Just Say No

Trials and Tribulations of Teaching Large Language Models to Understand Negation

Tereza Vrabcová
xvrabcov@fi.muni.cz

Faculty of Informatics, Masaryk University

November 7, 2024

Contents

- What *is* negation?
- Areas of interest
- Large language models vs Negation
  - Problems
  - Common approaches
  - More problems
- My research plan

Do you use negation in your day to day life?

Do you use negation in your day to day life?

trick question

Better question:
How often do you use negation in your day to day life?

What Is Negation?

Negation in Human Communication

- key component of human communication
- one of the basic tools for understanding the world
- philosophy: core part of individualism
  - creating bounds between self and the rest of the world
  - early in human development, 2 – 3 years of age
    - colloquially known as The Terrible Twos

Areas of Interest

Natural Language Processing

- as a part of natural language, we want to process it
- negation is a non-trivial problem
- number of ways to express negation
    - more than no, not, and n’t
    - negative adverbs – never, neither, rarely, barely
    - negative pronouns – nothing, none, nowhere
    - double negatives
    - partial vs total negation
        - She didn’t do it out of love.
        - She didn’t do it, out of love.
    - grammatical vs lexical negation

Areas of Interest

Large Language Models (LLMs)

- as a part of natural language, we want to use it to communicate
- its non-triviality causes problems for LLMs
- what problems?
    - let’s take a look

LLMs vs Negation

Loves Me, Loves Me Not I

- accuracy of LLM on cloze task

Allyson Ettinger (2019) [2]

A sparrow is a ___.
A sparrow is not a ___.

Nora Kassner, Hinrich Schutze (2020) [5]

Birds can ___.
Birds cannot ___.

Thinh Hung Truong et al. (2023) [7]

Paracetamol isn’t a kind of ___.

LLMs vs Negation

Loves Me, Loves Me Not II

- lack of accuracy of LLM on cloze task

Allyson Ettinger (2019) [2]

A sparrow is a *bird*.
A sparrow is not a *bird*.

Nora Kassner, Hinrich Schutze (2020) [5]

Birds can *fly*.
Birds cannot *fly*.

Thinh Hung Truong et al. (2023) [7]

Paracetamol isn’t a kind of *medicine*.

LLMs vs Negation

Common Approaches

- method of Reinforcement Learning with Human Feedback (RLHF)
    - used by big companies such as OpenAI and Microsoft [1, 4]
    - problem: data is not open-source, not easily reproducible
- modifying prompts
    - replacing words with antonyms [6]
    - adding more negation [3]
        - prepending the negative version of the prompt

LLMs vs Negation

More Problems

- LLM does not reason well with negation
- problem of misinformation, hallucinations
  - the model does not know what is and what is not true
  - further experiments – letting LLM know the prompt can be false [8]

Research Plan

- current methods focus on tackling the problem at the end
  - fine-tuning
  - prompt modification
- my goal:
  - start at the beginning
  - training data
    - different ratios of positive and negatives examples
    - different processing methods to enhance negation tokens
    - modification of the LLM architecture to boost negation
    - possible pathway to enable more complex reasoning in LLMs

Research Plan

- current methods focus on tackling the problem at the end
  - fine-tuning
  - prompt modification
- my goal:
  - start at the beginning
  - training data
    - different ratios of positive and negatives examples
    - different processing methods to enhance negation tokens
    - modification of the LLM architecture to boost negation
    - possible pathway to enable more complex reasoning in LLMs

Thank you for your attention

Bibliography I

[1] David Burch. *OpenAI on Reinforcement Learning With Human Feedback (RLHF)*. Arize AI, May 2023. URL: `https://arize.com/blog/openai-on-rlhf/` (visited on 10/17/2024).

[2] Allyson Ettinger. “What BERT Is Not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models”. In: *Transactions of the Association for Computational Linguistics* 8 (2020). Ed. by Mark Johnson, Brian Roark, and Ani Nenkova, pp. 34–48. DOI: `10.1162/tacl_a_00298`. URL: `https://aclanthology.org/2020.tacl-1.3`.

Bibliography II

[3] Md Mosharaf Hossain and Eduardo Blanco. "Leveraging Affirmative Interpretations from Negation Improves Natural Language Understanding". In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2022, pp. 5833–5847. DOI: `10.18653/v1/2022.emnlp-main.393`. URL: `https://doi.org/10.18653/v1/2022.emnlp-main.393`.

[4] Alyssa Hughes. *Learning from interaction with Microsoft Copilot (web)*. Microsoft Research, Mar. 2024. URL: `https://www.microsoft.com/en-us/research/blog/learning-from-interaction-with-microsoft-copilot-web/` (visited on 10/17/2024).

Bibliography III

[5] Nora Kassner and Hinrich Schütze. “Negated and Misprimed Probes for Pretrained Language Models: Birds Can Talk, But Cannot Fly”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020. DOI: `10.18653/v1/2020.acl-main.698`. URL: `http://dx.doi.org/10.18653/v1/2020.acl-main.698`.

[6] Izunna Okpala et al. “A Semantic Approach to Negation Detection and Word Disambiguation with Natural Language Processing”. In: *Proceedings of the 2022 6th International Conference on Natural Language Processing and Information Retrieval*. Vol. 1. NLPIR 2022. ACM, Dec. 2022, pp. 36–43. DOI: `10.1145/3582768.3582789`. URL: `https://doi.org/10.1145/3582768.3582789`.

Bibliography IV

[7] Thinh Hung Truong et al. “Language models are not naysayers: an analysis of language models on negation benchmarks”. In: *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*. Ed. by Alexis Palmer and Jose Camacho-collados. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 101–114. DOI: `10.18653/v1/2023.starsem-1.10`. URL: `https://aclanthology.org/2023.starsem-1.10`.

[8] Neeraj Varshney et al. *Investigating and Addressing Hallucinations of LLMs in Tasks Involving Negation*. 2024. DOI: `10.48550/ARXIV.2406.05494`. URL: `https://arxiv.org/abs/2406.05494`.

MUNI
FACULTY
OF INFORMATICS