

Improving the Generalizability of Contextualized Word Vectors by Exploiting Linguistic Structure

Motivation The broad goal of automatic language understanding breaks down into a constellation of subproblems, ranging from low-level tasks like part-of-speech tagging and language modeling (given a sequence of words, predict the word that comes next) to higher-level tasks like sentiment analysis and reading comprehension. In recent years, neural network models have achieved state-of-the-art results on virtually all of these tasks. At their base, these neural models use *word vectors*, which encode a word’s meaning as a vector in continuous space. Traditionally, these word vectors are static—a single vector is assigned to each word, which means they cannot handle words with multiple meanings. Recent work¹ has explored *contextualized* word vectors, which assign each word a vector that is a function of the entire input sequence; this enables them to model the use of words in context. Contextualized word vectors are extraordinarily effective—simply using them in place of traditional static word vectors within the latest models leads to instant improvements of up to 25% across the constellation of NLP tasks, a huge advance for the field.

While models augmented with contextualized word vectors enjoy improved performance, they remain brittle and fundamentally fail to generalize to new data domains. However, as different as these new domains may be, the data is still *human language*. Language, in all its varying uses and forms, has regular patterns and structure. **This linguistic structure underlies all of these differing data distributions, but the majority of neural NLP models do not explicitly consider this unifying feature.** On the contrary, modern neural models are often unconstrained by design, since it gives them the freedom to tune their millions of parameters toward optimizing an end objective. This freedom is also a curse, since these drastically overparameterized models overfit to their particular training dataset, biases and all². Reining in these currently unconstrained models by integrating appropriate linguistic biases will improve their generalization.

Contextualized word vectors are produced from neural network language models, and they thus exhibit the same lack of generalizability affecting other neural NLP models. This proposal seeks to improve the generalizability of neural language models and their derived contextualized word vectors in two parts: (1) analysis of what large-scale language models do and don’t learn about language, and (2) augmentation of language models with explicit notions of linguistic structure to improve their generalizability.

What Do Language Models Learn About Language? The success of language model-derived contextualized word vectors on a variety of downstream NLP tasks indicates that the vectors encode general features of language. However, there is little work in more thoroughly characterizing what specific features of language these vectors capture, and more critically, the aspects of language they miss. Such an analysis is a necessary first step towards enhancing their generalization by imbuing language models with linguistic information, since improving upon methods requires identifying and better understanding their weaknesses. The concrete subproblems here are:

1. Extending prior work that probes the representations learned by neural models³, I will build a software framework for easily analyzing the linguistic information captured by word vectors.

¹Matthew E. Peters et al. “Deep contextualized word representations.” In: *Proc. of NAACL*. 2018.

²Suchin Gururangan et al. “Annotation Artifacts in Natural Language Inference Data.” In: *Proc. of NAACL*. 2018.

³Xing Shi, Inkit Padhi, and Kevin Knight. “Does String-Based Neural MT Learn Source Syntax?” In: *Proc. of EMNLP*. 2016; Yonatan Belinkov et al. “What do Neural Machine Translation Models Learn about Morphology?” In: *Proc. of ACL*. 2017.

The package will make it simple to train configurable classifiers to predict linguistic labels from individual or pairs of word vectors. If a classifier can be trained to predict linguistic information about a word (e.g., its part-of-speech tag) or a pair of words (e.g., their semantic relation) from the word vector(s) alone, we can reasonably conclude that the vector(s) encode this information.

2. Using the toolkit produced in (1), I will analyze what linguistic information is encoded within state-of-the-art language models used for producing contextual word vectors. In particular, I am interested in what linguistic information is difficult for a classifier to predict from word vectors—this indicates that the information is unreliably (or not) learned by the language model and hints at avenues for augmentation.

The results of (2) are likely to motivate further investigation into *why* contextualized word vectors work well across a range of tasks.

Augmenting Language Model Structure with Linguistic Structure With knowledge of what linguistic information is not captured by language models, I can work toward improving them by explicitly incorporating linguistic biases that shore up their weaknesses. I intend to:

1. Incorporate linguistic inductive biases into language model structure, focusing on the weaknesses discovered in the first part. The details of this step are highly dependent on the results of the first part of the project, but I anticipate building upon initial work that augments model architectures with explicit notions of linguistic syntactic structure (recurrent neural network grammars⁴) and compositionality (neural module networks⁵).
2. Evaluate whether contextualized word vectors generated from the structure-augmented language model produced in (1) help downstream models better generalize to data domains unseen during training. I am especially interested in applications to model fairness, and plan to evaluate generalization to dialects used by minorities, such as African-American English.

Intellectual Merit Despite the prevalence of neural architectures in NLP, they are still poorly understood; this project will deeply examine neural language models and produce transferable insights that will help the community better understand neural NLP architectures as a whole. Furthermore, generalization is a fundamental problem of modern NLP systems—this method will produce general representations of text that work well across domains, thus improving the generalization and performance of a variety of downstream NLP models. In addition, this work will produce (1) an open-source toolkit for analyzing the linguistic information encoded within word vectors and (2) pre-trained models for easily adding our structure-enhanced contextualized word vectors to downstream models, from which the scientific community will benefit.

Broader Impacts NLP systems have great potential to positively impact society, especially in underdeveloped regions of the world—imagine if these communities had access to medical advice through a conversational interface in their native language! However, the lack of generalizability in modern NLP systems impedes their ethical deployment. For example, systems trained on standard newswire corpora fail to generalize to dialects used by minorities (e.g., African-American English), creating the potential for inequality. By improving the generalizability of a broad range of NLP systems, the proposed work moves us closer to realizing the positive real-world potential of NLP systems through ethical deployment of models that work equally well for all members of society.

⁴Chris Dyer et al. “Recurrent Neural Network Grammars.” In: *Proc. of NAACL*. 2016.

⁵Jacob Andreas et al. “Learning to Compose Neural Networks for Question Answering.” In: *Proc. of NAACL*. 2016.