Úvod do počítačové lingvistiky

Počítačová lingvistika

- od 60. let, NLP (Natural Language Processing)
- problematika analysy textů či jejich generování
- tvoří jazykové modely (gramatiky a tyhle blbosti)

Struktura jazyka

- info o slovech, jak se skládají do vět, z čeho se skládají, jak se z nich tvoří významy vět

Roviny analysy jazyka

- hierarchicky: fonetická → morfologická → syntaktická → sémantická → pragmatická

• *slide #4 in file plo2.pdf*

Fonetická

- rozlišování zvuků (fonémů)

Morfologická

- struktura slov, detekce (např.) slovních druhů na základě tvarů

Syntaktická

- struktura frází, vět
- analysátor analysuje na základě **gramatických pravidel** daného jazyka

Sémantická

- významy kombinací výrazů, závisí na zvolené **sémantické representaci**

Pragmatická

- významy v **kontextu**

Oronym/orofón

- fráze podobně znějící s jiným významem

Fonetická transkripce

- přepsané texty vstupem pro programy
- IPA
- SAMPA – strojově čitelná fonetická abeceda – speciální znaky zastoupeny znaky jako {, #, @ apod.

Synthesa řeči

- TTS – text to speech – **konverse psaného textu na mluvený**
- dva moduly:
    - jazykový (NLP) – in: text; out: fonémy + prosodické nfo (intonace, ...)
    - digit. zpracování signálu (DSP) – in: out z NLP; out: zvukový soubor

- čtyři fáze:

- normalisace textu – rozvinutí čísel, zkratek, členení na věty
- fonetická transkr.
- prosodická transkr. – tempo řeči, větný přízvuk, intonace – určuje, jak přirozeně to pak zní; může značně ovlivňovat porozumění
- akustické modelování – výsledný zvykový signál, DSP modul; dva přístupy:
    - synthesa v časové oblasti – konkatenativní – spojují se předem namluvené kusy řeči (fóny); deformace aplikací pravidel (intonace apod)

- synthesa ve frekvenční oblasti:
    - modelování hlas. ústrojí – počítačově se modeluje průchod vzduchu ústrojím a z toho se počítá, jak bude daný zvuk vypadat – děsně výpočetně náročné
    - formantová synthesa – základní zvuk se udělá jednoduchým modelem ústrojí a pak se to modifikuje pomocí filtrů; méně náročné, méně dat než u konkat. synt.

Rozpoznávání řeči (ASR – automatic speech recognition)

- řeč na text
- většinou vyhodí spoustu hypothes s pravděpodobnostmi (podle kontextu, p(toho_slova) v jazyce apod.)
- universální diktovací nástroje schopny rozpoznat cokoliv (jednotl. slova i plynulou řeč), ale musí se trénovat na mluvčím
- nástroje dedikované pro jeden účel (např. řízení auta – blinkr vlevo a par dalších příkazů, nemusí umět rozpoznat phil disputace) – určená regulární gramatika, nezávislé na mluvčím
- **tři fáze:**
    1. záznam signálu, vzorkování (digitalisace)
    2. vytvoření akustických charakteristik signálu
        - *wtf*
        • *slide 15 in file plo3.pdf*
        - vytvoří si záznam (vektor) o vlastnostech toho signálu
    3. porovnání vektorů parametrů
        - porovnávání vlastností získaného signálu s vlastnostmi naučených signálů
        - algoritmus borcení časové osy – ostraňuje čas. nerovnosti signálu – *wtf*
        - HMM – skryté Markovovy modely – konečné automaty; ústrojí může z jednoho stavu do jiného přejít jen s určitou pravděpodobností

Morfologie

- morfém – nejmenší jednotka **nesoucí význam**
• *slide 3 in file plo4.pdf*
- dělení morfémů se liší v analytických a flektivních jazycích
    - eng má dělení morf. **volné** × **funkční** a **obsahové** × **vázané**
    - cs dělí na **kořeny** a **afixy**, jež se pak dělí na milion dalších věcí podle funkcně a postavení ve slově (*viz slidy*)
- dělení morfologie: flektivní, derivační, komposicionální

- **fundace** (motivace) – zákl. slovotvorný vztah mezi slovy fundujícími (základními, motivujícími) a fundovanými (odvozenými, motivovanými)
- fundující slova nemají původ v jiných slovech jazyka (*voda, hlava*)
• *viz http://www.osu.cz/fpd/kcd/dokumenty/cestinapositi/tvoreni_slov.htm*
- **slovotvorná řada** – odvozování z odvozených slov (*ryba → rybník → rybníkář*)
- ~ svazek – slova odvozená od stejného fundovaného slova
- ~ čeleď – příbuzná slova (stejný kořen)

Morfologická analysa

- značkuje slova, klasifikuje slovní tvary
- kategorie analysy slov (PoS – Part of Speech):
    - lexikální – pojmenování věcí, dějů, … – subst, adj, pron, num, verb, adverb, interj
    - gramatické – vyjádření vztahů ve větě – prep, konj, part, articles
- slovní tvary se rozpoznávají morfologickým analysatorem – provádí lemmatisaci, správnost důležitá pro další vrstvy rozpoznávání
    - analysátor vyhodí všechny možné interpretace (analysy), tagger vybere nejpravděpodobnější a guesser si něco inteligetně vymyslí, když analysátor neví

- pro eng funguje **stemming** – odesekávání koncovek, 36 značek, <90% úspěšnost (Brillův značkovač)
- pro cs nutno mít **vzory** slov podle kmenů, **paradigma** – deklinační tvary slova (jak ho lze ohýbat)
- **vzor**: „reprezentace tvaroslovného paradigmatu paradigmatem určitého konkrétního slova." – prostě to, jak můžu ohýbat celou kategorii slov vyjádřím seznam tvarů jejího představitele (vzoru)

Popis vzorů

- segmentace od **konce slova** – rozdělí se na segmenty – kmen (neměnný), intersegment (proměnlivá část vzoru), množina koncovek (přípustné koncovky vzoru)
- segmentace od **začátku slova –** řeší se prefixy

trie – morfologický lexikon (ML) – *wtf*

• *slides 18-21 in plo4.pdf*

Syntaktická analysa, syntax

- **syntax** – správné tvoření vět pomocí gramatických pravidel
- ALGOL 60 – 1. progr. jazyk popsaný pomocí **Backus-Naurovy formy** (BNF) – *and da hell is dat about?*
    - zjistili (asi), že gramatika toho jazyka je stejná jako CFG a začali zkoumat progr. jazyky z hlediska přirozených

Gramatiky

- gramatika – pravidla generující správné řetězce jazyka

• *slide 4-5 in file plo5.pdf*

Termíny – viz slidy

• *slide 5-8 in file plo5.pdf*

Specifikace gramatik

- složkový přístup
    - skládají se jednotlivé fráze (složky): subst. součástí NP, přidá se preposice → PP, ...
    - struktura pomocí derivačního stromu

- závislostní přístup
    - jeden člen označen jako řídící, další na něm závisí
    - struktura pomocí závislostního stromu

- většinou se oba přístupy kombinují
- **syntaktický** (derivační nebo závislostní) **strom** – uzly jsou neterminály, označují roli slova ve větě, syntaktické funkce
    - gramatická role – vztahy mezi členy gramaticky
    - tematická role – vztahy sémantické (agens, patiens, ...)
- **příznaky** – syntaktické / sémantické informace (slovní druh, ...)

Kategoriální gramatiky

- slova jsou funkce, ktere urcuji, jak se budou kombinovat s jinými slovy
- význam složeného výrazu je složen z významů jednotlivých výrazů obsažených v něm

• *slide 4 in file plo6.pdf*

Závislostní gramatiky

- řeší závislosti slov, vztah mezi slovesem a jeho možnými doplněními
- využívá valence – vztah mezi slovem a jeho argumenty *wtf*
- vhodné pro popis jazyků s volným slovosledem

TAG gramatiky

- přímo stromu (ne řetězce)
- základ – **počáteční stromy** (pak nějaké pomocné)
    - počáteční stromy – základ, popisují složkovou strukt. jednoduchých vět (e.g. jmenná fráze), nerekursivní
        - listové uzly – terminály nebo neterm. určené k substituci [*wtf*]
        - nelistové ~ – neterminály

- pomocné ~ - rekursivní, popisují větné členy připojované k základním strukturám (e.g. přísl. určení)
    - navíc patový uzel – listový uzel neterminální, stejné označení jako kořenový uzel – slouží k připojení stromu k jinému uzlu
- operace v TAG:
    - substituce – nahrazení uzlu stromem, který má kořen stejného jména jako ten uzel
    - připojení – adjunction – vložení pomocného stromu… [*dál jsem to nepochopil]*
- TAG se definuje:
    - množ. konečných počátečních stromů
    - množ. pomocných stromů
    - neterminál S – věta

LTAG – lexikalisová verze, uzly mají lexikální kotvu na slovo ve slovníku
- větší síla než CF gramatiky, generují mirně kontextové jazyky

Lexical Functional Grammar

C-struktura

- organisace do frází, vnější, viditelná
- representována CFG stromem
- zachycuje frázovou dominanci [*wtf*]

F-struktura

- organisace podle gramatických funkcí, magie
- matice dvojic s věcmi typu **atribut-hodnota**
    - atributy jsou číslo, čas, rod, …

- **motivace**: ty obecné věci se objevují v mnoha jazycích, které se jinak třeba liší poskládáním vět apod.
- po interpretaci c-struktury získáme f-strukturu

[*následuje spousta strašně složitých metapavěcí*]

• *slides 24-34 in plo6.pdf*

HPSG

- modeluje gramatiku pomocí příznakových struktur, které korespondují s výrazy daného jazyka
- cílem teorie - které příznak. struktury jsou v daném jazyce přípustné
- založeno na omezeních
- *wtf wtf wtf wtf*
- není derivační – neodvozuje syntakt. úrovně se neodvozují od dalších
- uspořádané typované příznakové struct.

- každá věta (finitní sentence) má SUBJ noun [podmět] a hlavu (finitní slovesnou frázi), ta se dělí na HEAD → V(finitní sloveso) a COMP(object, předměty, komponenty)
- omezení jsou ty shity, které říkají, že tenhe shit se pojí s tímhle shitem, ale s tamtímhle shitem se nesmí pojit

Synt

Metagramatika

- asi soubor příkazů, které pracují jako příkazy gramatiky, v systému synt

Výstupy systému synt

- závislostni / syntaktické stromy
- seznamy frází
- zjednodušené morf. značky

Analysator typu chart

- dostává pravidla gramatiky a podle nich kontroluje, zda nějaké slovo je vytvořitelné tou gramatikou... *pokud to správně chápu*...

Princip komposicionality

- význam složeného výrazu je jednoznačně odvoditelný z významů výrazů, ze kterých se skládá
- omezení: idiomy a podobné idioty = **listémy**

Anaphora

- odkazování na něco v textu ("Honza šel pro penis, když v tom (**on**) uviděl...")

Indexické výrazy

- odkazování na něco mimo text ("Proč jsi **to** udělal?")

Pojem, výraz

- **výraz** representuje **pojem**, ten identifikuje **objekt**, který je označován **výrazem**

Intense je závislá na světě, extense nikoliv

Možný svět

- soubor myslitelných faktů
- největší konsistentní soubor takových faktů
- objektivní

~ jeden z nich je aktuální svět

Representace znalostí

- data ve formě logických formulí, vyvozování znalostí = důkazy

Sémantické sítě

- významy ve formě vztahů
- některé vztahy jsou dědičné, jiné ne ("kráva je hnědá"="všechny části krávy jsou hnědé", ale "kráva je šťastná" != "všechny části krávy jsou šťastné")
- uzly, spoje, hodnoty na druhém konci spoje

Rámce

- podobné jako sém. sítě, dědičnost, data v universálních **rámcích**
- objekty, sloty, hodnoty slotu

Pravidlové systémy

- if, then

Slovník × encyklopedie

- slovník uvádí **významy** listémů, encykl. **info o tom, co popisují**

_____

Zkratky:
(A)TN – (Augmented) Transition Network
(L)TAG – (Lexicalized) Tree Adjoining Grammar
AVM – atribut-value matrix (viz LFG)
AR – ??
BNF – Backus-Naurova forma
CCG – kombinatorické kategoriální gramatiky
CFG – context-free grammar (bezkontextová gramatika)
CG – categorial grammar
CKY – Cocke, Kasami, Younger (algoritmus)
DC gramatika – ??
FGD – Functional Generative Description
HG – Head Grammars – Pollard, 1984
HPSG – Head-driven Phrase Structure Grammar
IPA – international phonetic transcription
LFG – Lexical Functional Grammar
LFG – Lexical functional grammar
LIG – Linear Indexed Grammars – Gazdar, 1985
MCSL – mildly context-sensitive languages (mírně kontextové jazyky)
ML – morfologický lexikon
NLP – natural language processing
NP – noun phrase (jmenná fráze)
PJ – přirozený jazyk
PK – princip komposicionality
PoS – part of speech
PP – preposition phrase (předložková fráze)
TIL – transparentní intensionální logika
VP – verbal phrase
ZPJ – zpracování přirozeného jazyka