Natural Language Processing

PA153

Pavel Rychlý

23 Sep 2024

Natural Language Processing at FI

- Natural Language Processing Centre
  - around 10 PhD students
  - you can be part of it (PV173 - 3 credits each semester)

    bachelor/master thesis

    machine translation, AVER project
- Pavel Rychlý
  - head of NLP Centre
  - corpora, lexicography, machine translation

Technical information

- Study materials in IS
- Exam: written – max 10 questions
  - open books (offline)
  - max 60 points
- 30 point to pass (zk, k), (20 points for z)
- extra points (max 30) for homeworks, projects
  - correct typos in slides, improve slides
  - find good examples, illustrations to improve understanding
  - code, language, pictures
  - class competition is sentence boundary detenction, WSI
- exam, homeworks, … in English, Czech, Slovak

Previous knowledge

- no special requirements
  - reading mathematics
  - probabilities
- examples in Python
  - NumPy, PyTorch (matrix operations)
- complements
  - IB030: Introduction to Computer-based Natural Language Processing
  - IB047: Introduction to Corpus Linguistics and Computer Lexicography
  - PV021: Neural Networks
  - IA161: Natural Language Processing in Practice

Terminological remark

Used terms:

- Quantitative and statistical linguistics
- Algebraic linguistics (N. Chomsky)
- Mathematical linguistics
- computational (počítačová, komputační) linguistics
- Today Natural Language processing (ZPJ, NLP)
- Human language technology (HLT)
- speech processing (ASR, TTS)

Natural language (NL)

- Czech, English
- not formal languages (programming)
- 1000s different languages, sub-languages
- two different modalities
  - text: sentences, documents
  - speech: utterances, speakers

Motivation

Why to pay attention to natural language?

- Language behaviour represents one of the fundamental aspects of human behaviour.
- NL is an essential component of our life as a main tool of communication.
- In NL we express and record our knowledge, scientific findings, world understanding.
- Language texts serve as a memory of mankind for knowledge transfer between generations.
- NL is a base for human-computer communication.
- We want to know how **ChapGPT** works!

NLP – applications: MT

- Machine translation – testbed for NLP theory
- Georgetown–IBM experiment (1954) – demonstration
- ALPAC report (1966)
- Google Translator – first widely used
- Deep learning brings higher quality
- Human quality in many areas
- more in PV061 (Machine Translation)

NLP – applications: Text

- Text processing – spell checkers, grammar and style checkers
- Hyphenation, DTP
- Fulltext search (lemmatizaion, stemming)
- Semantic web – intelligent searching, exploiting metadata
- Information extraction
- Summarization

NLP – applications: Speech

- Speech communication with computers (robots)
- Synthesis – Text to speech systems
- Automatic speech recogition (ASR), dictating machines, smart phones
- Applications at courts, in Parliament, in medicine
- Can we have a chat with our computer? See PEPPER!

NLP – applications: AI

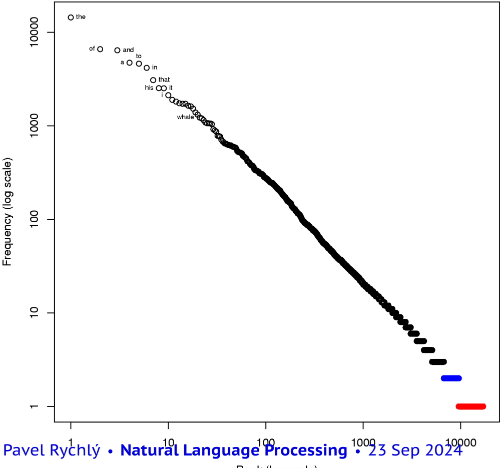
- Expert systems – e.g. Mycin (diagnostics in medicine)
- Dialogue and question-answering (QA) systems
- Turing test (Eliza, Loebner Prize)
- NL understanding in general, stories and messages
- Robotic applications – SHRDLU, 1971 (T. Winograd), the first system containing knowledge, inference and grammar
- Ontologies, semantic networks (WordNet)
- Robotic family NAO, **PEPPER**, ROMEO (Softbank)
- more in PV277 (Programming Applications for Social Robots)

Problems with NLP

- Zipf's law
  - high number of low frequent items (words, phrases, ...)
- Ambiguity
  - meaning depends on context
- Variability
  - languages evolve
  - new words/phrases
  - transfer from other areas

Problems: Zipf‘s law

- rank-frequency plot
- highly skewed distribution



Problems: Ambiguity

Many components in a natural language are ambiguous

- word meaning (*band*)
- wordforms (*he runs*, *my runs*)
    - basic form (lemma)
    - part of speech, morphological categories
- characters (l, L), different scripts
- names
- formal languages: unique identifiers

Problems: Variability

- languages evolve
  - old books are hard to read
  - different orthography, syntax, meaning
- new words/phrases
  - *mobile phone*
  - *Barbenheimer* (wikipedia page in 30 languages)

- transfer from other areas
- language is a live organism

Approaches to NLP

- symbolic
  - rules from experts
  - no data
- statistical
  - structure/model from experts
  - optimization of parameters from data
  - some data
- neural (deep learning)
  - everything from data
  - huge amount of data
- usually a combination

Example: sentence boundaries

Find rules to detect sentence boundaries.

- English: regular expression: `[.!?]`
- Is is good enough?
- Does it work in other languages?
- Is `[.!?] [A-Z]` better?

Outline of the semester

- morphology, syntax
- statistical NLP
- word embeddings
- neural networks
- recurrent networks
- transformers
- large languge models
- question answering, machine translation

Summary

- Problems with NLP
  - Zipf‘s law
  - Ambiguity
  - Variability
- Approaches
  - symbolic (rule-based)
  - statistical
  - neural (deep learning)