terms that belong to standard philosophical vocabulary, but that exactly because of this, paradoxically, are also highly ambiguous and context-dependent: the way they are used in various philosophical texts might require years of study. At the same time, we can rely on the fact that certain words will likely appear whenever a philosopher talks about epistemology, for instance *know*, *knowledge*, *warrant*, *justification*, *source* – we just do not know which ones exactly, which other possible technical neologisms will accompany them, and we need to be aware that these words are possibly quite different in meaning from philosopher to philosopher. It is this very fact that requires interpreters to fix explicit interpretive models for terms and concepts such as these at step 1.

Research questions RQ1-RQ3 are motivated by a debate in the philosophical literature regarding certain aspects of the concept of *naturalism* in Quine's work. *Naturalism* is an obscure term depending on context for its specification, and although the term comes compactly in one word, its meaning needs articulation in several complex components. The unpacking of the components of the meaning in context of terms such as this is usually given by analytic definitions or characterizations of the term to be explicated. Much philosophical business consists in writing up characterisations of terms and concepts in order to clarify the relations these terms have with other terms or concepts. In fact, the research questions above are formulated in such a way as

[13] https://github.com/YOortwijn/QuiNE-ground-truth

to create the chance to reveal the notion of naturalism that Quine adopts, by finding an answer through textual evidence. The questions stem from the following (fragment of a) model for the concept of naturalism set up by consensus at step 1 by the Quine experts in our team:

1. Epistemology is an autonomous enterprise with respect to science;

2. By 'science' in 1. only the natural sciences are meant.

This small twofold model makes it possible to elucidate an existing interpretive debate on the notion of naturalism that Quine adopts. One can look at 1. and 2. as claims one might agree or disagree with, and the combination of agreement and disagreement on these two questions can be used to explain what it means that someone is a naturalist or not. It is uncontroversial that Quine rejected 1. On Quine's stand on 2., however, there is controversy: according to Haack (1993), Quine's stance on 2. is ambiguous; according to Verhaegh (2017), Quine's stance on 2. is a rather clear rejection. In the Haack/Verhaegh debate, a few textual fragments relevant to 1. and 2. are used to settle this interpretive issue (Haack cites nine passages). By applying computational tools on our QUINE corpus we can however potentially retrieve *every* paragraph relevant to this debate, and build a ground truth$_{xc}$ to evaluate DSMs for further investigations into Quine's ideas. And so we proceeded to step 2-5 and generate term lists (functional synonyms, antonyms, hyponyms, hypernyms of potentially important terms and other types of related terms, ranked by the experts), and retrieved and annotated paragraphs.

> List 1 consists of 64 terms of immediate evidential weight for RQ1 — the term we model itself (*naturalism*) and those that are immediately related (e.g., *first philosophy*, *scientism*).
>
> List 2 consists of 70 terms that are related to epistemology (e.g., *evidence*, *know*). These terms are primarily relevant for RQ1.
>
> List 3 consists of 907 terms for all (scientific and scholarly) disciplines and their subfields. Among other resources, the HESA[14] academic coding system was used to get an overview of all academic subjects and their subfields. These terms are primarily relevant for RQ2.
>
> List 4 consists of 45 terms for the building blocks and the procedures of any scientific enterprise (e.g., *theory*, *truth*). These terms are relevant for RQ1 and RQ2.
>
> List 5 consists of 79 terms for pseudosciences (e.g. *parapsychology*) and what we called 'non-sciences' (e.g., *religion*) and their basic faculties (e.g., *clairvoyance*, *revelation*). These terms are relevant for RQ3.

For **RQ1**, we retrieved 179 paragraphs filtering for those with at least one occurrence of *naturali**. For **RQ2**, we retrieved 466 paragraphs by using four different restrictions: (i) 150 paragraphs containing *natural science* or *hard science*, (ii) 283 paragraphs containing *science* and at least one high-ranking (3) term from list 3, (iii) 10 paragraphs containing *humanities* or *soft sci** or *social science**, (iv) 23 paragraphs containing *science* and *boundar**. For **RQ3**, we retrieved 185 paragraphs with each at least one hit from list 5.

The paragraphs were annotated for relevance by three experts (Betti, Oortwijn and Ossenkoppele). First, all experts annotated one set of paragraphs independently, after which they discussed different results until they reached consensus (100% interannotator agreement). Disagreement resolution was done case by case, and mainly in five ways: (i) argumentative deliberation on the nature of the task, (ii) by increasing background knowledge to all annotators in case it was not common to all, (iii) by discussing an annotator's inconsistent scoring, (iv) by evaluating different readings of certain paragraphs and deciding on one by reaching interpretive consensus on the matter, or just (v) by pointing out to each other obvious mistakes. One example of (i) is

[14] https://www.hesa.ac.uk/support/documentation/jacs/jacs3-detailed

	RQ1	RQ2	RQ3
relevant	100	76	45
mildly relevant	38	70	53
not relevant	41	320	87
total	179	466	185

Table 1: Paragraph annotations per research questions

the case in which the annotators decided to exclude highly relevant paragraphs in which Quine reports, paraphrases or cites another author. Another example was a case in which the scoring depended on deciding how much expert background knowledge was admissible for a paragraph to still count as giving clear, explicit and direct evidence for a paragraph, and thus get the highest score. Before disagreement resolution, the initial interrater reliability (Cohen's Kappa) for RQ1 was $\kappa \approx 0.91$, for RQ2 $\kappa \approx 0.69$, for RQ3 $\kappa \approx 0.65$. After one round of disagreement resolution, interannotator agreement reached $\kappa = 1$. The function of the disagreement resolution was to ensure that each annotator used the same standards and rationale in annotating so that a second set of passages could be scored by a single annotator only. About 64% of the passages were annotated by multiple annotators. The results of the annotations are in Table 1.

6 Discussion

We have demonstrated how methodology from natural language processing research can be combined with methodology of the history of ideas to yield a concept-modeling ground truth controlled by domain experts. While ours is a resource-intensive approach, it can be broadly applied to tasks where language models make use of meaning representations that need to be grounded in an accurate conceptual model going beyond what is available in general resources such as WordNet. Specifically, in distributional semantic modeling, our approach can address methodological issues in evaluating the quality of representations trained on specialized, domain-specific text by serving as a gold standard. In future work, our approach can be used to evaluate the quality of meaning representations created by various DSMs from texts that are central to the history of specific scientific ideas. Our methodology for creating ground truths$_{xc}$ can also be further developed into one that can be used to compare differences among expressions of concepts, particularly when applied to diachronic data to study semantic shifts across time.

Ground truth$_{xc}$ construction As mentioned, the units to annotate are, ideally, paragraphs, as these tend to correspond to one thought. A single, isolated sentence is e.g. hard to annotate because such a short unit rarely gives enough evidence to answer a research question, while a section might contain a lot of irrelevant, and thereby distracting, data. Segmentation at paragraph level is however a challenging requirement: constructing or finding a sizeable corpus with acceptable quality for text analysis is already a major obstacle; getting a high-quality properly segmented corpus adds substantially to the difficulty. Moreover, although units segmented in a replicable and automatic way are necessary to construct and properly evaluate DSMs, philosophers and humanities experts are hardly used to build evidence up out of such segments: philosophers usually refer to the vague concept of a 'passage' of variable length as textual unit of evidence.

In step 4 of our method, double filtering takes place on the basis of term list hits: the experts select paragraphs among those already selected by HitPaRank. This move has pragmatic reasons and its results depend on contingent factors (how many paragraphs are retrieved with different filters, how the terms in the lists relate to each other, etc). It also remains methodologically quite risky business: this type of filtering has immediate bearing on the final ground truth, ending up favoring precision over recall. In corpora that are too big for fully manual annotation, however, it is unworkable to annotate every paragraph. In such cases, the use of a tool like HitPaRank and further expert filtering are mandatory, but the type of filtering needs special care. When

e.g. all terms are equally important to the research question, using a specific term as filter might not be sensible. In this case, purposive sampling (e.g. annotating a random sample out of the first HitPaRank selection) or a mix of methods might be the way to go.[15] In other cases, a trial and error approach with different filters might be fitting (by e.g. first checking whether at least some relevant units are retrieved with certain filters). The amount of annotated units should also match the purpose: training of certain DSMs would require a rather large sample, therefore, the terms used for filtering should be such to give the most relevant results without giving too many to annotate but enough for training purposes. If this is not attainable, then a different research question should be considered.

Use of the ground truth The last point brings us to stress that a ground truth$_{xc}$ is only suitable for the specific question it was designed to answer. The ground truth$_{xc}$ created by our method can be used in computational settings to evaluate what model or family of models is better suited to model the target semantic domain (i.e. epistemology in Quine). One way to do this would be to design an information retrieval task where the task is to retrieve the paragraphs that are relevant to the research question, out of the full corpus of texts from the ground truth. Such a task would serve as an extrinsic evaluation of an underlying DSM, or as direct evaluation of IR systems. In an IR task, a research question from the ground truth will have to be translated by experts to one or multiple queries fitting the IR system in question. We could e.g. use Doc2Vec (Le and Mikolov, 2014) or Ariadne (Wang and Koopman, 2017) to create vectors for each paragraph in the ground truth, and then turn the query into a vector in the same semantic space, e.g. by summing the vectors of the words in the query. In such a system, the paragraph vectors that are closest to the query vector are the query results, and they should be relevant. When performing this task with multiple models, we can say that the better system or model for this text domain is the one that retrieves the most paragraphs that are relevant in the ground truth on the basis of that query, while retrieving the fewest paragraphs that are irrelevant in the ground truth. We can also say that a model that performs this task well, is likely to have good semantic representations of the target concepts and is a good conceptual model for that corpus. While this may seem like a highly specific result, models or hyperparameter settings that perform well in the domain of epistemology in Quine may also perform well on other smaller corpora of concept-dense texts in specific fields. Furthermore, if our method is used to construct multiple ground truths covering different domains and different kinds of corpora, we may be able to draw more general conclusions about the performance of various DSMs used in NLP.

Future work When terms in the term lists are checked against the corpus in step 4 of the ground truth construction, the terms are lemmatised. In many cases lemmatisation is good since in concept-focused domains we are not interested in specific syntactical versions of words but in the concepts they express. However, there are cases in which one ideally only uses the full exact match of terms. For instance, the bigram *classical studies* is a part of list 3 as it is a scientific discipline. Treated as a lemma, this also gives us paragraphs containing the bigram *classical study*, which does not refer to a scientific discipline. Future work will try to improve on this by making it possible to choose between exact string matches or lemmas. Further, we plan to experiment with additional clustering of the term lists into classes of various levels of conceptual granularity. This would allow us to make a pattern analysis of highly relevant paragraphs and to investigate where this type of conceptual patterns correlates with relevance.

Acknowledgements

This research was supported by grants *e-Ideas* (VICI, 277-20-007) and *CatVis* (314-99-117), funded by the Dutch Research Council (NWO), and by the Human(e)AI grant *Small data, big challenges* funded by the University of Amsterdam.

[15] For a comparison of evaluation methods as to recall and precision, and the desirability of combining approaches, see Bloem (2016), proposing the use of lower representations (strings) to verify annotation of higher ones (concepts).

References

Charu C. Aggarwal. 2016. *Recommender systems*, volume 1. Springer.

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

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

Arianna Betti and Hein van den Berg. 2016. Towards a computational history of ideas. In Lars Wieneke, Catherine Jones, Marten Düring, Florentina Armaselu, and René Leboutte, editors, *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.

Arianna Betti, Hein van den Berg, Yvette Oortwijn, and Caspar Treijtel. 2019. History of philosophy in ones and zeros. In Mark Curtis and Eugen Fischer, editors, *Methodological Advances in Experimental Philosophy*, Advances in Experimental Philosophy, pages 295–332. Bloomsbury, London.

Jelke Bloem, Antske Fokkens, and Aurelie Herbelot. 2019. Evaluating the consistency of word embeddings from small data. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2019*, Varna, Bulgaria.

Jelke Bloem, Maria Chiara Parisi, Martin Reynaert, Yvette Oortwijn, and Arianna Betti. 2020. Distributional semantics for neo-Latin. In *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 84–93, Marseille, France. European Language Resources Association (ELRA).

Jelke Bloem. 2016. Evaluating automatically annotated treebanks for linguistic research. In *Proceedings of the 4th Workshop on Challenges in the Management of Large Corpora*, pages 8–14, Mannheim. Institut für Deutsche Sprache.

Stephen Clark. 2015. Vector space models of lexical meaning. In Shalom Lappin and Chris Fox, editors, *The Handbook of Contemporary semantic theory*, pages 493–522. Wiley Online Library.

Stéphane Clinchant and Florent Perronnin. 2013. Aggregating continuous word embeddings for information retrieval. In *Proceedings of the Workshop on Continuous Vector Space Models and their Compositionality*, pages 100–109, Sofia, Bulgaria. Association for Computational Linguistics.

Lieven Decock and Igor Douven. 2011. Similarity after Goodman. *Review of Philosophy and Psychology*, 2(1):61–75.

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

Christiane Fellbaum. 1998. *WordNet: An electronic lexical database*. Bradford Books.

Antske Fokkens, M. Erp, Marten Postma, Ted Pedersen, Piek T. J. M. Vossen, and Nuno Freire. 2013. Offspring from reproduction problems: What replication failure teaches us. In *ACL*.

Antske Fokkens, Serge Ter Braake, Niels Ockeloen, Piek Vossen, Susan Legêne, Guus Schreiber, et al. 2014. Biographynet: Methodological issues when NLP supports historical research. In *LREC*, pages 3728–3735.

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 Andreas de Block and Grant Ramsey, editors, *The Dynamics of Science: Computational Frontiers in History and Philosophy of Science*. Pittsburgh University Press, Pittsburgh.

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.

Susan Haack. 1993. The two faces of Quine's naturalism. *Synthese*, 94(3):335–356.

Zellig S. Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Johannes Hellrich. 2019. *Word embeddings: Reliability & semantic change*, volume 347 of *Dissertations in Artificial Intelligence*. IOS Press, Amsterdam. Google-Books-ID: 92OwDwAAQBAJ.

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.

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 Conference on Information and Knowledge Management*, pages 1191–1200.

Barbara A. Kipfer. 2019. *Roget's international thesaurus, 8th edition*. HarperCollins.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International Conference on Machine Learning*, pages 1188–1196.

Keqian Li, Hanwen Zha, Yu Su, and Xifeng Yan. 2018. Concept mining via embedding. In *2018 IEEE International Conference on Data Mining (ICDM)*, pages 267–276, Los Alamitos, CA, USA. IEEE Computer Society.

Pasquale Lops, Marco de Gemmis, and Giovanni Semeraro. 2011. Content-based recommender systems: State of the art and trends. In Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor, editors, *Recommender Systems Handbook*, pages 73–105. Springer US, Boston, MA.

Edouard Machery. 2010. Précis of doing without concepts. *Behavioral and Brain Sciences*, 33(2-3):195–206.

Eric Margolis and Stephen Laurence. 2019. Concepts. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2019 edition.

Claudio Masolo, Emilio Sanfilippo, Marion Lamé, and Perrine Pittet. 2019. Modeling concept drift for historical research in the digital humanities. In *1st International Workshop on Ontologies for Digital Humanities and their Social Analysis (WODHSA)*.

Letizia Mencarini. 2018. The potential of the computational linguistic analysis of social media for population studies. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, New Orleans, Louisiana, USA. Association for Computational Linguistics.

François Meyer, Yvette Oortwijn, Pia Sommerauer, Jelke Bloem, and Antske Fokkens. 2020. The semantics of meaning: distributional approaches for studying philosophical text. In *Proceedings of the Network Institute Academy Assistants programme 2018-2019*. Vrije University Amsterdam. In press.

Dong Nguyen, Maria Liakata, Simon DeDeo, Jacob Eisenstein, David Mimno, Rebekah Tromble, and Jane Winters. 2020. How we do things with words: Analyzing text as social and cultural data. *Frontiers in Artificial Intelligence*, 3:1–14.

Farhad Nooralahzadeh, Lilja Øvrelid, and Jan Tore Lønning. 2018. Evaluation of domain-specific word embeddings using knowledge resources. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

Thijs Ossenkoppele. 2019. Quine's naturalistic epistemology; A quantitative investigation. Unpublished BA thesis.

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

Martin Reynaert, Matje van de Camp, and Menno van Zaanen. 2014. OpenSoNaR: user-driven development of the SoNaR corpus interfaces. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*, pages 124–128, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Magnus Sahlgren. 2008. The distributional hypothesis. *Rivista di linguistica*, 20:33–53.

Jingbo Shang, Jialu Liu, Meng Jiang, Xiang Ren, Clare R. Voss, and Jiawei Han. 2018. Automated phrase mining from massive text corpora. *IEEE Transactions on Knowledge and Data Engineering*, 30(10):1825–1837.

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.

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.

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

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

Maarten van Gompel and Martin Reynaert. 2013. FoLiA: A practical XML Format for Linguistic Annotation - a descriptive and comparative study. *Computational Linguistics in the Netherlands Journal*, 3:63–81.

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

Sander Verhaegh. 2017. Quine on the nature of naturalism. *Southern Journal of Philosophy*, 55(1):96–115.

Shenghui Wang and Rob Koopman. 2017. Semantic embedding for information retrieval. In *BIR 2017: 5th Workshop on Bibliometric-enhanced Information Retrieval 2017*, pages 122–132. CEUR.

Hamed Zamani and W. Bruce Croft. 2017. Relevance-based word embedding. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 505–514.