Natural Language Processing (PA153)

Pavel Rychlý, Karel Pala

Technical information

- Exam: written – 11 questions
  - max 50 points
- 25 point to pass
- extra points (max 25) for homeworks, projects
  - find good examples, illustrations
  - code, language, pictures

Previous knowlege

- no special requirements
  - reading mathematics
  - probabilities
- examples in Python
  - NumPy, PyTorch (matrix operations)
- complements IB030, IB047

Natural language (NL)

- Czech, English
- not formal languages (programming)
- 1000s different languages, sub-languages
- text
- speech

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,
- ▶ NL is a starting point for artificial (formal) languages
- ▶ Language texts serve as a memory of mankind for knowledge transfer between generations
- ▶ NL is a base for human-computer communication

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)

What NLP includes?

NL study and research is interdisciplinary:

- ▶ In linguistics (structural, mathematical)
- ▶ In psychology and psycholinguistics
- ▶ In philosophy and logic – relations to the universe of discourse, reasoning (inference), basic units are truth functions (výroky) and propositions
- ▶ In algebraic (later computational) linguistics (in sixties) key role was played by N. Chomsky (Synt. Struct.)
- ▶ Language theory in the form of algorithms, and data structures, large empirical data (corpora)
- ▶ Relations to the Artificial Intelligence and Cognitive science
- ▶ Computer instruments for NL – language engineering

NLP – relation to computers

- ▶ Need for a two-way communication human-computer
- ▶ So far H-C communication is mainly one-way
- ▶ A richer H-C communication interface is necessary
- ▶ NL interface should be smarter and more flexible, especially for common users
- ▶ Distinct commercial consequences for the computer market
- ▶ Influence to the shape of operation systems
- ▶ Our knowledge about NL structure is incomplete
- ▶ Relevant role is played by the relation of theory (research) and applications

NLP – applications 1

- Machine translation – testbed for NLP theory
- Georgetown–IBM experiment (1954) – demonstration
- ALPAC report (1966)
- Google Translator – first widely used
- Hard task but human quality in some areas

NLP – applications 2

- ▶ Text processing – spell checkers, grammar and style checkers
- ▶ Hyphenation
- ▶ Fulltext search (lemmating, stemming)
- ▶ Semantic web – intelligent searching, exploiting metadata
- ▶ Morphological and syntactic analyzers, semantics
- ▶ Machine readable dictionaries, ontologies
- ▶ Information extraction
- ▶ summarization

NLP – applications 3 (speech)

- ▶ Speech communication with computers (robots)
- ▶ Synthesis – Text to speech systems (Demosthenes)
- ▶ 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 4 (relation to AI)

- ▶ Expert systems – e.g. Mycin (diagnostics in medicine)
- ▶ Dialogue and question-answering (QA) systems
- ▶ Turing test (Eliza, Loebner Prize, November 2019)
- ▶ NL understanding in general, stories and messages
- ▶ Robotic applications – SHRDLU, 1971 (T. Winograd), the first system containing knowledge, inference and grammar,
- ▶ Robotic family NAO, PEPPER, ROMEO (Softbank, demo)
- ▶ Ontology and concept systems for particular domains, sémantic networks (WordNet)

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

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

Levels of language analysis

- ▶ Phonetics and phonology, speech signal
- ▶ Morphology – flection (and word formation)
- ▶ Syntax – constituent, dependency
- ▶ Sémantics – lexical, logical
- ▶ Pragmatics – relations of users to the language expressions
- ▶ Discourse, anaphorical relations, reference

Levels – phonetics, phonology

- ▶ sounds of language (phones)
- ▶ physical properties of the speech signal
- ▶ Phonology – phonemes – abstractions on sounds
- ▶ The smallest units distinguishing meaning, pas – pás
- ▶ Phonological oppositions: long – short: vola/á
- ▶ Link to the ASR – automatic speech recognition
- ▶ TTS (text to speech) – speech synthesis, many systems
- ▶ ASR (dictation systems, (ARŘ, Newton DS 5, 6)
- ▶ Intensive research, IBM, Nuance, a lot of money in it

Summary

- Problems with NLP
  - Zipf's law
  - Ambiguity
  - Variability

- Approaches
  - symbolic (rule-based)
  - statistical
  - neural (deep learning)