Unified Unsupervised Grammar Induction for Typologically Diverse Languages

Byung-Doh Oh

Department of Linguistics, The Ohio State University

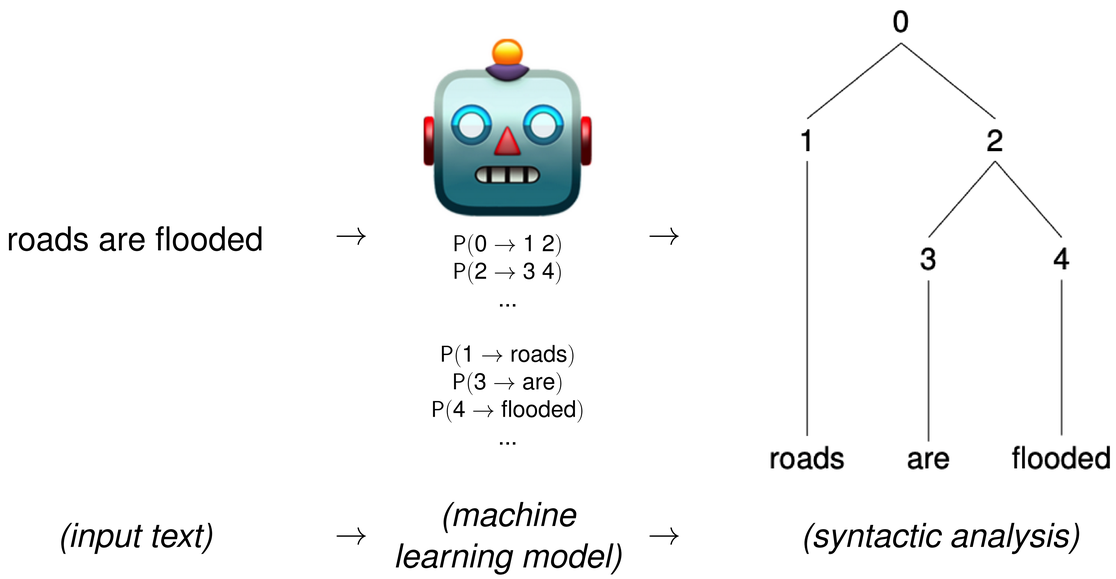
oh.531@osu.edu

byungdoh.github.io

THE OHIO STATE UNIVERSITY
COLLEGE OF ARTS AND SCIENCES
DEPARTMENT OF LINGUISTICS

Unsupervised Grammar Induction

- Rapid and accurate syntactic analysis of sentences is crucial for applications like machine translation and information extraction
- Unsupervised learning is attractive for use in scenarios where low-resource languages are spoken
- Still remains as a largely unsolved problem in natural language processing



Current Approaches

- Probability estimation using deep neural networks [2, 3, 5] by maximizing marginal probability of input sentence σ, which consists of rule probabilities from phrase to phrases ($c_\eta \rightarrow c_{\eta 1}\ c_{\eta 2}$) and from phrase to word ($c_\eta \rightarrow w_\eta$)

$$\mathsf{P}(\sigma) = \sum_{\tau \text{ for } \sigma} \prod_{\eta \in \tau \text{ s.t. } c_\eta \rightarrow c_{\eta 1}\ c_{\eta 2}} \mathsf{P}(c_\eta \rightarrow c_{\eta 1}\ c_{\eta 2}) \cdot \prod_{\eta \in \tau \text{ s.t. } c_\eta \rightarrow w_\eta} \mathsf{P}(c_\eta \rightarrow w_\eta)$$

- There is currently a high research focus on high-resource languages like English and Chinese [3, 5]
- English and Chinese have very simple word-internal structures, and represent one end of the typological spectrum
- The typical approach has been to model these probabilities using a classifier over word types, which works well for languages like English and Chinese, but not all languages

$$\mathsf{P}_{word}(c_\eta \rightarrow w_\eta) = \mathsf{P}(\mathsf{Term{=}1} \mid c_\eta) \cdot \underset{w_\eta}{\mathsf{SoftMax}}(\mathsf{ResNet}(\mathbf{v}_{c_\eta}))$$

- For more morphologically rich languages, sequence models can be used to capture word-internal patterns [2]

$$\mathsf{P}_{char}(c_\eta \rightarrow w_\eta) = \mathsf{P}(\mathsf{Term{=}1} \mid c_\eta) \cdot \prod_{l_i \in \{l_1, \ldots, l_n\}} \mathsf{P}(l_i \mid c_\eta, l_1, \ldots, l_{i-1})$$

Main Observations

- Multilingual induction experiments show favorable results for character-based model, with especially higher results on morphologically rich languages than word-based model [2]

Models	Individual languages										Average
	Arabic	Chinese	English	French	German	Hebrew	Japanese	Korean	Polish	Vietnamese	
word (RH)	23.0	20.8	29.7	29.8	33.8	21.6	29.8	11.7	**22.0**	**15.1**	23.7
char (RH)	**29.1**	**23.9**	**33.4**	**40.7**	**39.3**	**29.5**	**40.2**	**16.3**	21.0	12.8	**28.5**
word (F1)	36.9	41.3	44.4	41.5	44.4	40.0	42.4	23.3	**35.2**	**37.5**	38.7
char (F1)	**42.0**	**44.9**	**49.9**	**51.5**	**47.7**	**48.6**	**55.9**	**34.6**	33.1	28.7	**43.7**

- Highlights the importance of typological properties of the target language for deploying grammar induction systems
- However, the typology of the target language is likely to be **unknown** in an under-resourced setting
- Additionally, the amount and distribution of available text may result in more accurate syntactic analyses from one model over another

Future Work: Leveraging Representations From Neural Language Models

- These factors complicate *a priori* decisions about the model to be employed and motivate a **unification** of word-level and character-level information in a data-driven manner
- This can be achieved by leveraging vectorial representations from neural language models [1, 4], which are trained to predict the correct word given its neighboring context and therefore does not require manually annotated data
- These language models have been shown to yield high-quality representations that embody different levels of linguistic information
- Preliminary experiments posed a challenge in the stable estimation of $\mathsf{P}(c_\eta \rightarrow w_\eta)$ using these representations as input, as the same word can lie in vastly different areas of the vector space depending on the context
- Research into techniques to overcome this problem will further advance our understanding of learning syntactic structures as well as our ability to communicate in languages with limited resources

References

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint*, 2018.

[2] L. Jin, B.-D. Oh, and W. Schuler. 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, 2021.

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

[4] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. *OpenAI Technical Report*, 2019.

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