WP 1 : Identifying Compositionality-Related Biases

Compo Kick-Off

Guillaume Wisniewski

January 25th 2024

Goal(s)

What Benoit asked me to do

- talk about “bias” and “compositionality”

What I will talk about

- what we have promised
- what other have proposed
- what we (have done|could do)

What have we promised...

In the proposal:

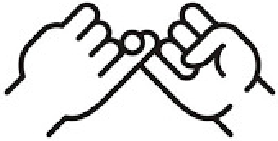
- goal of the WP: experimental setting to capture/describe compositionality
- tasks

↪ idioms and machine translation

↪ artificial languages (COGS like)

↪ the assignment task

⟹ all very good ideas



Related work (i) : COGS & SLOG

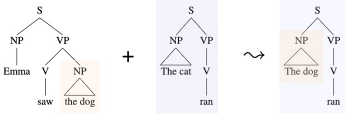
Benchmarks for Compositional Generalization

- lexical generalization: novel combination of known lexical items
- structural generalization: ability to combine known structures

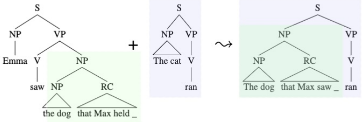
↪ evaluation on a semantic parsing task

How?

- CFG to model (a subset of English)
- OOD evaluation: different “structures” in the test and train sets



(a) Lexical generalization: object → subject (COGS)



(b) Structural generalization: RC object→RC subject (SLOG)

Related Work (2) : Idiomatic Expressions

E. Liu and G. Neubig Are Representations Built from the Ground Up? An Empirical Examination of Local Composition in Language Models, EMNLP'22

Motivating Example

il pleut des cordes

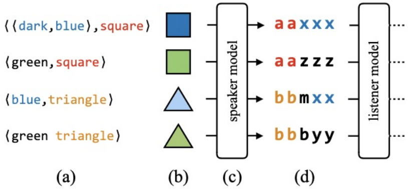
↓

it is raiing cats and dogs

- idiomatic expression: anti-compositional pattern
- meaning can not be built from its parts
- many works have discussed the importance of detecting such patterns

Related Work (3): Measuring Compositionality

J. Andreas, Measuring Compositionality in Representation Learning, ICLR'17



- artificial tasks
- simple idea:

↪ identify "part"

↪ identify a compositional operator (e.g.+)

↪ can you find representation of the part so that

- can we use it?

Proposal n°1: Extending COGS/SLOG with idioms

- everything is in the title 😄
- interesting point: “define” (non-)compositionality
- identified in the proposal
- difficulties: does it make sense?

↪ can idioms be used in several contexts?

Proposal n°ibis: Consider another task

- rather than semantic parsing → any other task

Proposal n^{o}2: The Assignment Task

Y. Zhang, A. Backurs, S. Bubeck, R. Eldan, S. Gunasekar & T. WagnerUnveiling Transformers with LEGO: a synthetic reasoning task, arXiv, 2022.

```
a = + 1; b = - a; c = + b; d = ?
a = + 1; b = - c; c = + a; d = - b
```

Why is it interesting?

- two ways to solve the task:
  1. maintain a mapping between variables and values at every position
  2. create a flow of information (through attention) to propagate values

⇒ easy (in theory) to see if and how a model solve the task

Proposal n°2: The gender assignment task

L. Conti and G. Wisniewski. Using Artificial French Data to Understand the Emergence of Gender Bias in Transformer Language Models. EMNLP'23

La chercheuse termine son travail.
Un artiste termine son travail.
L'artiste termine son travail.

Grammatical Gender in French

- can be marked by the determiner, the noun, the determiner and the noun or not marked at all.
- is the gender captured in the word representation or "computed" at inference time?
- artificial corpora: a word can be seen only in epicene context during training but not at test time
- ⊕ need to propagate the information (e.g. to translate *son* in English)

Proposal n°4: Word tokenization

```
So␣uve␣nt␣,␣pour␣s␣’␣am␣use␣r␣,␣les␣hommes␣d␣’␣équipage␣prennent␣des␣
alba␣tros␣,␣vaste␣s␣oiseaux␣des␣mer␣s␣,␣qui␣suivent␣,␣indo␣lent␣s␣
compagno␣ns␣de␣voyage␣,␣le␣navire␣gli␣ssant␣sur␣
les␣gouf␣fre␣s␣am␣ers␣.␣
```

- 2 levels of compositions:

↪ sub-words → words (considering neighborhood)

↪ word → components (considering grammatical structures)

Are the two mechanisms the same?