Oftentimes in natural language, one word can have several different meanings, or senses. From a cognitive linguistics perspective, how humans understand the distinctions between word sense is particularly interesting. Specifically, the human mind must make a choice to determine which sense of a word applies in a certain context, effectively resolving lexical ambiguities. Two major phenomena describe the nature of word sense: polysemy, when the senses of a word are semantically related, and homonymy, when this is not the case. Polysemy is often highly regular, with many words participating in the same usage pattern (Srinivasan & Rabagliati, 2015) in multiple languages. For instance, the word used to refer to a material is also used to refer to an item made of that material (i.e. glass, linen).

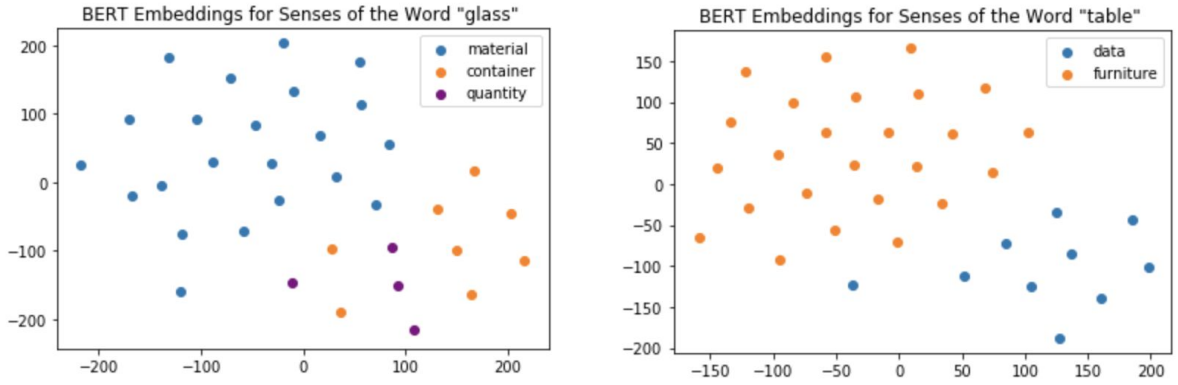
The most commonly used resource documenting word senses, WordNet(Miller, 1995), has notable limitations. The dataset is small and does not reflect senses known to many speakers. This is because oftentimes, an existing word can be used in slang, vernacular, or specialized technical settings. WordNet also often has *too many* senses, many of which are archaic, which leads the dataset to be too fine-grained for many natural language processing (NLP) tasks (Amamrami and Goldberg, 2018). Finally, WordNet does not consider polysemy and homonymy. With the growing deployment of NLP tools, the need for a data-driven, cognitively informed model of word sense becomes pressing for both NLP and cognitive science.

People can use linguistic context to effectively to deduce the sense of a word; computational models have been developed to imitate this ability. Many such models represent words as numerical vectors, or embeddings. Earlier methods, namely Word2Vec (Mikolov et al, 2013) and GloVe (Pennington, Socher, and Manning, 2015), are able to capture more semantic information than count-based models (Lilleberg, Zhu, and Zhang, 2015). 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. For example, the vector for the word "bank" could contain information about both the side of a river and a financial institution.

One promising direction for research is using more sophisticated representations of linguistic context. Recent advances in neural network architecture and pretraining have led to models that output individualized representations of words based on the sentences they are used in (Peters et al, 2018, Devlin et al, 2018). With such methods, the vectorized representation of the word "table" would be different in the phrases "in the table," referring to a set of data and "on the table," referring to a piece of furniture. Embeddings derived from BERT(Bidirectional Encoding Representations from Transformers, Devlin et al, 2018) perform better than other models on word sense disambiguation (Wiedemann et al, 2019). These models, namely BERT, are able to represent polysemy, but how they do so is relatively unexplored. As a result, we seek to explore the viability of BERT in creating a new, data-driven ontology of word sense in our work.

As an exploratory analysis of how BERT represents words with multiple senses, we show a proof-of-concept for a polysemous word, "glass," and a homonymous word, "table." There is a clearer distinction between the centroids of the senses for the homonymous word (cosine similarity of 0.54) than for the polysemous word (described below). For the polysemous word, two of the senses ("quantity"- as in "poured a glass of water" and "container"- as in "poured water into a glass"), have centroids that are closer together in the embedding space (cosine similarity of 0.89) than the sense referring to glass as a material (cosine similarity to "container" of 0.64, "quantity" of 0.59). This proximity between the centroids of the

sense vectors could reveal perceived degrees of similarity between senses of a word. The contextualized embeddings are based on the SEMCOR corpus (Langone et al, 2004), which contains 200,000 sense-annotated tokens, and we apply the t-SNE dimensionality reduction technique to the BERT embeddings to create the plots below.



The presence of patterns delineating word sense by inspecting the lower-dimensional space leads us to consider applying hierarchical clustering techniques to estimate the number of senses for a certain word, working toward creating a BERT-based ontology of word sense. One such technique, agglomerative clustering, begins with treating each vector as an individual cluster, and then fusing clusters based on their similarity, ultimately leading to a tree-like structure. At a certain level of a word's tree, we can analyze whether the number of branches corresponds to the number of senses, and if these clusters are split based on word sense. This would also be a useful extension of Amrami and Goldberg (2019), as they present a clustering algorithm for word sense induction, but do not relate it to WordNet as it is too fine-grained of a data source.

One possibility is that the geometry of the clusters corresponds to the distinction between polysemy and homonymy: homonymous clusters are more distinct that polysemous ones. We can test this by looking at distances between clusters for these two word types and also take basic supervised learning approaches to demonstrate how separable the sense clusters are for a particular word. For polysemous words, certain patterns could occur in the vector space that are analogous to common patterns in natural language as documented by Srinivasan and Rabagliati (2015). If this is the case, then it could be possible that certain components of the space could encode these relations, as Bolukbasi et al (2016) proved for gender using Word2Vec. We can determine the strength of various polysemous relationships exhibited in the usage of their constituent words through linear transformations in the vector space. Using the embeddings for word senses as well as the difference between their centroids, we hypothesize we will capture a noisy measurement of documented patterns in polysemy. In fact, higher and lower-level relationships between word sense might initiate at certain layers of the neural network, which can be verified by analyzing activations at these layers.

Advances in NLP offer us the opportunity to move towards a new, cognitively and empirically informed ontology of word senses that addresses some of the limitations with the current standard resource, WordNet. It also gives us an opportunity to explore both the viability and the challenges of using computational models to understand the properties of words and the lexicon in natural languages.

Works Cited:

Amrami, A., & Goldberg, Y. (2019). Towards better substitution-based word sense induction. *arXiv preprint arXiv:1905.12598*.

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

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems* (pp. 4349-4357).

Coenen, A., Reif, E., Yuan, A., Kim, B., Pearce, A., Viégas, F., & Wattenberg, M. (2019). Visualizing and Measuring the Geometry of BERT. *arXiv preprint arXiv:1906.02715*.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Helen Langone, Benjamin R. Haskell, and George A. Miller. 2004. Annotating wordnet. In Workshop On Frontiers In Corpus Annotation, pages 63–69. ACL, Boston.

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

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111-3119).

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, *38*(11), 39-41.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Srinivasan, M. & Rabagliati, H (2015). How concepts and conventions structure the lexicon: Cross-linguistic evidence from polysemy. *Lingua*, *157*, 124-152.

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

Wiedemann, G., Remus, S., Chawla, A., & Biemann, C. (2019). Does BERT Make Any Sense? Interpretable Word Sense Disambiguation with Contextualized Embeddings. *arXiv preprint arXiv:1909.10430*.

Dayu Yuan, Julian Richardson, Ryan Doherty, Colin Evans, and Eric Altendorf. 2016. Semi-supervised word sense disambiguation with neural models. In COLING 2016.