What counts as an exemplar model, anyway? A commentary on Ambridge (2020)

Kyle Mahowald, George Kachergis, Michael C. Frank
Stanford University

Abstract
Ambridge (2019) calls for exemplar-based accounts of language acquisition. Do modern neural networks such as transformers or word2vec – which have been extremely successful in modern natural language processing (NLP) applications – count? Although these models often have ample parametric complexity to store exemplars from their training data, they also go far beyond simple storage by processing and compressing the input via their architectural constraints. The resulting representations have been shown to encode emergent abstractions. If these models are exemplar-based then Ambridge's theory only weakly constrains future work. On the other hand, if these systems are not exemplar models, why is it that true exemplar models are not contenders in modern NLP?[1]

———

Ambridge (2019) calls for language acquisition researchers to take seriously the possibility that speakers do not possess any linguistic abstractions, and rely only on stored exemplars and fast analogical reasoning for comprehension and production. Part of Ambridge's argument against abstraction and in favor of exemplar models is the success of exemplar-based computational models in capturing empirical phenomena. Although the article cites exemplar-based models covering a range of phenomena, we were surprised to find a gap in the survey. In the last decade, there has been enormous progress in natural language processing (NLP) on a wide variety of tasks from speech recognition and language modeling all the way to question-answering and inference. The models that have enabled this progress are all variants of multi-layer neural networks, including neural word embedding models (Mikolov, et al. 2013) and transformer-based models like BERT (Devlin, et al. 2018) and GPT-2 (Radford, et al. 2019). These models are trained on very large data sets and have a very large number of parameters, allowing them to store and recode substantial summaries of their training data.

Under Ambridge's definition, are modern deep neural models of language examples of exemplar models? The answer to that question is not straightforward. One issue is that the definition of exemplar models on offer does not allow readers make that judgment. Additionally, there are open questions as to how fully modern language models encode the input they receive and the extent to which they retain it in their fitted parameters.

The current state of the art on many NLP tasks is achieved through transformer-based models, like BERT. In brief, a transformer model encodes a sequence (e.g., words in a sentence of English) into a high-dimensional vector as it is passed through a series of layers, before passing that encoded input back through a series of decoding layers that have been trained for a given task (e.g., translation or part-of-speech tagging). Transformers, sharing features such as many layers and self-attention with other modern language models, have proven capable of learning both nearby and long-distance dependencies, and yet these seemingly abstract rules are only incrementally learned as the connection weights gradually change during training on a large number of exemplars. Is a transformer an exemplar model? Because of the huge number of

[1] We thank Alexandra Carstensen and Urvashi Khandelwal for helpful comments.

parameters in these models and the complex ways in which they interact as information is passed through the layers, it is difficult to peer into the model and understand why it does what it does, what representations it learns, and how much information it stores. But using such models as a case study can illuminate aspects of Ambridge's argument.

On the one hand, if a modern language model *is* an exemplar model under Ambridge's definition, then what it means to be an exemplar model may be vacuous since these models appear to do some of the things that Ambridge considers outside the scope of exemplar models, such as representing abstract structures. Whereas NLP systems often made use of explicit, symbolic representations like probabilistic context-free grammars, neural NLP models are not explicitly designed to compute over abstract linguistic structures. So the way they store and manipulate abstract structures is more opaque. But that opacity does not mean they are free of abstraction.

In fact, there has been ample work showing that neural models capture aspects of abstract linguistic structure. For instance, neural networks capture syntactic generalizations necessary for long-distance agreement (Gulordava, et al., 2018). In some cases, the mechanism by which neural networks make such generalizations about long-distance agreement is well-understood. A study of one particular architecture, a long short-term memory (LSTM) neural network, showed that there are nodes in the network that respond selectively to abstract syntactic categories, such as subjecthood and number (Lakretz et al., 2019). There are also a number of linguistic structures that can be recovered from neural networks. A *probing* technique has shown that deep neural models of language encode syntactic tree distance (Hewitt et al., 2019). Investigations of BERT show that some parameters within the model appear to selectively identify syntactically relevant abstract categories like determiners for nouns and direct objects for verbs (Clark et al., 2019). Artificial neural networks can even be fruitfully analyzed using a technique that is often used as a paradigmatic case of abstract syntactic representation: syntactic priming (Prasad et al., 2019). If, under Ambridge's definition, models such as these are exemplar models, then we would argue that exemplar models do not provide evidence "against stored abstraction."

On the other hand, if modern NLP models are *not* exemplar models under Ambridge's definition – because they either do not store all the input or because they learn and store abstract structures – that would seem to create a distinction between what Ambridge, citing Chandler (2002), calls "de facto exemplar models" and full-scale exemplar models. Here, "de facto exemplar models" seem to be models with enough parameters to encode the full input[2]. But, in this case, Ambridge does not state why one should prefer a pure exemplar model to this class of *de facto* exemplar models. Moreover, this distinction undermines the argument that a major strength of exemplar models is their computational success. If BERT and cousins are outside

[2] Due to various mechanisms commonly used in neural networks such as regularization and incremental weight updating, typically not all training instances T are typically retrievable even if the number of parameters $P > T$ (see Arpit et al., 2017).

the space of what Ambridge would consider pure exemplar models, then such models do not approach state-of-the-art performance in NLP.

More broadly, there need not be a hard split between models that encode abstract structures and those that store a huge amount of information about the input and allow for fast analogical comparisons. Neural language models in all their variations provide a gradient across these dimensions. They are not explicitly trained to operate over syntactic trees or morphological hierarchies, but their representations may still encode and store those sorts of abstractions to varying degrees depending on their architecture and training.

Situating Ambridge's account of exemplar models within this space of modern models would help clarify what counts as an exemplar model anyway, just how radical Ambridge's proposal is, and how these ideas can guide future efforts to constrain the space of models for language acquisition and processing.

References

Ambridge, B. (2019). Against stored abstractions: A radical exemplar model of language acquisition. *First Language*. doi:10.1177/0142723719869731

Arpit, D., Jastrzebski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M. S., Maharaj, T., Fischer, A., Courville, A., Bengio Y., & Lacoste-Julien, S. (2017). *Proceedings of the 34th International Conference on Machine Learning*, Sydney, Australia, PMLR 70. https://arxiv.org/pdf/1706.05394.pdf

Chandler, S. (2002). Skousen's analogical approach as an exemplar-based model of categorization. In *Analogical modeling: An exemplar-based approach to language*. Eds. R. Skousen, D. Lonsdale, D.B, Parkinson. (Vol. 10). John Benjamins Publishing.

Clark, K., Khandelwal, U., Levy, O., & Manning, C.D. (2019). What Does BERT Look At? An Analysis of BERT's Attention. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. (pp. 276--286).

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.

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., & Baroni, M. (2018). Colorless green recurrent networks dream hierarchically. arXiv preprint arXiv:1803.11138.

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

Lakretz, Y., Kruszewski, G., Desbordes, T., Hupkes, D., Dehaene, S., & Baroni, M. (2019). The emergence of number and syntax units in LSTM language models. arXiv preprint arXiv:1903.07435.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Prasad, G., van Schijndel, M., & Linzen, T. (2019). Using Priming to Uncover the Organization of Syntactic Representations in Neural Language Models. arXiv preprint arXiv:1909.10579.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8).