Počítačové zpracování přirozeného jazyka – PA153 (Natural Language Processing)

K. Pala et al

NLP Centre FI MU

Autumn 2020

Podmínky hodnocení

- Exam – written – 11 questions

NLP (ZPJ) – motivation

- Why to pay attention to NL (PJ)?
- 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 (shrnující)
- computational (počítačová, komputační) linguistics
- Today Natural Language processing (ZPJ, NLP)
- Computer speech processing (ASR)
- Cognitive science (linguistics, psychology, philosophy, also logics, usually in USA)

What NLP includes?

- NL study and research is interdisciplinary:
- In linguistics (tradiční, 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
- Can we have OS with NL? – e.g. OS Merlin (IBM)
- Our knowledge about NL structure is incomplete
- Relevant role is played by the relation of theory

NLP – applications 1

- Text processing – spell checkers, grammar and style checkers
- Hyphenation, fulltext programs (lemmatizers)
- Morphological and syntactic analyzers: Majka, Morfodita,synt, SET, NTA (semantics – TIL)
- Browsers, editors – web, dictionary tools
- Machine readable dictionaries (MRD), platform DEB
- Dialogue and question-answering (QA) systems
- Turing test (Eliza, Loebner Prize, November 2019)
- Information extraction, summarization, abstracts, MUC
-
-

NLP – applications 2 (MT)

- Machine translation (SP) – testbed for NLP theory
- EU projekty – EuroMatrix, EUM+, Presemt etc.
- Systran – at the beginning the official MT system for EU
- Google Translator – the best usable product exploiting language models and neural networks
- Systems with translation memory – Trados (localization systems), based mainly on parallel corpora
- Systems working with sublanguages (Taum, Meteo)
- Voice MT – system Verbmobil (1992-2001, German, Japanese, English)
- Quality of MT? Google, IBM, neural networks, limits

NLP – applications 3 (speech)

- Speech communication with computers (robots)
- Synthesis – Text to speech systems (Demosthenes)
- Automatic speech recogition – ASR systems), dictating machines, smart phones
- Via Voice (IBM), Dragon (Nuance), En.,Fr.,Germ., It.
- For Czech – system Dictate 4.5, 6..., Newton Technologies (demo)
- Applications at courts, in Parliament, in medicine
- The accuracy of understanding of these applications is approximately – 90 %
- Can we have a chat with our computer? See PEPPER!

NLP – applications 4 (relation to AI)

- Expert systems – e.g. Mycin (diagnostics in medicine)
- Database systems with NLP interface
- NL understanding in general, stories and messages Abstracts from newspaper articles – MUC (Message Understanding Conference)
- Robotic applications – SHRDLU, 1971 (T. Winograd), the first systém containing knowledge, inference and grammar,
- Robotic family NAO, PEPPER, ROMEO (Softbank, demo)
- Semantic web – intelligent searching, exploiting metadata
- Ontology and concept systems for particular domains, sémantic networks (WordNet)
- Social networks, Facebook? Google? Seznam?
- AI is just a buzzword for industrial applications and speculations

Structure (levels) of language

- Nature of language system – for language levels and their formal description, various theories exist
- Phonetics and phonology, speech signal
- Morphology – flection (and word formation)
- Syntax (skladba) – constituent, dependency
- Sémantics – lexical, logical
- Pragmatics – relations of users to the language expressions
- Discourse, anaphorical relations, reference
- Algoritmic desriptions for all levels are built plus

Paradigms in NLP

- Introspective – Chomsky, notions of competence : performance, generative and transformational grammars
- Grammars are understood as finite sets of rules – their incompletenes is essential
- Empirical data – beginning of corpora: Brown Corpus, H. Kučera, N. Francis (1960-61), 1M
- Large collections of language data (text files), billions
- Rule vs. statistical approaches, advantages vs. disadvantages, K. Church (TSD 2018)
- Machine learning, language models – (statistical methods seem to take more interest ?)

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

Morphology

- Units – morphemes, the smallest units bearing meaning (usually smaller than words, uč-, prac-)
- Morpheme types – bearing lexical meaning, i.e. roots or stems, and morphemes carrying grammatical meanings
- Words and their segmentation – morphological analyzers – algorithms – *ne/u/věř/i/t/eln/ému*
- Inflectional (tvarosloví) vs. Derivational morhology
- Analyzers Ajka, Majka, also (MorfoDita - UFAL)
- Derivational morphology – a tool Derivancze

Syntax

- Captures between words in a sentence
- Units – sentence constituents, parts and sentence types
- Representat. of sentence structure (tree graphs) Formal grammars – N. Chomsky results
- Grammar hierarchy, languages and automata
- Syntax conceptions – dependency and constit.
- Syntactic analysis (parsing) and analyzers
- Tools for Czech – Synt, Set, (Va)Dis

Semantics

- It does not have its own units as such
- Key question – what is meaning?
- One can distinguish meaning of words and collocations – this is lexical semantics
- meaning of sentences – logical semantics
- Semantic representations of sentences
- Formalisms used – predicate calculus, TIL etc.
- Mixed techniques – valency verb frames – as the room without windows – we cannot see neither out nor inside (similarity with Platon‘s shadows)

Lexical semantics

- meaning of words and word collocations
- Lexicology – deals with word stock
- Lexicography – processing of word stock – presently in the form of electronic dictionaries (MRD)
- computer lexicography, types of dictionaries
- Software tools for building and process. dictionaries
- Description of word meanings in dictionaries – definitions, synonyms
- DebDict (https://deb.fi.muni.cz:8005/debdict/), approach, platform DEB, DebVisDic (WordNets)

Pragmatics

- Relations between language users and language expressions
- Internal – attitudes of users to a proposition: indicative, interrogative, commanding, wishing (sentence types)
- External – communication situation, its elements and relations to a proposition
- KS = (m, p, o1, …, on, t, l)
- Pragmatická function – (meaning of *Já mám žízeň.*)
- Deixis and deictic elements
- Their role in a communication situation

Discourse analyis

- Structure of discourse
- Anaphoric relations and their recognition
- Recognizing of the discourse parts
- Reference and corefence
- Structure of dialogue
- Models of discourse (Box model)

Knowledge representation and inference

- Semantic networks (WordNet, ontologies)
- Logical formalisms – PK1, TIL
- Valency frames – VerbaLex, Vallex, Verbnet argument structure of predicates
- Deduction, monotonic – nonmotonic
- Systems exploiting Common Sense
- Communication agents, model Belief-Desire-Intention (BDI)

Machine learning and NLP

- Presently popular techniques – subfield of artificial intelligence
- Přehled – samostatná prezentace
- Učení bez učitele
- Učení s učitelem
- Klasifikátory

History of NLP in ČSR and ČR 1

- Praha – FF UK, seminář SP, 1958
- B. Palek, vztah k N. D. Andrejevovi.
- P. Sgall, P. Novák, D. Konečná, L. Nebeský, E. Hajičová, J. Panevová, P. Piťha, K. Pala
- M. Těšitelová – autorka Frekvenčního slovníku češtiny, 1961 (1983)
- Odd. matem. lingvist. ÚJČ, vztahy Letenská (L. Doležel) vs. Malostranské nám., ÚFAL (P.Sgall)
- J. Štindlová – počátek počítačového zprac. PJ na děrných štítcích

History of NLP in ČSR and ČR 2

- V Praze - seminář SP na FF UK od r. 1958
- Brno – počátek ZPJ v 1964 (K. Pala)
- Ústav českého jazyka FF UJEP (MU)
- V 70. letech počítačové experimenty s českými generativními gramatikami – analýza a syntéza (OVC VUT)
- Implementace syntaktické a sémantické analýzy na počítači Tesla 200 (Čihánek, Palová)
- Havel, Machová, Pala, Sofsem 1978
- V 80. letech spolupráce s ÚVT UJEP, vytvoření

Historie ZPJ v Brně I

- ÚVT – Benešovský, Šmídek, Gerbrich, programovací jazyk Wander (1988-90)
- 1988-9 první PC na FF UJEP MU), vznik morfologického analyzátoru pro češtinu, Xantipa
- Franc, Osolsobě, Pala, gramatický korektor, generátor a analyzátor českých vět v Prologu
- Od r. 1995 dochází k přesunu výzkumu na FI MU
- V r. 1997 vzniká na FI MU Laboratoř ZPJ
- Umožnily to grantové proj. podporované MŠMT

ZPJ na FI MU II

- Budování korpusových nástrojů (Rychlý, 1997-8), korpusový manažer Bonito/Manatee
- Vytvoření české lexikální databáze WordNet, 1999
- Vytvoření nezávislého morfologického analyzátoru Ajka (Sedláček, 1999)
- Pokročilá syntaktická a sémantická analýza češtiny: systém Synt (Horák), Set (Kovář), (VA)Dis (Mráková)
- Budování slovesné databáze komplexních valenčních rámců – VerbaLex (Hlaváčková, Pala)
- Nový morfologický analyzátor Majka, systém Deriv (Šmerk), a Derivancze (derivační morfologie)

ZPJ na FI MU III

- New corpus tools – Word Sketches – slovní profily), Rychlý, Kilgarriff), Lexic. Computing Ltd.
- Building large web corpora
- Toolkit:

  - Justext – removing garbage (boilerplate) from web pages

  - Onion – cleaning duplicities from web

  - Chared – recognition languages at web

  - Word Sketch Engine, NoSketch, Skell (Kilgariff, Rychlý, Suchomel, Jakubíček)