Shallowly accurate but deeply confused – how language models deal with antonyms

Research question

- Antonyms are words that are semantically opposite of each other. We focus here on **antonymic adjectives**.
- Statistical models of language have previously been argued to give a poor treatment of antonyms, because:
  - they are based on the Distributional Hypothesis [1];
  - antonyms appear in similar environments [2, 4].
- Recent Large Language Models (**LLMs**) yet perform well in tasks which most probably require them to encode adjective polarity... **so, can they draw fine-grained inferences about antonyms?**

Background: adjective polarity & negation

The Inference Towards the Antonym (ITA) [3, 7, 8, 10]

(*not* A) $\Longrightarrow$ A', where A' is the antonym of A

ITA Pragmatic Mitigation Condition [7]

(*not* A) $\not\Longrightarrow$ A', if Cplx(*not* A) » Cplx(A')

Negative Adjectives Complexity Hypothesis [6, 5]

$\forall A^-.\ A^- = \text{Not-}A^+$, therefore:

♦ $\text{Cplx}(A^-) = \text{Cplx}(\text{Not-}A^+) \sim \text{Cplx}(\textit{not}\ A^+)$

⋆ $\text{Cplx}(\textit{not}\ A^-) = \text{Cplx}(\textit{not}\ \text{Not-}A^+) \gg \text{Cplx}(\text{-}A^+)$

- Based on Eq. ♦ and ⋆, Krifka [7] concludes that **inferring a negative adjective A^- from *not* A^+ is easier than inferring A^+ from *not* A^-.**
- As argued by Ruytenbeek et al. [8], this can be assessed empirically by comparing the felicity of (1a) vs. (1b).

(1) a. He is **not tall**$_{(A^+)}$. She <u>too</u> is **short**$_{(A^-)}$.

b. # He is **not short**$_{(A^-)}$. She <u>too</u> is **tall**$_{(A^+)}$.

- They also claim that the contrast between (1a) and (1b) should be smaller for morphologically opaque antonyms (**O-antonyms**), as opposed to morphologically transparent ones (**T-antonyms**).

Goal

- Building on the studies conducted by [8], we test if 3 recent LLMs (GPT-2 [13], XLNET [14] and BERT [9]):
  1. are more likely to "draw" an ITA when the adjective under negation is positive rather than negative (**H1**);
  2. show a higher discrepancy in ITA strength for T-antonyms as opposed to O-antonyms (**H2**).

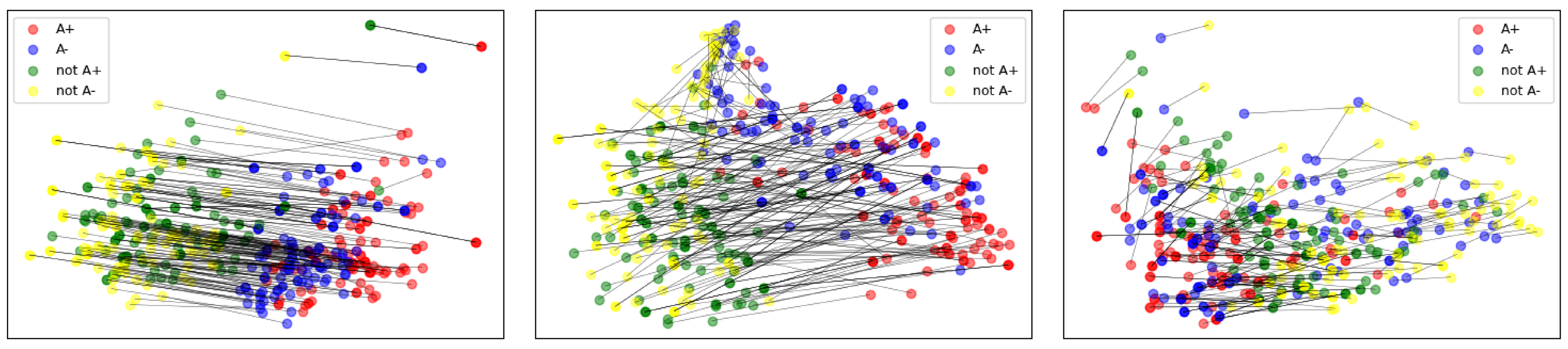
< .001
< .01
< .05
< .1
n.s.

	Task 1		Task 2	
	H1	H2	H1	H2
GPT-2	.14 (S) T+O	n.s.	.59 (L) T+O	n.s.
XLNet	.03 (N) T+O	.37 (M)	.42 (M) T+O	n.s.
BERT	.15 (S) T	.14 (S)	n.s.	n.s.

Table 1: Test results for Tasks 1 and 2.[1]

	H1	H2
GPT-2	.19 (S) T+O	.30 (S)
XLNet	.35 (M) T+O	.17 (S)
BERT	.22 (S) T+O	.31 (S)

Table 2: Test results for Task 3.



(a) GPT-2 (b) XLNet (c) BERT

Figure 1: 2D-PCA-reduced embeddings of the 3 models. Lines represent the effect of negation for any given adjective.

Task 1: Sentence surprisal ($\mathcal{S}_{sent}$)

Hypothesis 1: $\Delta\mathcal{S}_{sent} \triangleq \mathcal{S}_{sent}(\mathbf{1b}) - \mathcal{S}_{sent}(\mathbf{1a}) > 0$

- 111 pairs of antonyms (48 T, 63 O) tested using a one-sided paired-sample Wilcoxon test.
- 3/3 **LLMs captured the contrast** but with small effect sizes.
- Group-by-group testing (T vs. O) reveals that only for GPT-2 and BERT, *both* groups verify H1 (after corrections).

Hypothesis 2: $(\Delta\mathcal{S}_{sent})_{\text{O-group}} < (\Delta\mathcal{S}_{sent})_{\text{T-group}}$

- We compared the T- and O-groups using a Mann Whitney U-Test on the surprisal contrasts between (1a) and (1b).
- 2/3 **LLMs captured the effect of transparency**.

Task 2: Target-word surprisal ($\mathcal{S}_w$)

H1: $\Delta\mathcal{S}_A \triangleq (\mathcal{S}_{A^+}(\mathbf{1b}) - \mathcal{S}_{A^+}(\mathbf{2b})) - (\mathcal{S}_{A^-}(\mathbf{1a}) - \mathcal{S}_{A^-}(\mathbf{2a})) > 0$

- We compared the individual surprisals associated with the second adjective in (1a)/(1b).
- Because LLMs tend to assign morphologically complex words higher suprisals, we used (2a)/(2b) as baselines.

(2) a. He is **not tall**$_{(A^+)}$. She is **short**$_{(A^-)}$.

b. # He is **not short**$_{(A^-)}$. She is **tall**$_{(A^+)}$.

- 2/3 **LLMs captured the contrast.** For both models, individual groups (T/O) were also linked to a significant effect after corrections.

H2: $(\Delta\mathcal{S}_A)_{\text{O-group}} < (\Delta\mathcal{S}_A)_{\text{T-group}}$

- ***p*-values for H2 in this task were *not* significant**, i.e., no effect of morphological transparency was detected...

Task 3: Comparison in word embeddings

- We focus on the representation that LLMs assign to A^+, A^-, and their negations: $\overrightarrow{A^+}$, $\overrightarrow{A^-}$, $\overrightarrow{\textit{not } A^+}$, $\overrightarrow{\textit{not } A^-}$.
- Semantic closeness between those vectors is measured by cosine similarity (=angle between 2 vectors).

H1: $\Delta\text{Cos} \triangleq \text{Cos}(\overrightarrow{A^-}, \overrightarrow{\textit{not } A^+}) - \text{Cos}(\overrightarrow{A^+}, \overrightarrow{\textit{not } A^-}) > 0$

- We test if $\overrightarrow{\textit{not } A^+}$ is "closer" to $\overrightarrow{A^-}$ than $\overrightarrow{\textit{not } A^-}$ is to $\overrightarrow{A^+}$.
- 3/3 **embeddings captured the contrast**, suggesting that ITA strength translates into topological distance.

H2: $(\Delta\text{Cos})_{\text{O-group}} < (\Delta\text{Cos})_{\text{T-group}}$

- 3/3 **embeddings captured the effect of transparency**, although the effect sizes were small...

Discussion & outlook

- LLMs seem to distinguish positive and negative adjectives w.r.t their semantic closeness to their antonym (H1) and somewhat differentiate between T- and O-antonyms (H2).
- The relative weakness of this last result may be due to LLMs' tokenization strategy not aligning with human-like morphological segmentation.
- This should also be contrasted with two other results:
  - As already noted for older models [11], **LLMs represent *not* A^+ as closer to *not* A^- than to A^-, and vice-versa** (cf. Fig. 1)!
  - A refinement of BERT (RoBERTa-MNLI [12]) fine-tuned for Natural Language Inference was **more likely to conclude (*He not* A^- ⇒ *He is* A^+) than (*He not* A^+ ⇒ *He is* A^-)**, *contra* H1(!)

[1] *p*-values color-coded. Effect sizes are Cliff's Δ. N, S, M, L resp. mean 'negligible', 'small', 'medium', 'large'. Each cell also lists which subgroup(s) (T, O) drive(s) the effect in Task 1.

Adèle Hénot-Mortier

mortier@mit.edu

Massachusetts Institute of Technology

MIT Linguistics