Počítačové zpracování přirozeného jazyka – PA153 (Natural Language Processing)

K. Pala et al

Centrum ZPJ FI MU

Podzim 2018

Podmínky hodnocení

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

ZPJ (NLP) – motivace

- Proč si PJ zasluhuje naši pozornost?
- jazykové chování představuje jeden z fundamentálních aspektů lidského chování,
- PJ je podstatnou složkou našeho života jako hlavní nástroj komunikace,
- pomocí PJ vyjadřujeme a zachycujeme své znalosti, vědecké poznatky, vidění světa,
- PJ je východiskem pro umělé (formální) jazyky
- jazykové texty slouží jako paměť lidstva pro předávání znalostí z generace na generaci
- vztahy k technice a počítačům, komunikace h-c

Terminologická poznámka

- Používané termíny
- Kvantitativní a statistická lingvistika
- Algebraická lingvistika (Chomsky)
- Matematická lingvistika (shrnující)
- Počítačová (computational), komputační lingvistika
- Zpracování přirozeného jazyka (ZPJ, NLP)
- Počítačové zpracování mluvené řeči (ASR)
- Kognitivní věda (lingv., psychol., filos. i logika)

Co je předmětem ZPJ?

- PJ – studuje se a zkoumá interdisciplinárně:
- V lingvistice (tradiční, strukturní, matematická)
- V psychologii a psycholingvistice
- Ve filozofii a logice – vztahy k univerzu promluvy, usuzování (inference), pracuje se s výroky (propozicemi)
- V algebraické (komputační) lingvistice (60. léta min. stol.) je klíčová role N.
- Teorie jazyka ve formě algoritmů, dále jde o datové struktury, empirická data (korpusy)
- Vztahy ke kognitivní vědě a umělé inteligenci
- Počítačové modely PJ – jazykové inženýrství

ZPJ – vztah k počítačům

- Potřeba dvoucestné komunikace mezi čl. a počít.
- Zatím je komunikace člověk.-poč. jednocestná
- Potřeba komunikačně bohatšího rozhraní
- Rozhraní v PJ musí být chytřejší a pružnější – zejména pro nespecialisty – běžné uživatele
- Výrazné komerční důsledky pro počítač. trh
- Vliv na podobu operačních systémů
- Je možný OS s PJ? – pokusy s OS Merlin
- Naše znalosti o struktuře PJ jsou neúplné
- Roli hraje vztah teorie (výzkumu) a aplikací

ZPJ – aplikace 1

- Zpracování textů – korektory překlepů, gramatické, stylistické korektory
- Dělicí, fulltextové programy (lemmatizátory)
- Morfologické a syntaktické analyzátory: Majka, synt, SET, NTA (sémantika)
- Prohlížeče, editory – webové, slovníkové nástroje
- Strojově čitelné slovníky (MRD), platforma DEB
- Dialogové a otázkové (QA) systémy
- Turingův test (Eliza, Loebner Prize, listop. 2018)
- https://www.chatbots.org/ai_zone/viewthread/3129/ Extrakce informací, sumarizace, abstrakty, MUC

ZPJ – aplikace 2 (SP)

- Strojový překlad – max. snaha o využití v praxi
- EU projekty – EuroMatrix, EUM+, Presemt aj.
- Google Translator – aktuálně použitelný – neur. s.
- Systran – dříve oficiální systém SP v rámci EU
- Systémy s překladovou pamětí – Trados (lokalizační systémy), paralelní korpusy
- Systémy pracující s podjazyky (Taum Meteo)
- Hlasový SP – příklad: systém Verbmobil (1992-2001, němčina japonština, angličtina)
- Zlepšení SP? Firmy: Google, IBM, neuron. sítě

ZPJ – aplikace 3 (mluvená řeč)

- Hlasové ovládání počítačů (robotů)
- Syntéza – systémy TTS, Demosthenes (demo)
- Automatické rozpoznávání řeči (ASR), diktovací stroje, chytré mobily
- Via Voice (IBM), Dragon (Nuance), an.,fr.,něm., it.
- Pro češtinu – systém Dictate 4.5, 6..., Newton Technologies (demo)
- Aplikace na soudech, v parlamentu, v medicině
- Úroveň porozumění u těchto systémů – cca 90 %
- Můžeme si se svým notebookem popovídat?

ZPJ – další aplikace 4 (vztah k AI)

- Expertní systémy – např. Mycin (lék. diagnostika)
- Databázové systémy s PJ rozhraním
- Porozumění příběhům a porozumění PJ
- Abstrakty z novinových článků – konference MUC (Message Understanding Conference)
- Robotické aplikace – SHRDLU, 1971 (T. Winograd), první systém obsahující znalosti, inferenci, gram., NAO, PEPPER
- Sémantický web – chytré vyhledávání, uplatnění metadat
- Sociální sítě, Google? Seznam? IBM Watson?
- Ontologie a konceptuální systémy pro jednotlivé domény, sémantické sítě (WordNet)

Historie ZPJ v 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 – odd. matemat. lingvistiky, ÚJČ,

  - Frekvenční slovník češtiny, 1961, 1983
- Odd. matem. lingvistiky, ÚJČ, vztahy Letenská vs. Malostranské nám., ÚFAL
- J. Štindlová – počátek počítačového zprac. PJ na děrných štítcích

Struktura (roviny) jazyka

- 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

Paradigmata v 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

Morfologie

- 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

Sémantika

- 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

Historie ZPJ v ČSR a 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)