Testing for Grammatical Category Abstraction in Neural Language Models

Najoung Kim
Department of Cognitive Science,
Johns Hopkins University
n.kim@jhu.edu

Paul Smolensky
Department of Cognitive Science,
Johns Hopkins University
Microsoft Research, Redmond
psmo@microsoft.com

1 Introduction

The notion of grammatical categories is fundamental to human language. Humans abstract over individual lexical items to form grammatical categories, such as nouns and verbs in English, and this category membership (rather than lexical identity) governs the applicability of linguistic rules (e.g., nouns can be heads of subjects of a verb). Category membership of new words are rapidly inferred from their linguistic environment: if a speaker of English hears *I saw a blick*, it is immediately clear that *blick* is a noun. This knowledge about the novel word's grammatical category enables speakers to furthermore produce sentences such as *We like the blick* and *The blick jumped*, even though these new contexts have no lexical overlap with the context that *blick* was first observed in. Hence, the identification of a grammatical category allows application of rules that operate over that category, allowing for generalization outside of the context that the novel word has been observed in (Gómez and Gerken, 2000).

Can we find evidence of abstract grammatical categories and category-based generalization resembling humans in pretrained neural language models? From the perspective of Cognitive Science, category abstraction in pretrained neural models can provide an argument against the need for an innate bias towards categorization (and prespecification of the set of lexical categories) for learners of language. From the perspective of Natural Language Processing, it is known that contemporary neural models perform well (near 98% accuracy) on benchmarks for part-of-speech (POS) tagging (Bohnet et al., 2018; He and Choi, 2019), and that diagnostic classifiers for probing pretrained models also achieve similarly high performance on POS (Tenney et al., 2019). However, it still remains an open question whether pretrained models can perform category-based generalization using novel words learned from limited contexts, and without being explicitly trained to perform categorization. This is also in line with the problem of out-of-distribution generalization in neural models of language and efforts to develop benchmarks for linguistic generalization that humans are capable of (Kim and Linzen, 2020; Linzen, 2020, *i.a.*). To this end, we propose a new method inspired by human developmental studies to probe pretrained neural language models, and present experimental results on BERT-large (Devlin et al., 2019). Our method does not require training a separate classifier on top, which lets us bypass the methodological questions raised in the recent literature on the validity of using diagnostic classifiers as probes (Hewitt and Liang, 2019; Voita and Titov, 2020, *i.a.*).

2 Method

Our method is inspired by the experimental design of Höhle et al. (2004), in which infants were familiarized to contexts containing novel words, and were tested with new sentences that either obeyed or violated category-based co-occurrence restrictions using a head-turn preference procedure (Jusczyk and Aslin, 1995; Kemler Nelson et al., 1995). We reformulate this study into a probing task using cloze probabilities: if a masked language model (MLM) makes a valid category inference about a newly learned word, it should be able to assign a higher probability to that word in a novel context that obeys the co-occurrence restriction for that category, over a word of a different category. For example, if the model sees unseen words *blick* and *dax* in contexts that signal distinct category membership (1), they should be able to make a generalization that in (2-a), the masked word is more likely to be *blick* than *dax*. On the other hand, in (2-b), *dax* should be more likely. That is, we expect

Proceedings of the Society for Computation in Linguistics (SCiL) 2021, pages 467-470.
Held on-line February 14-19, 2021

$P(blick \mid I\ went\ to\ a\ \_) > P(dax \mid I\ went\ to\ a\ \_)$ in (2-a) and $P(dax \mid I\ \_\ with\ some\ friends) > P(blick \mid I\ \_\ with\ some\ friends)$ in (2-b).

(1) a. *The* ***blick*** (Category inference: N)
b. *They* ***dax*** (Category inference: V)

(2) a. *I went to a __.*
(N-expecting: should prefer *blick*)
b. *I __ with some friends.*
(V-expecting: should prefer *dax*)

Our method involves two steps. First, we finetune the MLM (with the same MLM objective) on two *signal contexts* like (1) that unambiguously signal the category of the novel word. Second, we test the finetuned model by comparing the probabilities of the two newly learned words on multiple *test contexts* like (2). We consider the model's category inference to be accurate if it assigns higher probability to the new word in the correct test context (e.g., higher probability for *blick* over *dax* in (2-a)).

3 Experiment

3.1 Data

We constructed the signal and test contexts from sentences in MNLI (Williams et al., 2018). This was to ensure that the contexts had different sources from the model's pretraining data (Wikipedia and BooksCorpus). Two signal contexts with one unseen word each ($w1$ and $w2$)—each context unambiguously signaling the unseen word's grammatical category—constituted a finetuning set. The two contexts matched in the number of words and the linear position of the unseen word from both left and right. For example:

(3) a. *A* w_1 *needs two people.* (N-signaling)
b. *She* w_2 *at the group.* (V-signaling)

For testing, we sampled sentences from MNLI that contained a word in the same grammatical category as *w1* and *w2*, respectively, and masked out that word as in (4). We only selected sentences that contained a different number of subword tokens from the signal contexts.

(4) a. *Keep everyone else company by sitting in the* [MASK]. (N-expecting)
b. *The colonel* [MASK] *us to a hotel.*
(V-expecting)

We applied the above generation method to generate 6 English datasets that test for the binary distinguishability between four open-class grammatical categories: Noun, Verb, Adjective and Adverb. Since we used the automatic parses provided in MNLI to determine the grammatical categories, we manually verified the sentences after generation to rule out parser errors and contexts that were ambiguous between the two categories being compared (e.g., *John is __* can be both verb- and adjective-expecting contexts). Each dataset included 2 signal contexts (for finetuning) and 400 test contexts (200 for each category; 50% used as a development set).

3.2 Model

We applied our method to BERT-large with whole word masking.[1] We used the unused tokens ('[unused*n*]') in the vocabulary of BERT to represent the novel words. We froze the entire model, except for the embeddings of the two unused tokens being used. Hence, we are asking the question: can novel words be placed in a space that enables category-based generalization? We finetuned the model for 70 epochs, and selected the test checkpoint based on development set performance.

3.3 Results

Table 1 shows the generalization accuracy for each category pair. All accuracy was significantly above chance ($p < .05$, one proportion z-test), suggesting that BERT can abstract over grammatical categories and generalize to novel contexts to an extent. However, this generalization capacity has limitations in the following two aspects. First, some distinctions were much weaker than others (e.g., N vs. Adv.). Second, BERT failed to display *rapid* category inference as competent English speakers often can from even a single exposure. The reported accuracy was only reached after many finetuning iterations—on average 51 epochs with an initial learning rate as high as 5.[2]

Effect of embedding initialization. For the experiment reported in Table 1, we had randomly selected two tokens from the 1000 '[unused*n*]' tokens available in the vocabulary of BERT to represent the novel words being learned. The embeddings of these unused tokens are randomly initialized

[1] We used model checkpoints and code provided by Wolf et al. (2020): https://github.com/huggingface/transformers.

[2] We used the AdamW optimizer (Loshchilov and Hutter, 2019) with a constant schedule.

Cat 1	Cat 2	Acc.	Acc.(1 > 2)	Acc.(2 > 1)
N	V	88.1	87.2	89.0
N	Adj.	83.1	86.2	80.0
N	Adv.	67.3	63.0	71.6
V	Adj.	87.3	85.4	88.4
V	Adv.	78.7	80.2	77.2
Adj.	Adv.	71.2	60.6	81.8

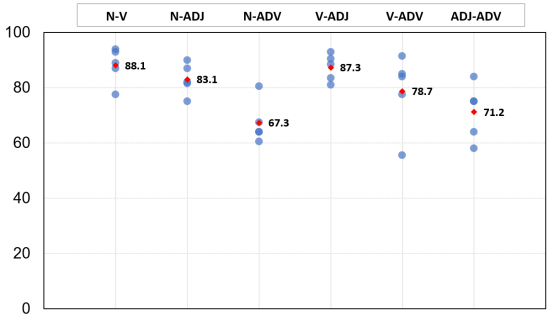

Table 1: Accuracy (%) of distinguishing two grammatical categories in novel contexts, averaged over five random seeds (individual accuracies are shown in the bottom figure). 'Acc.(1 > 2)' denotes the accuracy on the set of sentences where Category 1 should be preferred over Category 2 (e.g., assigning higher probability to a noun in a noun-expecting context for row 1), and vice versa. Column 'Acc.' lists the aggregate accuracy.

and remain unchanged during pretraining. To investigate the effect of the particular choice of the tokens representing the novel words (i.e., the initialization of the embeddings), we reran the whole experiment four additional times, selecting different '[unused*n*]' tokens for the novel words each time. The variation in the mean accuracy depending on the random selection of unused tokens is shown in Figure 1.

4 Remaining Questions and Future Work

We proposed a method for testing category-based generalization in neural language models and tested BERT-large with this method. Our results show that it achieves partial success on making such generalizations. However, the degree of generalization is still limited. In addition to the weak distinguishability of some open-class categories, BERT did not display rapid category inference; it was only able to achieve the reported performance after many repeated exposures to the finetuning examples. As an immediate next step, we will conduct a detailed error analysis to investigate whether certain subsets

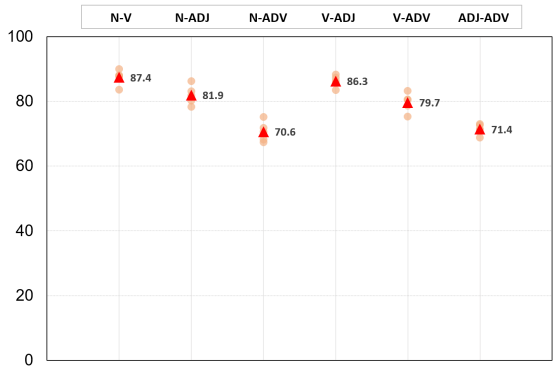

Figure 1: Variation in the mean accuracy across five experiment reruns with different '[unused*n*]' tokens to initialize the two novel words being learned. Each dot represents the *mean* accuracy of an experiment (i.e., a single dot corresponds to a single number under the 'Acc.' column in Table 1, which itself is an average over five random seeds).

of test cases pose greater challenges to the model (e.g., is generalizing across subcategories of verbs harder than within?).

While our method does test the generalization capacity of BERT regarding the usage of novel words outside of the contexts that they were observed in, further analysis is needed in terms of *how* the generalization is achieved to elucidate whether the partial success is driven by abstraction. In its current form, good performance on our task is a necessary condition for abstractive generalization but not sufficient. For instance, we could imagine a scenario where a model achieves success on generalization without abstraction by analogy to a single exemplar that is not part of a subspace representative of the relevant grammatical category. One way we could tease apart a true case of category-based abstraction would be by examining whether there exists a subspace (rather than a single point in space) of embeddings that gives rise to similar degrees of success on our generalization task. We plan to conduct such analyses in future work. More broadly, we will explore whether learning biases are needed for better category abstraction and generalization.

Acknowledgments

We thank the JHU Neurosymbolic Computation Lab, the JHU Computation and Psycholinguistics Lab, and the reviewers for their helpful feedback. This research project was conducted using computational resources at the Maryland Advanced Research Computing Center (MARCC).

References

Bernd Bohnet, Ryan McDonald, Gonçalo Simões, Daniel Andor, Emily Pitler, and Joshua Maynez. 2018. Morphosyntactic tagging with a meta-BiLSTM model over context sensitive token encodings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2642–2652, Melbourne, Australia. Association for Computational Linguistics.

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.

Rebecca Gómez and LouAnn Gerken. 2000. Infant artificial language learning and language acquisition. *Trends in Cognitive Sciences*, 4(5):178–186.

Han He and Jinho D. Choi. 2019. Establishing strong baselines for the new decade: Sequence tagging, syntactic and semantic parsing with BERT. *arXiv:1908.04943*.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. 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 2733–2743, Hong Kong, China. Association for Computational Linguistics.

Barbara Höhle, Jürgen Weissenborn, Dorothea Kiefer, Antje Schulz, and Michaela Schmitz. 2004. Functional elements in infants' speech processing: The role of determiners in the syntactic categorization of lexical elements. *Infancy*, 5(3):341–353.

Peter W. Jusczyk and Richard N. Aslin. 1995. Infants' detection of the sound patterns of words in fluent speech. *Cognitive Psychology*, 29(1):1–23.

Deborah G. Kemler Nelson, Peter W. Jusczyk, Denise R. Mandel, James Myers, Alice Turk, and LouAnn Gerken. 1995. The head-turn preference procedure for testing auditory perception. *Infant Behavior and Development*, 18(1):111–116.

Najoung Kim and Tal Linzen. 2020. COGS: A compositional generalization challenge based on semantic interpretation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9087–9105, Online. Association for Computational Linguistics.

Tal Linzen. 2020. How can we accelerate progress towards human-like linguistic generalization? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5210–5217, Online. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Sam Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? Probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

Elena Voita and Ivan Titov. 2020. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 183–196, Online. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

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.