上海科技大学
ShanghaiTech University

A Systematic Study of Compositional Syntactic Transformer Language Models

Yida Zhao, Hao Xve, Xiang Hu, Kewei Tu

School of Information Science and Technology, ShanghaiTech University

Ant Group

立志成才 报国裕民

Background

- Transformer lacks the inductive bias of **syntactic structures**, which is believed to improve generalization.
- Syntactic language models (SLMs) are developed to address this by **both modeling a linearized syntactic parse tree and the sentence**.

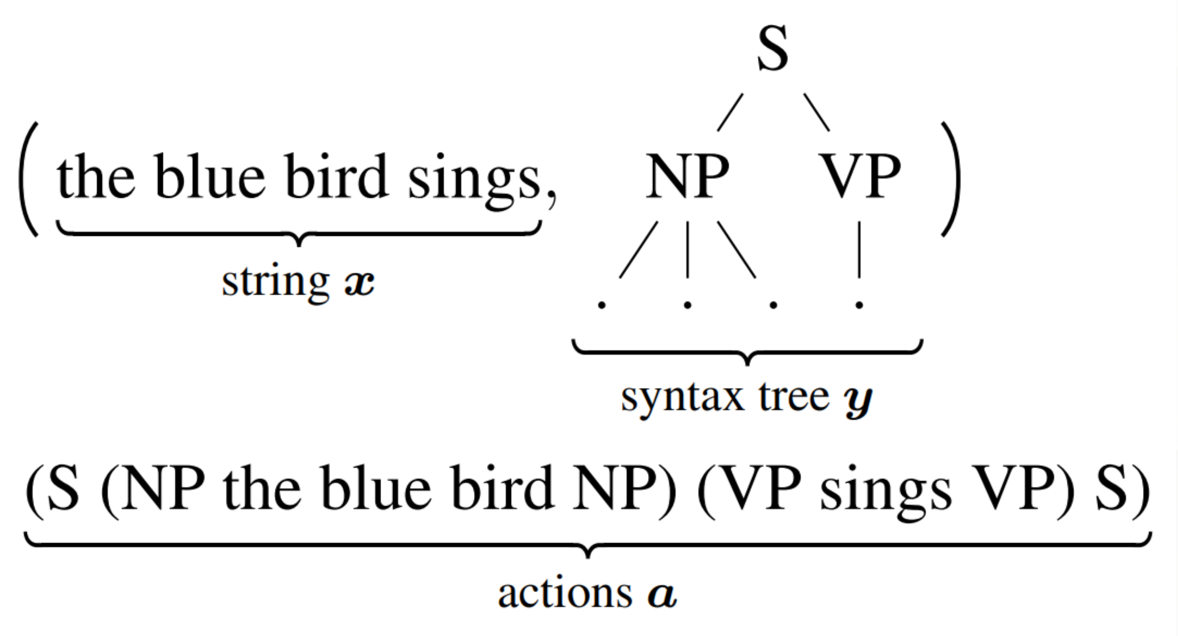


Figure from Transformer Grammars (Sartran et al., 2022).

Our Work

上海科技大学
ShanghaiTech University

- **Compositional SLMs** are based on constituency trees, forming constituent representations via **explicit composition** of sub-constituents.

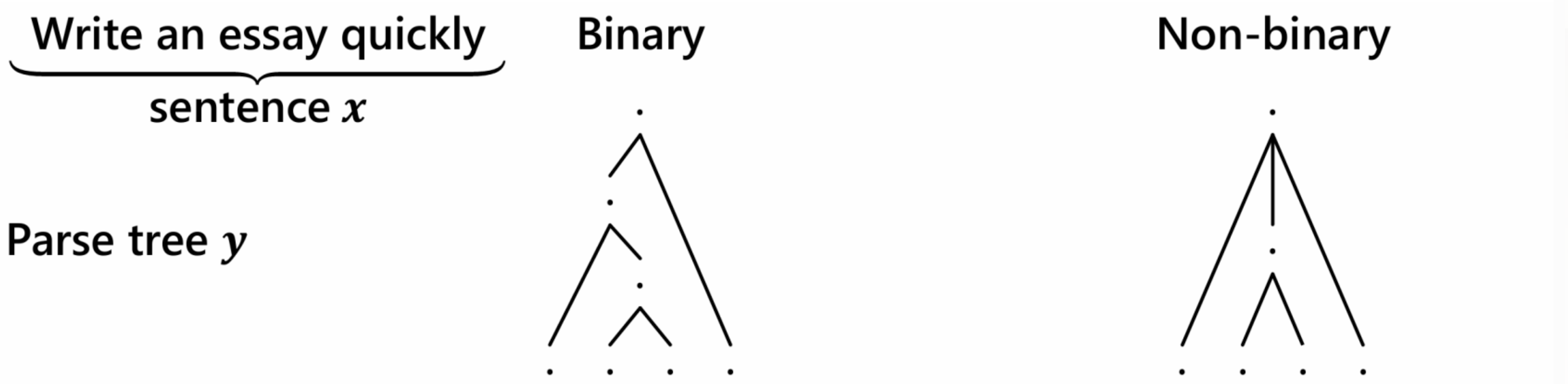
- Our work:
  - A unified framework of compositional SLMs
  - A comprehensive empirical evaluation of SLMs and multiple recommendations on design choices

A Framework for Compositional SLMs

- Our frame work encompasses four design choices:
  - Parse Tree Binarization
  - Linearization Method
  - Composition Function
  - Sub-Constituent Masking

Design Choices

1. **Parse Tree Binarization**: To use non-binary **(Nb)** trees or binary **(Bi)** trees.
2. **Linearization Method**: To linearize trees in a top-down **(Dn)** or a bottom-up **(Up)** manner.



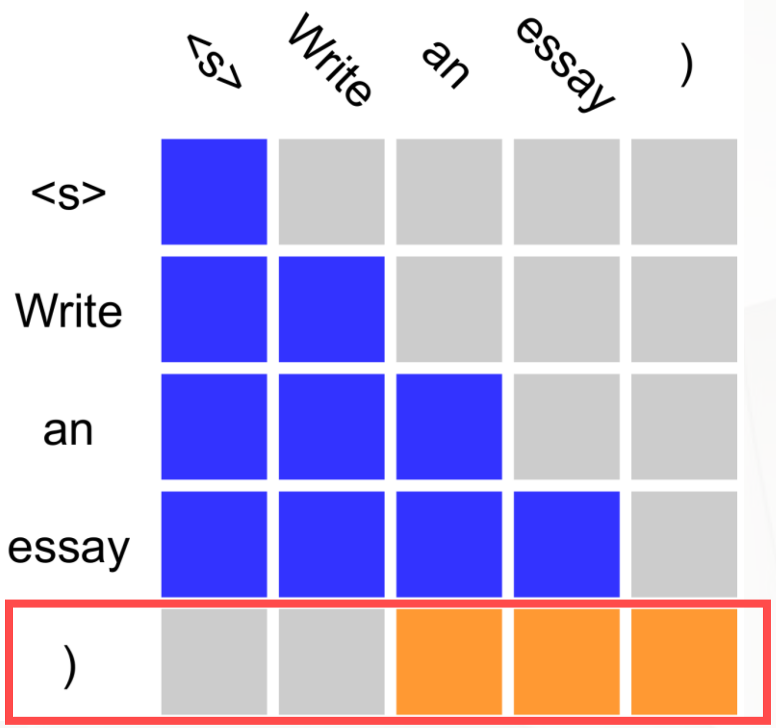
	Binary	Non-binary
Top-down	((Write (an essay)) quickly)	(Write (an essay) quickly)
Bottom-up	Write an essay)) quickly)	Write an essay) quickly)

special treatment

Design Choices

3. Composition Function:

- **Internal (In)**: Conduct composition **<u>within the Transformer</u>**. We duplicate the ")", with the first for composition and the second for next token prediction (Sartran et al., 2022).

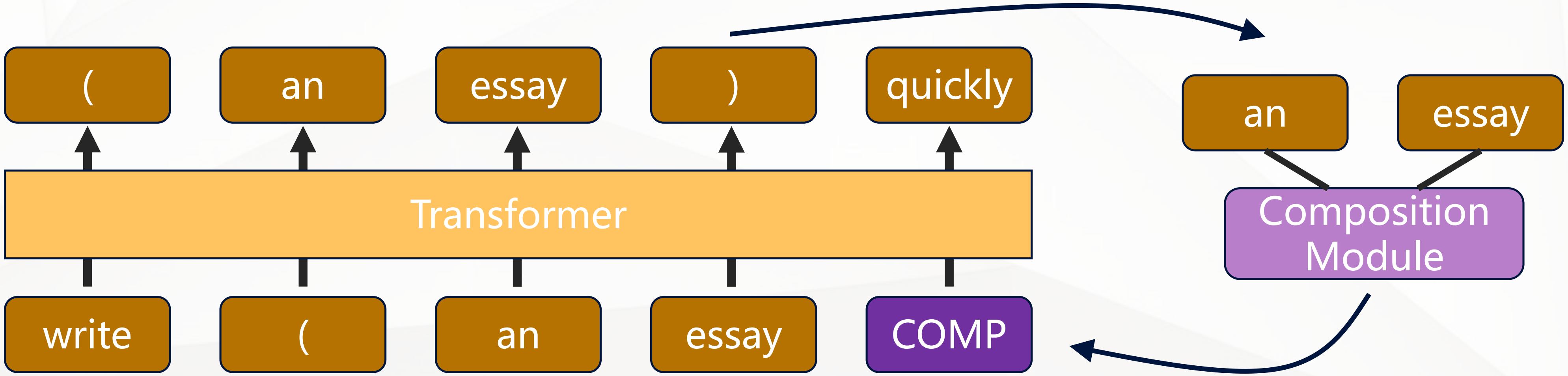


Design Choices

上海科技大学 ShanghaiTech University

3. Composition Function:

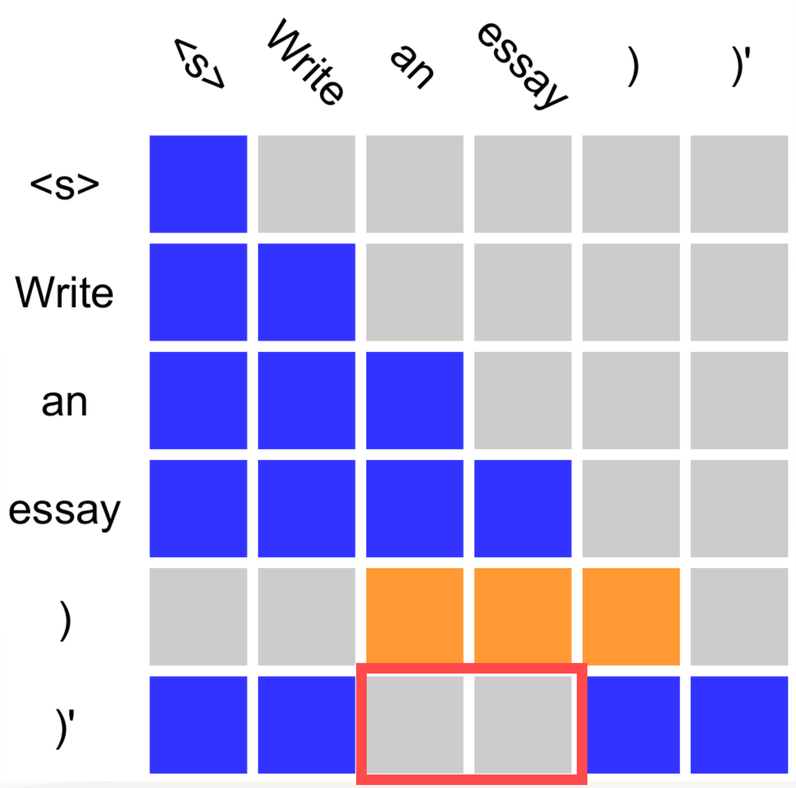
- **Internal (In)**: Conduct composition **within the Transformer**. We duplicate the ")", with the first for composition and the second for next token prediction (Sartran et al., 2022).
- **External (Ex)**: Conduct composition **with an external module**. The composed representation is fed back into the main Transformer.



Design Choices

上海科技大学 ShanghaiTech University

4. Sub-Constituent Masking: To mask out the already composed sub-constituents **(M)** or not **(Nm)**.



A Framework for Compositional SLMs

上海科技大学 ShanghaiTech University

- Two options for each of the four key aspects → sixteen distinct SLMs
  - Each variant is named on its configuration across four aspects, e.g., **Bi-Dn-In-M**.

- Three previous studies accommodated within our framework:
  - Transformer Grammars (Sartran et al., 2022): **Nb-Dn-In-M**.
  - Composition Attention Grammars (Yoshida and Oseki, 2022): **Nb-Dn-Ex-M**.
  - Generative Pretrained Structured Transformers (Hu et al., 2024): **Bi-Up-Ex-Nm**.

Experiments

- We compare the sixteen compositional SLMs from our framework with two Transformer baselines:
  - (i) **GPT2-token**, a traditional language model of token sequences.
  - (ii) **GPT2-tree**, a syntactic language model of linearized trees without explicit composition.
- All the models are trained from scratch on the BLLIP-LG dataset.
- We use an off-the-shelf CRF constituency to reparse the dataset and obtain silver constituency trees for training.

Language Modeling & Syntactic Generalization

- Language modeling (PPL) **may not benefit** from the inductive bias of syntax.
- Explicit modeling of syntax **improves the syntactic generalization** (SG) as expected.
- See our paper for more detailed analyses.

Model	PPL[†] (↓)	SG (↑)
GPT2-token	**17.31**	64.1
GPT2-tree	19.97	73.1
Bi-Up-Ex-Nm	20.51	80.1
Bi-Up-Ex-M	24.15	**82.4**
Bi-Up-In-Nm	19.99	77.5
Bi-Up-In-M	21.32	79.7
Bi-Dn-Ex-Nm	23.62	80.2
Bi-Dn-Ex-M	27.21	80.9
Bi-Dn-In-Nm	22.02	79.4
Bi-Dn-In-M	26.50	80.9
Nb-Up-Ex-Nm	23.85	40.8
Nb-Up-Ex-M	24.07	51.8
Nb-Up-In-Nm	19.36	79.6
Nb-Up-In-M	22.01	73.4
Nb-Dn-Ex-Nm	20.88	51.1
Nb-Dn-Ex-M	25.15	51.9
Nb-Dn-In-Nm	18.11	78.1
Nb-Dn-In-M	22.30	75.6

Downstream Tasks (Summarization & Dialogue)

上海科技大学 ShanghaiTech University

Model	Xsum				DailyDialog			
	R-1	R-2	R-L	R-AVG	R-1	R-2	R-L	R-AVG
GPT2-token	27.14	7.67	21.65	18.82	14.02	3.82	13.31	10.38
GPT2-tree	**29.59**	**9.47**	**23.58**	**20.88**	**14.99**	3.83	**14.31**	**11.04**
Bi-Up-Ex-Nm	29.04	8.95	23.01	20.33	14.14	**4.01**	13.61	10.59
Bi-Up-Ex-M	23.48	5.75	18.84	16.02	12.44	3.00	11.69	9.04
Bi-Up-In-Nm	28.93	8.97	22.97	20.29	13.01	3.33	12.19	9.51
Bi-Up-In-M	24.84	6.64	19.88	17.12	12.05	2.78	11.40	8.74
Nb-Up-In-Nm	29.05	9.06	23.21	20.44	13.81	3.68	13.09	10.19
Nb-Up-In-M	24.30	6.28	19.45	16.68	11.92	2.93	11.33	8.72
Nb-Dn-In-Nm	29.48	9.40	23.54	20.81	13.99	3.85	13.36	10.40
Nb-Dn-In-M	26.10	7.31	20.98	18.07	12.50	3.31	11.90	9.23

- Some SLMs can outperform the standard GPT-2 token baseline.
- Composition is not critical in generation tasks.
- See our paper for more detailed analyses of specific configurations.

Other Experiments and Findings

上海科技大学 ShanghaiTech University

- See our paper for more experiments and findings:
  - Efficiency evaluation
  - Recommendations on design choices
  - ……

Conclusion

- We propose a unified framework for compositional syntactic language models (SLMs).
- We evaluate all the variants on different tasks, providing a clearer understanding of the design trade-offs involved and offering some recommendations for future research and application.

THE END

上海科技大学
ShanghaiTech University

Thank you!

Paper

Code

立志成才 报国裕民