Introduction

PA154 Language Modeling (1.1)

Pavel Rychlý

pary@fi.muni.cz

February 21, 2024

PA154 – Technical Informations

- Slides in IS
  `https://is.muni.cz/auth/el/fi/jaro2025/PA154/`
- Final written exam, (open books, without interactive apps)
  60 points, 30 points for E
- optional individual projects
  up to 30 points

Individual projects

- presentation on a new research in language modeling
- small project as a part of bigger collaborative projects
  - neural machine translation
  - lexical acquisition
- small task
  - describe errors in ChatGPT
  - annotation of a langauge resource

Language model

- model
  - (mathematical) abstractions
  - similar/same behavior of modeled object
- language model
  - model a natural language

Language models—what are they good for?

- assigning scores to sequencies of words
- predicting words
- generating text

⇒

- machine translation
- automatic speech recognition
- optical character recognition
- chat, question answering

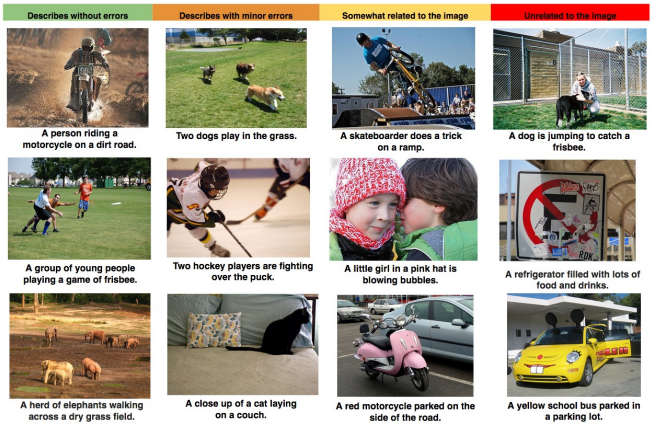
⇒

- ChatGPT and other LLMs are language models

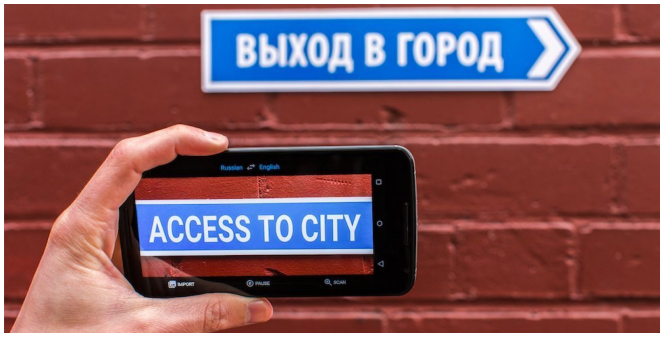
Predicting words

Do you speak ...
Would you be so ...
Statistical machine ...
Faculty of Informatics, Masaryk ...
WWII has ended in ...
In the town where I was ...
Lord of the ...

Generating text



MT + OCR



Language models – probability of a sentence

- LM is a probability distribution over all possible word sequences.
- What is the probability of utterance of s?

Probability of sentence

p_{LM}(Catalonia President urges protests)
p_{LM}(President Catalonia urges protests)
p_{LM}(urges Catalonia protests President)
...

Ideally, the probability should strongly correlate with fluency and intelligibility of a word sequence.