Stručný zápis témat a krátkých vysvětlivek k předmětu Úvod do počítačové lingvistiky vyučované doc. Kuboněm na MFF v ZS18/19. Zdrojem tohoto textu je samotná přednáška, ale také doplňková skripta prof. Hajičkové (Úvod do teoretické a počítačové lingvistiky - 1. svazek) a materiály poskytované k tomuto předmětu.

Přirozený jazyk

Funkce:

- popis věcí a jevů v reálném světě
- popis abstraktních struktur
- komunikace směřující k řešení
- popis samotného jazyka

Zásady:

- všeobecnost (nejčastější způsob komunikace)
- využitelnost (je agilní a nestárne)
- obsah (je úplný, včetně meta popisu)
- vágnost (jistá neurčitost je základem inovativního myšlení)
- vícevrstevnost (dokáže popisovat i různé úrovně komunikace)
- zkratkovitost (bývá kratší než dotazy v umělých jazycích)

Víceznačnost: Občas formální problém, ale i častý základ vtipu, nebo myšlenky.

Morfologie

Počátek už v -400 v popisu sanskrtu. Předmětem morfologie je **studium vnitřní struktury slov**.
Lexikologie - slova jako jednotky slovní zásoby
Lexikografie - sestavování slovníků
Morfém - nejmenší znaková jednotka jazyka nesoucí význam (lexikální|gramatický); skládá se ze sémat
Séma - nejmenší znaková jednotka jazyka vztahující se k formě
Foném - nejmenší zvuková jednotka

Za.hrad.ou
Předpona (prefix) . mofrém (lexikální) . morfém (gramatický, skládá se ze tří sémat (3. p., č. j., r. ž.))

Dom.ům morfém (lexikální) . morfém (gramatický - dvě sémata (plurál + dativ))

Skoňování - deklinace
Časování - konjugace
Tvaroslovné dublety - kolize v odvození z různých slovních základů | žena (1/5), tři (5/4), stát (5/1), už (5/6
Alternace - změna hlásek uvnitř kmene v tvarování | vůz - vozu, švec - ševce, prkno - prken
Alomorfy - varianty téhož morfému/kmene | -říkˇ-, -říc-, -řek-
Autosémantická slova - nesou význam | podstatná, přídavná jména, zájmena, číslovky, slovesa, příslovce, citoslovce (většinou ohebá - první čtyři se skloňují, slovesa časují a některá příslovce stupňuj)
Synsémantická slova - nemají samy o sobě plnohodnotný význam | obvykle předložky, spojky, částice (všechna neohebná)

Morfologická typologie

- **analytické (izolační)** - slovo = morfém | angličtina, čínština
- **syntetické (flektivní, aglutinační)** - slovo > morfém,
  u flektivních jazyků (slovanské) mají afixy více funkcí, zatímco u aglutinačních jazyků (maďarština) je zřejmá korespondence a řetězí se častěji
- **polysyntetické ("přehnané aglutinační")** - v případě, kdy třeba sloveso dostane takové množství afixů, že nese význam věty | eskymánština

Metody zpracování morfologie

Založeno na:

- slovo jako posloupnost morfémů
- slovo = $P_\alpha \cdots (P_b(P_a(\text{kořen})) \cdots)$, kde P_i je pravidlo
- je dána tabulka vzorů + všechny jejich tvary; u každého slova určíme vzor a z toho odvodíme gramatické kategorie

Dvojúrovňová morfologie

Místo funkce, která generuje výsledné slovo se provádí dva výpočty zároveň: lexikální a povrchový.

`b a b y + 0 s`
`b y b i 0 e s`
`0` obvykle značí prázdné místo (protějšek nemá žádnou realizaci)
spojuje se do zápisu: `b:b a:y b:b y:i +:0 0:e s:s`
některé stavy ohlašují gramatické kategorie (třeba `s:s` → `plural`)

Poměrně nový přístup. Doposud byl problém ten, že dva různé generovací průběhy mohly skončit ve stejném slově a tudíž zpětná analýza není jednoznačná.

Česká morfologie

Každé slovo má 13 (+2 rezerva) značek, které mají dle pozice svůj význam a popisují morfologické kvality.

nejnezajímavější

`AAFP3--3N--` adjective . regular . feminine . plural . dative . - . superlative . negated . -
no poss. gender, no poss. number, no person, no tense, no voice, base variant

Lemma - slovníková reprezentace základního tvaru | lesům, lesy, lesích → les
stát/slov. → stát-1, šel → jít

Morfologická analýza

→
`Prezident rezignoval na svou funkci.`
←

```
<csts>
<f cap>Prezident<MMl>prezident<MMt>NNMS1---A--
<f>rezignoval<MMl>rezignovat_:T<MMt>VpYS--XR-AA--
<f>na<MMl>na<MMt>RR-4-----<MMt>RR-6-----
<f>svou<MMl>svůj-1_^(přivlast.)<MMt>P8FS4-----1<MMt>P8FS7-----1
<f>funkci<MMl>funkce<MMt>NNFS3---A--<MMt>NNFS4---A--
<MMt>NNFS6---A--
<D>
<d>.<MMl>.<MMt>Z:-------
</csts>
```

Použití:

- **morfologická analýza** - výsledek je seznam lemmat a značek (klidně více dvojic)
- **morfologické značkování** - výběr správné dvojice lemma-značka
- **částečná morfologická desambiguace** - pomocí pravidel jazyka zahazujeme značky, které by v daném postavení nemohly reprezentovat správnou větu
- **lemmatizace** - výběr správného lemmatu, ze kterého byl odvozen vstupní tvar (třeba pro vyhledávání v textu)
- **stemming** - odříznutí koncovky, dostaneme kořen slova
- **generování** - výběr správného slovního tvaru, pokud známe lemma a dostatek gramatických kategorií

Kontrola překlepů

1. nalézt všechny výskyty a opravit je
2. opravená verze musí sedět do kontextu
3. neznámá slova nejsou chyby
4. no false positive
5. co nejvíce automatická korekce
6. vše počítat co nejrychleji

Dva triviální přístupy:

1. porovnání řetězců se slovy ve slovníku
   - seznam všech možných slovních tvarů daného jazyka / seznam lemmat + morfologická analýza
   - výhoda: spolehlivé a jednoduché
   - nevýhoda: závislé na kvalitě slovníku, který se musí udržovat; neznámá slova jsou chybná
2. srovnání skupin znaků (dvojice, trojice, .. n-gramy) a hledání nedovolených kombinací

- výhoda: výpočetně rychlé, nezávislé na slovníku
- nevýhoda: spousta chybných slov se skládá ze správných kombinací znaků

Vylepšení:

- počítat se vznikem chyb (časté chyby, blízkost kláves, ..)
- pravopisné chyby, ne jen překlep (mně x mě, shoda podmětu s přísudkem)
- heuristicky (či strojové učení) na neznámá slova
- vzít v potaz i kontext slova
- na základě confidence chyby jeden z možných návrhů:
  1. nic nenabízet
  2. chcete A, nebo B
  3. potvrďte B
  4. auto oprava na B

ASIMUT

Automatická Selekce Informací Metodou Úplného Textu

Vyhledávací modul

Jazyk pro vyhledávání, např. *vzdálenost!, odstup! -3- rodinný! -1- domek!*. Vlastně docela triviální.

Jazykový modul

- pracuje s předpokladem, že slova se stejnou koncovkou mají stejné skloňování
- složen z retrográdního slovníku, seznamu vzorů a výjimek
- algoritmus
  - odzadu porovnávej vstupní slovo až do doby, než je jasné, jak se dané slovo skloňuje (popř. výjimka)
  - speciální kódování české diakritiky (pomocí čísel)
  - lze sepsat do pravidel v binárním stromu
- výhoda: dobrej nápad
- nevýhoda:
  - počet výjimek může být příliš velký
  - není vždy jasné určit vzor na základě jen koncovky a málo pravidel
  - příliš hrubá klasifikace → příliš mnoho možných koncovek → přegenerování
  - verze pro slovesa funguje ještě hůř (časování)
- **negativní slovník** - obsahuje slova, která nejsou důležitá pro dotazování (spojky, citoslovce apod.) a jsou v první fázi odstraněna z textu
- **konkordance** - všem důležitým slovním tvarům se přiřadila adresa a frekvence výskytu a pak se hledalo jen na konkordanci; slova v negativním slovníku měla adresu, ale nulovou frekvenci pro určení vzdálenosti mezi významovými slovy

MOZAIKA

- MOSAIC (Morphemic Oriented System of Automatic Indexing and Condensation)
- indexace obvykle řešena pomocí slovníku klíčových slov + v dokumentu spočtena četnost
- MOZAIKA řeší relevanci + více pojmů pro jeden denotát
- založeno na pozorování části gramatiky: *-or, -er* je konatel děje, *-ity, -ness* vlastnost apod. (en)
- je třeba pro každou tématickou oblast (např. elektrické inženýrství) vytvořit slovník takových přípon

Algoritmus:

- vstup je text i s formátováním
- projde to lemmatizací a morfologickou analýzou
- pokud je slovo irelevantní k danému tématu, tak se vyhodí (negativní slovník)
- kondenzace jmenných skupin pomocí jednoduché gramatiky (operační zesilovač TESLA KC 415 → zesilovač (s vyšší váhou)
- váhy na základě pozice výskytů v textu (uprostřed nejméně důležité, na konci a začátku textu nejvíce)
- normalizace dle délky dokumentů
- výstup je seznam deseti nejvýznamnějších termínů v dokumentu → další zpracování pro vnější účely

Výhody:

- není nutné vytvářet masivní slovník klíčových slov
- kondenzace jmenných skupin je chytré

Nevýhody:

- pracné vytváření tématických slovníků a pravidel
- zájmena by měla zvyšovat četnost, ale to MOZAIKA nedetekuje

Syntax

Závislostní strom

- rozumný pro jazyky s volným slovosledem (zejména slovanské)
- přehledný, zachycuje zřejmé syntaktické vztahu mezi členy
- neříká však nic o tom, jak sestrojit
- ne všechny vztahy jsou řídící a podřízený (např. Petr a Pavel)
- lze zploštit

Složkový strom

- vhodný pro jazyky s pevným slovosledem
- méně přehledný, obsahuje uzly, které nejsou větné členy, předpokládá bezkontext
- lze automaticky zpracovat lépe (lze třeba uzávorkovat)
- lze zploštit

Neprojektivní konstrukce

- obvykle uspořádáváme vrcholy tak, jak jsou ve větě
- strom nad danou větou je neprojektivní, pokud obsahuje neprojektivní závislost
- **neprojektivní závislost** - závislost mezi dvěma slovy oddělenými ve větě třetím slovem, které (ani nepřímo) nezávisí na žádném z nich
- při splácnutí nelze nakreslit šipky bez křížení, nelze uzávorkovat
- existují v mnoha jazycích

Transformační gramatika

Historie:

- Deskriptivismus: popis a klasifikace faktů, ale nikoliv vysvětlení (povrchové)
- Analytická syntax
- Logický přístup: surface & deep structure (různé výrazy mohou reprezentovat stejný význam, ale zároveň význam může mít více reprezentujících výrazů)

Noam Chomsky

3 komponenty:

- **Báze** - bezkontextová pravidla generující složkové stromy (phrase markers)
- **Transformační komponent** - pravidla operující na celých úrovních, vytváří povrchovou strukturu věty
- **Fonologický komponent** - pravidla generující fonetické interpretace

Tree Adjoining Grammars

- podobná myšlenka přepisování, ale přepisují se celé stromy, nikoliv jen řetězce
- při substituci se musí vkládaný kořen a nahrazovaný list shodovat
- algoritmus končí až když už nelze nic substituovat (nebo není třeba)
- síla bezkontextových gramatik, ale lze udělat i silnější (s modifikacemi)

Lexial Functional Grammar

- c-structures: spojují slova do frází
- f-structures: reprezentace funkčních vztahů ve větě (např. shoda), matice atribut-hodnota
- každá c-struktura spojena s pouze jednou f-strukturou, ale ne naopak

Kategoriální gramatiky

- každý slovní tvar má přiřazen kategorii (např. [S NP]/NP)
- malé množství pravidel typu: [X/Y => X]

Unifikační gramatiky

- každý objekt má seznam vlastností/atributů (v neuspořádané množině)
- unifikace je pak spojování tohoto seznamu; je dovolena pouze pokud nejsou konfliktní atributy (jinak sporná sestava)
- spojují se však vlastnosti, které spolu nesouvisí, takže výrazy mají u sebe nerelevantní údaje řešení: typované sestavy rysů (typ určuje možné vlastnosti)
- sestavy rysů pak obsahují kombinaci rysů, jenž popisují konkrétní jev (např. shoda)

HPSG

Hot Potato Soup with Garlic
Generování řetězců pomocí pozice v hiearchii

Augmented Transition Networks

3 typy hran, přechod mezi sítěmi

Q-systémy

- transformace grafů (vět)
- stromy jsou linearizovány (s neprojektivními konstrukcemi se dá nějak vypořádat)
- přesně definované typy objektů: atomy, stromy a seznamy stromů
- jednopísmenné proměnné, logické operátory
- spojování vrcholů, gramatická pravidla

Funkční generativní popis, valence

- ještě větší štěpení: 5 rovin (fonetická, fonologická, morfématická, povrchová, tektogramatická)
- tektogramatická rovina nejvyšší, závislosti
- valence (povolené kombinace různých členů - aktantů), Vallex
  - vytváří formy (Lojban)
  - TG rovina má členy: konator/aktor, patient, adresát, origo, efekt (každý pouze jednou, ale lze triviálně koordinovat)
  - každý aktant může být ještě obligatorní/fakultativní (obligatorní nesmí chybět)
  - aktant je obligatorní, pokud nemůžeme odpovědět na otázku pomocí *nevím*
  - *Moji přátelé přijeli.* Kam? **Nevím**, Odkud? Nevím. *Kam* tedy má obligatorního aktanta

Kontrola gramatiky

- dobrá heuristika je řešit pouze nejčastější chyby
- nechceme žádné false positive (precision blízko 1, na recall kašleme)
- jazyky s volným slovosledem jsou problematické

RFODG

Robust Free-Order Dependency Grammar je použití ručně psaných sjednocovacích gramatik. Různé fáze výpočtu (pozitivní projektivní, negativní projektivní/pozitivní neprojektivní, negativní neprojektivní). Různé gramatiky mohou popisovat i chybné věty.

LanGR

Pravidla opět psaná ručně. Nejsou aplikována systematicky (neuspořádaně), ale cyklicky.

Automatický překlad přirozených jazyků

Kromě samotného slovníku slov je třeba znát tvarosloví (morfologie), pravopis (syntax), ustálená spojení (idiomy). Slovníky však nejsou 1:1, neboť reflektují kulturu. Závisí taky na kontextu z předchozích vět (podobný problém lemmatizace).
Obvykle se posunujeme v analýze směrem k interlingvě a v nějakém bodě provedeme transfer a vygenerujeme cílový text.

Historie

Už od války - první vyvrcholení v USA utnuto ALPACem, pokračování v Evropě. Žádně extra dobré výsledky. Řešením bylo omezit doménu. Systém TAUM - METEO (vytvořili k tomu Q-systémy) první úspěšný. Pak pokusy s SYSTRAN a EUROTRA. VERBMOBIL chtěl kromě překladu dělat rozpoznávání a syntézu řeči (rok 2000).

Konec

Nějak tak jsem si uvědomil, že pěkněji zpracované poznámky jsou na mff wiki. Obsah přednášek se moc nemění během let, tudíž je asi rozumnější řídit se tím.