Contextualized Word Embeddings Encode Aspects of Human-Like Word Sense Knowledge

Sathvik Nair, Mahesh Srinivasan
Cognitive Science Program,
Department of Psychology,
University of California, Berkeley
Berkeley, CA, United States
`sathviknair,srinivasan@berkeley.edu`

Stephan Meylan
Department of Brain and Cognitive Sciences,
Massachusetts Institute of Technology
Cambridge, MA, United States
`smeylan@mit.edu`

Abstract

Understanding context-dependent variation in word meanings is a key aspect of human language comprehension supported by the lexicon. Lexicographic resources (e.g., WordNet) capture only some of this context-dependent variation; for example, they often do not encode how closely senses, or discretized word meanings, are related to one another. Our work investigates whether recent advances in NLP, specifically contextualized word embeddings, capture human-like distinctions between English word senses, such as polysemy and homonymy. We collect data from a behavioral, web-based experiment, in which participants provide judgments of the relatedness of multiple WordNet senses of a word in a two-dimensional spatial arrangement task. We find that participants' judgments of the relatedness between senses are correlated with distances between senses in the BERT embedding space. Homonymous senses (e.g., bat as mammal vs. bat as sports equipment) are reliably more distant from one another in the embedding space than polysemous ones (e.g., chicken as animal vs. chicken as meat). Our findings point towards the potential utility of continuous-space representations of sense meanings.

1 Introduction

A key challenge in natural language understanding is grasping the range of meanings that a word can take as a function of linguistic and non-linguistic context. Successful linguistic comprehension involves constantly resolving lexical ambiguity of this nature (Frazier and Rayner, 1990; Klepousniotou, 2002). This re-use of word forms by speakers — relying on listeners to choose the appropriate meaning depending on context — confers language with higher communicative efficiency than it would otherwise have, and may pose a solution to the problem of limited memory (Piantadosi et al., 2012). The re-use of word forms also allows speakers to extend the lexicon to new communicative situations, for example to refer to new objects, entities or processes using existing words in the language (Ramiro et al., 2018; Xu et al., 2020; Srinivasan et al., 2019). How these context-specific meanings are understood and represented pose critical open questions regarding the lexicon.

Following the conventions of lexicographic resources like natural language dictionaries, variation in word meaning is often treated in a categorical fashion: lexical types contain clusters of related meanings, or *word senses*. Under this treatment, word tokens (instances of a word type that are used in context) can be categorized into these word senses. Different sub-types of lexical ambiguity are thus reflected in different relations between word senses: pairs of word senses are *polysemous* if they are semantically related (Pustejovsky, 1998), or *homonymous* if this is not the case (Apresjan, 1974; Tuggy, 1993; Lyons, 1995). For example, using "bottle" to refer to a container as well as to the liquid it contains is an example of polysemy. By contrast, using "bank" to refer to a riverside or a financial institution — two semantically unrelated meanings — constitutes homonymy. Polysemous relations are often *regular* (Apresjan, 1974), such that many word types exhibit the same alternation (e.g., the container-for-contents relation is also exemplified by "box" and "glass"); such patterns can also be generalized to new words (Srinivasan et al., 2019).

Because of variation in the perceived relationship between senses across language users, theories in psycholinguistics assert that humans treat polysemy and homonymy as falling onto a continuous gradient (Tuggy, 1993; Crossley et al., 2010). This is corroborated by psycholinguistic experiments which show that both adults and children represent and process polysemous and homonymous senses differently (Frazier and Rayner, 1990; Rodd et al., 2004; Klepousniotou et al., 2008; Rabagliati and Snedeker, 2013; MacGregor et al., 2015). However, these theories of word sense representation lie in stark contrast with many lexicographical approaches, most notably WordNet (Miller et al., 1990). Although WordNet contains a vast store of word senses—and has historically been considered the gold standard ontology of word senses (Jurafsky and Martin, 2014)—it does not encode relations among those senses, and thus does not distinguish between polysemy and homonymy.

Models based on distributional semantics can help bridge the gap between psychological theories of word senses and existing lexicographical resources. Contextualized word embeddings (CWEs), especially BERT (Devlin et al., 2019), which is derived from a Transformer-based architecture (Vaswani et al., 2017), have the potential to capture fine-grained distinctions in relatedness between word senses, because they offer a continuous measure of relatedness between individual uses of words in context (Lake and Murphy, 2020). While fine-tuned BERT models are known to perform exceptionally well in word sense disambiguation (WSD) tasks (Wiedemann et al., 2019; Loureiro and Jorge, 2019; Blevins and Zettlemoyer, 2020), there are still open questions of whether BERT representations capture fine-grained distinctions in sense relatedness.

In the present work, we evaluate whether contextualized embeddings from BERT capture relationships between word senses similar to English speakers. We collect human judgments of the relatedness among different senses of 32 English words through a web-based two-dimensional spatial arrangement task (Goldstone, 1994). We then compare the experimental data to BERT vectors for the same set of words. This is done through extracting and analyzing BERT embeddings for word tokens in the Semcor corpus (Miller et al., 1993), which has been annotated with word sense identifiers from WordNet. Assessing whether relatedness in BERT representations corresponds with human judgments of sense relatedness is an important test of whether CWEs can be used to develop more realistic computational models of word meanings.

2 Background

The distributional hypothesis proposes that a word's meaning can be represented by the lexical context in which it occurs (Harris, 1954). This insight underlies the success of computational models that represent words as continuous-valued vectors representing their surrounding lexical context. Latent semantic analysis (Dumais et al., 1988), and neural models like Word2Vec(Mikolov et al., 2013) encode information about all of a word's senses in a single vector. By contrast, newer models include contextualized embeddings: vector space representations of specific word uses (tokens), reflecting the context of their use. These richer word representations from contextualized embeddings have allowed researchers to ask which aspects of linguistic knowledge are incorporated in their vector spaces (See Rogers et al. (2020) for a review; see also Ethayarajh (2019)). For example, one of the most notable discoveries from this line of work suggests that a sentence's dependency parse tree can be reconstructed from BERT embeddings (Hewitt and Manning, 2019; Reif et al., 2019; Jawahar et al., 2019). However, despite progress investigating how BERT encodes syntactic information, relatively little work has explored how it encodes semantic information such as the relations between word senses.

While older models like Word2Vec may encode some aspects of word sense (Arora et al., 2018), we focus here on Transformer architectures (Vaswani et al., 2017), which may be able to find sense-distinguishing context words at larger distances, and thus perform better for WSD. BERT-based approaches have led to state of the art performance on word sense disambiguation tasks (Wiedemann et al., 2019; Loureiro and Jorge, 2019; Blevins and Zettlemoyer, 2020), but approaches analyzing the model's representation of word senses remain exploratory (Reif et al., 2019; Lake and Murphy, 2020). Evidence from Reif et al. (2019) shows that BERT embeddings for the same word in a large text corpus are largely clustered based on their meanings, but the authors do not investigate the model's ability to encode canon-

ically homonymous or polysemous relationships. Mickus et al. (2020) note that BERT embeddings for the same word type vary as a function of the position in the sentences they occur in. Lake and Murphy (2020) assess BERT's capacity for homonym resolution by comparing relatedness of words that are similar to a homonym in highly constraining contexts. Ettinger (2020) points out the need for more psycholinguistic diagnostics of neural language models like BERT, and compares BERT to human data from tasks such as commonsense inference. Our work falls under this framework, focusing on comparing human judgements of word sense relatedness to BERT representations.

3 Methods

We conducted a metalinguistic experiment (vs. a processing task testing implicit knowledge) where participants used a web interface to assess the relatedness of WordNet senses in a two-dimensional spatial arrangement task. We then obtained BERT embeddings for word types in the Semcor corpus (Miller et al., 1993) and compared them to the experimental data, looking at both cosine distance in the embedding space as well as the accuracy of a sense classifier using BERT's contextualized word embeddings as input. All stimuli, code and visualizations are available at `https://osf.io/fm78w`.

3.1 Data

We select a sample of word types for analysis from Semcor (Miller et al., 1993), which has WordNet sense annotations for 235,000 tokens from the Brown corpus (Francis and Kucera, 1979). We use the corpus reader from the Python Natural Language Toolkit (NLTK) (Bird and Loper, 2004) to access the corpus. The fact that syntax is represented within BERT embeddings (Hewitt and Manning, 2019) implies that BERT embeddings can easily capture distinctions in part of speech, so we focus on lexical ambiguity within part-of speech (specifically nouns and verbs) as a more challenging test case.

For the behavioral experiment, we selected word types across a range of sense entropy values (i.e., uniform to highly peaked sense distributions). We created a multinomial distribution over the senses of each lemma, i.e., (word type, part of speech) pair, and computed its entropy as follows: $-\sum_{s \in L} \frac{c_s}{c_L} \log(\frac{c_s}{c_L})$, where L is a lemma, s is a sense and c corresponds to a frequency, or count in the corpus. We removed stopwords and lemmas with zero entropy (79.3% of all lemmas). This yielded 444 word types.

For experimental stimuli, we chose 32 word types. 20 word types came from Semcor, including 11 low/medium entropy word types and 9 high entropy word types. We defined high entropy to be greater than 1.5 when rounded to the nearest tenth, and medium/low entropy to be less than this value, based on the distribution of sense entropy in the available Semcor data (in Figure 1). To account for variability in participants' placement of tokens in the spatial arrangement task, we needed to normalize measurements in the interface on a per-trial basis, so we selected words with three senses or more. We chose word types with varying entropy values to determine if this quantity had an effect on the correlation between BERT vectors and human judgements; relatedness between embeddings for words with more unpredictable sense distributions may be less consistent with results from the experiment. To avoid overwhelming participants, we selected words with fewer than eight senses. We also chose six additional words of theoretical importance that exemplify patterns of regular polysemy that have been observed across languages from Srinivasan and Rabagliati (2015), and manually chose six words with three senses, one of which was less semantically related to the other two. Because we expected these words to exhibit the greatest differences in pairwise similarity, we elicited judgments from all participants on these items. More details about how these stimuli were presented and exclusion criteria are in Appendix B.

3.1.1 Participants

105 undergraduate students from a major research university participated in the experiment and were compensated with class credit. Upon providing informed consent, participants reported their experience with English and other languages to ensure that data were collected from proficient English speakers (defined as at least 50% of daily language use); one participant was excluded using this criterion.

3.2 Experiment

From the set of 32 word types, each participant received 14 word types as stimuli. We used a two-dimensional spatial-arrangement task (Goldstone, 1994), because it allowed us to efficiently capture psychological judgments, and because tasks of this nature effectively capture relatedness in a high-dimensional semantic space (Richie et al., 2020). Participants were told to place less related sense tokens further apart from one another on the canvas, and more closely related tokens near each other. The experimental interface is shown in Appendix A. To place the sense tokens, participants were given definitions (from WordNet) and example sentences (from Semcor; in some cases shortened for brevity). The task was untimed, and participants were encouraged to introspect for as long as necessary about the meanings of the presented senses before making their placements. Partcipants were encouraged to adjust their placements to reflect all senses.

3.3 Relatedness Matrices

For each word, we collected spatial (x,y) positions of each of the word's senses from each participant. To derive an estimate of the relatedness of senses for each word type, we normalized the distances such that they were rescaled according to the largest reported pairwise distance. This controlled for variation in the absolute amount of space in the canvas that different participants used when making their placements. After excluding participants with unreliable responses (Appendix B), we then averaged relatedness matrices for each word type across the remaining participants. This yielded a single "aggregate relatedness matrix" for each word type which could then be directly compared with model results for a word type. To account for the larger number of participants reporting data for shared stimuli, we selected data from a random set of participants whose size corresponded to the average amount of test items ($n = 29$). Data from a minimum of 21 participants and a maximum of 37 participants were used to construct the aggregate relatedness matrix for each word type.

3.4 Modeling

We retrieved BERT embeddings for labeled tokens from the Semcor corpus (Miller et al., 1993) and ran classifiers to distinguish between different senses of individual words based on BERT vectors. We compared these data to results from the experiment based on two metrics: distance in the embedding space and classification accuracy, assessing their similarity through correlation (Budanitsky and Hirst, 2006).

Given a word type and a part of speech, we derived word embeddings from BERT as follows. For each sense of the (word type, part of speech) pair, we retrieved sentences from Semcor, and tokenized them using rules specified by the BERT authors (Devlin et al., 2019). We loaded a pre-trained BERT model `BERT-base` from Wolf et al. (2019) and ran the forward pass on each sentence, extracting the activations corresponding to the type, and storing the summed activations of the final four layers, which has produced strong results in word sense disambiguation (Loureiro and Jorge, 2019).

3.5 Word Sense Classifier

As a matter of due diligence, we first evaluated the effectiveness of BERT representations for word sense disambiguation (classification) in the Semcor corpus. While WSD-focused work has used the Semcor dataset for training and tested on other datasets (Raganato et al., 2017), we confirmed that BERT demonstrates similar levels of performance on Semcor. To this end, we assessed performance of multiclass logistic regression models to predict word senses for 401 word types from Semcor. Senses were omitted from this classification whenever there were fewer than 10 tokens. Each logistic regression model was trained to predict WordNet sense labels for each instance of a sense based on its contextual word embedding from BERT. We conducted 5-fold cross validation and applied L1 regularization to prevent overfitting. To confirm consistency with prior work on WSD (Loureiro and Jorge, 2019; Reif et al., 2019; Blevins and Zettlemoyer, 2020), we report the average F1 score (weighted for each sense) across the runs for each word type. Comparable F1 scores relative to this prior work would suggest that it is appropriate to further investigate the geometry implicit in these contextual word embeddings.

Method	All Types in Semcor	Types in Behavioral Expt.
Number of Available Types	401	32
Random Sense	0.480	0.423
Majority Sense	0.441	0.403
Logistic Regression on BERT Embeddings	**0.757**	**0.797**

Table 1: Average F1 scores (across types and train-test splits) for BERT-based word sense disambiguation in the Semcor corpus.

3.5.1 Extracting Cosine Distances Between Word Sense Centroids

After qualitatively reproducing previous word sense disambiguation results using BERT embeddings, we then tested whether the relationship between BERT embeddings of senses parallels human judgments of relatedness between senses. Because BERT generates one vector for each use of a word token, we computed the centroids of the BERT embeddings corresponding to each sense, and then compared different senses by using the cosine distance of these centroids to one another. In a broad range of models, cosine distances between vectors corresponding to word types can encode the degree of semantic relatedness (Dumais et al., 1988; Mikolov et al., 2013; Bojanowski et al., 2017), so we can use this same metric with BERT embeddings to compare word senses to one another. Following the same procedure as the relatedness matrices, we define relatedness as (1 - cosine distance), such that the largest cosine distance (least related pair) takes the value of 0 and the smallest cosine distance (most closely related pair) takes the value of 1 for each word type. We stored these relatedness measures between sense centroids in a matrix analogous to the aggregate relatedness matrices from the behavioral experiment. To evaluate their fit to human data, we computed the Spearman rank correlation coefficient between the upper triangular entries in these two distance matrices. A nonparametric measure of correlation is appropriate given that the relationship is not necessarily linear.

4 Results

First, we verified that BERT can discriminate between senses in Semcor comparably to other test datasets. We then compared relatedness estimates derived from distances in the BERT embedding space to the aggregate relatedness measures from the experiment. We then conduct two exploratory analyses, the first comparing the relatedness of homonymous and polysemous sense pairs, and the second examining the utility of pairwise sense confusion as a measure of relatedness.

4.1 Classification Results for All Semcor Data

We compared the distribution of F1 scores of the classifiers on the 401 words from Semcor (BERT failed to process 43 of the words) with two baselines: random choice and selecting the most frequent sense (Figure 1). Classification performance on all 401 words, including 25 words used in the behavioral experiment, were similar to those reported by Blevins and Zettlemoyer (2020), Loureiro and Jorge (2019), and Reif et al. Reif et al. (2019) (0.739, 0.754, and 0.711, respectively), which used test sets from Raganato et al. (2017). We find that type-wise F1 score is negatively correlated with sense entropy computed from Semcor ($\beta_{sense\_entropy} = -0.074, R^2 = 0.06$; Fig. 1).

4.2 Comparing Human Relatedness with BERT Word Sense Geometry

We computed the Spearman's rank correlation coefficient between human judgments of sense relatedness (from the aggregate relatedness matrices) and measures of word sense relatedness from the centroids of BERT contextualized word embeddings. This yielded a positive correlation (Spearman's $r = 0.565, p < 0.001$), with a 95% confidence interval of (0.459, 0.655). To confirm that this correlation was above that expected by random chance, we compared it to a random baseline established by 1000 draws of randomly generated sense placements for 29 participants (the average number of participants per item). This correlation was much lower (Spearman's $r = 0.062, p < 0.001$, 95% CI: -0.065 - 0.208). We report the correlation between relatedness measures from the embedding space and human judgements by part of speech and high vs. low/medium entropy in Table 2. Correlations between human relatedness
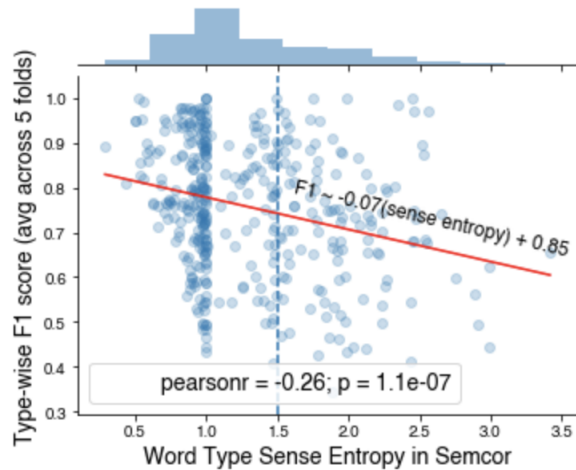
133

Figure 1: Classifier performance for word types with respect to their sense entropy in Semcor. F1 score for each type is averaged over cross validation folds and weighted by sense frequency. Vertical line represents boundary between low/medium and high entropy word types.
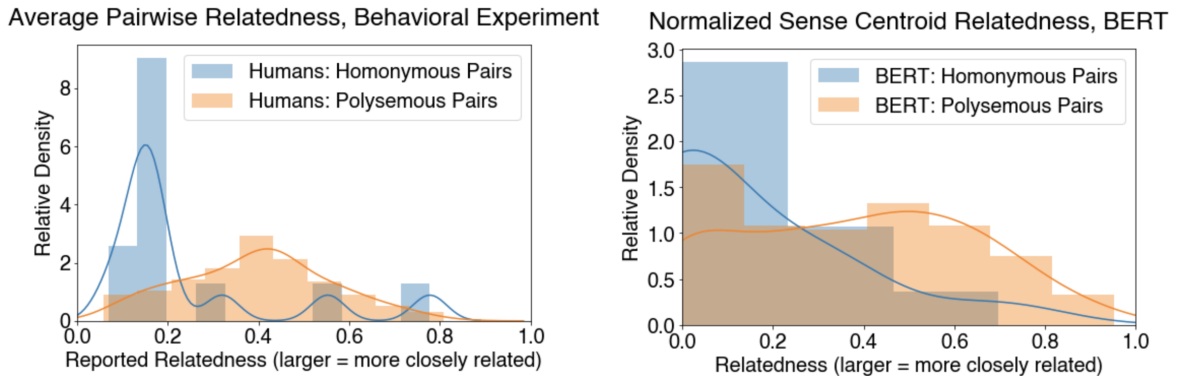

Figure 2: Distribution of distances between pairs of senses from the human data (left) and from BERT (right), across word types.

judgments and BERT relatedness were substantially higher among low to medium than high entropy words(Table 2). We also find that correlations among human and model relatedness are higher among verbs than nouns.

4.2.1 Polysemous vs. Homonymous Sense Pairs

We conducted an exploratory analysis to see if human relatedness judgments differed between homonymous and polysemous relations, and whether the geometry of BERT embeddings captured this distinction between word senses. In the absence of gold-standard datasets labeling polysemous and homonymous word sense relationships,[1] we tagged pairs of senses as polysemous or homonymous ourselves for the set of words in the behavioral experiment. For human judgments, we took the set of average distances between polysemous and homonymous pairs of senses. For BERT embeddings, we took the cosine distances between pairs of polysemous and homonymous sense centroids.

The density plots in Figure 2 show that participants in the behavioral experiment judged the polysemous pairs to be more similar (i.e., less distant) than homonymous pairs (Mann-Whitney $U = 516, p < 0.001$). BERT-based representations reproduced this basic pattern (Mann-Whitney $U = 493, p < 0.001$), though many relatedness estimates were more dispersed for both polysemous and homonymous pairs.

To illustrate the differences between how BERT represents polysemous and homonymous relationships, we show t-SNE and dendrogram visualizations in Figure 3. For `table.n` (Figure 3, top), we

---

[1]We could in principle use natural language dictionaries, some of which list polysemous and homonymous senses under separate entries. However, such definitions would not necessarily map to WordNet labeled data.
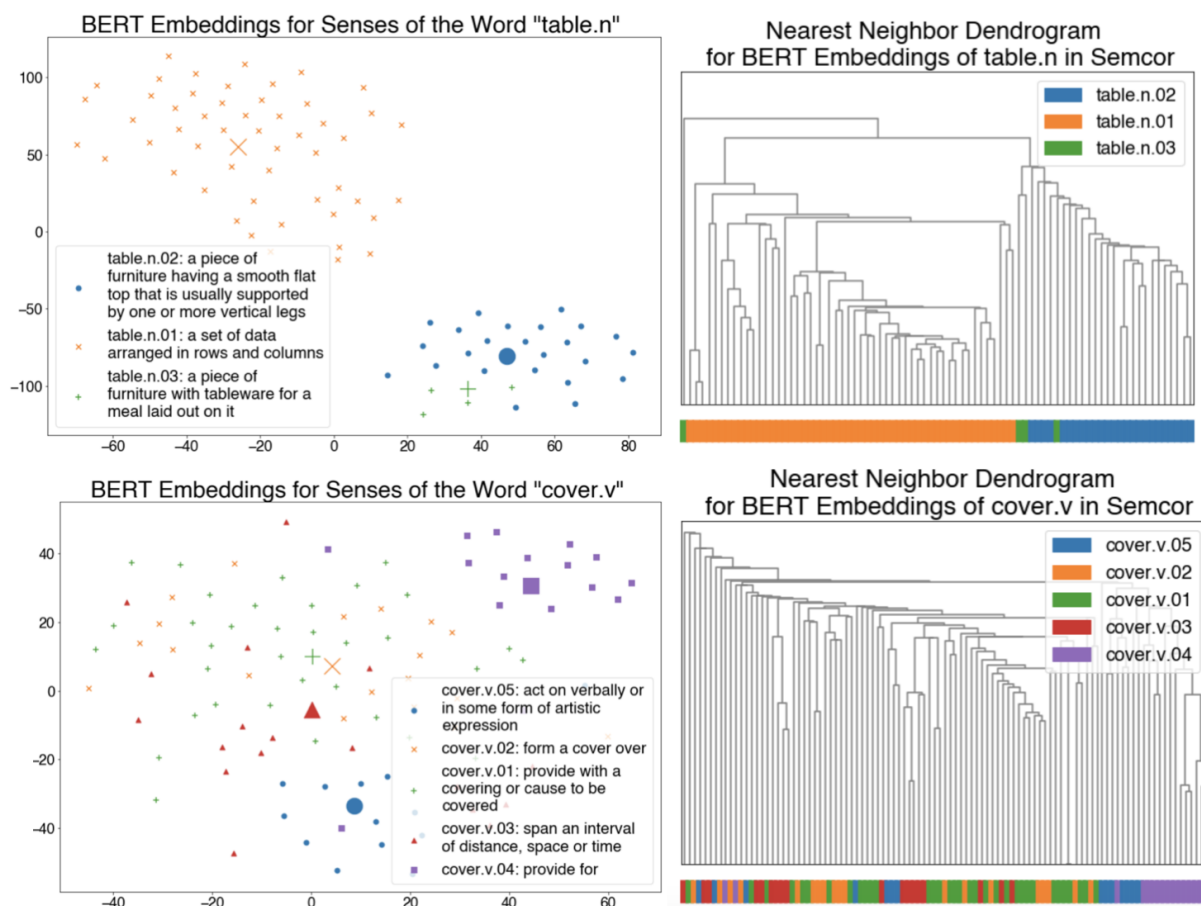
134

Figure 3: Visualizations of example BERT embeddings for stimuli `table.n cover.n`. Left column contains results of t-SNE dimensionality reduction with small points indicating word tokens and large points indicating sense centroids. Right column contains results of single linkage agglomerative clustering.

show that two senses related by polysemy (piece of furniture vs. a tablesetting) are judged to be closer together than homonymous pairs (either of the first two senses vs. "a set of data arranged in rows and columns"). For `cover.v`, all sense pairs were polysemous. Instances of the same sense were still closer to one another in the embedding space, but tokens corresponding to different senses were much less clearly distinguished except for `cover.v.04`, which may be well-separated from other senses because it refers to a metaphorical meaning of the lemma. We find that the BERT centroids' cosine distances and the aggregated relatedness judgements are strongly correlated (Spearman's $r = 0.851, p < 0.001$), with a 95% confidence interval of (0.638, 0.943).

4.3 Predicting Human Relatedness Judgments with Pairwise Confusion

One possibility is that the cosine distance computed over all dimensions of the BERT embedding space overlooks the possibility that some dimensions may be more or less useful for discriminating word senses depending on the word type. One alternative is to use the pairwise confusion probabilities from multiclass logistic regression (described in Section 3.5) as a measurement of pairwise relatedness in the model. This choice invokes substantive theoretical questions: in principle, sense relatedness and discriminability could be orthogonal. For example, two senses could be judged as closely related by humans, but they could in principle be able to discriminate between instances of those senses without errors. In the absence of human sense discrimination performance data, we leave this question to future work. Nonetheless, we investigated the utility of pairwise sense confusion as an alternative predictor for human judgments of sense relatedness.

135

Metric	Part of Speech		Word Sense Entropy	
	Nouns	Verbs	High	Low/Medium
Relatedness (1 - Normalized Cosine Distance)	0.487	0.623	0.345	0.570
Pairwise Confusion	0.609	0.627	0.518	0.731

Table 2: Spearman rank correlations between relatedness judgements and confusion matrices in BERT embedding space, split over part of speech and entropy level
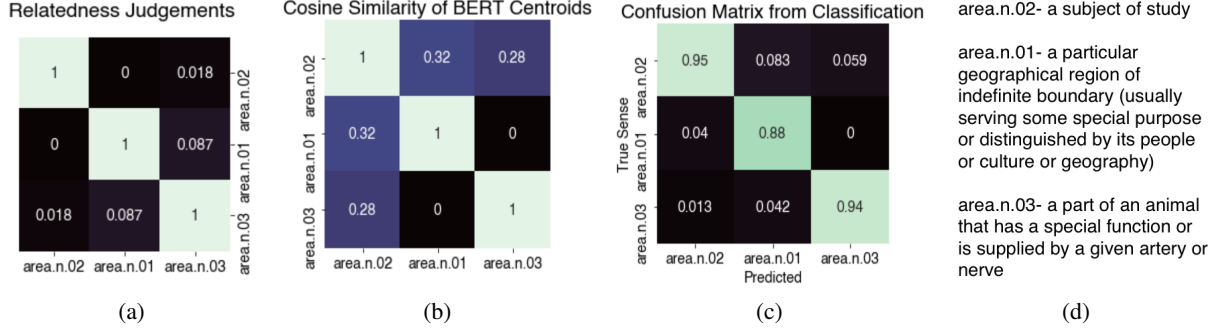

Figure 4: Mean human relatedness judgements, cosine similarity matrix, and confusion matrix for senses of `area.n`.

To this end, we summed the confusion matrices when senses from the stimuli are classified during each iteration of cross-validation from Section 4.1. We normalized 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. Across all word types with available data (20 out of 32 types, including 150 out of 189 sense pairs), we found a positive Spearman's rank correlation between entries in the confusion matrices and matrices of the corresponding relatedness judgements ($r = 0.649, p < 0.001$), with a 95% confidence interval of (0.592, 0.7). For this measure, we considered all entries in both matrices, as the confusion matrices are not necessarily symmetric. Because this analysis omits word senses with fewer than 10 tokens in Semcor, this evaluation reflects a smaller set of items than the one reported in Section 4.2 for cosine-based relatedness. When we computed the correlation between cosine-based relatedness and human aggregated relatedness matrices on this smaller set, the correlation remains lower (Spearman's $r = 0.518, p < 0.001$), with a 95% confidence interval of (0.391, 0.627). This suggests that pairwise confusion in BERT may be a stronger predictor of human relatedness judgments. We also report correlations stratified by part of speech and type sense entropy in Table 2. To illustrate the approach, we present aggregate relatedness matrices and confusion matrices for the word type `area.n` in Fig. 4. In this case, the experimental data (Fig 4, A) was more closely aligned with the confusion matrix (C) the than the cosine similarity of centroids (B).

We also evaluated sense classification accuracy on the basis of the homonymous and polysemous sense pairs from the preceding exploratory analysis. Among homonymous sense pairs, the classifiers achieved an average F1 score of 0.992. Among the polysemous sense pairs, the classifier achieved an average F1 score of 0.752.

5 Discussion

We investigated the ability of an artificial neural network model that represents word tokens with contextualized word embeddings to capture human-like relations among English word senses. On a subset of word types from the Semcor corpus, we reproduced previous word sense disambiguation results. We then showed that these same BERT embeddings capture a significant amount of information regarding the relationship between word senses, and are able to at least partially reproduce human relatedness judgments. An exploratory analysis revealed that BERT-based measures of the relatedness

between pairs of homonymous senses are much lower than for pairs of polysemous senses, matching human intuitions for the same set of senses. Pairwise confusion from the sense classifiers provided a slightly better predictor of human judgments of sense relatedness compared to distance in the embedding space, although this may be unsurprising given that pairwise confusion reflects some additional degree of supervision. Analyzing the error rates of classifiers trained on BERT embeddings of the stimuli, we also provided evidence that their results are more accurate for homonymous sense pairs compared to polysemous sense pairs.

Following in the approach of comparing BERT vectors to experimental data from human participants (Ettinger, 2020), the present research addresses how BERT represents word senses, bridging the gap between human and computational models of lexical ambiguity resolution. Recent progress showing BERT's performance on word sense disambiguation (Wiedemann et al., 2019; Loureiro and Jorge, 2019; Blevins and Zettlemoyer, 2020) indicates that BERT-based models perform better than past approaches, and our work specifies that the representations they use to accomplish this are relatively consistent with human intuitions. Existing exploratory work aims to analyze how BERT represents word senses (Reif et al., 2019; Lake and Murphy, 2020), but we systematically evaluate these claims with both WordNet senses and human data. Our work corroborates claims by Lake and Murphy (2020) that BERT captures relations between homonymous senses. Indeed, poor performance in discriminating polysemous pairs of words senses suggests that more work needs to be done on capturing polysemous relations; one possibility is to explore how BERT could be fine-tuned to capture relations exemplified by regular polysemy. One limitation from these findings is that the dataset only covers 32 word types, which we hope to address in future work.

More generally, our findings suggest that a continuous measure of sense relatedness derived from neural language models could potentially be used to augment existing lexicographic resources like WordNet, in this case providing a measure of relatedness between senses. We encourage other researchers developing sense ontologies, especially those found through word sense induction, to consider representing the relationships among word senses in addition to discovering new sense inventories.

6 Conclusion

Through comparing results from a behavioral experiment with data from contextualized word embeddings, we demonstrate that new Transformer-based neural network architectures may reflect human intuitions about the relationships between word senses, at least in English. Exploratory analyses suggest that these measures of relatedness between sense pairs reflect a distinction between polysemous and homonymous relationships between word senses, but that these models are much more effective in discriminating homonymous sense pairs than polysemous ones. By demonstrating basic levels of consistency with human judgements, we hope to stimulate further research that combines contextualized word embedding techniques with discrete, symbolic ontologies to develop a more cognitively informed model of the lexicon.

References

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

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

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

Terra Blevins and Luke Zettlemoyer. 2020. Moving down the long tail of word sense disambiguation with gloss informed bi-encoders. In *Proceedings of the 58th Association for Computational Linguistics*.

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

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

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

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

Susan T Dumais, George W Furnas, Thomas K Landauer, Scott Deerwester, and Richard Harshman. 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.

Kawin Ethayarajh. 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, November. Association for Computational Linguistics.

Allyson Ettinger. 2020. What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

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

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

Robert L Goldstone. 1994. Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: General*, 123(2):178.

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

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

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy, July. Association for Computational Linguistics.

Dan Jurafsky and James H Martin. 2014. Speech and language processing. vol. 3.

Ekaterini Klepousniotou, Debra Titone, and Carolina Romero. 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.

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

Brenden M. Lake and Gregory L. Murphy. 2020. Word meaning in minds and machines. *arXiv preprint, arXiv:2008.01766*.

Daniel Loureiro and Alípio Jorge. 2019. Language modelling makes sense: Propagating representations through WordNet for full-coverage word sense disambiguation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5682–5691, Florence, Italy, July. Association for Computational Linguistics.

John Lyons. 1995. *Linguistic semantics: An introduction*. Cambridge University Press.

Lucy J MacGregor, Jennifer Bouwsema, and Ekaterini Klepousniotou. 2015. Sustained meaning activation for polysemous but not homonymous words: Evidence from eeg. *Neuropsychologia*, 68:126–138.

Timothee Mickus, Denis Paperno, Mathieu Constant, and Kees van Deemter. 2020. What do you mean, bert? assessing BERT as a distributional semantics model. *Proceedings of the Society for Computation in Linguistics*, Vol. 3, Article 34.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

George A Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J Miller. 1990. Introduction to wordnet: An on-line lexical database. *International journal of lexicography*, 3(4):235–244.

George A Miller, Claudia Leacock, Randee Tengi, and Ross T Bunker. 1993. A semantic concordance. In *Proceedings of the workshop on Human Language Technology*, pages 303–308. Association for Computational Linguistics.

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

James Pustejovsky. 1998. *The generative lexicon*. MIT press.

Hugh Rabagliati and Jesse Snedeker. 2013. The truth about chickens and bats: Ambiguity avoidance distinguishes types of polysemy. *Psychological science*, 24(7):1354–1360.

Alessandro Raganato, Jose Camacho-Collados, and Roberto Navigli. 2017. Word sense disambiguation: A unified evaluation framework and empirical comparison. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 99–110.

Christian Ramiro, Mahesh Srinivasan, Barbara C Malt, and Yang Xu. 2018. Algorithms in the historical emergence of word senses. *Proceedings of the National Academy of Sciences*, 115(10):2323–2328.

Emily Reif, Ann Yuan, Martin Wattenberg, Fernanda B Viegas, Andy Coenen, Adam Pearce, and Been Kim. 2019. Visualizing and measuring the geometry of bert. In *Advances in Neural Information Processing Systems*, pages 8594–8603.

Russell Richie, Bryan White, Sudeep Bhatia, and Michael C Hout. 2020. The spatial arrangement method of measuring similarity can capture high-dimensional semantic structures. *Behavior Research Methods*, pages 1–23.

Jennifer M Rodd, M Gareth Gaskell, and William D Marslen-Wilson. 2004. Modelling the effects of semantic ambiguity in word recognition. *Cognitive science*, 28(1):89–104.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how bert works. *arXiv preprint, arXiv:2002.12327*.

Mahesh Srinivasan and Hugh Rabagliati. 2015. How concepts and conventions structure the lexicon: Cross-linguistic evidence from polysemy. *Lingua*, 157, 01.

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

David Tuggy. 1993. Ambiguity, polysemy, and vagueness. *Cognitive linguistics*, 4(3):273–290.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Gregor Wiedemann, Steffen Remus, Avi Chawla, and Chris Biemann. 2019. Does bert make any sense? interpretable word sense disambiguation with contextualized embeddings. *arXiv preprint, arXiv:1909.10430*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint, arXiv:1910.03771*.

Yang Xu, Khang Duong, Barbara C Malt, Serena Jiang, and Mahesh Srinivasan. 2020. Conceptual relations predict colexification across languages. *Cognition*, 201:104280.
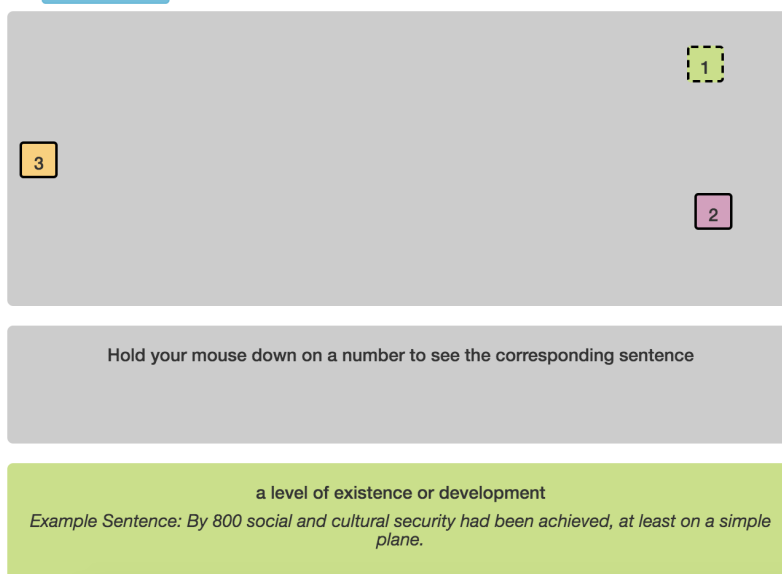
Appendices

A   Experiment Details

Figure 5: A sample trial in the experiment

B   Trial Types and Exclusion Criteria

The first two of the eighteen trials participants received were presented as *training* trials to ensure that participants were familiar with the interface; data from these trials were discarded. Participants received a mixture of *shared* trials completed by all participants, and *test* trials where only a subset of participants provided judgments. The inclusion of both trial types gave us a consistent set of word types for which all participants contributed data (allowing us to evaluate which participants gave unreliable responses compared to all other participants), while at the same time characterizing a broad set of word types with more sparse responses. Test trials were drawn from the set of 26 lemmas consisting of words from Srinivasan and Rabagliati (2015). To identify participants who provided low-quality data, we computed hold-one-out correlations for each participant using their relatedness matrices for the shared trials, and excluded participants whose data had a rank correlation with the hold-one-out averages that was lower than 0.4. This threshold corresponds roughly to the 92nd percentile of scores if sense tokens are placed randomly in the interface.

In addition to the shared and test trials, participants also saw two *repeat* trials, drawn from the same set as the test trials, to evaluate the reliability of their responses within the same testing session. Participants with rank correlations in reported distances between their original and repeat trials lower than 0.2 were excluded from analysis, corresponding to the 70th percentile of correlations if sense tokens are placed randomly (this threshold was lower because of variation in both the number of senses and the recency of the test trial). Ten participants who failed to meet both criteria were excluded from further analysis.