Unified Unsupervised Grammar Induction for Typologically Diverse Languages

Byung-Doh Oh (Department of Linguistics, The Ohio State University)

The unsupervised learning of syntactic grammars is a difficult problem that has been a central focus of natural language processing research. Part of this impetus is due to the fact that rapid and accurate syntactic analyses of free text can greatly improve the performance of downstream tasks such as machine translation and information extraction. Additionally, the unsupervised nature of such learning models makes them attractive for use in low-resource settings, in which human-annotated linguistic resources may not be available for the target language.

Recent work on unsupervised grammar induction has employed deep neural networks to learn probabilistic grammars [2, 3] and has shown promising results in discovering syntactic structures solely from word sequences. However, these systems have been typically geared towards achieving accurate performance on high-resource languages such as English and Chinese. A recent evaluation of these systems on multilingual grammar induction demonstrated their limitations on typologically different languages that have richer morphological systems [1]. In addition, a comparison of neural grammar induction models that differ minimally in how word emission probabilities are learned (i.e. words are treated as symbols or as a sequence of characters) showed that incorporating character-level information favors morphologically rich languages, resulting in grammars that match up to 55.9% of labeled brackets on manually annotated Japanese data [1]. This further highlights the importance of considering the typological properties of the target language for deploying grammar induction systems.

Although the grammar induction models proposed in [1] broadened the coverage for typologically diverse languages, an explicit decision regarding which model to employ nonetheless has to be made in an under-resourced setting in which not much is known about the typology of the target language. Additionally, the amount and distribution of text that is available may result in more accurate syntactic analyses from one model over another. This work first presents a grammar induction model that addresses these issues by interpolating word-based emission probabilities and character-based emission probabilities that are at the core of [1] in a data-driven manner. Subsequently, results will be presented from experiments that test the robustness of this system across typologically different languages and scenarios in which training data is especially limited.

Additionally, potential future directions and challenges for unsupervised grammar induction systems will be discussed. More specifically, contextualized word representations, which are vectors from neural network models that typically incorporate information from both directions of the sentence and are central to many natural language processing applications, can provide semantic information that is useful for disambiguating polysemous words and learning more accurate and fine-grained grammars. However, the presence of potentially infinite vector representations for the same word presents challenges for probability estimation. Research into grammar induction techniques that can overcome this problem will yield fruitful practical results and also further advance our understanding of learning syntactic structures as well as our ability to communicate in languages with limited resources.

References

[1] Jin, L., Oh, B.-D., and Schuler, W. (2021). Character-based PCFG induction for modeling the syntactic acquisition of morphologically rich languages. In *Findings of the 2021 Conference on Empirical Methods in Natural Language Processing.*, pages 4367–4378.

[2] Kim, Y., Dyer, C., and Rush, A. (2019). Compound probabilistic context-free grammars for grammar induction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2369–2385.

[3] Zhu, H., Bisk, Y., and Neubig, G. (2020). The return of lexical dependencies: Neural lexicalized PCFGs. *Transactions of the Association for Computational Linguistics*, 8:647–661.