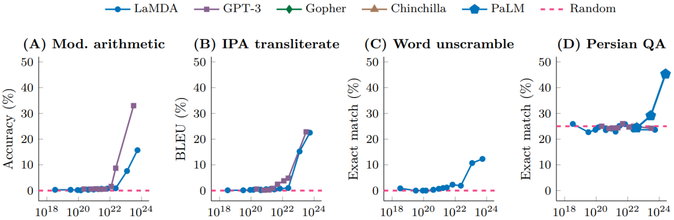
On Synonymy and Language Models

Ioana Ivan

Encadrants: Alexis Nasr et Carlos Ramisch

June 19, 2024

Motivation



[1]

- Simple capabilities that underlie all tasks ?

[1] Wei et al., Emergent Abilities of Large Language Models, 2022

Motivation

What can constitute a simple capability that can be evaluated ?

Motivation

What can constitute a simple capability that can be evaluated ?
Linguistic capabilities.

Motivation

What can constitute a simple capability that can be evaluated ?
Linguistic capabilities.
Which linguistic concepts ?

Motivation

What can constitute a simple capability that can be evaluated ?
Linguistic capabilities.
Which linguistic concepts ?
Lexical semantics.

Motivation

What can constitute a simple capability that can be evaluated ?
Linguistic capabilities.
Which linguistic concepts ?
Lexical semantics.

Research questions

1. Do pre-trained causal language models recognize semantic linguistic concepts such as synonymy ?

Research questions

1. Do pre-trained causal language models recognize semantic linguistic concepts such as synonymy ?
2. What type of test would be the most appropriate ?
   Does a single type of test suffice ?
   Should the test be constructed automatically or by hand ?
   Should the test be validated by a human ?

Research questions

1. Do pre-trained causal language models recognize semantic linguistic concepts such as synonymy ?
2. What type of test would be the most appropriate ?
   Does a single type of test suffice ?
   Should the test be constructed automatically or by hand ?
   Should the test be validated by a human ?
3. How is the LM's performance correlated with its characteristics ? (what enters into play in this performance)
   Size of training data
   Content of training data
   Tokenisation
   Model architecture

Tests - preview

1. Substitution-based

word	**example obtained**	**metric**
character	She plays the **character** of the factory worker.	pp1
role	She plays the role of the factory worker.	pp2
quality	She plays the quality of the factory worker.	pp3

Metric: $pp_2 < pp_3$?

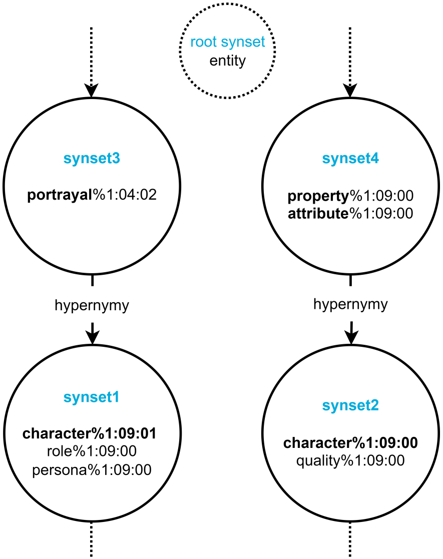
Data

- **SemCor**
  She plays the character lexsn=”1:09:01” of the factory worker.
- **WordNet**
  character:
  ”1:09:00” quality, lineament (a characteristic property)
  ”1:07:01” fiber, fibre (the complex of attributes that determines a persons morals)
  ”1:09:01” role, theatrical role, part, persona (an actor’s portrayal of someone in a play)

WordNet

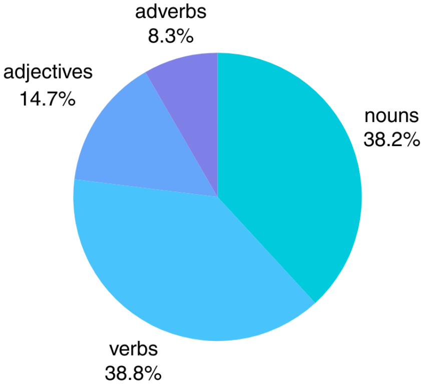
Underlying concepts:

- **synset**
  set of synonyms sharing a sense
- **lexical relations**
  synsets are linked using relations such as hypernymy



SemCor

documents	sentences	words total	words annotated
352	37 176	778 587	229 517



It should work, but . . .

target - **synonym** - **other**	**example**
power - powerfulness - ability	Constitutional government, [. . .] and the veto **power** in world councils are but a few examples.
bit - spot - moment	But I'm not one damned **bit** sorry I went out to question the people I know . . .
rabbit - coney - hare	We come upon a **rabbit** that has been caught in one of the brutal traps in common use.
amount - sum - quantity	Multiply the result obtained in item 3 above by the **amount** used for each State in item 1 above.

- the example is ambiguous (4)
- substitution does not work (1) (2)
- WordNet accuracy error (3)

Evaluator	**Accuracy (185 total)**
Annot 1	76,76%
Annot 2	75,14%
OLMo-1B	76,22%
OLMo-7B	70,81%
Amber	73,51%

Manual annotation

Manual choice

character:

"1:09:00" quality, lineament (a characteristic property)

"1:07:01" fiber, fibre (the complex of attributes that determines a persons morals)

"1:09:01" role, theatrical role, part, persona (an actor's portrayal of someone in a play)

Result

target	synonym	other	example
character	role	quality	She plays the **character** of the factory worker.

Human feedback

Three instances of human intervention or feedback :

1. WordNet and SemCor
2. hand-picking the triples (word, synonym, other)
3. human evaluation of final dataset

Measure	**Annot 1**	**Annot 2 (native)**	Agree
Weird item	28/149 (18.8%)	31/200 (15.5%)	6/150
Accuracy	134/149 (89.9%)	153/169 (90.5%)	107/150 (71.3%)

Human feedback

Can the scores between LMs and humans be compared? NO

Humans	**LMs**
see all three sentences	see one sentence at a time
choose the 2nd or 3rd	no choice, give probability
might use reference sentence	does not see reference
target word highlighted	no highlights

Can a test be designed to be applicable to both? MAYBE

- we cannot be sure what heuristics humans use to perform the test
- prompting or acceptability judgements for humans ?

Methodology

1. perplexity

$$PP(t_1, \ldots, t_n) = exp\left(-\frac{1}{t}\sum_{1}^{t}\log p_\theta(t_i|t_{<i})\right),$$

   where $t_1, \ldots, t_n$ is a sequence of n tokens and θ represents our model

   - the same as the model was trained
   - can be applied to both pre-trained only and fine-tuned models

2. minimal pairs
   - preserve length (excl. tokenisation)

Tests summary

Multiple aspects of *synonymy*:
- property of having the same meaning (sharing the same contexts)

1. (implicit) substitute one word by another in a context
2. (implicit) reference one word with another to avoid repetition
3. (explicit) a relation between two words named 'synonymy'

test type	**substitution**	**reference**	**relation**
1	X		
2			X
3		X	X
4		X	

Tests

1. Substitution-based

word	example obtained	metric
character	She plays the **character** of the factory worker.	pp1
role	She plays the role of the factory worker.	pp2
quality	She plays the quality of the factory worker.	pp3

Metric: $pp_2 < pp_3$?

Tests

2 Explicit relation

word	**relation**	**metric**
role	**Character** is a synonym of role.	pp1
quality	**Character** is a synonym of quality.	pp2
random1	**Character** is a synonym of random1.	pp3
. . .		
random10	**Character** is a synonym of random10.	pp12

Metric: $pp_1 < min(pp_3, \ldots, pp_{12})$ AND $pp_2 < min(pp_3, \ldots, pp_{12})$?

Tests

3. Explicit relation and context

word	relation	metric
role	She plays the **character** of the factory worker. **Character** is a synonym of role.	pp1
quality	She plays the **character** of the factory worker. **Character** is a synonym of quality.	pp2

Metric: $pp_1 < pp_2$?

Tests

4. Reference and context

word	**relation**	**metric**
role	She plays the **character** of the factory worker. This role	pp1
quality	She plays the **character** of the factory worker. This quality	pp2

Metric: $pp_1 < pp_2$?

Tests variations

Test of the **synonymy relation** using multiple constructions :

type	**relation**
explicit	A is a synonym of B
paraphrase 1	A means B
paraphrase 2	A is the same as B

test	**test type**	**substitution**	**reference**	**syn**	**par1**	**par2**
1	1	X				
2a	2			X		
2b	2				X	
2c	2					X
3a	3		X	X		
3b	3		X		X	
3c	3		X			X
4	4		X			

Limitations and bias

We test (only):

- polysemous words
- nouns
- one negative example (in most tests)
- no compound words

Bias:

- the triples are chosen according to test 1

Models

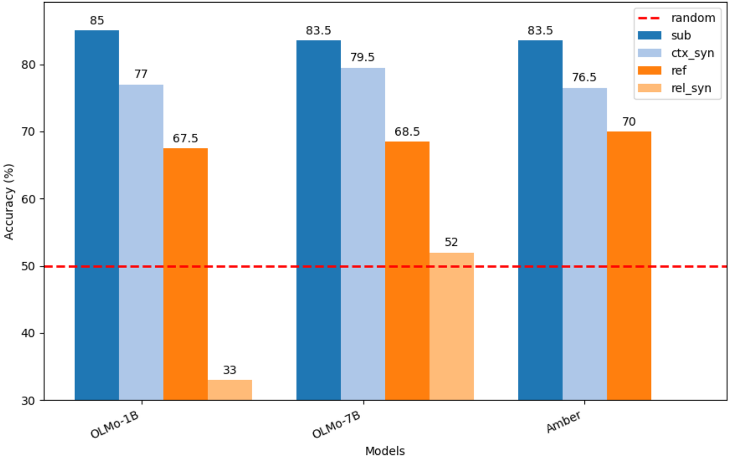
In preparation for the next steps - fully open models :

- training dataset open, accessible
- exact order and content as used in training
- models parameters open (with access to multiple checkpoints)
- monolingual (English)
- research paper present

OLMo (1B, 7B), Amber (7B)

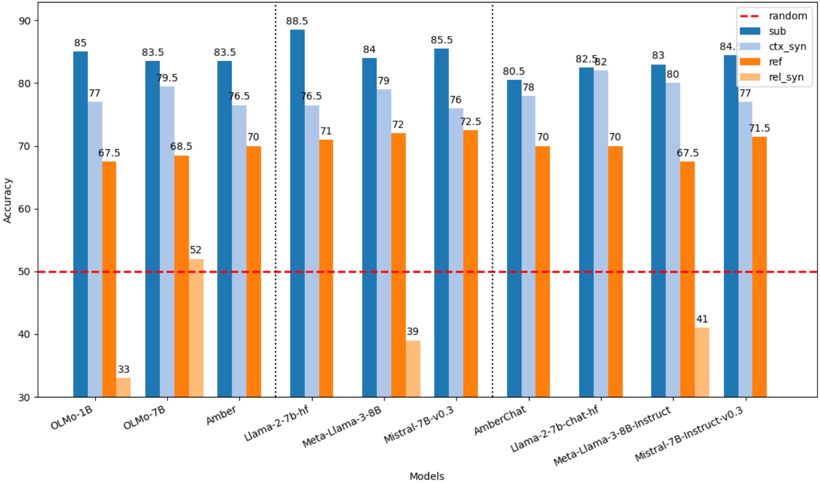
Results

Do LMs recognize the concept of *synonymy* ?
Are multiple tests needed or one suffice ?



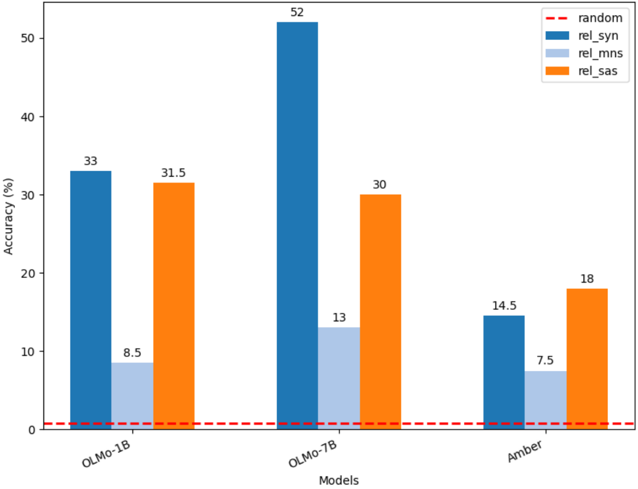
Results

How do the selected LMs fare when compared to their competitors ?



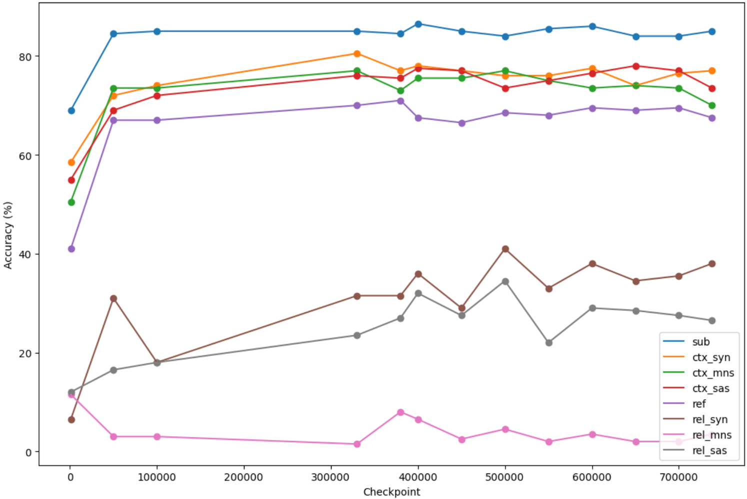
Results

How do LMs best represent the synonymy relation ?
Explicit or paraphrase ?



Results

How much data is needed to learn synonyms ?



Conclusions so far

1. Relation tests seem difficult for language models
   Soil is a synonym of emancipation.
   Illusion is a synonym of duplicity.
   Idea is a synonym of tyrannosaurus.
2. LMs have good performance in binary tests (substitution, reference) that include **context** - as expected
3. The preferred formulation for the relation between the three seems to be *is a synonym of*
4. Curated tests improve the accuracy in humans and language models by 10% - 15%
5. LMs seem to attain peak accuracy in some tests after (only) 50000 checkpoints (200 billion tokens)

Future work

(Near)

1. compute correlation / statistical significance between the tests
2. inspect more closely the learning curve on the first 50000 checkpoints
3. inspect frequency bias in training data
4. inspect the role of tokenisation
5. random candidates for the other tests as well

(Less near)

Is it legitimate to expect an LM to be coherent ? (it does a good job without)

Why is it not coherent (from our experiments) ? Different training ?

Can we modify the data to make it more coherent ? Can we modify / analyze the fine-tuning already present in the data ?

On Synonymy and Language Models

Ioana Ivan

Encadrants: Alexis Nasr et Carlos Ramisch

June 19, 2024