EMNLP 2021

Grounded Graph Decoding improves Compositional Generalization in Question Answering

Yu Gai*, **Paras Jain***, Wendi Zhang
Joseph Gonzalez, Dawn Song, Ion Stoica

UC Berkeley RISELab

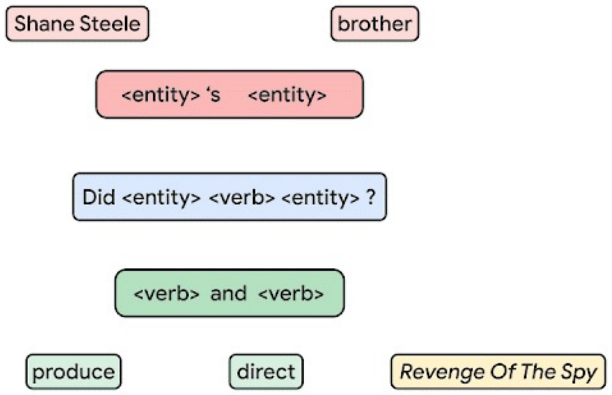
ucbrise.github.io/graphdecoder

Can neural networks "make infinite use of finite means" with language? (Chomsky and Lightfoot 2002)

Compositionality is a critical component of complex language

Can neural networks "make infinite use of finite means" with language? (Chomsky and Lightfoot 2002)

Humans can form novel compound sentences from simple constructs!



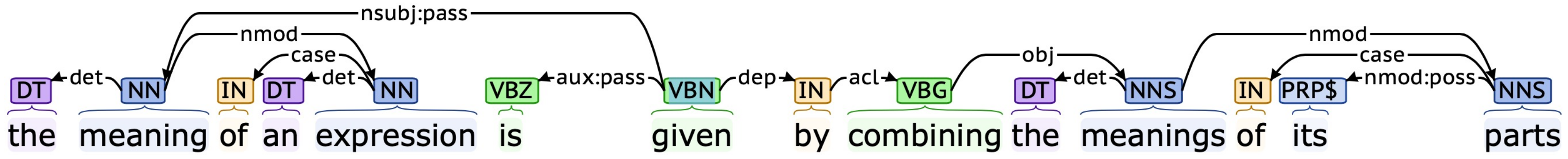
From Keysers et al, 2020

Compositionality is a critical component of complex language

Can neural networks “make infinite use of finite means” with language? (Chomsky and Lightfoot 2002)

Humans can form novel compound sentences from simple constructs!

Compositionality: the meaning of an expression is given by combining the meanings of its parts (Montague 1970, Frege 1884)



Benchmarking Compositional Generalization

Compositional Freebase Questions (CFQ): successor to the SCAN dataset to evaluate real-world performance of question answering models

Freebase™

Who directed and produced Goldfinger and Inception?

$$\exists x.\ \mathrm{direct}(x, \mathrm{Goldfinger}) \wedge \mathrm{produce}(x, \mathrm{Goldfinger})$$
$$\wedge \mathrm{direct}(x, \mathrm{Inception}) \wedge \mathrm{produce}(x, \mathrm{Inception})$$

Maximum Compound Divergence splits (MCD1, MCD2, MCD3) test **generalization to unseen compositions** at test-time

"Measuring Compositional Generalization: A Comprehensive Method on Realistic Data", Keysers et al 2020

Grounded Graph Decoding improves compositional generalization

Challenge 1: Challenging to retain syntax structure from complex inputs → **Grounding via Attention**

Challenge 2: SPARQL output domain is group and permutation invariant → **Conjunctive graph decoding**

Results: SOTA results (+36% MCD1 accuracy) with smaller models and no pre-training

Method	# Params.	Accuracy per-split MCD1	MCD2	MCD3
LSTM w/ attention (Keysers et al., 2020)		$28.9 \pm 1.8\%$	$5.0 \pm 0.8\%$	$10.8 \pm 0.6\%$
Transformer (Keysers et al., 2020)		$34.9 \pm 1.1\%$	$8.2 \pm 0.3\%$	$10.6 \pm 1.1\%$
Universal Transformer (Keysers et al., 2020)		$37.4 \pm 2.2\%$	$8.1 \pm 1.6\%$	$11.3 \pm 0.3\%$
Evolved Transformer (Keysers et al., 2020)		$42.4 \pm 1.0\%$	$9.3 \pm 0.8\%$	$10.8 \pm 0.2\%$
T5-base (Furrer et al., 2020)	220M	$57.6 \pm 1.4\%$	$19.5 \pm 1.0\%$	$16.6 \pm 1.5\%$
T5-large (Furrer et al., 2020)	770M	$63.3 \pm 0.6\%$	$22.2 \pm 1.5\%$	$18.8 \pm 2.6\%$
T5-11B (Furrer et al., 2020)	11000M	$61.4 \pm 4.8\%$	$30.1 \pm 2.2\%$	$31.2 \pm 5.7\%$
T5-11B (modified) (Furrer et al., 2020)	11000M	$61.6 \pm 12.4\%$	$31.3 \pm 12.8\%$	$33.3 \pm 2.3\%$
Grounded Graph Decoding	0.3M	$\mathbf{97.9 \pm 0.2\%}$	$\mathbf{47.1 \pm 10.4\%}$	$\mathbf{50.8 \pm 17.2\%}$

⬅️ Pre-training
+19.2% MCD1

⬅️ Ours
+36% MCD1 vs. T5

Grounded Graph Decoding improves Compositional Generalization in Question Answering

Yu Gai*, **Paras Jain***, Wendi Zhang, Joseph Gonzalez, Dawn Song, Ion Stoica

https://github.com/ucbrise/graphdecoder

- **Compositional generalization** enables expressing complex concepts from simple constructs
- **Large + pretrained** models perform poorly at compositional generalization
- We introduce (a) **grounding** and (b) **graph decoding** to mitigate common compositional failures.
- We improve MCD1 accuracy on the CFQ dataset by **36%**.