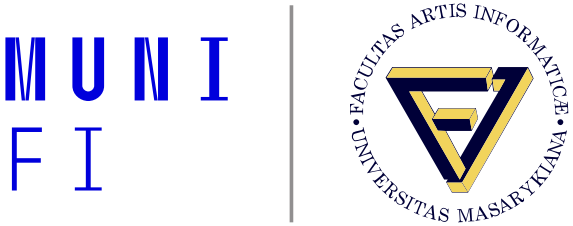


Introduction

PA154 Language Modeling (1.1)

Pavel Rychlý

pary@fi.muni.cz

February 20, 2024

PA154 – Technical Informations

- Slides in IS
  `https://is.muni.cz/auth/el/fi/jaro2024/PA154/`
- Final written exam (online)
  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

$\Rightarrow$

- statistical machine translation
- automatic speech recognition
- optical character recognition

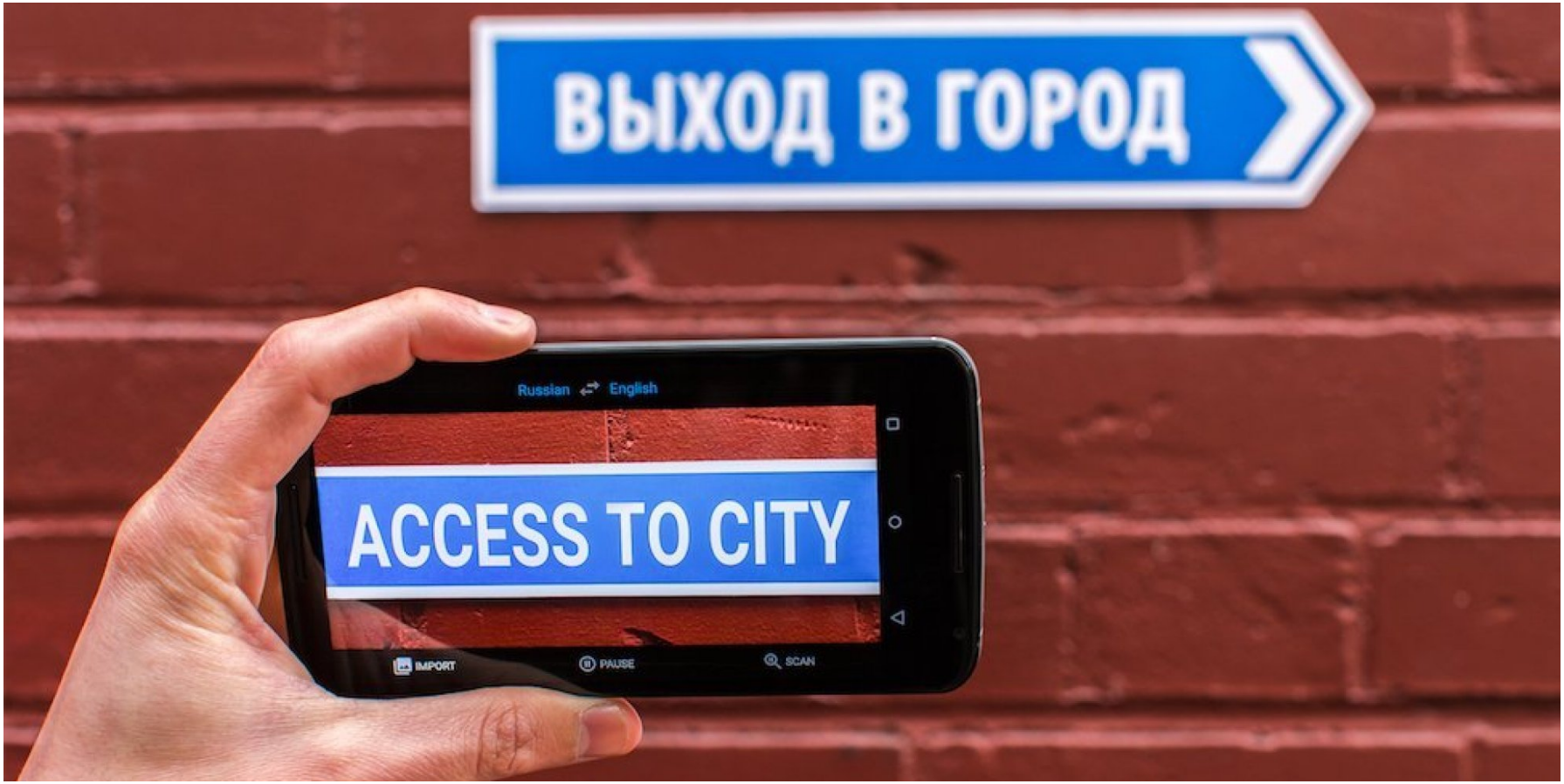
Predicting words

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

Describes without errors	Describes with minor errors	Somewhat related to the image	Unrelated to the image
			
A person riding a motorcycle on a dirt road.	Two dogs play in the grass.	A skateboarder does a trick on a ramp.	A dog is jumping to catch a frisbee.
			
A group of young people playing a game of frisbee.	Two hockey players are fighting over the puck.	A little girl in a pink hat is blowing bubbles.	A refrigerator filled with lots of food and drinks.
			
A herd of elephants walking across a dry grass field.	A close up of a cat laying on a couch.	A red motorcycle parked on the side of the road.	A yellow school bus parked in a parking lot.

MT + OCR


ВЫХОД В ГОРОД
Russian
English
ACCESS TO CITY
IMPORT
PAUSE
SCAN

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.