Počítačové zpracování přirozeného jazyka – PA153 (Natural Language Processing)

K. Pala et al

NLP Centre FI MU

Autumn 2019

Podmínky hodnocení

- Exam – written – 10 questions
- (Prezentace (výběrově) – na určené téma
- Rozšíření pohledu na probíranou problematiku
- Prez. není součástí zkoušky, ale přihlíží se k ní)
-
-

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 (Syn. Struct.)
- Language theory in the form of algorithms, and data structures, large empirical data (corpora)
- Relations to the Artificional Intelligence and cognitive science
- Computer instruments for NL – language engineering

NLP – relation to computers

- Need for a two-way communication h-c.
- 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 oper. 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, 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 neural networks
- System with translation memory – Trados (localization systems), based mainly on paralel 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 – TTS systems (Demosthenes)
- Automatic speech recogition – SR 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 level 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 conference (Message Understanding Conference)
- Robotic applications – SHRDLU, 1971 (T. Winograd), the first systém containing knowledge, inference, grammar,
- Robotic family NAO, PEPPER, ROMEO (Softbank, demo)
- Semantic web – intelligent searching, exploiting metadata
- Ontologie a konceptuální systémy pro jednotlivé domény, sémantické sítě (WordNet)
- Social networks, Facebook? Google? Seznam?
- AI is just a buzzword, industrial applications, speculations

Structure (levels) of language

- Povaha jazykového systému – jazykové roviny a jejich formální popis – existuje řada teorií
- Fonetika a fonologie, řečový signál
- Morfologie – flexe (ohýbání) a tvoření slov
- Syntax (skladba) – složková, závislostní
- Sémantika – lexikální, logická
- Pragmatika – vztahy uživatelů k jazyk. výrazům
- Promluva, anaforické vztahy, reference
- Na všech rovinách se budují algoritmické popisy a k nim vhodné počítačové aplikace

Paradigms in NLP

- Introspektivní – Chomsky, pojmy kompetence : performance, generativní a transformační gramatiky
- Gramatiky jsou chápány jako konečné množiny pravidel – jejich neúplnost je klíčová
- Empirická data – počátek korpusů: Brown Corpus, H. Kučera, N. Francis (1960-61), 1M
- Velké počítačové soubory jazykových dat, mld.
- Pravidlové vs. statistické přístupy, výhody vs. nevýhody, K. Church (TSD 2018)
- Strojové učení, jazykové modely – (kdo vede?)

Roviny – fonetika, fonologie

- Zvuková stránka jazyka – hlásky (fóny)
- Fyzikální vlastnosti řečového signálu
- Fonologie – fonémy – abstrakce nad hláskami
- Nejmenší jednotky rozlišující význam, *pas –pás*
- Fonologické protiklady: délka – krátkost: *vola/á*
- Vazba na zpracování mluvené řeči
- TTS (text to speech)– syntéza řeči, Demosthenes
- ASR (automatic speech recognition, ARŘ, demo)
- Intenzivní výzkum, IBM, Nuance, hodně peněz

Morphology

- Jednotky – morfémy, nejmenší jednotky nesoucí význam (obvykle menší než slova, uč-)
- Typy morfémů – nesoucí lexikální význam, kořeny či kmeny, morfémy nesoucí gramatické významy
- Slova a jejich segmentace – morfologické analyzátory – algoritmy - *ne/u/věř/i/t/eln/ému*
- Flexe (tvarosloví) vs. derivační morfologie
- Čeština je jazyk s bohatou morfologií proti angličt,
- Analyzátory Ajka, Majka, další (Morče) pro češtin.
- Derivační morfologie – nástroj Derivancze

Syntax

- Zachycuje vztahy mezi slovy ve větě
- Jednotky – větné složky, větné členy, typy vět
- Reprezentace větné struktury (grafy stromy)
- Formální gramatiky – výsledky N. Chomského
- Hierarchie gramatik, jazyků a automatů
- Koncepce syntaxe – závislostní a složková
- Syntaktická analýza (parsing) a analyzátory
- Pro češtinu nástroje – Synt, Set, (Va)Dis
- Statistické nástroje (MALT, Collins), n-gramy

Sémantics

- Nemá vlastní jednotky jako takové
- Klíčová otázka – co je to význam?
- Můžeme rozlišovat význam slov a slovních spojení
  – lexikální význam – lexikální sémantika
- Význam vět – větná či logická sémantika
- Sémantické reprezentace vět
- Používané formalismy – PK1, TIL aj.
- Kombinované techniky – valenční rámce
- Význam – jako místnost bez oken – nevidíme ven ani dovnitř (podobnost s Platonovými stíny)

Lexikální sémantika

- Významy slov a slovních spojení
- Lexikologie – nauka o slovní zásobě
- Lexikografie – zpracování slovní zásoby – nyní v podobě elektronických slovníků
- Počítačová lexikografie, typy slovníků
- Softwarové nástroje pro práci se slovníky
- Popis významu slov ve slovnících – definice, synonyma,
- DebDict (https://deb.fi.muni.cz:8005/debdict/), přístup, platforma DEB, DebVisDic

Pragmatika

- Vztahy mezi uživateli jaz. a jazykovými výrazy
- Interní – postoje uživ. k propozici: oznamovací, tázací, rozkazovací, přací (typy vět)
- Externí – komunikační situace a její prvky, vztahy k propozici
- KS = (m, p, o1, …, on, t, l)
- Pragmatická funkce – (*Já mám žízeň.*)
- Deixe a deiktické prvky
- Jejich role v komunikační situaci
-

Analýza promluvy

- Struktura promluvy
- Anaforické vztahy a jejich rozpoznávání
- Rozpoznávání částí promluvy
- Reference a korefence
- Krabicový model
- Struktura dialogu
-

Reprezentace znalostí a inference

- Sémantické sítě (WordNet, ontologie)
- Logické formalismy – PK1, TIL
- Valenční rámce – VerbaLex, Vallex, argumentová struktura predikátů
- Dedukce, monotonní - nemotonní
- Systémy využívající Common Sense
- Komunikační agenti, model Belief-Desire-Intention (BDI)

Strojové učení a NLP

- V současnosti populární techniky - podoblast umělé inteligence
- Přehled – samostatná prezentace
- Učení bez učitele
- Učení s učitelem
- Klasifikátory

History of NLP in ČSR a 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

- Nové korpusové nástroje – slovní profily (Word Sketches) (Rychlý, Kilgarriff), LCL – ukázat
- Budování velkých webových korpusů
- Soubor nástrojů:

  - Justext – odstraňování smetí z webových stránek (boilerplate)

  - Onion – čištění duplicit z webu

  - Chared – rozpoznávání jazyků na webu

  - WSE, NoSketch, Skell (Suchomel, Jakubíček)