Adjudicating LLMs as PropBank Annotators

Julia Bonn
University of Colorado Boulder

Harish Tayyar Madabushi
University of Bath

Jena D. Hwang
Allen Institute for AI

Claire Bonial
DEVCOM Army Research Lab

LLMs

- Larger models may be able to sort short sentences by semantic similarity based on constructional semantics (*she blinked the tears off her eyelashes* & *She wiped the flour off the table)*, but smaller models sort by lexical semantics (Li et al., 2022).
- Even the largest models are unable to recognize the semantic similarity of events in argument structure constructions (*He yelled himself hoarse*) (Bonial & Tayyar Madabushi, 2024).
- Can readily achieve surface-level semantic analysis such as locating the main predicate and its core arguments (*who-did-what-to-whom*), but struggle with more complex analysis, such as AMR, even with in-context examples (Ettinger at al., 2023).

Can GPT-3.5 and GPT-4, which excel in language generative capabilities, produce viable PropBank annotation?

We assess the meta-linguistic capabilities of LLMs to annotate **PropBank** information by:

1. Designing three prompts relating to ***transitive***, ***intransitive***, and ***middle voice*** constructions
2. Dissecting LLMs' abilities with respect to the PropBank tasks of ***argument annotation*** and ***roleset annotation***.

Evaluation set includes 35 sentences with 7 verbs from 7 VerbNet classes.

PropBank

- A powerful resource that provides simple, explicit mappings between particular syntactic patterns of argument expression and the semantic roles of those arguments, enabling a shallow semantic analysis facilitated by clearly recognizable syntactic patterns.
- Human annotators able to make judgments easily and consistently. High IAA in the upper 80%s: exact match = 84.4%, core-arg match = 88.3% (Bonial et al., 2017).
- If there's any semantic annotation that GPT might be able to accomplish, it seems like PropBank would be it.

Training Conditions

Zero-Shot

Harder than human annotation:

GPT asked to provide sense and role annotation.

Given the following verb and sentence, produce PropBank annotations of the verb sense and its arguments. Limit your annotation to the sentence provided.

Annotate this:
Sentence: *Ty poured some syrup in his pan and got out a piece of bread.*
Verb: *pour*

3-Shot

Similar to human annotation:

GPT asked to provide sense and role annotation, and given examples in **transitive**, **intransitive**, and **middle voice**.

+

Example 1:
Sentence: ***They*** *went to India and Nepal, stayed in hostels and* ***hiked*** ***mountains.***

Verb: hike
Sense: hike.01 (walk for pleasure or exercise)
Arguments:
Arg0: *They*
Rel: *hiked*
Arg1: *mountains*

Example 2:
Sentence: ***Connor Kobal hikes*** *regularly in Boulder Mountain Park.*

Verb: hike
Sense: hike.01 (walk for pleasure or exercise)
Arguments:
Arg0: *Connor Kobal*
Rel: *hikes*
ArgM-TMP: *regularly*
ArgM-LOC: *in Boulder Mountain Park*

Example 3:
Sentence: ***This trail hikes*** *through a portion of the historic area and then up to a ridge overlooking Stone Valley.*

Verb: hike
Sense: hike.01 (walk for pleasure or exercise)
Arguments:
Arg1: *This trail*
Rel: *hikes*
ArgM-DIR: *through a portion of the historic area and then up to a ridge overlooking Stone Valley.*

3-Shot+Roleset

Easier than human annotation:

GPT given sense, examples, and asked to provide role annotation.

+

Use the roleset information provided to produce the annotation.

Roleset:
Arg0: Causer of motion
Arg1: path of motion; location

Gold-Standard

Sentence: *Lightweight and compact, the Inka pen writes upside down, underwater and at any altitude.*

Verb: *write*

Verb Sense: write.01, *set pen to paper*
Arg0: writer
Arg1: thing written
Arg2: benefactive

Annotations:
ArgM-PRD: *Lightweight and compact*
Arg0: *The Inka pen*
ArgM-MNR: *upside down, underwater and at any altitude*

Evaluation Metric

Exact Match:

Annotations:
Verb Sense: write.01

ArgM-PRD: *Lightweight and compact*
Arg0: *The Inka pen*
ArgM-MNR: *upside down, underwater and at any altitude*

Perfect match to gold for verb sense, numbered arguments, and ArgMs, including function tags like 'MNR', as well as text span.

Core-Arg Match:

Annotations:
Verb Sense: write.01

ArgM-ADV: *Lightweight and compact*
Arg0: *The Inka pen*
ArgM-LOC: *upside down, underwater and at any altitude*

All the arguments annotated in gold are identified by the model. Match for verb sense and numbered args (full text span). ArgMs are correctly identified as 'ArgM', but one or more function tags do not match.

Number-Arg Match:

Annotations:
Verb Sense: write.01

-
Arg0: [*the … pen*]
ArgM-LOC: [*underwater*]

All numbered arguments are correctly identified by model, but not necessarily with the correct span of text. ArgMs may or may not be identified.

No Match:

Annotations:
Verb Sense: write.02

ArgM-PRD: *Lightweight and compact*
Arg1: *The Inka pen*
ArgM-MNR: *upside down, underwater and at any altitude*

Verb sense is incorrect, or incorrect numbered arguments are assigned.

Results

Model	Setting	Match Types: Exact	Core-Arg	Num-Arg
GPT-3.5	0-shot	8.6%	8.6%	17.1%
	3-shot	11.4%	17.1%	37.1%
	3-shot+rs	2.9%	2.9%	20.0%
GPT-4	0-shot	8.6%	17.1%	34.3%
	3-shot	14.3%	20.0%	42.9%
	3-shot+rs	**22.9%**	**22.9%**	**48.6%**

Table 1: positive matches for GPT-3.5 and GPT-4 over three prompt settings: 0-shot, 3-shot, and 3-shot with roleset (3-shot+rs).

Construction	N	Match Types: Exact	Core-Arg	Num-Arg
Transitive	14	50.0%	50.0%	85.7%
Intransitive	13	7.7%	7.7%	23.1%
Middle	8	12.5%	12.5%	25.0%

Table 2: percentage of positive matches for the **best-performing prompt and model** combination: **GPT-4** with the **3-shot+roleset** prompt. *N* refers to the number of instances available to each construction.

Even the best-performing model & prompt fail miserably to identify the correct semantic roles for intransitives and middle voice constructions.

Conclusions

- Zero-shot knowledge of PropBank annotation is almost nonexistent.
- The largest model evaluated, GPT-4, achieves the best performance in the setting where it is given both examples and the correct roleset in the prompt, demonstrating that larger models can ascertain some meta-linguistic capabilities through in-context learning.
- However, even in this setting, which is simpler than the task of a human in PropBank annotation, the model achieves only 48% accuracy in marking numbered arguments correctly.
- GPT's relatively poor performances stems from its apparent inability to generalize semantics across various syntactic realizations.

GPT-3.5 and GPT-4 do not make good PropBank annotators, failing at the task of identifying* who-did-what-to-whom, *especially in intransitive and middle voice constructions. Valuable semantic resources like PropBank still need human annotation!