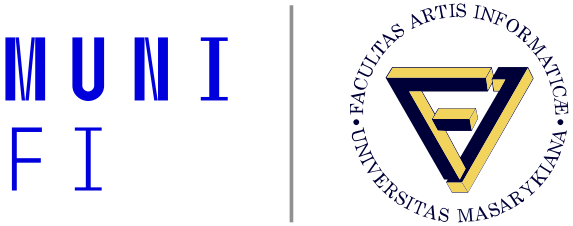

Negation Disrupts Compositionality in Language Models

The Czech Usecase

Tereza Vrabcová, Petr Sojka
{xvrabcov, sojka}@mail.muni.cz

Faculty of Informatics, Masaryk University

December 6, 2024

Motivation

To be, or not to be,
that is the question.

Language Models

- Text is represented as tokens
  - On the level of words / subwords
- Estimates the probability of the token in the output sequence
- In Slavic languages: negation often as token ne
- Question: Does the token carry the meaning of negation sufficiently?

Behaviour of LMs

Brief Look At English

- English is one of the most well studied languages in NLP [3]
- Truong et al. (2023) – analysis of language models on negation benchmarks [5]
  - Shows problems LMs have dealing with negation
  - LLMs ability to reason under negation worsens
  - Example of a cloze task with negation

Iburofen is a type of _____.
Iburofen is not a type of _____.

Behaviour of LMs

Brief Look At English

- English is one of the most well studied languages in NLP [3]
- Truong et al. (2023) – analysis of language models on negation benchmarks [5]
  - Shows problems LMs have dealing with negation
  - LLMs ability to reason under negation worsens
  - Example of a cloze task with negation

Iburofen is a type of medicine.
Iburofen is not a type of _____.

Behaviour of LMs

Brief Look At English

- English is one of the most well studied languages in NLP [3]
- Truong et al. (2023) – analysis of language models on negation benchmarks [5]
  - Shows problems LMs have dealing with negation
  - LLMs ability to reason under negation worsens
  - Example of a cloze task with negation

Iburofen is a type of medicine.
Iburofen is not a type of medicine.

Behaviour of LMs

Slavic Languages

- Fusional language family
  - Utilises affixes to denote syntactic / semantic features
  - Looser word order than in English

Behaviour of LMs

Slavic Languages

- Fusional language family
  - Utilises affixes to denote syntactic / semantic features
  - Looser word order than in English

Máma mele maso.
Máma mele maso?
Mele máma maso?
Maso mele máma.

Methodology

NLI Task

- Natural Language Inference (Textual Entailment)
- Given a premise and a hypothesis, the model evaluates the truth of the hypothesis based on the premise
- The hypothesis can entail / contradict / be neutral


```
{
 "premise": "Robert J. 'Rob' O'Neill
 (narozen 10. dubna 1976) je bývalý námořník
 námořnictva Spojených států amerických.",
 "hypothesis": "Robert J. O'Neill se narodil
 10. dubna 1976.",
 "entailment": "entails"
}
```

Methodology

Instruction LM – Prompting

```
[
  {
     "role": "system",
   "content": "You are a fact checker for queries in the Czech
   language. You will be given a premise, which you know is
   factually correct, and a hypothesis. You will return the
   truth value of the hypothesis, based on the premise.
   Return True if the hypothesis is correct and False if
   the hypothesis is incorrect."
  },
  {
    "role": "user",
    "content": [
      "Premise: Antigua a Barbuda. Jméno zemi dal Kryštof Kolumbus
      v roce 1493 po objevení ostrova na počest Panny Marie La
      Antigua v sevillské katedrále.",
      "Antigua a Barbuda nebyla rodištěm Kryštofa Kolumba."
    ]
  }
]
```

Methodology

Evaluation

- Hypothesis: Inclusion of negation in hypothesis should not hinder the model's ability to correctly identify entailment / contradiction
- Comparison of model's accuracy on hypotheses of positive and negative polarity

Data

Selection

- CsFEVER (Czech Fact Extraction and VERification) [6]
  - Localization of the English FEVER dataset
  - Using the test dataset for evaluation
  - Filtering: removal of neutral hypotheses to simplify it to a binary problem
  - However, majority of claims have positive polarity
- Goal: Creation of parallel hypotheses with opposite polarities
  - Creation of a simple pipeline that negates one verb in the sentence

Data

Negation Pipeline – UDPipe2 Tagger [4]

Máma mele maso.

Data

Negation Pipeline – UDPipe2 Tagger [4]

Máma mele maso.

↓

```
# text = Máma mele maso.
1       Máma        máma        NOUN
2       mele        melit       VERB
3       maso        maso        NOUN
4       .           .           PUNCT
```

Data

Negation Pipeline – `majka` Analyser [2]

mele

↓

mela:k1gFnSc3wH

mela:k1gFnSc6wH

mlít:k5eAaImIp3nSrD

Data

Negation Pipeline – `majka` Analyser [2]

mele

↓

mela:k1gFnSc3wH

mela:k1gFnSc6wH

mlít:k5eAaImIp3nSrD

mlít:k5eAaImIp3nSrD

↓

mlít:k5eNaImIp3nSrD

↓

nemele

Data

Modified CsFEVER


```
{
"id": 10567,
"premise": "Google Search, běžně označovaný jako
Google Web Search nebo jednoduše Google, je
internetový vyhledávač vyvinutý společností Google.
Patří sem synonyma, předpovědi počasí, časová pásma,
burzovní kotace, mapy, údaje o zemětřesení, časy
promítání filmů, letiště, seznamy domů a sportovní
výsledky.",
"positive_hypothesis": "Vyhledávač Google zobrazuje
informace o domovech.",
"negative_hypothesis": "Vyhledávač Google nezobrazuje
žádné informace o domovech.",
"correct_polarity": "P"
}
```

Models

- Multilingual models
- Transformer architecture
- Open-source, all available on Hugging Face [1]
- Evaluated models:
  - *Mistral-Nemo-Instruct-2407* (12.2 B, Mistral AI)
  - *Qwen2.5-7B-Instruct* (7.62 B, Alibaba Cloud)
  - *Llama-3.1-8B-Instruct* (8.05 B, Meta)

Models

Prediction Accuracy

Table: Model accuracy (P = positive hypotheses, N = negative hypotheses).

Model Name	P Accuracy	N Accuracy
Mistral-Nemo	79.81% (2075)	38.73% (1007)
Qwen2.5	87.38% (2272)	76.15% (1980)
Llama-3.1	71.35% (1855)	51.35% (1335)

Future Work

- Further exploration of negation in LM
  - What are the exact causes?
- Further development of the negation pipeline
  - Possible more complex forms of negation

Bibliography I

[1] *Hugging Face – The AI community building the future*. URL: https://huggingface.co/.

[2] Pavel Šmerk. “Fast Morphological Analysis of Czech”. In: *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2009*. 2009, pp. 13–16. URL: https://nlp.fi.muni.cz/raslan/2009/papers/13.pdf.

Bibliography II

[3] Anders Søgaard. “Should We Ban English NLP for a Year?” In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Ed. by Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 5254–5260. DOI: 10.18653/v1/2022.emnlp-main.351. URL: https://aclanthology.org/2022.emnlp-main.351.

[4] Milan Straka and Jana Straková. “Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe”. In: *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Ed. by Jan Hajič and Dan Zeman. Vancouver, Canada: Association for Computational Linguistics, Aug. 2017, pp. 88–99. DOI: 10.18653/v1/K17-3009.

Bibliography III

[5] 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, July 2023, pp. 101–114. DOI: 10.18653/v1/2023.starsem-1.10.

[6] Herbert Ullrich et al. “CsFEVER and CTKFacts: acquiring Czech data for fact verification”. In: *Language Resources and Evaluation* 57.4 (May 2023), pp. 1571–1605. ISSN: 1574-0218. DOI: 10.1007/s10579-023-09654-3. URL: https://doi.org/10.1007/s10579-023-09654-3.

Thank You for Your Attention!

MUNI
FACULTY
OF INFORMATICS