

Language Modeling and Hierarchical Sentence Structure

How do Language Models (LMs) process hierarchical syntactic structure?

Sentence structure is hierarchical and recursive

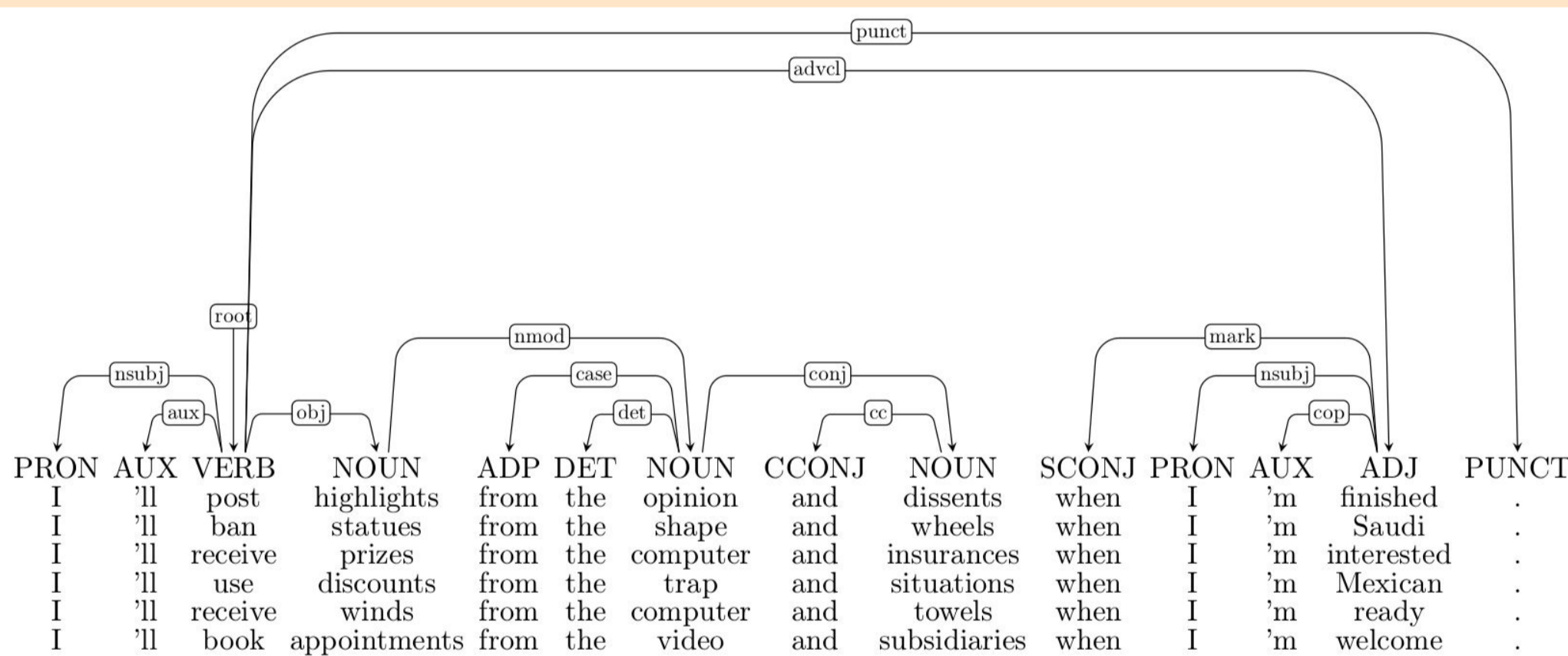
- Various approaches to modeling tree-based structure
- Differences in annotation
 - Format (constituency and dependency trees)
 - Content (tree properties, linguistic assumptions)

LMs are trained on sequential tasks

- Self-supervised learning
- Autoregressive LMs, $p(w_t | w_{<t})$ *predict next*
- Masked LMs, $p(w_t | w_{<t}, w_{>t})$ *word in [MASK] middle*

What do sequential LMs learn about syntax?

- Arps et al. (2022, 2024): Understanding how LMs represent and process nonsensical sentences
- Nonce data: separate syntactic and semantic information
- LM architectures behave differently: Autoregressive LM perplexity increases more than masked LM scores
- LM representations encode dependency trees for nonce data, shown via probing

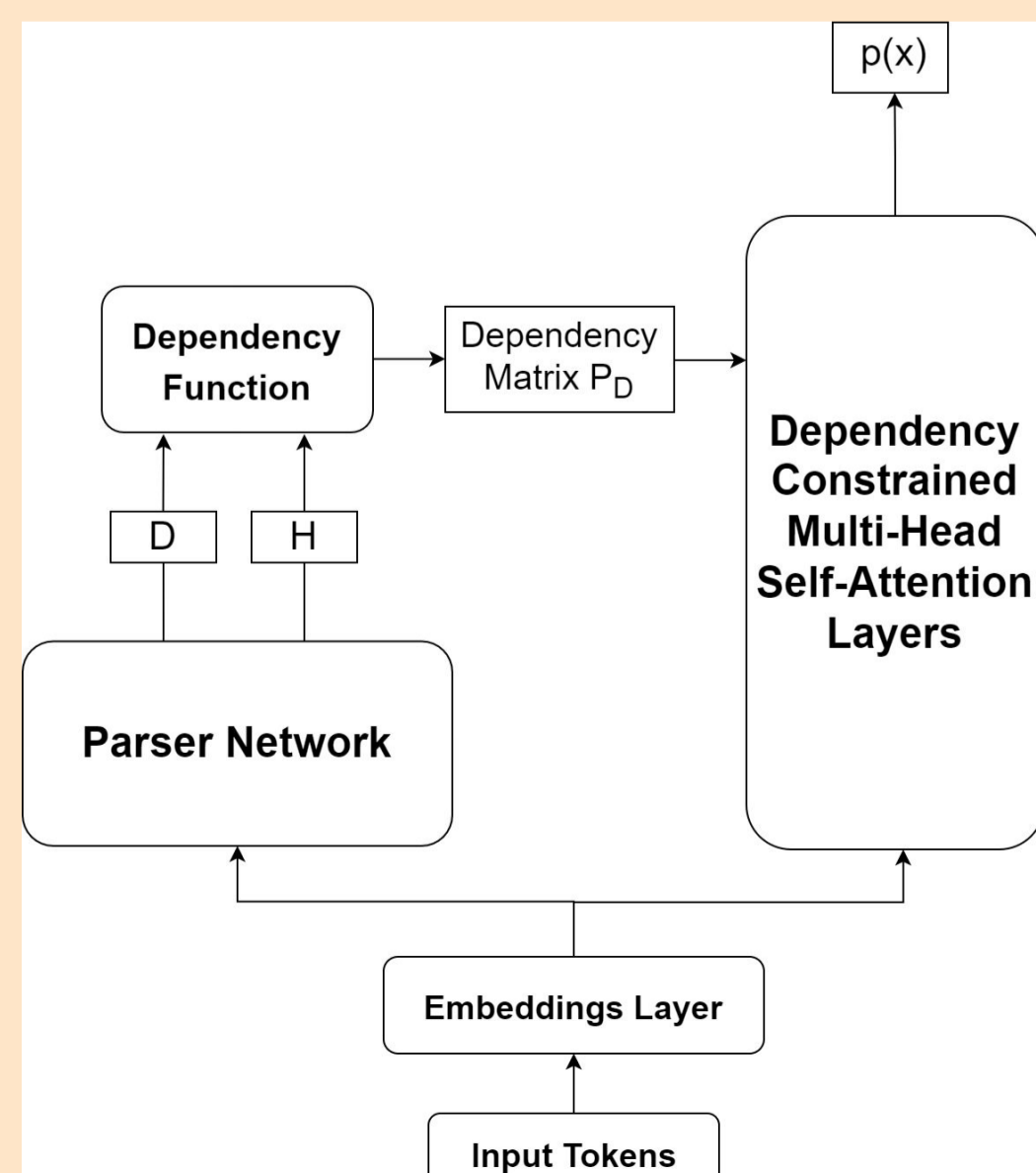


	RelAcc		UAS		LAS		RelAcc		UAS		LAS	
	orig	Δ	orig	Δ	orig	Δ	orig	Δ	orig	Δ	orig	Δ
ar	82.8	5.5	63.1	7.3	55.5	8.3	85.2	7.7	59.5	12.6	53.4	13.5
de	92.7	1.7	84.4	3.4	80.2	4.0	92.6	3.6	84.2	6.8	79.9	8.1
en	87.1	2.3	74.2	3.8	67.9	4.1	82.8	5.2	58.9	5.5	52.2	6.5
fr	89.3	2.0	76.0	4.4	70.4	4.3	87.5	2.8	69.6	6.2	64.0	6.7
ru	88.2	1.5	75.4	3.2	69.2	3.1	88.7	3.0	75.6	5.4	69.6	6.0

mBERT (masked) mGPT (autoregressive)
Probing results for two multilingual models.
Δ indicates performance drop on nonce data

Building LMs that use hierarchical structure

- Adapt neural architecture to latently model trees
- Keep self-supervised learning objectives: Build joint LM and unsupervised parser
- Evaluate on parsing, linguistic and NLP tasks
- **Bonus:** Hierarchy is usually abstract, we make it explicit!
- Momen et al. (2023) train StructFormer (Shen et al., 2021)
- Induced dependency trees constrain transformer self-attention
- Similar performance to transformer baseline, but higher interpretability



Future Research Goals

- Which forms of syntactic inductive bias...
 - work best on different languages?
 - have the best scaling behavior?
- Do different architectures induce similar trees when trained on the same data?
- How do the induced structures relate to existing treebanks?
- How useful are the emerging structures for NLP applications?
- Focus on constituency structure and masked LMs

References

David Arps, Younes Samih, Laura Kallmeyer, Hassan Sajjad. 2022. Probing for Constituency Structure in Neural Language Models. In Findings of EMNLP.
David Arps, Laura Kallmeyer, Younes Samih, Hassan Sajjad (2024). Multilingual Nonce Dependency Treebanks: Understanding how Language Models Represent and Process Syntactic Structure. NAACL.
Omar Momen, David Arps, Laura Kallmeyer. 2023. Increasing The Performance of Cognitively Inspired Data-Efficient Language Models via Implicit Structure Building. BabyLM Shared Task at CONLL.
Yikang Shen, Yi Tay, Che Zheng, Dara Bahri, Donald Metzler, Aaron Courville. 2021. StructFormer: Joint Unsupervised Induction of Dependency and Constituency Structure from Masked Language Modeling. ACL-IJCNLP.