

Úvod do počítačové lingvistiky

- Počítačová lingvistika = obor, jehož cílem je rozvoj vlastnosti přirozeného jazyka pomocí matematiky
PL
 - cíl: pochopit jak přirozené jazyky fungují
- Počítačové reprezentace přirozeného jazyka = blízké aplikace metod abstraktní NLP něm na velká jazyková slova (korpusy)
 - můžeme si něčeho o funkci tohoto jazyka
- Oblasti PL
 - Morfologie
 - Syntaxe
 - Semantika
 - Formalismus
 - Korpusová lingvistika
 - Statistická lingvistika
 - Strojový překlad
 - Rozpoznávání a generování mluvené řeči
 - Vyhledávání v textu
 - Dialogové systémy
- Problemy
 - přirozený jazyk je nelineární
 - slovo má své číslo, rod, pád... jednoznačné

Morfologie

= studium vnitřní struktury slov

morfém = nejméně znaková jednotka jazyka menejší než samostatný význam

za - hrad - on = preféra

preféra ↑ ↑
prefix lexikální, gramatický morfém

prefix ↳ určuje 3 sémata: žánr, číslo a rod
slovoráni = deklinace
časováni = konjugace

Problemy

- Transliterace dublety = jiné tvary, stejný význam: na hrade / hradu
- homonymum / tvary < stejný tvary, různé rozdělení: stát (země, slavos, sloves)
- alternace = random zmena klasických rena (podstatné, přechodné)
- plurisemantika (byl) × formální slova určitě význa: něk → roční

Morfologické typy jazyků

- Analytické: slovo = morfém → čínština
- Synetické: slovo > morfém → slovanské jazyky
- Pojsynetické: slovo = věta → indiánské jazyky

Přístupy spracování morfologie

→ ráckod slov bude

anglická

• morfemy: slovo = říkají morfémů

?

• lexemy: slovo = nejdoklešší aplikace pravidel, co méně ráckod a význam

slovník

• slova: hlavní roli hrají vzory

↓

→ když mám ráckod tvor + vzor, tak umím

breaks
broke
broken

vygenerovat ostatní tvary podle těchto vzorů

→ tohle se nazývá formálně

"
I break

→ věstivě je asi 250 jednoznačných vzorů

Two-Level Morphology - 1980

- první obecný model správného přezkoušení jazyka
- → jazyk svůj slovík a pravidla, ale mechanismus morfologie obecný
- 2 úrovně
 - lexikální
 - funkcionální
- pravidla se uplatňují parallelně, nikoli sekvencně
- funkcionálně se mohou vztahovat k 1 úrovni nebo k obouv různých
- lexikální rámec \leftrightarrow funkcionální rámec \leftrightarrow kontextové rámce
- lexikální + vyhledávací + trii a morfológická analýza probíhají současně
- Ceská morfologie
 - princip znaky, když řeči ve znaci má nejaly význam
 - 1) kategorie: slov' druh, rod, číslo, osoba, čas, ...
 - lemma = jednoznačný identifikátor toho slova = radikál + korekta + index
- Morfologická analýza: dativne slovo p. j. stál \rightarrow stát-1
 - vytvoří sestavu lemma a znak, která popisuje jednotlivé možnosti, což slovo může být vymezeno
 - např. pokud vše jde o má slejšího bratra \Rightarrow vše znak
 - pokud to slovo má vše významy \rightarrow vše lemma
 - + → lemma má nejaky jiná pravidla na delší znak
- Ačkoli label může být spousty
- Morfologické znakování = Tagging
 - z nich mohou znaky vybíráme buď správnou v daném kontextu
 - statistické metody (nasázíme se oto)
- Ceská morfologická disambiguační - Odstraňování
 - řadě gramatických pravidel jsou některé konfigurace nelegální
 - ↳ shoda podmínek s pravidlem, shodny příklasky
 - ↳ z. j. řád po předložce s
 - odstraňuje pouze ty znaky, které jsou 100% špatné

Lemnaceae

- proces výběru správného rozhodujícího zákonu

• Stemming

- sdílené konverg.
 - na rozdíl od demokracie je zákonem určen even slova

• Generation

- ručné lemma a kombinaci gramatických kategórií
⇒ černe správny slovní stav

• Environnement

- hledáme nejpravděpodobnější posloupnost znaků pro daný rečník
 - rovníkové Bayesův vzorec $P(x|y) = \frac{P(y|x) P(x)}{P(y)}$
 - $P(\text{značka}|\text{slovo}) = \frac{P(\text{slovo}|\text{značka}) \cdot P(\text{značka})}{P(\text{slovo})}$, ale $P(\text{slovo})$ je pro všechny znaky stejná
 - ⇒ čení je maximální
 - čím více po sobě jdoucích slov a jejich znaků bude v náhodě, tím lepsi, ale je to drahé a nejsou daleko
 - ⇒ kritikou jsou na fakt znáčka a fakt výsledek několika znaků na poslední znáčku
 - nej znáčka = $\arg \max_{\text{značky}} \prod_{i=1}^m P(\text{slovo}_i | \text{značka}_i) \cdot P(\text{značka}_i | \text{značka}_{i-1})$

Skrysti markovny modely HMM

- metoda analyzy řad (posloupnosti následků v čase)
 - my se aplikujeme na posloupnost morf. znaků v řeči
 - je to posloupnost rozhodnutí, co má sobě nejč. navázání (gramat. gramida)
 - Marcovova hypotéza: kontext = tři první znaky, co s nimi souvisí; je možné ztráctit má nejdůležitější znaky a řeči budou delší
→ kontext délky 2 = bigramy, délky 3 = trigramy
 - Skryté, protože některé vlastnosti se posloupnosti nejsou než vidit - jsou tam slova, ne znaky
 - je to basically stochastický konecny automat
(↳ náhodný)

Jak se HMM používají

1) Rozpoznávání (ohodnocení) statistického modelu

→ jsou dány parametry HMM, cíl je specifikovat jst.
že je rozpoznaná posloupnost X .

→ použití: rozpoznávání drásek - registracích rucek aut

2) Dekódování

= hledání nejpravděpodobnější posloupnosti s�edých stavů
↳ máme model a posloupnost rozpoznání

3) Včlenění statistického modelu

→ máme struktuuru modelu a rozpoznanou drášku

→ chceme najít jst. přesnosti mezi stavy a jst. nich stavů

Kontrola předlepu

Předavky:

- mališ a opravil všechny předlepy
 - Jen výsledek by neměl být nesmysl
 - Edyž nejde slovo neznám, tak to není chyba
 - Řádné false positives
 - co nejméně automaticky
 - co nejrychleji
- } samozřejmě
nereálné

Metrika náspěšnosti

$$\text{Precision } P = \frac{\# \text{ true positives}}{\# \text{ positives}} \dots \text{kolik kohor hlasům doleje}$$

$$\text{Recall } R = \frac{\# \text{ true positives}}{\# \text{ chyb v textu}} \dots \text{kolik \% chyb jsem odhalil}$$

→ Precision je reálně důležitější, když nerodi false positives

$$F\text{-measure } F = \frac{2PR}{P+R}$$

Metody kontroly překladu

1) Porovnání rěčíků se slovy ve slovnici

seznám všechny možné slovníky slovnic = word lists
slovník lemmat + morfologická analýza

(+) spolehlivé a simple

(-) pomale, může být na krok dřív slouzen, neznáma slova = chybou slova
+ slepění musí do slovníku přidat všechny (ab so vše je nec. m.-antonym)

2) Slovnářování skupin znaků (digram, trigram) + nejake rázovane kombinace

(+) rychle na slovníku, rychle

(-) hnedě chybíto nevhodné - chybou slova mohou být různé kombinace

Možná výhledy

- možnost ohraničení výběru chyb - blízké klávesy
- rohlednutí statistická chyby - lišta souboru podm. s pris. je. hodnoty
- rohlednutí pravopisné chyby - nne x me, jsem x jste
- heuristika na oddílení chyb a neznámých slov
- zapojení gramatiky a semantiky
- pracovat s kontextem - formovat korpusy

Jak se to reálně dělá

- správné slovo → jak vybrat možné správné slovy?
→ Levenshteinova vzdálenost rěčíků
- důležitá je přesnost (Precision) a rychlosť
- kontrola na pozadí, rádce upozorní kde se nachází chyba
- výběr výčtu do slovníku jen konkrétní slova - nenech morfologie

Systém ASIMOVÍ

= Automatická Selece Informací Metodom Úplného Textu

- 2 moduly

- Jazykový modul

- nemí rozsáhlý slovník - používá řádky
- mnoho slv, která mají v rozkladním tvaru stejný koncový segment, se stejně sklonuje
 - ⇒ retrográdní slovník
 - ⇒ pravidla kam jít podle rodu + čísla + pádu a když zadávaný koncovec je využití nejde suffix, což dává tvar tvaru
 - ⇒ jsou rádově až jenom slovy vyznamenávající

- Vyhledávací modul

- basically to vyhledávání regexy
- následující z podst. a příd. jmen v rozkladním tvaru + pravidly
 - ! rozdílnost slv = jazykové tvar
 - 1 - obě slva vedle sebe
 - 2 - mezi nimi slvy obsahují nejméně 2 slva
 - 3 - obě slva ve stejném rázku
 - 4 - - 5 - ve stejném odslouzení
- vyhledávání v nejzákladnějších významených důležitostech
Důležitosti: rozdílnost!, odslouzení! - 3 - rodinný! - 1 - domě!
 - ... jak delero od sebe mají být domy
- problémy jazykového modulu
 - proč nemí vždy určen jednoznačně
 - příliš hrubá klasifikace
 - malý rozsah retrográdního slovníku ⇒ výběrová vyznamenávání
 - nefunguje to tak spolehlivě pro slva
 - ↳ koncové segmenty zad. tvaru slv jsou významná!

- Negativní slovík

- obsahyé nedůležitá slova pro vyhledávání (spojky, čísla, jazyk)
- ↳ odstranění se při preprocessingu textu

- Kondence

- + slovík, které co nemají v negativním slovníku dostane
 - ↳ adresu
 - ↳ frekvenci výskytu
- náčl: urychlit hledání

Systém MOZAICKA

= Morphemic Oriented System of Automatical Indexing and Condensation

→ hledá nejdůležitější termíny z nějaké oblasti (technika, průmysl) a textu

1) Standardní přístup k indexaci

- slovík s hledaných slov, dokumenty indexovány slovy
- význam se bere celostní výskytu

2) Přístup MOZAICKY

→ řada koncových přípon mívají různé funkce

- ič, -ac, -čka, -ér, -or, -dro, -metr, -graf, -fon, -shop ... naškápnout
- ace, - ece, -ák ... procesy
- ost, -ila, -mce ... vlastnosti
- ač, -ec ... náčl.

anglická: - er, -or = koncové deje

- sion = činnost
- isty, -ness = vlastnost

→ pro polohy semantické oblasti elektrických obvodů ~ 800 přípon

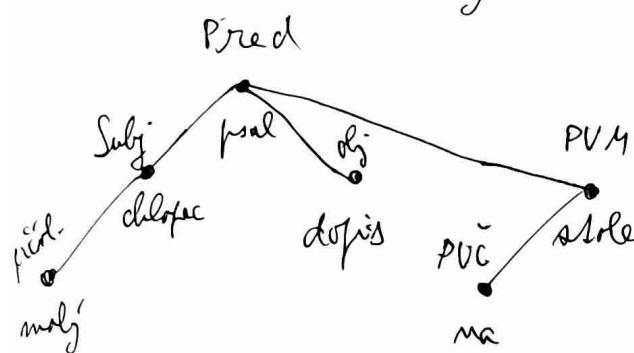
Algoritmus MOSAIKY

- 1. - lemmatizace + morfologická analýza vstupního textu → knací
 2. projde nálezená lemmata a všechnu ty, ježichž člen nemá vzhled k dané semantické oblasti (např. člověk) → vznikne malý negativní slovník
 3. Chci odhodit několikrátovně termíny
operací resilnací TESLA KC 415 → resilnací
 4. priřadím ráhy na základě místa výskytu v textu
popis je nejvíce
 1. odstavec > dobbí odstavce
 1. věta v odstavci > random věta
 5. udělam několik normalizací vah vzhledem k délce dokumentu
↳ aby se mohla formovat relevance různě slabých dokumentů
⇒ výstup = 10 nejvýznamnějších termínů sčítanou výskytu
- (+) nepřebírá slovník odbočujících termínů, jen možna přípona členů
lokalní syntaktická analýza tvoří ještě hledat termíny
+ negativní slovíčka / pravidla
- (-) vymyslel si přípony a pravidla je právě
nepřesná / náročná
- ↳ Jen resilnací TESLA sam můžu napsat 1x
a pak se na něj 10x odkazovat takovým
⇒ 1 výskyt

Syntax

právadek = hojen

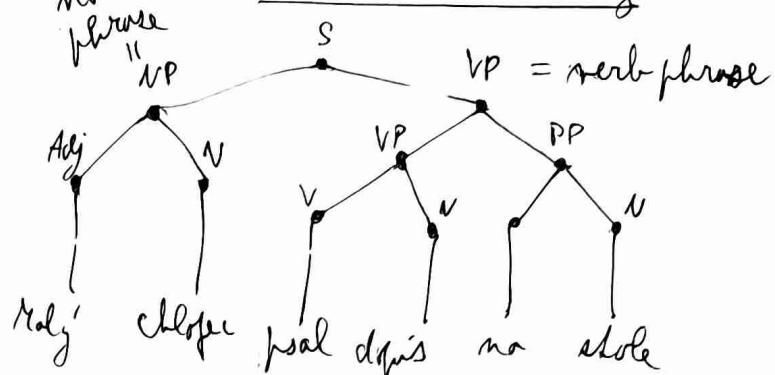
závislostní stromy



- dobré reprezentuje vztahy
- strom nevyžaduje poskyt výpočtu
⇒ nemá jasné, jak ho řešit
- nemá jednoznačný
- nevšechny vztahy se něčem
se zohledí dají popsat →

Petra Pavel

složkové stromy

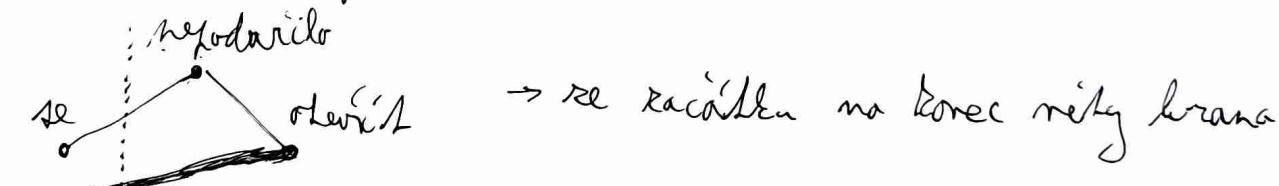


- derivacní strom kontekstové gramotnosti
- méně přehledný, slyšitelně vzdálen
- převrácení jazyk nemá kontekstový
- spojuje slovní druh do fráze

$$V + N \rightarrow VP$$

Neprojektivní konstrukce

= dloně brány ... za brana → závislostním stromu jde dopřejc
víta: Sonbor se nepodařilo okénka



→ neprojektivní = vertikálně má 2 průseče
→ složkový strom má problém : by slova spolu nesouhodí

Kontekstová gramotnost

X

Derivativní gramotnost

- nejádá pravidla
- nevíte by měl byť
formět a přísudek

- pravidla na spojování enclitics
- dokle terminální znaky

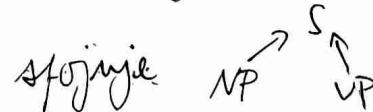
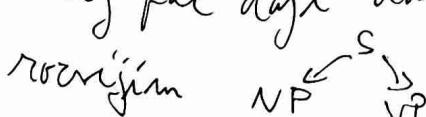
Generativní

X

Analyza

- myroba stromu, drážekho se na
značky pak dají doradit slova

- z něčí délka strom



Transformační gramatika - Noam Chomsky ~ 1965

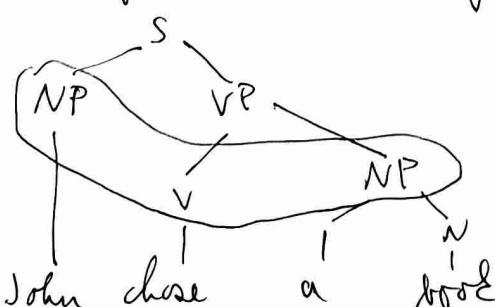
- převratný způsob formálního popisu přirozeného jazyka
- princip: kontextuálně vlnitá syntaktická struktura
 - parckary sstrom - Elastický složkohý strom
 - blonking strom - zadává pouze důležité rady
 - ↳ změna pořadí slov ho neuním
- komponenty
 - báky - kontextuální pravidla pro generaci složkých stromů
 - = frázové ukratky
 - Transformační L. - Transf. pravidla operující na frázových ukratkách
 - spojování / mísí až shromáždily
 - pravidly obligatorické
 - fokalizace
 - fonologická L. - regexy co dávají režim morfémna fonek. intonač.
- Generační procedura - jak generovat stromy

$$VP \rightarrow \left\{ \begin{array}{l} \text{mínimál 1} \\ \text{mínimál 2} \end{array} \right\} / NP$$

→ min. 1 je přesně jen tedy predikát NP

→ nedokáže zachytit různé varianty vět → syntaxická ornamenace

→ transf. složka má pravidla jak k fokalizacím ch
fon. ukratku vytváří funkce struktury věty



$$NP_1 - V - NP_2 \Rightarrow NP_2 - was - V + - en - by + NP_1$$

Book was chosen by John

- Ta teorie se hojně měnila, častož vydával další verze
- 1965 - Standard Theory - dobré
 - 1990 - Teorie minimalistické - pouze základ logické funkce a funkční konstrukce

Tree Adjoining Grammars - subtree sharing

- stromecíky jako ~~el.~~ struktury
 - řífka & rovní možnost substituční
 - strom ~ syn. struktura výšky

```

graph TD
    S --- N1[John]
    S --- VP[VP]
    VP --- V[loves]
    VP --- NP2[NP]
    NP2 --- N2[Mary]

```

The diagram shows a mic substitutional transformation. The root node S branches into N (John) and VP . The VP node branches into V (loves) and NP . The NP node branches into N (Mary).

Logical Functional Grammar

- C-structures ~ språkain' slav do frås
f-structures ~ referensiell funktion' voktby ve veile (vokt slaves)

↳ notice	[al ri 'br 'sy]	hov du nr sy]
----------	--------------------------	------------------------

→ hoofdzaar waarbij i jiná f-structura

• C-structure má pláne 1 f-structure

ale 1 f.s. misse figs nemme c.s.

Kategorialná gramatika

Kategorie slow dastane kategorii,

→ Kategorie fysieke symbolische relaties sehr schwer

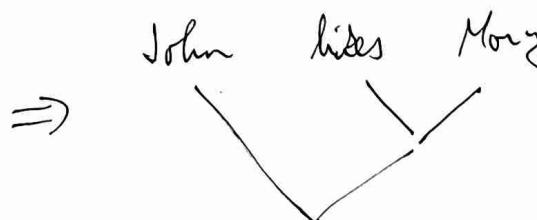
→ obecny formal 2/3 nebo 3/2 ... komisie rika vlovo (opera)

- likes má' kategórii ($NP \backslash S$)/ NP \rightarrow díra na névo agym
 - 2 hiberní formilla: $X / Y + Y \rightarrow X$, $Y + Y \backslash X \rightarrow X$

```

graph TD
    NP_S[NP / S] --> NP1[ ]
    NP_S --> NP2["(NP/S) \ NP"]
    NP_S --> NP3[ ]
    NP1 --> John[John]
    NP2 --> likes[likes]
    NP3 --> Mary[Mary]

```



- ④ neplatí vždy obecná gramatická pravidla → správna rýma
 - ⑤ strasné moc kategórií

Unifikácia gramatiky

- popis vlastností díječku

→ objekt ~ možná vlastnosti (marker vč: hodnota vč.) = sestava rysu

→ sestavy rysu se dají unifikovať

↳ tře to, když vlastnosti 1. sestavy neodpovídají sestavě 2.

→ sestava rysu něčím popisuje nějaký sign. znam. (shoda f. v.)

→ hodnota vlastnosti může být dletoč sestava neb počítat na ni

Shoda funkček + případkem

$$\hookrightarrow \begin{bmatrix} \text{podmínka:} & \begin{bmatrix} \text{roba 2} \\ \text{rod ž} \end{bmatrix} = 1 \\ \text{případek:} & 1 \end{bmatrix}$$

Problém: tře unifikovat nesouvisející rysy

→ sestavy mají syst → ten určuje ježí vlastnosti

↳ Tyto samé sestavy rysu

↳ např: sloveso, funkce, jméno, ...

HPSG

- kombinuje principy gramatiky pravidla a slovníkové pravidly

- slouží k tomu strukturování, spracování info

- el. jednotka = znak (sign)

- slova a fráze jsou ručně podloženy znaku

→ slova má 2 rysy

- fonetický ... znak

- syntaktické a semantické informace

Nástroje na syntaktickou analýzu

Augmented Transition Networks

- vymírájí slov' fraze co potřebuje

$$S \rightarrow NP \cup VP$$

$$NP \rightarrow Det \cup N \quad \dots \text{Det} = \text{člen "a", "the"}$$

$$VP \rightarrow V [NP]$$

The girl saw a boy ... kožen = S a S potřebuje NP a VP

\Rightarrow SEEK NP ... NP potřebuje Det a N

\hookleftarrow CAT Det ... Pokud následuje Det, jdi dle dalšího stromu
CAT N ... Pokud následuje N, jdi dle dalšího stromu

\Rightarrow SEEK VP ... potřebuje V a následně NP

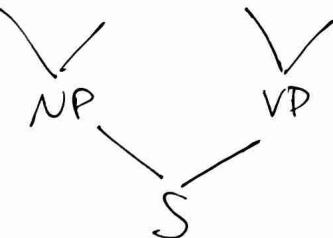
\hookleftarrow CAT V ... pokud následuje V, horeací

SEEK NP

JUMP ... přejdi dle dalšího stromu aniž bys měl zhlédnout

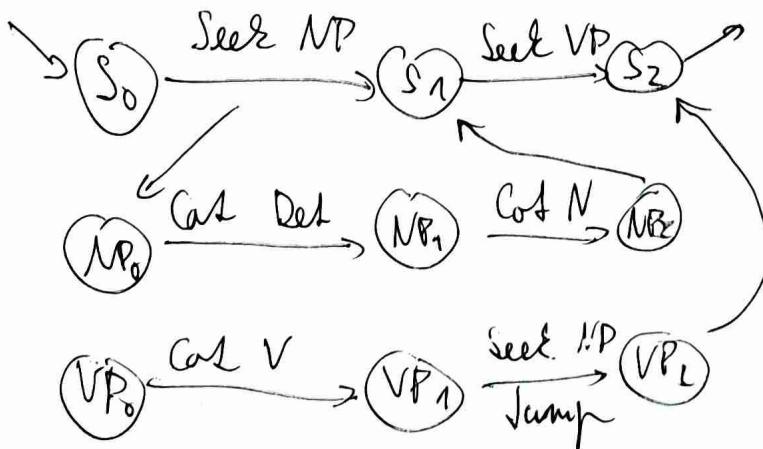
The girl saw a boy

Det N V in NP



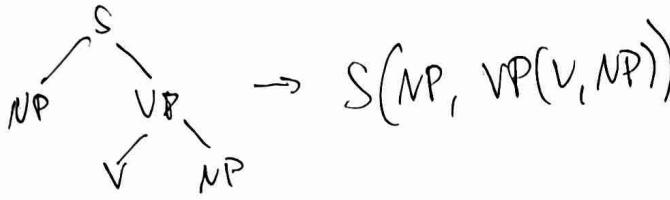
I. Synt. bran

- SEEK = přechod dle pravidla
- CAT = přechod dle dalšího stromu
Pokud najde co hledáš
- JUMP = přechod dle bez hledání



Q-Systém

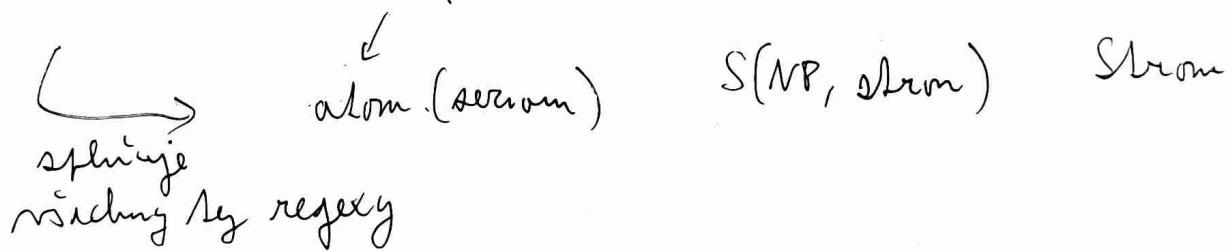
- formolismus pro vyslovaci grafu
- strong vlastido linearne



3 typy objektu

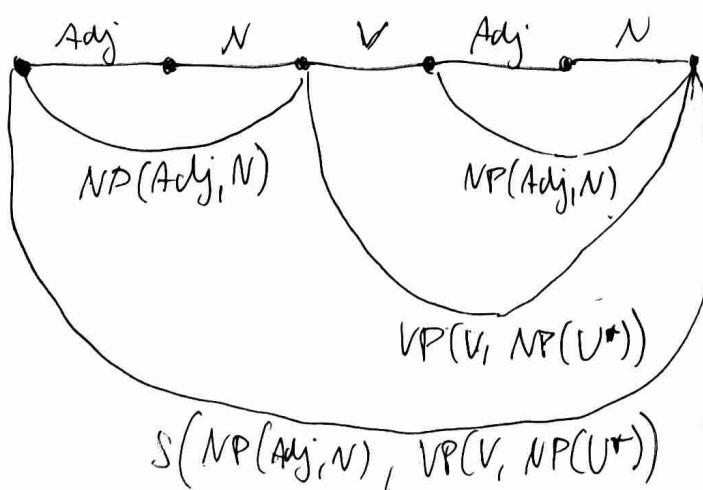
- atom = konstanty ... písmenka A - Z
- strong = strong ... písmenka L - N
- serny strum ... písmenka V - Z
- + má nějaké operátory na nich objektech
- zapisuje strom jinou reprezentací
- *. = proměnná

$$S(NP, VP(V, NP)) \sim A^*(V^*) \text{ nebo } S(NP, L^*) \text{ či } N^*$$



pravidla

- Adj + N → NP(Adj, N)
- V + NP(V*) → VP(V, NP(V*))
- NP(V*) + VP(V*) → S(NP(V*), VP(V*))



ten výsledek můžeme poslat do dalšího Q-Systému

- číslování stran

- odstranění neschybnejší brany - dílčí mezinásobek
- vždy upříoji brany ... písmenka / výrazy = brany (na obou stranách)
- 1. odstranění vždy brany, co jsou psány na levé straně nejdeho pravidla
- 2. odstranění slevy někdy

Funkce generativního popisu - množné projekty

- stratifikacíová teorie - 5 rovin

- jednotka na výšším rovině reprezentuje funkci
jednotky na nižší rovině

- teorie valence

↳ varby jsou využívány zdroje proložení různými slovy

Telogramatická TG
→ Porchová
→ Hofmystická
→ Fonologická
→ Fonetická

- TG rovina zachycuje 3r funkce

→ kniha byla vyd. vyd. = vyd. vyd. knihy

→ 2 druhý člen objekt domén

• Aktanty: konzult (agens), Příjemec, Adresát, Origo, Efekt
↳ je možné být v 1 řeči jen 1x

• volná doplnění: libidne vlivná

→ členy dle téma

• obligatorní - nemusí na TG chybět

↳ podmínka tam musí být, i když na funkci roviny může chybět

• fakultativní

→ obligatorní členy jsou funkce sloužit jiné!

→ Moji přátele přijeli

• kam? ... kde je důležité ⇒ obligatorní

• odkud? ... fakultativní

Valenciový rámcem = sečením aktantu (i fakultativních)
a obligatorních volných doplňků

Kontrola gramatickej správnosti

Takínek ňe do práce. ← Co do kam dnes?

- problemy ~ výzviny

- neprav. konštrukcie, shoda form. s príčekou
- interpunkcia + rájmena mē / mne

• chybore' roviny

- prav. jazyk s pravým slovopredelom
 - ↳ chybore' konštrukcie sú blisko u sebe

• gramatika

- kontrola podľa pravidel
- nelič rovinnas, rda je sú chyba,
nebo rda je naše gramatika neplna'

• Metoda redukčnej analýzy

- vieda redukciu tiež, že výzvy odstraním niečo, čo je 100% správny
⇒ neviem chybnou viedu opraviť ani správnu robiť

1. Nasí hokejiste' viera zo viede' súve opäť vyhraly.

2. hokejiste' viera zo snaze vyhraly.

3. hokejiste' viera vyhraly.

4. hokejiste' vyhraly.

→ napr. velké snaze, ... zdroby pôvodstva ... kontrola shody
... je sú dobré → odstraniť

⇒ chybu odstrana blisko & súbe ⇒ chybore' roviny

- musí se obehovať speciálne prípady

výsledok: mínimálna chybore' konfigurácia ... vysvetlana vieda

! funké hľadanie chybore' roviny slovitejších konštrukcií nefunguje

> Které dielčia chýbajú dočasť návinie? *

↳ min. chyb. konf., ale dlonha brana

} 2 prístupy

! homonymie ... číselník jaro organ x číselník jaro mezikáři
rod ž rod m

RFGODG = Robust Free-Order Dependency Grammar

- 1 gramatické pravidlo může popisovat správnou i chybnou konstrukci různě
 - výsledek ne je řízen, interpret gramaticky rozhoduje, jak se bude stýkat gramatické pravidlo s jazykem
 - příklad pravidla: pro podřízení A je 7. fází shoda shodného pravidla
- fáze: positionální: - trvalé forminky → musí se splnit
negativní: - mikré forminky → nemusí se splnit
projektivní = vícenásobně sebe

→ fáze: z fáze následují typicky následující les

1, positionální projektivní

- pokud je ta věta správná a bez neg. kroků sice je všechno celé slovo

2, neg. proj. nebo froz., neg. proj.

3, neg. neg. proj.

- když máte fokalická všechno

→ výhoda je možnost struktury, jak to může být

* shlyky a násilnictví → negativní struktura

↓ struktura s chybou posouzenou

• Lan GR

- redukuje snadky pro morf. analyzu \rightarrow disambiguace
- fuzízová a negativní disambiguace formulu
- formulu mohou mít několik kontextů
- jsou vždyco nejvíce, nezávislé a jsou uplatňovány v cyklu
- \rightarrow části: kontext, desabiguace části, report; ale
 \Rightarrow cont, disamb₁ cont₂ disamb₂... cont_n disamb_n cont_{n+1} report action
- \Rightarrow odhalí podletole' výzvy - nějaké slovo na konci mnoha rashnov ruky

Korpusová lingvistika

korpus = velké množství textů s nějakou přidánou informací
(např. morf. nebo syn. analýza)

• Studie struktury jazyka

- \rightarrow identifikace jednotek a tříd jazyka (morfing, slova, fraze...)
a tyto jednotky se mohou kombinovat na velké

• Studie významu jazyka

- \rightarrow když máme hodně rámcového významu vybraných textových dílčích
tak mi to dává nějakou objektivní informaci o tom,
jak se ten jazyk reálně používá

- \rightarrow měl by to být reprezentativní význam jazyka

• výběr významu a reprezentativnost

- jazyk je neonejednoznačný, ale korpusy konečné
 \Rightarrow význam musí být nějak vyrovnán

• konečná velikost

- s rozšířením monitorovacích korpusů, kam se počítají nově přidávané
mají korpusy fixní velikost
- \rightarrow jsemlik se jednou vydá, než se změní

• strojení i jeho forma

- měly by se dodržovat nějaké standardy - UD

Brown Corpus of Standard American English

- fo moderní elektronický korpus
- ~1 milion slov (textů) v americké angličtině vystřílených v roce 1961
- náhodný vzětí z různých, reprezentativních
- morfologické knacky

Penn Treebank

- 1. a nejznámější syntakticky anotovaný korpus
- ~1 milion slov
- články z Wall Street Journal
- ⇒ uplně & mimoř., je to barevný slang, vzhledem k rozdílu anglicky

• British National Corpus ~ 100 mil. slov, morfologická anotace

• American National Corpus ~ 22 mil. slov, morfologická anotace

• Corpus of Contemporary American English ~ 910 mil. slov →
↳ 20 mil. slov za každý rok mezi lety 1990 a 2010

• Oxford English Corpus ~ 2 miliardy slov →
↳ gramatické výroby Anglie nejakej

• Negra Corpus - německá - syntaktická anotace

• EUROPARL - 11 evropských jazyků, ~80 mil. slov / jazyk

↳ jednoují paralelní korpus

→ dobré pro srovnání překladů

→ díky tomu lze překlad sestavit pro všechny jazyky, ne jenom alež
↳ nezávisle na kontextu

- Český národní korpus - VK, MUNI, Ústav pro jazyk český
 - morfologická anotace
 - dneska ~ 5 miliard slov
 - když někdo bude vydávat nový korpus se sloužit na něm může být i s tím
 - ↳ může být SYN2020 = výsledek 2020, slova z 2015-2019

• Právěký rámcový korpus

- chtěli syntaktickou anotaci jako TreeBank
- Secretory rámcového je Funkční generativní jazyk FGP
- 100 000 vět ~ 1.25 mil slov → malá funkcionálnost ČNK
- anotace
 - morfologická
 - syntaktická pomocí rámcových struktur
 - ↳ chtěli udržet, že jsou lepší než složitější
- FGP ⇒ sebe 5 rovin
 - ↳ anotace různých jazykůch detailů
 - ↳ v TG rovině se kromě struktury vlastní i různé vzdálenostní reference, např. naco od rozdílu každém

• Prague Arabic Dependency Treebank

- ⇒ aby dokázali, že rámcový korpus je super, tak pomocí svého události korpus arabského

• Universal Dependencies

- dřív různé jazyky anotovaly stejně konstrukce různě
 - ⇒ snaha anotovat stejně různy různobojich jazykůch stejně
 - kvůli tomu vzniklo experimentální rámcový korpusy, ale když jinak
 - PEK byl moc detailní → UD výchozí schéma
 - dneska se tisknou pouze UD
- LINDAT ← všechny korpusy Jana Sam

Semantika

- syntaxe je v říčce, jestliže je gramaticky OK
 - semantika říčce, jestliže dává smysl

Když obdobněk zelení mosírovou kouří mohou mít různý význam

 - význam ≠ producents
 - ↳ nepravidelná sdílení mají svýj význam
 - ↳ ne vždy je možné většinu producents
 - nemají vždy jasné, kdo jde o věty se stejným významem
 - Pozornali ho dle vzhledu
 - Byl jimi pozornán dle vzhledu
 - vyplývání - producent věty mají dle vzhledu
 - ↳ ! Animaci jsem počítal \Rightarrow mají křídla a letojí
 - formální jazyk \times fiktivní jazyk

Freyd's principle of compositionality

- význam slovnického výrazu je ovlivněn myšlenkou jeho částí a spůsobem jejich kombinací.

• Lexikalni semantiken

Přirozený - stejný nebo jiný - jeho reálném - lepsi'

! mykhan reidy rainisi' nma korkelam

→ Problem: na "rysuémém" mohou byt potíže dleží neznáma slova

Redit: na počítač se přidalo Nášbor jen 2000 nejběžnějších slov jazyka

🕒 významy slov je možné rozdělit na kontextuální pojistit
formu významůvých kódů

- Ontologie = množina kódů objektů na klasifikaci objektů univerza

objekty ∈ příroda ∈ kachynské maškary ∈ maškary ∈ fyzické objekty

→ reální seřídké → několik druhů

→ problém, že význam slov není jednoznačný
↳ lehota, hlaška

} krankosti
reality
větování
vlastnosti
:

ontologie ↗ nechávej - obecné významy

↓ důležité - mimo specifického

↳ struktura ontologie vlastnosti

- Semantické sítě

+ příloha

- WordNet → databáze anglických fráziček a přidružených jmen. a sloves

→ dletem na výzvy ekvivalence = Synonymy

↳ f. synony vyjadřuje vztahy Koncept

→ rámec mimo synonymy - semantický a lexicální význam

→ slovo může patřit do několika synonym

~ 155 000 hesel

~ 117 000 synonym

- Euro WordNet

- měkké jazyky

→ nechávej ontologie - 63 nejdůležitějších jazykově závislostí konceptů

→ méně základních konceptů - 1000 základních konceptů

↳ tvoří jazykovou síť slov, jazykové závislosti

- dobré pro autoritativní přehled

- dobré pro práci se semantickými vztahy - hlavní synonymy

Reprezentace významu něky

Predikátová logika 1. rádu

- jakoby se fogyje, ale něky nejsou formule

... nevlastní tři možné pravdivosti

↳ possible (F), believe (x, F), true-at-some-point-in-future (F)

Presupozice

= předpoklad, co musí být pravidlo, aby věta mohla mít pravdiv. hodnotu

"Jupiter má rovné prahy."

↳ Jupiter musí mít první 7 měsíců

• nemocnost = Falseness - T/F nesouhlasí ... Paralel je myšlení.

Existence or intenčnost

- Intenčnost = definice

- Existence = reči, co splňují danou definici

• Cína Big Mac je 20 Kč.

Nefrakce
||

polení do vztahům za intenčnost → 125 Kč je 20 Kč.

• Myslím, že cena Big Mac je 20 Kč.

↳ nemí ekvivalentní k "Myslím, že 125 Kč je 20 Kč."

Příslušny k semantice

světu

1) Modelově-theoretická semantika

↑

pravidelní funkce jsou vztaheny k konkrétním modelům

→ např. montagueova gramatika

2) Kompozičních semantika

- výchozí z principu kompozičnosti = význam složeného výroku
závisí na význame jeho částí a na spjatých významech

• Montagueova gramatika

- rádkov je funkční logika
- Montague byl přesvědčen, že semantika přirozených jazyků se mohou vypočítat
- além: semantická formula je tak strukturalně se syntaktickými
- do jisté míry se mu to podařilo
- formula připomínají kategorialní gramatiku
- ↳ slovesa co se chvojují podobně se seskupují do kategorií a rachácky se s nimi stejně

• TIL = Transfenzivní intencionální logika

- logika 1. řádu věsticí
- vyhodnocuje formou semantického množinového sčítání
 - ↳ věta může mít více významů v různých sčítacích
- totéž zohledňuje čas, když se to vzdělává
- formální lambda kalkul - náročná logika
- $w = \text{sčítka}$
- $T = \text{čas}$
- $\Delta w \sqcup T (\vdash_{\omega T}^{\star} F_{\omega T} : H_{\omega T})$
- ↳ tohle pojíždí všechny věci

→ kromě množinového zohledňuje i věta individuální

- studentka Alena - nejlepší individuální
- ministr zahraničí - individuální koncept
- běžící metr - všechny metry jsou individuální
- ministr financí

→ věta může mít více významů v důsledku Aleny si myslí, že MF \geq MZ

↳ je to nejlepší objekt $O_{\omega T}$

↳ když máme sčítka a času přirozeného množinového (jednotlivé) počet

Rozdílnostní vztahy v textu - Anafory

→ odvozování rájmena - problém při překladu

↳ všechny názvy se mohou překládat na { množ: m → he
sklid: m → is

⇒ potřebujeme uvedít název na co rájmeno odvozuje

Anafora = rájmena jež interpretace závisí na kontextu

dlelem:

• Exofora = odvozování mimo text. → "Viděl ho?"

• Endofora = odvozování v rámci textu

• Anafora (zpětné) ... to, fakt se to jenž objeví

↳ "Petru vyrobil Rajenští. To neměl diktat"

• Kofora (dopředu): "Když se sloví, není s Petrem rádha' řeč..."

• Anafora - všichni bláznové ona, rebab předchůdce & následník

1) Zájmena a nevyjadřené větve čluz (franšíz)

Petr si komplik vstupoval. Pot ji dr žasy. Byla divná.

2) Vrácení jmeny shaping

Elektrový zasilovač Tesla 1000. Toto rázivě...

3) Elipsa

Petr přinesl dva stoly. Převlék a krovny.

Petra pojde do kina. Sírka stoly.

4) Spojovací výrazy

například, jednotk - druhak, nejdříve - potom

⇒ velmi obtížné: p třeba aplikovat několik různých

Využití:

• rájmena informací v textu - MOSAIKA

• automaticky překlad - správný překladem rájmen "m" → is/he

• dialogové systémy

Rézimí anafory

1) morfologické znaky - mpx. v rámci shoda v rodi

2) syntaktická struktura výzvy

↳ pro určení možných kandidátů na předchádce

3) statistické metody

4) abstraktní členění - viz alg. schéma znaků

5) rozšířené použití znaků - ontologie, semantické sítě, ...

Zásoba schílených enalogií

- modeluje zásobu znaků, co mluví předpoklad, že schéma je shodné
↳ mezi se prodele hovor, co je správná r, centru povnosti

- jednoduchá pravidla

→ věta má nějaké ohnisko a jádro

→ aⁿ znaková, že a má dimenznost n

příklad pravidel:

pokud referuje přes rájmeno: n se nemění

pokud a je v oblasti: aⁿ → a⁰

pokud a je v jádru: aⁿ → a¹

pokud aⁿ → a^m: objekty asociovány s a obdobně buď a^{m+2}

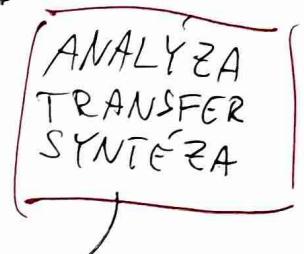
pokud a nemá smysl ani asociovat: aⁿ → a^{m+2}

Slovojazyčný překlad

- problemy:
- potřebujeme znat náplň jazyku, nejdé překlady slova od slova
 - pravidla propisů: jeden jazyk = čízky
 - ustálená sfinčení: všechny maz = mazehn
 - ne vždy je možné namapovat slova mezi jazyky 1-1
 - homonyma
 - slovenčiny výraz: airport long term car park courtesy vehicle pickup point

Metody první generace

- dvojjazyčný slovník: překlad slova od slova
- ve slovníku nejsou všechna slova, ale morfem
- freedicting → překlad → postdicting
- jeden jazyk se zrovnaže pravidly pro překlad do druhého jazyka?
- Georgelovský experiment 1958
 - velmi skrblivé ořezení domén a používajících jazyků (překladač)
 - ale fungovalo to
- 1955-1965 bouřlivý rozvoj
- 1966 Zpráva ALPAC
 - je nutné dle obdobobějšího instinktu teoretičkého hing. výkladu
 - obnášedlo: římské podoby v USA
 - ↳ ne Francii, SSSR, Kanadě ani podobně



- TAVM-METO - překlad meteorologických zpráv A → F
 - ↳ dobrě definovaná a hustá sít mezi překládajícími jazyky
 - ↳ použitá Q-Systemy
- SYSTRAN - překlad dokumentů EU mezi ~20 jazyků → fungování A-F-N
- EUROTRA - projekt EU, megolomacie: 72 jazyků - negativní efekt pro ALPAC
- VERBMOBIL - překlad mluvené řeči, nájemce mluvíce EUROTRE
 - ↳ sít mezi řečemi: plánované překlady schvály drom ochranných

- S trojím podporaným překladem = CAT
 - komerčné náhledy nejsou jisté ani autorizované - už od 80.-let
 - = překladová řízení
 - ↳ obsahuje páry překladeckých segmentů - typicky vět
 - + metadate, minimálně efektivní správna funkci
 - nové vstupy vět se mohou dát do funkce
 - funkce sami jsou, překladačel dostane na vše překlady se svým vlastním oddělením a procesy probíhají
 - ⇒ vhodné pro opravu překlady mnoha aktuálních textů - technika může být mnoho různých

- 2 přístupy

1) překlad věta po větě

- překladačel má kontext toho co překládá
- když výsledek se musí překládat znova
- IBM Translation Manager

2) překladová Atalka

→ pokračující překlady

- kontext nemá vliv, ale tří věta se překládají jen jednom
- Déjà Vu

Cesté překladače

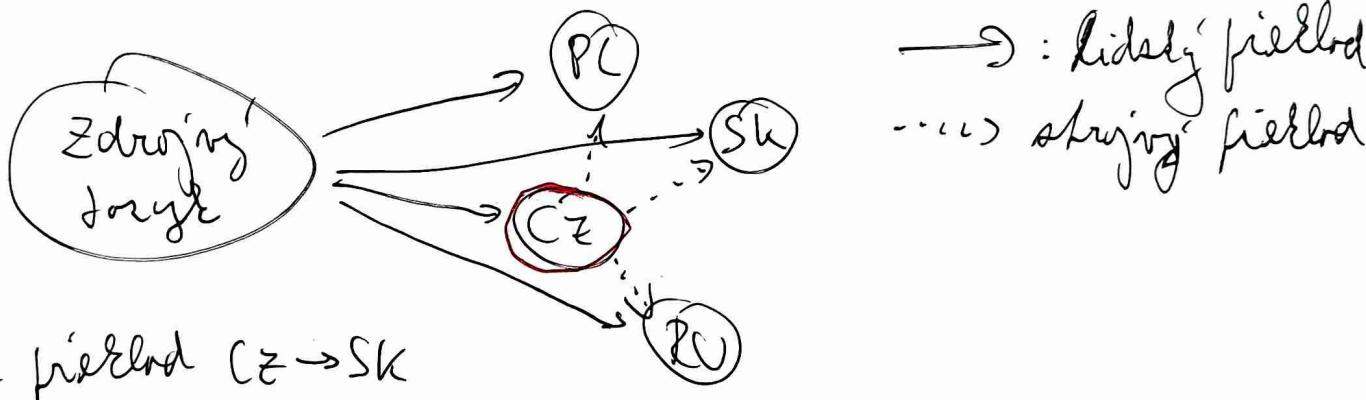
RUSLAN - překlad mnoha milionů fragmentů ~ 1985

- 1 věta ~ 4 minuty
- dvojjazyčný slovník, klasický přístup: analýza, transfer, syntéza
- Q-systemy
- transduktivní slovník

↳ některá slova s recto-rotundým základem je možné převzít pravou
 -ace → -acia: industrializare → industrializacia

• Systém Česillo

- lokalizace velkých sw. systému
- nahrazena lokalizací r. jazyka sýmbolem překladem r. jazyka během
- překládá se mezi velmi blízkými jazyky - vysoké "vodivé" překlad
- slovíčkové znacitelní cíistiky
hlavní autentický
- dvojjazyčný shéf



- překlad CZ → SK

Cs stejná symboly

→ většinou shodné pořadí slov neříše

→ různé slovní a morfológie

• PC Translator 2003

- tradiční překlad, novější překlady
- asi nejlepší česky domácí systém

Složitý strojní řešení

- věci se dojí řešením mle řípnutí : řípnutí $\text{im}(EN) \rightarrow F$
- relatiní cestou má řešení
- Bayesov řešení $P(x|y) = \frac{P(y|x) P(x)}{P(y)}$

$$P(A \cap B) = P(A|B) \cdot P(B)$$

$$P(A \cap B \cap C) = P(A|B \cap C) \cdot P(B|C) \cdot P(C)$$

Model námi řezen

úkol: řešení řeči dle řeči v běžném textu

→ na základě historie h řešení řeči dle řeči : $P(w|h)$

cíl: specifický řešení řeči $P(\text{řeš}|w) = P(w_1 \cap w_2 \cap \dots \cap w_n)$

N-gram:

$$\begin{aligned} P(\text{n-řeš}) &= P(w_1 \cap \dots \cap w_n) = P(w_n | w_1 \cap \dots \cap w_{n-1}) \cdot P(w_{n-1} | w_1 \cap \dots \cap w_{n-2}) \\ &\quad \cdot P(w_{n-2} | w_1 \cap \dots \cap w_{n-3}) \cdot \dots \cdot P(w_2 | w_1) \cdot P(w_1) \end{aligned}$$

problem = složitá historie \Rightarrow fakticky obecně

→ reálné řešení řeči bigramy, předchozí bigramy

$$P(W) = P(w_1 | w_2 \cap w_3) \cdot P(w_2 | w_1) \cdot P(w_3)$$

Vyhlearání



- problem je velkost dat

→ slovník V , $|V| = 40\,000$ slov \Rightarrow velikost modelu $|V|^2 = 6,4 \cdot 10^{13}$

↳ standardné řešení řeči na $\sim 10^8$ řešení

\Rightarrow skutečně moc nulových řešení = řešení řeči

řešení: nahradit nulovou řešení řeči malou řečí řeči

↳ řešení řeči nulovou řečí řeči neexistující kombinací

Princip statistického překladu

- paraleký korpus = trenovací množina dobie překlazujích něk
- vede nedávna překládající metoda, dneska spíš novinky

Metoda racionálního kanónu

→ čereme překlady $F \rightarrow A$

⇒ hledáme model $P(a|f)$, což znamená f a a můžou být různé

$$P(a|f) = \frac{P(f|a) \cdot P(a)}{P(f)}$$

→ $P(f)$ je stejná pro f a

⇒ obecně jsme de-facto smír překladu

⇒ hledáme 2 modely: $P(f|a)$, $P(a)$

→ vlasti překládání, tedy jsme doslova něčím překlazem z A do F a hledáme její správný original

→ fázkový model $P(a)$ může být trigramový model

založený na mnohem rozsáhlějším korpusu až do fázky

→ překladový model $P(f|a)$ je založený na paralekém korpusu

(s mnohem menší)

podstata:

- překladový model buduje v oznámeném směru

- fázkový model odfiltruje nepraktické překlady a vybrá dleby překladního modelu

→ fázová, jestliže anglická verze dává smysl

→ vybírá totiž formu "herce" někdy, nemá vzhled originální

- hledání překladových hypotéz (deklonací) je totiž svý problem

dostat něčemu

- mám na výber s mých překladech
- pro f specifický $P(a)$

} závislost → max

• Evaluace systémů automatického překladu

→ metrika BLUE

- přesnost překladu v m-gramech

⇒ máme 2 kandidáty na překlad ... tedy je lepší

⇒ pro tu nebo máme rozsáhlé měřit správných překladů

• unigramová přesnost

$$= \frac{C}{N}, \quad N = \# \text{ slov v daném kandidátu}$$

$C = \# \text{ slov z kandidáta, co se vyskytly}$

→ užívání jednom referenčním překladu

• m-gramová přesnost

→ dvojice, trojice ...

⇒ $N = \# m\text{-tic v daném kandidátu}$ (co má sebe násobit)

$C = \# m\text{-tic z kandidáta co jsou v nějakém referenčním překladu}$

• BLUE = sítové score

• penalizace za skríniny = Brevity penalty (BP)

- pokud bychom brali v úvahu jen $\# m\text{-gramů}$,
akdyž by to favorizovalo krátké výrazy, pejsek
m-gramy by byly v překladech, i když by
byly překlady reálně lepší byly mnohem lepší

$$\text{BLUE} = \text{DP} \cdot \sqrt[4]{p_1 \cdot p_2 \cdot p_3 \cdot p_4} \in [0, 1]$$

↳ G. průměr m-gramové přesnosti pro $m=1, 2, 3, 4$

(+) - rychlý výpočet, relativně objektivní míra

- obecně přijímaný standard

(-) - následný dostatečně rozsáhlý referenční překlad je druhé

- favorizuje statistické systémy, co se náleží m' také'

- špatně zachází s různými varianty - slvozdroje, morfologie

- přecenívá se, nemá až tak univerzální