Alexis Cooper

University of Colorado Boulder

Tree-Planted Translation for Free-Order, Case-Marking Languages

Motivation

- **Free-order, case-marking languages** tend to require more for adequate machine translation.[1][2]
- Many languages do not have the scale of data required to **implicitly** pick up this more fluid morphosyntactic structure.
- **Can we explicitly teach syntactic structure?**

Method

- Supervised attention through **Tree-Planting**[3]
- Model dependency graph and train attention head on distance
- Claim: **training efficiency** of syntactic language models with **inference efficiency** of transformers

[1] Arianna Bisazza, Ahmet Üstün, Stephan Sportel; On the Difficulty of Translating Free-Order Case-Marking Languages. *Transactions of the Association for Computational Linguistics* 2021; 9 1233–1248.

[2] Gabriele Sarti, Arianna Bisazza, Ana Guerberof-Arenas, and Antonio Toral. 2022. DivEMT: Neural Machine Translation Post-Editing Effort Across Typologically Diverse Languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7795–7816, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

[3] Ryo Yoshida, Taiga Someya, and Yohei Oseki. 2024. Tree-Planted Transformers: Unidirectional Transformer Language Models with Implicit Syntactic Supervision. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5120–5134, Bangkok, Thailand. Association for Computational Linguistics.

[4] Gil Rosenthal. 2023. Machina cognoscens: Neural machine translation for latin, a case-marked free-order language.

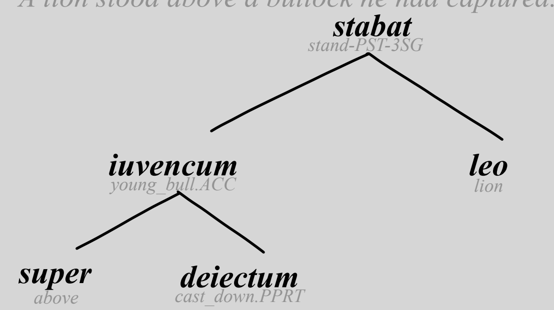
[5] Milan Straka, Jana Straková, and Federica Gamba. 2024. ÚFAL LatinPipe at EvaLatin 2024: Morphosyntactic Analysis of Latin. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 207–214, Torino, Italia. ELRA and ICCL.

[6] Shikhar Murty, Pratyusha Sharma, Jacob Andreas & Christopher D. Manning. 2023. Characterizing intrinsic compositionality in transformers with tree projections. In ICLR 2023: The eleventh International Conference on Learning Representations. Kigali.

Example

Super iuvencum stabat deiectum leo.

A lion stood above a bullock he had captured.



	super	*iuvencum*	*stabat*	*deiectum*	*leo*
super	0.59	0.21	0.081	0.081	0.029
iuvencum	0.16	0.44	0.164	0.164	0.06
stabat	0.06	0.18	0.498	0.067	0.183
deiect	0.08	0.21	0.081	0.592	0.029
leo	0.03	0.08	0.23	0.031	0.624

↓ (softmax)

0.592	0.217	0.081	0.081	0.029
0.164	0.446	0.164	0.164	0.06
0.067	0.183	0.498	0.067	0.183
0.081	0.217	0.081	0.592	0.029
0.031	0.084	0.23	0.031	0.624

Construction of an attention supervision matrix for a short example sentence.

We convert the tree to an $\ell \times \ell$ **matrix** capturing the **syntactic distance between all word pairs.** We then convert the rows to a **probability distribution** using softmax.

The subword attention native to the model is converted to word-level attention by **averaging over the attention of all tokens within a word.**

This produces **two $\ell \times \ell$ matrices** which can then be **directly compared** using their KL Divergence - our loss function!

Results

†: baseline comparison

	BLEU	METEOR
Google Translate†	19.4	**0.467**
Rosenthal (2024)†	**22.43**	-
Finetune (10 epochs)	17.855	0.387
Finetune (30 epochs)	15.950	0.366
Tree-Planted	14.070	0.341

Experiments

- **Data:**
  - **~100k Classical Latin-English parallel sentences**[4] and automatically-generated[5] **dependency parses**
- **Baseline Model:**
  - **Helsinki-NLP it-en** finetuned for 30 epochs on **parallel sentences only**
- **Tree-Planted Model:**
  - **Helsinki-NLP it-en** finetuned for 10 epochs on **parallel sentences**, 20 epochs **both parallel sentences and trees**

Discussion

Tree-Planting seemed to **impede performance.** Potential causes:

- Syntactic knowledge may be implicitly encoded across neurons rather than within one head.
  - **Next step:** Probe model weights for implicit tree structure[6] - is it impeding hierarchies, or are hierarchies not useful?
- Implementation may not be optimal. Italian tokenizer may not capture Latin morphology.
  - **Next step:** Train a tokenizer on Latin text specifically. Tune hyperparameters and tree-planted head configuration.