From Word Types to Tokens and Back: A Survey of Approaches to Word Meaning Representation and Interpretation

Marianna Apidianaki*
University of Pennsylvania
Department of Computer and Information Science
marapi@seas.upenn.edu

Vector-based word representation paradigms situate lexical meaning at different levels of abstraction. Distributional and static embedding models generate a single vector per word type, which is an aggregate across the instances of the word in a corpus. Contextual language models, on the contrary, directly capture the meaning of individual word instances. The goal of this survey is to provide an overview of word meaning representation methods, and of the strategies that have been proposed for improving the quality of the generated vectors. These often involve injecting external knowledge about lexical semantic relationships, or refining the vectors to describe different senses. The survey also covers recent approaches for obtaining word type-level representations from token-level ones, and for combining static and contextualized representations. Special focus is given to probing and interpretation studies aimed at discovering the lexical semantic knowledge that is encoded in contextualized representations. The challenges posed by this exploration have motivated the interest towards static embedding derivation from contextualized embeddings, and for methods aimed at improving the similarity estimates that can be drawn from the space of contextual language models.

1. Introduction

Word representation in vector space lies in the core of distributional approaches to language processing. The idea that words' collocations describe their meaning (Harris 1954; Firth 1957) underlies Distributional Semantic Models (DSMs) and the structure of the semantic space built by neural language models. Different approaches, however, address different units of meaning representation. DSMs represent words by aggregating over their usages in a corpus of documents (Landauer and Dumais 1997; Lund and Burgess 1996). Similarly, word embedding approaches such as word2vec, GloVe, and fastText generate a static vector per word type, which groups its different senses (Mikolov et al. 2013a; Pennington, Socher, and Manning 2014; Bojanowski et al. 2017).

*Part of the work was accomplished when the author was affiliated with the University of Helsinki.

Action Editor: Ekaterina Shutova. Submission received: 15 February 2022; revised version received: 20 October 2022; accepted for publication: 5 November 2022.

https://doi.org/10.1162/coli_a_00474

Contextual language models, on the contrary, generate dynamic representations that change for every new occurrence of a word in texts and directly encode the contextualized meaning of individual tokens (Peters et al. 2018; Devlin et al. 2019; Liu et al. 2019). Contrary to a static embedding model which would propose a single vector for a polysemous word like *bug*, a contextual model would generate different representations for instances of the word in context (e.g., "There is a *bug* in my soup", "There is a *bug* in my code").

Contextualized representations constitute a powerful feature of state-of-the-art language models, and contribute to their impressive performance in downstream tasks. Their flexibility confers them an undeniable advantage over static embeddings which, by aggregating information from different contexts in the same word vector, often lead to the "meaning conflation" problem (Pilehvar and Camacho-Collados 2020). Additionally, the dynamic nature of contextualized vectors provides a more straightforward way for capturing meaning variation than previous sense representation methodologies (Reisinger and Mooney 2010; Iacobacci, Pilehvar, and Navigli 2015; Camacho-Collados and Pilevar 2018). In DSMs, this type of contextualization was performed through word vector composition, where the basic vector for a word was adapted to a new context of use by being combined with the vectors of the words in the context (Mitchell and Lapata 2008; Erk and Padó 2008; Thater, Fürstenau, and Pinkal 2011; Dinu and Lapata 2010; Dinu, Thater, and Laue 2012). In deep contextual language models, every word is influencing every other word in a sequence and all the representations are getting updated in different layers of the model based on this distributional information.

The dynamic character of contextualized representations also poses some challenges for meaning representation. Although modeling word usage is one of their recognized merits and a highly useful methodological tool for studying linguistic structure (Linzen, Dupoux, and Goldberg 2016; Hewitt and Manning 2019), the observed context variation makes the study of the encoded semantic knowledge challenging (Ethayarajh 2019b; Mickus et al. 2020; Timkey and van Schijndel 2021). We, thus, witness in recent work a resurgence of interest towards more abstract, higher (word type) level, representations, deemed to provide a more solid basis for meaning exploration. Naturally, this trend is mainly observed in the lexical semantics field where the notion of lexical concept is central (Lauscher et al. 2020; Liu, McCarthy, and Korhonen 2020; Bommasani, Davis, and Cardie 2020; Vulić et al. 2020b; Garí Soler and Apidianaki 2021a).

The prevalence of contextual models in the field of computational linguistics has also brought about a shift from out-of-context word similarity and analogy tasks—used for evaluating static embedding quality (Mikolov et al. 2013b)—to interpretation tools common in human language learning studies (such as cloze tasks and probes) (Linzen, Dupoux, and Goldberg 2016; Kovaleva et al. 2019; Tenney et al. 2019; Ettinger 2020). These serve to assess the linguistic and world knowledge encoded in contextualized vectors, and are often complemented with methods that explore the models' inner workings (Voita, Sennrich, and Titov 2019; Hewitt and Manning 2019; Clark et al. 2019; Voita et al. 2019; Tenney, Das, and Pavlick 2019). In lexical semantics, probing is used to explore the knowledge that the models encode about the semantic properties of words and their relationships (Petroni et al. 2019; Bouraoui, Camacho-Collados, and Schockaert 2020; Ravichander et al. 2020; Apidianaki and Garí Soler 2021), or their understanding of semantic scope and negation (Ettinger 2020; Lyu et al. 2022). Nevertheless, evaluations that rely on probing are not always indicative of the knowledge that is encoded by the models. Language models are brittle to small changes in the used prompts, and the output strongly depends on prompt quality and naturalness (Ettinger 2020; Ravichander et al. 2020; Apidianaki and Garí Soler 2021; Jiang et al.

2020). Furthermore, the output of semantic probes is difficult to evaluate, since there might be multiple valid answers and possible fillers (e.g., *red*, *tasty*, and *fruits* would all be good fillers for the masked slot in the query "*Strawberries are* [MASK]"). These issues have brought attention back to word similarity and analogy tasks, considered to be more established and mature for exploring the concept-related knowledge encoded in language model representations (Vulić et al. 2020b; Bommasani, Davis, and Cardie 2020).

Survey Goal. The goal of this survey is to provide an overview of word meaning representation methodologies and evaluation practices. It will put current developments into perspective with respect to previous representation and evaluation paradigms, discuss their specificities, and highlight the issues that have been addressed and the challenges that remain. Special focus will be put to word type (static) and word token (dynamic) embedding approaches. We will also discuss methods for deriving word type-level vectors from contextualized representations. This back-and-forth between representation types and evaluation strategies nourishes active discussions in the community. Our goal is to clarify their respective strengths and weaknesses, and to open up perspectives for future research.

The overview of the methods proposed in this survey is not intended to be exhaustive. Our main concern has been to include work that is representative of the evolution and trends on the topic of word meaning representation in the past years. Nevertheless, given the pace in which the field evolves and the actual space constraints this publication needs to abide by, it is practically impossible to include a full account of existing work. Furthermore, the majority of the methods and datasets that will be presented have been developed for the English language. We include a discussion of results obtained in other languages when needed in order to highlight the cross-lingual generalization potential of the presented methods—or their limitations in this respect—as well as the methodological differences and design choices that apply in a multilingual setting.

Survey Outline. An overview table of the survey contents is given in Figure 1. Section 2 presents methodologies that generate embeddings at the level of word types and word

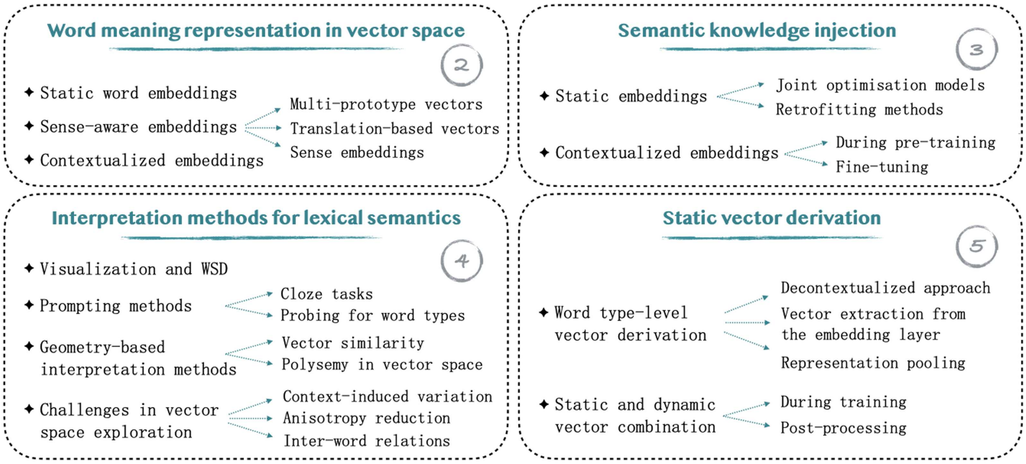

Word meaning representation in vector space
2
✦ Static word embeddings
✦ Sense-aware embeddings
✦ Contextualized embeddings
Multi-prototype vectors
Translation-based vectors
Sense embeddings
Semantic knowledge injection
3
✦ Static embeddings
Joint optimisation models
Retrofitting methods
✦ Contextualized embeddings
During pre-training
Fine-tuning
Interpretation methods for lexical semantics
4
✦ Visualization and WSD
✦ Prompting methods
Cloze tasks
Probing for word types
✦ Geometry-based interpretation methods
Vector similarity
Polysemy in vector space
✦ Challenges in vector space exploration
Context-induced variation
Anisotropy reduction
Inter-word relations
Static vector derivation
5
✦ Word type-level vector derivation
Decontextualized approach
Vector extraction from the embedding layer
Representation pooling
✦ Static and dynamic vector combination
During training
Post-processing

Figure 1
Overview table of the survey contents. The numbers refer to Sections 2 to 5.

tokens. We discuss their strengths and limitations, as well as solutions that have been proposed to address the latter, including the generation of embeddings at the level of senses. The section includes a critical presentation of benchmarks commonly used for evaluating word embedding quality. Section 3 presents methods that specialize static and contextualized word embeddings for semantic relationships during (pre-)training or at a post-processing stage. Section 4 presents interpretation and evaluation methods aimed at exploring the semantic knowledge that is encoded in contextual embedding representations. We discuss the challenges posed by probing methodologies for lexical semantic analysis. We also explain how the geometry of the vector space that is built by contextual language models can provide insights into the quality of the representations, and highlight factors that might complicate the derivation of high quality similarity estimates. In Section 5, we present methods that generate word type-level representations from contextualized vectors, and methods that combine static and dynamic embeddings in order to leverage their respective strengths and address their limitations. The Conclusion includes a discussion of perspectives for future work in word meaning representation.

2. Word and Meaning Representation in Vector Space

This section provides an overview of word and meaning representation methodologies that rely on language models. We present approaches that generate distributed representations (embeddings) at the level of word types, senses, and tokens. Links with distributional approaches are established when needed in order to better understand the evolution of embedding representations, or to explain their advantages over count-based distributional models. For a full account of distributional approaches and their origins, we point the reader to the survey paper by Turney and Pantel (2010). The interaction between distributional and formal semantics is explained in Boleda and Herbelot (2016). For a thorough look into embeddings generated by different types of language models, we refer the reader to the book by Pilehvar and Camacho-Collados (2020).

2.1 Static Word Embeddings

2.1.1 Vector Creation. Word embedding models leverage neural networks to learn low-dimensional word representations from corpora (Bengio et al. 2003; Collobert and Weston 2008; Collobert et al. 2011; LeCun, Bengio, and Hinton 2015; Mikolov et al. 2013a). These "self supervision" models are trained on raw text and rely on the optimization of a language modeling objective. The vector estimation problem is framed directly as a supervised task, where the weights in a word vector are set to maximize the probability of the contexts in which the word is observed in the corpus used for training. The popular Continuous Bag-of-Words (CBOW) word2vec model architecture (Mikolov et al. 2013a) is based on the feedforward neural network language model (Bengio et al. 2003); the task is to predict the current word (w_t) using its surrounding context ($W_t = w_{t-n}, \ldots, w_t, \ldots, w_{t+n}$) minimizing a loss function. In word2vec Skip-gram, on the contrary, the goal is to predict the words in the surrounding context given the target word (w_t).

The idea underlying word embedding models is that contextual information can provide a good approximation to word meaning since semantically similar words tend to have similar contextual distributions (Harris 1954; Firth 1957; Miller and Charles 1991). This is also the guiding principle of DSMs (Turney and Pantel 2010; Erk 2012;

Clark 2015), while the idea of meaning as distribution goes back to Wittgenstein (1953) who wrote that 'the meaning of a word is its use in the language'.[1] In count-based methods, vectors keep track of the contexts where words appear in a large corpus (i.e., their co-occurrences) as proxies for meaning representation. In both cases, word similarity can be measured by applying geometric techniques (e.g., cosine similarity or Euclidean distance) to the obtained (embedding or count-based) vectors. Additionally, similar to DSMs, the self-supervised embedding learning approach requires no manual annotations and the models can be trained on raw text. Both methodologies can thus be applied to different languages given that large-scale unannotated corpora are available.

Low dimensionality is considered an advantage of word embeddings over count-based vectors (Baroni, Dinu, and Kruszewski 2014). In DSMs, vector dimensions correspond to words in the vocabulary (V) so their number can easily reach hundreds of thousands or even millions, depending on the corpus the vectors are trained on. Storing each word $w \in V$ in a $|V|$-dimensional vector results in a very large matrix with $|V|^2$ cells. Additionally, the generated vectors are sparse, containing a small number of non-zero elements. The high dimensionality and sparseness of distributional vectors challenge both the scalability of the models and their computational efficiency.

A common approach to alleviate the sparseness of distributional representations and improve their performance in semantic tasks is to apply some type of transformation to the raw vectors. This involves reweighting the counts for context informativeness and smoothing them with dimensionality reduction techniques (e.g., Singular Value Decomposition [SVD]) (Turney and Pantel 2010). The applied optimization process is generally unsupervised and based on independent (for example, information-theoretic) considerations (Baroni, Dinu, and Kruszewski 2014). Such transformations are not needed with word embedding techniques which involve a single supervised learning step. Word embedding models generate low-dimensional vectors which are more compact than count-based vectors. Consequently, similarity calculations and other operations on these vectors are fast and efficient.[2] Much of the power of word embedding models derives from their ability to compress distributional information into a lower-dimensional space of continuous values.

Strengths and Limitations. Pretrained word embeddings outperform count-based representations in intrinsic evaluations (i.e., word similarity and relatedness tasks) (Mikolov, Yih, and Zweig 2013; Baroni, Dinu, and Kruszewski 2014), and can be successfully integrated in downstream applications due to their high generalization potential. They also present limitations. Models like word2vec (Mikolov et al. 2013a), GloVe (Pennington, Socher, and Manning 2014), and fastText (Bojanowski et al. 2017), for example, are by design unable to model polysemy, since they build a single representation for each word in the vocabulary of a language. The contextual evidence for different word meanings is thus conflated into a single vector.

Modeling a word type as a single point in the semantic space is considered as a major deficiency of static embedding models. Not distinguishing between different meanings of a polysemous word (e.g., *plant*, *mouse*, *bug*) can negatively impact the semantic understanding of NLP systems that rely on these representations. Additionally, meaning conflation has consequences on the structure of the obtained semantic

[1] In Wittgenstein (1953), use is perceived as the situational context of communication. Firth (1957) views words' habitual collocations as their context of use.

[2] Tools that make the manipulation of word embeddings faster and more efficient have also been developed (Patel et al. 2018).

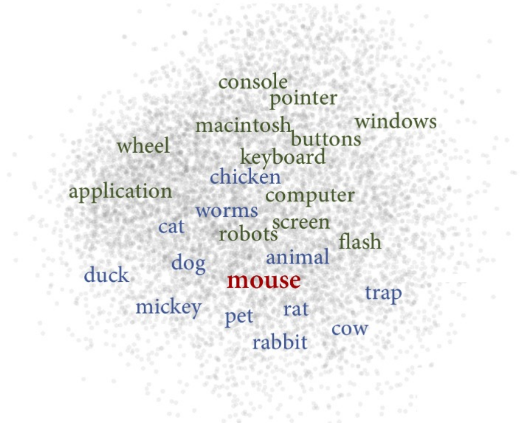


Figure 2
Illustration of word embeddings' meaning conflation deficiency in a 2D semantic space. Representing an ambiguous word (*mouse*) as a single point in space pulls together semantically unrelated words (e.g., *keyboard*, *chicken*, *screen*) (Camacho-Collados and Pilevar 2018).

space and on semantic modeling accuracy, since the vectors of unrelated words are pulled closer together (Neelakantan et al. 2014; Chen, Liu, and Sun 2014; Camacho-Collados and Pilevar 2018). This is illustrated in Figure 2 by the proximity of *rat*, *cat*, and *keyboard*, due to their similarity to different senses of the noun *mouse*. A careful analysis shows that multiple word senses reside in linear superposition within word2vec and GloVe word embeddings, and that vectors that approximately capture the senses can be recovered using simple sparse coding (Arora et al. 2018). In the distributional semantics literature, context-specific representations for words were generated through vector composition (Schütze 1998; Mitchell and Lapata 2008; Baroni and Zamparelli 2010; Zanzotto et al. 2010), sometimes taking into consideration the syntactic role and selectional preferences of words in the sentence (Padó and Lapata 2007; Erk and Padó 2008; Thater, Fürstenau, and Pinkal 2011).[3]

Another shortcoming of the dense continuous-valued vector representations that are learned by word embedding models is that they lack interpretable dimensions, limiting our understanding of the semantic features they actually encode (Chersoni et al. 2021; Petersen and Potts 2022). This is in contrast to co-occurrence-based distributional vectors, where features can deliver direct and interpretable insights. In spite of their low dimensionality, word embeddings are still able to capture word similarity due to the objective used for training, which makes them create similar vectors for similar words.

The next section describes the methodology most commonly used for evaluating word type-level embeddings.

2.1.2 Static Embedding Evaluation. Word embeddings have often been intrinsically evaluated against manually compiled word analogy, similarity, and relatedness datasets, which test their capability to represent word meaning. This section presents the most common approaches and datasets used in this goal. Although these datasets are not

[3] A model that does not account for syntax would, for example, generate the same representation for the noun *school* in "*law school*" and in "*school law*".

perfect and their use as a test bed for evaluation has often been criticized, they still remain interesting and relevant for this survey.[4]

Word Analogy. Word analogy has been extensively used for evaluating the quality of static word embeddings. It is usually framed as a relational similarity task, and models the idea that pairs of words may hold similar relations to those that exist between other pairs of words (Turney 2006). In the equation $a : b :: c : d$ (which reads as *"a is to b as c is to d"*), the first three terms (*a*, *b*, *c*) are given and the tested model needs to predict the word that stands for *d*. Mikolov, Yih, and Zweig (2013) showed that such relations are reflected in vector offsets between word pairs.[5] In the famous example *"man is to king as woman is to X"*, the embedding for the word *queen* can be roughly recovered from the representations of *king*, *man*, and *woman* using the following equation: $\overrightarrow{queen} \approx \overrightarrow{king} - \overrightarrow{man} + \overrightarrow{woman}$. Benchmarks commonly used for this type of evaluation include the Google analogy set (Mikolov et al. 2013a),[6] the Microsoft Research Syntactic (MSR) analogies dataset (Mikolov, Yih, and Zweig 2013), and the SemEval 2012 Task 2 "Measuring Degrees of Relational Similarity" dataset (Jurgens et al. 2012).

In spite of their popularity, word analogies have been progressively discredited as a test bed for evaluation due to numerous concerns regarding their validity. First, the accuracy of the vector offset method depends on the proximity of the target vector to its source (e.g., $\overrightarrow{queen}$ and $\overrightarrow{king}$), limiting its applicability to linguistic relations that happen to be close in the vector space (Rogers, Drozd, and Li 2017). Reliance on cosine similarity also conflates offset consistency with largely irrelevant neighborhood structure (Linzen 2016). Linzen also notes that results are inconsistent when the direction of the analogy is reversed, even though the same offset is involved in both directions. Moreover, linguistic relations might not always translate to linear relations between vectors but to more complex correspondence patterns (Drozd, Gladkova, and Matsuoka 2016; Ethayarajh 2019b). The classic implementation of the analogy task is also problematic; examples are structured in such a way that given the first three terms, there is one specific correct fourth term. This might be the case with factual queries involving morpho-syntactic and grammatical alternations (e.g., *high* : *higher* :: *long* : *X*), but for semantic queries there might be several equally plausible correct answers (e.g., *man:doctor :: woman:X*) (Nissim, van Noord, and van der Goot 2020).[7] The usual implementation of this type of evaluation, which excludes premise vectors from predictions, is also problematic (Schluter 2018).[8] Finally, the queries often reflect subjective biases that compromise the value of analogies as a bias detection tool.

Semantic Similarity and Relatedness. Another way to evaluate the quality of word representations is to compare their similarity against human semantic similarity and relatedness judgments. A high correlation between human judgments on word pairs and the cosine of the corresponding vectors is perceived as an indication of the quality of

[4] Word embeddings can also be evaluated in downstream applications. However, the complexity of these tasks might blur aspects that matter for assessing embedding quality. We thus focus on intrinsic evaluations in this article.

[5] The answer is represented by hidden vector *d*, calculated as $argmax_{d \in V}(sim(d, c - a + b))$. V is the vocabulary excluding words *a*, *b*, and *c*, and *sim* is a similarity measure.

[6] This comprises "syntactic" analogies (e.g., PLURAL: *banana - bananas*, GERUND: *scream - screaming*) and lexico-semantic analogies (e.g., GENDER: *boy - girl*, COMMON CAPITALS: *Athens - Greece*).

[7] Various terms could be used for completion depending on the implied relation, which might be unspecified in the query (Turney 2012).

[8] In the unconstrained setting where input words are allowed, large drops in performance are observed.

the constructed space. Similarity describes a tighter relationship between synonyms or words linked with "IS-A" (hypernymy) relations (e.g., *a car* IS-A *vehicle*), while related words have some other type of connection (they might be meronyms or holonyms) or are topically associated (Agirre et al. 2009; Bruni, Tran, and Baroni 2014). For example, *house* is similar to *building*, and is also related to *brick* and *garden*.

The bulk of these datasets have been compiled in the context of linguistic and psycholinguistic studies (Rubenstein and Goodenough 1965; Miller and Charles 1991; Hodgson 1991; Finkelstein et al. 2001; Bruni et al. 2012; Hill, Reichart, and Korhonen 2015; Gerz et al. 2016; Pilehvar et al. 2018; Vulić et al. 2020a). Such datasets also serve to assess the proficiency of English language learners (e.g., the TOEFL dataset), and to evaluate distributional models in dedicated shared tasks (Jurgens et al. 2012).[9] However, there are some issues with this type of evaluation too. First, the same word pairs may be rated differently in similarity and relatedness datasets (Bruni et al. 2012; Hill, Reichart, and Korhonen 2015). Second, judgments for related word classes (*cat-dog*) are more reliable than for unrelated words (*cat-democracy*) (Kabbach and Herbelot 2021). Another downside of this type of evaluation is that similarity scores are assigned to pairs of words in isolation. Consequently, the comparison of static embeddings to these scores does not allow to assess the capability of the models to capture polysemy and word meaning in context.

2.2 Sense-aware Embeddings

2.2.1 Multi-prototype Embeddings. Multi-prototype methods were proposed as a solution to the meaning conflation problem of static word embeddings. These methods generate separate vectors for the different senses of a word, which are often discovered from text corpora using unsupervised Word Sense Induction methods. The contexts where a word occurs are clustered, and cluster centroids are used as prototype vectors. The multi-prototype method of Reisinger and Mooney (2010) is illustrated in Figure 3.

Multi-prototype methods vary with respect to the vector representations, the clustering algorithm, and the context used. Reisinger and Mooney (2010) use count-based vectors composed of features that correspond to unigrams in a 10-word context window around a target word w_t, while Huang et al. (2012) and Neelakantan et al. (2014) use word embeddings. For clustering, Reisinger and Mooney apply a mixture of von Mises-Fisher distributions (movMF) clustering method. Huang et al. (2012) use the K-means algorithm to decompose word embeddings into multiple prototypes. In the Multiple-Sense Skip-Gram (MSSG) method of Neelakantan et al. (2014), clustering and sense embedding learning are performed jointly during training. The multi-prototype Skip-gram model of Tian et al. (2014) has fewer parameters and is trained using the Expectation-Maximization algorithm. In contrast to methods where senses are induced from words' local context, Liu et al. (2015) propose Topical Word Embeddings (TWE). This method allows each word to have different embeddings under different topics computed globally using latent topic modeling (Blei, Ng, and Jordan 2003).

Multi-prototype embedding methods offer a way to capture and represent senses, but also face a number of challenges. In early methods, the number of clusters (or senses) k was a parameter that had to be pre-defined. This number was sometimes chosen arbitrarily and used for all words, independently of their polysemy (Huang

[9] Table A1 in the Appendix provides an overview of the available datasets alongside information about the number of word pairs they contain, their grammatical category, the range of similarity scores used, and the number of annotators who provided the similarity judgments.

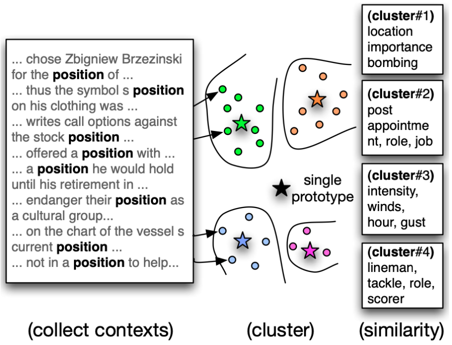
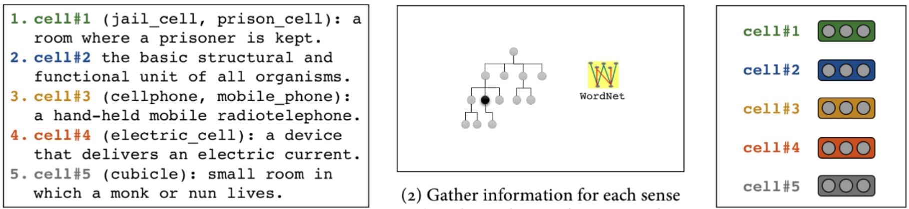

... chose Zbigniew Brzezinski for the position of ...
... thus the symbol s position on his clothing was ...
... writes call options against the stock position ...
... offered a position with ...
... a position he would hold until his retirement in ...
... endanger their position as a cultural group...
... on the chart of the vessel s current position ...
... not in a position to help...
single prototype
(cluster#1) location importance bombing
(cluster#2) post appointme nt, role, job
(cluster#3) intensity, winds, hour, gust
(cluster#4) lineman, tackle, role, scorer
(collect contexts)
(cluster)
(similarity)

Figure 3
Illustration of the multi-prototype approach (Reisinger and Mooney 2010).

et al. 2012). Moreover, these methods are generally offline and difficult to adapt to new data and domains, or to capture new senses (Chen, Liu, and Sun 2014). An alternative has been to use non-parametric clustering which allows to dynamically adjust the number of senses to each word. The method of Neelakantan et al. (2014) precisely relies on the notion of "facility location" (Meyerson 2001); a new cluster is created online during training with probability proportional to the distance from the context to the nearest cluster (sense). The bigger this distance, the higher the probability that the context describes a new sense of the word. Similarly, the method of Li and Jurafsky (2015) learns embeddings for senses of a word induced using the Chinese Restaurant Processes (Blei et al. 2003), a practical interpretation of Dirichlet Processes (Ferguson 1973) for non-parametric clustering. In this approach too, a word is associated with a new sense vector when evidence in the context (its neighboring words) suggests that it is sufficiently different from its previously identified senses.

Other concerns that have been expressed with respect to multi-prototype methods are that the clusters are not always interpretable (i.e., it is difficult to identify the senses they correspond to), and the representations obtained for rare senses are unreliable (Pilehvar and Collier 2016). Finally, the usefulness of using this type of sense embeddings in downstream tasks is unclear. These have been shown to outperform previous word embedding representation methods in intrinsic evaluations, but when tested in real NLP applications they seem to benefit some tasks (part-of-speech tagging and semantic relation identification) and harm others (sentiment analysis and named entity extraction) (Li and Jurafsky 2015).

2.2.2 Translation-based Embeddings. Seeking a more stable criterion than clustering for sense identification, several studies have proposed to use translations as proxies for senses. This idea dates back to work by Gale, Church, and Yarowsky (1992), where it was put forward as a solution to the knowledge acquisition bottleneck, and has since been adopted in numerous word sense induction and disambiguation approaches (Dagan and Itai 1994; Dyvik 1998, 2002, 2005; Resnik and Yarowsky 1999; Ide, Erjavec, and Tufis 2002; Resnik 2004; Diab and Resnik 2002; Apidianaki 2008, 2009; Lefever, Hoste, and De Cock 2011; Carpuat 2013). The underlying assumption is that the senses of a polysemous word in a source language (w_s) are translated with different words ($T = t_1, \ldots, t_n$) in other languages. Clustering is still relevant in this context since sets of synonymous translations may describe the same sense of word w_s (Apidianaki 2008, 2009).


1. cell#1 (jail_cell, prison_cell): a room where a prisoner is kept.
2. cell#2 the basic structural and functional unit of all organisms.
3. cell#3 (cellphone, mobile_phone): a hand-held mobile radiotelephone.
4. cell#4 (electric_cell): a device that delivers an electric current.
5. cell#5 (cubicle): small room in which a monk or nun lives.
WordNet
cell#1
cell#2
cell#3
cell#4
cell#5

(1) Get senses as defined by a sense inventory (e.g., WordNet)

(2) Gather information for each sense (e.g., by exploiting the structural properties of sense inventory's semantic network, and (optionally) then from text corpora)

(3) Compute sense representation

Figure 4
Illustration of the sense embedding approach (Camacho-Collados and Pilevar 2018).

Translations have also served to create embeddings for word senses. Guo et al. (2014) project clusters of English translations describing senses onto Chinese words in a parallel corpus, in order to create the labeled data needed for training a neural network model that generates sense embeddings. The sense embedding method of Šuster, Titov, and van Noord (2016) also exploits monolingual and translation information. Their model consists of an encoding part which assigns a sense to a given word (called "pivot"), and a decoding (or reconstruction) part that predicts context words based on the pivot word and its sense. Parameters of encoding and reconstruction are jointly optimized, the goal being to minimize the error in recovering context words based on the pivot word and its assigned sense. Finally, methods that form rich context-aware features and vectors for source language words and phrases have also served to improve translation quality in Phrase-Based Statistical Machine Translation and Neural Machine Translation systems (Carpuat and Wu 2007; Apidianaki et al. 2012; Liu, Lu, and Neubig 2018).

2.2.3 Sense Embeddings. Sense embedding methods produce vectors for senses found in lexicographic resources, sometimes combining this knowledge with information from large text corpora. A merit of this approach is that the generated sense vectors are more interpretable than clustering-induced senses (Camacho-Collados and Pilevar 2018). A typical sense embedding procedure is illustrated in Figure 4.

The SENSEMBED method of Iacobacci, Pilehvar, and Navigli (2015) and the Senses and Words to Vector (SW2V) method of Mancini et al. (2017) both learn sense representations from disambiguated texts.[10] A difference between the two approaches is that the former produces sense representations only, while the latter jointly learns word and sense embeddings which share the same unified vector space. In both methods, the quality of the generated sense representations strongly depends on the success of the disambiguation step. The method of Chen, Liu, and Sun (2014) alleviates this dependence by learning representations from sense definitions (glosses) in WordNet (Fellbaum 1998). Each sense is represented by the average of the vectors of the content words in the gloss that are most similar to the target word. The training objective of Skip-gram is then modified in order to obtain vectors that are good at predicting not only a word's context words, but also its senses. Similarly, Rothe and Schütze (2017)

[10] SENSEMBED uses Babelfy, a knowledge-based Word Sense Disambiguation algorithm (Moro, Raganato, and Navigli 2014), while SW2V relies on a shallow word-sense connectivity algorithm.

propose a model called AutoExtend which learns embeddings for WordNet synsets. The embedding for a word corresponds to the sum of the embeddings of its senses, and the vector for a synset corresponds to the sum of the senses contained in the synset.

Sense embedding approaches provide a clear solution to the meaning conflation problem of word embeddings, but they are tied to an external semantic lexicon. State of the art contextual language models, on the contrary, produce vectors that capture the meaning of individual tokens in a more straightforward way. The next section describes different contextual language models with special focus on the widely used Transformer-based BERT model.

2.3 Contextualized Embeddings

Contextual language models constitute a new representation paradigm where the generated embeddings encode the meaning of individual word tokens (Peters et al. 2018; Devlin et al. 2019; Liu et al. 2019). Contrary to static embeddings which describe word types (e.g., there is only one word2vec vector for the noun *bug*), contextual models assign different vectors to different instances of the same word depending on the context of use (e.g., "There is a *bug* in my soup", "There is a *bug* in my code"). These vectors are dynamic and can capture subtle meaning nuances expressed by word instances, alleviating, at the same time, the meaning conflation problem of static embeddings and sense embeddings' reliance on lexicographic resources.

Vector contextualization has been extensively studied with respect to DSMs, well before the appearance of contextual language models. This was achieved using vector composition methods, which build representations that go beyond individual words to obtain word meanings in context (Mitchell and Lapata 2008; Erk and Padó 2008; Dinu and Lapata 2010; Thater, Fürstenau, and Pinkal 2011). Specifically, the contextualized meaning of a target word w_t in context c was obtained by creating a vector that combined the vectors of w_t and of the words $\{w_1, \ldots, w_n\}$ in c, using some operation such as component-wise multiplication or addition. Some models also use latent semantic dimensions. The model of Dinu and Lapata (2010), for example, represents word meaning as a probability distribution over a set of latent senses reflecting the out-of-context likelihood of each sense. The contextualized meaning of a word is then modeled as a change in the original sense distribution.[11] The model of Van de Cruys, Poibeau, and Korhonen (2011) exploits the latent space to determine the features that are important for a particular context and adapts the out-of-context (dependency-based) feature vector of the target word accordingly, allowing for a more precise and more distinct computation of word meaning in context. Thater, Fürstenau, and Pinkal (2011), on the contrary, use no explicit sense representation but rather derive a contextualized vector from the basic meaning vector of a target word by reweighting its components on the basis of the context of occurrence.[12] They observe that retaining only the dimensions that correspond to the word's syntactic neighbors results in an extremely sparse vector (with zero values for most of its dimensions). They thus propose to leverage semantic similarity information about the context words and to also retain dimensions that are distributionally similar to them, weighted by their similarity score.

[11] The latent senses are induced using non-negative matrix factorization (NMF) (Lee and Seung 2000) and Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003).

[12] The dimensions of the basic and contextualized vectors represent co-occurring words in specific syntactic relations.

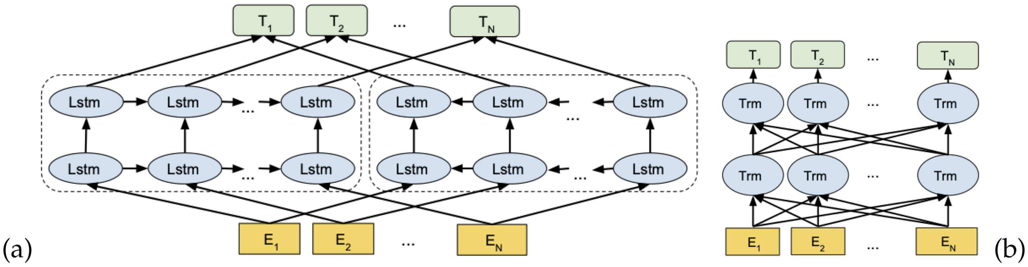


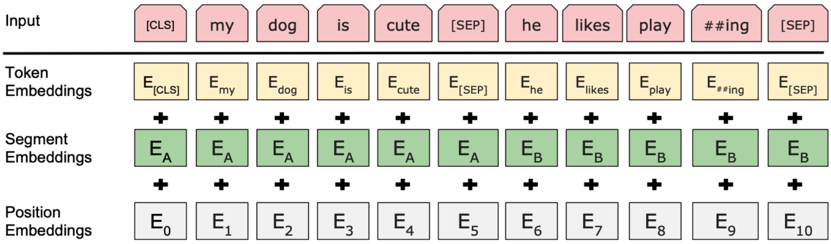
Figure 5
The figures illustrate (a) the architecture of the ELMo language model and (b) that of the Transformer-based BERT model (Devlin et al. 2019).

2.3.1 ELMo (Embeddings from Language Models). ELMo (Peters et al. 2018) relies on a bidirectional LSTM (biLSTM) (Hochreiter and Schmidhuber 1997; Graves and Schmidhuber 2005) trained on a large corpus with a language modeling objective. The original model consists of three layers: A character *n*-gram convolutional layer which is followed by a two-layer bidirectional LSTM, as shown in Figure 5 (a). ELMo representations are a linear combination of the internal layers of the model. The representation that is generated for each token is a combination of the hidden states of the two BiLSTM layers which encode the context-sensitive representation of the word, and the static representation of the word which is character-based. When ELMo is integrated into task-specific architectures, the task and the linear combination of different layers are simultaneously learned in a supervised way.

2.3.2 BERT (Bidirectional Encoder Representations from Transformers). BERT (Devlin et al. 2019) is a very widely used contextual language model. It relies on the Transformer architecture (Vaswani et al. 2017) which was initially developed for sequence-to-sequence (seq2seq) tasks such as machine translation. The goal was to simplify the Recurrent Neural Network (RNN) and Convolutional Neural Network (CNN) architectures previously used. These encoder-decoder models typically used an attention mechanism. The Transformer removed recurrence and convolutions, and relied entirely on the "self-attention" mechanism. This fully attention-based approach, where the representation of a sequence is computed by relating different words (positions) in the same sequence, shows improved performance compared to previous architectures in numerous NLP tasks. Additionally, attention is a useful interpretation tool which shows how the model assigns weight to different input elements when performing specific tasks (Raganato and Tiedemann 2018; Voita, Sennrich, and Titov 2019; Kovaleva et al. 2019; Rogers, Kovaleva, and Rumshisky 2020).

Contrary to ELMo, where a forward and a backward language model are separately trained (cf. Figure 5 (a)), BERT relies on a bidirectional model which jointly conditions on the left and right context in all layers (cf. Figure 5 (b)). BERT is pre-trained using two objectives, Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). MLM is similar to a Cloze task (Taylor 1953). In MLM, a portion of the input tokens is masked at random (e.g., *The cat* [MASK] *on the mat*) and the model has to predict them based on the context.[13] In NSP, the model needs to predict whether two segments

[13] The portion of words to mask is a parameter that needs to be set for model training. In BERT, it is 15% of the token positions.



Input	[CLS]	my	dog	is	cute	[SEP]	he	likes	play	##ing	[SEP]
Token Embeddings	$E_{[CLS]}$	E_{my}	E_{dog}	E_{is}	E_{cute}	$E_{[SEP]}$	E_{he}	E_{likes}	E_{play}	$E_{\#\#ing}$	$E_{[SEP]}$
	+	+	+	+	+	+	+	+	+	+	+
Segment Embeddings	E_A	E_A	E_A	E_A	E_A	E_A	E_B	E_B	E_B	E_B	E_B
	+	+	+	+	+	+	+	+	+	+	+
Position Embeddings	E_0	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9	E_{10}

Figure 6
BERT input representation for the sequence "*My dog is cute. He likes playing*". The input embeddings are the sum of the token, segment, and position embeddings.

follow each other in the original text. The goal of this objective is to improve performance in downstream tasks that require reasoning about the relationships between pairs of sentences (e.g., NLI or Question Answering). Sentence pairs are grouped into a single sequence and separated with a special token (`[SEP]`).

BERT receives as input a combination of token embeddings, position embeddings, and segment embeddings, as shown in Figure 6. The position embedding shows where the token occurs in the input string, and the segment embedding indicates whether it occurs in the first or the second sentence (A or B). These three vectors are added element-wise to deliver the representation of the word which will be passed through the Transformer layers. Each word in the sentence influences every other word, and the representations are updated based on this contextual information due to the dense interconnections inside the Transformer. The first token of every sequence is a special token (`[CLS]`), the final hidden state of which is used as the aggregate sequence representation. BERT can be fine-tuned for different tasks by simply adding a classification or regression head on top of the `[CLS]` token.

Two pre-trained English BERT models are available ($\mathbf{BERT_{BASE}}$ and $\mathbf{BERT_{LARGE}}$) which were trained on the BooksCorpus (800M words) (Zhu et al. 2015) and the English Wikipedia (2,500M words).[14] BERT models are trained with a specific kind of tokenization where words are split into smaller units called WordPieces (Wu et al. 2016). For example, the word *playing* in Figure 6 is split into two pieces, *play* and *##ing*. Word vectors can be derived from these subword-level representations using different mechanisms, described in the next section.

2.3.3 Subword Pooling. Word embedding methods often operate at the subword level. The embedding for a word corresponds, in this case, to the average of its subword embeddings. This subword pooling operation is a common and necessary first step for generating a representation for a word that has been broken into smaller pieces, but is not always explicitly stated as a separate step in research papers.

BERT-like models specifically use WordPiece tokenization (Wu et al. 2016). The most frequent words in the training corpus are represented as a single token, while other less frequent words might be split into multiple wordpieces. This process yields $w^1, \ldots, w^k$ pieces for a word w, which can be concatenated in order to form the word

[14] The two models differ in terms of number of layers (L=12 vs. L=24), hidden size (H=768 vs. H=1,024), number of self-attention heads (A=12 vs. A=16), and total number of parameters (11M and 340M).

($cat(w^1, \ldots, w^k) = w$) (Bommasani, Davis, and Cardie 2020; Vulić et al. 2020b). The final representation for the word is constructed by taking the average over the subword encodings, further averaged over $n \leq N$ layers where N is the number of Transformer layers.[15] Apart from the arithmetic mean ($\texttt{mean}(\cdot)$), other mechanisms for aggregating these vectors include element-wise min or max pooling ($\texttt{min}(\cdot)$, $\texttt{max}(\cdot)$). It is also possible to use the last vector of a word ($\texttt{last}(\cdot)$), discarding the representations of earlier layers (Bommasani, Davis, and Cardie 2020).

Subword units were used in earlier models as well, since they provide the flexibility needed to account for rare, unknown or out-of-vocabulary (OOV) words (i.e., not seen in the training data). This subword information allows the models to improve the representation of morphologically complex words (formed via compounding, affixation, or inflection) and to capture the explicit relationship among morphological variants (Luong and Manning 2016). This is especially important in the case of morphologically rich languages where a word (verb or noun) might have a high number of inflected forms or cases, the majority of which occur rarely in the corpora used for model training (Bojanowski et al. 2017). In the context of NMT, Sennrich, Haddow, and Birch (2016) proposed to encode unknown words as sequences of subword units in order to enable open-vocabulary translation. The idea was that the morphemes of morphologically complex words can be translated separately, and that character-level translation rules can be used for cognates and loanwords with a common origin (Tiedemann 2012). The segmentation into subword units allows the model to take into account morphology when learning word representations, and to learn translations that it can generalize to unseen words. Their segmentation techniques included simple character n-gram models and a variant of the byte pair encoding (BPE) compression algorithm, which merges frequent character n-grams into a single symbol (Gage 1994).

Character-level embedding models like fastText (Bojanowski et al. 2017) learn representations directly from characters, which also allows to form robust representations for OOV tokens. Similarly, the CHARAGRAM model (Wieting et al. 2016) embeds a character sequence (word or sentence) by adding the vectors of its character n-grams. ELMo representations are also character-based (Peters et al. 2018; Jozefowicz et al. 2016). The model uses a character CNN. The produced contextualized representations are a function of the internal states of a deep bidirectional language model (biLM). In the model of Kim et al. (2016), C is the vocabulary of characters, d is the dimensionality of character embeddings, and $\mathbf{Q} \in \mathbb{R}^{d \times |C|}$ is the matrix of character embeddings. If a word $k \in V$ of length l is made up of a sequence of characters $[c_1, \ldots, c_l]$, then the character-level representation of k is given by the matrix $\mathbf{C}^k \in \mathbb{R}^{d \times l}$, where the j-th column corresponds to the character embedding for c_j (i.e., the c_j-th column of $\mathbf{Q}$). Words are then represented as the sum of their character n-gram vectors followed by an elementwise nonlinearity.[16] A character-based variant of BERT has also been proposed as an alternative to re-training BERT for specialized domains, where the general-domain wordpiece vocabulary might not be optimal (El Boukkouri et al. 2020). CharacterBERT uses a Character-CNN module (Peters et al. 2018) to produce a single embedding representation for a word, which is then added to position and segment embeddings. Pre-training is carried out as in BERT

[15] L_0 is the embedding layer, L_1 is the bottom layer, and L_N is the final (top) layer. Vulić et al. (2020b) showed that excluding higher layers from the average may result in stronger vectors in different languages, since lexical information is predominantly concentrated in lower Transformer layers.

[16] In their implementation, they append start-of-word and end-of-word characters to each word in order to better represent prefixes and suffices, hence $\mathbf{C}^k$ has $l + 2$ columns. For batch processing, $\mathbf{C}^k$ is zero-padded so that the number of columns is constant for all words in V (i.e., equal to the max word length).

but in MLM, the model predicts entire words instead of wordpieces. Each input token is assigned a single final contextual representation by the model.

Multilingual Models. BERT-type models trained on monolingual text exist in several languages (e.g., Martin et al. 2020; Le et al. 2020; Cañete et al. 2020; Koutsikakis et al. 2020; Virtanen et al. 2019). The multilingual BERT (mBERT) model was (pre-)trained on the 104 languages with the largest Wikipedias.[17] mBERT uses a 110k shared WordPiece vocabulary which is mostly English-driven, often resulting in arbitrary partitionings in other languages. This suboptimal tokenization has a negative impact on the quality of the lexical knowledge that is encoded in the representations (Garí Soler and Apidianaki 2021a,b). Language-specific monolingual models generally perform better and contain more linguistic information for a particular language than their multilingual counterparts (Vulić et al. 2020b). This is due to the trade-off that is observed when the number of languages scales up but model capacity remains fixed, also described as the "curse of multilinguality". As noted by Conneau et al. (2020a), encompassing more languages leads to better cross-lingual performance in low-resource languages up to some point, after which the overall performance on both monolingual and cross-lingual benchmarks degrades. In other words, the models tend to sacrifice monolingual information coverage for a wider language coverage. Still, very large multilingual models (e.g., XLMR-L) perform on par with language-specific BERT models in some tasks such as multilingual WSD (Pasini, Raganato, and Navigli 2021), mainly because of the difference in model size.[18]

2.3.4 Other Transformer-based Models. Lighter BERT-inspired models also exist. DistilBert (Sanh et al. 2019) and ALBERT (A Lite BERT) (Lan et al. 2020) have significantly fewer parameters than BERT, and still yield high performance in Natural Language Understanding (NLU) tasks. RoBERTa (Liu et al. 2019) is trained longer and with larger batches than BERT, over more data and on longer sequences. The NSP objective is removed, and a dynamic masking pattern is applied to the training data. SpanBERT (Joshi et al. 2020) masks random contiguous spans of variable length instead of individual tokens. It replaces BERT's MLM objective by a span-boundary objective, where the model learns to predict the entire masked span from the observed tokens at its boundary. Also, SpanBERT is pre-trained on single segments, allowing the model to learn longer-range features. AMBERT (Zhang, Li, and Li 2021) adopts a multi-grained tokenization approach and generates representations for words, sub-word pieces, and phrases. Fine- and coarse-grained representations are learned in parallel using two encoders with shared parameters and MLM. The model is fine-tuned for classification using the `[CLS]` representations created by both encoders. Fine-tuning is defined as optimization of a regularized loss of multi-task learning.

Other high performing Transformer-based models are the OpenAI GPT-2 and GPT3 models (Radford et al. 2019) which deliver high performance on several benchmarks in a zero-shot setting. Finally, the ELECTRA model (Clark et al. 2020) is trained using the "replaced token detection" procedure, which corrupts the input by replacing some tokens with plausible alternatives sampled from a small generator network. A discriminative model is then trained that predicts whether a token in the corrupted

[17] The languages with the largest Wikipedias were under-sampled, and the ones with lower resources were over-sampled.

[18] The XLMR-Large model (Conneau et al. 2020a) has roughly 200M more parameters than most of the language-specific models applied to this task.

input was replaced by a generator sample or not, instead of predicting masked tokens as in MLM.

2.3.5 Evaluation of Contextualized Representations. New datasets aimed at evaluating the capability of contextual models to capture in-context similarity and the meaning of individual word instances have been created. Such datasets existed since the era of DSMs but their coverage was limited. The Usage Similarity (Usim) dataset (Erk, McCarthy, and Gaylord 2009, 2013), for example, contains ten instances of 56 target words manually annotated with graded pairwise usage similarity judgments on a scale from 1 to 5 (from less to more similar). The Stanford Contextual Word Similarity (SCWS) dataset (Huang et al. 2012) includes pairs of sentences that contain instances of different target words, or of homographs with different part of speech (e.g., *pack* as noun and verb).[19] The Concepts in Context (CoInCo) corpus (Kremer et al. 2014) contains substitute annotations for all content words in a sentence. The similarity of word instances is modeled through the overlap of their substitutes, similar to the SemEval-2007 lexical substitution dataset (McCarthy and Navigli 2007).[20] Datasets with automatically assigned substitute annotations have also been created. The ukWaC-subs dataset (Garí Soler and Apidianaki 2020b), for example, contains sentences automatically annotated with lexical substitutes from the Paraphrase Database (PPDB) (Ganitkevitch, Van Durme, and Callison-Burch 2013; Pavlick et al. 2015) using the context2vec model (Melamud, Goldberger, and Dagan 2016).

Word-in-Context (WiC) (Pilehvar and Camacho-Collados 2019) is another automatically created dataset which contains binary similarity judgments for pairs of word instances in context.[21] Sentences are labeled as true (T) or false (F) based on whether they are listed under the same sense in WordNet. Automatic pruning was applied in order to remove related instances with subtle sense distinctions, avoid replicating the fine sense granularity of WordNet and reduce errors, but some noisy annotations still remain (Garí Soler, Apidianaki, and Allauzen 2019). An analysis of the target word and context biases in this and other in-context similarity datasets has been performed by Liu, McCarthy, and Korhonen (2022). A smaller and more focused dataset for studying regular (or systematic) polysemes (Apresjan 1974), the Contextualized Polyseme Word Sense Dataset, has been proposed by Haber and Poesio (2020, 2021). The dataset covers ten types of regular metonymic polyseme alternations (e.g., ANIMAL/MEAT LAMB: chicken, pheasant; FOOD/EVENT: lunch, dinner).[22] It contains three measures of word sense similarity including graded similarity judgments for the word instance pairs used, co-predication acceptability judgments,[23] and categorical word class judgments.

Benchmarks addressing in-context similarity in a multilingual setting also exist. The XL-WiC dataset addresses twelve languages (Raganato et al. 2020).[24] It was

[19] SCWS contains 1,328 noun-noun, 399 verb-verb, 140 verb-noun, 97 adjective-adjective, 30 noun-adjective, and 9 verb-adjective pairs.

[20] The SemEval dataset contains 10 sentences for each of 201 target words. CoInCo covers around 35K tokens of running text from two domains of the MASC corpus (newswire and fiction) (Ide et al. 2008, 2010) where all 15.5K content words are labeled with in-context synonyms.

[21] Sentences come from WordNet (Fellbaum 1998) (23,949 examples), VerbNet (Kipper Schuler 2006) (636 examples), and Wiktionary (10,564 examples).

[22] It addresses 10 systematic polysemes (e.g., *newspaper*, *school*, *chicken*), 15 homonyms, and 15 synonyms.

[23] As an example of co-predication, consider the sentence "The *newspaper* wasn't very interesting and got wet from the rain" decomposed into "The *newspaper* wasn't very interesting" and "The *newspaper* got wet from the rain".

[24] Bulgarian, Chinese, Croatian, Danish, Dutch, Estonian, Farsi, French, German, Italian, Japanese, and Korean.

automatically created by leveraging information from Multilingual WordNet (Bond and Paik 2012) and Wiktionary. In XL-WiC, positive (True) examples correspond to the same sense in the underlying resource, and negative (False) examples correspond to different senses. The MCL-WiC dataset (Martelli et al. 2021) was manually annotated using lexemes from the multilingual BabelNet network (Navigli and Ponzetto 2010) and addresses five languages.[25] MCL-WiC enables a multilingual and a cross-lingual evaluation scenario. The cross-lingual AM2iCo dataset addresses 14 language pairs where English is paired with a target language (Liu et al. 2021).[26] AM2iCo was created using Wikipedia's cross-lingual links which served to identify cross-lingual concept correspondences. A sample of examples for each language pair was then validated through crowdsourcing.[27]

2.4 Conclusion

The representations that have been presented in this section describe the meaning of words at the level of word types, senses, and individual instances. The advantages of each representation method have been discussed as well as their shortcomings, which often incite the development of new approaches. The majority of the presented methods derive meaning representations from text data in an unsupervised or self-supervised way. This data-driven knowledge is refined or occasionally combined with sense information from external lexicons.

Another type of methods aims at improving the information that is learned from corpora by injecting different types of external knowledge in the representations during pre-training or fine-tuning. These semantic specialization methods integrate external knowledge in the form of linguistic constraints, and improve the quality of word representations to better reflect word meaning compared to vanilla word vectors. This makes them highly relevant for this survey. We devote the next section to these knowledge integration methods.

3. Semantic Knowledge Injection into Word Embeddings

3.1 Motivation

Semantic specialization methods infuse knowledge about different types of lexical relationships into word embeddings. The motivation behind these methods is that the rich information that is present in knowledge graphs and other handcrafted resources can complement the incomplete, and sometimes ambiguous, information that is extracted from texts (Xu et al. 2014). The linguistic and factual information used is often difficult to capture with conventional distributional training. Semantic specialization methods can also serve to adapt generic embedding representations to a specific task, by feeding into them information from resources constructed for that task (Yu and Dredze 2014). We can distinguish semantic specialization methods across three axes:

(i) the type of embeddings they modify: static or contextualized;

[25] Arabic, Chinese, English, French, and Russian.

[26] German, Russian, Japanese, Chinese, Arabic, Korean, Finnish, Turkish, Indonesian, Basque, Georgian, Bengali, Kazakh, and Urdu.

[27] The annotators were native speakers of the target language and fluent in English.

(ii) the stage in which they intervene: during training or post-processing;

(iii) whether they address knowledge about individual words ("intra-word" approaches) or about the relationships between words ("inter-word" approaches);

(iv) the type of knowledge (e.g., linguistic, factual) used for vector enrichment.

We choose to organize the presentation of these methods according to the first dimension (i.e., the type of embeddings being modified) for the sake of coherence with the previous section, where we showed the evolution from static to contextualized embeddings. Inside each subsection, we will address the other three dimensions that we consider equally important. Regarding the types of knowledge used, the focus will be put on methods that address knowledge aimed at improving the representation of word meaning and lexical relationships. We will, however, also provide pointers to studies addressing factual knowledge for the interested reader.

3.2 Knowledge Injection into Static Embeddings

3.2.1 Joint Optimization Models. Joint models integrate external linguistic constraints into the training procedure of word embedding algorithms. The RC-NET framework of Xu et al. (2014) leverages relational and categorical knowledge extracted from the Freebase graph (Bollacker et al. 2008). This knowledge is transformed into two separate regularization functions that are combined with the original objective function of Skip-gram (Mikolov et al. 2013b). The combined optimization problem is solved using back propagation neural networks, and the generated word representations encode the knowledge found in the graph. The Relation Constrained Model (RCM) of Yu and Dredze (2014) uses a learning objective that incorporates the CBOW objective and linguistic constraints from WordNet and the Paraphrase Database. The combined objective maximizes the language model probability and encourages embeddings to encode semantic relations present in the resources.

Kiela, Hill, and Clark (2015) propose a joint learning approach where they supplement the Skip-gram objective with additional contexts (synonyms and free-associates) from external resources in order to direct embeddings towards similarity and relatedness.[28] Given a sequence of words $\{w_1, w_2, w_3, \ldots, w_n\}$ and c the size of the training context, the Skip-gram objective is to generate representations that are useful for predicting the surrounding words in a sentence by maximizing the average log probability:

$$\frac{1}{T}\sum_{t=1}^{T} J_\theta(w_t) = \frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0} \log p(w_{t+j}|w_t) \tag{1}$$

In Equation (1), T is the number of training examples and θ is the word weight (embedding) matrix that needs to be optimized by maximizing the probability of predicting context words given a center word (w_t). The weight matrix is passed into the cost

[28] They use English synonyms from the MyThes thesaurus developed by the OpenOffice.org project and the University of South Florida (USF) free association norms (Nelson, Mcevoy, and Schreiber 2004).

function J as a variable and optimized (J_θ). The basic Skip-gram formulation defines $p(w_{t+j}|w_t)$ using the softmax function

$$p(w_{t+j}|w_t) = \frac{exp^{u_{w_{t+j}}^\top v_{w_t}}}{\sum_{w'=1}^{V} exp^{u_{w'}^\top v_{w_t}}} \tag{2}$$

where u_w and v_w are the context and target vector representations for word w, respectively, and w' ranges over the full vocabulary V. Additional contexts from the external resources are used to supplement this objective:

$$\frac{1}{T}\sum_{t=1}^{T}\left(J_\theta(w_t) + \sum_{w^a \in A_{w_t}} \log p(w^a|w_t)\right) \tag{3}$$

The set of additional contexts A_{w_t} contains the relevant contexts for a word w_t, i.e., its synonyms in the external resource.[29]

Some specialization methods address both synonymy and antonymy. This distinction is hard to make for word embedding models that have been trained using co-occurrence information, given that antonyms (e.g., *east/west*, *happy/sad*, *interesting/boring*) appear in near-identical contexts and get highly similar distributional vectors. Pham, Lazaridou, and Baroni (2015) propose an extension of the Skip-gram method in their multi-task Lexical Contrast Model which optimizes embedding vectors on the joint tasks of predicting corpus contexts and making the representations of WordNet synonyms closer than that of WordNet antonyms. Similarly, Ono, Miwa, and Sasaki (2015) combine information from WordNet and the Roget's thesaurus (Kipfer 2009) with distributional data to train embeddings specialized for capturing antonymy, by modifying the objective function of Skip-gram. Nguyen, Schulte im Walde, and Vu (2016) also integrate lexical contrast information in the objective function of the Skip-gram model in order to capture antonymy. Contrary to the above described methods which apply lexical contrast information from WordNet to each of the target words, the model of Nguyen, Schulte im Walde, and Vu (2016) applies lexical contrast to every single context of a target word in order to better capture this semantic relationship. The method of Schwartz, Reichart, and Rappoport (2015) uses automatically discovered symmetric patterns indicative of antonymy such as "from X to Y" and "either X or Y" (e.g., *"from bottom to top"*, *"either high or low"*) to assign dissimilar vector representations to antonyms.

Embeddings augmented by joint models generally perform better than vanilla Skip-gram embeddings on intrinsic semantic tasks such as word similarity and analogy, and in downstream settings. A limitation of joint models is that they are tied to the distributional objective of the embedding model. As a result, any change to the underlying distributional model induces a change to the joint model too.

3.2.2 Retrofitting Methods. Retrofitting methods tune word vector spaces post-hoc according to external linguistic constraints (Faruqui et al. 2015; Mrkšić et al. 2016; Vulić and Mrkšić 2018; Vulić 2018). An advantage of retrofitting methods compared to joint

[29] They implement a "sampling" and an "all" condition. In the former, the objective is modified to include an additional context w^a sampled uniformly from the set of additional contexts A_{wt}. In the latter, all additional contexts for a target word are added at each occurrence.

models—which inject information during training—is that they are more versatile and applicable to any input space (Glavaš and Vulić 2018). Some approaches, however, combine joint learning and retrofitting techniques in order to specialize embeddings for semantic constraints (Kiela, Hill, and Clark 2015).

Early retrofitting methods (Faruqui et al. 2015; Jauhar, Dyer, and Hovy 2015; Wieting et al. 2015; Rothe and Schütze 2017) used synonymy constraints to bring the vectors of semantically similar words closer in the vector space. Specifically, the methods encouraged words that were linked in a semantic resource (e.g., paraphrases, synonyms, or hypernyms) to have similar vector representations. Let $V = \{w_1, w_2, \ldots, w_N\}$ be the vocabulary or set of word types, $\{\vec{w_1}, \vec{w_2}, \ldots, \vec{w_N}\}$ the corresponding vectors in the original vector space, and D a resource (lexicon, ontology, or dictionary) which encodes semantic relationships between words in V. The goal of retrofitting methods is to produce new word vectors $\{\vec{w'_1}, \vec{w'_2}, \ldots, \vec{w'_N}\}$ that observe the constraints present in D. These are often sets of synonymous words (w_i, w_j), or they can describe other relationships such as antonymy and entailment. In the case of synonymous word pairs, the goal is to bring the vectors of these words closer together in the vector space. A way to achieve this is to reduce their cosine distance $(d(\vec{w_i}, \vec{w_j}) = 1 - cos(\vec{w_i}, \vec{w_j}))$ (Mrkšić et al. 2016).

For methods that leverage antonymy relations, the goal is to push the vectors of antonymous words away from each other in the transformed vector space V' (i.e., increase their cosine distance), similar to the objective in the Lexical Contrast model of Pham, Lazaridou, and Baroni (2015). Using both similarity and dissimilarity constraints further improves performance, compared to methods that only inject similarity constraints in the vector space built by language models. The counter-fitting approach of Mrkšić et al. (2016) and the ATTRACT-REPEL algorithm (Mrkšić et al. 2017) use both synonymy and antonymy constraints. ATTRACT-REPEL makes it possible to embed vector spaces of multiple languages into a single vector space using constraints from the multilingual BabelNet network (Navigli and Ponzetto 2010). This allows to exploit information from high-resource languages in order to improve the word representations in lower-resource ones.

The above described retrofitting and counter-fitting methods address symmetric relationships (synonymy and antonymy). The LEAR (Lexical Entailment Attract-Repel) and GLEN (Generalized Lexical ENtailment) models (Vulić and Mrkšić 2018; Glavaš and Vulić 2019) address the asymmetric lexical entailment relationship. This relationship was considered by joint models that used hypernymy constraints for learning hierarchical embeddings (Yu et al. 2015; Nguyen et al. 2017), instead of injecting knowledge to modify an existing vector space. Contrary to similarity-focused specialization models that tune only the direction of distributional vectors (Mrkšić et al. 2017; Glavaš and Vulić 2018; Ponti et al. 2018), a specialization procedure that addresses hierarchical relationships between concepts needs to also rescale the vector norms. Two words can be differentiated by means of their Euclidean norms so that the norm of the hypernym is larger than that of the hyponym (Nguyen et al. 2017). By injecting external linguistic constraints ('IS-A' WordNet links) into the original vector space, LEAR brings true hyponymy-hypernymy pairs closer together in the transformed Euclidean space, and adjusts their norms to reflect the actual hierarchy of concepts in WordNet. At the same time, a joint objective enforces semantic similarity using the symmetric cosine distance, yielding a vector space specialized for both lexical relations at once.

A limitation of early retrofitting methods is that they specialize only the vectors of words seen in the constraints, leaving unchanged the vectors of unobserved words. More recent methods address this issue by specializing the full vocabulary of the original embedding space. The adversarial post-specialization method of Ponti et al.

(2018), for example, propagates the external lexical knowledge to the full distributional space. Words seen in the resources serve as training examples for learning a global specialization function, by combining a distance loss with an adversarial loss. The Explicit Retrofitting (ER) models of Glavaš and Vulić (2018, 2019) also learn a global specialization function which can specialize the vectors of words unseen in the training data. The model proposed in the earlier study accounts for symmetric relations (synonymy and antonymy), while the more recent one (GLEN) addresses asymmetric relations (entailment). An important advantage of these models is that they can be extended to specialize vector spaces of languages unseen in the training data, by coupling ER with shared multilingual embedding spaces. Word vectors of a new language can be projected to the space of the language for which the specialization function has been learned using some cross-lingual projection method (e.g., Artetxe, Labaka, and Agirre 2018; Lample et al. 2018).

3.3 Knowledge Injection into Contextualized Embeddings

Contextual language models (e.g., ELMo, BERT, GPT-2, GPT-3) are trained on large amounts of raw text using self-supervision. Consequently, although more powerful than static embeddings, they also encode only the distributional knowledge that is available in the corpora used for training. Without explicit grounding to real-world entities, they have difficulty recovering factual knowledge (Peters et al. 2019; Logan et al. 2019). Numerous knowledge injection methods aimed at augmenting the knowledge that is encoded in contextualized vectors have been proposed.

3.3.1 Knowledge Injection during Pre-training. Similar to joint optimization methods for static embeddings, knowledge about semantic relationships can be integrated during contextual language models' pre-training. This can be done by adding an auxiliary word relationship classification task, as in the Lexically-Informed BERT (LIBERT) model of Lauscher et al. (2020). In LIBERT, knowledge about semantic similarity (i.e., synonymy and hypernymy) is infused into BERT vectors in a multi-task learning setting, where the MLM and NSP objectives are coupled with an auxiliary binary word relation classification task.

The SenseBERT model (Levine et al. 2020) injects information about senses into contextualized representations using an auxiliary masked word sense prediction task, alongside BERT's usual training tasks (MLM and NSP). The language model that predicts the missing words' sense is trained jointly with the standard word form-level language model, without need for sense-annotated data. Specifically, information from WordNet serves as weak supervision for self-supervised learning; the masked word's supersenses form a set of possible labels for the sense prediction task.[30]

The joint learning of words and knowledge from pre-crafted resources has also been attempted in other studies (Wang et al. 2014; Toutanova et al. 2015; Han, Liu, and Sun 2016; Cao et al. 2017; Zhang et al. 2019). These mainly address entities and relationships found in large knowledge graphs such as the Freebase (Bollacker et al. 2008), or in in-domain specific resources (Liu et al. 2020). The proposed models generally embed

[30] When a single supersense is available, the network learns to predict this supersense given the masked word's context. When multiple supersenses are available (e.g., *bass*: noun.food, noun.animal, noun.artifact, noun.person), the model is trained to predict any of these, leading to a simple soft-labeling scheme.

words, entities, and their relationships in the same continuous latent space. Operating simultaneously on relations observed in text and in pre-existing structured databases permits to reason about structured and unstructured data in mutually-supporting ways (Riedel et al. 2013). Such knowledge-enhanced language models are created by integrating fixed entity embeddings into pre-trained language representation models, or by jointly optimizing a knowledge embedding objective and a language modeling objective during pre-training.

In KnowBert (Peters et al. 2019), an integrated entity linker is used to retrieve relevant entity embeddings, and contextual word representations are updated via a form of word-to-entity attention. In this model, the entity linker and self-supervised language modeling objective are jointly trained end-to-end in a multi-task setting. The KEPLER (**K**nowledge **E**mbedding and **P**re-trained **L**anguag**E** **R**epresentation) model of Wang et al. (2021) encodes entity descriptions as entity embeddings, and jointly optimizes a knowledge embedding objective and a masked language modeling objective during pre-training by means of a multi-task loss. ERNIE (**E**nhanced Language **R**epresentatio**N** with **I**nformative **E**ntities) (Zhang et al. 2019) uses a pre-training task where token-entity alignments are masked and the system is required to predict both. The reported experimental results show that jointly embedding this information improves performance in downstream tasks involving fact and entity relation prediction. Yamada et al. (2020) propose the **LUKE** (**L**anguage **U**nderstanding with **K**nowledge-based **E**mbeddings) model, which is trained on a large entity-annotated corpus retrieved from Wikipedia, and integrates an "entity-aware" self-attention mechanism which considers the types of tokens (words or entities) when computing attention scores. LUKE achieves impressive empirical performance on a wide range of entity-related tasks.

The above-mentioned approaches use hand-crafted knowledge resources. The "Symbolic knowledge distillation" paradigm is a conceptual framework where large general language models (e.g., GPT-3) author commonsense knowledge graphs to train commonsense models (Hwang et al. 2021; West et al. 2022). The approach is motivated by knowledge distillation (Hinton, Vinyals, and Dean 2014) where a larger teacher model transfers knowledge to a compact student model.

These knowledge enrichment methods mainly address factual knowledge about entities and relations, not lexical semantic relationships (such as synonymy, hyponymy, or entailment). However, our proposed overview addresses approaches that enrich embeddings with lexico-semantic knowledge, with the aim to improve the representations of words and their relationships. This section will thus not go deeper into this analysis. The interested reader may consult the references given above for a more thorough presentation of the proposed methods.

3.3.2 Knowledge Injection through Fine-tuning. These methods retrofit external semantic knowledge into contextualized embeddings through a process similar to fine-tuning. Shi et al. (2019) tune ELMo embeddings (Peters et al. 2018) using the Microsoft Research Paraphrase Corpus (MRPC) (Dolan, Quirk, and Brockett 2004). They propose an orthogonal transformation for ELMo that is trained to bring representations of word instances closer when they appear in meaning-equivalent contexts (paraphrases). The idea is the same as the one underlying retrofitting methods. Arase and Tsujii (2019) use paraphrase data for fine-tuning BERT, and then fine-tune the model again for the tasks of interest (paraphrase identification and semantic equivalence assessment). The model that is first exposed to paraphrase data is better at performing these tasks, compared to a model directly fine-tuned on task data. Garí Soler and Apidianaki (2020b) also show that BERT fine-tuned on usage similarity and paraphrasing datasets (Erk, McCarthy, and Gaylord 2013; Kremer et al. 2014; Creutz 2018) performs better on the Graded Word Similarity in Context (GWSC) task (Armendariz et al. 2020).

The LexFit model of Vulić et al. (2021) relies on dual-encoder network structures in order to extract the lexical knowledge that is stored in pre-trained encoders, and turn language models into static decontextualized word encoders. The procedure involves fine-tuning pre-trained language models on lexical pairs from an external resource. These comprise positive pairs where a particular relationship holds between two words, and negative pairs, where the first word is paired with a random word with which it is not related.

3.4 Conclusion

The knowledge enrichment methods described in this section aim at improving the quality of the lexical semantic knowledge that is encoded in language model representations. All studies report improved results in intrinsic and extrinsic evaluations compared to vanilla models with no access to external knowledge. Naturally, these methods rely on the assumption that some lexical semantic knowledge is already encoded in language model representations and can be refined. The next section presents methods that attempt to decipher the encoded knowledge which is generally acquired through exposure to large volumes of data during language model pre-training.

4. Interpretation Methods for Lexical Semantics

4.1 Motivation

Interpretation studies demonstrate that language model representations encode rich knowledge about language and the world (Voita, Sennrich, and Titov 2019; Clark et al. 2019; Voita et al. 2019; Tenney, Das, and Pavlick 2019; Linzen 2019; Rogers, Kovaleva, and Rumshisky 2020). The intense exploration of language models' inherent knowledge has been motivated by their impressive performance in NLU tasks. Interpretation studies attempt to understand what this high performance is due to, by deciphering the knowledge that is encoded inside the representations. The bulk of this interpretation work relies on probing tasks which serve to predict linguistic properties from the representations that are generated by vanilla models, before integration of any external knowledge. Success in these tasks indicates that the model's representations encode the addressed linguistic knowledge.

Early probing studies explored surface linguistic phenomena pertaining to grammar and syntax, which are directly accessible in contextualized (token-level) representations (Linzen, Dupoux, and Goldberg 2016; Hewitt and Manning 2019; Hewitt and Liang 2019). The first studies addressing semantic knowledge explored phenomena in the syntax-semantics interface such as semantic role labeling and coreference (Tenney, Das, and Pavlick 2019; Kovaleva et al. 2019), and the symbolic reasoning potential of LM representations (Talmor et al. 2020). Lexical polysemy is more challenging to study using token-level representations since it is encoded at a higher level of abstraction than individual instances, that of word types. Representations extracted from pools of sentences allow to abstract away from individual context variation. They are more informative about words' semantic properties (Vulić et al. 2020b), and can serve to model abstract semantic notions (e.g., intensity) (Garí Soler and Apidianaki 2020a, 2021b). Semantic relationships like hypernymy and entailment are also usually encoded at the word type-level (e.g., *cat* $\models$ *animal*, *tulip* $\models$ *flower*), although they are context-dependent in the case of polysemous words.[31]

We hereby present methods that have been proposed for exploring the semantic information that is encoded in contextualized representations. These include visualization and probing tasks, as well as studies that investigate the semantic properties of the words in the constructed space by relying on its geometry.

4.2 Visualization and WSD

The capability of language models to represent polysemy was initially studied by visualizing sense distinctions found in lexicons. Reif et al. (2019) and Wiedemann et al. (2019) generate BERT representations from Wikipedia sentences and the SemCor corpus (Miller et al. 1993). They show that the representations of polysemous words' usages are organized in the semantic space in a way that reflects the meaning distinctions present in the data. They also demonstrate BERT's WSD capabilities when leveraging sense-related information from these resources. Similarly, in a more recent study, Loureiro et al. (2021) show that BERT representations are precise enough to allow for effective disambiguation. Illustrating the semantic distinctions made by the BERT model in a visualization experiment, they show that it is able to group instances of polysemous words according to sense. The dataset used in these experiments is CoarseWSD-20, a benchmark that targets the ambiguity of 20 nouns and illustrates easily interpretable sense distinctions.

All these studies rely on sense annotated data and do not directly address the semantic knowledge that is encoded in contextualized representations. This type of investigation can be done using probing.

4.3 Prompting Methods

Prompting methods are widely used for few-shot and zero-shot learning, but have also been quite popular in interpretability studies that explore the linguistic and common sense knowledge which is encoded in pre-trained LMs. This framework does not involve any model training or fine-tuning. Instead, prompting methods use a template that contains some unfilled slots, which serves to modify the original input x into a textual string prompt x' (Liu et al. 2022). In a sentiment analysis task, for example, the input *"I love this movie"* may be transformed into *"I love this movie. Overall, it was a* [Z] *movie"* using the template "[X] Overall, it was a [Z] movie". The language model is then used to fill the slot [Z] in prompt x' with the highest scoring answer $\hat{z}$. This answer, or the obtained final string ($\hat{x}$), can then be mapped to a class (e.g., *positive* or *negative*) in a classification setting.

A correct filler selected by the model can also be seen as an indication of the existence of a specific type of knowledge in the pre-trained LM representations. In the above example, it would be that the model knows *"love"* is a verb with positive sentiment. Consequently, prompting methods are also commonly used in interpretation studies (Petroni et al. 2019; Bouraoui, Camacho-Collados, and Schockaert 2020; Ravichander et al. 2020; Ettinger 2020; Apidianaki and Garí Soler 2021). The input is again transformed into a prompt that contains a slot to be filled by the model. Cloze-style prompts

[31] For example, the two instances of *bug* "There is a *bug* in my soup" and "There is a *bug* in my code" entail *"insect"* and *"error"*, respectively.

that contain blanks that need to be filled (e.g., *I love this movie, it is a* [Z] *movie.*) are a good fit for exploring masked LMs because they closely match the form of the pre-training task. Prefix prompts, which continue a string prefix (e.g., "*I love this movie. What's the sentiment of the review?* [Z]"), can also be used. Naturally, due to their form, prefix prompts are a better fit for generation tasks.

In this section, we present prompting methods that have been used for lexical semantics. For an extensive overview of work on prompt-based methods, we refer the reader to the survey of Liu et al. (2022). Query reformulation methods are also relevant from a semantics perspective. These methods modify the input (using query mining and paraphrasing) in order to produce semantically similar prompts that serve to promote lexical diversity and improve knowledge extraction from the language model (Jiang et al. 2020). These may also involve prompt ensembling methods that combine multiple prompts, as well as end-to-end re-writing models for generating rephrased queries (Haviv, Berant, and Globerson 2021). Lastly, there is evidence that prompt-based models do not understand the meaning of the prompts used (Webson and Pavlick 2022), which opens up new interesting research avenues for semantics.

4.3.1 Cloze-based Probing for Semantic Knowledge. Cloze task queries (Taylor 1953) are a good fit for BERT-type models which are trained using the MLM objective. These prompts contain a "[MASK]" token at the position that needs to be filled and serve to query the model for different types of knowledge, such as encyclopedic knowledge (e.g., "*Dante was born in* [MASK]") (Petroni et al. 2019), relational knowledge (e.g., "*Recessions are caused by* [MASK]") (Bouraoui, Camacho-Collados, and Schockaert 2020), hypernymy relationships (e.g., "*A car is a* [MASK]") (Ravichander et al. 2020), noun properties (e.g., "*Strawberries are* [MASK]") (Apidianaki and Garí Soler 2021), and others.

Cloze task probing is often criticized as an interpretation method because language models are brittle to small changes in the used prompts (Jiang et al. 2020). For example, plural queries (*Strawberries are* [MASK]) are more efficient than singular queries (*A strawberry is* [MASK]) in retrieving noun properties (Apidianaki and Garí Soler 2021). The naturalness of the query is also important. There are higher chances that the model has seen natural statements in the training data and can thus handle them better than unnatural queries (Ettinger 2020). Concerns are also expressed regarding the systematicity of the knowledge that is identified using probes. Ravichander et al. (2020) showed that affirmative factual or relationship knowledge extracted from BERT does not systematically generalize to novel items. Consequently, BERT's capabilities as discovered through probes may not correspond to some systematic general ability.

Importantly, some types of information are difficult to retrieve using cloze tasks due to the "reporting bias" phenomenon which poses challenges to knowledge extraction (Gordon and Van Durme 2013; Shwartz and Choi 2020). According to this phenomenon, the frequency with which people write about actions and properties is not necessarily a reflection of real-world frequencies, or of the degree to which a property is characteristic of a class of individuals. Hence, exceptional actions or properties (e.g., *A person was killed*) are over-represented in texts (for example, in newspaper articles) and amplified by the models that are trained on these data, at the expense of more common or trivial ones which are obvious to the participants in the communication (e.g., *A person is breathing*). As a consequence, low results in a probing experiment might suggest either that the tested model has encoded marginal knowledge about the studied phenomenon, or that it has not seen the relevant information during training. A model might encode lexical and encyclopedic knowledge which is available in Wikipedia texts used for training (e.g., *a banana is a fruit*, *Obama was President of the United States*), and miss other types of perceptual knowledge (e.g., *bananas are yellow*, *silk is soft*). Finally, another issue with semantic cloze task evaluations is that—contrary to queries addressing structural properties (e.g., number agreement or syntactic dependencies)—there might be multiple correct answers for a query. These are only partially covered by the resources used for evaluation which are often developed in (psycho-)linguistic studies following different annotation protocols (McRae et al. 2005; Devereux et al. 2014). Misra, Rayz, and Ettinger (2022) highlight the risk of taking the absence of evidence in the resource used for evaluation as evidence, and stress the need for more comprehensive evaluation datasets.[32]

4.3.2 Probing for Word Type-level Information. In the work of Aina, Gulordava, and Boleda (2019), probing serves to explore the word type information that is present in the representations of a biLSTM language model, and how this interacts with contextual (token-level) information in the hidden representations of the model. They specifically train diagnostic classifiers on the tasks of retrieving the input embedding for a word (Adi et al. 2017; Conneau et al. 2018) and a representation of its contextual meaning, as reflected in its in-context lexical substitutes. The results show that the information about the input word is not lost after contextualization.

More recent probing methods for lexical semantics rely on word type-level embeddings that are derived from contextualized representations using vector aggregation techniques (Vulić et al. 2020b; Bommasani, Davis, and Cardie 2020; Garí Soler and Apidianaki 2021a).[33] Using this type of word type-level embeddings has become standard practice in studies that address the models' knowledge about lexical meaning. Reasons for this are the strong impact of context variation on the quality of the representations and the similarity estimates that can be derived from them (Ethayarajh 2019a; Mickus et al. 2020). This situation is problematic given that vector similarity calculations are key in lexical semantic tasks. In the next paragraph, we explain that the similarity estimates which are drawn from the semantic space of contextual models are not that reliable for reasons mainly related to the geometry of the vector space.

4.4 Interpretation Methods Based on Space Geometry

4.4.1 Vector Similarity. Ethayarajh (2019a) proposes to rely on the geometry of the vector space in order to investigate the degree of contextualization in representations extracted from different layers of the BERT, ELMo, and GPT-2 models. Context-specificity is approximated through vector similarity using the self-similarity (*SelfSim*) metric, the intra-sentence similarity (*IntraSim*) metric, and the maximum explainable variance (*MMEC*) metric. We explain here the *SelfSim* metric in more detail. Let w be a word that occurs in sentences $\{s_1, \dots, s_n\}$ at indices $\{i_1, \dots, i_n\}$, such that $w = s_1[i_1] = \dots = s_n[i_n]$. Let $f_l(s, i)$ be a function that maps $s[i]$ to its representation in layer l of model f. The *SelfSim* of w in layer l is given by Equation 4

$$SelfSim_l(w) = \frac{1}{n^2 - n} \sum_j \sum_{k \neq j} cos(f_l(s_j, i_j), f_l(s_k, i_k)) \tag{4}$$

[32] According to the CSLB dataset (Devereux et al. 2014), for example, only six animals "can breathe" and "have a face", because annotators did not propose obvious features for other animals present in the resource.

[33] We present these derivation techniques in detail in Section 5.1.

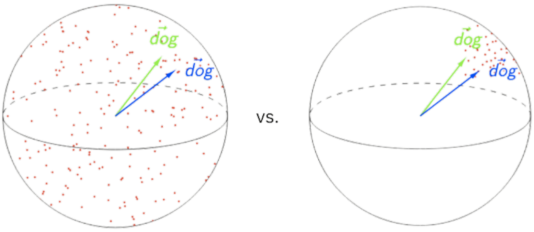


Figure 7
Comparison of an isotropic vector space (left) where embeddings are uniformly distributed in all directions, with a highly anisotropic space (right) (Ethayarajh 2019a).

where cos denotes the cosine similarity. This means that the *SelfSim* for a word w in layer l is the average cosine similarity between its contextualized representations across its n contexts. The more varied, dissimilar or contextualized the representations for w are in l, the lower its self-similarity is expected to be. If they are identical (i.e., no contextualization occurs in layer l), then $SelfSim_l(w) = 1$.

The results reported by Ethayarajh (2019a) highlight the high anisotropy of the space that is constructed by contextual language models, which has a strong negative impact on the quality of the similarity estimates that can be drawn from it. Anisotropic word representations occupy a narrow cone in the vector space instead of being uniformly distributed with respect to direction (Gao et al. 2019; Cai et al. 2021; Rajaee and Pilehvar 2021). In this anisotropic cone, even unrelated words have excessively positive correlations (Ethayarajh 2019a; Rajaee and Pilehvar 2021). This is illustrated in Figure 7. A highly isotropic space is shown on the left, where the similarity of instances of a monosemous word (the noun *dog*) is high compared to other words represented in the space. The anisotropic space on the right suggests the opposite; since all word instances are found in a narrow cone, any two word instances have high cosine similarity. As a result, similarity estimates are distorted because of the geometry of the space, so the high similarity of *dog* is not important.[34] For example, instance representations obtained with GPT-2 for randomly sampled words are as close in the space as instances of the same word.

4.4.2 Measuring Polysemy in the Vector Space. The Self Similarity (*SelfSim*) metric introduced by Ethayarajh (2019a) for measuring contextualization and exploring the geometry of the embedding space (presented in the previous section) is a useful tool for studying words' semantic properties. Garí Soler and Apidianaki (2021a) use *SelfSim* to study lexical polysemy. They form four sentence pools from the SemCor corpus (Miller et al. 1993) that reflect different sense distributions:

- `mono` groups instances of a monosemous word (e.g., *camping*, *aircraft*);
- `poly-bal` contains a balanced distribution of each sense of a polysemous word;
- `poly-same` groups instances of a single sense of the polysemous word;
- `poly-rand` is composed of randomly sampled instances of the word.

[34] This issue affects all models tested by Ethayarajh (2019a) but seems to be extreme in the last layer of GPT-2, where two random words could have almost perfect cosine similarity.

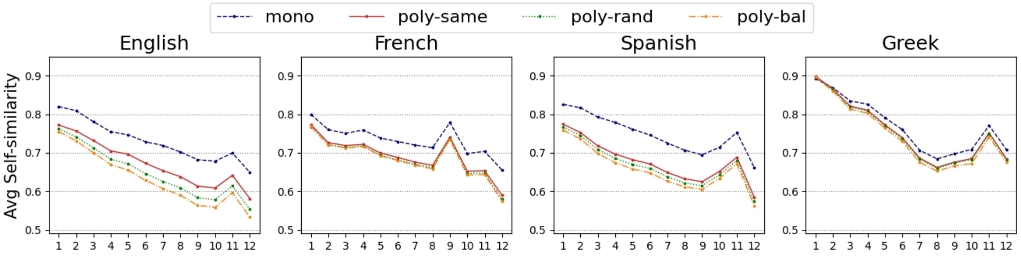


Figure 8
Distinctions established by means of average self-similarity (y axis) between sentence pools representing monosemous and polysemous words with different sense distributions (Garí Soler and Apidianaki 2021a). Distinctions are clear across all 12 layers (x axis) of monolingual BERT models in four languages.

The `poly-rand` pool is expected to be highly biased towards the most frequent sense due to the skewed frequency distribution of word senses (Kilgarriff 2004; McCarthy et al. 2004), and to reflect natural occurrence in texts. The controlled composition of the pools offers the possibility to explore (a) the knowledge about lexical polysemy that is acquired by BERT during pre-training, and (b) the influence of new contexts on token representations. The answer to question (a) is given by comparing the `mono` and `poly-same` pools, which contain instances of the same sense. Question (b) is answered by observing whether it is possible to distinguish the pools that contain different sense distributions for polysemous words using contextualized representations.

The four pools are compared by means of the average pairwise *SelfSim* of the ten instances they each contain. Figure 8 shows the results obtained using monolingual BERT-type (uncased) models in English, French, Spanish, and Greek (Devlin et al. 2019; Le et al. 2020; Cañete et al. 2020; Koutsikakis et al. 2020). The distinctions that are established between the pools are significant in all layers of the models, and show that BERT representations encode an impressive amount of knowledge about polysemy.[35] The distinction between `mono` and `poly-same` is particularly important. Both pools contain instances of a single sense of the studied words (i.e., there is no meaning-related variation across instances inside each pool). Hence, their distinction across layers shows that information about polysemy is acquired by the models during pre-training. Polysemous words should be encountered in more varied contexts than monosemous words, and this variation must be encoded with BERT's language modeling objective when the model is exposed to large amounts of data during training.

This prior knowledge about lexical meaning is combined with information from new contexts of use, as shown by the finer-grained distinctions made between different `poly` pools. As expected, average *SelfSim* is higher between instances of the same sense (`poly-same`) than in `poly-bal` and `poly-rand` which contain instances of different senses. The `poly-rand` pool naturally has higher average *SelfSim* than `poly-bal` due to the larger representation of the most frequent sense in randomly sampled sentences. The decreasing trend observed in the plots and the peak in layer 11 confirm the phases of context encoding and token reconstruction observed by Voita, Sennrich,

[35] The same ranking of pools is observed with the multilingual BERT model in these languages, although the distinctions are less clear than with monolingual models. The multilingual version of BERT was trained on multiple languages and not optimized on each language individually, resulting in poor approximations in languages with less resources (Pimentel et al. 2020).

and Titov (2019). Other results reported by Garí Soler and Apidianaki (2021a) show that average *SelfSim* is higher for monosemous words and words with low polysemy than for highly polysemous words, even when controlling for grammatical category and word frequency. Additionally, BERT representations can serve to determine the partitionability of a word's semantic space into senses (McCarthy, Apidianaki, and Erk 2016).

These findings about BERT's knowledge of lexical polysemy are consistent with results reported by Pimentel et al. (2020), who investigate the relationship between lexical ambiguity and contextual uncertainty using an information theoretic approach. This approach relies on the assumption that speakers compensate lexical ambiguity by making contexts more informative in order to facilitate disambiguation and communication (Piantadosi, Tily, and Gibson 2011). As a corollary, the ambiguity of a word type should correlate with how much information the context provides about it, and negatively correlate with contextual uncertainty. Furthermore, they expect the contextualized representations of polysemous wordforms to be more spread out and to occupy larger regions of the meaning space built by BERT than the representations of wordforms with fewer senses. With respect to the mBERT model, they also observe that it can serve to estimate lexical ambiguity in English, but the quality of the estimates deteriorates in lower-resource languages. A similar type of investigation is performed by Xypolopoulos, Tixier, and Vazirgiannis (2021), who examine the geometry of the ELMo embedding space (Peters et al. 2018). Using multiresolution grids, the authors observe the volume covered by the cloud of points that correspond to different instances of a word, which they consider to be representative of its polysemy. Specifically, they construct a hierarchical discretization of the space where each level corresponds to a different resolution, and the same number of bins are drawn along each dimension. The polysemy score for a word is based on the volume (i.e., the proportion of bins) covered by its vectors at each level.

4.5 Challenges in Vector Space Exploration

4.5.1 Variation Due to Contextualization. Word representations are more dissimilar in upper layers of the models, where contextualization is higher. Voita, Sennrich, and Titov (2019) investigate the evolution of token representations in Transformers trained with different training objectives (LM, MLM, and a Machine Translation objective) using an information-theoretic approach. They explore the flow of information inside the Transformer by specifically estimating the mutual information between a token representation at a certain layer and the input token. Their results show that, with the MLM objective, information about the input token is initially lost during a "context encoding" phase, but is recovered at the last layer (just before prediction) during a "token reconstruction" phase.

Similarly, Mickus et al. (2020) show that BERT representations contain information about the word type,[36] but are also strongly impacted by their position in the sentence and in the input sequence (i.e., whether they occur in the first or the second segment). As explained in Section 2.3.2, BERT distinguishes between tokens that appear in the first and the second sentence by learning two feature vectors called "segment encodings" ($\vec{\texttt{seg}}_{\texttt{A}}$ and $\vec{\texttt{seg}}_{\texttt{B}}$), which are added to all tokens in the two sentences.[37] Information regarding the index i of a token in the sentence is also added using a position embedding

[36] This is demonstrated by their natural clustering according to type.

[37] These vectors serve to mark the two sentences in the input sequence for the NSP pre-training task.

$p(i)$. Thus, the actual input for the sentence "*My dog barks. It is a pooch.*" in the Mickus et al. (2020) paper corresponds to the following sequence of vectors:

$$\begin{gathered}\vec{[\texttt{CLS}]} + p(\vec{0}) + \vec{\texttt{seg}}_{\texttt{A}}, \vec{My} + p(\vec{1}) + \vec{\texttt{seg}}_{\texttt{A}}, \vec{dog} + p(\vec{2}) + \vec{\texttt{seg}}_{\texttt{A}},\\ \vec{barks} + p(\vec{3}) + \vec{\texttt{seg}}_{\texttt{A}}, \vec{.} + p(\vec{4}) + \vec{\texttt{seg}}_{\texttt{A}}, \vec{[\texttt{SEP}]} + p(\vec{5}) + \vec{\texttt{seg}}_{\texttt{A}},\\ \vec{It} + p(\vec{6}) + \vec{\texttt{seg}}_{\texttt{B}}, \vec{is} + p(\vec{7}) + \vec{\texttt{seg}}_{\texttt{B}}, \vec{a} + p(\vec{8}) + \vec{\texttt{seg}}_{\texttt{B}},\\ \vec{pooch} + p(\vec{9}) + \vec{\texttt{seg}}_{\texttt{B}}, \vec{.} + p(\vec{10}) + \vec{\texttt{seg}}_{\texttt{B}}, \vec{[\texttt{SEP}]} + p(\vec{11}) + \vec{\texttt{seg}}_{\texttt{B}}\end{gathered}$$

Vector aggregation differentiates the representations for instances of the same word that occur in similar context, because their segment and position embeddings will be different.

Luo, Kulmizev, and Mao (2021) consider position embeddings responsible for the presence of outliers which they view as a major cause of anisotropy. The outliers are a very small number of dimensions that regularly appear in the same position in the pre-trained encoder layers of a Transformer model (e.g., BERT, RoBERTa, GPT-2, BART, XLNet, ELECTRA). As shown by Kovaleva et al. (2021), disabling these outliers dramatically disrupts model performance. It degrades the quality of the language model as reflected in the MLM loss and in the quality of masked predictions. Luo, Kulmizev, and Mao (2021) train RoBERTa–base models from scratch without using position embeddings, and show that the outliers disappear.

4.5.2 Anisotropy Reduction. Isotropy is a desirable property of word embedding spaces (Huang et al. 2018; Cogswell et al. 2006). As explained in Section 4.4.1, from a geometric point of view, a space is called isotropic if the vectors within that space are uniformly distributed in all directions (Rajaee and Pilehvar 2021). Low isotropy affects the expressiveness of the embedding space and the optimization procedure (models' accuracy and convergence time).

Methods for reducing the anisotropy of the vector space and improving the quality of the obtained similarity estimates have been proposed, initially addressing static embedding representations. Mu and Viswanath (2018) showed that the static word vectors generated by algorithms such as GloVe, Skip-gram, and CBOW, share a large common vector and the same dominating directions. After removing the common mean vector during a post-processing operation, the representations become more "isotropic", that is, more distinct and uniformly distributed within the vector space. By eliminating the top principal components of all words, word representations become stronger and deliver improved performance in intrinsic and downstream tasks.[38] Raunak, Gupta, and Metze (2019) also apply this procedure as the first step in their vector dimensionality reduction method. The resulting embeddings are lower dimension and perform on par or better than the original embeddings in similarity and classification tasks.

In more recent work, Bihani and Rayz (2021) also propose a method that renders off-the-shelf representations isotropic and semantically more meaningful, called "Low Anisotropy Sense Retrofitting" (LASeR). Their method also resolves the representation degeneration problem at a post-processing stage, and conducts sense-enrichment of contextualized representations. Rajaee and Pilehvar (2021) combine the method of Mu

[38] They test the vectors in intrinsic lexical-level tasks (word similarity, concept categorization, and analogy), as well as in sentence-level tasks involving semantic textual similarity and text classification (e.g., sentiment analysis, subjectivity detection, and question classification) (Socher et al. 2013; Pang and Lee 2004, 2005; Maas et al. 2011; Li and Roth 2002).

and Viswanath (2018) with a cluster-based approach, in order to mitigate the degeneration issue that occurs in contextual embedding spaces and increase their isotropy. Their approach relies on the observation that the embedding space of contextual models is extremely anisotropic in all non-input layers from a global sight, but significantly more isotropic from a local point of view (i.e., when embeddings are clustered). Isotropy is measured in their experiments using the method of Arora et al. (2016). They apply PCA to embedding clusters in order to find the principal components (PCs) that indicate the dominant directions for each specific cluster. Dominant directions in clusters of verb representations, for example, encode tense information. As a result, representations for different instances of a verb with the same tense and different meaning are closer to each other in the space than representations for instances with the same meaning and different tense. Removing these directions increases the isotropy of the BERT and RoBERTa spaces, makes them more suitable for semantic applications, and improves performance on semantic tasks.

Rajaee and Pilehvar (2022) show that the spaces that are created by mBERT in different languages are also massively anisotropic.[39] However, unlike monolingual models, there is no dominant dimension that has high contribution to the anisotropic distribution.[40] This contradicts the finding of Luo, Kulmizev, and Mao (2021) about the role of position embeddings in the emergence of outliers, since monolingual and multilingual spaces are constructed using the same training procedure which involves position embeddings. Methods aimed at increasing the isotropy of a multilingual space could, however, significantly improve its representation power and performance.

Most often, representation similarity is estimated using the cosine or the Euclidean distance. Interestingly, Timkey and van Schijndel (2021) call into question the informativity of these measures for contextual language models. They show that these measures are often dominated by 1–5 "rogue dimensions" in GPT-2, BERT, RoBERTa, and XLNET, regardless of the pre-training objective. They explain that it is this small subset of dimensions that drives anisotropy, low self-similarity, and the drop in representational quality in later layers of the models. Their presence can cause cosine similarity and Euclidean distance to rely on less than 1% of the embedding space. Finally, they show that there is a striking mismatch between these dimensions and those that are important to the behavior of the model.

4.5.3 Inter-word Relation Exploration. In the "intra-word" approaches described in the previous section, the context varies across sentences but the target word remains the same. The polysemy pools in the study of Garí Soler and Apidianaki (2021a), for example, contain ten instances of each target word in different contexts. Additionally, the *SelfSim* score for a word corresponds to the average cosine similarity between its contextualized representations across a number of contexts (Ethayarajh 2019a). The exploration of "inter-word" relationships using contextualized representations is more challenging because the target words are also different. A way to control for the imprint of context on the representations would be to substitute the words to be compared in the same sentence (Garí Soler and Apidianaki 2020a, 2021b). Naturally, this process can only be applied to words whose substitution leads to natural sentences. This is often

[39] They analyze spaces created by mBERT in English, Spanish, Arabic, Turkish, Sundanese, and Swahili.

[40] In order to investigate the existence of rogue dimensions, or outliers Kovaleva et al. (2019), they average over 10,000 randomly selected representations and calculate the mean and standard deviation (σ) of dimensions' distribution. Dominant dimensions are dimensions with values at least 3σ larger or smaller than the mean of the distribution.

the case with synonyms (e.g., "*I am* [*glad, happy, excited, delighted*] *to let you know that your paper will be published*"), but can also be the case with hypernyms (e.g., "I like *begonias/flowers*]").[41] By keeping the context stable, the comparison of the contextualized vectors of the words can reveal their similarities and differences. Cloze task probing has also been used for exploring this type of lexical relationships (Ravichander et al. 2020).

An alternative for probing contextualized representations for inter-word relationships is to create some sort of word type-level embedding by aggregating over the representations of individual instances (Vulić et al. 2020b; Bommasani, Davis, and Cardie 2020). This transformation reduces the strong imprint of context variation on the obtained vectors and improves the quality of similarity estimates. Using this type of transformation, Vulić et al. conduct extensive experiments and show that monolingual pre-trained LMs store rich type-level lexical knowledge. Other alternatives include feeding the word without context, or using the embedding before contextualization. Methods that derive word-level representations have, thus, started gaining popularity since they provide a more solid basis for meaning exploration. A detailed account of such methods is given in Section 5.

4.6 Conclusion

In this section, we presented methods that have been proposed for probing contextual language model representations for lexical semantic knowledge. We explained how visualization has been used to this purpose, and we presented prompting methods for semantics and other interpretation methods that serve to analyze the geometry of the vector space. We dedicated a section to the challenges for semantic exploration that are posed by the space itself mainly due to its high anisotropy, and discussed solutions that have been proposed for improving the quality of the similarity estimates that can be drawn from it.

The next section provides an overview of methods aimed at deriving static embeddings from contextualized ones, and at combining the two types of vectors in order to leverage their respective strengths and address their limitations.

5. Deriving Static from Contextualized Representations

Contextualized representations pose challenges to the investigation of lexical semantic knowledge that is situated at the level of word types (cf. Section 4.5). Additionally, static word embeddings are more interpretable than their contextualized counterparts, and the knowledge they encode is easier to analyze due to the wide availability of evaluation datasets. Word type-level embeddings present several advantages over token-level ones in application settings as well, specifically in terms of speed, computational resources, and ease of use (Bommasani, Davis, and Cardie 2020).

In general, continuous real-valued representations give rise to a large memory footprint and slow retrieval speed. This hinders their applicability to low-resource (memory and computation-wise) platforms, such as mobile devices (Shen et al. 2019; Gupta and Jaggi 2021). Tools for utilizing and processing static embeddings efficiently and in a more lightweight fashion have also been proposed (Patel et al. 2018). However, the situation becomes even more complicated with contextualized embeddings. Gupta and

[41] In this case, it is necessary to define some criteria (e.g., language model perplexity or the score of a lexical substitution model such as context2vec [Melamud, Goldberger, and Dagan 2016]) in order to determine whether the generated text is natural.

Jaggi (2021) report that even when ignoring the training phase, the computational cost of using static word embeddings is typically tens of millions times lower than using contextual embedding models. Furthermore, Strubell, Ganesh, and McCallum (2019) characteristically highlight the environmental cost of contextualized representations; training a BERT model on GPU is roughly equivalent to a trans-American flight in terms of carbon emissions.

Given the above observations and the need to integrate embeddings in low-resource settings and devices, word type-level representations are progressively being brought back to the foreground of representation learning. They are also becoming popular in probing studies, since they serve to study the encoded lexical semantic knowledge (Vulić et al. 2020b; Bommasani, Davis, and Cardie 2020). However, this recent interest does not involve the use of "traditional" static embeddings (i.e., word2vec, fastText, or GloVe), but a new type of word type-level embeddings that are derived from contextual language models. The possibility to derive a lexicon at a higher level of abstraction than individual word instances is a valuable tool for linguistic and semantic analysis. Different strategies have been proposed for performing this operation.

5.1 Word Type-level Vector Derivation

5.1.1 Decontextualized Approach. A simple way to generate a decontextualized vector for a word w is to feed it "in isolation" (i.e., without any context) into the pre-trained language model, and to use the output vector as its representation. If the word is split into multiple pieces ($w^1, \ldots, w^k$), these can be concatenated to form a representation for the word ($cat(w^1, \ldots, w^k) = w$) (Bommasani, Davis, and Cardie 2020; Vulić et al. 2020b). Vulić et al. (2021) prepend and append the special tokens `[CLS]` and `[SEP]` to the word or subword sequence before they pass it through BERT.

The decontextualized approach is simple but it presents an unnatural input to the pre-trained encoder which is not trained on out-of-context words. Furthermore, the approach is not very efficient compared with methods that perform pooling over different contexts, described hereafter. Recent work, however, shows that this approach can serve to turn contextual language models into effective decontextualized (static) word encoders through contrastive learning techniques (Vulić et al. 2021). Static type-level vectors can be extracted from any Transformer-based LM, directly from the representations that are generated by the model or after fine-tuning using linguistic constraints from an external resource.

5.1.2 Representation Extraction from the Embedding Layer. Another simple strategy for extracting context-agnostic representations is to use BERT's embedding layer (L_0) before contextualization (Vulić et al. 2020b; Conneau et al. 2020b). Contrary to the decontextualized approach described in Section 5.1.1—where the word is fed to the model in isolation—in this case, the input is a contextualized word instance but its representation is extracted from the embedding layer (L_0) before the context influences the representation. Word vector parameters from this layer can be directly used as static embeddings. However, this method produces suboptimal representations when tested in a wide range of setups and languages (Vulić et al. 2020b; Wang, Cui, and Zhang 2020).

5.1.3 Representation Pooling. Aggregating representations across multiple contexts is the most common approach for creating word type-level embeddings from contextualized representations. In this case, the encoder receives more natural input than with the decontextualized approach. Context pooling for a word w basically involves sampling

Their eyes betrayed too much of their **emotions** , she thought sadly .
In our attempt to interpret the **emotions** in their physiological and pathological range , we emphasized the […]
When she returned to life in the big house she felt shriveled of all **emotion** save dedication to duty .
He wanted no part of the **emotions** of the exchange , no memory of the joy and gratitude that other men felt .
Although in both **emotions** sympathetic symptoms are present , different autonomic somatic patterns underlie […]
[…] the hypothalamic balance would influence emotional_state and behavior but that **emotion** itself would act likewise .
[…] hypothalamic balance plays a crucial role at the crossroads between physiological and pathological forms of **emotion**
[…] she accepted , and so great was my **emotion** that all I could think_of saying was , `` You 're amazing , you_know '' ?
[…] in the hypothalamus with respect to autonomic and somatic functions which are closely associated with the **emotions** .
[…] with others who share this faith , and by opportunities in religious acts for giving_vent to **emotions** and energies .

Figure 9
Sentences collected from SemCor for the noun *emotion*.

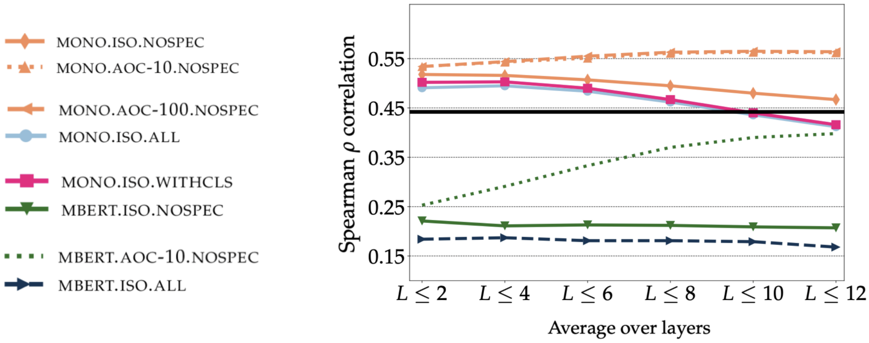


Figure 10
Spearman's ρ correlation scores for different types of representations on the English lexical semantic similarity task of Multi-SimLex (Vulić et al. 2020b). The figure shows results with the English monolingual BERT model (MONO) and with Multilingual BERT (MBERT); for words encoded in isolation (ISO) or using the average over their encodings from 10 or 100 contexts (AOC-10, AOC-100); excluding the [CLS] and [SEP] special tokens (NOSPEC), including them (ALL), and only including the [CLS]. The thick horizontal line denotes fastText vectors' performance. The x axis shows average representations over Transformer layers up to the n-th layer (L_n).

a number n of sentences that contain instances of w, as shown in Figure 9. Vectors are computed for each of these contextualized instances $(w_{c1}, \ldots, w_{cn})$ and a pooling strategy ($f \in \{\texttt{min}, \texttt{max}, \texttt{mean}\}$) is applied to yield a single representation for the word ($w = f(w_{c1}, \ldots, w_{cn})$) (Bommasani, Davis, and Cardie 2020). The number of contexts n to be pooled is a parameter that needs to be set. Interestingly, Vulić et al. (2020b) observe marginal albeit consistent performance gains in all their evaluation tasks when using a larger over a smaller number of contexts (100 vs. 10).[42] This suggests that a few sentences are sufficient for building a good quality word type-level representation. Figure 10 illustrates the results obtained by different static vector derivation approaches and the BERT (MONO) and Multilingual BERT (MBERT) models on the English lexical semantic similarity task of the Multi-SimLex benchmark (Vulić et al. 2020a).[43] The representations that are built from 10 contexts capture similarity better on this task than those derived from 100 contexts, and they both perform better than MBERT and fastText embeddings. The results for words encoded in isolation are shown with the

[42] These include lexical semantic similarity, word analogy, bilingual lexicon induction, cross-lingual information retrieval, and lexical relation prediction.
[43] Multi-SimLex covers 1,888 word pairs in 13 languages.

[MONO | MBERT].ISO.[NOSPEC | WITHCLS | ALL] lines in Figure 10. Interestingly, the representations built using words in isolation are only marginally outscored by their counterparts which aggregate representations across different contexts in several tasks.

Representation pooling has also been applied to ELMo embeddings (Peters et al. 2018). Schuster et al. (2019) showed that "static anchors" created by pooling over ELMo embeddings can serve for aligning cross-lingual spaces. This is not straightforward in the case of token-level embeddings because context variation makes the alignment of cross-lingual spaces difficult. Operating at the anchor level compresses the space and makes possible the use of a dictionary for cross-lingual space alignment.

5.2 Static and Dynamic Vector Combination

Methods that combine static and contextualized representations aim to build on their respective strengths and address their limitations. Similar to knowledge injection methods (cf. Section 3), such combination methods can be applied during vector training or at a post-processing stage. The combined vectors are high quality and tend to perform better than static (word2vec, GloVe, fastText) embeddings and contextualized representations. They thus indicate a viable option for replacing compute-heavy contextual embedding models in application settings where resources are limited. Combination is directional with contextualized representations being used to enhance the quality of static embeddings, and vice versa.

5.2.1 Embedding Combination during Training. These methods integrate contextualized representations into the training process of static embeddings. Wang, Cui, and Zhang (2020) precisely use words' BERT representations for training a Skip-gram model. Contextualized vectors serve, in this case, to resolve ambiguities, and to inject rich syntactic and semantic information in the generated vectors.

Gupta and Jaggi (2021) also propose a distillation method (called X2STATIC) that generates word type-level representations from contextualized vectors by extending CBOW training. Their method relies on the SENT2VEC embedding method (Pagliardini, Gupta, and Jaggi 2018) which uses the entire sentence to predict the masked word instead of a fixed-size context window (used by CBOW), and allows to learn higher *n*-gram representations. Contrary to SENT2VEC, which encodes context as the sum of static vectors, in X2STATIC context is represented as the average of all embeddings returned by BERT. This more refined context representation accounts for word order and interaction.

5.2.2 Embedding Combination during Post-processing. Static embeddings have also been used for enhancing contextualized representations for lexical semantics. The method of Liu, McCarthy, and Korhonen (2020) learns a transformation of contextualized vectors through static anchors. The anchors might be word type-level embeddings (such as word2vec Skip-gram, GloVe, and fastText), or the average of the contextualized representations for a word across instances (similar to the representation pooling methods described in Section 5.1.3).[44]

The contextualized embeddings are represented as a source matrix and the static model representations as the target matrix. In order to combine them, an orthogonal alignment matrix is found which rotates the target space to the source space by solving

[44] The method can also align representations produced by different contextual language models.

Table 1
Performance of RoBERTa on context-aware lexical semantic tasks before and after adjustment to static embeddings (blue rows) and other contextualized embeddings (red rows). Results for a static embedding baseline (fastText) are also given for comparison.

	Within Word			**Inter Word**	
	Usim (ρ)	**WiC (acc%)**	**CoSimlex-I (*r*)**	**CoSimlex-II (ρ)**	**SCWS (ρ)**
fastText	0.1290	56.21	0.2776	0.4481	0.6782
RoBERTa	0.6196	68.28	0.7713	0.7249	0.6884
→ BERT	0.6529	68.21	**0.7814**	0.7087	0.6938
→ XLNet	0.6371	67.50	0.7622	0.6977	0.6689
→ fastText	0.6544	69.00	0.7794	**0.7344**	**0.7159**
→ SGNS	0.6473	**70.07**	0.7761	0.7140	0.7009
→ GloVe	**0.6556**	67.85	0.7783	0.7254	0.6763

the least squares linear regression problem. A linear mapping is learned that serves to transform the source space towards the average of source and the rotated target space. Three contextual models are used in the experiments, BERT (Devlin et al. 2019), RoBERTa (Liu et al. 2019), and XL-Net (Yang et al. 2019). The quality of the transformed vectors is evaluated on three "Within Word" tasks that address the similarity of same-word instances: Usage Similarity (Usim) (Erk, McCarthy, and Gaylord 2013), Word-in-Context (WiC) (Pilehvar and Camacho-Collados 2019), and Cosimlex-I (Armendariz et al. 2020); and two "Inter Word" tasks which address the similarity of different words: Stanford Contextual Word Similarity (SCWS) (Huang et al. 2012) and Cosimlex-II (Armendariz et al. 2020). The evaluation results show that the proposed transformation improves the performance of contextual model representations in all tasks, with the largest and most consistent gains coming from aligning them towards static (especially fastText) embeddings. Overall, the transformation results in better within word contextualization by increasing the similarity of same word instances.[45] It also contributes to a better overall inter-word semantic space, as shown by the improvements in Inter Word tasks. The contextualization power of the original space is preserved and enhanced by the transformation.

Table 1 presents the results reported by Liu, McCarthy, and Korhonen (2020) for the RoBERTa model before and after the adjustment. We show results for the combination of RoBERTa representations with static embeddings (fastText, SGNS, GloVe) and with contextualized embeddings generated by other models (BERT, XLNet), and compare to a fastText embedding baseline. Performance on Usim, CoSimlex-II, and SCWS is measured using Spearman correlation (ρ). Accuracy is used for WiC and uncentered Pearson correlation for Cosimlex-I. The combination of static and contextualized embeddings has also been shown to benefit downstream tasks, such as Bilingual Lexicon Induction (BLI) (Zhang et al. 2021a) and social media categorization (Alghanmi, Espinosa Anke, and Schockaert 2020).

[45] It manages, for example, to correct erroneous predictions on the WiC dataset by bringing closer instances of monosemous words (e.g., *daughter*). These often have low similarity in the original contextual space due to the models' over-sensitivity to context variation.

5.3 Conclusion

The methods that have been presented in this section are aimed at deriving word type-level embeddings from contextualized representations, or at mutually enriching static and contextualized vectors through their combination. The former type of transformation is useful for reducing the impact of context variation, and for exploring the word type-level information that is encoded in contextualized representations. As for vector combination, it serves to improve the quality of each type of embeddings, allowing them to benefit from the strengths of the representations with which they are being combined.

6. Conclusion and Perspectives

The goal of this survey has been to trace the evolution of word meaning representations from static to contextualized word embeddings, and to present current approaches for vector transformation and static embedding derivation. Despite the superiority of contextualized over word type-level vectors in Natural Language Understanding tasks, static representations present several advantages in application settings in terms of speed, computational resources, and ease of use. From a theoretical standpoint, the resurgence of strategies aimed at obtaining vectors at the word type-level from vectors built for individual tokens has been motivated by the distorted similarity estimates that are derived from contextual models' space. Observations about the intense contextualization of token representations—which complicates reasoning at a higher, more abstract, level—have also contributed to this direction.

Several questions around word meaning representation remain open for exploration. Further investigation of the degeneration issue in contextual embedding spaces is needed, as well as the development of mitigation methods aimed at increasing the isotropy of contextual spaces and improving the quality of the similarity estimates that can be derived from them (Ethayarajh 2019a; Mu and Viswanath 2018; Rajaee and Pilehvar 2022). It is also important to revisit the metrics commonly used for measuring semantic similarity which might not be very informative in highly anisotropic contextual spaces (Timkey and van Schijndel 2021). Furthermore, although contextualized representations contain rich information about words and their context, the representation of longer text sequences remains a challenge. The BERT `[CLS]` token, which is commonly used to represent sentences in classification settings, hardly captures their meaning (Reimers and Gurevych 2019). Notably, it produces representations of inferior quality compared to the ones created by averaging over sentences' subword tokens (Vulić et al. 2020b). Such a simple method, which represents a sentence by the weighted average of the pre-computed word vectors (slightly modified using PCA/SVD), was also shown to work particularly well for static embeddings (Arora, Liang, and Ma 2017). However, one drawback of this approach is that it does not account for word order and for the interaction between words in a sentence, which are important for generating a proper interpretation. An interesting new research avenue would be to explore strategies that account for these factors, possibly drawing inspiration from methods that have been proposed in the distributional semantics literature for capturing compositionality (Mitchell and Lapata 2008, 2009; Baroni, Bernardi, and Zamparelli 2014; Boleda and Herbelot 2016; Baroni 2019). A proper representation of the meaning of longer text segments is crucial for improving the models' language understanding and reasoning capabilities.

As far as interpretability is concerned, the initial goal of "BERTology" studies has been to explain the high performance of language models in NLU tasks. It has, however, been shown that the models might not actually use the rich knowledge that is encoded in their representations for performing these tasks (Ravichander, Belinkov, and Hovy 2021; Feder et al. 2021; Belinkov 2022). An interesting and challenging topic for further exploration is to identify the semantic knowledge that is actually used to understand and reason about language. This could be done using counterfactual and adversarial approaches that study the impact of specific types of information on model performance (Elazar et al. 2021; Goodfellow, Shlens, and Szegedy 2015; Jia and Liang 2017; Alzantot et al. 2018).

Adversarial methods constitute a valuable tool for assessing the knowledge that language models encode, their understanding of it, and the extent to which it is used for accomplishing specific tasks. Generating adversarial examples for text data is, however, challenging due to the discrete nature of word tokens (as opposed to the continuous nature of image pixel values). In addition, a natural language attacking system is expected to satisfy three utility-preserving properties: human prediction consistency, semantic similarity, and language fluency (Jin et al. 2020). Future work should focus on developing adversarial methods that satisfy these utility-preserving constraints, in order to increase the usefulness of such methods. Adversarial methods addressing semantics would be highly useful, especially given the brittleness of language models to small changes in the ways they are queried (Jiang et al. 2020).

An additional limitation of models that are trained with the Masked Language Modeling objective is that they learn to predict a single word (or wordpiece) that is missing from a query. As a result, they cannot propose fillers for more than one slot. Combining the probabilities given by BERT to individual wordpieces for tokens composed of multiple subword tokens (or multi-word expressions) is non-trivial, since it would require running BERT several times with the same number of masked tokens as are the wordpieces, increasing computational requirements (Pimentel et al. 2020). In order to estimate the probability distribution over the entire vocabulary, an arbitrary number of MASKs would be needed at each position and the probability values would have to be normalized. Although methods that generate contextual representations at different granularity exist (e.g., SpanBERT and AMBERT), the lexical knowledge encoded in these representations has not yet been analyzed. This type of investigation could provide useful insights regarding the models' understanding of the meaning of longer sequences, and about compositionality processing in contextual language models.

Finally, the semantic knowledge that language models encode depends on the data they were exposed to during training. The impact of reporting bias on the amount and quality of this knowledge is important, since people do not tend to state trivial perceptual or commonsense information in texts (Gordon and Van Durme 2013; Shwartz and Choi 2020). This can, however, be captured using different modalities. Language grounding is a challenging and highly interesting perspective in this respect that could enhance the models' commonsense reasoning potential. Knowledge drawn from images (Lazaridou, Pham, and Baroni 2015; Li et al. 2020, 2021; Zhang et al. 2021b; Yang et al. 2022) could serve to complement the often incomplete information derived from texts, and to reduce the biases that are present, and often amplified, in language model representations.

Appendix A. Word Embedding Evaluation Datasets

Out-of-context Similarity Datasets. Table A1 contains information about existing word similarity datasets that are commonly used for static (word type-level) embedding evaluation. We give the number of word pairs contained in each dataset, the grammatical category of the words included in the dataset, the scale of similarity scores used, and the number of annotators. For most of these datasets, annotators were asked to assign an absolute similarity score to each word pair, while for MEN (Bruni et al. 2012) they were asked to make comparative judgments between words. Given two candidate word pairs (e.g., *wallet-moon* and *car-automobile*), they had to pick the pair whose words were most related in meaning. Each pair was rated against 50 comparison pairs, resulting in a final score on a 50-point scale.

Most of these datasets are in English, except for Multi Simlex which is multilingual and addresses twelve typologically diverse languages: Chinese Mandarin, Welsh, English, Estonian, Finnish, French, Hebrew, Polish, Russian, Spanish, Kiswahili, Yue Chinese. The dataset contains 1,888 semantically aligned concept pairs in each language.

In-context Similarity Datasets. Table A2 contains examples extracted from English in-context similarity datasets commonly used for evaluating contextualized representations. CoInCo (Kremer et al. 2014) and Usim (Erk, McCarthy, and Gaylord 2009) contain manual annotations, while ukWaC-subs (Garí Soler and Apidianaki 2020b) and WiC (Pilehvar and Camacho-Collados 2019) have been automatically created. We provide examples of sentence pairs from each dataset with their annotation and the target word highlighted. For CoInCo, the annotations are in-context substitutes proposed by the annotators. In ukWaC-subs, instance pairs are categorized as true (T) or false (F) depending on the overlap of substitutes that have been automatically assigned to them using context2vec (Melamud, Goldberger, and Dagan 2016). ukWaC-subs contains sentence pairs where a word *w* is replaced by (a) a good substitute, (b) a synonym of a different sense of *w* (i.e., not a good in-context substitute), (c) a random word of the same part of speech. The WiC dataset contains T/F labels that have been automatically assigned to instance pairs based on their proximity in WordNet. Finally, Usim and the

Table A1
Semantic similarity and relatedness datasets. The table shows the number of word pairs present in each resource, the grammatical categories (parts of speech: PoS) covered, the similarity scale used for annotation, and the number of annotators who participated in the annotation task. All datasets are in English, except for Multi-SimLex which includes word pairs in twelve languages.

Dataset	Pairs	PoS	Scale	Annotators
RG-65 (Rubenstein and Goodenough 1965)	65	N	0–4	51
MC-30 (Miller and Charles 1991)	30	N	0–4	38
MC-28 (Resnik 1995)	28	N	0–4	38
WordSim-353 (Finkelstein et al. 2001)	353	N, V, Adj	0–10	13–16
The MEN dataset (Bruni et al. 2012)	3,000	N, V, Adj	0–50	Crowdworkers
SimLex-999 (Hill, Reichart, and Korhonen 2015)	999	N, V, Adj	0–10	500
Multi-SimLex (Vulić et al. 2020a)	1,888	N, V, Adj, Adv	0–6	145
SimVerb-3500 (Gerz et al. 2016)	3,500	V	0–6	843
Stanford Rare Word Similarity (RW) (Luong, Socher, and Manning 2013)	2,034	N, V, Adj, Adv	0–10	10
CARD-660 (Pilehvar et al. 2018)	660	N, Adj, Adv	0–5	8

Table A2
Examples of sentence pairs from English in-context similarity evaluation datasets.

Dataset	Sentence pair	Annotation
CoInCo (Kremer et al. 2014)	A mission to end a war	calling, campaign, dedication, devotion, duty, effort, goal, initiative, intention, movement, plan, pursuit, quest, step, task
	In his heart, he holds the unshakable conviction that the mission he helped organize saved America from disaster.	campaign, initiative, plan, military mission, offensive, operation, project
ukWaC-subs (Gar Soler and Apidianaki 2020b)	(a) For neuroscientists, the message was clear. For neuroscientists, the message was unambiguous	T
	(b) Need a present for someone with a unique name? Need a moment for someone with a unique name?	F
	(c) Overdue tasks display on the due date. Overdue tasks display on the due heritage.	F
WiC (Pilehvar and Camacho-Collados 2019)	The circus will be in town next week . He ran away from home to join the circus .	T
	The political picture is favorable . The dictionary had many pictures .	F
Usim (Erk, McCarthy, and Gaylord 2009)	While both he and the White House deny he was fired [...] At the Cincinatti Enquirer, reporter Mike Gallagher was fired for stealing voice mail messages [...]	4.875/5
	While both he and the White House deny he was fired [...] They shot more blobs of gelfire, fired explosive projectiles .	1.125/5
SCWS (Huang et al. 2012)	[...] Named for the tattoos they decorated themselves with and bitter enemies of encroaching Roman legions [...] [...] and Diana was extremely resentful of Legge-Bourke and her relationship with the young princes . [...]	7.35/10
	[...] Andy 's getting ready to pack his bags and head up to Los Angeles tomorrow [...] [...] who arrives in a pickup truck and defends the house against another pack of zombies [...]	2.1/10

Stanford Contextual Word Similarity (SCWS) dataset (Huang et al. 2012) contain manual usage similarity annotations with scores in the range from 1 (low similarity) to 5 (high similarity) for Usim, and 1 to 10 for SCWS. Usim targets instances of the same word. SCWS contains annotations for instance pairs corresponding to different words, and for homographs with different part of speech (e.g., *pack* as a verb and a noun).

Acknowledgments

This research is based on work supported in part by the European Research Council (ERC) FoTran project (grant agreement 771113), the NSF (award No. 1928631), and the Office of the Director of National Intelligence (ODNI) via the IARPA HIATUS Program (contract No. 2022-22072200005). The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, NSF, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Adi, Yossi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *Proceedings of the International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=BJh6Ztuxl

Agirre, Eneko, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa. 2009. A study on similarity and relatedness using distributional and WordNet-based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 19–27. https://doi.org/10.3115/1620754.1620758

Aina, Laura, Kristina Gulordava, and Gemma Boleda. 2019. Putting words in context: LSTM language models and lexical ambiguity. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3342–3348. https://doi.org/10.18653/v1/P19-1324

Alghanmi, Israa, Luis Espinosa Anke, and Steven Schockaert. 2020. Combining BERT with static word embeddings for categorizing social media. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 28–33. https://doi.org/10.18653/v1/2020.wnut-1.5

Alzantot, Moustafa, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. 2018. Generating natural language adversarial examples. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2890–2896. https://doi.org/10.18653/v1/D18-1316

Apidianaki, Marianna. 2008. Translation-oriented word sense induction based on parallel corpora. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. http://www.lrec-conf.org/proceedings/lrec2008/pdf/822_paper.pdf

Apidianaki, Marianna. 2009. Data-driven semantic analysis for multilingual WSD and lexical selection in translation. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 77–85. https://aclanthology.org/E09-1010

Apidianaki, Marianna and Aina Garí Soler. 2021. ALL dolphins are intelligent and SOME are friendly: Probing BERT for nouns' semantic properties and their prototypicality. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 79–94. https://doi.org/10.18653/v1/2021.blackboxnlp-1.7

Apidianaki, Marianna, Guillaume Wisniewski, Artem Sokolov, Aurélien Max, and François Yvon. 2012. WSD for n-best reranking and local language modeling in SMT. In *Proceedings of the Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 1–9. https://aclanthology.org/W12-4201

Apresjan, Juri D. 1974. Regular polysemy. *Linguistics*, 12:5–32. https://doi.org/10.1515/ling.1974.12.142.5

Arase, Yuki and Jun'ichi Tsujii. 2019. Transfer fine-tuning: A BERT case study. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5393–5404. https://doi.org/10.18653/v1/D19-1542

Armendariz, Carlos Santos, Matthew Purver, Senja Pollak, Nikola Ljubešić, Matej Ulčar, Ivan Vulić, and Mohammad Taher Pilehvar. 2020. SemEval-2020 task 3: Graded word similarity in context. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 36–49. https://doi.org/10.18653/v1/2020.semeval-1.3

Arora, Sanjeev, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2016. A latent variable model approach to PMI-based word embeddings. *Transactions of the Association for Computational*

Linguistics, 4:385–399. https://doi.org/10.1162/tacl_a_00106

Arora, Sanjeev, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2018. Linear algebraic structure of word senses, with applications to polysemy. *Transactions of the Association for Computational Linguistics*, 6:483–495. https://doi.org/10.1162/tacl_a_00034

Arora, Sanjeev, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings. In *Proceedings of the International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=SyK00v5xx

Artetxe, Mikel, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798. https://doi.org/10.18653/v1/P18-1073

Baroni, Marco. 2019. Linguistic generalization and compositionality in modern artificial neural networks. *Philosophical Transactions of the Royal Society B*, 375, 7 pages. https://doi.org/10.1098/rstb.2019.0307

Baroni, Marco, Raffaella Bernardi, and Roberto Zamparelli. 2014. Frege in space: A program for composition distributional semantics. In *Linguistic Issues in Language Technology, Volume 9, 2014 - Perspectives on Semantic Representations for Textual Inference*, pages 241–346. https://doi.org/10.33011/lilt.v9i.1321

Baroni, Marco, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 238–247. https://doi.org/10.3115/v1/P14-1023

Baroni, Marco and Roberto Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1183–1193. https://aclanthology.org/D10-1115

Belinkov, Yonatan. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219. https://doi.org/10.1162/coli_a_00422

Bengio, Yoshua, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A Neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155.

Bihani, Geetanjali and Julia Rayz. 2021. Low Anisotropy Sense Retrofitting (LASeR) : Towards isotropic and sense enriched representations. In *Proceedings of Deep Learning Inside Out (DeeLIO): The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 81–95. https://doi.org/10.18653/v1/2021.deelio-1.9

Blei, David M., Michael I. Jordan, Thomas L. Griffiths, and Joshua B. Tenenbaum. 2003. Hierarchical topic models and the nested Chinese restaurant process. In *Proceedings of the 16th International Conference on Neural Information Processing Systems (NIPS'03)*, pages 17–24. https://proceedings.neurips.cc/paper/2003/file/7b41bfa5085806dfa24b8c9de0ce567f-Paper.pdf

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022. https://doi.org/10.1162/jmlr.2003.3.4-5.993

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146. https://doi.org/10.1162/tacl_a_00051

Boleda, Gemma and Aurélie Herbelot. 2016. Formal distributional semantics: introduction to the special issue. *Computational Linguistics*, 42(4):619–635. https://doi.org/10.1162/COLI_a_00261

Bollacker, Kurt, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD '08, pages 1247–1250. https://doi.org/10.1145/1376616.1376746

Bommasani, Rishi, Kelly Davis, and Claire Cardie. 2020. Interpreting pretrained contextualized representations via reductions to static embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4758–4781. https://doi.org/10.18653/v1/2020.acl-main.431

Bond, Francis and Kyonghee Paik. 2012. A survey of WordNets and their licenses. In *Proceedings of the 6th International Global*

Wordnet Conference (GWC 2012), pages 64–71.

Bouraoui, Zied, José Camacho-Collados, and Steven Schockaert. 2020. Inducing relational knowledge from BERT. In the *Thirty-Fourth AAAI Conference on Artificial Intelligence*, pages 7456–7463.

Bruni, Elia, Gemma Boleda, Marco Baroni, and Nam-Khanh Tran. 2012. Distributional semantics in technicolor. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 136–145. https://aclanthology.org/P12-1015

Bruni, Elia, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *Journal of Artificial Intelligence Research (JAIR)*, 49:1–47. https://doi.org/10.1613/jair.4135

Cai, Xingyu, Jiaji Huang, Yuchen Bian, and Kenneth Church. 2021. Isotropy in the contextual embedding space: Clusters and manifolds. In *Proceedings of the International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=xYGNO86OWDH

Camacho-Collados, José and Mohammad Taher Pilevar. 2018. From word to sense embeddings: A survey on vector representations of meaning. *Journal of Artificial Intelligence Research*, 63:743–788. https://doi.org/10.1613/jair.1.11259

Cao, Yixin, Lifu Huang, Heng Ji, Xu Chen, and Juanzi Li. 2017. Bridge text and knowledge by learning multi-prototype entity mention embedding. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1623–1633. https://doi.org/10.18653/v1/P17-1149

Carpuat, Marine. 2013. NRC: A machine translation approach to cross-lingual word sense disambiguation (SemEval-2013 Task 10). In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 188–192. https://aclanthology.org/S13-2034

Carpuat, Marine and Dekai Wu. 2007. Context-dependent phrasal translation lexicons for statistical machine translation. In *Proceedings of Machine Translation Summit XI: Papers*, 7 pages. https://aclanthology.org/2007.mtsummit-papers.11

Cañete, José, Gabriel Chaperon, Rodrigo Fuentes, and Jorge Pérez. 2020. Spanish pre-trained BERT model and evaluation data. In *Proceedings of the Practical ML for Developing Countries Workshop at ICLR 2020*.

Chen, Xinxiong, Zhiyuan Liu, and Maosong Sun. 2014. A unified model for word sense representation and disambiguation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1025–1035. https://doi.org/10.3115/v1/D14-1110

Chersoni, Emmanuele, Enrico Santus, Chu-Ren Huang, and Alessandro Lenci. 2021. Decoding word embeddings with brain-based semantic features. *Computational Linguistics*, 47(3):663–698. https://doi.org/10.1162/coli_a_00412

Clark, Kevin, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? An analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286. https://doi.org/10.18653/v1/W19-4828

Clark, Kevin, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *Proceedings of the International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=r1xMH1BtvB

Clark, Stephen. 2015. Vector space models of lexical meaning. In *The Handbook of Contemporary Semantic Theory*. John Wiley & Sons, Ltd, pages 493–522. https://doi.org/10.1002/9781118882139.ch16

Cogswell, Michael, Faruk Ahmed, Ross B. Girshick, Larry Zitnick, and Dhruv Batra. 2016. Reducing overfitting in deep networks by decorrelating representations. In *Proceedings of the 4th International Conference on Learning Representations, ICLR 2016, Conference Track Proceedings*, pages 1–12. http://arxiv.org/abs/1511.06068

Collobert, Ronan and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 160–167. https://doi.org/10.1145/1390156.1390177

Collobert, Ronan, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(76):2493–2537. http://jmlr.org/papers/v12/collobert11a.html

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020a. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451. https://doi.org/10.18653/v1/2020.acl-main.747

Conneau, Alexis, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136. https://doi.org/10.18653/v1/P18-1198

Conneau, Alexis, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020b. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034. https://doi.org/10.18653/v1/2020.acl-main.536

Creutz, Mathias. 2018. Open subtitles paraphrase corpus for six languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. https://aclanthology.org/L18-1218

Dagan, Ido and Alon Itai. 1994. Word sense disambiguation using a second language monolingual corpus. *Computational Linguistics*, 20(4):563–596. https://aclanthology.org/J94-4003

Devereux, Barry J., Lorraine K. Tyler, Jeroen Geertzen, and Billi Randall. 2014. The Centre for Speech, Language and the Brain (CSLB) concept property norms. *Behavior Research Methods*, 46(4):1119–1127. https://doi.org/10.3758/s13428-013-0420-4, PubMed: 24356992

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. https://doi.org/10.18653/v1/N19-1423

Diab, Mona and Philip Resnik. 2002. An unsupervised method for word sense tagging using parallel corpora. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 255–262. https://doi.org/10.3115/1073083.1073126

Dinu, Georgiana and Mirella Lapata. 2010. Measuring distributional similarity in context. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1162–1172. https://aclanthology.org/D10-1113

Dinu, Georgiana, Stefan Thater, and Soeren Laue. 2012. A comparison of models of word meaning in context. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 611–615. https://aclanthology.org/N12-1076

Dolan, Bill, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 350–356. https://doi.org/10.3115/1220355.1220406

Drozd, Aleksandr, Anna Gladkova, and Satoshi Matsuoka. 2016. Word embeddings, analogies, and machine learning: Beyond king - man + woman = queen. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3519–3530. https://aclanthology.org/C16-1332

Dyvik, Helge. 1998. Translations as semantic mirrors. In *Proceedings of Workshop W13: Multilinguality in the lexicon II. The 13th biennial European Conference on Artificial Intelligence ECAI 98*, pages 24–44.

Dyvik, Helge. 2002. Translations as semantic mirrors: From parallel corpus to WordNet. *Language and Computers*, 49:311–326.

Dyvik, Helge. 2005. Translations as a semantic knowledge source. In *Proceedings of the Second Baltic Conference on Human Language Technologies*, pages 27–38.

El Boukkouri, Hicham, Olivier Ferret, Thomas Lavergne, Hiroshi Noji, Pierre Zweigenbaum, and Jun'ichi Tsujii. 2020. CharacterBERT: Reconciling ELMo and BERT for word-level open-vocabulary representations from characters. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6903–6915. https://doi.org/10.18653/v1/2020.coling-main.609

Elazar, Yanai, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175. https://doi.org/10.1162/tacl_a_00359

Erk, Katrin. 2012. Vector space models of word meaning and phrase meaning: A survey. *Language and Linguistics Compass*, 6(10):635–653. https://doi.org/10.1002/lnco.362

Erk, Katrin, Diana McCarthy, and Nicholas Gaylord. 2009. Investigations on word senses and word usages. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 10–18. https://aclanthology.org/P09-1002

Erk, Katrin, Diana McCarthy, and Nicholas Gaylord. 2013. Measuring word meaning in context. *Computational Linguistics*, 39(3):511–554. https://doi.org/10.1162/COLI_a_00142

Erk, Katrin and Sebastian Padó. 2008. A structured vector space model for word meaning in context. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 897–906. https://doi.org/10.3115/1613715.1613831

Ethayarajh, Kawin. 2019a. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65. https://doi.org/10.18653/v1/D19-1006

Ethayarajh, Kawin. 2019b. Rotate king to get queen: Word relationships as orthogonal transformations in embedding space. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3503–3508. https://doi.org/10.18653/v1/D19-1354

Ettinger, Allyson. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48. https://doi.org/10.1162/tacl_a_00298

Faruqui, Manaal, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. 2015. Retrofitting word vectors to semantic lexicons. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1606–1615. https://doi.org/10.3115/v1/N15-1184

Feder, Amir, Nadav Oved, Uri Shalit, and Roi Reichart. 2021. CausaLM: Causal model explanation through counterfactual language models. *Computational Linguistics*, 47(2):333–386. https://doi.org/10.1162/coli_a_00404

Fellbaum, Christiane, editor. 1998. *WordNet: An Electronic Lexical Database*. Language, Speech, and Communication. MIT Press, Cambridge, MA. https://doi.org/10.7551/mitpress/7287.001.0001

Ferguson, Thomas S. 1973. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230. https://doi.org/10.1214/aos/1176342360

Finkelstein, Lev, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: The concept revisited. *ACM Transactions on Information Systems - TOIS*, 20:406–414. https://doi.org/10.1145/371920.372094

Firth, John R. 1957. A synopsis of linguistic theory 1930-1955. In *Studies in Linguistic Analysis*, Philological Society, Oxford. Reprinted in Palmer, F. (ed. 1968) *Selected Papers of J. R. Firth*, Longman, Harlow.

Gage, Philip. 1994. A new algorithm for data compression. *C Users Journal*, 12(2):23–38.

Gale, William A., Kenneth W. Church, and David Yarowsky. 1992. Using bilingual materials to develop word sense disambiguation methods. In *Proceedings of the Fourth Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*, pages 101–112. https://aclanthology.org/1992.tmi-1.9

Ganitkevitch, Juri, Benjamin Van Durme, and Chris Callison-Burch. 2013. PPDB: The paraphrase database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 758–764. https://aclanthology.org/N13-1092

Gao, Jun, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tieyan Liu. 2019. Representation degeneration problem in training natural language generation models. In *International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=SkEYojRqtm

Garí Soler, Aina and Marianna Apidianaki. 2020a. BERT knows Punta Cana is not just beautiful, it's gorgeous: Ranking scalar adjectives with contextualised representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7371–7385. https://doi.org/10.18653/v1/2020.emnlp-main.598

Garí Soler, Aina and Marianna Apidianaki. 2020b. MULTISEM at SemEval-2020 Task 3: Fine-tuning BERT for lexical meaning. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 158–165. https://doi.org/10.18653/v1/2020.semeval-1.18

Garí Soler, Aina and Marianna Apidianaki. 2021a. Let's play mono-poly: BERT can reveal words' polysemy level and partitionability into senses. *Transactions of the Association for Computational Linguistics*, 9:825–844. https://doi.org/10.1162/tacl_a_00400

Garí Soler, Aina and Marianna Apidianaki. 2021b. Scalar adjective identification and multilingual ranking. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4653–4660. https://doi.org/10.18653/v1/2021.naacl-main.370

Garí Soler, Aina, Marianna Apidianaki, and Alexandre Allauzen. 2019. LIMSI-MULTISEM at the IJCAI SemDeep-5 WiC challenge: Context representations for word usage similarity estimation. In *Proceedings of the 5th Workshop on Semantic Deep Learning (SemDeep-5)*, pages 6–11. https://aclanthology.org/W19-5802

Gerz, Daniela, Ivan Vulić, Felix Hill, Roi Reichart, and Anna Korhonen. 2016. SimVerb-3500: A large-scale evaluation set of verb similarity. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2173–2182. https://doi.org/10.18653/v1/D16-1235

Glavaš, Goran and Ivan Vulić. 2018. Explicit retrofitting of distributional word vectors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 34–45. https://doi.org/10.18653/v1/P18-1004

Glavaš, Goran and Ivan Vulić. 2019. Generalized tuning of distributional word vectors for monolingual and cross-lingual lexical entailment. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4824–4830. https://doi.org/10.18653/v1/P19-1476

Goodfellow, Ian, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*.

Gordon, Jonathan and Benjamin Van Durme. 2013. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction (AKBC '13)*, pages 25–30. https://doi.org/10.1145/2509558.2509563

Graves, Alex and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, 18(5):602–610. https://doi.org/10.1016/j.neunet.2005.06.042, PubMed: 16112549

Guo, Jiang, Wanxiang Che, Haifeng Wang, and Ting Liu. 2014. Learning sense-specific word embeddings by exploiting bilingual resources. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 497–507. https://aclanthology.org/C14-1048

Gupta, Prakhar and Martin Jaggi. 2021. Obtaining better static word embeddings using contextual embedding models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5241–5253. https://doi.org/10.18653/v1/2021.acl-long.408

Haber, Janosch and Massimo Poesio. 2020. Assessing polyseme sense similarity through co-predication acceptability and contextualised embedding distance. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 114–124. https://aclanthology.org/2020.starsem-1.12

Haber, Janosch and Massimo Poesio. 2021. Patterns of polysemy and homonymy in contextualised language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2663–2676. https://doi.org/10.18653/v1/2021.findings-emnlp.226

Han, Xu, Zhiyuan Liu, and Maosong Sun. 2016. Joint representation learning of text and knowledge for knowledge graph completion. *ArXiv*, 1611.04125.

Harris, Zellig. 1954. Distributional structure. *Word*, 10(2–3):146–162. https://doi.org/10.1080/00437956.1954.11659520

Haviv, Adi, Jonathan Berant, and Amir Globerson. 2021. BERTese: Learning to speak to BERT. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3618–3623. https://doi.org/10.18653/v1/2021.eacl-main.316

Hewitt, John and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743. https://doi.org/10.18653/v1/D19-1275

Hewitt, John and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138. https://doi.org/10.18653/v1/N19-1419

Hill, Felix, Roi Reichart, and Anna Korhonen. 2015. SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695. https://doi.org/10.1162/COLI_a_00237

Hinton, Geoffrey, Oriol Vinyals, and Jeffrey Dean. 2014. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*, 9 pages.

Hochreiter, Sepp and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735, PubMed: 9377276

Hodgson, James M. 1991. Informational constraints on pre-lexical priming. *Language and Cognitive Processes*, 6:169–205. https://doi.org/10.1080/01690969108406942

Huang, Eric, Richard Socher, Christopher Manning, and Andrew Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 873–882. https://aclanthology.org/P12-1092

Huang, Lei, Dawei Yang, Bo Lang, and Jia Deng. 2018. Decorrelated batch normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 791–800.

Hwang, Jena D., Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. (Comet-) Atomic 2020: On symbolic and neural commonsense knowledge graphs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(7):6384–6392. https://doi.org/10.1609/aaai.v35i7.16792

Iacobacci, Ignacio, Mohammad Taher Pilehvar, and Roberto Navigli. 2015. SensEmbed: Learning sense embeddings for word and relational similarity. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 95–105. https://doi.org/10.3115/v1/P15-1010

Ide, Nancy, Collin Baker, Christiane Fellbaum, Charles Fillmore, and Rebecca Passonneau. 2008. MASC: The manually annotated sub-corpus of American English. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. http://www.lrec-conf.org/proceedings/lrec2008/pdf/617_paper.pdf

Ide, Nancy, Collin Baker, Christiane Fellbaum, and Rebecca Passonneau. 2010. The manually annotated sub-corpus: A community resource for and by the people. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 68–73. https://aclanthology.org/P10-2013

Ide, Nancy, Tomaz Erjavec, and Dan Tufis. 2002. Sense discrimination with parallel corpora. In *Proceedings of the ACL-02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, pages 61–66. https://doi.org/10.3115/1118675.1118683

Jauhar, Sujay Kumar, Chris Dyer, and Eduard Hovy. 2015. Ontologically grounded multi-sense representation learning for semantic vector space models. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 683–693. https://doi.org/10.3115/v1/N15-1070

Jia, Robin and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031. https://doi.org/10.18653/v1/D17-1215

Jiang, Zhengbao, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438. https://doi.org/10.1162/tacl_a_00324

Jin, Di, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. Is BERT really robust? A strong baseline for natural language attack on text classification and entailment. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(5):8018–8025. https://doi.org/10.1609/aaai.v34i05.6311

Joshi, Mandar, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77. https://doi.org/10.1162/tacl_a_00300

Jozefowicz, Rafal, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. 2016. Exploring the limits of language modeling. *arXiv:1602.02410*.

Jurgens, David, Saif Mohammad, Peter Turney, and Keith Holyoak. 2012. SemEval-2012 Task 2: Measuring degrees of relational similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 356–364. https://aclanthology.org/S12-1047

Kabbach, Alexandre and Aurélie Herbelot. 2021. Avoiding conflict: When speaker coordination does not require conceptual agreement. *Frontiers in Artificial Intelligence*, 3:523920. https://doi.org/10.3389/frai.2020.523920, PubMed: 33733196

Kiela, Douwe, Felix Hill, and Stephen Clark. 2015. Specializing word embeddings for similarity or relatedness. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2044–2048. https://doi.org/10.18653/v1/D15-1242

Kilgarriff, Adam. 2004. *How dominant is the commonest sense of a word? Lecture Notes in Computer Science* (vol. 3206), *Text, Speech and Dialogue*. Sojka Petr, Kopeek Ivan, Pala Karel (eds.). Springer, Berlin, Heidelberg, pages 103–112. https://doi.org/10.1007/978-3-540-30120-2_14

Kim, Yoon, Yacine Jernite, David Sontag, and Alexander M. Rush. 2016. Character-aware neural language models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, pages 2741–2749. https://doi.org/10.1609/aaai.v30i1.10362

Kipfer, Barbara Ann. 2009. *Roget's 21st Century Thesaurus*, 3rd edition. Philip Lief Group.

Kipper Schuler, Karin. 2006. *VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon*. Ph.D. thesis, University of Pennsylvania.

Koutsikakis, John, Ilias Chalkidis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2020. GREEK-BERT: The Greeks visiting Sesame Street. In *11th Hellenic Conference on Artificial Intelligence*, pages 110–117. https://doi.org/10.1145/3411408.3411440

Kovaleva, Olga, Saurabh Kulshreshtha, Anna Rogers, and Anna Rumshisky. 2021. BERT busters: Outlier dimensions that disrupt transformers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3392–3405. https://doi.org/10.18653/v1/2021.findings-acl.300

Kovaleva, Olga, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the dark secrets of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4364–4373. https://doi.org/10.18653/v1/D19-1445

Kremer, Gerhard, Katrin Erk, Sebastian Padó, and Stefan Thater. 2014. What substitutes tell us - analysis of an "all-words" lexical substitution corpus. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 540–549. https://doi.org/10.3115/v1/E14-1057

Lample, Guillaume, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *International Conference on Learning Representations (ICLR'18)*. https://openreview.net/forum?id=H196sainb

Lan, Zhenzhong, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A Lite BERT for self-supervised learning of language representations. In *International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=H1eA7AEtvS

Landauer, Thomas K. and Susan T. Dumais. 1997. A Solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240. https://doi.org/10.1037/0033-295X.104.2.211

Lauscher, Anne, Ivan Vulić, Edoardo Maria Ponti, Anna Korhonen, and Goran Glavaš. 2020. Specializing unsupervised pretraining models for word-level semantic similarity. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1371–1383. https://doi.org/10.18653/v1/2020.coling-main.118

Lazaridou, Angeliki, Nghia The Pham, and Marco Baroni. 2015. Combining language and vision with a multimodal skip-gram model. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 153–163. https://doi.org/10.3115/v1/N15-1016

Le, Hang, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoît Crabbé, Laurent Besacier, and Didier Schwab. 2020. FlauBERT: Unsupervised language model pre-training for French. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2479–2490. https://aclanthology.org/2020.lrec-1.302

LeCun, Yann, Y. Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature*, 521:436–444. https://doi.org/10.1038/nature14539, PubMed: 26017442

Lee, Daniel and H. Sebastian Seung. 2000. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems*, volume 13, MIT Press. https://proceedings.neurips.cc/paper/2000/file/f9d1152547c0bde01830b7e8bd60024c-Paper.pdf

Lefever, Els, Véronique Hoste, and Martine De Cock. 2011. ParaSense or how to use parallel corpora for word sense disambiguation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 317–322. https://aclanthology.org/P11-2055

Levine, Yoav, Barak Lenz, Or Dagan, Ori Ram, Dan Padnos, Or Sharir, Shai Shalev-Shwartz, Amnon Shashua, and Yoav Shoham. 2020. SenseBERT: Driving some sense into BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4656–4667. https://doi.org/10.18653/v1/2020.acl-main.423

Li, Jiwei and Dan Jurafsky. 2015. Do multi-sense embeddings improve natural language understanding? In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1722–1732. https://doi.org/10.18653/v1/D15-1200

Li, Liunian Harold, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2020. What does BERT with vision look at? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5265–5275. https://doi.org/10.18653/v1/2020.acl-main.469

Li, Liunian Harold, Haoxuan You, Zhecan Wang, Alireza Zareian, Shih-Fu Chang, and Kai-Wei Chang. 2021. Unsupervised vision-and-language pre-training without parallel images and captions. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5339–5350. https://doi.org/10.18653/v1/2021.naacl-main.420

Li, Xin and Dan Roth. 2002. Learning question classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics*, 7 pages. https://doi.org/10.3115/1072228.1072378

Linzen, Tal. 2016. Issues in evaluating semantic spaces using word analogies. In *Proceedings of the First Workshop on Evaluating Vector-Space Representations for NLP*, pages 13–18. https://doi.org/10.18653/v1/W16-2503

Linzen, Tal. 2019. What can linguistics and deep learning contribute to each other? Response to Pater. *Language*, 95(1):99–108.

Linzen, Tal, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535. https://doi.org/10.1353/lan.2019.0015

Liu, Frederick, Han Lu, and Graham Neubig. 2018. Handling homographs in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1336–1345. https://doi.org/10.18653/v1/N18-1121

Liu, Pengfei, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2022. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 34 pages. https://doi.org/10.1145/3560815

Liu, Qianchu, Diana McCarthy, and Anna Korhonen. 2020. Towards better context-aware lexical semantics: Adjusting contextualized representations through static anchors. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4066–4075. https://doi.org/10.18653/v1/2020.emnlp-main.333

Liu, Qianchu, Diana McCarthy, and Anna Korhonen. 2022. Measuring context-word biases in lexical semantic datasets. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2699–2713. https://aclanthology,org/W13-3512

Liu, Qianchu, Edoardo Maria Ponti, Diana McCarthy, Ivan Vulić, and Anna Korhonen. 2021. AM2iCo: Evaluating word meaning in context across low-resource languages with adversarial examples. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7151–7162. https://doi.org/10.18653/v1/2021.emnlp-main.571

Liu, Weijie, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2020. K-BERT: Enabling language representation with knowledge graph. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)*. https://doi.org/10.1609/aaai.v34i03.5681

Liu, Yang, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. 2015. Topical word embeddings. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 2418–2424. https://doi.org/10.1609/aaai.v29i1.9522

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT pretraining approach. *arXiv:1907.11692*.

Logan, Robert, Nelson F. Liu, Matthew E. Peters, Matt Gardner, and Sameer Singh. 2019. Barack's wife Hillary: Using knowledge graphs for fact-aware language modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5962–5971. https://doi.org/10.18653/v1/P19-1598

Loureiro, Daniel, Kiamehr Rezaee, Mohammad Taher Pilehvar, and Jose Camacho-Collados. 2021. Analysis and evaluation of language models for word sense disambiguation. *Computational Linguistics*, 47(2):387–443. https://doi.org/10.1162/coli_a_00405

Lund, Kevin and Curt Burgess. 1996. Producing high-dimensional semantic space from lexical co-occurrence. *Behavior Research Methods Instruments & Computers*, 28:203–208. https://doi.org/10.3758/BF03204766

Luo, Ziyang, Artur Kulmizev, and Xiaoxi Mao. 2021. Positional artefacts propagate through masked language model embeddings. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5312–5327. https://doi.org/10.18653/v1/2021.acl-long.413

Luong, Minh Thang and Christopher D. Manning. 2016. Achieving open vocabulary neural machine translation with hybrid word-character models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1054–1063. https://doi.org/10.18653/v1/P16-1100

Luong, Thang, Richard Socher, and Christopher Manning. 2013. Better word representations with recursive neural networks for morphology. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 104–113. https://aclanthology.org/W13-3512

Lyu, Qing, Zheng Hua, Daoxin Li, Li Zhang, Marianna Apidianaki, and Chris Callison-Burch. 2022. Is "My Favorite New Movie" my favorite movie? Probing the understanding of recursive noun phrases. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5286–5302. https://

doi.org/10.18653/v1/2022.naacl-main.388

Maas, Andrew L., Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150. https://aclanthology.org/P11-1015

Mancini, Massimiliano, Jose Camacho-Collados, Ignacio Iacobacci, and Roberto Navigli. 2017. Embedding words and senses together via joint knowledge-enhanced training. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 100–111. https://doi.org/10.18653/v1/K17-1012

Martelli, Federico, Najla Kalach, Gabriele Tola, and Roberto Navigli. 2021. SemEval-2021 Task 2: Multilingual and cross-lingual word-in-context disambiguation (MCL-WiC). In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 24–36. https://doi.org/10.18653/v1/2021.semeval-1.3

Martin, Louis, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: A tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219. https://doi.org/10.18653/v1/2020.acl-main.645

McCarthy, Diana, Marianna Apidianaki, and Katrin Erk. 2016. Word sense clustering and clusterability. *Computational Linguistics*, 42(2):245–275. https://doi.org/10.1162/COLI_a_00247

McCarthy, Diana, Rob Koeling, Julie Weeds, and John Carroll. 2004. Finding predominant word senses in untagged text. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 279–286. https://doi.org/10.3115/1218955.1218991

McCarthy, Diana and Roberto Navigli. 2007. SemEval-2007 Task 10: English lexical substitution task. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 48–53. https://doi.org/10.3115/1621474.1621483

McRae, Ken, George Cree, Mark Seidenberg, and Chris Mcnorgan. 2005. Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, 37(4):547–59. https://doi.org/10.3758/BF03192726, PubMed: 16629288

Melamud, Oren, Jacob Goldberger, and Ido Dagan. 2016. context2vec: Learning generic context embedding with bidirectional LSTM. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 51–61. https://doi.org/10.18653/v1/K16-1006

Meyerson, Adam. 2001. Online facility location. In *Proceedings of the 42nd IEEE Symposium on Foundations of Computer Science (FOCS '01)*, pages 426–431. https://doi.org/10.1109/SFCS.2001.959917

Mickus, Timothee, Denis Paperno, Mathieu Constant, and Kees van Deemter. 2020. What do you mean, BERT? Assessing BERT as a distributional semantics model. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 279–290. https://aclanthology.org/2020.scil-1.35

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26, Curran Associates, Inc. https://proceedings.neurips.cc/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf

Mikolov, Tomas, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751. https://aclanthology.org/N13-1090

Miller, George A. and Walter G. Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6:1–28.

Miller, George A., Claudia Leacock, Randee Tengi, and Ross T. Bunker. 1993. A semantic concordance. In *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*, pages 303–308. https://doi.org/10.3115/1075671.1075742

Misra, Kanishka, Julia Rayz, and Allyson Ettinger. 2022. A property induction framework for neural language models. In *Proceedings of the 44th Annual Conference of the Cognitive Science Society*, pages 1977–1984.

Mitchell, Jeff and Mirella Lapata. 2008. Vector-based models of semantic composition. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL'08: HLT)*, pages 236–244. https://aclanthology.org/P08-1028

Mitchell, Jeff and Mirella Lapata. 2009. Language models based on semantic composition. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 430–439. https://aclanthology.org/D09-1045

Moro, Andrea, Alessandro Raganato, and Roberto Navigli. 2014. Entity linking meets word sense disambiguation: A unified approach. *Transactions of the Association for Computational Linguistics*, 2:231–244. https://doi.org/10.1162/tacl_a_00179

Mrkšić, Nikola, Diarmuid Ó. Séaghdha, Blaise Thomson, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2016. Counter-fitting word vectors to linguistic constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 142–148. https://doi.org/10.18653/v1/N16-1018

Mrkšić, Nikola, Ivan Vulić, Diarmuid Ó. Séaghdha, Ira Leviant, Roi Reichart, Milica Gašić, Anna Korhonen, and Steve Young. 2017. Semantic specialization of distributional word vector spaces using monolingual and cross-lingual constraints. *Transactions of the Association for Computational Linguistics*, 5:309–324. https://doi.org/10.1162/tacl_a_00063

Mu, Jiaqi and Pramod Viswanath. 2018. All-but-the-top: Simple and effective postprocessing for word representations. In *International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=HkuGJ3kCb

Navigli, Roberto and Simone Paolo Ponzetto. 2010. BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225. https://aclanthology.org/P10-1023

Neelakantan, Arvind, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. 2014. Efficient non-parametric estimation of multiple embeddings per word in vector space. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1059–1069. https://doi.org/10.3115/v1/D14-1113

Nelson, Douglas, Cathy Mcevoy, and Thomas Schreiber. 2004. The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, 36(3):402–407. https://doi.org/10.3758/BF03195588, PubMed: 15641430

Nguyen, Kim Anh, Maximilian Köeper, Sabine Schulte im Walde, and Ngoc Thang Vu. 2017. Hierarchical embeddings for hypernymy detection and directionality. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 233–243. https://doi.org/10.18653/v1/D17-1022

Nguyen, Kim Anh, Sabine Schulte im Walde, and Ngoc Thang Vu. 2016. Integrating distributional lexical contrast into word embeddings for antonym-synonym distinction. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 454–459. https://doi.org/10.18653/v1/P16-2074

Nissim, Malvina, Rik van Noord, and Rob van der Goot. 2020. Fair is better than sensational: Man is to doctor as woman is to doctor. *Computational Linguistics*, 46(2):487–497. https://doi.org/10.1162/coli_a_00379

Ono, Masataka, Makoto Miwa, and Yutaka Sasaki. 2015. Word embedding-based antonym detection using thesauri and distributional information. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 984–989. https://doi.org/10.3115/v1/N15-1100

Padó, Sebastian and Mirella Lapata. 2007. Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199. https://doi.org/10.1162/coli.2007.33.2.161

Pagliardini, Matteo, Prakhar Gupta, and Martin Jaggi. 2018. Unsupervised learning of sentence embeddings using compositional *n*-gram features. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 528–540. `https://doi.org/10.18653/v1/N18-1049`

Pang, Bo and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 271–278. `https://doi.org/10.3115/1218955.1218990`

Pang, Bo and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 115–124. `https://doi.org/10.3115/1219840.1219855`

Pasini, Tommaso, Alessandro Raganato, and Roberto Navigli. 2021. XL-WSD: An extra-large and cross-lingual evaluation framework for word sense disambiguation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13648–13656. `https://doi.org/10.1609/aaai.v35i15.17609`

Patel, Ajay, Alexander Sands, Chris Callison-Burch, and Marianna Apidianaki. 2018. Magnitude: A fast, efficient universal vector embedding utility package. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 120–126. `https://doi.org/10.18653/v1/D18-2021`

Pavlick, Ellie, Pushpendre Rastogi, Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2015. PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 425–430. `https://doi.org/10.3115/v1/P15-2070`

Pennington, Jeffrey, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543. `https://doi.org/10.3115/v1/D14-1162`

Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237. `https://doi.org/10.18653/v1/N18-1202`

Peters, Matthew E., Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. 2019. Knowledge enhanced contextual word representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 43–54. `https://doi.org/10.18653/v1/D19-1005`

Petersen, Erika and Christopher Potts. 2022. Lexical semantics with large language models: A case study of English *break*. Manuscript, Stanford University. Available at `https://ling.auf.net/lingbuzz/006859`.

Petroni, Fabio, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473. `https://doi.org/10.18653/v1/D19-1250`

Pham, Nghia The, Angeliki Lazaridou, and Marco Baroni. 2015. A multitask objective to inject lexical contrast into distributional semantics. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 21–26. `https://doi.org/10.3115/v1/P15-2004`

Piantadosi, Steven, Harry Tily, and Edward Gibson. 2011. Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences of the United States of America*, 108:3526–3529.

Pilehvar, Mohammad Taher and Jose Camacho-Collados. 2019. WiC: The Word-in-Context dataset for evaluating context-sensitive meaning representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short*

Papers), pages 1267–1273. https://doi.org/10.18653/v1/N19-1128

Pilehvar, Mohammad Taher and Jose Camacho-Collados. 2020. *Embeddings in Natural Language Processing: Theory and Advances in Vector Representations of Meaning*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, Cambridge, MA. https://doi.org/10.1007/978-3-031-02177-0

Pilehvar, Mohammad Taher and Nigel Collier. 2016. De-conflated semantic representations. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1680–1690. https://doi.org/10.18653/v1/D16-1174

Pilehvar, Mohammad Taher, Dimitri Kartsaklis, Victor Prokhorov, and Nigel Collier. 2018. Card-660: Cambridge Rare Word Dataset - a reliable benchmark for infrequent word representation models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1391–1401. https://doi.org/10.18653/v1/D18-1169

Pimentel, Tiago, Rowan Hall Maudslay, Damian Blasi, and Ryan Cotterell. 2020. Speakers fill lexical semantic gaps with context. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4004–4015. https://doi.org/10.18653/v1/2020.emnlp-main.328

Ponti, Edoardo Maria, Ivan Vulić, Goran Glavaš, Nikola Mrkšić, and Anna Korhonen. 2018. Adversarial propagation and zero-shot cross-lingual transfer of word vector specialization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 282–293. https://doi.org/10.18653/v1/D18-1026

Radford, Alec, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. Technical Report. Available at https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf.

Raganato, Alessandro, Tommaso Pasini, Jose Camacho-Collados, and Mohammad Taher Pilehvar. 2020. XL-WiC: A multilingual benchmark for evaluating semantic contextualization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7193–7206. https://doi.org/10.18653/v1/2020.emnlp-main.584

Raganato, Alessandro and Jörg Tiedemann. 2018. An analysis of encoder representations in transformer-based machine translation. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 287–297. https://doi.org/10.18653/v1/W18-5431

Rajaee, Sara and Mohammad Taher Pilehvar. 2021. A cluster-based approach for improving isotropy in contextual embedding space. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 575–584. https://doi.org/10.18653/v1/2021.acl-short.73

Rajaee, Sara and Mohammad Taher Pilehvar. 2022. An isotropy analysis in the multilingual BERT embedding space. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1309–1316. https://doi.org/10.18653/v1/2022.findings-acl.103

Raunak, Vikas, Vivek Gupta, and Florian Metze. 2019. Effective dimensionality reduction for word embeddings. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 235–243. https://doi.org/10.18653/v1/W19-4328

Ravichander, Abhilasha, Yonatan Belinkov, and Eduard Hovy. 2021. Probing the probing paradigm: Does probing accuracy entail task relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3363–3377. https://doi.org/10.18653/v1/2021.eacl-main.295

Ravichander, Abhilasha, Eduard Hovy, Kaheer Suleman, Adam Trischler, and Jackie Chi Kit Cheung. 2020. On the systematicity of probing contextualized word representations: The case of hypernymy in BERT. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 88–102. https://aclanthology.org/2020.starsem-1.10

Reif, Emily, Ann Yuan, Martin Wattenberg, Fernanda B. Viegas, Andy Coenen, Adam Pearce, and Been Kim. 2019. Visualizing and measuring the geometry of BERT. In *Advances in Neural Information Processing Systems*, pages 8592–8600.

https://proceedings.neurips.cc/paper/2019/file/159c1ffe5b61b41b3c4d8f4c2150f6c4-Paper.pdf

Reimers, Nils and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992. https://doi.org/10.18653/v1/D19-1410

Reisinger, Joseph and Raymond J. Mooney. 2010. Multi-prototype vector-space models of word meaning. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 109–117. https://aclanthology.org/N10-1013

Resnik, Philip. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1*, IJCAI'95, pages 448–453. https://doi.org/10.1007/978-3-540-24630-5_35

Resnik, Philip. 2004. Exploiting hidden meanings: Using bilingual text for monolingual annotation. In *Computational Linguistics and Intelligent Text Processing*, pages 283–299. Springer Berlin Heidelberg, Berlin, Heidelberg.

Resnik, Philip and David Yarowsky. 1999. Distinguishing systems and distinguishing senses: new evaluation methods for word sense disambiguation. *Natural Language Engineering*, 5:113–133. https://doi.org/10.1017/S1351324999002211

Riedel, Sebastian, Limin Yao, Andrew McCallum, and Benjamin M. Marlin. 2013. Relation extraction with matrix factorization and universal schemas. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 74–84. https://aclanthology.org/N13-1008

Rogers, Anna, Aleksandr Drozd, and Bofang Li. 2017. The (too many) problems of analogical reasoning with word vectors. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 135–148. https://doi.org/10.18653/v1/S17-1017

Rogers, Anna, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866. https://doi.org/10.1162/tacl_a_00349

Rothe, Sascha and Hinrich Schütze. 2017. AutoExtend: Combining word embeddings with semantic resources. *Computational Linguistics*, 43(3):593–617. https://doi.org/10.1162/COLI_a_00294

Rubenstein, Herbert and John B. Goodenough. 1965. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633. https://doi.org/10.1145/365628.365657

Sanh, Victor, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *ArXiv*, 1910.01108.

Schluter, Natalie. 2018. The word analogy testing caveat. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 242–246. https://doi.org/10.18653/v1/N18-2039

Schuster, Tal, Ori Ram, Regina Barzilay, and Amir Globerson. 2019. Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1599–1613. https://doi.org/10.18653/v1/N19-1162

Schütze, Hinrich. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123. https://aclanthology.org/J98-1004

Schwartz, Roy, Roi Reichart, and Ari Rappoport. 2015. Symmetric pattern based word embeddings for improved word similarity prediction. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 258–267. https://doi.org/10.18653/v1/K15-1026

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725. https://doi.org/10.18653/v1/P16-1162

Shen, Dinghan, Pengyu Cheng, Dhanasekar Sundararaman, Xinyuan Zhang, Qian Yang, Meng Tang, Asli Celikyilmaz, and Lawrence Carin. 2019. Learning compressed sentence representations for on-device text processing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 107–116. https://doi.org/10.18653/v1/P19-1011

Shi, Weijia, Muhao Chen, Pei Zhou, and Kai-Wei Chang. 2019. Retrofitting contextualized word embeddings with paraphrases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1198–1203. https://doi.org/10.18653/v1/D19-1113

Shwartz, Vered and Yejin Choi. 2020. Do neural language models overcome reporting bias? In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6863–6870. https://doi.org/10.18653/v1/2020.coling-main.605

Socher, Richard, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642. https://aclanthology.org/D13-1170

Strubell, Emma, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650. https://doi.org/10.18653/v1/P19-1355

Šuster, Simon, Ivan Titov, and Gertjan van Noord. 2016. Bilingual learning of multi-sense embeddings with discrete autoencoders. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1346–1356. https://doi.org/10.18653/v1/N16-1160

Talmor, Alon, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. 2020. oLMpics-on what language model pre-training captures. *Transactions of the Association for Computational Linguistics*, 8:743–758. https://doi.org/10.1162/tacl_a_00342

Taylor, Wilson L. 1953. "Cloze procedure": A new tool for measuring readability. *Journalism quarterly*, 30(4):415–433. https://doi.org/10.1177/107769905303000401

Tenney, Ian, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601. https://doi.org/10.18653/v1/P19-1452

Tenney, Ian, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Sam Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? Probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=SJzSgnRcKX

Thater, Stefan, Hagen Fürstenau, and Manfred Pinkal. 2011. Word meaning in context: A simple and effective vector model. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 1134–1143. https://aclanthology.org/I11-1127

Tian, Fei, Hanjun Dai, Jiang Bian, Bin Gao, Rui Zhang, Enhong Chen, and Tie-Yan Liu. 2014. A probabilistic model for learning multi-prototype word embeddings. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 151–160. https://aclanthology.org/C14-1016

Tiedemann, Jörg. 2012. Character-based pivot translation for under-resourced languages and domains. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 141–151. https://aclanthology.org/E12-1015

Timkey, William and Marten van Schijndel. 2021. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4527–4546. https://doi.org/10.18653/v1/2021.emnlp-main.372

Toutanova, Kristina, Danqi Chen, Patrick Pantel, Hoifung Poon, Pallavi Choudhury, and Michael Gamon. 2015. Representing text for joint embedding of text and knowledge bases. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509. https://doi.org/10.18653/v1/D15-1174

Turney, Peter and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188. https://doi.org/10.1613/jair.2934

Turney, Peter D. 2006. Similarity of semantic relations. *Computational Linguistics*, 32(3):379–416. https://doi.org/10.1162/coli.2006.32.3.379

Turney, Peter D. 2012. Domain and function: A dual-space model of semantic relations and compositions. *Journal of Artificial Intelligence Research*, 44(1):533–585. https://doi.org/10.1613/jair.3640

Van de Cruys, Tim, Thierry Poibeau, and Anna Korhonen. 2011. Latent vector weighting for word meaning in context. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1012–1022. https://aclanthology.org/D11-1094

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper-Paper.pdf

Virtanen, Antti, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. 2019. Multilingual is not enough: BERT for Finnish. *arXiv: 1912.07076*.

Voita, Elena, Rico Sennrich, and Ivan Titov. 2019. The bottom-up evolution of representations in the transformer: A study with machine translation and language modeling objectives. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4396–4406. https://doi.org/10.18653/v1/D19-1448

Voita, Elena, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808. https://doi.org/10.18653/v1/P19-1580

Vulić, Ivan. 2018. Injecting lexical contrast into word vectors by guiding vector space specialisation. In *Proceedings of The Third Workshop on Representation Learning for NLP*, pages 137–143. https://doi.org/10.18653/v1/W18-3018

Vulić, Ivan, Simon Baker, Edoardo Maria Ponti, Ulla Petti, Ira Leviant, Kelly Wing, Olga Majewska, Eden Bar, Matt Malone, Thierry Poibeau, Roi Reichart, and Anna Korhonen. 2020a. Multi-SimLex: A large-scale evaluation of multilingual and crosslingual lexical semantic similarity. *Computational Linguistics*, 46(4):847–897. https://doi.org/10.1162/coli_a_00391

Vulić, Ivan and Nikola Mrkšić. 2018. Specialising word vectors for lexical entailment. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1134–1145. https://doi.org/10.18653/v1/N18-1103

Vulić, Ivan, Edoardo Maria Ponti, Anna Korhonen, and Goran Glavaš. 2021. LexFit: Lexical fine-tuning of pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5269–5283. https://doi.org/10.18653/v1/2021.acl-long.410

Vulić, Ivan, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020b. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240. https://doi.org/10.18653/v1/2020.emnlp-main.586

Wang, Xiaozhi, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. 2021. KEPLER: A unified model for knowledge embedding and pre-trained language representation. *Transactions of the Association for Computational Linguistics*, 9:176–194. https://doi.org/10.1162/tacl_a_00360

Wang, Yile, Leyang Cui, and Yue Zhang. 2020. How can BERT help lexical semantics tasks? *arXiv: 1911.02929*.

Wang, Zhen, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge graph and text jointly embedding. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1591–1601. https://doi.org/10.3115/v1/D14-1167

Webson, Albert and Ellie Pavlick. 2022. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2300–2344. https://

doi.org/10.18653/v1/2022.naacl-main.167

West, Peter, Chandra Bhagavatula, Jack Hessel, Jena Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 2022. Symbolic knowledge distillation: From general language models to commonsense models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4602–4625. https://doi.org/10.18653/v1/2022.naacl-main.341

Wiedemann, Gregor, Steffen Remus, Avi Chawla, and Chris Biemann. 2019. Does BERT make any sense? Interpretable word sense disambiguation with contextualized embeddings. In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)*, pages 161–170.

Wieting, John, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. From paraphrase database to compositional paraphrase model and back. *Transactions of the Association for Computational Linguistics*, 3:345–358. https://doi.org/10.1162/tacl_a_00143, https://doi.org/10.1162/tacl_a_00246

Wieting, John, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2016. Charagram: Embedding words and sentences via character n-grams. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1504–1515. https://doi.org/10.18653/v1/D16-1157

Wittgenstein, Ludwig. 1953. P*hilosophical Investigations*. Published 2001, Third ed. Blackwell Publishing Ltd, Oxford, England. Translated by G.E.M. Anscombe.

Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv:1609.08144*.

Xu, Chang, Yalong Bai, Jiang Bian, Bin Gao, Gang Wang, Xiaoguang Liu, and Tie-Yan Liu. 2014. RC-NET: A general framework for incorporating knowledge into word representations. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, pages 1219–1228. https://doi.org/10.1145/2661829.2662038

Xypolopoulos, Christos, Antoine Tixier, and Michalis Vazirgiannis. 2021. Unsupervised word polysemy quantification with multiresolution grids of contextual embeddings. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3391–3401. https://doi.org/10.18653/v1/2021.eacl-main.297

Yamada, Ikuya, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. LUKE: Deep contextualized entity representations with entity-aware self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6442–6454. https://doi.org/10.18653/v1/2020.emnlp-main.523

Yang, Yue, Artemis Panagopoulou, Marianna Apidianaki, Mark Yatskar, and Chris Callison-Burch. 2022. Visualizing the obvious: A concreteness-based ensemble model for noun property prediction. In *Findings of the Association for Computational Linguistics: EMNLP 2022*. https://aclanthology.org/2022.findings-emnlp.45

Yang, Zhilin, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R. Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32, Curran Associates, Inc. https://proceedings.neurips.cc/paper/2019/file/dc6a7e655d7e5840e66733e9ee67cc69-Paper.pdf

Yu, Mo and Mark Dredze. 2014. Improving lexical embeddings with semantic knowledge. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 545–550. https://doi.org/10.3115/v1/P14-2089

Yu, Zheng, Haixun Wang, Xuemin Lin, and Min Wang. 2015. Learning term embeddings for hypernymy identification. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1390–1397.

Zanzotto, Fabio Massimo, Ioannis Korkontzelos, Francesca Fallucchi, and Suresh Manandhar. 2010. Estimating linear models for compositional distributional semantics. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1263–1271. https://aclanthology.org/C10-1142

Zhang, Jinpeng, Baijun Ji, Nini Xiao, Xiangyu Duan, Min Zhang, Yangbin Shi, and Weihua Luo. 2021a. Combining static word embeddings and contextual representations for bilingual lexicon induction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2943–2955. https://doi.org/10.18653/v1/2021.findings-acl.260

Zhang, Xinsong, Pengshuai Li, and Hang Li. 2021. AMBERT: A pre-trained language model with multi-grained tokenization. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 421–435. https://doi.org/10.18653/v1/2021.findings-acl.37

Zhang, Yizhen, Minkyu Choi, Kuan Han, and Zhongming Liu. 2021b. Explainable semantic space by grounding language to vision with cross-modal contrastive learning. In *Advances in Neural Information Processing Systems*, volume 34, pages 18513–18526, Curran Associates, Inc. https://proceedings.neurips.cc/paper/2021/file/9a1335ef5ffebb0de9d089c4182e4868-Paper.pdf

Zhang, Zhengyan, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451. https://doi.org/10.18653/v1/P19-1139

Zhu, Yukun, Ryan Kiros, Richard S. Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, pages 19–27. https://doi.org/10.1109/ICCV.2015.11