Attention-based Neural Networks Encode Aspects of Human Word Sense Knowledge

A Cognitive Science Honors Thesis

Sathvik Nair

University of California, Berkeley

May 2020

Advisors: Dr. Stephan Meylan and Professor Mahesh Srinivasan

Language and Cognitive Development Lab, University of California, Berkeley

Second Reader: Professor Steven Piantadosi

Computation and Language Lab, University of California, Berkeley

Abstract

How humans understand variation in a word's senses is key to explaining the structure of the lexicon, but formal models categorizing word senses like WordNet (Miller et al., 1990) do not capture this cognitive phenomenon. Our work specifically determines if neural models that make use of attention to represent words in context, or word embeddings, can encode human-like distinctions between word senses, particularly for polysemous (semantically related) and homonymous relations (semantically unrelated) between senses. We collect data from a behavioral web-based experiment, in which English speakers provide judgements of the relatedness of multiple senses of several words. We compare human judgements for these senses to contextualized word embeddings from BERT (Bidirectional Encoder Representations from Transformers, Devlin et al. (2018)), run on a corpus tagged with WordNet senses (Miller et al., 1993). We demonstrate human participants have shared intuitions about the relatedness of word senses, and BERT embeddings often capture homonymous relationships between word senses, but fail to do so for highly polysemous cases, most notably metaphorical polysemy. Through our work, we present an empirical justification that attention-based neural models of word representation can be used to provide a cognitive backing for formal models of word sense.

1 Introduction

Natural language is a critical aspect of human cognition, in that proficient speakers are able to effortlessly combine words to express complete thoughts. When describing a new situation, it is communicatively efficient for speakers to reuse a given string rather than creating a new word form entirely, if real-world context allows speakers of a language to disambiguate between different meanings for the same string (Piantadosi et al., 2012). This process gives way to a key feature of natural language: different possible meanings of words can be represented by a set of typically mutually exclusive categories, or senses. The lexicon can therefore be defined as *generative*(Pustejovsky, 1991), so new meanings of words can be inferred from a core mental representation. How these meanings vary, and how their distinctions are learned, are key problems in understanding natural language. A phenomenon known as polysemy describes the case that the senses of a word are related (Bréal, 1897) to one another semantically. Polysemy is shown to be regular in many situations (Apresjan, 1974), and commonly exhibits the same patterns across multiple word forms. A case of regular polysemy in English is the use of the same word form to represent both an item and a material made from that item, like the words "glass," "fur," and "iron." Conversely, if the senses of a word are not semantically related to one another, and if the same string is used in two unrelated contexts, this is a case of homonymy. A canonical example of homonymy would be "bank," because the word can describe both the side of a river and a financial institution. Dictionaries usually list polysemous word senses under the same entry, and homonymous senses are listed as separate definitions (Rodd et al., 2002).

Because words with multiple senses occur so frequently in the lexicon, how humans understand word sense is important for language learning and processing. Proficient speakers of a language are able to determine which senses of a word apply in a certain context, effectively resolving lexical ambiguity (Klepousniotou, 2002). Learning to make sense distinctions is also a key stage in language acquisition. Once a child learns one sense of a word, they are able to use the relatedness between the senses and real-world context to infer new senses of that word, as demonstrated by Srinivasan et al. (2019). This is critical evidence that the lexicon can be extended to novel scenarios from a core understanding of certain tokens. Although the emergence of different meanings of the same word is a cognitive process, many formal theories of word sense have their basis in lexicography, suggesting a binary between polysemy and homonymy. WordNet (Miller et al., 1990), the gold standard inventory of word senses, does not distinguish between polysemy and homonymy (Veale, 2004), despite how frequently it is used (Budanitsky and Hirst, 2006a; Hearst, 1998; Jurafsky and Martin, 2014; Pedersen et al., 2004; Vincent-Lamarre et al., 2016). Due to contradictory experimental evidence (Lopukhina et al., 2018) pointing to both separate and core representations of word senses in the mental lexicon, specific details about the underlying cognitive basis of word sense are less clear. One general theory about how humans view word sense is that the different levels of relatedness between senses point to a gradient between polysemy and homonymy, rather than a binary (Apresjan, 1974; Crossley et al., 2010). Our work seeks to determine if computational models of word representations that have led to record performance on many tasks in natural language processing (NLP) (Devlin et al., 2018; Peters et al., 2018) can also shed some light on on the architecture of the lexicon through comparing their data with human participants.

These models of word representation, or word embeddings, encode words in context as numerical vectors. The distributional hypothesis (Harris, 1954) states that the meaning of a word can be inferred by its lexical context, or other words in proximity to it, and word embeddings reflect this idea through computational models. Mcdonald and Ramscar (2001) found that lexical context is critical in distinguishing semantic similarity among human participants as well, so we have reason to believe that word embeddings are approximating some cognitive features of language. Previous work (Dumais et al., 1988; Mikolov et al., 2013), encode an individual token using a single vector which contains information about all of its senses, which makes it difficult to identify how a single word sense is represented in their vector space. Recent models of word embeddings (Devlin et al., 2018; Peters et al., 2018) make use of sophisticated neural network architectures like attention, dynamically creating different vectors depending on different lexical contexts of a word.

These models, or contextualized word embeddings, may be able to represent word sense much more effectively than previous methods. By analyzing patterns in the vector space representations of different senses of words from contextualized embeddings, it is possible for us to improve upon existing formal models of word sense taking inspiration from NLP techniques.

Our work revolves around determining if contextualized word embeddings can be used to lay the foundations for a cognitively informed, data-driven model of word sense. These models are performing well on NLP tasks, but research on how they represent linguistic features is rather limited, and research on comparing them to human judgements even more so. Because they represent individual uses of a word in context as a point in a high dimensional vector space, we can relate this to the idea of an exemplar space (Tversky and Kahneman, 1973), where vectors for each word type represent exemplars. Additionally, because these vector space models encode semantic relatedness (Dumais et al., 1988), they may be able to represent polysemous and homonymous senses continuously. However, before we can test this theory and evaluate contextualized word embeddings from a cognitive perspective, we must first determine whether they encode human-like distinctions between word senses. We assess human intuitions about the relatedness between different meanings of several words by collecting judgements through a web-based experiment. We then compare the experimental data to representations from contextualized word embeddings for the same set of words. This is done through extracting embeddings for word tokens in Semcor, (Miller et al., 1993) a corpus that has been annotated with word senses from WordNet, and analyzing the outputs of the model. If the outputs from the models compare well with human knowledge about word senses, so if they show evidence they may accurately represent phenomena like polysemy and homonymy, we can claim attention-based contextualized word embedding models can be used as a starting point for a new ontology of word sense.

2 Background

We begin by discussing cognitive theories of polysemy and homonymy and describe how formal models often fall short. Then, we examine both experimental work on how humans understand word senses, and speculate why contextualized word embeddings may provide insight into cognitive models of word sense. Finally, we provide a framework that could define the lexicon using our knowledge of human cognition and these models, and describe expected outcomes under our hypotheses from both our experimental task and modeling work.

2.1 The Structure of Polysemy and Homonymy

Homonymy and polysemy have already been mentioned as descriptions of the ways in which the same word can be used to express different meanings. Polysemy, the case where multiple senses of a word are related to one another semantically, itself is highly regular and divided into different categories. Additionally, certain words may have both homonymous and polysemous senses. “Bank”, for instance, has the polysemous senses referring to a financial institution and a store of goods (as money is stored in financial institutions) but both senses are homonymous with respect to the use of “bank” to describe the side of a river. Although polysemy occurs across different parts of speech (“shower” is both a noun and a verb), we restrict our analysis to variation of word sense within parts of speech, as distributional statistics allow a word’s part of speech to be easily inferred in context (Søgaard, 2010).

One of the simplest applications of polysemy is expressing a synecdoche. For instance, “door” could be used to refer to both a piece of furniture, and the doorway that contains it. This relationship, or metonymy, can be extended further to many other semantic domains, including, but not limited to: animals for their meat, items for their contents, and artists’ names for their works. Metonymic senses and metaphorically induced senses are often closely related (Lakoff and Johnson, 1980). For example, “raise” can be used to refer to physically lifting an object, or to describe any quantity increasing. Because of varying degrees of relatedness among word senses, cognitive theories of word sense propose a gradient between polysemy and homonymy (Apresjan, 1974).

2.2 The Documentation of Word Sense and its Applications

One of the most familiar collections of word meanings is a dictionary, categorizing the definitions of different words. Dictionaries have separate entries for homonyms and shared entries for polysemes(Rodd et al., 2002). This suggests a binary between polysemy and homonymy instead of a gradient, which places lexicography in contrast with cognitive theories of word sense. The lexical database WordNet (Miller et al., 1990) extends the concept of a dictionary, grouping words and their senses into semantic categories, or synsets, that express similar concepts. WordNet also enforces a structure to connect its word senses, based on whether one sense is an instance of another sense (i.e. "dog" and "animal" are linked to one another).

Besides being a useful tool linking language and cognition (Vincent-Lamarre et al., 2016), WordNet has also been influential in natural language processing, making tasks like word sense disambiguation and induction possible at scale (Jurafsky and Martin, 2014) and serving as a gold-standard inventory of word senses. Word sense disambiguation (WSD) refers to inference of the sense of a word used in a particular context, and is done through supervised learning. Word sense induction (WSI), on the other hand, is an unsupervised task, involving inferring which uses of a word correspond to different senses. WSI does not use prelabeled WordNet senses, but WordNet is influential in designing and evaluating several algorithms for the task (Amrami and Goldberg, 2019; Khodak et al., 2017)).

Despite the ubiquity of WordNet (Budanitsky and Hirst, 2006a; Jurafsky and Martin, 2014; Yuan et al., 2016; Zhou et al., 2019), the resource is not without its problems. The dataset does not reflect senses known to many speakers, and it does not account for novel senses of a word. For instance, an existing word can be adapted for use in slang, vernacular, or specialized technical settings. As a specific example, WordNet does not account for “text” being used to refer to digital messaging. WordNet also often has too many senses for individual words (“take” has 44 senses alone) and many of the senses it does define are archaic. This leads the dataset to be too fine-grained for many NLP tasks (Amrami and Goldberg, 2019). In some cases, WordNet can also be coarse-grained. Looking at the senses for “table,” there are three senses describing a piece of furniture, but WordNet does not list a sense for the metaphorical use of the word (“This one of our options on the table.”) Finally, although WordNet does link senses of words with different forms, it does not consider the relationships between senses of the same word. This means that WordNet neither distinguishes polysemy and homonymy, nor does it provide relatedness metrics between senses of a word, even if humans perceive some senses to be more related than others (Budanitsky and Hirst, 2006b) and are able to easily distinguish polysemous and homonymous senses.

2.3 Experimental Studies on Polysemy

Polysemy has received a significant amount of attention in the psycholinguistics literature, but many experimental studies conducted show contradictory evidence, specifically relating to whether senses are stored separately or together. Klein and Murphy (2001) point out that senses are more likely to be stored separately and states that there are limited differences between polysemy and homonymy through a sensicality judgement task. Their work also makes a key point about how related senses of a word are not necessarily semantically similar, offering a refutation to the theory of a core representation of word sense. For instance, one might refer to an electronic copy of a work of literature as a “book” despite the fact that it is not conceptually similar to a physical book. This distinction between similarity and relatedness arises for both homonymy and polysemy, especially metaphorical polysemy. Frazier and Rayner (1990) presented participants with pairs of sentences for both homonymous words and polysemous words. Participants’ reaction time was measured as they read pairs of sentences like “Of course the pitcher pleased Mary, being so elegantly designed,” and “Of course the pitcher pleased Mary, throwing many curveballs.” This was done so the first part of the sentence was the same, but the participants received disambiguating information in the second part of the sentence to show which sense of the target word ("pitcher" in this case) was being referred to. Eye tracking evidence from the study suggests that homonymous senses are perceived as unrelated to polysemous ones– participants' reaction time was much longer when they were processing pairs of sentences where homonymous senses were used than polysemous senses. In addition to comprehending homonymy different from polysemy, humans may understand different polysemous relationships based on how much of the senses of the target word overlap with one another (Klepousniotou et al., 2008). We have described some behavioral evidence for how humans process word sense, but many of these patterns also persist at the neural level. EEG work shows differences in N400 event-related potentials when participants reacted to polysemous and homonymous stimuli in a lexical decision task(Klepousniotou et al., 2012; MacGregor et al., 2015). This result is significant because it points to evidence for polysemous words having a core meaning in the mental lexicon, and separate storage of homonymous senses, due to how the brain physically reacts when presented with these stimuli.

The experiments we presented focused on words in English, perhaps because much of the early lexicographical work showing regular polysemy was done on a select few languages, namely English (Lakoff and Johnson, 1980) and Russian (Apresjan, 1974). Many of the patterns shown by polysemous words were aggregated and analyzed by Srinivasan and Rabagliati (2015), showing that systematic polysemy persists across multiple languages, but the words that lend themselves to these patterns like metonymy differ from language to language. From this work, we find that conceptual structure is important to polysemy, but individual senses themselves are inferred as conventions of the language they are used in.

Word embeddings are trained on corpora that have evidence of linguistic conventions, or how certain senses are used, so they may offer some potential for explaining polysemy. However, studies that aim to provide a computational cognitive model of word sense, backed with experimental data, are few in number. The work of Lopukhina et al. (2018) combines a behavioral experiment and an embedding-based corpus analysis, but offers fixed categories for participants to categorize senses of a word, as opposed to extracting continuous metrics describing how related participants say they are; the latter methodology would be more consistent with existing theories of how humans understand word sense. This study, as well as many other experimental studies with polysemy, show whether an example sentence containing a word uses a certain sense of that word better than other senses, rather than how related the senses are themselves. Additionally, Lopukhina et al. (2018) present a computational model, but the way this model creates representations of words is much less sophisticated than contextualized models (Ethayarajh, 2019).

2.4 Word Embedding Methods: From word2vec to BERT

People often use linguistic context to effectively deduce the sense of a word (Mcdonald and Ramscar, 2001); computational models have been developed to imitate this ability. Many of these models are rooted in the distributional hypothesis, that a word is defined by its lexical context, (Harris, 1954). One of the first models to take advantage of this claim was latent semantic analysis, which created low-dimensional word representations by factorizing a matrix of co-occurrence statistics (Dumais et al., 1988). The seminal work of Mikolov et al. (2013) uses a neural network to create distributed, low-dimensional representations of words as individual vectors, where information about the context of the words is encoded in the vector. Words that appear in similar contexts would have similar vectors, or embeddings. Earlier methods, namely Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014), are able to capture more semantic information than count-based models (Lilleberg et al., 2015), because they learn the relationships between words through a shallow neural network. However, they do not represent homonymy and polysemy accurately (Arora et al., 2018), as all the senses for a word are represented as a single vector without further architectural elaboration (e.g., Sense2Vec, (Trask et al., 2015)). For example, the vector for the word *bank* could contain information about both the side of a river and a financial institution, because embeddings resulting from these methods were not contextualized.

One promising direction for research is using more sophisticated representations of linguistic context. With contextualized word embeddings, the vectorized representation of the word “table” would be different in the sentences “The results for the experiment are presented in the table,” referring to a set of data and “They served many meals on the sturdy wooden table,” referring to a piece of furniture. These models could yield more cognitively useful representations of words, because individual uses of words could be represented as exemplars in the embedding space.

Recent advances in neural network architecture have led to models that output contextualized representations of words based on the sentences they are used in (Devlin et al., 2018; Peters et al., 2018). Rather than just predicting whether certain words belong in the proximity of other words, these methods make use of probabilistic language models that are used to encode contextual information into a word’s vector by way of both deeper, bidirectonal neural networks as well as novel architectures. Peters et al. (2018) use a character-based recurrent language model, and Devlin et al. (2018) use an attention-based architecture (described in detail below), which captures more semantically relevant information than recurrent neural networks. The use of contextualized word embeddings to represent textual data has drastically improved on previous results on downstream NLP tasks, such as textual entailment, semantic role labeling, coreference resolution, and sentiment analysis (Peters et al., 2018) as well as question answering, machine translation, and named entity recognition (Devlin et al., 2018). This is because words are represented much more richly with contextualized embeddings than Word2Vec and GloVe, making use of information from the text the words are used in when they are inputted into the models that complete these tasks. This progress demonstrates that contextualized representations may have access to aspects of linguistic structure not captured by previous models.

Embeddings derived from BERT (Bidirectional Encoding Representations from Transformers, (Devlin et al., 2018)) are more powerful than embeddings created with other techniques (Ethayarajh, 2019). Using principle components of representations from only the lower layers of the BERT network, Ethayarajh (2019) demonstrates results on NLP tasks similar to those when static embeddings (Mikolov et al., 2013; Pennington et al., 2014) were used, suggesting that the final result of the outputs from BERT, a summation over the later layers, are more accurate than those of static embeddings. Our work involves word senses, so we specifically choose to work with BERT because its performance on word sense disambiguation is superior to other contextualized embedding models (Wiedemann et al., 2019).

The main reason BERT has been successful is that it makes use of a novel neural network architecture known as the transformer (Vaswani et al., 2017). Before the transformer architecture, attention was always used in combination with recurrent layers (Bahdanau et al., 2014), while transformers only use attention mechanisms, "dispensing of recurrence and convolutions entirely"(Vaswani et al., 2017). Attention allows weights for each word in a sentence to be influenced by every other word, allowing the neural network to learn which other words should be "attended to" when creating the vector for a particular word. One reasons transformers perform better on NLP tasks than recurrent models, such as those used by (Peters et al., 2018), is that disambiguating cues are at different distances for different words. Regarding generating word embeddings, this means that tokens that are key to learning the in-context meaning of the target word are not necessarily close to it in the sentence. This is important because long-range dependencies between tokens are common in natural language texts. Learning these dependencies is difficult for recurrent networks, but transformers are able to model them more efficiently (Vaswani et al., 2017). The progress on NLP tasks achieved by using BERT representations leads us to ask the question of whether BERT encodes useful linguistic information.

2.5 BERT and Polysemy

Contextualized word embeddings, namely BERT, seem to be able to represent more information about word sense than static embeddings, but how they do so is an active area of research we hope to contribute to. There are several approaches that can be taken to evaluate how BERT takes polysemy and homonymy into consideration. Many of these rely on the idea of a *probe*, a downstream classification task used to evaluate whether deep neural language models encode linguistically relevant features (Hewitt and Liang, 2019). The work of Hewitt and Manning (2019) and Coenen et al. (2019) focus on evaluating the neural network architecture of BERT.

Significant progress has been made regarding how BERT encodes syntax, but finding semantic information in contextualized embeddings is still an active area of research. Through a "structural probe" that applies linear transformations to contextualized embeddings of words in a sentence, Hewitt and Manning (2019) are able to reconstruct the dependency parse tree of a sentence. Further use of probing techniques by Jawahar et al. (2019) demonstrates that syntactic information is represented at the middle layers of BERT's architecture, and semantic information, such as tense, is represented at higher level layers. Probing the attention matrices of BERT, shows that attention vectors for pairs of words exhibit dependency relations between them, advancing the claims of Hewitt and Manning (2019) (Coenen et al., 2019). Neural models do not have a definitive intuition of which words are important in a sentence, unlike humans, but attention, at a higher level, is able to quantify these relations of importance between words. Because of the importance of context in human inference of a word's sense (Mcdonald and Ramscar, 2001), we can extend this work into understanding if contextualized embeddings effectively represent word senses. Evidence from Coenen et al. (2019) suggests that word senses may lie in a lower-dimensional space, via both exploratory analysis and evaluation on WSD tasks. The authors give examples of words whose embeddings are clustered around their senses, and homonymous senses are clustered further apart from polysemous ones. This is especially key considering our approach with polysemy and homonymy, but the authors make their claims about word sense based on heuristics inferred from the context of the word in question, and neither investigate them empirically nor mention how BERT represents polysemous and homonymous senses.

The work we discussed points to BERT being a more effective representation of word sense than previous word embedding techniques. This could be due to the fact that word sense may lie in a lower-dimensional subspace of the BERT embedding space. If this is the case, then the nature of how models like BERT represent word sense becomes much more interpretable; it will be easier to derive sense relations from transformations in the vector space. We strive to take this work further by trying to empirically determine if BERT distinguishes homonymous senses of a word more clearly than its polysemous senses, comparing BERT's results to both English speakers' intuition about words with multiple senses as documented in WordNet, and determining if a lower-dimensional subset of the BERT vectors is consistent with human judgements.

2.6 Present Work

In the present work we propose a cognitive, data-driven model of word sense based on exemplar theory. This theory proposes humans categorize novel stimuli by comparing them to examples stored in their memory (Tversky and Kahneman, 1973). Given the productivity of language, fluent speakers are capable of interpreting novel linguistic utterances by comparing them with similar sounds, phrases, or words. Exemplar theory has often been used to describe this aspect of human language processing (Batali, 1999; Bod and Cochran, 2007). With our proposed model, individual tokens representing words are the exemplars, and a word sense can be inferred through the relationship between sets of tokens, forming a cluster in this high-dimensional space. Based on a token's position in relation to other tokens, it can be categorized as a use of a particular word sense, or part of the "sense cluster." As a result, the relationship to both tokens and senses, as well as senses to one another, is implicit in the geometry of this vector space.

Prototypical sense usages, which are useful in creating a core mental linguistic representation, can be found near the centroid of the cluster of tokens corresponding to a sense. Token clusters that are closer together in this space are representative of word senses that are closely related to one another. Clusters for homonymous senses can be easily distinguished from one another, while clusters for polysemous senses often overlap (Figures 9 and 10). Our goal is to conduct an experiment collecting prototypical judgements of sense relatedness from human participants, then test if exemplars derived from contextualized word embeddings capture these same relations.

First, we seek experimental evidence that words have a stable set of relationships between their senses; that language users share a similar set of intuitions about which senses of words are more related to one another (Budanitsky and Hirst, 2006b; Lopukhina et al., 2018; Mcdonald and Ramscar, 2001). We select a set of English words with polysemous and homonymous senses from WordNet, and collect speakers' judgments of relatedness using a web experiment. Second, we investigate whether representations derived from recently developed NLP models – contextualized word embeddings from attention-based neural networks (BERT) – are similar to the relatedness judgments gathered from participants in the experiment. These models may encode significantly more sense-specific information about words than static word embeddings; our goal is to find whether they encode general distinctions between word senses. If in fact representative, the continuous, high dimensional representations from these models may serve as a significant advance over existing discrete, symbolic resources approximating the lexicon.

3 Experiment

We perform an experiment where participants use a web interface to gauge the relatedness of senses for different words as documented in WordNet. Even if we believe contextualized word embeddings are capturing more information about word sense than previous models, there are not many ways they can be evaluated, besides performance on downstream NLP tasks. Additionally, WordNet senses alone are not enough to capture people's internal representations of words because they do not describe relations between word senses, but we are able to extract relations between embeddings from BERT's vector space model. Thus, we need human-labeled data consisting of relatedness judgements between WordNet senses to test our hypotheses described in the previous section.

3.1 Methods

We are able to efficiently obtain relatedness judgements between word senses by having participants place tokens representing different senses of a word in a 2-dimensional space (Figure 4). This follows the work of Goldstone (1994), which notes using spatial judgements of psychological similarity is more efficient than asking for simple numerical ratings of the strength of the relationship between two senses of a word. First, we can efficiently compute pairwise ratings. When presenting participants with n senses for a given word, we can obtain $\frac{n(n-1)}{2}$ ratings (each sense is compared with $n-1$ other senses, and we divide by 2 because the judgements are symmetric), and are therefore able to obtain ratings for more words per participant than if we ask for individual relatedness judgements. Second, this ensures that the distances between word senses are in a metric space, such that they can be compared easily to the BERT vectors. We revisit the implications of the assumption that senses lie in a metric space in the discussion. Therefore, our experimental design allows us to find perceived relationships between prototypical views of word senses.

3.1.1 Selection of Stimuli

We choose 30 possible words to serve as stimuli for the task. We needed to ensure that each sense had an adequate amount of data(at least 10 instances) that was annotated in Semcor (Miller et al., 1993) to make a comparison to BERT embeddings. Additionally, we needed to have the set of stimuli cover both polysemous and homonymous word types. To account for participant subjectivity in token placement, we would need to normalize measurements. This means that no information would be gained from just placing two senses on a canvas, so we choose words with 3 senses or more.

Lemmas in Semcor 20 of the lemmas were selected using estimates of their entropy (relative unpredictability of senses) calculated from their frequency in Semcor. The entropy of a given lemma L in Semcor is expressed as $-\sum_{s \in L} \frac{c_s}{c_L} \log(\frac{c_s}{c_L})$, where s is a sense and c corresponds to a frequency, or count in the corpus. We did this such that we could assess whether the performance of BERT when compared to human judgements would be worse for more complex words. Once we computed entropy, we selected stimuli based on a combination of the following features: part of speech, frequency, number of senses, and entropy.

Part of Speech	Number of Lemmas
Verb	105
Noun	48
Adverb	6
Adjective	5

Table 1: Occurrence in Semcor for High to Medium Entropy Lemmas with 3 or More Senses

From Table 1, we limited the selected stimuli to 10 nouns and 10 verbs because other parts of speech were not adequately covered. Because of the skewed nature of the distribution (Figure 1), we ensured that 10 of the words selected would have 3 senses, and 10 would have more than 3 (4 to 7).

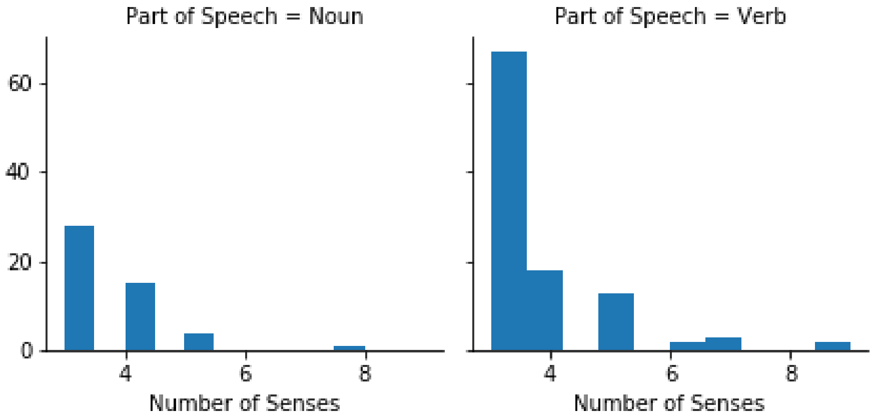


Figure 1: Number of WordNet Senses With Respect to Lemmas in Semcor

From Figure 2, we ensured that half the words would have an entropy of below 1.5 and half above 1.5. When selecting words, we place an emphasis on higher frequency words to ensure that there would be sufficient data representing them when we extract BERT embeddings. All the words we selected from Semcor are referenced in Table 4 in the Appendix.

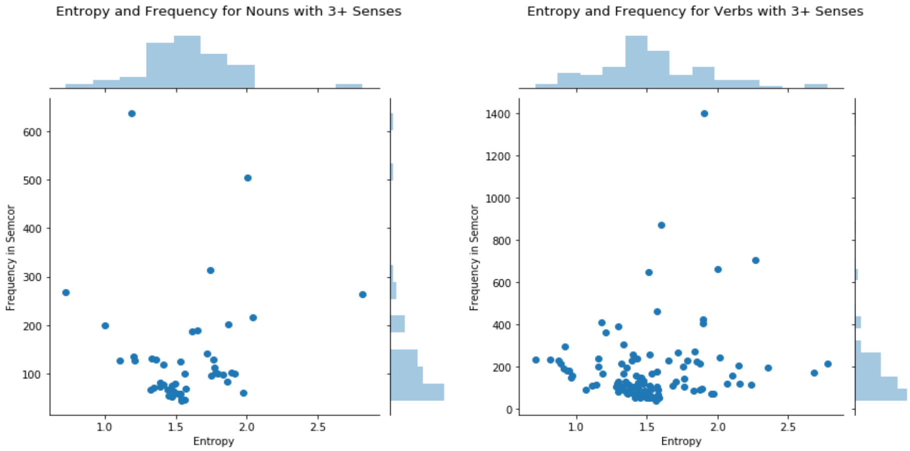


Figure 2: Distribution of entropy and frequency for nouns and verbs with 3 or more senses in Semcor

Words showing Patterns in Polysemy We selected 6 lemmas from Srinivasan and Rabagliati (2015) because they showed specific patterns in polysemy that persisted across a set of 14 languages. These patterns were drawn from a broad review of literature, so did not necessarily pick out senses documented in WordNet, which is what our analysis focuses on. Additionally, many senses that Srinivasan and Rabagliati (2015) covered that were in WordNet had an insufficient amount of tokens in Semcor (below 10 across all senses); this would present difficulties in the modeling tasks. We select the following word types representing systematic patterns in polysemy from the paper: `book.n` (container for representational contents), `glass.n` (material for artifact), `door.n` (figure for ground), `school.n` (building for people in the building), `face.n` and `heart.n` (body part for object part).

Homonymous Words We selected 6 additional words having a high degree of homonymy, so one of the senses would not have a strongly systematic relationship with the others. [1] There were often only two senses with at least 10 tokens in Semcor, the same threshold used to select words from Srinivasan and Rabagliati (2015) and Semcor. Because we would need to normalize distances

[1] There is no gold standard resource specifying whether words are homonymous or polysemous (dictionaries often disagree). The goal here was to choose a set of words that were likely to have both polysemous and homonymous meanings to make it easier to discern differences among human judgements.

to account for individual differences in placing the senses, we added a third sense so subjects could anchor their judgements, and place the two polysemous senses close to one another and the homonymous sense further away. Based on frequency in Semcor, we selected `foot.n`, `table.n`, `plane.n`, `degree.n`, `right.n`, and `model.n`.

Data Preparation For each sense not used in the training task, we provided a definition from WordNet and randomly selected an example sentence from Semcor to be shown with the interface. Once we found the sentences, we inspected them to ensure that they were concise, adequately explained the definition of the sense, and did not contain objectionable content.

3.1.2 Participants

Human judgements were collected in a web-based experiment on 105 undergraduates, compensated with academic credit through UC Berkeley's Research Participation Program. Upon accepting a consent form, participants were given the option of entering demographic information, and were required to input their experience with English and other languages. This information was used to ensure that proficient English speakers were working on the task. We used data from participants who reported they used English at least 50% of the time; one participant was excluded using this criterion. They were told that they would be participating in a study on language and cognition, they would be compensated with academic credit, and that they would not be penalized if they did not complete the experiment.

3.1.3 Experimental Interface

Once participants provided information about their linguistic experience, they were directed to a site that hosted the experimental interface. This site provided participants with a second consent page, that briefly described the nature of the task as arranging a group of definitions of a set of words depending on how related they perceived the definitions were.

Before doing the task, participants received a more specific set of instructions. They were notified: “Tokens for closely related definitions should be placed close together, while tokens for definitions that are not strongly related to one another should be placed far apart” (Appendix B). Afterwards, they were given example layouts for how they might place senses for the lemmas “bank.n” and “chicken.n”(Figure 3). Finally, they were told that if they saw senses they believed were equally related, they should place the boxes in a pile such that they were equidistant from one another rather than a line (Appendix B).

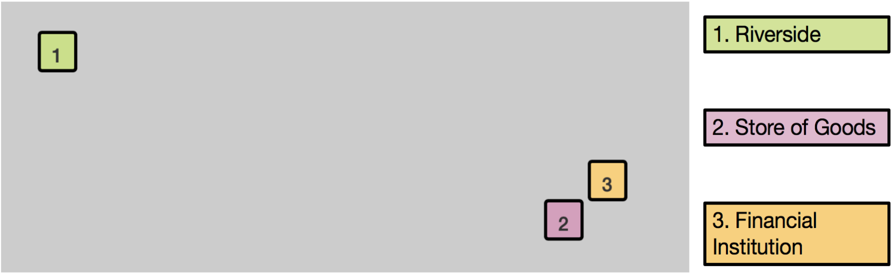
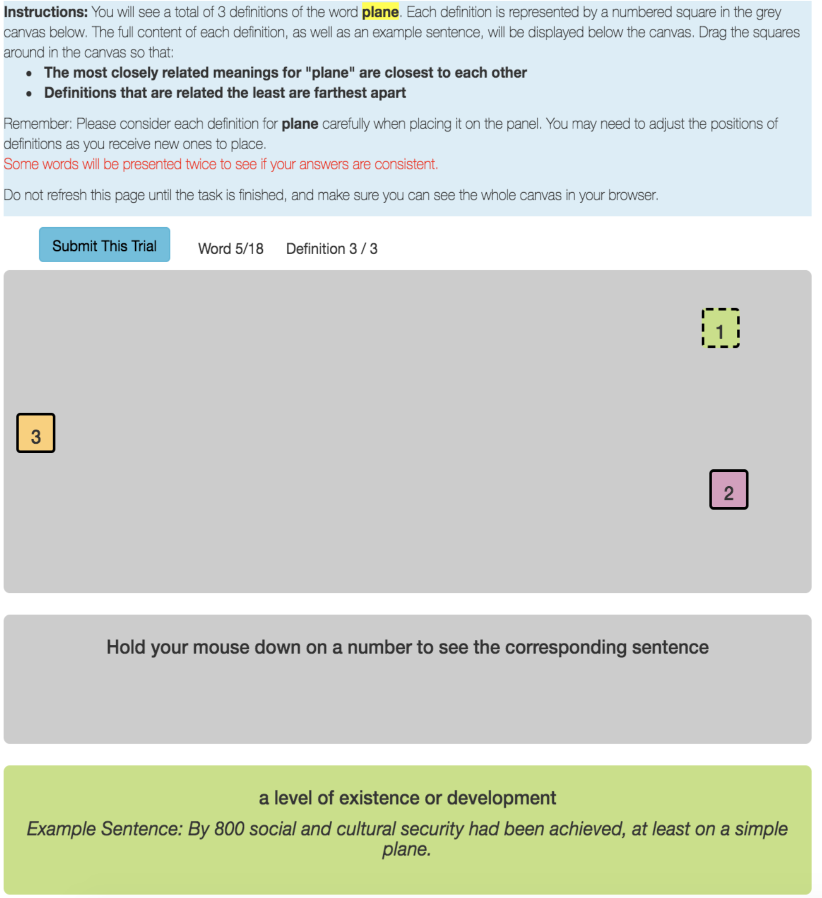


Figure 3: Layout of canvas presented to participants, describing an example placement of senses for `bank.n`

After reading the instructions, each participant completed 18 trials. Participants saw instructions for the task above the canvas, as well as their progress through the experiment on each trial (Figure 4). They were not timed during the task and were encouraged to introspect for as long as necessary about the meanings of word senses. Upon clicking on or hovering over a token, participants could see the sense’s definition in WordNet and an example sentence from Semcor demonstrating how the sense could be used. They were also able to see the definition of a previously placed sense if they dragged the current definition token next to one that had been already placed.



Instructions: You will see a total of 3 definitions of the word **plane**. Each definition is represented by a numbered square in the grey canvas below. The full content of each definition, as well as an example sentence, will be displayed below the canvas. Drag the squares around in the canvas so that:

- **The most closely related meanings for "plane" are closest to each other**
- **Definitions that are related the least are farthest apart**

Remember: Please consider each definition for **plane** carefully when placing it on the panel. You may need to adjust the positions of definitions as you receive new ones to place.
Some words will be presented twice to see if your answers are consistent.

Do not refresh this page until the task is finished, and make sure you can see the whole canvas in your browser.

Hold your mouse down on a number to see the corresponding sentence

a level of existence or development

Example Sentence: By 800 social and cultural security had been achieved, at least on a simple plane.

Figure 4: A sample trial in the experiment

Because the order participants received the senses could affect their judgements of relatedness (judgements are order dependent), the experimental interface allowed them to rearrange the tokens upon receiving a new word. For instance, a participant may have initially believed that the following senses for `case.n` might be similar: “a comprehensive term for any proceeding in a court of law whereby an individual seeks a legal remedy” and “a special set of circumstances.” However, once they were presented with the sense “an occurrence of something” they might have moved the box corresponding to “a special set of circumstances” closer to the sense they most recently received. We record how many times participants rearranged previously placed senses to determine if subjects took order dependency into consideration.

3.1.4 Training, Shared, Test, and Repeat Trials

The first two of the eighteen trials participants receive are presented as *training* trials to ensure they are familiar with the interface for test trials. The word types for these trials are `bank.n` and `bass.n`, and participants place two polysemous senses and one homonymous sense for both trials on the canvas. We ensure that every participant receives all six words with one homonymous sense and two polysemous senses for the next set of trials These will be referred to as *shared* trials. We provide these stimuli to all participants because we expect that the homonymous sense should be judged as less related to the polysemous senses, consistent with past work, and that people's shared intuitions about the relatedness between word senses may be reflected most clearly in these trials. The next eight test trials are drawn from the remaining set of 26 lemmas, consisting of words from Semcor and Srinivasan and Rabagliati (2015). This is done to cover a broader set of word types which have different sense relationships. Finally, participants are asked to repeat two randomly sampled test trials from the set of eight test trials.

We are able to evaluate the quality of participants' responses in several ways. First, we have a set of training trials, and mention a possible layout for the senses of `bank.n`, one of the training words, in the instructions (Figure 3). Second, we use results from the shared trials to compare participants' judgements for canonically homonymous senses. Third, we gauge the consistency of participants' responses by asking them to repeat trials. We describe how we derive quantitative measures of the latter two criteria below.

3.2 Results

We first summarize some of the metadata from the experiment, then describe the metrics we used to validate participants' ratings, and finally demonstrate evidence of how participants placed homonymous definitions compared with polysemous ones.

3.2.1 Metadata

Participants were told that the experiment would take around 25 minutes, and 99 out of the 105 participants took fewer than 30 minutes to complete it. The distribution of completion times is shown in the Appendix (Figure 24). One of the reasons we preferred a two-dimensional layout is that participants are allowed to change the placement of the word senses. We find that this feature appeared to be used quite often, because on average, participants changed the layout of the tokens after they initially placed them for around half the trials (almost 9 out of 18) ($M = 8.58, SD = 6.12$).

3.2.2 Overview of Metrics

We are able to assess the quality of participants' ratings through analyzing two sets of trials. First, we consider repeat trials, when participants received two words randomly selected from previous trials. Second, we consider the shared trials, performed on a series of words with homonymous senses, where we have data for each participant. For all tasks, we compute a distance matrix for the senses of each word a participant reported. We divide by the largest distance reported to standardize the values in the matrix between 0 and 1. This is because different participants might take up different amounts of space in their arrangement of the sense tokens. We also compare the metrics for both sets of trials to random baselines so we can derive our exclusion criteria (Figures 5 and 6).

3.2.3 Self-Consistency for Repeat Trials

To assess a participant's performance on the repeat trials, we consider how consistently they placed the senses during both tasks. We use the Spearman rank correlation between the participants' reported distances from the first trial they saw the word and the repeated trial. When computed over all words, this metric will be referred to *self-consistency.* We compute the correlation over the upper triangular portion of the participant's distance matrices for the two words they received in both test and repeat trials. We compare the results participants provided to the Spearman rank correlation computed over 1000 sets of two pairs of distance matrices derived from random placements of tokens, one for each repeated word type, to determine our exclusion criteria (discussed below).

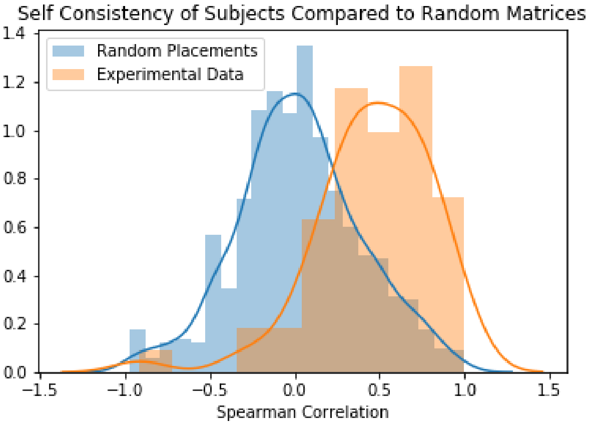


Figure 5: Spearman correlation for each participant's repeat trials compared to 1000 sets of two randomly generated pairs of distance matrices.

Group Consistency for Shared Trials The previous metric, shared consistency, measured how reliable individual participants' responses were, but the metric discussed here compares the performance of one participant to the rest of the group. Based on the previous research reviewed (MacGregor et al., 2015), we provided participants the same set of words with one homonymous sense and two polysemous senses, because we expected judgements of relatedness to be fairly consistent. We compute a hold-one out correlation to assess a participant's performance on the shared trials. This is defined as a participant's *group consistency* and is computed as follows. For each shared word, we compute the average distance matrix when the participant's data is not included. We then take the Spearman rank correlation between these matrices and the participant's responses for each word.

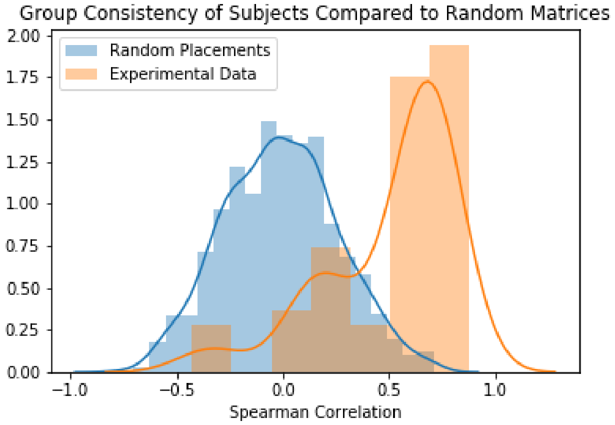


Figure 6: Hold-one out Spearman correlation between each participant's shared trials and the group averages compared to 1000 randomly generated sets of token placements against the group averages

3.2.4 Exclusion Criteria

We construct our exclusion criteria based on group and self consistency from the previous analyses. There are two sources of variance with self consistency– the interval between the first trial and the repeated trial and the number of the senses of the words that were repeated. We exclude participants with a self consistency below 0.2 or a group consistency below 0.4. After applying these criteria, we are able to use data from 94 participants.

3.2.5 Analyses

Once we obtain distance matrices for each word, we compute their average and visualize the aggregate reported distances with multidimensional scaling (MDS) after applying the exclusion criteria described above. Figure 7 demonstrates that participants placed homonymous senses of words further away from one another than polysemous senses. Looking at the entire set of data, we manually tag pairs of stimuli as polysemous and homonymous (none of the test items had any homonymous senses) based on whether there was a clear semantic relationship, and plot the normalized distances individual participants reported in Figure 8. WordNet definitions of the senses used in the shared trials are detailed in Table 5. The median reported distance was 0.881 for homonymous sense pairs and 0.631 for polysemous sense pairs. Because the collected data were not normally distributed (Figure 8), we use a non-parametric test to show a statistically significant difference between the distributions of polysemous and homonymous distances (Mann-Whitney $U = 1959176.5, n_1 = 940, n_2 = 5584, p < 0.001$).

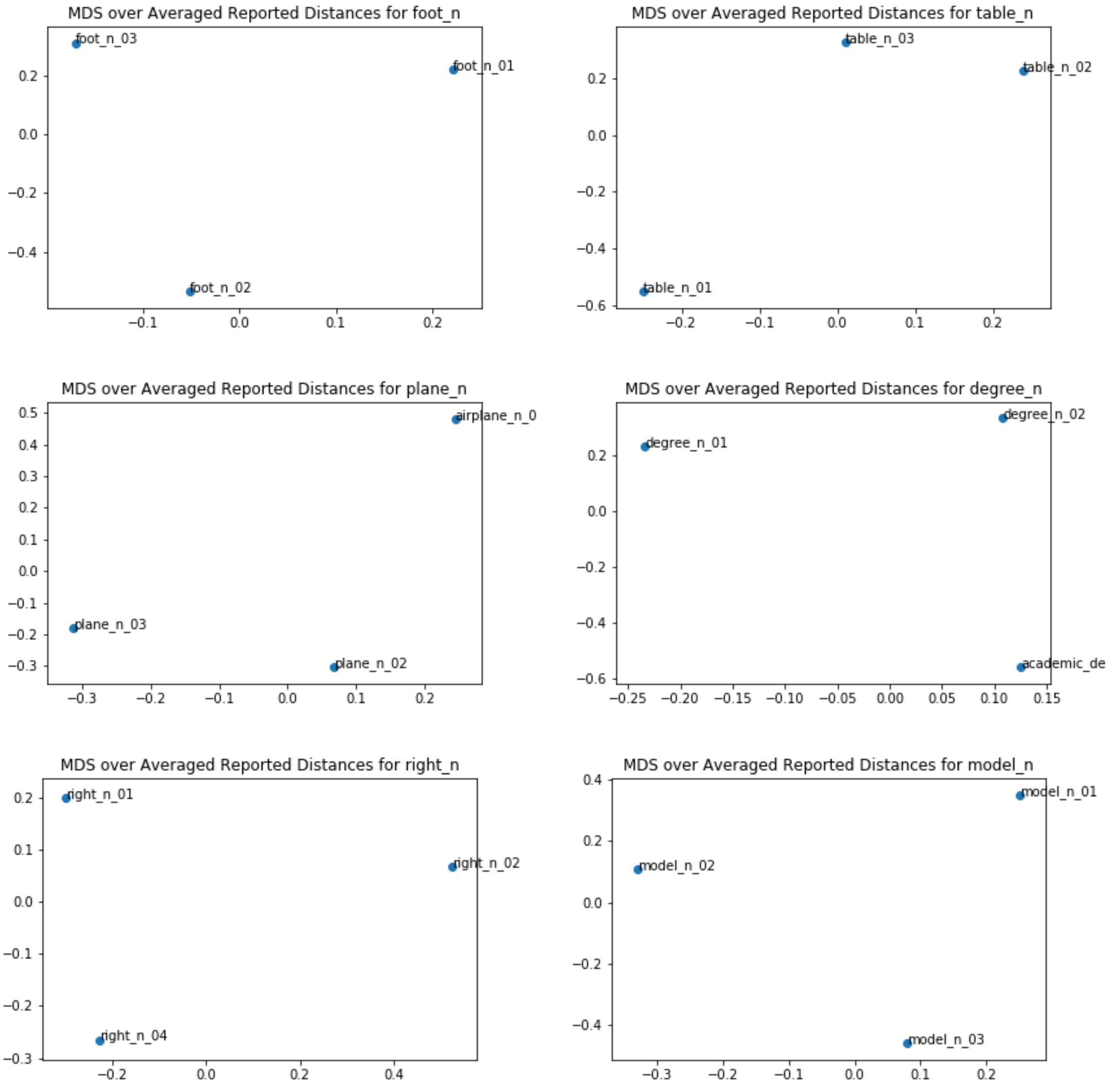


Figure 7: MDS on shared words. Note that homonymous senses like `foot.n.02` (unit of measure) and `table.n.01` (arrangement of data) are positioned further away from the two polysemous senses (Definitions in Table 5).

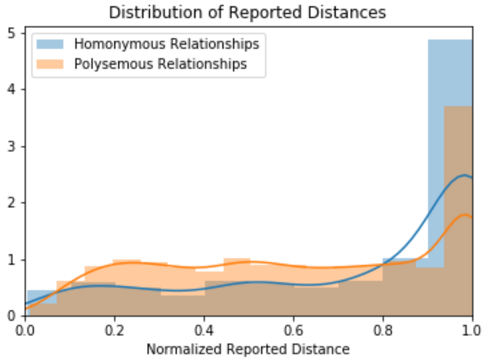


Figure 8: Distributions of normalized reported distances between polysemous and homonymous pairs of word senses

3.3 Discussion

Based on the experimental results, we demonstrate that humans have shared intuitions regarding the relations between of senses of words, where polysemous senses are judged to be more closely related than homonymous ones. We demonstrate evidence that relatedness judgements are consistent because the group (hold-one out) correlations for most participants is noticeably higher than that of the random baseline, as reported in Figure 6. Results on the individual trials from MDS plotted in Figure 7 demonstrate that participants believe that polysemous senses are closely related to one another, but homonymous senses are placed further apart. We confirm this by comparing participants' judgements of the two relationships (Figure 8), and show a statistically significant difference between how they rated polysemy and homonymy. We next turn our attention to whether contextualized word embeddings capture these same relationships between senses.

4 Modeling

We have obtained human data that capture structural patterns in the relationship between word senses for a large set of word types. As a result, we can now assess whether vector space models capture some of this linguistic structure. We analyze how BERT represents words with different senses through an analysis of the corpus Semcor (Miller et al., 1993), which consists of words tagged with WordNet senses. We conduct exploratory analyses and visualizations, compare our data to the experimental results, and assess the performance of a classifier between different senses of individual words to further understand how BERT distinguishes word sense. This classifier is effectively a probe for word sense discrimination, based on the definition from Hewitt and Liang (2019) which describes a probe as a classification task used to predict certain aspects of linguistic structure from potentially black box models of linguistic representation. We discuss how the embeddings were extracted, several techniques we used for exploratory data analysis, and classification tasks. All code and visualizations are available at `https://github.com/sathvikn/bert-wordsense/`.

4.1 Selection of Words

Like in the experimental section, we select words for this analysis based on their entropy in Semcor (Miller et al., 1993). This corpus provides WordNet sense annotations for 235,000 lemmas selected from the Brown corpus (Francis and Kucera, 1979). We use the corpus reader built into the Python Natural Language Toolkit(NLTK) (Bird and Loper, 2004) to access all lemmas in Semcor.

The fact that syntax is represented within BERT embeddings (Hewitt and Manning, 2019) implies that information about parts of speech is also captured, so in our analysis, we focus on polysemy and homonymy where the part of speech is the same.

As in our method of selecting stimuli, we chose to extract BERT embeddings from tokens based on their entropy in Semcor. This is so we can focus on words which do not have a unimodal sense distribution, and analyze whether the amount of uncertainty in a word's sense frequency distribution affects the quality of BERT representations when compared with the experimental data. We find the total number of instances for each sense, and create a multinomial distribution over the senses of each lemma (name, part of speech pair) by dividing the frequency of each sense by the total amount of times that lemma occurs, and compute the entropy over this distribution by the formula in Section 3.1.1. Senses that occur fewer than 10 times in the corpus are excluded from this analysis because it would bias the entropy estimate. We remove stopwords and lemmas with zero entropy (79.3% of all lemmas), and generate embeddings associated with the tokens for each of the remaining 406 types with BERT. For words in the experiment, the threshold is reduced to 4 for word senses that were presented to account for all experimental stimuli that were taken from Srinivasan and Rabagliati (2015) or were homonymous.

4.2 Data Pipeline

We create a data pipeline that takes in a word and a part of speech, referred to as a lemma, or type, and outputs word embeddings from BERT. Once we have a specified lemma, we search for it in Semcor. We then count the number of instances of each sense that lemma contains, and choose senses with 10 or more instances in Semcor. For each sense, we get sentences containing that sense, and tokenize them according to rules specified by the BERT authors (Devlin et al., 2018), such as padding sentences with start and stop tokens. We load a pre-trained BERT model (Wolf et al., 2019) that has 12 layers and outputs a 768 dimensional vector for each word, and run the forward pass on each sentence, extracting the activations that correspond to the word in question. This model was trained on the BooksCorpus (800 million words) Zhu et al. (2015) and English Wikipedia (2.5 billion words) (Devlin et al., 2018). For a single word, the model outputs 12 sets of activations, so to create the final embeddings, we sum the activations of the final four layers, as recommended by the authors of the BERT paper (Devlin et al., 2018). As the annotations are for word forms, the pipeline is able to return representations for different morphemes of a word. For example, if the pipeline received type `indicate.v` as input, it would output word embeddings for sentences containing “indicated,” “indicating,” and so on.

4.3 Exploratory Data Analysis

For exploratory data analysis purposes, we visualize the embeddings in a lower dimensional space using t-distributed Stochastic Neighbor Embeddings (t-SNE) (Maaten and Hinton, 2008). To demonstrate how BERT may cluster instances of senses together, we run a single linkage hierarchical clustering algorithm using cosine similarity as its distance metric. The algorithm begins with individual BERT vectors, and combines them into clusters at higher levels using their cosine similarity, to create a tree-based representation of how a certain word is divided into senses and sense-based clusters.

4.4 Extracting Cosine Distances of Centroids

The exploratory analyses can give us a qualitiative idea of how BERT represents different word senses, so we now take a quantitative approach. This is similar to our analysis for the experimental data where we compared reported distances between pairs of polysemous and homonymous senses (Figure 8). For all 189 sense pairs, we load the embeddings from the pipeline, compute their centroids through adding their values, and report the cosine similarity of the centroid of each pair of senses. We plot the distributions of polysemous and homonymous sense pairs (Figure 11), and conduct a Mann-Whitney test to determine if the cosine distances for homonymous sense pairs and polysemous sense pairs come from the same distribution.

Just like we constructed distance matrices out of the relatedness judgements from the human experiment, we now compute the pairwise cosine distances of the BERT centroids by adding vectors representing uses of the senses of each word in Semcor, and store them in a matrix (Figure 12). This matrix is normalized to be between 0 and 1 by dividing by the largest distance, such that it can be compared with the matrices from the experiment. Similar to how we compared the correlations of the relatedness matrices from the experiment to one another through deriving consistency metrics, we now compare the Spearman correlation between these matrices from the previous experiment and the BERT distance matrices, and compare these correlations to a random baseline to determine if BERT is capturing additional information.

4.5 Probe for Word Sense Discrimination

For each lemma where BERT embeddings are saved (based on the entropy estimates from Semcor), we construct a logistic regression model to classify the senses of individual words. To do this, we train a model to predict WordNet senses from the embeddings for each lemma. We conduct 5-fold cross validation and report both the average accuracy and F1 score(Van Rijsbergen, 1979) across the runs for each word. We also combine the confusion matrices during each iteration of cross-validation, and normalize each item in the matrix by the number of true labels, such that it represents the probability an item was predicted given its true class.

We also fit models for all the embeddings for each word, applying $\ell 1$ regularization ($\lambda = 1$) to shrink the weights of values that are not useful in discriminating the senses toward 0 (Ng, 2004). We do this to test the claim of Coenen et al. (2019) that "word sense information may be contained in a lower-dimensional space." The trained multiclass models output more than one set of weights, because each set of weights is used discriminating one sense from all other senses; for a model trained over s senses, we thus obtain s sets of weights. We identify the nonzero weights from the models trained on the entire dataset of embedding-sense pairs for each type, and then extract the values of the embeddings at these positions in each set of weights. As a result, we are now able to determine if the lower-dimensional embeddings capture an amount of signal similar to those of the full BERT vectors when compared with human relatedness judgements.

5 Results

First, we describe results from visualizing BERT embeddings with dendrograms and t-SNE plots. Afterwards, we compare the distances reported by BERT to the judgements from the experiment. Finally, we evaluate the human relatedness judgements against classification accuracy from the probing task.

5.1 Visualizing BERT Embeddings of WordNet Senses

We show a clean separation between homonymous senses in the dendrograms and scatterplots for items given in the shared task (Figure 9).

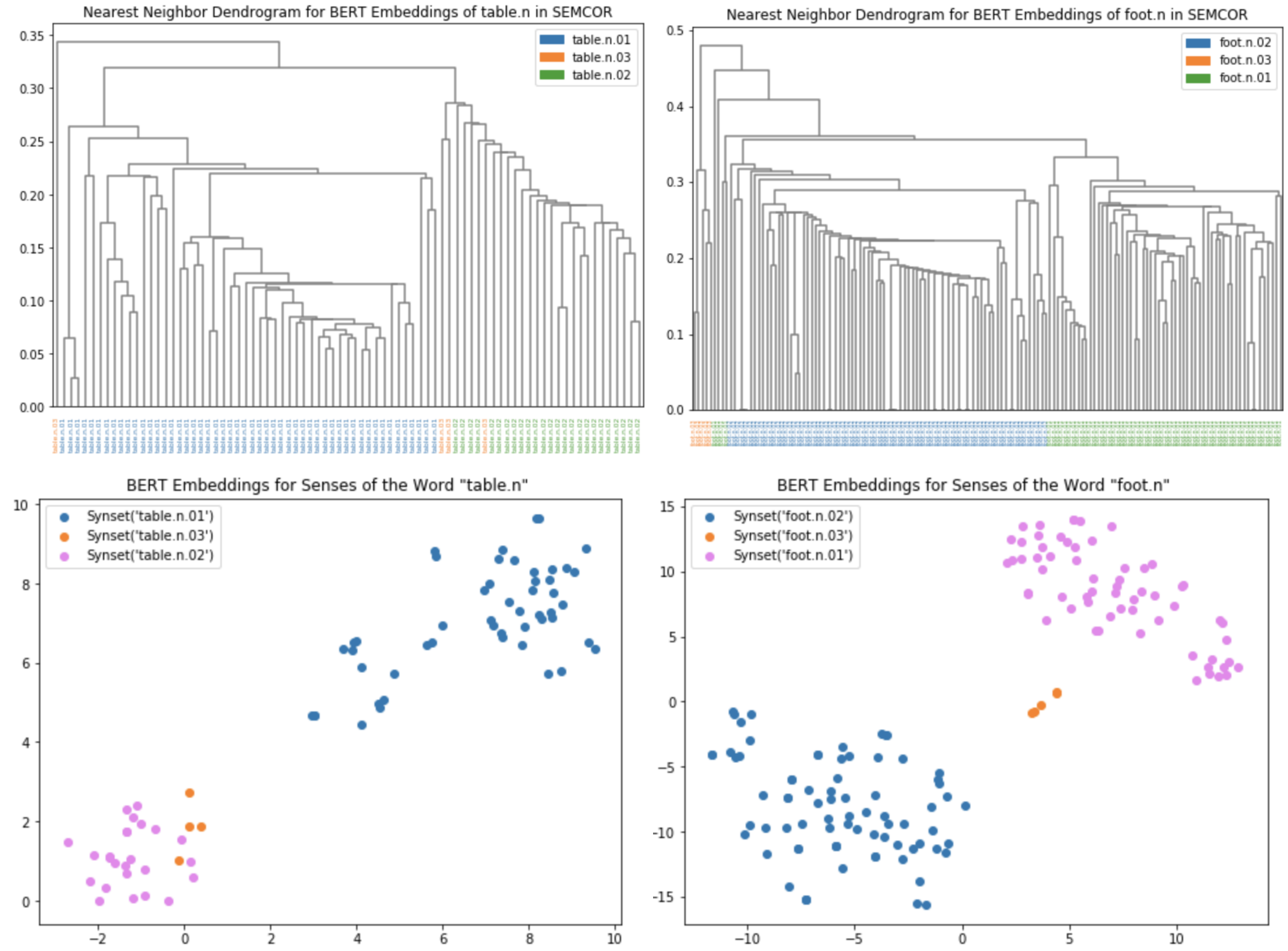


Figure 9: Dendrograms and t-SNE scatterplots for BERT embeddings of homonymous stimuli `degree.n`, `foot.n`, `table.n`, and `plane.n`. Sense definitions are in Table 5.

Although our visualizations are useful in determining if BERT cleanly separates senses of homonymous words, it is evident that they do not discriminate high-frequency, polysemous words well. We visualize how BERT represents these words in Figure 10. These results demonstrate BERT vectors may form possible clusters representing senses, but many of them overlap with one another.

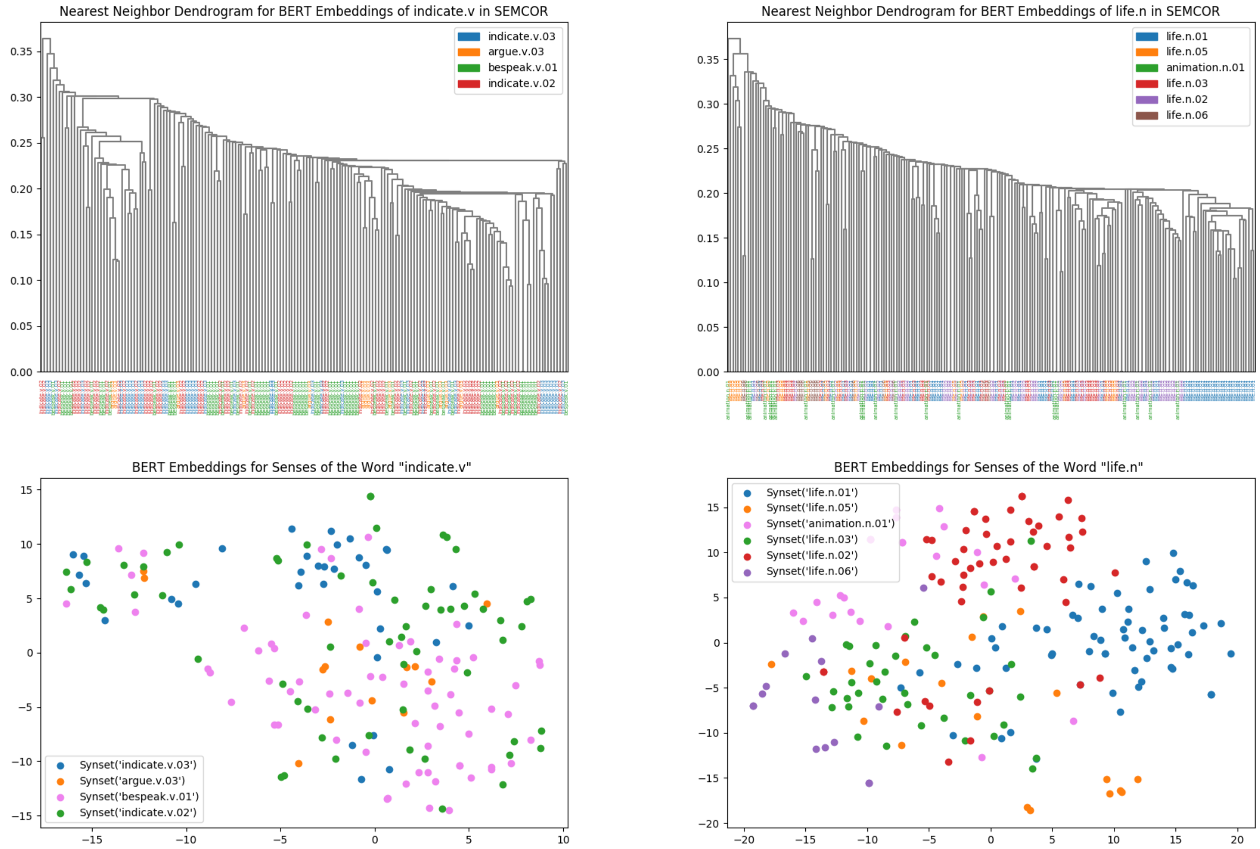


Figure 10: Dendrograms and t-SNE scatterplots for BERT embeddings of polysemous stimuli `lead.v`, `indicate.v`, `world.n`, and `life.n`. Sense definitions are listed in Tables 6 and 7, respectively.

5.2 Comparison of BERT Embeddings to Experimental Stimuli

Cosine distances between word vectors can demonstrate semantic relatedness (Dumais et al., 1988; Mikolov et al., 2013), so we can use this metric with the BERT embeddings to compare tokens to one another. Because BERT generates one embedding for each usage of a word, we compute the centroids of the BERT embeddings to get a representation for each sense, and compare different senses by using the cosine distance of these centroids. This allows us to test whether prototypical uses of a word senses occur close to the centroids of the points used to represent them. We plot the cosine distances of polysemous and homonymous sense pairs, using the centroids of the BERT vectors representing each sense in Figure 11. The median distance is 0.965 between pairs of centroids for homonymous senses, and 0.623 for polysemous sense pairs, and their distributions differ significantly (Mann-Whitney $U = 488.5, n_1 = 10, n_2 = 179, p < 0.01$).

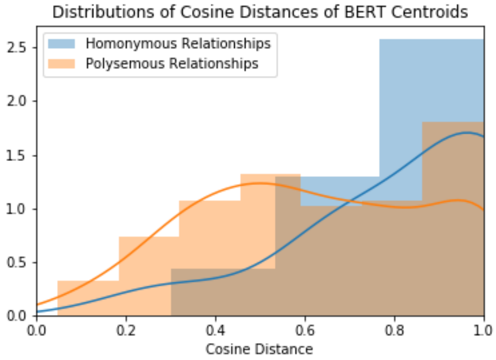


Figure 11: Distributions of cosine distance between pairs of centroids representing polysemous and homonymous relationships.

Now that we have evidence that BERT represents polysemous and homonymous relationships differently, we compare the distances derived from the BERT embeddings to the experimental data. Based on both the previous analysis and by comparing distance matrices from BERT and the experiment, we find evidence that both BERT and human representations place senses that are less semantically related from one another further than ones that are more semantically related. For type `degree.n` and `table.n`, note how both humans and BERT represent the two polysemous senses close to one another in Figure 12, and the homonymous sense far from both the polysemous senses.

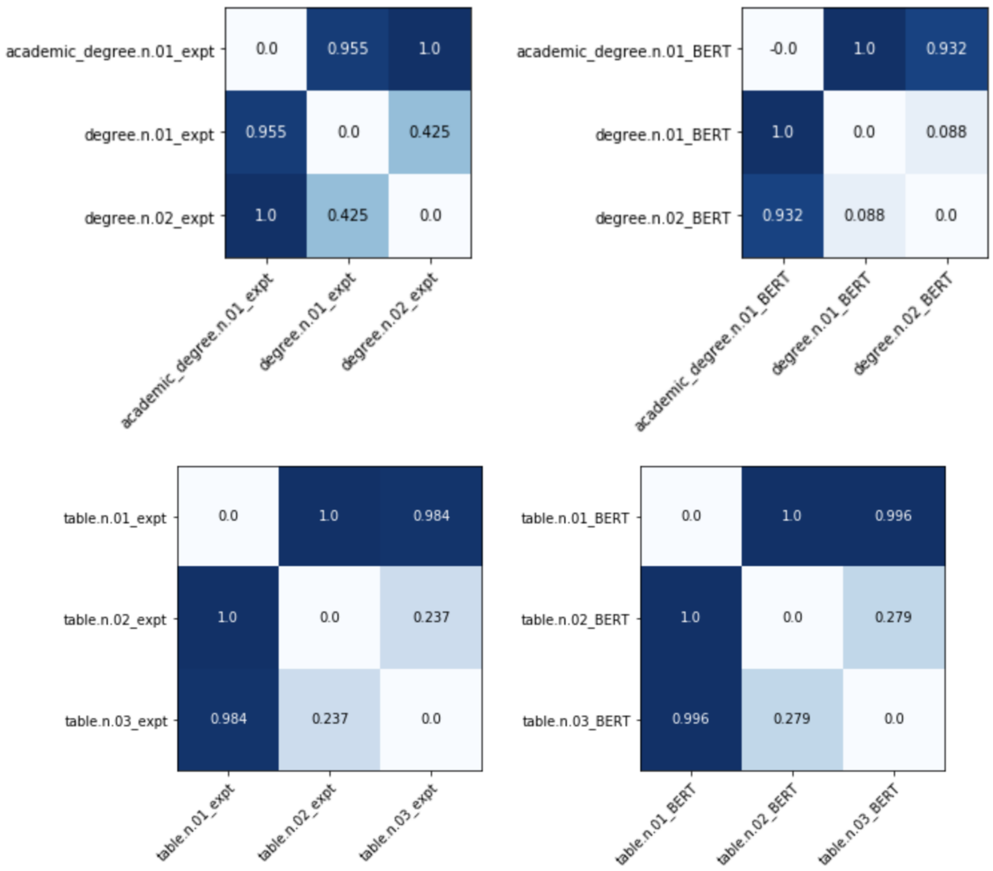


Figure 12: Distance matrices from experiment and BERT centroids for `degree.n.` and `table.n` (Senses in Table 5)

According to the analysis of the previous sections as well as inspection of the matrices, we seek further evidence that BERT is capturing many of the relationships between word senses reported by humans. For the word types used in the shared trials, we find that the BERT centroids' cosine distances and the relatedness judgements are highly correlated ($r = 0.784, p < 0.001$). When considering all stimuli, the cosine distances and human relatedness judgements are less correlated than the shared stimuli, but are still predictive ($r = 0.545, p < 0.001$). These global measures present evidence in the direction that BERT is capturing certain aspects of how humans perceive relations between some senses. To confirm that this relationship does not emerge as an artifact, we compare these correlations to a random baseline (Figure 13). This is done by simulating trials for 1000 hypothetical participants, constructing distance matrices from randomly generated relatedness judgements, and taking their Spearman rank correlation with the reported distance matrices.

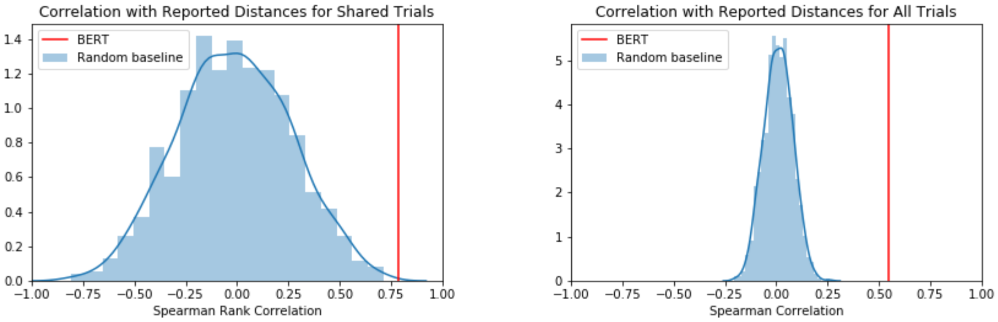


Figure 13: Comparing the correlation between words' reported relatedness judgements and their cosine distance matrices from BERT to their correlation with a baseline from 1000 randomly generated trials. For shared trials, $r = 0.74, p < 0.001$, and for all trials, $r = 0.542, p < 0.001$.

To determine if information about word sense lies in a lower dimensional space, we trained $\ell 1$ logistic regression models, where values in the embeddings that are unimportant in discriminating between the senses shrink to zero. This allows us to compare the experimental data with both the full BERT embeddings, as well as a reduced set, only considering the values in the vectors that were important to classification in our cosine distance computation. Correlations with experimental data are reported in Table 5.2. Using this information, we now stratify our analyses across three major categories: parts of speech, number of senses, and entropy, comparing the full set of embeddings to the reduced set. We define high entropy to be greater than 1.5 when rounded to the nearest tenth, and medium/low entropy to be less than this value (Figure 14), based on the values reported in Figure 2.

Table 2: Spearman Correlations of BERT Embeddings with Experimental Data

	Shared Stimuli	All Stimuli
Full Embeddings	0.742	0.541
Embeddings with Nonzero Weights	0.694	0.513

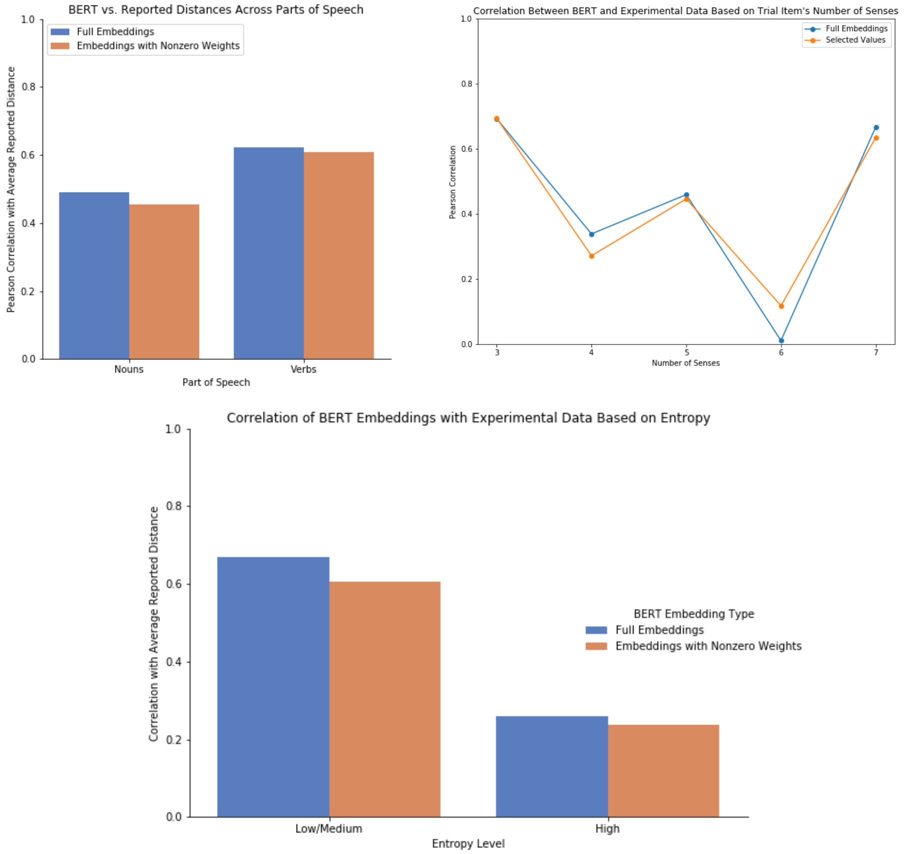


Figure 14: Differences between correlation with human data for full BERT embeddings and selected values with nonzero weights, across part of speech, number of senses, and entropy.

5.3 Evaluating the Probe

The main metric we use when evaluating performance of the multiclass logistic regression classifiers is the F1 score, computed as a weighted average of the precision and recall over all sense labels. Because we run 5-fold cross validation to ensure that each sense is part of a test set, we report an average of the F1 scores. We consider the F1 scores of the classifiers on all the words we computed entropy for (Figure 15), reporting an average of 0.759 across the entire dataset. The distribution of the F1 scores peaks around the average, but has a slight left tail, which suggests that the probe discriminates between BERT vectors for certain words better than others.

Afterwards, we plot the performance of the classifiers on all 406 words we have access to in

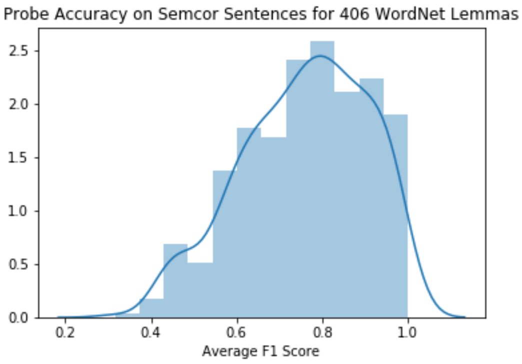


Figure 15: Classification accuracy of multiclass logistic regression models for lemmas in Semcor. Averages of F1 scores taken across 5-fold cross validation.

Semcor segmented by entropy and the number of senses in Figures 16 and 17, respectively.

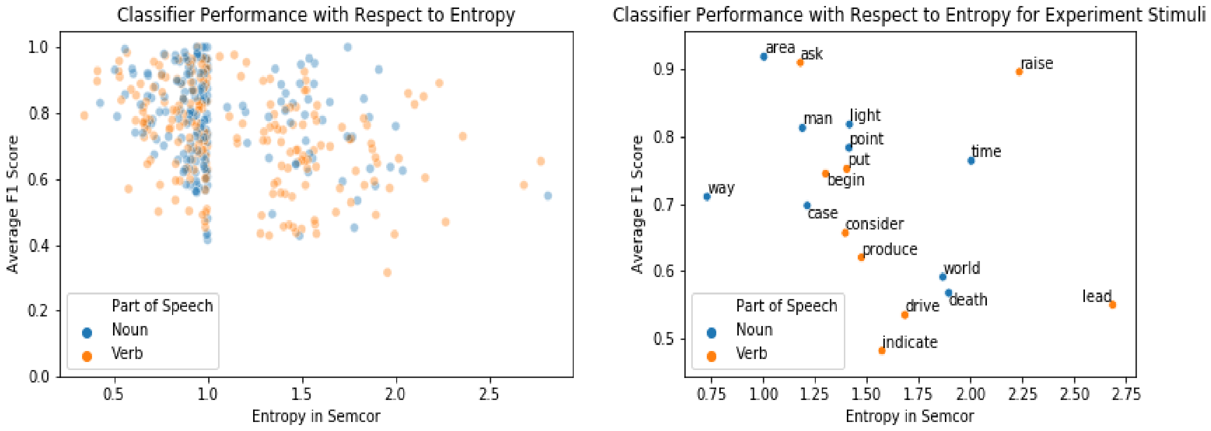


Figure 16: Classification accuracy roughly decreases as entropy increases. Left plot shows 406 lemmas from Semcor, right plot shows only experimental stimuli. F1 score is averaged over 5 runs of k-fold cross validation.

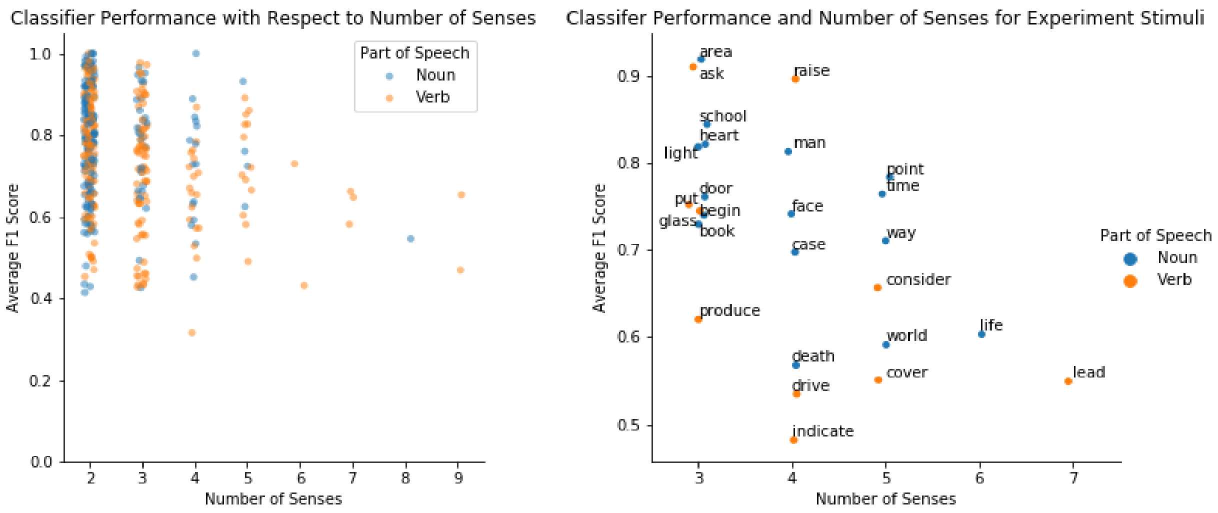


Figure 17: Classification accuracy is plotted compared to number of senses. Left plot shows 406 lemmas from Semcor, right plot shows only experimental stimuli. F1 score is averaged over 5 runs of k-fold cross validation.

A reason why we conducted the probing task was to use the results of the models to determine if a human-like relationship was present between the senses. We train binary (rather than multi-class) logistic classifiers to distinguish between polysemy and homonymy on shared trials, because there were not enough instances of the polysemous sense (fewer than 10 tokens) in the shared trials to train a classifier. These yield average F1 scores of 1 on all shared words besides `right.n` and `model.n` (0.962 and 0.982, respectively).

If BERT is representing words like humans are, we would expect patterns of misclassification more often among senses that humans judged to be highly related. We present the results of the classification with confusion matrices (Figures 18, 19, and 20). Results like those for `put.v` and `raise.v` (Figure 18), demonstrate what we would expect assuming BERT embeddings perfectly captured human representations of sense relatedness. The generally high degree of accuracy of `raise.v` demonstrated that BERT was able to effectively discriminate between senses. When the probe misclassified uses of `put.v`, its results were senses that human participants reported as similar to the true sense. For the most part, however, the patterns in misclassification were not as consistent with human intuition. Although general patterns were similar to the experimental data, we analyze specific word types with lower performance to help understand the different errors made by the probe.

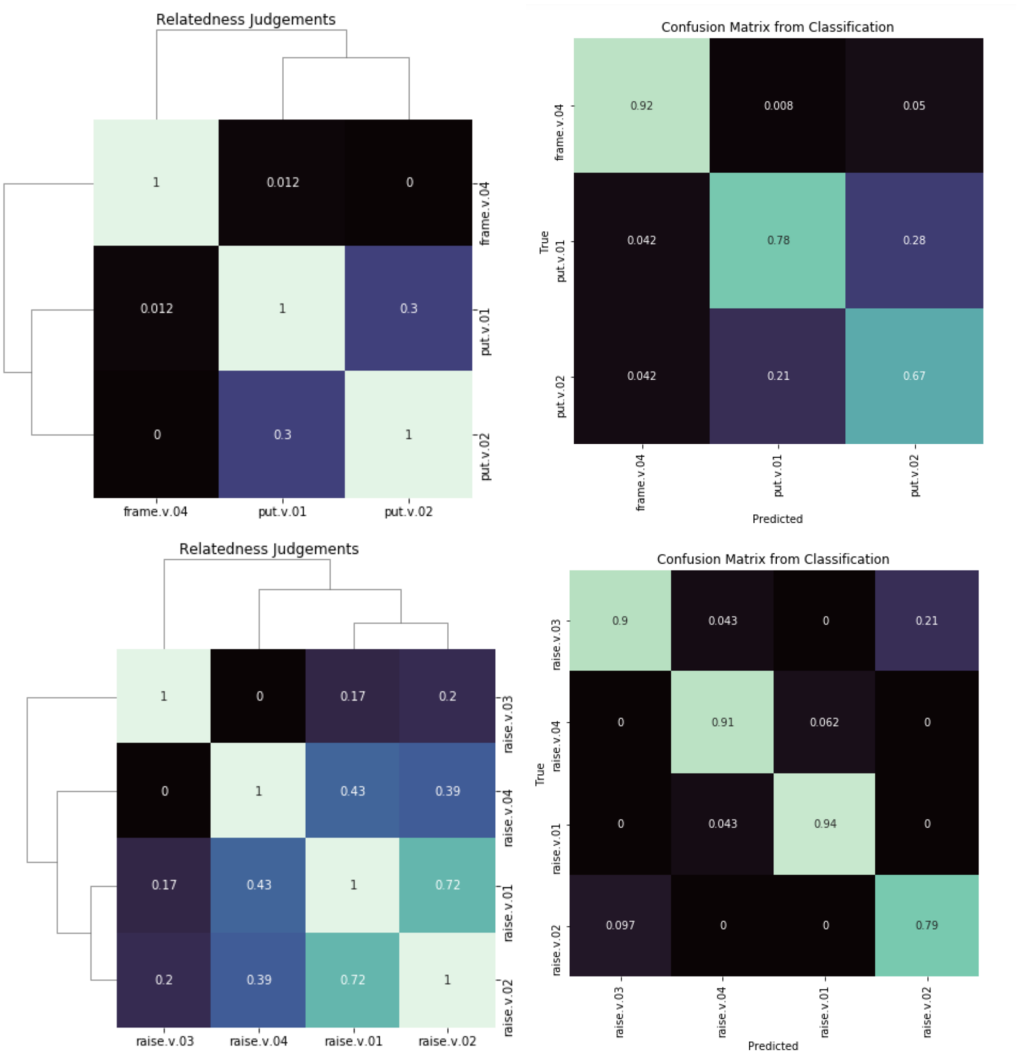


Figure 18: Relatedness judgements and confusion matrices for words that are classified accurately and reflect some similarity to human judgements, `put.v` and `raise.v`. (Sense definitions in Tables 8 and 9)

For the most part, BERT's performance was worse on abstract and metaphorical senses of words, but it sometimes picked up on cluster-like patterns in the data (Figures 19 and 20). Even if experimental results for the types `indicate.v` and `consider.v` demonstrated clear patterns in averaged human relatedness judgements, BERT did not classify its senses according to this pattern (Figure 19). In fact, `indicate.v` had the lowest F1 score of all stimuli (Figures 16 and 17), possibly because it had highly abstract senses (Table 6). However, even if many other words were not classified perfectly, we are still able to find discernible patterns in the probe's classification errors, and the inference of possible sense clusters (Figure 20). Certain words reflect patterns that showed sense categories being divided depending on whether they were concrete or abstract, demonstrating high performance for physical senses and lower performance for the set of metaphorical senses (Figure 20). Out of the 32 stimuli, 16 had metaphorical senses. The sense with the lowest accuracy was metaphorical for 9 out of the 16 word types that had metaphorical senses. For the remaining words, the metaphorical sense was classified with low accuracy, and in the cases where these senses had high accuracy, other senses were frequently misclassified as metaphorical.

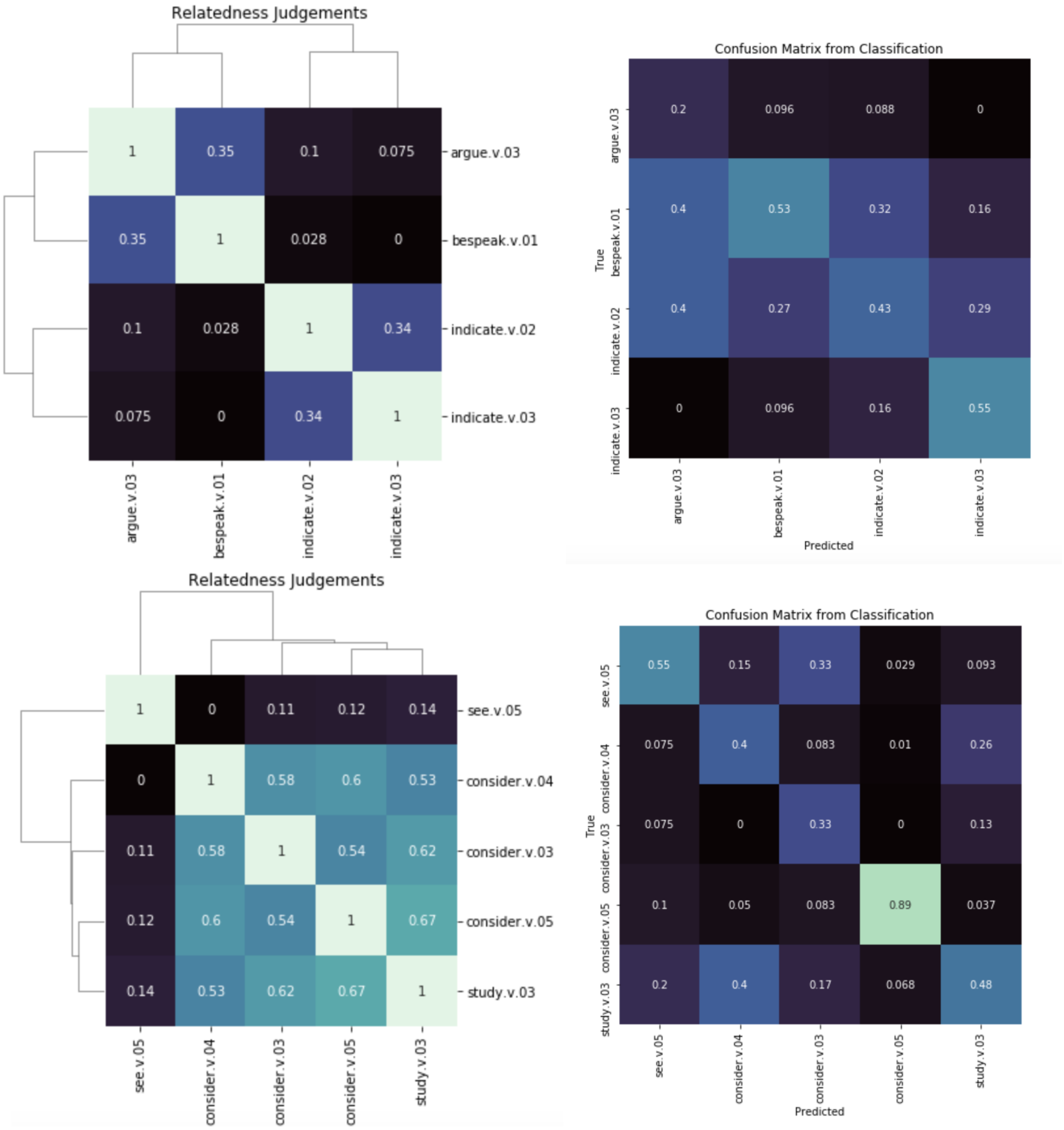


Figure 19: Relatedness judgements and confusion matrices for words the probe poorly performed on, `indicate.v` and `consider.v`. (Sense definitions in Tables 6 and 10)

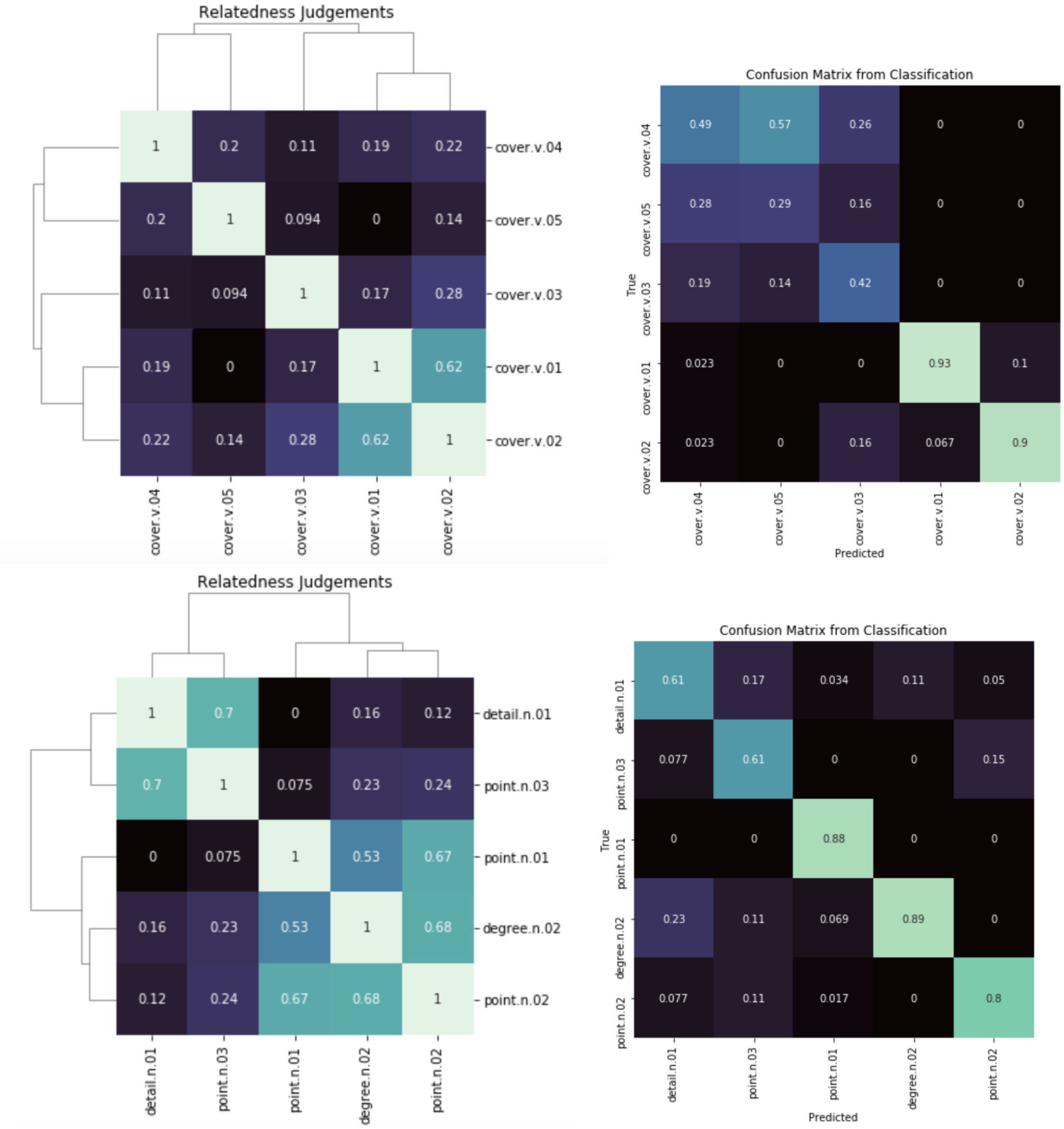


Figure 20: Relatedness judgements and confusion matrices for words with metaphorical senses that the probe was able to effectively isolate, but not effectively classify, `cover.v` and `point.n`. (Sense definitions in Table 11 and 12)

To compare the classification task with experimental data, we take the Spearman rank correlation between entries in the confusion matrices and matrices of the corresponding relatedness judgements. We use the Spearman correlation because there may not necessarily be a linear relationship between these values, and classification accuracy is not necessarily between zero and one like the normalized relatedness judgements. Across all test stimuli, we find a positive correlation ($r = 0.629, p < 0.001$). Splitting stimuli across the groups demonstrated in Figure 14, we report correlations for entropy levels and part of speech in Table 3, and for number of senses in Figure 21. Although both nouns and verbs have similar correlations to the global value, reported distances and cosine distances for high entropy words are substantially less correlated than for low to medium entropy words, suggesting that BERT is not encoding human-like distances relating to sense for words whose sense distributions are relatively less predictable. The correlation also does not reduce for words with many senses, showing that human judgements are relatively consistent with BERT regardless of the number of senses.

Table 3: Spearman Correlations of Confusion Matrices with Relatedness Judgements

Category	Correlation
Nouns	0.645
Verbs	0.6
High Entropy	0.517
Low/Medium Entropy	0.774

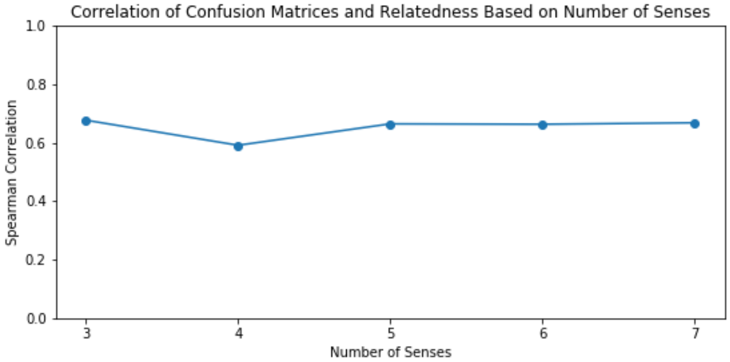


Figure 21: Spearman correlation of confusion matrices and relatedness judgements for test stimuli, stratified based on part of speech, entropy level, and number of senses

6 Discussion

We find that BERT can represent words with homonymous senses differently from polysemous senses, as demonstrated by both exploratory visualizations and classification accuracy. If we compare these embeddings to data from the experiment, where humans labeled homonyms as drastically different from polysemes, we find that BERT's representations are often highly predictive of these data. We provide evidence of a statistically significant difference between representations of polysemous and homonymous senses, both for human relatedness judgements and cosine distances of the centroids of BERT vectors (Figures 8 and 11). Afterwards, we use logistic regression as a probe for seeing how accurately BERT is able to discriminate between word senses. If the probe had high degrees of accuracy and it misclassified senses humans perceived to be related to the true sense, it would confirm our hypothesis that BERT vectors are capturing human-like intuitions about word senses. Because logistic regression infers a linear decision boundary, we can conclude the following: near-perfect degrees of accuracy on the classification task show that the homonymous senses can be linearly separable in the embedding space. Additionally, we show that BERT's encodings of homonyms are largely consistent with human intuition by comparing the correlations between cosine distance matrices and experimental data for the shared stimuli with a random baseline (Figure 13),

With polysemous words however, the story is quite different. The cosine distances between centroids of BERT vectors and classification accuracy were less correlated with human data for these stimuli, although they showed some degree of consistency (Figure 13). Our analysis lies along several dimensions: part of speech, number of senses, and entropy. The main finding is that the human judgements and BERT representations for high-entropy words are substantially less consistent (both in terms of cosine distance and classification accuracy) when compared to low and medium entropy words (Figures 14 and 16). This finding is supported by the overall decrease in classification accuracy with increasing entropy plotted in Figure 16. We might expect the correlations between BERT embeddings and the human judgements to be lower for words with more senses. However, this does not necessarily appear to be the case when comparing classification accuracy with relatedness judgements in Figure 17. These correlations are particularly robust compared to the number of senses the words have. If a classifier uniformly assigned sense names to embeddings at random, then its accuracy would decrease as the number of senses increases, and possibly become less consistent with human judgements. The fact that this is not the case indicates that BERT may be encoding key linguistic information about words with many senses, and could also tie back to the idea of word senses facilitating communicative efficiency (Piantadosi et al., 2012). This is because the comparing entropy to classification accuracy demonstrates a clearer pattern than comparing it with number of senses. Finally, an analysis of the confusion matrices for the individual senses of words suggests that it is difficult for BERT to distinguish metaphorical senses, like those in Figure 20. This is evident because the probe predicts metaphorical senses of words with much lower accuracy than their other senses.

Our work takes existing experimental work further by asserting that fluent speakers of a language have similar judgements of how related different meanings of a word are. While this seems intuitive, there exists little, if any empirical work (Lopukhina et al., 2018), hampering progress on formal models of word sense. We also provide evidence that lexical embeddings obtained from attention-based neural networks encode somewhat relevant information about human word sense knowledge, advancing the claims of (Coenen et al., 2019) and further suggesting that they can be used to develop a cognitively informed model of word sense. This is important because WordNet, the gold-standard resource for word sense, encodes discrete relations between senses of a word, but other cognitive theories of word sense assert that people treat definitions of a word as a continuous gradient between homonymy and various types of polysemy. We provide cosine similarity in the vector space as a continuous degree of relatedness between senses that somewhat approximates human judgements. This suggests that words that have senses that are easily distinguishable are homonymous and that words that have overlapping senses are polysemous. Our work thus demonstrates the potential utility of an exemplar-theoretic understanding of word sense, by comparing prototypes from both human judgements and BERT. The fact that centroids from BERT, representing prototypes in the embedding space, are reflected similarly to human judgements, is critical evidence that this model can be taken further.

We demonstrate that state of the art results on word sense disambiguation tasks could happen because certain words that are distinguished well by BERT are outweighing ones that are not discriminated as easily, rather than all words performing around the same in Figure 15. Additionally, we provide corroborating evidence for the claim from Coenen et al. (2019) that word sense is contained in a lower dimensional subspace of the BERT embedding space, through reducing the BERT vectors to only include the dimensions important in discriminating senses achieves similar performance to the full BERT embeddings when compared with our experimental data (Figure 14). This means that further work exploring the geometric representations of word meaning can be more interpretable, and potentially demonstrate that systematic types of polysemy can be expressed as transformations in the vector space.

7 Further Work and Conclusion

Now that we demonstrate that BERT encodes knowledge about word senses that is largely consistent with human judgements, logical next steps would be understanding how BERT represents polysemy and further investigating homonymous relations. First, our work uses a linear classifier that effectively distinguishes homonymous senses from one another, but we can determine if more sophisticated classifiers can learn higher-order, nonlinear decision boundaries, and also develop a classifier to determine if a word is homonymous or polysemous based on how BERT represents its senses. Second, we can explore transformations of the vector space to determine if metaphorical polysemy or other patterns, like item-for-material or objects-for-contents, are implicit in its geometry, or if the architecture needs to be fine-tuned to recognize these distinctions. We can also revisit the probe using methods like principle components analysis to provide more evidence that models like BERT can represent word senses in a lower-dimensional space. This work could be extended to cross-linguistic studies, as patterns of polysemy in one language may persist in others. Another line of work that could be used to better understand the results would be comparing the attention matrices of individual tokens, to understand what aspects of the context contribute the most to generating the embeddings. Due to issues with data sparsity in Semcor, we could collect more data with sense annotations, emphasizing spoken and conversational corpora. This would especially be useful to strengthen our claims about senses of homonymous words

Finally, we can use this study's experimental paradigm to further research human perceptions of word sense. Because we impose a metric space for human judgements, it is possible that they may not obey the triangle inequality, so this could explain why BERT is not capturing all the distinctions reported by participants. Tversky (1977) note that similarity judgements for different words do not obey the triangle inequality, but these judgements were taken out of context. If the triangle inequality does hold with in-context judgements of word senses, we can strengthen the claim that regular polysemy is accounted for through transformations in the embedding space. Because our model of word senses is exemplar theoretic but the experiment uses prototypes, future studies could take an exemplar-focused experimental approach. This could involve participants arranging sentences containing individual uses of a word, and then comparing these results to BERT embeddings. Now that we have seen that BERT's judgement of WordNet senses are similar to human intuitions, we can run experiments on human participants where WordNet senses for a word are compared with the results of BERT-based word sense induction (Amrami and Goldberg, 2019), so participants can judge if the senses learned by the model are truly more intuitive than WordNet.

We provide critical evidence for an exemplar-theoretic model of word sense through our work. Through the behavioral experiment, we demonstrate that English speakers share similar judgements about the relatedness of different meanings of words. By comparing their understanding to prototypical sense usages derived from exemplars modeled by attention-based contextualized word embedding models, we demonstrate that these models encode some human-like distinctions between word senses. They effectively model the senses of homonymous words and provide a continuous measure of relatedness for senses of polysemous words. Our work serves as evidence in the direction that continuous representations from these models have the potential to describe the structure of the lexicon, and can be an improvement over discrete, symbolic resources. Now that we demonstrate basic levels of consistency with human judgements, we can inspire further research to create formal, quantitative models for a critical problem in language acquisition and processing.

Acknowledgements

This work would not have been possible without the guidance and useful criticism of many people. Special thanks to my mentor, Dr. Stephan Meylan, for investing an inordinate amount of time cultivating my growth as a researcher for the past three years, and encouraging me to pursue this as my first independent project. I greatly appreciate the feedback throughout the project from my first reader, Professor Mahesh Srinivasan, who provided extremely useful theoretical input, and from my second reader, Professor Steven Piantadosi, who pushed me to think beyond technical tools and about the bigger questions. I would also like to thank members of the Language and Cognitive Development Lab, specifically Ruthe Foushee for her feedback during the early stages of the project and Jon Wehry for ensuring the online experiment ran smoothly, and the Computation and Language Lab, for sparking insightful discussions. Although practicing research is one thing, it would not happen without foundational knowledge, so I owe a lot to Professors Celeste Kidd, Yang Xu, Peter Jenks, Jonathan Shewchuk, David Bamman, and Ani Adhikari. On a more personal note, thanks to my friends who never lost faith in my abilities, especially when I doubted my own; I really wish we could all celebrate this achievement in person. Finally, I express deep gratitude for my family, it is fitting that they are witnessing the final stretches of this journey, when they inoculated me with a curiosity about language in the first place.

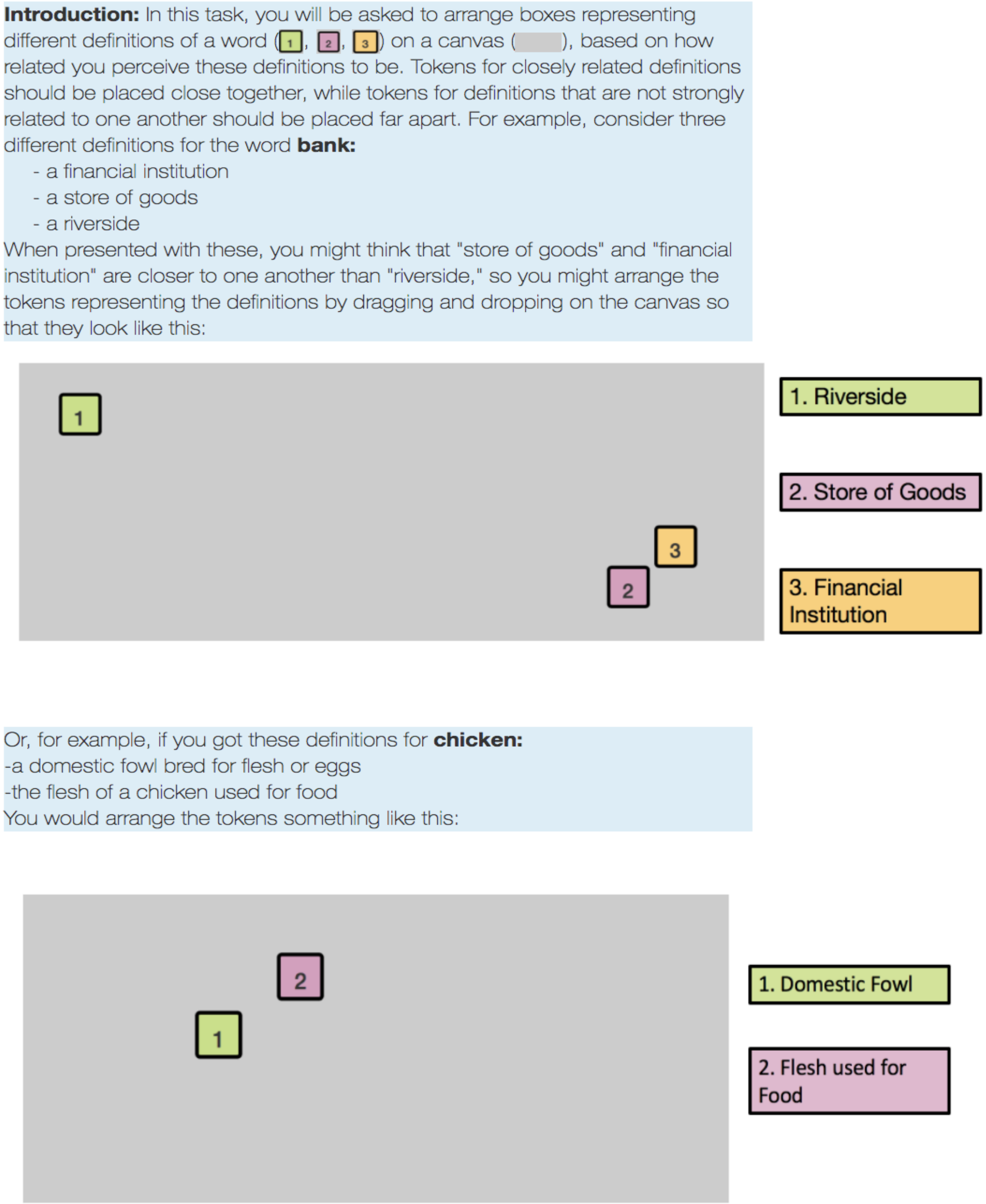
Appendices

A Experimental Stimuli from Semcor

Lemma	**Entropy**	**Number of Senses**	**Frequency in Semcor**	**Part of Speech**
thing.n	2.812868508	8	264	Noun
lead.v	2.68498204	7	170	Verb
raise.v	2.234669101	5	111	Verb
time.n	2.002841765	5	505	Noun
find.v	1.9969499819999998	6	663	Verb
death.n	1.894324902	4	103	Noun
world.n	1.8656682230000001	4	202	Noun
drive.v	1.683292156	4	106	Verb
indicate.v	1.572228763	3	177	Verb
produce.v	1.473220252	3	130	Verb
light.n	1.4156332059999999	3	77	Noun
point.n	1.4134178480000001	3	118	Noun
put.v	1.402340119	3	257	Verb
consider.v	1.394830833	3	229	Verb
begin.v	1.300172387	4	390	Verb
case.n	1.211936446	3	127	Noun
man.n	1.1876221390000001	4	638	Noun
ask.v	1.178507665	3	408	Verb
area.n	1.002967627	3	200	Noun
way.n	0.728520948	3	269	Noun

Table 4: All experimental stimuli from Semcor, with entropy, part of speech, and number of senses

B Instructions Given to participants Before Task



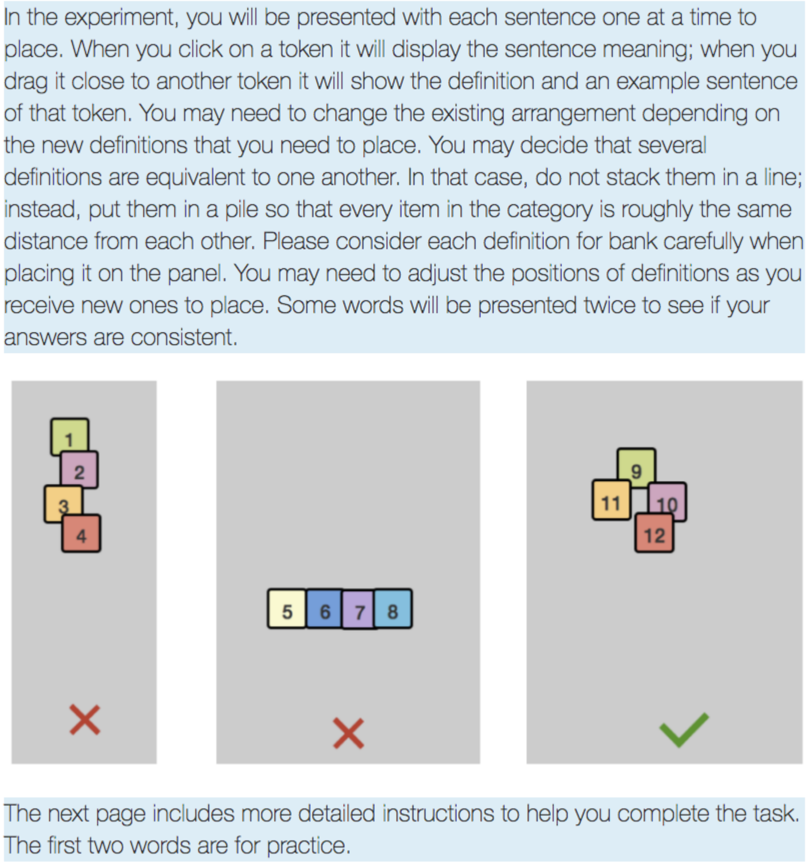
Introduction: In this task, you will be asked to arrange boxes representing different definitions of a word (1, 2, 3) on a canvas (▨), based on how related you perceive these definitions to be. Tokens for closely related definitions should be placed close together, while tokens for definitions that are not strongly related to one another should be placed far apart. For example, consider three different definitions for the word **bank:**

- a financial institution
- a store of goods
- a riverside

When presented with these, you might think that "store of goods" and "financial institution" are closer to one another than "riverside," so you might arrange the tokens representing the definitions by dragging and dropping on the canvas so that they look like this:

Or, for example, if you got these definitions for **chicken:**
-a domestic fowl bred for flesh or eggs
-the flesh of a chicken used for food
You would arrange the tokens something like this:

Figure 22: Participants were presented with instructions and examples on how to place tokens for each word sense on a canvas



In the experiment, you will be presented with each sentence one at a time to place. When you click on a token it will display the sentence meaning; when you drag it close to another token it will show the definition and an example sentence of that token. You may need to change the existing arrangement depending on the new definitions that you need to place. You may decide that several definitions are equivalent to one another. In that case, do not stack them in a line; instead, put them in a pile so that every item in the category is roughly the same distance from each other. Please consider each definition for bank carefully when placing it on the panel. You may need to adjust the positions of definitions as you receive new ones to place. Some words will be presented twice to see if your answers are consistent.

The next page includes more detailed instructions to help you complete the task. The first two words are for practice.

Figure 23: Further instructions issued to participants

C Sense Definitions of Words in Figures

Sense Name	WordNet Definition
foot.n.01	the part of the leg of a human being below the ankle joint
foot.n.02	a linear unit of length equal to 12 inches or a third of a yard
foot.n.03	the lower part of anything
table.n.01	a set of data arranged in rows and columns
table.n.02	a piece of furniture having a smooth flat top that is usually supported by one or more vertical legs
table.n.03	a piece of furniture with tableware for a meal laid out on it
airplane.n.01	an aircraft that has a fixed wing and is powered by propellers or jets
plane.n.02	(mathematics) an unbounded two-dimensional shape
plane.n.03	a level of existence or development
academic degree.n.01	an award conferred by a college or university signifying that the recipient has satisfactorily completed a course of study
degree.n.01	a position on a scale of intensity or amount or quality
degree.n.02	a specific identifiable position in a continuum or series or especially in a process
right.n.01	an abstract idea of that which is due to a person or governmental body by law or tradition or nature
right.n.02	location near or direction toward the right side; i.e. the side to the south when a person or object faces east
right.n.04	those who support political or social or economic conservatism; those who believe that things are better left unchanged
model.n.01	a hypothetical description of a complex entity or process
model.n.02	a type of product
model.n.03	a person who poses for a photographer or painter or sculptor

Table 5: WordNet senses for shared types

Sense	Type	Definition
argue.v.03	indicate.v	give evidence of
bespeak.v.01	indicate.v	be a signal for or a symptom of
indicate.v.02	indicate.v	indicate a place, direction, person, or thing; either spatially or figuratively
indicate.v.03	indicate.v	to state or express briefly

Table 6: WordNet Sense Definitions for `indicate.v` (Figures 10 and 19)

Sense	Type	Definition
animation.n.01	life.n	the condition of living or the state of being alive
life.n.01	life.n	a characteristic state or mode of living
life.n.02	life.n	the experience of being alive; the course of human events and activities
life.n.03	life.n	the course of existence of an individual; the actions and events that occur in living
life.n.05	life.n	the period during which something is functional (as between birth and death)
life.n.06	life.n	the period between birth and the present time

Table 7: WordNet Sense Definitions for `life.n` (Figure 10)

Sense	Type	Definition
frame.v.04	put.v	formulate in a particular style or language
put.v.01	put.v	put into a certain place or abstract location
put.v.02	put.v	cause to be in a certain state; cause to be in a certain relation

Table 8: WordNet Sense Definitions for `put.v` (Figure 18)

Sense	Type	Definition
raise.v.01	raise.v	raise the level or amount of something
raise.v.02	raise.v	raise from a lower to a higher position
raise.v.03	raise.v	cause to be heard or known; express or utter
raise.v.04	raise.v	collect funds for a specific purpose

Table 9: WordNet Sense Definitions for `raise.v` (Figure 18)

Sense	Type	Definition
consider.v.03	consider.v	take into consideration for exemplifying purposes
consider.v.04	consider.v	show consideration for; take into account
consider.v.05	consider.v	think about carefully; weigh
see.v.05	consider.v	deem to be
study.v.03	consider.v	give careful consideration to

Table 10: WordNet Sense Definitions for `consider.v` (Figure 19)

Sense	Type	Definition
cover.v.01	cover.v	provide with a covering or cause to be covered
cover.v.02	cover.v	form a cover over
cover.v.03	cover.v	span an interval of distance, space or time
cover.v.04	cover.v	provide for
cover.v.05	cover.v	act on verbally or in some form of artistic expression

Table 11: WordNet Sense Definitions for `cover.v` (Figure 20)

Sense	Type	Definition
degree.n.02	point.n	a specific identifiable position in a continuum or series or especially in a process
detail.n.01	point.n	an isolated fact that is considered separately from the whole
point.n.01	point.n	a geometric element that has position but no extension
point.n.02	point.n	the precise location of something; a spatially limited location
point.n.03	point.n	a brief version of the essential meaning of something

Table 12: WordNet Sense Definitions for `point.n` (Figure 20)

D Experiment Completion Times

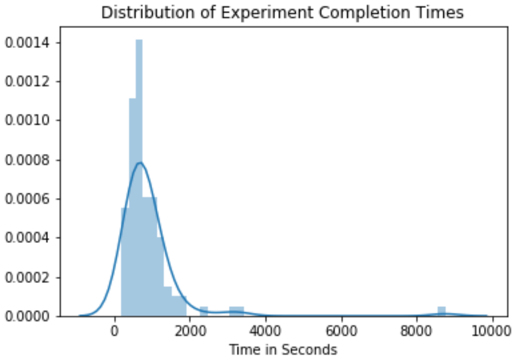


Figure 24: Distribution of experiment completion times for all participants with the exclusion of one outlier (20 hours were recorded for this participant). This time may have been recorded because they could have had the task open on their computer even if they were not working on it. Their data was ultimately included based on their consistency scores.

References

Amrami, A. and Goldberg, Y. (2019). Towards better substitution-based word sense induction. *CoRR*, abs/1905.12598.

Apresjan, J. D. (1974). Regular polysemy. *Linguistics*, 12(142):5–32.

Arora, S., Li, Y., Liang, Y., Ma, T., and Risteski, A. (2018). Linear algebraic structure of word senses, with applications to polysemy. *Transactions of the Association for Computational Linguistics*, 6:483–495.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate.

Batali, J. (1999). The negotiation and acquisition of recursive grammars as a result of competition among exemplars. In *Linguistic evolution through language acquisition: Formal and computational models*. Citeseer.

Bird, S. and Loper, E. (2004). NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

Bod, R. and Cochran, D. (2007). Introduction to exemplar-based models of language acquisition and use. In *Proceedings of the ESSLLI Workshop "Exemplar-Based Models of Language Acquisition and Use"', ESSLLI.*

Bréal, M. (1897). Essai de sémantique: Science des significations (hachette, paris).

Budanitsky, A. and Hirst, G. (2006a). Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47.

Budanitsky, A. and Hirst, G. (2006b). Evaluating wordnet-based measures of lexical semantic relatedness. *COMPUTATIONAL LINGUISTICS*, 32(1):13–47.

Coenen, A., Reif, E., Yuan, A., Kim, B., Pearce, A., Viégas, F., and Wattenberg, M. (2019). Visualizing and measuring the geometry of bert. *arXiv preprint arXiv:1906.02715.*

Crossley, S., Salsbury, T., and McNamara, D. (2010). The development of polysemy and frequency use in english second language speakers. *Language Learning*, 60(3):573–605.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Dumais, S. T., Furnas, G. W., Landauer, T. K., Deerwester, S., and Harshman, R. (1988). Using latent semantic analysis to improve access to textual information. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 281–285.

Ethayarajh, K. (2019). 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, Hong Kong, China. Association for Computational Linguistics.

Francis, W. N. and Kucera, H. (1979). Brown corpus manual. *Letters to the Editor*, 5(2):7.

Frazier, L. and Rayner, K. (1990). Taking on semantic commitments: Processing multiple meanings vs. multiple senses. *Journal of memory and language*, 29(2):181–200.

Goldstone, R. (1994). An efficient method for obtaining similarity data. *Behavior Research Methods, Instruments, Computers*, 26:381–386.

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

Hearst, M. A. (1998). Automated discovery of wordnet relations. *WordNet: an electronic lexical database*, 2.

Hewitt, J. and Liang, P. (2019). Designing and interpreting probes with control tasks. *arXiv preprint arXiv:1909.03368*.

Hewitt, J. and Manning, C. D. (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.

Jawahar, G., Sagot, B., and Seddah, D. (2019). What does bert learn about the structure of language?

Jurafsky, D. and Martin, J. H. (2014). Speech and language processing. vol. 3.

Khodak, M., Risteski, A., Fellbaum, C., and Arora, S. (2017). Automated wordnet construction using word embeddings. In *Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and their Applications*, pages 12–23.

Klein, D. E. and Murphy, G. L. (2001). The representation of polysemous words. *Journal of Memory and Language*, 45(2):259–282.

Klepousniotou, E. (2002). The processing of lexical ambiguity: Homonymy and polysemy in the mental lexicon. *Brain and language*, 81(1-3):205–223.

Klepousniotou, E., Pike, G. B., Steinhauer, K., and Gracco, V. (2012). Not all ambiguous words are created equal: An eeg investigation of homonymy and polysemy. *Brain and language*, 123(1):11–21.

Klepousniotou, E., Titone, D., and Romero, C. (2008). Making sense of word senses: The comprehension of polysemy depends on sense overlap. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(6):1534.

Lakoff, G. and Johnson, M. (1980). Metaphors we live by. *Chicago, IL: University of Chicago.*

Lilleberg, J., Zhu, Y., and Zhang, Y.-Q. (2015). Support vector machines and word2vec for text classification with semantic features. *2015 IEEE 14th International Conference on Cognitive Informatics Cognitive Computing (ICCI*CC)*, pages 136–140.

Lopukhina, A., Laurinavichyute, A., Lopukhin, K., and Dragoy, O. (2018). The mental representation of polysemy across word classes. *Frontiers in psychology*, 9:192.

Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.

MacGregor, L. J., Bouwsema, J., and Klepousniotou, E. (2015). Sustained meaning activation for polysemous but not homonymous words: Evidence from eeg. *Neuropsychologia*, 68:126–138.

Mcdonald, S. and Ramscar, M. (2001). Testing the distributional hypothesis: The influence of context on judgements of semantic similarity. In *In Proceedings of the 23rd Annual Conference of the Cognitive Science Society*, pages 611–6.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. J. (1990). Introduction to wordnet: An on-line lexical database. *International journal of lexicography*, 3(4):235–244.

Miller, G. A., Leacock, C., Tengi, R., and Bunker, R. T. (1993). A semantic concordance. In *Proceedings of the workshop on Human Language Technology*, pages 303–308. Association for Computational Linguistics.

Ng, A. Y. (2004). Feature selection, l 1 vs. l 2 regularization, and rotational invariance. In *Proceedings of the twenty-first international conference on Machine learning*, page 78.

Pedersen, T., Patwardhan, S., Michelizzi, J., et al. (2004). Wordnet:: Similarity-measuring the relatedness of concepts. In *AAAI*, volume 4, pages 25–29.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proc. of NAACL*.

Piantadosi, S. T., Tily, H., and Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition*, 122(3):280–291.

Pustejovsky, J. (1991). The generative lexicon. *Computational linguistics*, 17(4):409–441.

Rodd, J., Gaskell, G., and Marslen-Wilson, W. (2002). Making sense of semantic ambiguity: Semantic competition in lexical access. *Journal of Memory and Language*, 46(2):245–266.

Søgaard, A. (2010). Simple semi-supervised training of part-of-speech taggers. In *Proceedings of the ACL 2010 Conference Short Papers*, ACLShort '10, page 205–208, USA. Association for Computational Linguistics.

Srinivasan, M., Berner, C., and Rabagliati, H. (2019). Children use polysemy to structure new word meanings. *Journal of Experimental Psychology: General*, 148(5):926.

Srinivasan, M. and Rabagliati, H. (2015). How concepts and conventions structure the lexicon: Cross-linguistic evidence from polysemy. *Lingua*, 157.

Trask, A., Michalak, P., and Liu, J. (2015). sense2vec - A fast and accurate method for word sense disambiguation in neural word embeddings. *CoRR*, abs/1511.06388.

Tversky, A. (1977). Features of similarity. *Psychological review*, 84(4):327.

Tversky, A. and Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive psychology*, 5(2):207–232.

Van Rijsbergen, C. J. (1979). Information retrieval.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Veale, T. (2004). A non-distributional approach to polysemy detection in wordnet.

Vincent-Lamarre, P., Massé, A. B., Lopes, M., Lord, M., Marcotte, O., and Harnad, S. (2016). The latent structure of dictionaries. *Topics in Cognitive Science*, 8(3):625–659.

Wiedemann, G., Remus, S., Chawla, A., and Biemann, C. (2019). Does bert make any sense? interpretable word sense disambiguation with contextualized embeddings.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., and Brew, J. (2019). Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Yuan, D., Richardson, J., Doherty, R., Evans, C., and Altendorf, E. (2016). Semi-supervised word sense disambiguation with neural models. *arXiv preprint arXiv:1603.07012*.

Zhou, W., Ge, T., Xu, K., Wei, F., and Zhou, M. (2019). Bert-based lexical substitution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3368–3373.

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., and Fidler, S. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.