Contextual Embeddings Can Distinguish Homonymy from Polysemy in a Human-Like Way

Kyra Wilson [1] and **Alec Marantz** [1,2]
[1] Research Institute, New York University Abu Dhabi
[2] Departments of Psychology and Linguistics, New York University
{kyra.wilson, marantz} @ nyu.edu

Abstract

Lexical ambiguity is a pervasive feature of natural language, and a major difficulty in understanding language is selecting the intended meaning when more than one are possible. Despite this difficulty, many studies of single word recognition have found a processing advantage for ambiguous words compared to unambiguous ones. This effect is not homogeneous however–studies find consistent advantages for polysemes (words with multiple related meanings), and inconsistent results for homonyms (words with multiple unrelated meanings). Complicating this is the fact that most measures of ambiguity are derived from human- annotated or curated lexicographic resources, and their use is not consistent between studies. Our work investigates whether contextualized word embeddings are able to capture human-like distinctions between senses and meanings, and whether they can predict human behavior. We reanalyze data from previous experiments reporting ambiguity (dis)advantages using the lexical decision times reported in the English Lexicon Project. We find that our method does replicate the polyseme advantage and homonym disadvantage previously reported, and the predictors are superior to binary distinctions derived from lexicographic resources. Our findings point towards the benefits of using continuous-space representations of senses and meanings over more traditional measures. Additionally, we make our code publicly available for use in future research.

1 Introduction

Distributed representations of meaning (word embeddings) have brought great advancements to many natural language processing tasks including sentiment analysis, text summarization, and translation, to name just a few. Outside of computational applications, these embeddings have also been used successfully in psycholinguisitcs to predict semantic priming data (Ettinger and Linzen, 2016), eye-tracking data (Søgaard, 2016), and even neural activations (Honari-Jahromi et al., 2021). Despite their success in a wide variety of fields and tasks, these static representations' performance is greatly limited because each orthographic wordform is limited to only one vector representation.

This is problematic because ambiguity is quite pervasive in natural language. Words that have the same spelling can have multiple senses (polysemes) and/or meanings (homonyms). For example, in Wordsmyth online dictionary, *slam* has two entries (meanings). Under the first entry there are multiple senses–a noun meaning "sharp criticism", a verb meaning "shut something loudly", an additional noun meaning "the sound made by shutting something loudly", and others. Under the second there are additional senses unrelated to the first entry's senses–a noun meaning "winning of all tricks in a card game" and another noun meaning "a poetry reading event." Despite all these different usages, *slam* has just a single unchanging spelling and pronunciation.

To successfully use language, both humans and language models must somehow be able to select a single meaning from a set of multiple candidates for ambiguous words. For models based on static representations, this was not a straightforward task because all of the possible senses and meanings were collapsed into a single representation that was used invariantly across any possible context. There were no senses or meanings for the model to choose among. This flaw impacts a model's ability to understand the true meaning of words when they are used in changing contexts.

One recent advancement to address this problem is the use of contextualized word embeddings such as ELMo and BERT. Instead of having a single representation per wordform, these systems produce embeddings that change dynamically based on the surrounding context of a single word occurrence.

This way, *slam* used in the sentences *When I'm mad I slam the door* and *I attended the poetry slam last night* will have two distinct representations despite sharing the same orthographic form. The use of these contextualized embeddings have provided even further progress in a wide variety of computational applications because they successfully address the ambiguity problem.

While contextualized embeddings and transformer architectures have begun to be adopted in the analysis of human language processing (Jain and Huth, 2018; Kumar et al., 2022; Heilbron et al., 2021), this body of work has largely focused on language processing in context, rather than at the single-word level. In this work, we attempt to show that contextual embeddings can also be useful for analyzing human language processing even in the absence of context by looking specifically at the "ambiguity advantage". This is a widely studied psycholinguistic phenomenon in which ambiguous words are recognized faster than unambiguous ones in lexical decision experiments[1].

In previous work investigating the ambiguity advantage, senses and meanings have often been distinguished based on how they are listed in lexicographic resources (Rodd et al., 2002). Meanings generally correspond to dictionary entries, while senses will correspond to the various distinguished uses within those entries. (Following the previous example, *slam* would have two meanings, and at least five senses). Additionally, senses are generally assumed to be related and share some semantic core between them, while meanings have no shared semantics and are unrelated. For example, two of *slam*'s senses both have something to do with shutting something and making a noise–one is the action and one is the resulting sound; clearly there is a shared semantic core here. However, it is not clear that there is a semantic core shared between these senses and the senses of the other meaning, such as winning tricks in a card game.

Even though lexicographic resources do distinguish senses and meanings, using them to study the ambiguity advantage is challenging because they typically lack explicit criteria or explanations to why particular distinctions of relatedness are made. For example, a *door slam* and a *slam of the production* are considered related to each other according to Wordmsyth online dictionary (even though the latter doesn't necessarily have any meaning related to shutting something or a noise), but neither are related to a *poetry slam*.

Understandably, extensive discussions of the nature of semantic relatedness is typically outside the scope of most lexicographic resources, but this means they are not very well-suited to psycholinguistic research where such distinctions are of great importance. For this purpose, a more useful measure of a word's senses and meanings would be derived from the way speakers use the words at present as opposed to lexicographer categorizations and would have clear criteria for what makes something a sense versus a separate meaning.

In this study, we used BERT to derive a new measure of a word's numbers of senses and meanings [2], and we apply this measure to previously gathered lexical decision data. We compare our results to those from a previous study which quantified ambiguity using lexicographic resources and find that ours perform at least as well. This points to the benefits of further adopting contextualized embeddings for use in psycholinguistic research.

2 Related Work

Comparing the way that ambiguous and unambiguous words are processed can give information about the organization of the mental lexicon and ways in which different kinds of words may be retrieved and recognized. This is primarily tested in lexical decision experiments, where a mix of target stimuli and non-words are presented one at a time, and participants respond as quickly as possible with whether or not they recognize the presented string. Their reaction times on the target stimuli are then analyzed to determine what variables make word recognition easier (faster response times) or harder (slower responses times). These experiments generally reveal that there is, in fact, a difference in the way words with multiple senses and/or meanings are processed as compared to unambiguous words.

Most studies find that multiple senses facilitate recognition, as evidenced by a faster reaction time for words with multiple senses in a lexical decision paradigm (Borowsky and Masson, 1996; Hino and Lupker, 1996) compared to unambiguous words.

[1] This finding has been observed in both single word recognition tasks (Rodd et al., 2002; Borowsky and Masson, 1996; Hino and Lupker, 1996) and sentence presentation contexts (Frazier and Rayner, 1990; Klepousniotou, 2002). In this work, we focus exclusively on advantages for single word recognition.

[2] The tools developed for this study are available at `https://github.com/kyrawilson/word-senses-from-CWE`.

This lends support to a model of word recognition where words with multiple senses have multiple semantic representations, leading to easier recognition as a result of increased semantic activation compared to words with fewer senses.

However, the same result is not as consistent for words with multiple meanings. Some studies find increased reaction time compared to unambiguous words (Rodd et al., 2002; Beretta et al., 2005) (suggesting that having multiple unrelated meanings may make recognition more difficult because of competing activations), while others find an equivalent advantage for both multiple senses and multiple meanings (Hino et al., 2010; Pexman et al., 2004). Because the results are mixed, it is unclear whether words with multiple meanings are stored and accessed in a way similar to those with multiple senses, or whether they are different in some critical way.

It has been proposed that the contradictory results for words with multiple meanings are a consequence of differing methodologies in selecting ambiguous stimuli (Haro and Ferré, 2018). Namely, experimenters use a variety of sources for selecting ambiguous words because there is no gold standard resource for differentiating between related senses and related meanings; thus differences arise not only in what sources are used, but also in what the individual sources classify as ambiguous words since they are curated by different groups using varied techniques. This paper shows that that with advances in distributed representations of meanings, previous measures that relied on lexicographic sources can be exchanged for measures derived from contextual representations (specifically contextualized meaning vectors from BERT, a transformer-based language model) without reference to any outside resources, and these measures will perform at least as well as traditional ones.

There have been previous attempts to identify information about word senses from BERT embeddings. Reif et al. (2019) sampled sentences from Wikipedia and found that similar contextual usages of words tended to cluster together in meaning vector space and that the spatial location of a word could be changed by altering the context sentence. This suggests that BERT is able to represent meaningful semantic information within a subset of its vector dimensions [3]. Following Reif et al. (2019), there have been multiple attempts to use BERT for word sense disambiguation, including some which also use lexicographic resources to interpret the disambiguated senses (Wiedemann et al., 2019; Du et al., 2019; Vial et al., 2019).

In addition, there has also been research investigating how BERT represents words with different numbers of senses and meanings. Garí Soler and Apidianaki (2021) investigated both whether BERT could distinguish words with a single versus multiple senses and whether the senses cluster in interpretable ways. First, they found that usages of words with a single sense (according to WordNet) had a higher similarity than words with multiple senses. Furthermore, they used a k-means algorithm to cluster senses of ambiguous words, and they found that the quality of this clustering was high and correlated with annotator judgements about sense similarities. Although this study did demonstrate the potential of using clustering to analyze BERT embeddings, the use of the k-means algorithm is suboptimal because the number of clusters must be known a priori, and thus does not extend well to applications in which human annotations are unavailable or contradictory.

There has also been work investigating how BERT's representations of polysemy may correspond to humans'. Nair et al. (2020) collected human judgements of meaning relatedness for homonyms and polysemes and compared them to distances in BERT embedding space. They found that homonym meanings were more reliably distant than related senses. This suggests that the way BERT represents information is somewhat consistent with human intuitions. However, the experimental task in this study was metalinguistic: people were asked about how they use language, which may or may not be consistent with actual language use.

An additional test of how well BERT corresponds with human language would be to use it to predict actual human behavior rather than intuitions. Therefore in our study, we explore BERT's similarities to human language knowledge, analyzing behavioral reaction time data to potentially ambiguous and polysemous words and correlating human reaction times to the numbers of senses and meanings derived from BERT embeddings.

[3] A similar result was observed by Thompson and Mimno (2020) in the topic modeling domain.

3 Methods

3.1 Data

For the 182 words (124 ambiguous and 58 unambiguous) used in the first experiment of Rodd et al. (2002), we retrieved their mean reaction time in a visual lexical decision experiment from the English Lexicon Project (ELP) (Balota et al., 2007). These reaction times were used as the response variable in a linear regression model.

The words used in this experiment were selected by Rodd et al. (2002) to amplify the differences between ambiguous and unambiguous words. Of the 124 ambiguous words, 113 were taken from the Twilley et al. (1994) homograph norms, and the remaining 11 were judged to have similar properties. Most of these words were judged to have two or three meanings according to the original annotations, where meanings and senses were conflated, and half of them had two distinct entries in the Wordsmyth dictionary (corresponding to two meanings). The other half only had a single entry; the other "meaning" was annotated by Wordsmyth as a sense instead. This difference in the two groups allowed for a comparison of meaning relatedness. The words with two Wordsmyth entries were considered ambiguous (homonyms) while the remaining 58 words in the stimuli set were identified as being unambiguous (polysemes, since they were judged to have multiple senses) and had only one meaning.

We also included a number of control variables in our analysis in line with Rodd et al. (2002), including log word frequency, length, orthographic neighborhood, and concreteness. These were also collected from ELP.

3.2 Number of Senses

Our method for deriving the number of senses for a word assumes that same senses will be used in similar contexts, and therefore the contextual embeddings for a word in a particular sense will also be similar to each other. Furthermore, other senses will have dissimilar enough contexts that we can derive a measure of the number of senses by applying a clustering algorithm (HDBSCAN) to the BERT embeddings, where the identified clusters will correspond to individual senses of a word.

HDBSCAN (Campello et al., 2013) is a hierarchical clustering algorithm which uses the stability and persistence of clusters in order to select an optimal clustering from the hierarchy. It works by first identifying areas of high and low density points and deriving a distance (mutual reachability) metric that amplifies the distance to sparse points. Next, a minimum spanning tree is constructed using the mutual reachability distance and then converted into a hierarchy by sorting the edges in increasing order and creating a new cluster for each edge. Finally, a single clustering is selected from the hierarchy by selecting the clusters with the greatest stability, meaning that for a large range of distance values the cluster remains as a whole and does not split into two smaller clusters.

The use of HDBSCAN is particularly suited to the clustering of word senses for two reasons [4]. First, the algorithm allows extreme outlier points to be categorized as noise rather than coercing them into a cluster. This is good for our application because of the flexibility of language. Even though words have a generally standard and accepted set of meanings, there is nothing to prevent novel usages of a word in a new context. For our purposes, we would like to avoid including very low-frequency senses or meanings which are unlikely to be known by a majority of speakers.

Additionally, the only hyperparameter of the algorithm is the minimum number of points a cluster must contain, in contrast to other clustering algorithms in which the number of clusters must be specified a priori. We are interested in deriving the number of different senses from an unlabelled corpus rather than simply identifying the sense clusters which correspond to entries in lexicographic resources. Another side effect of this is that we are able to specify how many usages a particular sense must have in order to be considered well-known and avoid contaminating our clusters with too many "noise usages". We specified that our clusters should contain, at minimum, at least one percent of the points in the total number of embeddings for a given word.

Following Reif et al. (2019), we first sampled 1,000 occurrences of each word in Rodd et al. (2002)'s stimuli set from English Wikipedia[5], and used the publicly available pre-trained BERT$_{\text{BASE}}$ model (Devlin et al., 2019) in combination with the Hugging Face (Wolf et al., 2020) and Flair li-

[4] A related algorithm, DBSCAN (Ester et al., 1996), has also been shown to have success in clustering word embeddings (Mohammed et al., 2020). We chose to use HDBSCAN due to its increased flexibility over DBSCAN.

[5] For one word (*poach*), there were only 578 occurrences in Wikipedia. We used all of the occurrences in this case.

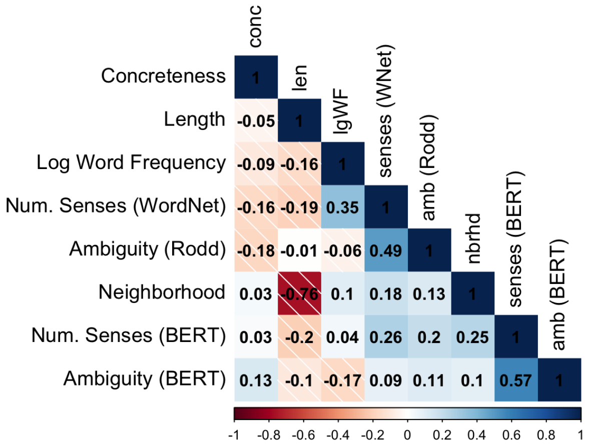

Figure 1: Spearman's rank correlation between all predictors.

braries (Akbik et al., 2019) to encode the word in their context sentences[6]. The word token of interest was then extracted from each context sentence and its layers were averaged, resulting in a single 768-dimension embedding for each sentence [7].

Finally, the embedding dimensions were reduced from 768 to two using t-SNE (Van der Maaten and Hinton, 2008). HDBSCAN is not guaranteed to perform well for high-dimensional data, so we chose to have it operate over embeddings that were also used for visualization in order to aid with interpretability of the clustering results. For each word, the minimum cluster size was one percent of the total number of embeddings for that word.

3.3 Ambiguity

Since there can be multiple senses of a word within a single meaning, we were interested in identifying any superstructure amongst the clusters which might correspond to different meanings. Broadly, to identify meanings, we are now aiming to cluster the senses of words themselves rather than the individual usages as an attempt to join senses that are most similar to each other. We do this by only clustering a subset of the points used in the the number of senses calculation as well as increasing the minimum cluster size hyperparameter in the HDBSCAN algorithm. This way, we are able to use the same algorithmic approach to derive unique measures for number of senses and ambiguity.

To begin, we select a subset of points to use for identifying meaning clusters. This is done in order to make the data sufficiently different to avoid recreating identical senses clusters as well as eliminating possible noise usages from the meaning clustering. The subset of points we used were those identified as "exemplars" by HDBSCAN within each of the identified sense clusters. In this implementation, exemplar points are those which persist in their cluster for the largest range of distance values and which are generally centrally located in their respective clusters. In other words, the exemplars are the points which are identified as being the strongest members of the cluster and least likely to be noise.

After identifying the set of exemplar points for each cluster, we used HDBSCAN clustering again in order to identify any potential higher order clusters. In contrast to the number of senses clustering, in this iteration we allowed the clustering algorithm to assign all the exemplar points to a single cluster, under the assumption that some subset of the stimuli are unambiguous and should thus have only one meaning.

Another difference between the ambiguity clustering and the number of senses clustering is the minimum cluster size. It has been observed that there is interpretable structure even within sense clusters (Reif et al., 2019). For example, for the word *die*, Reif and colleagues found that within a single sense cluster there was a separation relating to the number of people who died. We wanted to avoid the formation of even more granular sense clusters, so in this iteration we set the minimum cluster size to be the size of the smallest set of exemplar points from a single sense cluster. Finally, if the clustering procedure still resulted in a larger value for ambiguity (number of meanings) than the number of senses, we assigned the number of meanings to be equal to the number of senses post-hoc.

4 Results

4.1 Qualitative Analysis

An example of the clustering of senses and meanings can be seen in Figure 2 for the word *tent*, which has three senses and one meaning according to our proposed method. The three different shapes indicate that there were three senses identified–one that has to do with *tent* as a physical object used

[6] We only selected from sentences in which the target word appeared a single time.

[7] Multiple studies have shown that semantic representations differ depending on the BERT layer (Garí Soler and Apidianaki, 2021; Jawahar et al., 2019). While we averaged all layers together, it is possible that selecting a single layer would yield higher performance. We leave this investigation for future work.

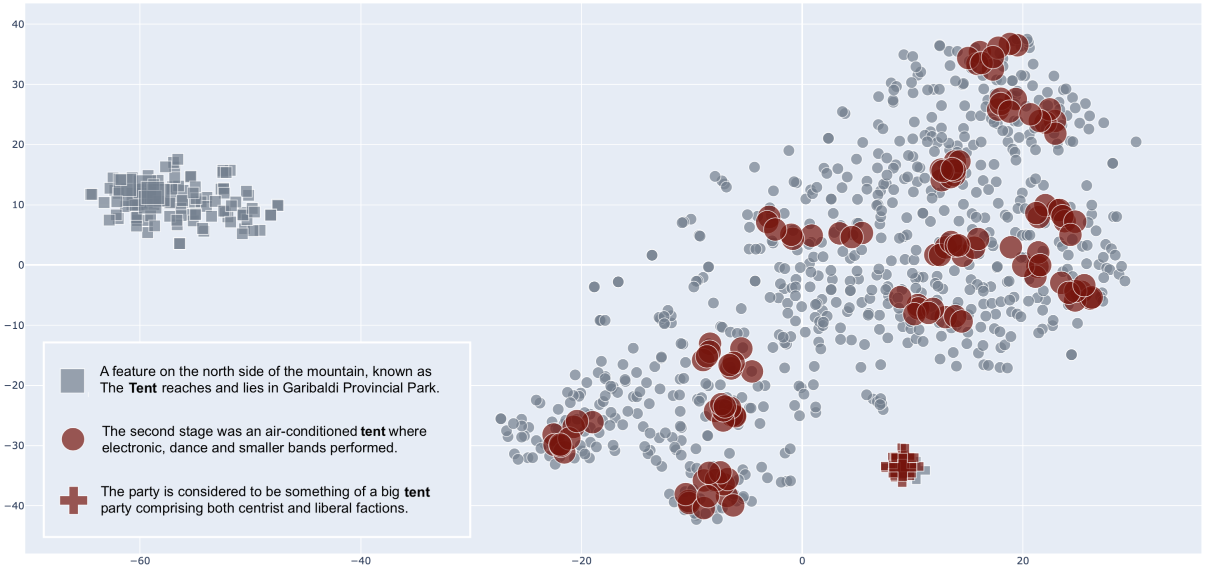

Figure 2: Example of sense and meaning clustering for *tent*, including example usages in sentences from Wikipedia. Different senses are indicated by different shapes, exemplar points used to cluster meanings are larger, and colored points indicate meaning groups.

for shelter, one that is a part of the phrase *big tent party*, and a third where *tent* is part of a title or used as a proper noun. The senses are accurately separated into groups that have internal cohesion, but separated from other groups with slightly different semantics.

The single group of red points indicate that *tent* has only one meaning combining the "physical object" and "political party" uses. Although these senses are not interchangeable, they are clearly related. Just as people might congregate under tents at a concert, they also metaphorically congregate in a *big tent party*. The third sense, however, is both unrelated to this meaning and also not cohesive enough to form its own separate meaning. The cluster contains titles and other proper nouns usages, so *tent* is both unlikely to have a single shared context corresponding to a new meaning in this cluster or a context close enough to the other two senses that it should be included in the first meaning. Therefore, it is correctly identified as "noisy" usages of *tent* and not analyzed as an additional meaning.

4.2 Number of Senses

To begin, we compared the BERT-derived number of senses to the number of senses as indicated in WordNet[8]. There was a weak positive correlation between the BERT-derived number of senses and the number of senses reported by WordNet (ρ = 0.26), as shown in Figure 1. We entered both predictors into a linear regression model with response time in a lexical decision task as reported in the ELP as the dependent variable. The full model results are shown in Figure 3.

Only the number of senses as derived from BERT was a significant predictor of reaction time, and the effect replicated what has been reported in previous studies. Words with more senses were generally recognized faster than those with fewer senses. This effect can be seen in Figure 4. Next we performed an ANOVA to assess whether additional variance is explained by our predictor. As expected, the ANOVA indicated that including the number of senses derived from the contextual embeddings did improve the model fit ($F = 3.78$, $p = 0.05$).

4.3 Ambiguity

We compared the binary ambiguity variable used by Rodd et al. (2002) with our continuous variable derived from contextualized embeddings. There was low correlation between the binary ambiguity variable and our BERT-derived variable ($\rho = 0.11$). In the model with none of our predictors, we did not replicate the ambiguity effect reported by Rodd et al. (2002). In fact, we found the opposite; Rodd et al. (2002) reported an inhibitory effect where ambiguous words were recognized more slowly than unambiguous words, but our analysis showed that ambiguity made reaction time faster (just as

[8] The Rodd et al. (2002) study did not indicate how the number of senses was calculated, so we used WordNet as an approximation of their metric.

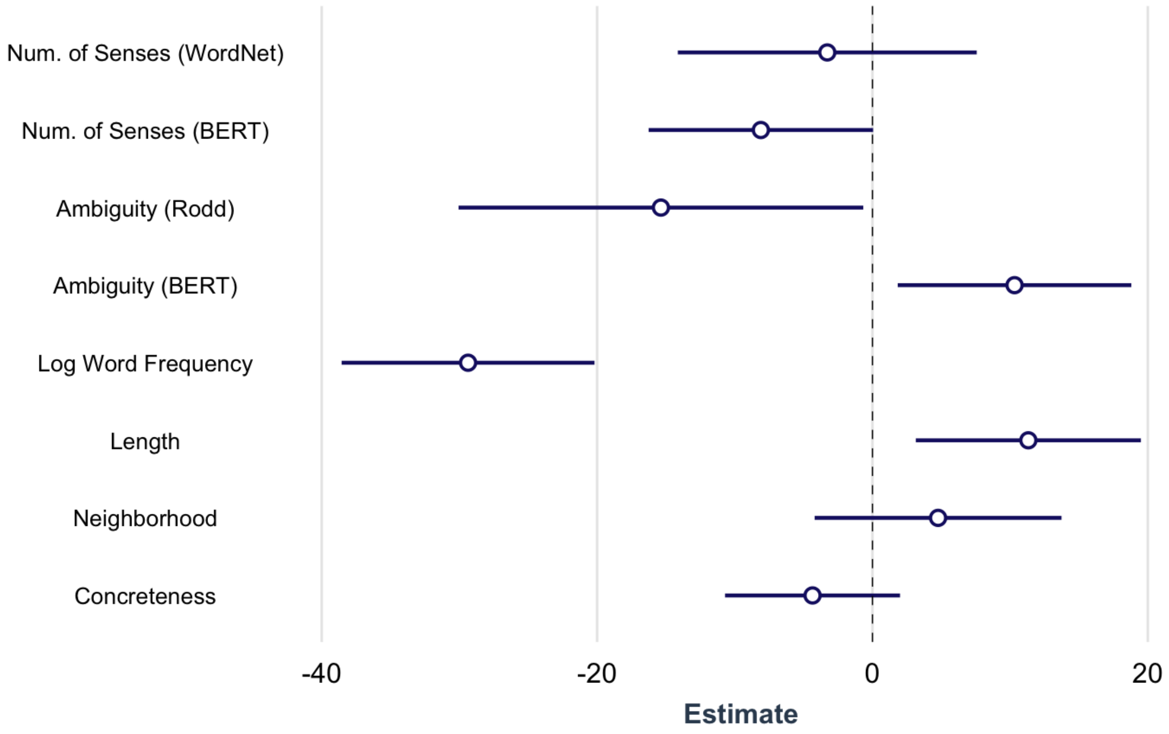

Figure 3: Estimates of linear regression coefficients predicting reaction time. The significant predictors are the number of senses and meanings derived from contextual embeddings, the binary ambiguity variable from Rodd et al. (2002), log word frequency, and length.

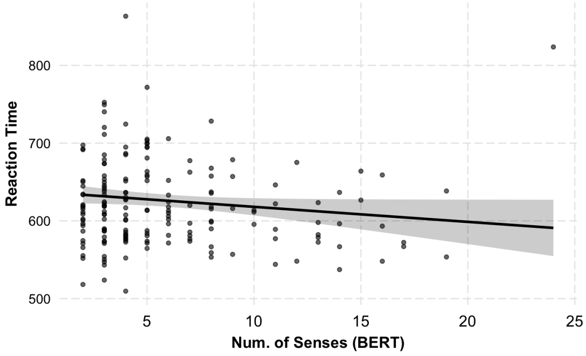

Figure 4: Regression line showing inverse relationship between number of senses and reaction time.

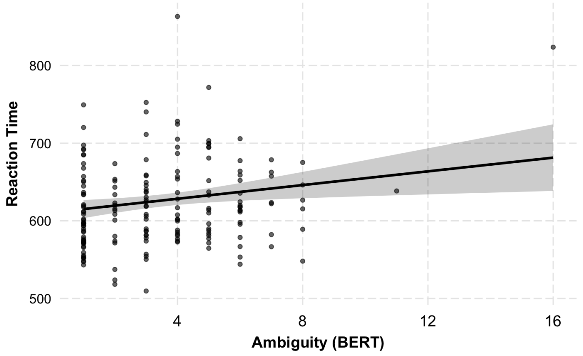

Figure 5: Regression line showing direct relationship between ambiguity and reaction time.

multiple senses facilitate recognition).

However, when we included our predictors, we found that the ambiguity variable as derived from BERT did produce an inhibitory effect as originally reported, as shown in Figure 5. Compring models with and without the ambiguity in ANOVA showed that our predictors also significantly improved the fit of the model ($F = 6.61$, $p = 0.01$).

5 Discussion

There are multiple important results from this investigation. First, we found that the contextual embeddings not only correspond to human judgements as previously reported (Nair et al., 2020), but also to human behavior. Our number of senses measure replicated the well-reported finding that having multiple senses is facilitatory in word recognition (more senses lead to faster identification). In fact, for this particular set of stimuli, our measure outperformed the more traditional measure derived from WordNet in predicting reaction times in a lexical decision task.

The results for ambiguity (number of meanings) are slightly more complex. First, our analysis did not replicate the original results when using the binary ambiguous/unambiguous variable as computed by Rodd et al. (2002). We instead found an additional facilitatory effect for this variable where multiple meanings correspond to faster recognition as compared to single meanings. However, our number of meanings variable, derived from clustering senses using sense exemplars, did result in words with more meanings having slower reaction times as previously reported, above and beyond the effects of the binary variable. For theories of word recognition, it is not immediately apparent why these two variables should have opposite effects, but as our measure has consistent criteria and clear definitions for deriving predictor values, further experiments should be able to investigate this in depth using a wider variety of stimuli.

Finally, another interesting outcome worth further investigation is that our results were obtained using only two-dimensional embeddings derived from BERT. Previous experiments investigating the representations of polysemy and ambiguity within BERT have done so using all 768 dimensions of the embeddings (Reif et al., 2019; Garí Soler and Apidianaki, 2021), while our experiment suggests similar information can be represented using far fewer dimensions. Determining the optimal number of dimensions for representing polysemy and ambiguity using BERT remains an open question worth further study.

The replication of the previous ambiguity advantage results show how contextual embeddings such as BERT can be useful in the analysis of experimental data. For the number of senses advantage, we showed a stronger effect than more traditional predictors relying on lexicographers. For the number of meanings, we also replicated previous findings and found that our predictor performed just as well as traditional ones. However, because our predictor was derived from unlabelled corpora without resorting to any human annotation (which may introduce bias) we find it methodologically superior to predictors derived from lexicographic resources such as dictionaries and WordNet. We think that continuing to use contextual embeddings to derive predictors will facilitate transparency and replicability across many different areas of linguistic research as well as allowing for more flexibility in what words and languages are able to be studied.

Finally, another potential benefit of this methodology is the possibilities of extending it to languages other than English. Generating high-quality lexicographic resources is very time- and labor-intensive, so current research into the ambiguity advantage is limited to those languages which already have such resources. Our methodology, on the other hand, could theoretically be extended to any language which has a pre-trained model able to produce contextual embeddings (or for a slightly higher cost, any language for which a new contextual embedding model could be trained and deployed), and further research should be done to verify that the properties of BERT embeddings observed in this experiment would also be present in models trained on other languages.

6 Conclusion

This study further supports work which indicates that contextualized embeddings contain information which is able to predict human language processing. We extended the approaches of earlier work by not only deriving a measure of how many senses a word has, but also finding how many distinct meanings a word has by clustering those senses. We used these numbers to replicate the finding that multiple senses facilitate recognition in a lexical decision experiment and add support to the finding that multiple meanings inhibit word recognition. This is an important result because it suggests this method can be used as a replacement for traditional ways of deriving measures of ambiguity and polysemy, allowing for standardization of variable predictors across experiments in order to facilitate comparison and minimize conflicting results.

Acknowledgements

This research was supported by the NYUAD Research Institute under Grant G1001 and was carried out on the High Performance Computing resources at New York University Abu Dhabi. We additionally thank Tal Linzen and anonymous reviewers for their guidance and feedback on earlier versions of this paper.

References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for*

Computational Linguistics (Demonstrations), pages 54–59.

David A Balota, Melvin J Yap, Keith A Hutchison, Michael J Cortese, Brett Kessler, Bjorn Loftis, James H Neely, Douglas L Nelson, Greg B Simpson, and Rebecca Treiman. 2007. The english lexicon project. *Behavior research methods*, 39(3):445–459.

Alan Beretta, Robert Fiorentino, and David Poeppel. 2005. The effects of homonymy and polysemy on lexical access: An MEG study. *Cognitive Brain Research*, 24(1):57–65.

Ron Borowsky and Michael EJ Masson. 1996. Semantic ambiguity effects in word identification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(1):63.

Ricardo JGB Campello, Davoud Moulavi, and Jörg Sander. 2013. Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 160–172. Springer.

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. Association for Computational Linguistics.

Jiaju Du, Fanchao Qi, and Maosong Sun. 2019. Using BERT for word sense disambiguation. *arXiv preprint arXiv:1909.08358*.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231.

Allyson Ettinger and Tal Linzen. 2016. Evaluating vector space models using human semantic priming results. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 72–77, Berlin, Germany. Association for Computational Linguistics.

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.

Aina Garí Soler and Marianna Apidianaki. 2021. Let's Play Mono-Poly: BERT Can Reveal Words' Polysemy Level and Partitionability into Senses. *Transactions of the Association for Computational Linguistics*, 9:825–844.

Juan Haro and Pilar Ferré. 2018. Semantic ambiguity: Do multiple meanings inhibit or facilitate word recognition? *Journal of Psycholinguistic Research*, 47(3):679–698.

Micha Heilbron, Kristijan Armeni, Jan-Mathijs Schoffelen, Peter Hagoort, and Floris P de Lange. 2021. A hierarchy of linguistic predictions during natural language comprehension. *BioRxiv*, pages 2020–12.

Yasushi Hino, Yuu Kusunose, and Stephen J Lupker. 2010. The relatedness-of-meaning effect for ambiguous words in lexical-decision tasks: When does relatedness matter? *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, 64(3):180.

Yasushi Hino and Stephen J Lupker. 1996. Effects of polysemy in lexical decision and naming: An alternative to lexical access accounts. *Journal of Experimental Psychology: Human Perception and Performance*, 22(6):1331.

Maryam Honari-Jahromi, Brea Chouinard, Esti Blanco-Elorrieta, Liina Pylkkänen, and Alona Fyshe. 2021. Neural representation of words within phrases: Temporal evolution of color-adjectives and object-nouns during simple composition. *PloS one*, 16(3):e0242754.

Shailee Jain and Alexander Huth. 2018. Incorporating context into language encoding models for fMRI. *Advances in Neural Information Processing Systems*, 31.

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. Association for Computational Linguistics.

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

Sreejan Kumar, Theodore R Sumers, Takateru Yamakoshi, Ariel Goldstein, Uri Hasson, Kenneth A Norman, Thomas L Griffiths, Robert D Hawkins, and Samuel A Nastase. 2022. Reconstructing the cascade of language processing in the brain using the internal computations of a transformer-based language model. *bioRxiv*.

Shapol M Mohammed, Karwan Jacksi, and Subhi RM Zeebaree. 2020. Glove word embedding and dbscan algorithms for semantic document clustering. In *2020 International Conference on Advanced Science and Engineering (ICOASE)*, pages 1–6. IEEE.

Sathvik Nair, Mahesh Srinivasan, and Stephan Meylan. 2020. Contextualized word embeddings encode aspects of human-like word sense knowledge. In *Proceedings of the Workshop on the Cognitive Aspects of the Lexicon*, pages 129–141, Online. Association for Computational Linguistics.

Penny M Pexman, Yasushi Hino, and Stephen J Lupker. 2004. Semantic ambiguity and the process of generating meaning from print. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30(6):1252.

Emily Reif, Ann Yuan, Martin Wattenberg, Fernanda B Viegas, Andy Coenen, Adam Pearce, and Been Kim. 2019. Visualizing and measuring the geometry of BERT. *Advances in Neural Information Processing Systems*, 32.

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

Anders Søgaard. 2016. Evaluating word embeddings with fMRI and eye-tracking. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 116–121, Berlin, Germany. Association for Computational Linguistics.

Laure Thompson and David Mimno. 2020. Topic modeling with contextualized word representation clusters. *arXiv preprint arXiv:2010.12626*.

Leslie C Twilley, Peter Dixon, Dean Taylor, and Karen Clark. 1994. University of alberta norms of relative meaning frequency for 566 homographs. *Memory & Cognition*, 22(1):111–126.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Loïc Vial, Benjamin Lecouteux, and Didier Schwab. 2019. Sense vocabulary compression through the semantic knowledge of WordNet for neural word sense disambiguation. In *Proceedings of the 10th Global Wordnet Conference*, pages 108–117, Wroclaw, Poland. Global Wordnet Association.

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. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

A Polysemy and Homonymy Values

Word	**Ambiguity (Rodd)**	**Senses (WordNet)**	**Meanings (BERT)**	**Senses (BERT)**
admit	Amb.	8	1	2
advance	Amb.	20	8	15
affair	Amb.	3	1	14
alone	Unamb.	6	2	5
amuse	Unamb.	2	1	3
apple	Unamb.	2	3	3
arms	Amb.	10	8	11
article	Amb.	5	5	6
baby	Unamb.	8	2	3
badger	Amb.	4	3	3
bark	Amb.	9	3	3
batter	Amb.	5	4	4
Bible	Unamb.	2	1	3
blind	Amb.	10	2	14
bonnet	Amb.	3	2	3
bowl	Amb.	12	4	6
boxer	Amb.	4	1	4
brain	Unamb.	7	5	5
bridge	Amb.	12	1	8
broke	Amb.	60	1	17
brutal	Unamb.	4	1	2
bulb	Amb.	6	5	5
bus	Unamb.	7	7	8
cabinet	Amb.	4	1	3
calf	Amb.	4	4	4
can	Amb.	8	1	2
cane	Amb.	4	6	6
case	Amb.	22	1	3
cattle	Unamb.	1	3	3
chance	Amb.	9	5	5
charm	Amb.	8	4	4
chest	Amb.	4	3	3
China	Amb.	4	1	2
cider	Unamb.	1	3	3
cigar	Unamb.	1	3	3
citizen	Unamb.	1	5	5
clay	Unamb.	5	3	6
clog	Amb.	9	4	4
coal	Unamb.	5	3	3
company	Amb.	10	3	4
craft	Amb.	6	6	10
cricket	Amb.	3	1	8

Word	**Ambiguity (Rodd)**	**Senses (WordNet)**	**Meanings (BERT)**	**Senses (BERT)**
custard	Unamb.	1	3	3
deed	Amb.	2	3	5
degree	Amb.	7	7	9
dense	Amb.	4	1	4
destroy	Unamb.	4	1	2
diamond	Unamb.	6	6	10
digit	Amb.	3	6	6
dollar	Unamb.	4	1	3
dozen	Unamb.	2	1	2
dry	Amb.	19	1	7
express	Amb.	13	6	6
fee	Unamb.	3	2	3
feet	Amb.	11	5	5
fence	Amb.	7	1	2
firm	Amb.	14	3	3
fling	Amb.	7	1	2
forest	Unamb.	3	5	5
fraud	Unamb.	3	1	2
free	Amb.	22	1	7
frog	Unamb.	4	4	4
fun	Unamb.	4	3	4
glare	Amb.	6	3	3
glass	Amb.	12	4	4
glove	Unamb.	3	5	8
goat	Unamb.	4	4	4
grain	Amb.	15	2	3
grief	Unamb.	2	1	2
grow	Unamb.	10	1	2
hamper	Amb.	4	1	3
hill	Unamb.	6	2	7
horn	Amb.	12	6	8
hotel	Unamb.	1	4	5
interest	Amb.	10	2	2
item	Unamb.	6	1	2
jumper	Amb.	8	6	15
kid	Amb.	7	3	3
kind	Amb.	4	7	13
kingdom	Unamb.	6	1	2
lake	Unamb.	3	5	8
last	Amb.	21	1	12
late	Amb.	11	1	2
lean	Amb.	10	3	9
left	Amb.	24	1	3
letter	Amb.	8	1	13
lie	Amb.	10	6	14
like	Amb.	11	1	3
limp	Amb.	5	2	5
lobby	Amb.	4	4	5
lung	Unamb.	1	8	11
marble	Amb.	4	3	3
march	Amb.	14	6	10
maroon	Amb.	6	4	4
metal	Unamb.	4	4	4
might	Amb.	1	1	3
misery	Unamb.	2	4	5
nail	Amb.	10	6	11
net	Amb.	12	1	14
novel	Amb.	4	1	9
ocean	Unamb.	2	4	4
odd	Amb.	6	8	8
organ	Amb.	6	4	4
palm	Amb.	5	6	6
panel	Amb.	10	5	6
park	Amb.	8	5	9
patient	Amb.	3	3	4
peer	Amb.	3	11	19
picket	Amb.	8	5	5
pine	Amb.	3	4	4
pitcher	Amb.	5	6	13
poach	Amb.	2	16	24
poet	Unamb.	1	1	6
poker	Amb.	2	3	4
pole	Amb.	13	5	11
prayer	Unamb.	5	1	2
pride	Amb.	6	5	7
pupil	Amb.	3	3	4
rabbit	Unamb.	4	3	3
ram	Amb.	9	4	8
rare	Amb.	6	1	17
rate	Amb.	7	6	13
reflect	Amb.	7	2	3
refrain	Amb.	3	3	3
river	Unamb.	1	3	6
ruler	Amb.	2	1	2
sack	Amb.	13	6	7
safe	Amb.	7	1	7
sage	Amb.	5	6	16
sane	Unamb.	2	7	11

Word	**Ambiguity (Rodd)**	**Senses (WordNet)**	**Meanings (BERT)**	**Senses (BERT)**
scrap	Amb.	7	3	3
screen	Amb.	16	2	15
seal	Amb.	15	4	8
season	Amb.	6	5	5
second	Amb.	15	1	19
seek	Unamb.	6	1	8
sense	Amb.	9	4	5
sentence	Amb.	4	1	2
shed	Amb.	6	3	4
sign	Amb.	20	6	6
spade	Amb.	4	7	8
speaker	Amb.	3	5	5
spell	Amb.	10	5	5
stable	Amb.	7	2	3
staff	Amb.	8	2	2
stag	Amb.	5	4	4
stalk	Amb.	8	3	3
stamp	Amb.	18	2	4
staple	Amb.	7	3	3
static	Amb.	5	1	3
stern	Amb.	7	5	5
store	Amb.	6	1	2
strand	Amb.	9	5	5
straw	Amb.	7	7	7
swallow	Amb.	11	5	5
swear	Amb.	5	1	2
task	Unamb.	4	4	4
temple	Amb.	4	2	4
tend	Amb.	3	1	2
tense	Amb.	8	2	3
tent	Unamb.	3	1	3
term	Amb.	8	2	3
terror	Unamb.	4	5	6
thief	Unamb.	1	3	3
throat	Unamb.	4	1	2
throw	Unamb.	20	6	7
tiger	Unamb.	2	5	6
toast	Amb.	6	2	4
travel	Unamb.	9	3	3
trial	Amb.	6	4	4
trust	Amb.	12	6	8
uniform	Amb.	6	1	2
unite	Unamb.	6	6	8
urban	Unamb.	2	4	5
vent	Amb.	7	8	12
vote	Unamb.	10	1	16
warn	Unamb.	4	1	2
watch	Amb.	13	8	16
weapon	Unamb.	2	1	2
winter	Unamb.	2	1	4
yard	Amb.	9	7	13
lorry	Unamb.	2	3	3