Počítačové zpracování přirozeného jazyka (vývoj)

K. Pala
Centrum zpracování přirozeného jazyka
FI MU
(podzim 2013)

Terminologická poznámka 1

- Kvantitativní a statistická lingvistika
- Už v 30. letech min. století
- Statistické vlastnosti jazykových jednotek
- Zkoumání frekvencí hlásek, znaků, slov
- Frekvenční slovník č. Jelínek, Bečka,Těšitelová
- Teorie informace – Shannon, Weawer, 1948
- Množství informace, entropie, redundance
- Stylostatistika, sporné autorství, G. Herdan (1956-66), G. Yule (1986, Study of Language),
-

Terminologická poznámka 2

- Počítačová lingvistika začíná s příchodem počítačů – text napsaný čínsky je vlastně anglický, jen je v jiném kódu (W. Weawer, 1947)
- Pokusy se strojovým překladem – cca od r. 1950
- USA – Georgetown University: P. Toma, Systran
- SSSR – Peterburg: N. D. Andrejev, Moskva: O. Kulagina, Ju. Rozencvejg, I. Mel'čuk
- ČSR – FF UK, B. Palek, P. Novák, P. Sgall
- ALPAC report, 1966, negativní hodnocení
- USA, UK, Francie, později Japonsko, Eurotra

Terminologická poznámka 3

- Strukturalismus – pražský (Jakobson, Vachek, Havránek, Mathesi Karcevskij, Firbas)

  - americký: Harris, Bloomfield, Wells, složková an.

  - dánský: Hjelmslev, Ulldal, glosématika, 50. léta
- Matematická a algebraická lingvistika
- N. Chomsky, Syntaktické struktury 1964, formální jazyky, grama hierarchie gramatik a automatů
- Využití matematických struktur pro popis jazyka
- Kvantitativní a kvalitativní (strukturní) matematika
- Generativní gramatiky, transformační g., sémantika
- Katz, Fodor, The structure of a semantic theory, 1963
- Ch. Fillmore, The case for case,1968 (Framenet)

Vznik ZPJ jako discipliny

Výsadní role PJ v komunikaci mezi lidmi

- jazykové chování je jedním z fundamentálních aspektů lidského chování

- PJ je podstatnou složkou našeho života jako hlavní nástroj komunikace a vlastně i myšlení

- V PJ vyjadřujeme a zachycujeme své znalosti,

- PJ je východiskem pro umělé (formální) jazyky

- texty v PJ představují paměť lidstva - předávání znalostí z generace na generaci

- PJ jsou východiskem pro vědecké poznání světa

Východiska ZPJ

PJ se studuje a zkoumá

- V lingvistice (tradiční, strukturní, matematické)
- V psychologii a psycholingvistice
- Ve filozofii a logice – vztahy k univerzu promluvy
- V počítačové (komputační) lingv. (60. léta min.stol.)
- Pojmy jako algoritmus, datové struktury, teorie ve formě algoritmů
- Vztahy ke kognitivní vědě a umělé inteligenci
- Počítačové modely PJ – výzkum vs. jaz. inžen.

Východiska ZPJ 2

- Nová disciplína – jazykové inženýrství
- Potřeba dvoucestné komunikace mezi čl. a poč.
- Zatím je komunikace čl.-poč. jednocestná
- Potřeba komunikačně bohatšího rozhraní
- Rozhraní v PJ musí být chytřejší a pružnější – hlavně pro nespecialisty
- Výrazné komerční důsledky pro OS
- Je možný OS s PJ? - pokus s OS Merlin
- Nejen textové, ale i hlasové ovládání

Východiska ZPJ 3

Softwarové aplikace v oblasti PJ

- Zpracování textů – korektory překlepů, gramatické, stylistické
- Dělicí, fulltextové programy (lemmatizátory)
- Morfologické analyzátory
- Prohlížeče (editory) – webové, slovníkové
- Strojově čitelné slovníky (MRD)
- Dialogové a dotazovací systémy
- Vztahy k UI -Turingův test (Eliza, Loebner Prize)
- Extrakce informací, sumarizace, abstrakty

Východiska ZPJ 4

- Strojový překlad – snaha o využití v praxi
- EU projekty – EuroMatrix, Presemt aj
- Google Translator – je prakticky použitelný?
- Systran – oficiální EU systém SP
- Systémy s překladovou pamětí – Trados (lokalizační systémy)
- Systémy pracující s podjazyky (Meteo, Taum)
- Hlasový SP – systém Verbmobil
- Kvalita a budoucnost SP?

Zpracování mluvené řeči

- Hlasové ovládání počítačů (OS Merlin)
- Syntéza – systémy TTS, Demosthenes (demo)
- Automatické rozpoznávání řeči (ASR), diktovací stroje
- Via Voice (IBM), Dragon (Nuance), ang., něm. …
- Pro češtinu – systém Dictate 2.5, Newton Technologies (demo)
- Aplikace na soudech, v parlamentu, v medicině
- Úroveň porozumění u těchto systémů
- Můžeme si se svým nb. popovídat (chatboxy)?

ZPJ – další systémy

- Expertní systémy – např. Mycin
- db. systémy s rozhraním v PJ
- Porozumění příběhům a porozumění PJ
- Abstrakty z novinových článků – konference MUC (Message Understanding Conference)
- Robotické aplikace
- Sémantický web
- Doplnění webu o metadata
- Ontologie a konceptuální systémy

ZPJ v ČSR a ČR 1

- Praha – FF UK, seminář SP, 1959
- B. Palek, vztah k N. D. Andrejevovi, SP.
- 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, 1983 (1960)
- L. Doležel, vedoucí odd. matem. lingvistiky, ÚJČ - šedesátá léta, metodologické konflikty
- J. Štindlová – mechanografická laboratoř v ÚJČ (60.-70. léta), počítače a děrné štítky

ZPJ v ČSR a ČR 2

- Brno – počátek ZPJ v r. 1964 (K. Pala)
- Ústav českého jazyka FF UJEP (MU), formální syntax češtiny
- V 70. letech experimenty s českými generativ. gramatikami – analýza a syntéza (OVC VUT)
- Implementace syntaktické a sémantické analýzy na počítači Tesla 200 (P. Čihánek, I.Palová) - TIL
- Havel, Machová, Pala, Sofsem 1978
- V 80. letech spolupráce s ÚVT UJEP, vytvoření

  českých gramatik v Prologu (počítač PDP 11), generování vět a syntaktická analýza

ZPJ v Brně 1

- ÚVT MU – Benešovský, Šmídek, Gerbrich, programovací jazyk Wander (1980-87)
- první PC na FF UJEP MU - 1988-9, vznik morfologického analyzátoru pro češtinu, Xantipa
- Franc, Pala, Osolsobě, gramatický korektor (editor T602), generátor a analyzátor českých vět v Prologu
- V r. 1995 dochází k přesunu výzkumu v oblasti ZPJ z FF na FI MU – obor počítačová lingvistika
- V r. 1997-8 vzniká na FI MU Laboratoř ZPJ
- Grantové projekty ve spolupráci s MFF UK – Čeština ve věku počítačů
- Další projekty – v rámci EU, EuroWordNet, 1999

ZPJ na FI MU (od 1995)

- Budování korpusových nástrojů (Rychlý, 1997-8), Bonito/Manatee, Word Sketches – slovní profily)
- Vznik české lexikální databáze WordNet, 1999
- Vytvoření nezávislého morfologického analyzátoru Ajka (Sedláček,1999, Lemma, Lingea, korektor)
- Pokročilá syntaktická a sémantická analýza češtiny – systém Synt (Horák), Set (Kovář), Dis, VaDis (Mráková), Grác (parciální analyzátor IOBBER – též polština)
- Budování slovesné databáze komplexních valenčních rámců – Verbalex (Hlaváčková, Horák)
- Nový mf. analyzátor Majka, systém Deriv (Šmerk)
- Nástroj Desamb a problematika značkování

ZPJ na FI MU 2

- Nové korpusové nástroje – slovní profily (Word Sketches) (Rychlý, Kilgarriff)
- Paralelní slovní profily – ukázat
- Budování velkých webových korpusů (1mld a více – ukázat) – Czes a CzTenTen12 (5,5 mld tokenů)
- Soubor nástrojů - Spiderling, JusText, Onion, Chared
- Spolupráce s Lexical Computing Ltd. – ukázky
- Syntaktické analyzátory – Synt a Set – vazby na gramatiky pro WS (slovní profily)

ZPJ na FI MU 3

- Bonito/Manatee – P. Rychlý – 1998-9
- Slovní profily (Word Sketches) s A. Kilgarriffem (2000)
- Objektivní zkoumání kontextů a významů na základě korpusových dat – podklady pro lexikografy
- Kombinace statistických technik (parametr logdice – vztah k MI, tj. k míře těsnosti kolokací (Hanks, Church)
- A technik strukturních (pravidlových) – korpus musí být označkován a je potřeba mít gramatiku pro sl. prof. WS
- Funguje to dobře na velkých korpusech – CzTenTen12
- GDEX – Kilgarriff, Husák – nástroj pro tvorbu slovníků
- Euralex 2008 (Barcelona)

ZPJ na FI MU 4

- Budování velmi velkých korpusů z webu
- Web je nejlepší korpus (Google)?
- Ukazuje se, že ne – webové stránky obsahují mnoho smetí (boilerplate)
- Nástroje pro automatické budování korpusů z webu
- Spiderling - plazivec (crawler), prochází www stránky
- JusText – odstraňuje smetí, čistí www stránky
- Onion – odstraňuje duplicity na www stránkách
- Chared – prochází www stránky a rozpoznává jazyky a jejich kódování

ZPJ na FI MU 5

- WordNet – G.A.Miller, Ch. Fellbaum 1995 (Princeton)
- Psycholingvistická východiska – zkoumání lidské lexikální paměti
- Experimenty s reakcemi na podněty (výrazy jako canary – bird) – rozdíly v rychlosti
- Hierarchické uložení – vztah hyperonymie/hypon.
- Miller navrhl slovník WordNet – slovní síť
- Vztahy H/H, antonymie, holo/meronymie a další
- EuroWordNet I (1997-8), EWN II (1998-9), P. Vossen
- CzWn – 1999, chytrý nápad ILI – mezijazykový index

ZPJ na FI MU 6

- VerbaLex

ZPJ na ÚFALu v Praze

Počátek v r. 1968-9, rozdělení

- Největší skupina v ČR, (Sgall) Hajičová, Hajič, Panevová, Žabokrtský, Kuboň, Bojar aj.)
- Morfologický analyzátor pro češtinu
- Závislostní syntax – PDT (vytvořen manuálně)
- Strojový překlad – účast na projektech EU (Euromatrix)
- Teorie tektogramatiky
- Publikace

ÚTKL v Praze

- Byl zřízen 1990 (P. Sgallem)
- Vedoucím je doc. V. Petkevič, dále A. Rosen, Z. Skoumalová, T. Jelínek
- Česká morfologie – pravidlový přístup
- Zpracování českých valencí na základě dat z databáze Brief
- Spolupráce s ÚČNK, projekt NovaMorf, pokus sjednotit pražskou a brněnskou variantu české morfologie (poziční vs. atributový přístup)
- Podíl na českém gramatickém korektoru

ÚČNK na FF UK

- Čermák, Cvrček, Křen, vznikl 1994 na FF UK
- Český národní korpus, nyní je vedoucím Cvrček http://korpus.cz/co_je_korpus.php
- SYN2000, SYN500 aj.
- Používají korpusový manažer Manatee/Bonito
- Intercorp – paralelní korpusy pro cca 15 jazyků - http://www.korpus.cz/intercorp/
- Tradiční způsob budování korpusů
- Snaha o vyváženost, brání se webovým korpusům
- Značkování nástroji z ÚFAlu MFF UK – problémy
- Přesnost je cca 96 %, to je v SYN2000 4 mil. chyb

ZPJ na Slovensku

- Ján Horecký –1960-2000
- V. Benko – počítačová lexikografie
- SNK – M. Šimková, R. Garabík – v JÚĽŠ, Bratislava
- http://korpus.juls.savba.sk/
- Používají korpusový manažer Manatee/Bonito
- Aktuální verze slov. webového korpusu 2013 - obsahuje cca 900 mil. tokenů, je značkován
- Vztah k ZPJ – značkování
- M. Rusko, Ivanecký – zpracování mluvené řeči
- Spolupráce v oblasti korpusových nástrojů - WSE

Světový kontext

- USA: N. Chomsky, generativní a transf. gr., jeho škola, výsledky a následky (25.9. ...?)
- ČR: ÚFAL (2.10. ...?)
- UK: korpusy, BNC (9.10. ...?)
- Rusko: Kulagina, Apresjan, Mel'čuk, další (16.10. ...?)
- Francie: Grenobl, Colmerauer (Prolog)(23.10. ...?)
- Německo: M. Bierwisch a další (30.10. ...?)
- Japonsko (6.11. ...?)
- Jednotlivé školy (13., 20., 27.11. ...?)
- Připravit vždy cca 20. min. prezentace

Chomsky – generativní gramatiky

- Kompetence vs. Performance
- Internalizovaná znalost jazyka
- Soubor konečných pravidel pro nekonečné jaz.
- Hierarchie gramatik – regulární, nekontextové, kontextové, přepisovací systémy (typy automatů, konečný, zásobníkový,Turingův stroj)
- Nekontextové gr. - G = (Vt, Vn, R, S), kde
- Vn – terminální slovník, Vn – množina neterm.s
- R – množina pravidel A → omega, A patří do Vn, omega je řetěz symbolů z Vt a Vn

Jednoduchá nekontextová gr.

- S → Np Vp

  Np → N/A Np

  N → {pes, kočka, tygr, maso, ...}

  A → {hladový, neviditelný, rezavá, divoká, …}

  Vp → V

  V → {žere, trhá, …}
- Neadekvátnost této gramatiky pro češtinu, jak ji upravit?
- Morfologie – pády, čísla, osoby

Transformační pravidla

- Nekontextová pravidla nestačí – pasivní, tázací konstrukce: *Maso je žráno psem*
- Větné struktury se reprezentují pomocí grafů-stromů
- Transformační pravidla přiřazují stromy stromům, jsou složitější než nekontext. pravidla
- V Syntakt. str. Chomsky navrhl gen. gramatiku tvořenou jádrovou a transformační složkou (kolem r. 1960)
- Transformační gr. v této podobě neobsahuje sémantiku – Chomsky se sém. nevyrovnal, stand. a rozšířená teorie GG
- The Structure of a Semantic Theory, Katz a Fodor, 1963
- Komponentová analýza, amalgamační pravidla, slovník, sémantická interpretace, aplikace transf. Pravidel
- The Case for Case, Ch. Fillmore, 1968, vlivná teorie, objev hloubkových pádů a pádových rámců, vztah k valencím

Transform. gr. a reakce na ně

- transf. gr. angličtiny včetně pokusů o implementace se v zásadě neprosadily
- Konkurenční teorie: závislostní (dependency) gr.
- HPSG - head-driven phrase structure grammar - Carl Pollard, Ivan Sag.
- Lexikálně funkční gramatika – Bresnan, Kaplan kolem 1970
- Kategoriální grammatika – Bar Hillel, 60. léta
- tree-adjoining grammar - Joshi
- Funkčně generativní přístup - Sgall, Hajičová, systém zásobníkových převodníků mezi rovinami, tektogram. r.
- Daneš, Hlavsa: větné vzorce – ne generativní (1981)

Situace v UK

- BNC vznikl1994 – poslední verze je BNC XML Edition, publikovaná v r. 2007.
- http://www.natcorp.ox.ac.uk/
- Rozsah 100 mil. tokenů, 90 % psaný text, 10 % mluvené (transkribované) texty
- Značkování je podle Guidelines of the Text Encoding Initiative (TEI)
- Hlavní podíl mají tři pracoviště: OUP (Hanks), Uni. of Birmingham (Sinclair), Uni of Lancaster (Leech)
- Významná osoba v UK je A. Kilgarriff (LCL)
- Největší korpus angličtiny je v Centru ZPJ – ClueWeb
- Cca 80 miliard tokenů (webové korpusy Wac, TenTen)

Počítačová lingvistika v Rusku (SSSR)

- Historický exkurs – od 30. let min. stol. - N. Marr a jeho následovník I. Meščaninov – ovl. ruskou lingvist.
- čtyřelementová jafetická teorie (Jáfet, Noeho syn)
- Nová věda o jazyce – všechny jazyky se vyvinuly ze čtyř elementů, sal, roš, jon, a ber
- Jazyky mají třídní, nikoli národní původ
- Marrismus hrozil zničit obecnou lingvistiku v SSSR, ovládal významné pozice a likvidovali oponenty fyzicky
- V r. 1950 vyšel Stalinův článek v čas. Voprosy jazyk., v němž marrismus odsoudil
- Ve skutečnosti článek pro Stalina napsal gruzínský lingvista Čikobava
- Proti tomu zde byla jména Jakobson, Karcevskij z PLK

Počítačová lingvistika v SSSR

- Počátek v r. 1956 – strojový překlad z ruš. do franc.
- V. Ju. Rozencvejg, O. Kulagina, I. Melčuk, J. D. Apresjan, A. Žolkovskij – v Moskvě
- N. D. Andrejev v Leningradě (vazba přes B. Palka) – statistický přístup (1956)
- Melčuk a Žolkovskij: teorie význam – text, 1960
- Dále Tolkovo-Kombinatornyj slovar' sovremennogo russkogo jazyka, 1984
- Rozencvejg, Mašinnyj perevod i prikladnaja lingv. 1974
- Apresjan – práce v oblasti lexikální sémantiky - Leksičeskaja semantika, 1974
- Institut russkogo jazyka AN, Institut Informelektro aj.

Počítačová lingvistika v Rusku – pokr.

- V. Zacharov. G.J.Martynenko – Peterburg, korpusový systém Linda, Pet. státní univerzita
- I. Boguslavskij – Moskva, Institut perenosa informaciji, systém ETAP 1,2,3
- Navazuje na práce Melčuka, Apresjana
-

Historický exkurs II

- 25. srpna 68 – demonstrace 7 osob na Rudém n.
- mj. Konstantin Babickij – kolega Melčuka a Apresj.
- Jako jediní demonstrovali proti invazi v Československu
- Zatčeni KGB, odsouzeni – Babický šel na 3 roky do vyhnanství (v rep. Komi na severu)
- Po návratu nesměl pracovat v Institutu ruského jazyka, živil se jako řezbář
- V r. 1993 byl přijat V. Havlem
- V r. 2013 demonstrace k výročí invaze – účastníci opět rozehnáni a pak zatčeni

NLP ve Francii

- Bernard Vauquois, strojový překlad, 1972, Grenoble,
- Marcel Schuetzenberger, 1967, formální gramatiky
- Maurice Gross, Laboratoire d'Automatique Documentaire et Linguistique (LADL), Paříž, 1968
- A. Colmerauer, programovací jazyk PROLOG, 1984, Marseille, protějšek v USA – jazyk LISP
- Prolog vychází z predikátové logiky, z její podmnožiny, Hornových klauzulí
- Deklarativní jazyk, data jsou oddělena od výpočetní procedury, jazyk pro počítačovou lingvistiku a UI
- Dokazování teorémů, expertní systémy, hry, dotazovací systémy, ontologie, parsing

PROLOG a syntaktická analýza

- Philippe Roussel, programování v logice
- Gramatiky konečných klauzulí (Definite clause grammars, DCG) (s. 55 demo)
- PROLOGové systémy fungují jako syntaktické analyzátory (parsery) a generátory
- Analýza shora, využívá se unifikace, rekurze a backtrackingu
- Prologové systémy – SWI, SICSTUS (ukázat)
- Česká morfologie v PROLOGU
- Česká syntax v PROLOGu, přepis nekontextových pravidel do PROLOGu

NLP v Německu

- M. Bierwisch, Humboldtova universita, od 1985
- Modern linguistics: its development, methods and problems. Mouton, The Hague 1971
- komponentová analýza – lexikální sémantika
- Popis významu slov pomocí sémantických rysů
- Soubor sémantických rysů – svého druhu formální jazyk
- *muž*: +Hum +Mask +Adult
- *žena*: +Hum +Fem +Adult
- *chlapec*: +Hum +Mask -Adult
- *Děvče*: +Hum +Fem -Adult
- Vhodné pro popis výrazů označujících rodinné vztahy

ZPJ v Německu 2

- Saarland University – nejsilnější skupiny NLP v Německu
- M. Pinkal, Dept. Of Computational Ling. and Phonetics
- H. Uszkoreit, Language Technology, Meta-Forum, Net
- DFKI - Deutsches Forschungzentrum fuer Kuenstliche Intelligentz, W. Wahlster
- Stuttgart University, U. Heid, korpusy, korpusové nástroje, CQP (1996), CQL, počítačová lexikografie
- Tuebingen University, E.Hinrichs, Kunze, GermaNet, Clarin
- Institut fuer Deutsche Sprache, Mannheim, 1964, největší korpusy němčiny, korpusové nástroje (Grammis, Cosmos)
- Deutschen Referenzkorpus, 3,2 miliardy tokenů

Japonsko I

- Communications Research Laboratory (CRL)
- Japan Electronic Dictionary Research Institute (EDR)
- Japan Advanced Institute of Science and Technology (JAIST), Hokuriku
- Kyushu Institute of Technology, Department of Artificial Intelligence, Lizuka, Fukuoka
- ATR Interpreting Telecommunications Research Laboratories, Kyoto
- Kyoto University, Language Media Laboratory
- Nara Institute of Science and Technology (NAIST), Computational Linguistics Laboratory

Japonsko II

- Tokyo
- National Institute of Informatics (NII) (formerly NACSIS)
- Tokyo Institute of Technology
- University of Tokyo, Department of Information Science, Natural Language Processing
- Electrotechnical Laboratory, Tsukuba