Probing for Referential Information in Language Models

Ionut-Teodor Sorodoc [1] Kristina Gulordava Gemma Boleda[1 2]

[1]Universitat Pompeu Fabra [2]ICREA

Introduction

Background

Language models: computational models that learn to predict the next word considering the past context.

These models develop representations that encode different types of linguistic information.

Introduction

Background

Language models: computational models that learn to predict the next word considering the past context.

These models develop representations that encode different types of linguistic information.

Goal

We want to understand to what extent they capture referential information.

Mary went with John. She → She / She

Hypothesis

Language models encode:

- ☺ **grammatical** properties of anaphora
- ☹ **semantico-referential** information

Mary went with John. She → She / She

Hypothesis

Language models encode:

- ☺ **grammatical** properties of anaphora
- ☹ **semantico-referential** information

Method

Probe model: small classifier to predict a feature of interest, in this case anaphoric coreference, given the model's hidden representations as input.

Resources

Pretrained Models

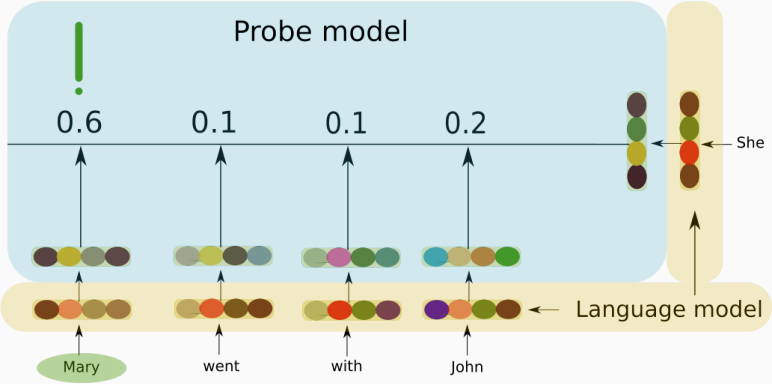
- AWD-LSTM
- Transformer-XL

Both models are trained on the same corpus with a comparable vocabulary.

Dataset: Ontonotes

he$_1$ was elected to be president of the People's Republic of China, and chairman of the Central Military Commission$_2$. Yeping Wang$_3$ was born in Shanghai in 1926. She$_3$ studied in Shanghai Foreign Language College, and started working in 1949. For a long time, she$_3$

Model



- Task: predict which element the target (**she**) refers to
- Correct if it points to any element in the coreference chain

Results

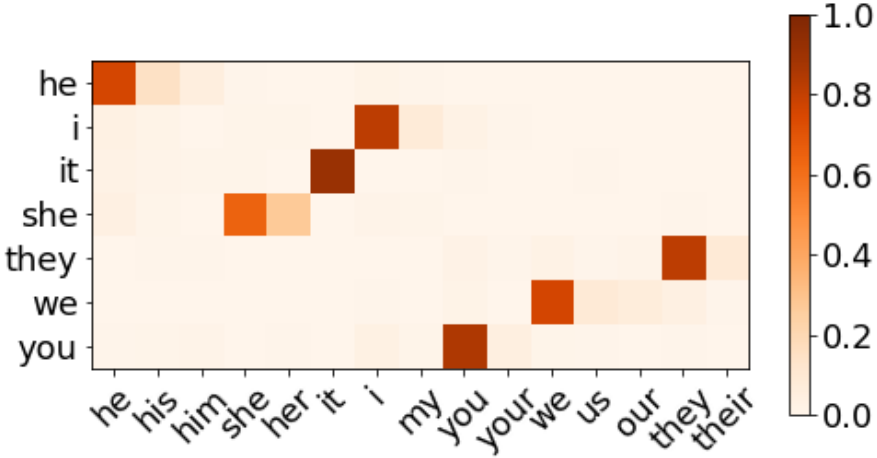
Model	Accuracy
closest gold entity	56.1
closest same-form token	61.3
LSTM	64.8
Transformer	**75.9**

Morphosyntactic factors: Noun phrases

Syntax

- Pronouns refer to nominal elements
  - Refers to element in a chain: 92.6%
  - Even when it doesn't, refers to a nominal element 82% of the time
- Pronouns agree in gender and number
- Across sentences

Morphosyntactic factors: Gender tendencies



Current findings

Findings
Language models encode:

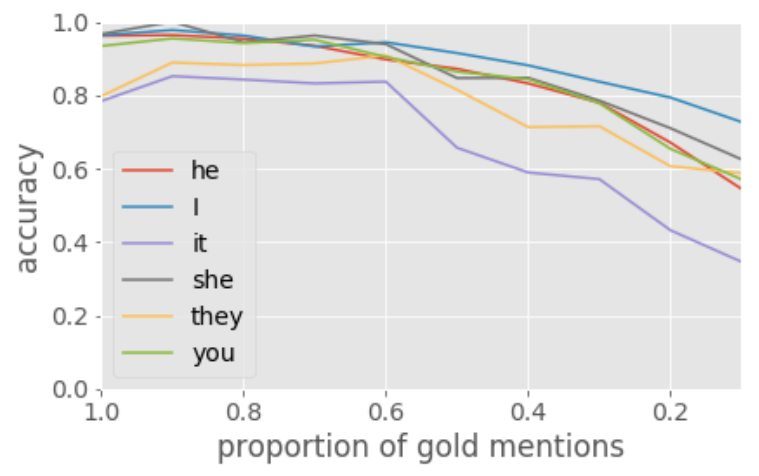
- ☺ **grammatical** properties of anaphora

Current findings

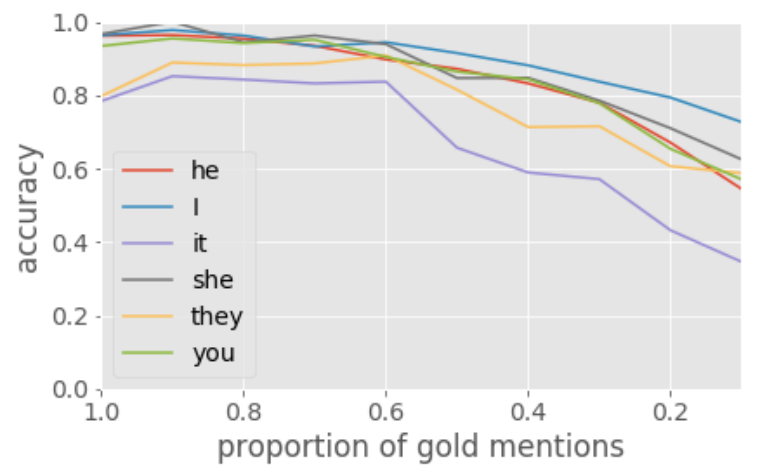
Findings
Language models encode:

- ☺ **grammatical** properties of anaphora
- ? **semantico-referential** information

Semantic factors: Distractors



Semantic factors: Distractors



DISTRACTORS CONFUSE THE MODEL, BUT THEY DO NOT FOOL IT COMPLETELY

Semantic features: Distractor types

Type	T Acc.
No distractor	**81.8**
Distractor(s)	73.8
= number	65.3
= gender	48.6
= pron.	49.1

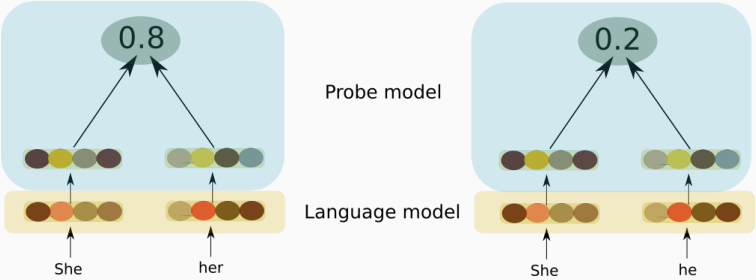
Accuracy of the model with different types of distractors

Semantic features: Distractor types

Type	T Acc.	Baseline
No distractor	81.8	100
Distractor(s)	73.8	32.0
= number	65.3	26.6
= gender	48.6	15.7
= pron.	49.1	20.3

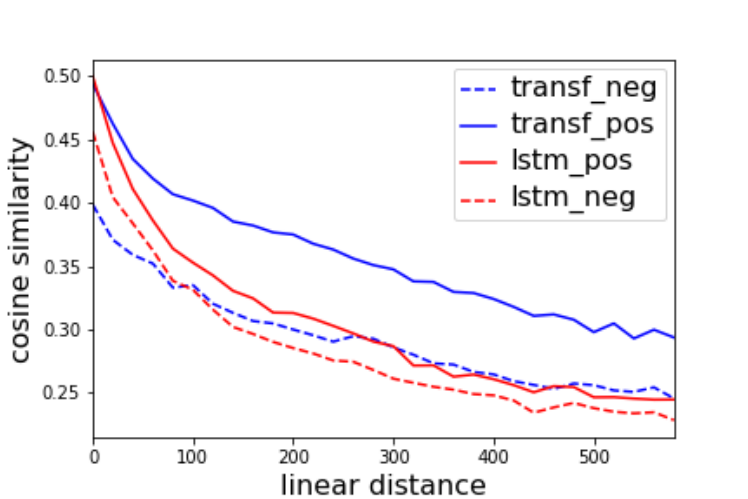
Accuracy of the model with different types of distractors

Document level: Model



- Take all pronominal mentions in a document
- Task: distinguish those that refer to the same entity from those that refer to different entities

Semantic features: Linear distance vs cosine distance



Conclusions

Hypothesis

Language models encode:

☺ **grammatical** properties of anaphora

- Gender
- Number
- Part of speech

😐 **semantico-referential** information

- Confusion when there are other mentions in the context
- But limited ability to distinguish mentions that have the same form but are in different chains

Probing for Referential Information in Language Models

Ionut-Teodor Sorodoc [1] Kristina Gulordava Gemma Boleda[1 2]

[1]Universitat Pompeu Fabra [2]ICREA