Syntax and Semantics Meet in the "Middle":
Probing the Syntax-Semantics Interface of LMs through Agentivity

NEULAB

Carnegie Mellon University
Language Technologies Institute

Lindia Tjuatja, Emmy Liu, Lori Levin, Graham Neubig
lindiat@andrew.cmu.edu

PDF

How do LMs deal with interactions between word-level meaning and the meaning of larger syntactic structures?

To answer this question, we focus on the linguistic concept of **agentivity** as a case study. We utilize the **unique properties of some optionally transitive English verbs** to create our test suite.

Q1: Do models display sensitivity to agentivity at the lexical level?

Is an entity typically more agent-like or patient-like?

Q2: Can models use word-level meaning to disambiguate the role of a noun?

Same surface form structure, different semantic roles

This **car** sells easily.

This **salesman** sells easily.

Q3: Can models disregard word-level priors in appropriate contexts?

Deterministic form-to-meaning mapping: subject = agent, object = patient

This **car** sells something easily.

Something sells this **salesman** easily.

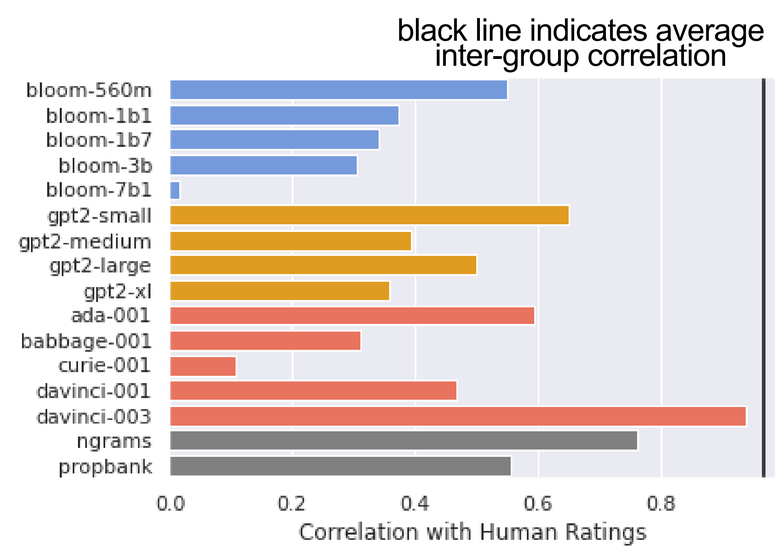
Exp 1: Agentivity at the lexical level

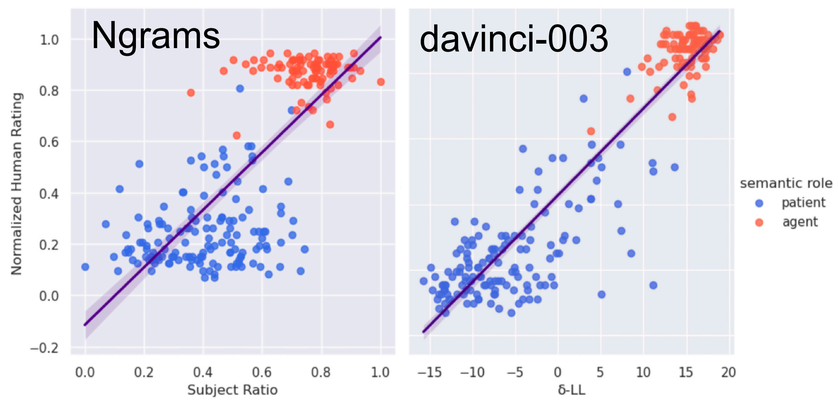
Exp 1: noun (lexical level)

```
noun: John
agent/patient: agent

noun: vase
agent/patient: patient

noun: nurse
agent/patient: agent

noun: mango
agent/patient: patient

noun: <noun>
agent/patient:
```

We calculate **delta in probability** of predicting "agent" and "patient" and take the **correlation** between delta and **human judgements** for how typically agent-like a noun is.

We also compare with **corpus statistics** from **Google Syntactic Ngrams** and **Propbank**.





Out of all models tested, **GPT-3 davinci-003 is best correlated with human ratings**, nearing the value of average inter-group correlation.

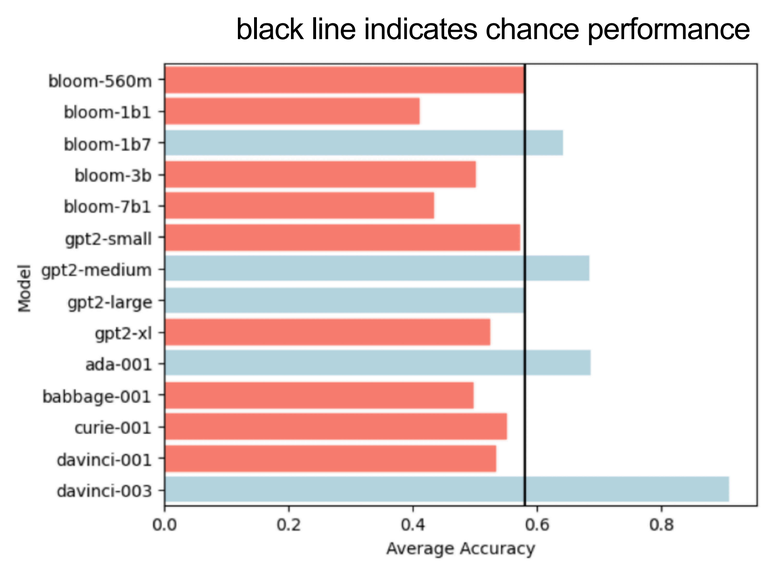
Furthermore, it is **better correlated with humans than frequency statistics** from both syntactic and semantic annotated corpora.

Exp 2 : Disambiguating intransitives

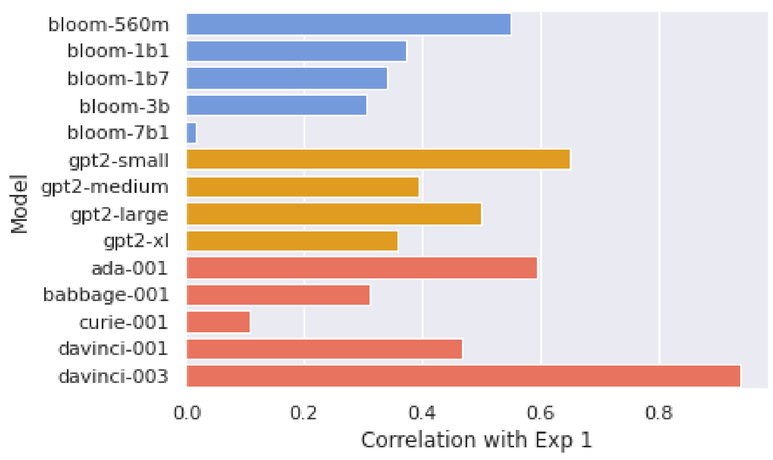
Exp 2: intransitive (ambiguous mapping)

```
Sentence: John walks quickly.
Is John an agent or a patient?: agent

Sentence: This vase breaks easily.
Is vase an agent or a patient?: patient

Sentence: This nurse works swiftly.
Is nurse an agent or a patient?: agent

Sentence: This mango blends well.
Is mango an agent or a patient?: patient

Sentence: <intr-agent/intr-patient>
Is <noun> an agent or a patient?:
```

- Can models correctly predict the role in an ambiguous form-to-meaning mapping?
- Are the predictions from this experiment correlated with Exp 1?



Most models are **unable to accurately predict** the role of the noun in this setting. However, **davinci-003 is able to do so with high accuracy** and demonstrates sensitivity to agentivity in context that is **well-correlated with agentivity of the noun in isolation**.

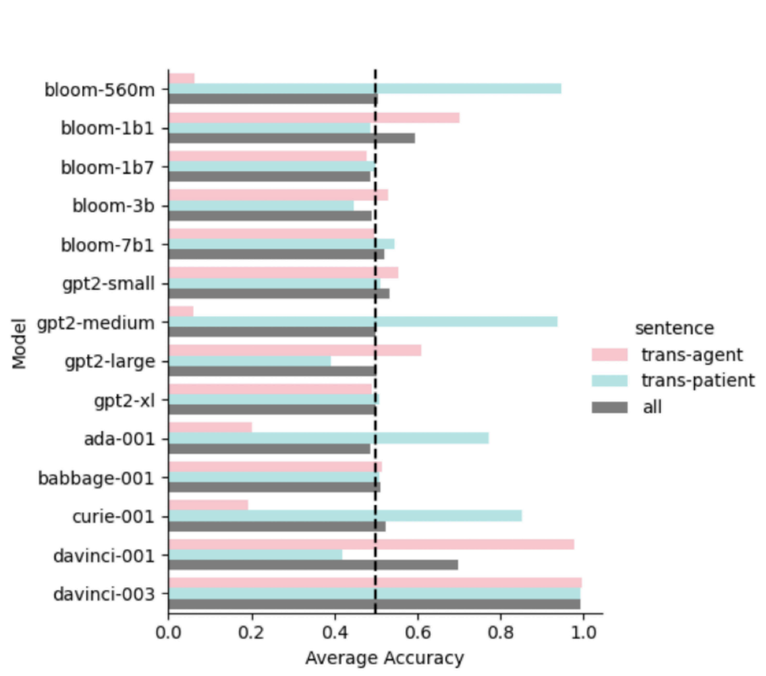


Exp 3: Overriding priors with transitives

Exp 3: transitive (deterministic mapping)

```
Sentence: Jack throws something easily.
Is Jack an agent or a patient?: agent

Sentence: Something hires the nurse swiftly.
Is nurse an agent or a patient?: patient

Sentence: The hammer breaks something quickly.
Is hammer an agent or a patient?: agent

Sentence: Something blends the mango well.
Is mango an agent or a patient?: patient

Sentence: <trans-agent>/<trans-patient>
Is <noun> an agent or a patient?:
```

Can models correctly predict the role in a deterministic form-to-meaning mapping?



Like Exps 1 and 2, **davinci-003 outperforms all other models by far**. Despite the form-to-meaning mapping being deterministic, most models are unable to pick up this pattern.

A closer look at the performance of davinci-003 on examples in Exp 2 shows:

- davinci-003 does worse on **nouns with a "patient" label that are more agent-like**
- Though animate nouns are the most agent-like to humans, **vehicles also act as "pseudo-animates"**
- Examples with vehicles comprise the **most frequent errors**

Overall, we find that:

- GPT-3 davinci-003 performs much better than other models and is better correlated with humans than corpus statistics
- There is no monotonic increase in performance with size
- Many of the examples davinci-003 gets incorrect involve a number of linguistic confounders that make them more ambiguous to humans as well