Assessing the ability of neural language models to abstract syntactic representations

JTT, April 14, 2022

Bingzhi Li Guillaume Wisniewski Benoît Crabbé
bingzhi.li@yahoo.com

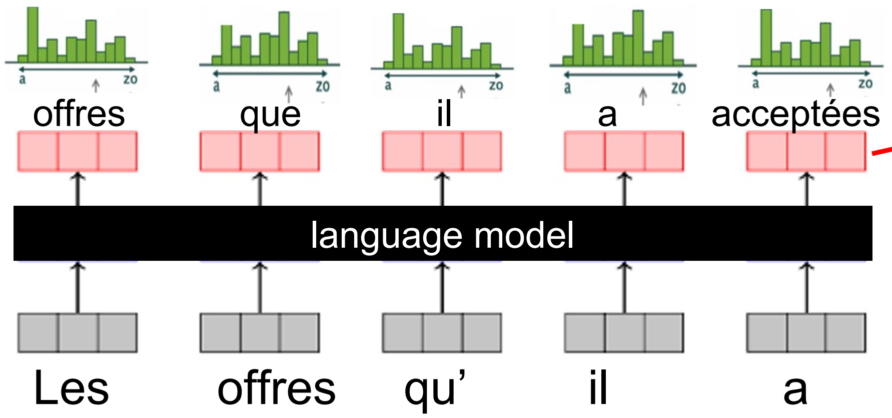
Université Paris Cité

Laboratoire de linguistique formelle

cnrs

Deep Learning Tsunami

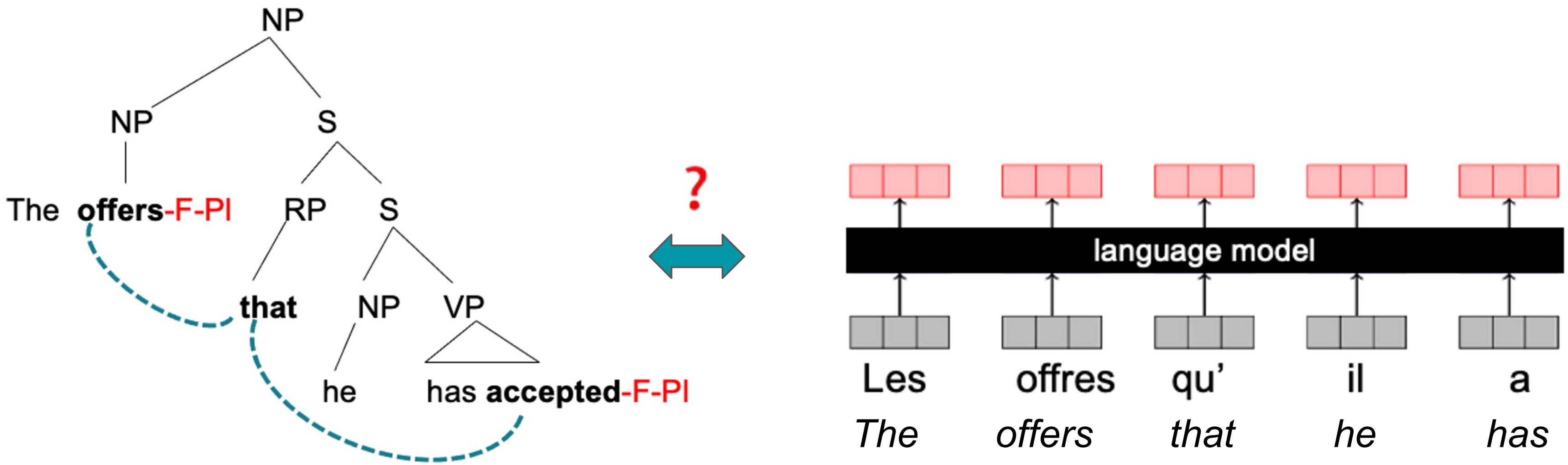
language model : a probability distribution over sequences of tokens



- **NLP tasks**
- Text generation
- Machine translation
- Summarization
- •••

Research question

★ Are neural networks essentially recognizing **superficial patterns** or are they capable of **abstracting** more **general rules** ?



Motivation

Characterize representations

Study prediction mechanisms

Improve training

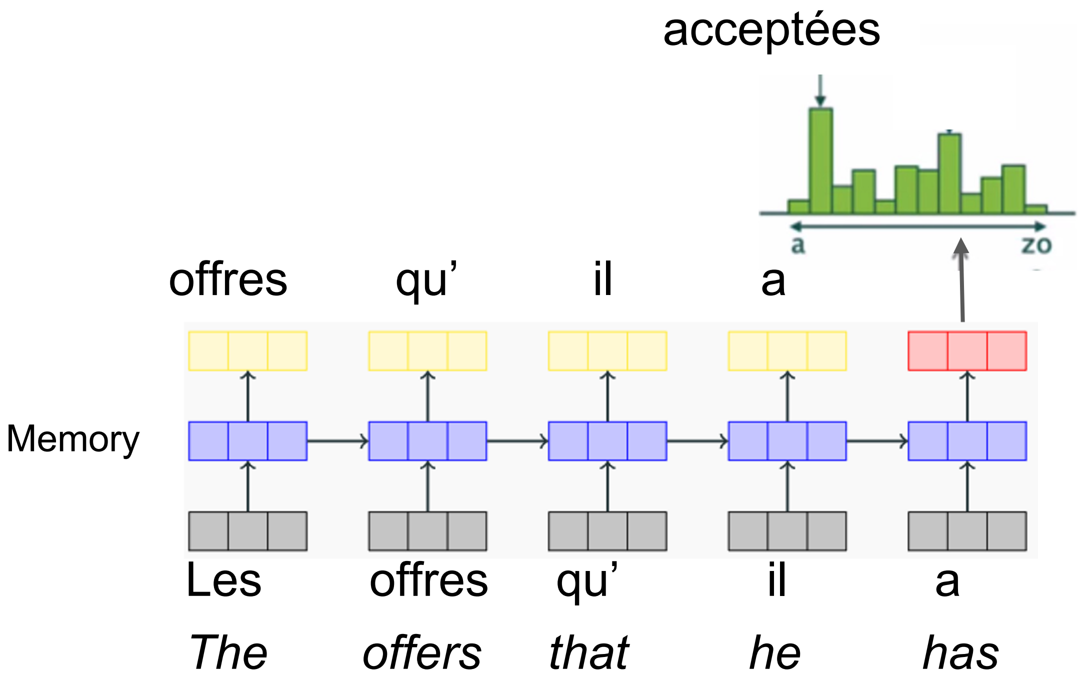
★ More **interpretable** and **accountable** models

★ Novel insights into **linguistic theory** and **human generalization**

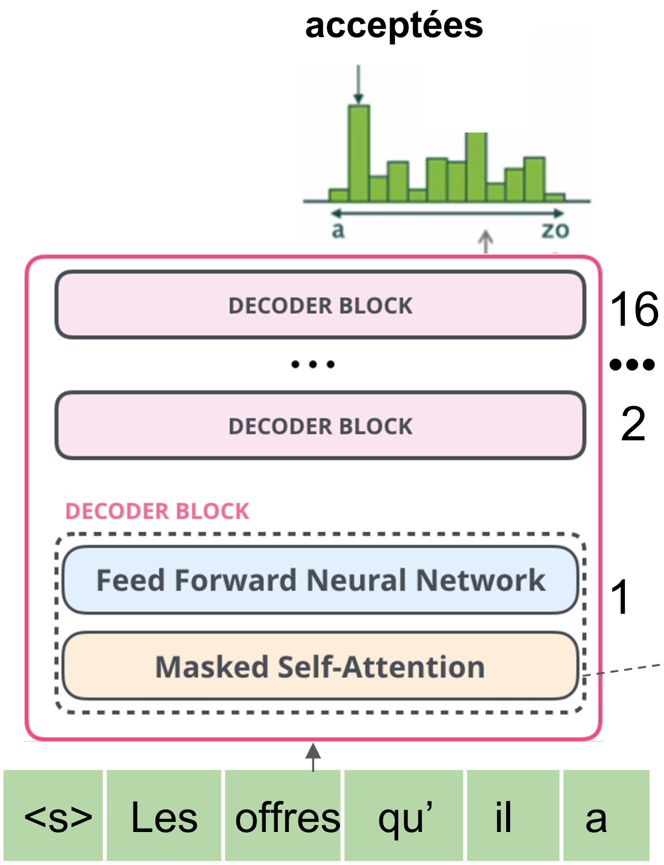
Part I

Do language models learn syntax?

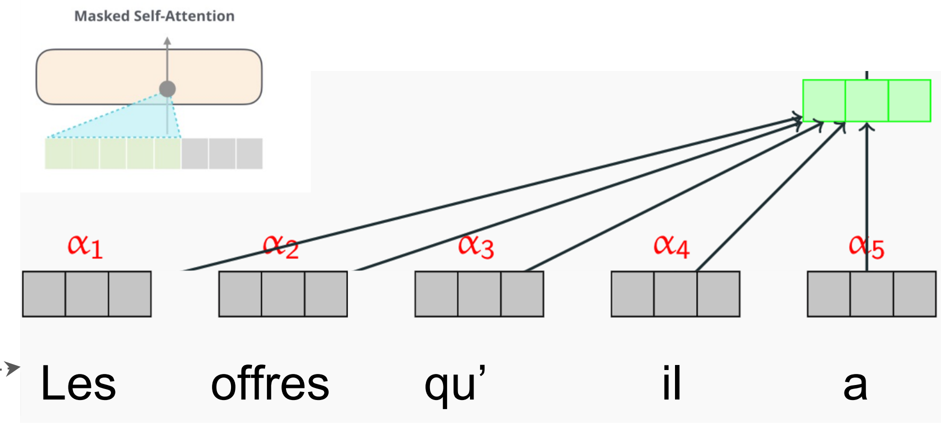
RNN LSTM (Long short-term memory)



Incremental Transformers



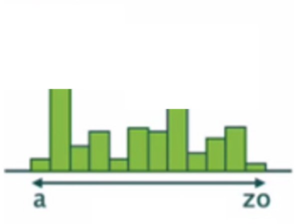
Masked self-attention

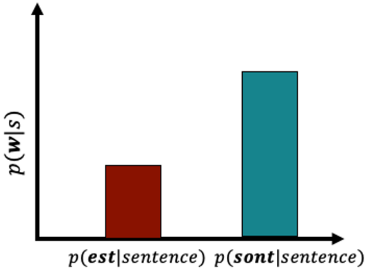


Methodology: Number Prediction Task

Les **bureaux**-pl en <u>métal</u>
qu'il aime ______

Language model





sont-pl

Long-distance agreement as **evidence for syntactic structure**

(Linzen et al 2016; Gulordava et al.2018; Goldberg 2019)

Number Agreement Tasks in French

Les **bureaux** en métal [qu'il a vus dans la rue]$_{\mathrm{RC}}$ **sont** …
Les **bureaux** en métal [qu'il a **vus** dans la rue]$_{\mathrm{RC}}$ sont …

cue target-pp target-V

Two agreement tasks:

- ➢ **subject**-verb across object relative clause
- ➢ Object-past participle

Number Agreement Tasks in French

Les **bureaux** en métal [qu'il a vus dans la rue $]_{RC}$ **sont** …
Les **bureaux** en métal [qu'il a **vus** dans la rue$]_{RC}$ sont …

- Surface form: very similar
- Distinct linguistic analysis:
  subject-verb: ignore embedded clause
  object-past participle: detect filler-gap pattern

Overall Results

LSTM* & Incremental Transformers trained on 80M words of Wikipedia

	#sentences	LSTMs	Transformers
object–verb	68,794	**82.1%**	**94.6%**
subject–verb	27,582	**94,3%**	**98,9%**

High prediction accuracies across the board

- Transformers: consistently much better
- Accuracy subject-V > object-pp

* Be careful: LSTM has much less number of parameters

But … Right for the Wrong reason?

Les **bureaux** en métal [qu’il a **vus** dans la rue]$_{\mathrm{RC}}$ sont beaux.

cue target-pp target-V

→ Identifying “bureaux” as **agreement controller**

→ Memorising “bureaux” as **1st noun**

But … Right for the Wrong reason?

Les **bureaux** en métal [qu'il a **vus** dans la rue]$_{RC}$ sont beaux.

Heuristic	Precision	
	obj-pp	sub-verb
first noun in the prefix	69.5%	83.7%
last noun in the prefix	88.6%	77.5%
last token with a mark of number in the prefix	60.3%	66.9%
majority number expressed in the prefix	70.0%	75.9%
noun before *que*	95.7%	91.6%
LSTM	82.1%	94.3%
Transformer	94.6%	98.9%

Do models learn **more than heuristics**?

Task difficulty measured by count of heuristics

Principle

- More heuristic matching the target number, easier the prediction
- Count of heuristics as a proxy of task difficulty

Examples

5 heuristics (4)Si les **idées**(1)(5) que ces mots(2) représentent(3) **est*/sont**

easiest

•

•

•

0 heuristic Ce soir les **hommes** que j'ai postés sur la route que doit suivre le roi **prendra*/prendront** …

hardest

Results: Subject-Verb Agreement

Constructions	size in sentences	LSTMs	Transformers
Subject-verb across object relative clause			
overall	27,582	$94.9_{\pm 0.8}$	$98.9_{\pm 0.1}$
5 heuristics	14,708	$98.9_{\pm 0.2}$	$99.6_{\pm 0.1}$
4 heuristics	3,799	$96.2_{\pm 0.7}$	$98.9_{\pm 0.1}$
3 heuristics	4,189	$92.4_{\pm 1.3}$	$98.4_{\pm 0.1}$
2 heuristics	3,166	$86_{\pm 2}$	$97.6_{\pm 0.3}$
1 heuristic	1,451	$81.4_{\pm 3.2}$	$96.6_{\pm 0.3}$
0 heuristic	269	$74.1_{\pm 4.3}$	$94.2_{\pm 1.2}$

Results: Object-Verb Agreement

Constructions	size in sentences	LSTMs	Transformers
Object past participle			
overall	68,497	81.7 ±1.4	94.6 ±0.4
5 heuristics	32,149	96.6 ±0.6	99.3 ±0.05
4 heuristics	12,711	85.2 ±1.7	96.3 ±0.3
3 heuristics	9,159	69.9 ±3.0	91.7 ±0.5
2 heuristics	10,621	58.7 ±3.9	87.4 ±0.8
1 heuristic	2,870	36.2 ±5.1	76.9 ±2.3
0 heuristic	987	38.8 ±4.4	73.8 ±2.3

- Performance **degrades** with task difficulty
- accuracy subject-verb **>** object-verb
- **Transformers generalize beyond heuristics**

Part II

How Distributed are Distributed Representation?

Locality of Syntactic Information

Observation: Transformers generalize beyond superficial heuristics

Question: How transformers represents the required syntactic information?

→ Distributed **all over** the sentence (Klafka and Ettinger 2020)

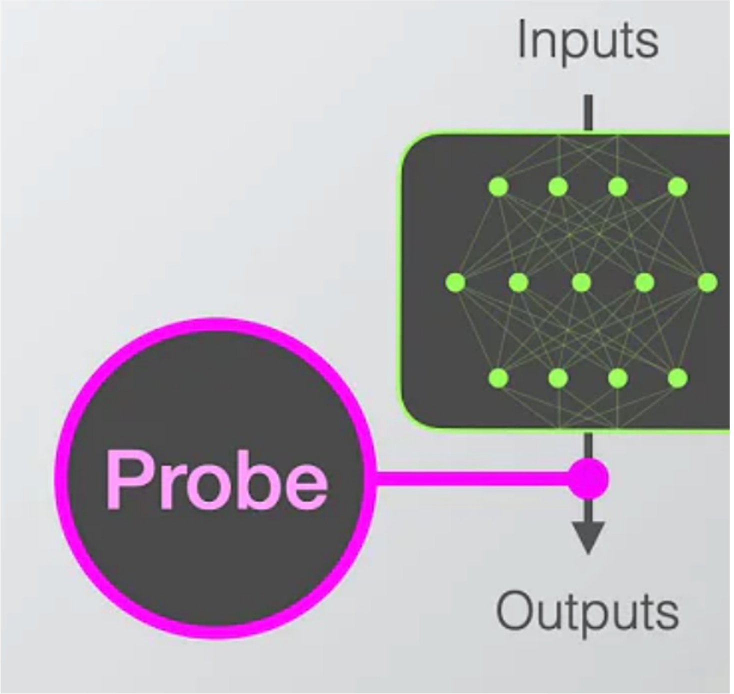
→ Distributed **locally** around the agreement elements

Method: Probing classifiers

Hypothesis: Model's internal representations encode a linguistic property (the number of the cue)

Task: Train a classifier to predict the property from the representations

Evaluation: Accuracy reflects how well the property is encoded



(Veldhoen et al., 2016; Conneau et al., 2018)

Probing for agreement information

plural Les **bureaux** en métal [qu'il a **vus** dans la rue]$_{RC}$ **sont** beaux.

- Each sentence is labelled with the `cue`'s number: sing/plur
- Observation: token representation built by transformers (last layer)
- Train a logistic regression classifier:

  mapping token representation → sentence label

Probing Results

Sans doute les bureaux qu' il a vus dans la rue ... .

No doubt the desk that he has seen in the street ...

	Averaged Accuracy	
	object-past participle	subject-verb
prefix	59.4%	62.1%
context	94.4%	93.6%
suffix	71.6%	79.5%

- Similar pattern: number information encoded locally when relevant

Part III

Is the Agreement Decision Syntactically-motivated?

How transformers used the encoded Information?

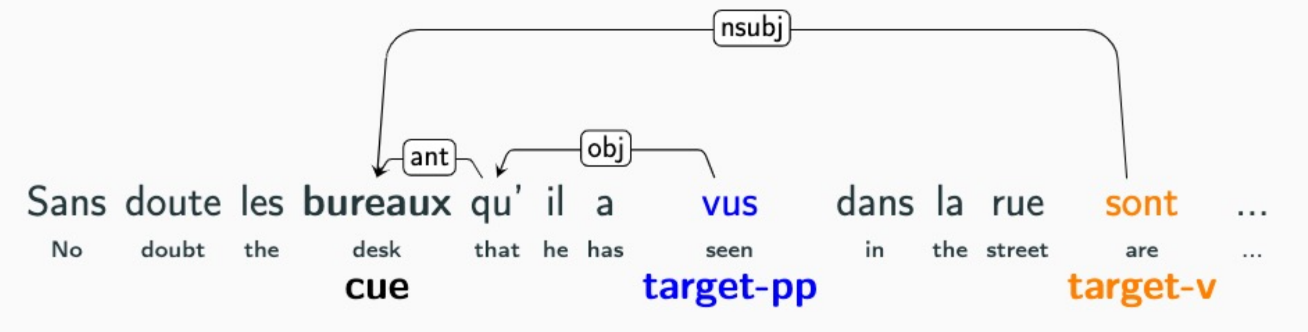
Observation:

- Transformers are able to predict number agreement with high accuracy
- Number information is mainly distributed between the cue and target

Question: How transformers actually used the encoded information?

➢ correlation ≠ causation

Are Transformers Decisions Syntactically Motivated?



Two hypotheses (not mutually exclusive)

Transformers based its prediction on:

- ➢ Tokens involved in the agreement rules
  - cue and 'que'
  - cue
- ➢ All tokens between the cue and the target

Causal Relations?

How to answer such a question:

- Many tools/ approach (Woodward 2003, Pearl et Mackenzie 2018):
  → counterfactual reasoning/intervention

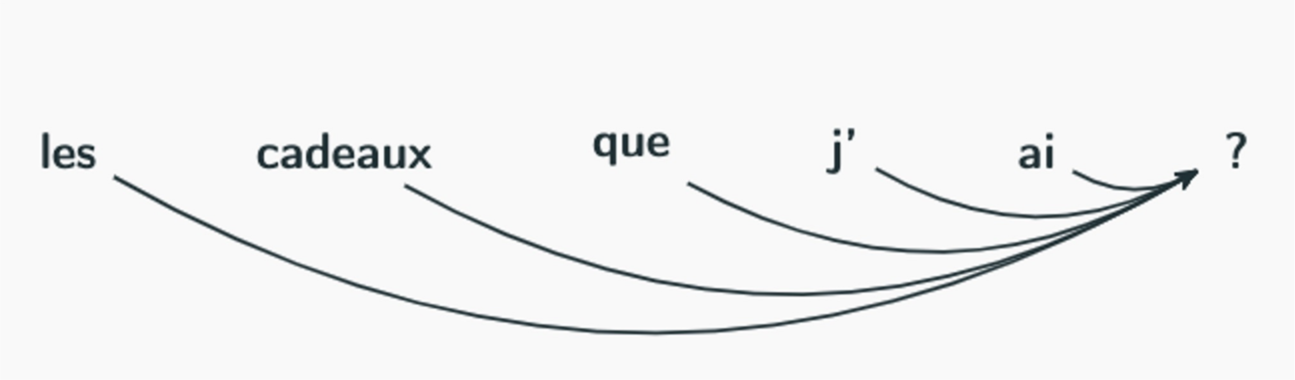
Main idea

Causal explanation between if X and Y:

changing X will change Y

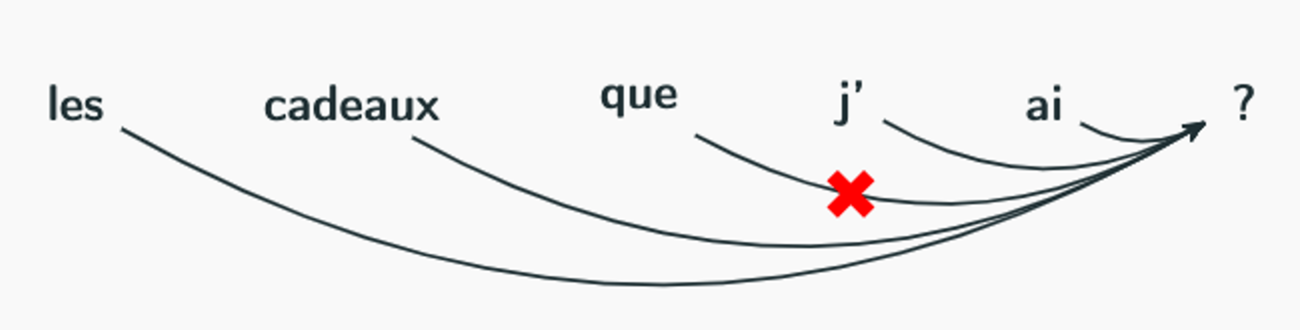
→ causal intervention on self attention

Method: Causal Intervention on Self-attention



➢ Contextualized representation: when predicting the target, model attend all previous words

Method: Causal Intervention on Self-attention



- ➢ Contextualized representation: when predicting the target, Model attend all previous words
- ➢ **Counterfactual** representation: when predicting the target, do not look at token 'que'

Measure change in behavior after intervention

Masking Results for Object-Verb Agreement

Masked tokens	Precision
none	**94.6%**
cue $\oplus$ *que*	66.9%
cue	71.4%
que	78.8%
all context tokens except cue$\oplus$*que*	87.1%

- number information present across all tokens of context
- but Transformers decisions rely mainly on **cue** and **que**

Compared with Subject-Verb Agreement

Masked tokens	Precision	
	obj-pp	sub-verb
none	**94.6%**	**98.9%**
cue ⊕ *que*	66.9%	86.0%
cue	71.4%	89.1%
que	78.8%	96.7%
all context tokens except cue⊕*que*	87.1%	82.0%

- number information present across all tokens of context
- masking 'que' has **little** effect on subject-verb agreement

Masking ‘que’ intervention

Subsets	subj-verb		obj-pp	
	original	mask-que	original	mask-que
5 h	99.6%	99.1%	99.3%	95.9%
4 h	99.0%	96.5%	96.3%	85.3%
3 h	98.4%	93.4%	91.7%	63.9%
2 h	97.7%	91.5%	87.4%	49.4%
1 h	96.8%	94.1%	76.9%	32.1%
0 h	94.1%	89.3%	73.8%	30.8%

- masking ‘que’ **cause** below than chance level performance on object-pp agreement

Masking 'que' intervention

Subsets	subj-verb		obj-pp	
	original	mask-que	original	mask-que
5 h	99.6%	99.1%	99.3%	95.9%
4 h	99.0%	96.5%	96.3%	85.3%
3 h	98.4%	93.4%	91.7%	63.9%
2 h	97.7%	91.5%	87.4%	49.4%
1 h	96.8%	94.1%	76.9%	32.1%
0 h	94.1%	89.3%	73.8%	30.8%

- same distribution of 'number information'

  different prediction mechanism

Conclusions

Discussion and Conclusions

Incremental transformers LM are capable of

- abstracting syntactic representations (though not perfectly)
- basing their predictions on linguistically motivated cues

Open Questions

➢ Frequency bias

➢ Semantic cues

➢ …