Úvod do počítačové lingvistiky

- Počítačová lingvistika = obor, jehož cílem je popsat vlastnosti přirozeného jazyka pomocí matematiky
  PL
  → cíl: pochopit jak přirozené jazyky „fungují"
- Počítačové zpracování přirozeného jazyka = hlavně aplikace metod strojového učení na velká jazyková data (korpusy)
  NLP
  ⇒ nic nám to neříká o funkci toho jazyka
- Oblasti PL
  - Morfologie
  - Syntaxe
  - Sémantika
  - Formalismy
  - Korpusová lingvistika
  - Statistická lingvistika
  - Strojový překlad
  - Rozpoznávání a generování mluvené řeči
  - Vyhledávání v textu
  - Dialogové systémy
- Problémy
  - přirozený jazyk je víceznačný
  - slovo nevyjadřuje svůj číslo, rod, pád... jednoznačně

• Morfologie

= studium vnitřní struktury slov

morfém = nejmenší znaková jednotka jazyka nesoucí význam

kmen
za - hrad - ou = přípona

předpona / prefix, lexikální, gramatický morfém

↳ určuje 3 sémata: pád, číslo a rod

- skloňování = deklinace
- časování = konjugace

Problémy

- tvaroslovné dublety = jiné tvary, stejný význam: na hradě / hradu
- homonymní tvary = stejný tvar, více základů: stát (země, sloveso, sloveso), žena (podst. jména, přechodník)
- alternace = random změna hlásky; určití kmene: vůz → vozu; matka → matce, mateřin, matek
- plnovýznamová (byl) × pomocná slovesa (jsem doma)

Morfologické typy jazyků

- Analytické: slovo = morfém → čínština
- Syntetické: slovo > morfém → slovanské jazyky
- Polysyntetické: slovo = věta → indiánské jazyky

Přístupy zpracování morfologie

→ základ jsou buď

- morfémy: slovo = řetězec morfémů
- lexémy: slovo = výsledek aplikace pravidel, co mění základ a vytváří slovní tvar (angličtina → breaks, broke, broken = 1 lexém)
- slova: hlavní roli hrají vzory

→ když mám základní tvar + vzor, tak umím vygenerovat ostatní tvary podle toho vzoru

→ tohle se reálně používá

→ v češtině je asi 250 jednoznačných vzorů

- **Two-Level Morphology** - 1980
  - první obecný model zpracování přirozeného jazyka
  - V jazyk svůj slovník a pravidla, ale mechanismus morfologie obecný
  - 2 úrovně
    - lexikální
    - povrchová
  - → pravidla se uplatňují paralelně, nikoli sekvenčně
  - → podmínky se mohou vztahovat k 1 z úrovní nebo k oběma zároveň

    lexikální rovina ⟷ mezi rovina ⟷ povrchová rovina
  - → lexikální vyhledávání v trii a morfologická analýza probíhají současně

Česká morfologie

- **poziční značky**, každá pozice ve značce má nějaký význam
  - → 13 kategorií: slovní druh, rod, číslo, osoba, čas, ...

lemma = jednoznačný identifikátor daného slova = základní tvar/kořen + index

p.j. stát → stát-1

- **Morfologická analýza**: dostane slovo
  - → vyrobí seznam lemmat a značek, která popisují jednotlivé možnosti, co to slovo může znamenat
  - → např. pokud více pádů má stejný tvar ⇒ více značek
  - → pokud to slovo má více významů → více lemmat
    - + k lemma má nejaká jiná pravidla na dělání značek
  - → těch labelů může být spousta
- **Morfologické značkování** = tagging
  - → z těch mnoha značek vybíráme tu 1 správnou v daném kontextu
  - → statistické metody (značíme se to)
- **Částečná morfologická disambiguace** - Oliva
  - podle gramatických pravidel jsou některé konfigurace nelegální
    - ↳ shoda podmětu s přísudkem, shoda přívlastek
    - ↳ 7. pád po předložce s
  - → odebírají pouze ty značky, které jsou 100% špatně

• Lemmatizace
- proces výběru správného základního tvaru

• Stemming
- odříznutí koncovky
- na rozdíl od lemmatizace je základním tvarem kmen slova

• Generování
- známe lemma a kombinaci gramatických kategorií
⇒ chceme správný slovní tvar

• Značkování
- hledáme nejpravděpodobnější posloupnost značek pro celou větu
→ využíváme Bayesův vzorec $P(x|y) = \frac{P(y|x)\,P(x)}{P(y)}$

$P(\text{značka}|\text{slovo}) = P(\text{slovo}|\text{značka}) \cdot \frac{P(\text{značka})}{P(\text{slovo})}$, ale $P(\text{slovo})$ je pro všechny značky stejné

⇒ chci to maximalizovat

→ čím více po sobě jdoucích slov a jejich značek beru v úvahu, tím lepší, ale je to drahé a nejsou data

⇒ závislost jsou na jeho značce a počtu výskytu této značky na poslední značce

→ $\text{nej značka} = \underset{\text{značky}}{\operatorname{argmax}} \prod_{i=1}^{n} P(\text{slovo}_i|\text{značka}_i) \cdot P(\text{značka}_i|\text{značka}_{i-1})$

• Skryté markovovy modely HMM
- metoda analýzy řad (posloupností událostí v čase)
→ my to aplikujeme na posloupnost morf. značek v textu
→ je to posloupnost rozhodnutí, co na sobě nějak závisí (gramat. pravidla)
→ Markovova hypotéza: kontext = # po sobě jdoucích věcí, co spolu souvisí, je možné zkrátit na nějakou pevnou délku
→ kontext délky 2 = bigramy, délky 3 = trigramy
- Skryté, protože některé vlastnosti té posloupnosti nejsou nebo nelze vidět – jsou tam slova, ne značky
- je to basically stochastický konečný automat
↳ náhodný

Jak se HMM používají

1) Rozpoznávání (ohodnocení) statistického modelu
   → jsou dány parametry HMM, cíl je spočítat pst, že je pozorována posloupnost X.
   → použití: rozpoznávání obrázků – registračních značek aut

2) Rekódování
   = hledání nejpravděpodobnější posloupnosti skrytých stavů
   ↳ máme model a posloupnost pozorování

3) Učení statistického modelu
   → máme strukturu modelu a trénovací data
   → chceme najít pst. přechodů mezi stavy a pst. těch stavů

Kontrola přeřlepů

Požadavky:

- nalézt a opravit všechny přeřlepy
- ten výsledek by neměl být nesmysl
- když nějaké slovo neznám, tak to není chyba
- žádné false positives
- co nejvíc automatizace
- co nejrychlejší

} samozřejmé nereálné

Metrika úspěšnosti

Precision $P = \frac{\#\text{ true positives}}{\#\text{ positives}}$ ... kolik slov hlásím dobře

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

→ Precision je reálně důležitější, lidi nerodí false positives

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

Metody kontroly překlepů

1) Porovnávání řetězců se slovy ve slovníku

- seznam všech možných slovních tvarů = word list
- slovník lemmat + morfologická analýza

⊕ spolehlivé a simple

⊖ pomalé, náročné na kvalitu slovníku, neznámá slova = chybná slova
* slepení musí do slovníku přidat uživatel (ale to už je také m.-analýza)

2) Srovnávání skupin znaků (trojic, trojic) + nějaké zakázané kombinace

⊕ nezávislé na slovníku, rychlé

⊖ hodně chyb to nezachytí – chybná slova mohou být z obvyklých kombinací

Možná vylepšení

- zvažovat důvody vzniku chyby – blízké klávesy
- zohlednit statistiku chyby – která schoda pohn. s přís. je hodně těžká
- zohlednit pravopisné chyby – mně × mě, jsem × jsme
- heuristika na oddělení chyb a neznámých slov
- zapojení syntaxe a sémantiky
- pracovat s kontextem – porovnávat korpusy

Jak se to reálně dělá

- špatné slovo → jak vybrat možné správné tvary?
  → Levenshteinova vzdálenost řetězců
- důležitá je přesnost (Precision) a rychlost
- kontrola na pozadí, žádná upozornění že se našla chyba
- uživatel nahlásí do slovníku jen kontrolní slovní tvar – není morfologie

Systém ASIMUT

= Automatická Selekce Informací Metodou Úplného Textu

- 2 moduly

- **Jazykový modul**
  - není rozsáhlý slovník - pouští brzk
  - mnohr slov, která mají v základním tvaru stejný koncový segment, se stejně skloňuje
    ⇒ retrográdní slovník
    → pravidla lam jít podle roch + čísla + páchu a konce zákl. tvaru
      ↳ nakonec to vyplivne nějaký suffix, co dává ten tvar
    → jsou zřídka asi jenom slovy výjimek

- **Vyhledávací modul**
  - basically to vyhledává regexy
  - výrazy z podst. a příd. jmen v základním tvaru + operátory
    ! vzdálenost slov = jakýkoli tvar
    -1- obě slova vedle sebe
    -2- mezera mezi slovy obsahuje nejvýše 2 slova
    -3- obě slova ve stejné větě
    -4- -5- ve stejném odstavci
  - vyhledávání v nějakých úředních dokumentech

  Dotaz: vzdálenost!, odstup!-3- rodinný!-1-domek!
  ... jak dlelko od sebe mají být domky

- **problémy jazykového modulu**
  - proč není vždy určen jednoznačně
  - příliš hrubá klasifikace
  - malý rozsah retrográdního slovníku ⇒ všetřování výjimek
  - nefunguje to tak úplně pro slovesa
    ↳ koncové segmenty zákl. tvarů sloves jsou různácná

- Negativní slovník
  - obsahuje nedůležitá slova pro vyhledávání (spojky, číslovce)
    ↳ odstraní se při preprocessingu textu

- Kondenzace
  * slovní tvar co není v negativním slovníku dostane (adresu, frekvenci výskytu
  účel: urychlit hledání

Systém MOZAIKA

= Morphemic Oriented System of Automatic Indexing and Condensation
→ hledá nejdůležitější termíny z nějaké oblasti (technické obrody) v textu

1) Standardní přístup k indexaci
   - slovník klíčových slov, dokumenty indexovány slovy
   - v návrhu se bere četnost výskytu

2) Přístup MOZAIKY
   - řada kmenů a přípon nese význam
     -ič, -ač, -ička, -éř, -or, -dlo, -metr, -graf, -fon, -skop ... nástroje/přístroje
     -ace, -ece, -áče ... procesy
     -ost, -ita, -nce ... vlastnosti
     -ací, -ecí ... účel
   - angličtina: -er, -or = kmenem děje
     -tion = činnost
     -ity, -ness = vlastnost
   → pro potřeby témat. oblasti elektrických obvodů ~ 800 přípon

Algoritmus MOSAIKY

1. lemmatizace + morfologická analýza vstupního textu → značky
2. projdu nalezená lemmata a odstraním ty, jejichž kmen nemá vztah k dané sémantické oblasti (například)
   → vznikne malý negativní slovník
3. Chci odhalit několikaslovné termíny
   operační zesilovač TESLA KC 415 > zesilovač
   ↳ podtermíny se také započítají
   ⇒ Syntaktická analýza pomocí jednoduché gramatiky v Q-systému (pozdější)
4. přiřadím váhy na základě místa výskytu v textu
   nadpis je nejvíc
   1. odstavec > další odstavec
   1. věta v odstavci > random věta
5. udělám nějakou normalizaci vah vzhledem k délce dokumentu
   ↳ aby se mohla porovnávat relevance různě dlouhých dokumentů

⇒ výstup = 10 nejzajímavějších termínů spolu s četností výskytu

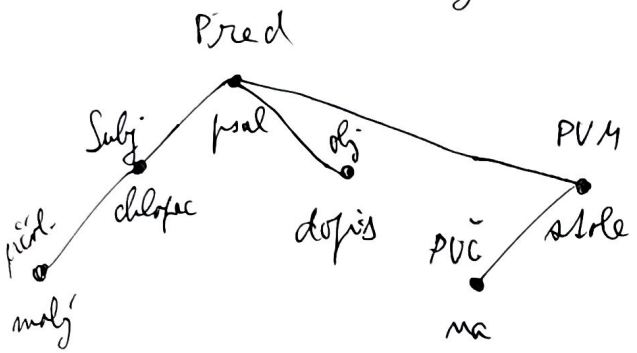
(+) nepotřebuje slovník odborných termínů, jen množina přípon, koncovek + negativní slovník / pravidla
lokální syntaktická analýza umožňuje líp hledat termíny

(−) vymyslet ty přípony a pravidla je pracné
nezapočítává zájmena
↳ ten zesilovač TESLA tam můžou napsat 1x
a pak se na něj 10x odkazovat zájmeny
⇒ 1 výskyt

Syntax

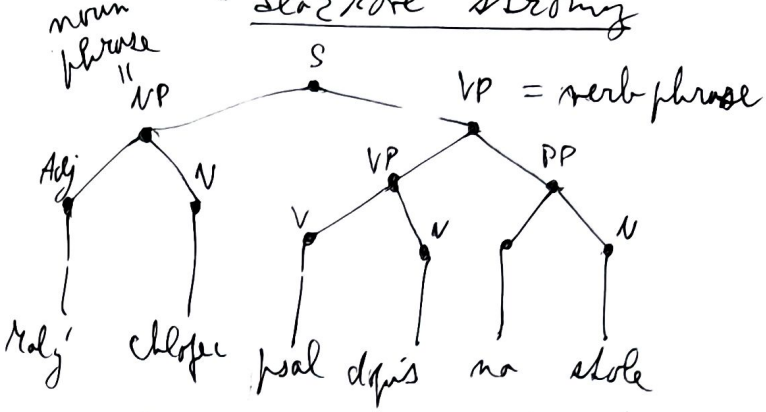
přísudek = kořen

• závislostní stromy



- dobře zachycují vztahy
- strom nezachycuje pořadí výpočtu
  ⇒ není jasné, jak ho získat
- není jednoznačný
- ne všechny vztahy se vždy se všemi dají popsat → Petra a Pavel

• složkové stromy



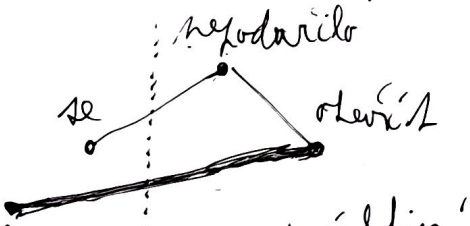
- derivační strom bezkontextové gramatiky
- méně přehledný, složitější vztahy
- přirozený jazyk není bezkontextový
- → spojuje slovní druhy do frází

V + N → VP

Nepojektivní konstrukce

= dlouhé hrany ... za hrana v závislostním stromu jde dopryč

věta: Soubor se nepodařilo otevřít



→ se začátku na konec věty hrana

→ nepojektivní ≡ vertikála má 2 průsečíky

→ složkový strom má problém :(ty slova spolu nesousedí

Kontextová gramatika × Bezkontextová gramatika

- nějaká pravidla
- ve větě by měl být hodnět a přísudek

- pravidla na spojování značek
- dole terminální značky

Generování × Analýza

- výroba stromu, do kterého se za značky pak dají dosadit slova

rozvíjím NP ← S → VP

- z věty dělá strom

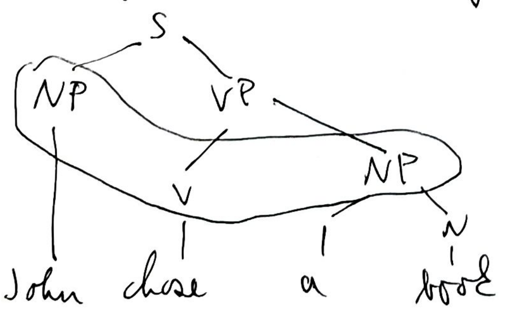
spojuje NP → S ← VP

Transformační gramatika – Noam Chomsky ~ 1965

- převratný způsob formálního popisu přirozeného jazyka
- <u>princip:</u> povrchová × hloubková syntaktická struktura
  - <u>povrchový strom</u> – klasický složkový strom
  - <u>hloubkový strom</u> – zachycuje pouze důležité vazby
    ↳ změna pořadí slov ho nemění
- <u>komponenty</u>
  - <u>báze</u> – bezkontextová pravidla pro generaci složkových stromů
    = <u>frázové ukazatele</u>
  - <u>transformační k.</u> – transf. pravidla operující na frázových ukazatelích
    → spojují / mění ty stromečky
    pravidla – obligatorní / fakultativní
  - <u>fonologická k.</u> – regexy co dávají řetězcům morfémů fonetic. interpr.
- <u>Generativní procedura</u> – jak generovat stromy

$$VP \rightarrow \begin{Bmatrix} \text{možnost 1} \\ \text{možnost 2} \end{Bmatrix} / \; NP \; \_\_$$

→ můžu to přepsat jen když předtím je NP
→ nedokáže zachytit různé varianty vět ← tázací, oznamovací
→ transf. složka má pravidla jak z podkladových fráz. ukazatelů vytvořit povrchovou strukturu věty



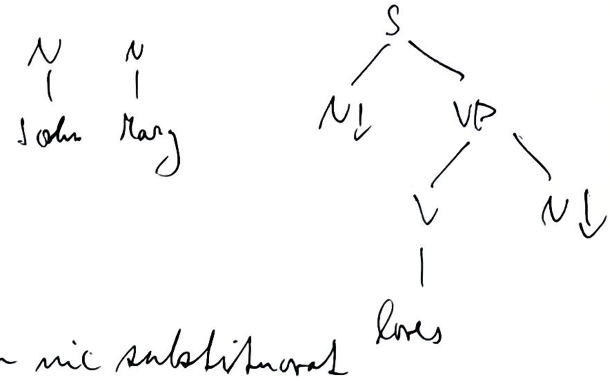
$NP_1 - V - NP_2 \Rightarrow NP_2 - was - V + -en - by + NP_1$

Book was chosen by John

→ ta teorie se hodně měnila, chomsky vydával další verze
- 1965 – Standard Theory – tohle
- 1990 – Teorie minimalismu – pouze rovina logické formy a fonetická rovina

Tree Adjoining Grammers – substituce stromů

- stromečky jako el. struktury
- šipka ↓ značí možnou substituci
- strom ~ syn. struktura věty
  - → vznik postupnou substitucí
- → proces končí, když už nejde nikam nic substituovat



Logical Functional Grammer

- c-structures ~ spojování slov do frází
- f-structures ~ reprezentuje funkční vztahy ve větě (vztahy slovies)
  - ↳ matice $\begin{bmatrix} \text{at} & \text{ho} \\ \text{ri} & \text{dn} \\ \text{bu} & \text{no} \\ \text{ty} & \text{ty} \end{bmatrix}$

→ hodnota může být i jiná f-struktura

∀ c-struktura má právě 1 f-strukturu

ale 1 f s. může být ve více c s.

Kategoriální gramatiky

∀ vstupní slovo dostane kategorii,

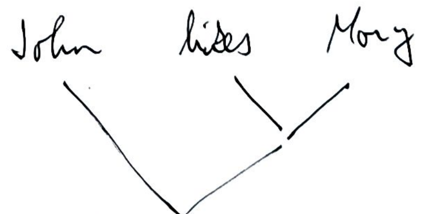
→ kategorie popisuje syntaktické vlastnosti toho slova

→ obecný formát α/β nebo $\beta\backslash\alpha$ ... lomítko říká vlevo/vpravo

- likes má kategorii (NP\S)/NP → díra na něco typu Y
- 2 hlavní pravidla: $X/Y + Y \Rightarrow X$, $Y + Y\backslash X \Rightarrow X$

John likes Mary
NP (NP/S)\NP NP
→ NP/S
← S

⇒



⊕ nepotřebuju obecná gramatická pravidla ∃ spousta výjimek

⊖ strašně moc kategorií

Unifikační gramatiky

- popis vlastností objektů
→ objekt ~ množina vlastností (název vl : hodnota vl.) = sestava rysů
→ sestavy rysů se dají unifikovat
  ↳ lze to, když vlastnosti 1. sestavy neodpovídají těm 2.
→ sestava rysů většinou popisuje nějaký syntagmatický (shoda podmětu)
  → hodnota vlastnosti může být další sestava nebo pointer na ni

Shoda podmětu s přísudkem

↳ $\begin{bmatrix} \text{podmět} : \begin{bmatrix} \text{osoba} & 2 \\ \text{rod} & \check{z} \end{bmatrix} = 1 \\ \text{přísudek:} \quad 1 \end{bmatrix}$

Problém: lze unifikovat nesouvisející rysy
→ sestavy mají typ → ten určuje její vlastnosti
  ↳ Typované sestavy rysů
  ↳ např: sloveso, podstatné jméno, ...

HPSG

- existuje principy, gramatická pravidla a slovníkové položky
- slovník bohatě strukturován, spousta info
- el. jednotka = znak (sign)
- slova a fráze jsou různé podtypy znaků
→ slova mají 2 rysy
  - fonetický ... znak
  - syntaktické a sémantické informace

Nástroje na syntaktickou analýzu

- **Augmented Transition Networks**

– vím jaká slovní fráze co potřebuje

S → NP a VP

NP → Det a N .. Det = cílen "a", "the"

VP → V [NP]

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

⇒ SEEK NP ... NP potřebuje Det a N

↳ CAT Det ... pokud následuje Det, jdi do dalšího stavu

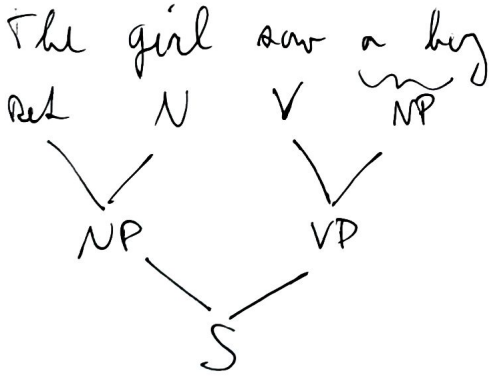
CAT N ... pokud následuje N jdi do dalšího stavu

⇒ SEEK VP ... potřebuje V a volitelně NP

↳ CAT V ... pokud následuje V, pokračuj

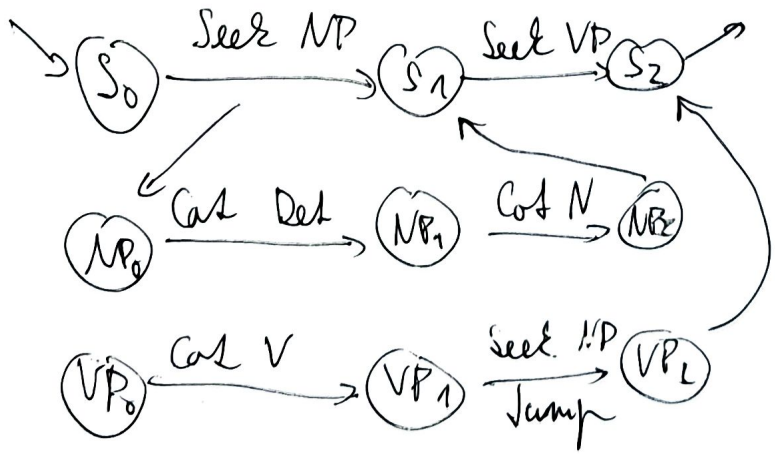
SEEK NP

JUMP ... přejdi do dalšího stavu aniž bys nic nehledal

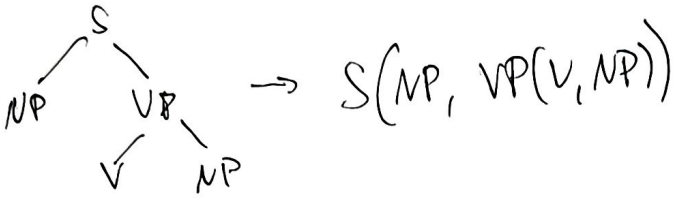


3 typy hran

- SEEK = přechod do podstromu
- CAT = přechod do dalšího stavu pokud najde co hledá
- JUMP = přechod dál bez hledání



- **Q-Systémy**
  - formalismus pro transformaci grafů
  - stromy ukládá lineárně



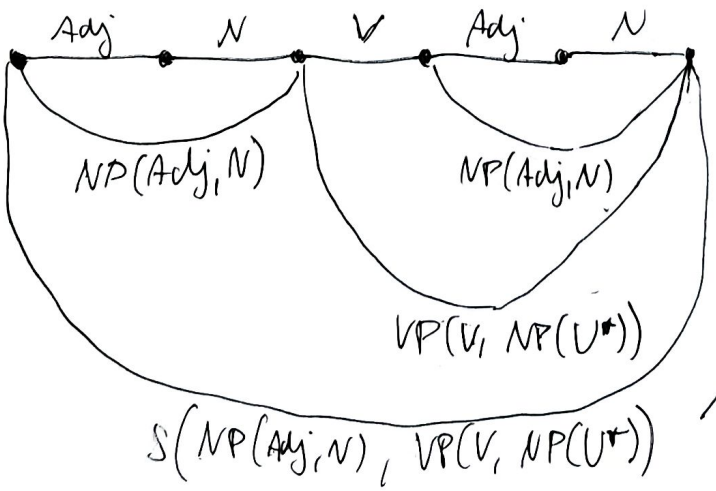
  - 3 typy objektů
    - atomy = konstanty ... písmenka A – J
    - stromy = stromy ... písmenka L – N
    - seznamy stromů ... písmenka V – Z
  - má nějaké operátory na těch objektech
  - zapisuje strom jako regex

    *. = proměnná

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

    $A^*(U^*)$ → splňuje všechny ty regexy

    atom (seznam) — $S(NP, strom)$ — Strom
  - pravidla
    - $Adj + N \rightarrow NP(Adj, N)$
    - $V + NP(U^*) \rightarrow VP(V, NP(U^*))$
    - $NP(U^*) + VP(V^*) \rightarrow S(NP(U^*), VP(V^*))$



Sem výstup můžeme poslat do dalšího Q-Systému

- čištění stromu
  - odstraníme nadbytečné hrany – dílčí mezivýsledky
  → vždy spojuje hrany ... písmenka / výrazy = hrany (na obrázku)
  1. odstraním všechny hrany, co jsem použil na levé straně nějakého pravidla
  2. odstraním slepé uličky

Funkční generativní popis – matfyz projekt

- stratifikační teorie – 5 rovin
  - Tektogramatická TG
  - Povrchová
  - Morfémotická
  - Fonologická
  - Fonetická
- jednotka na vyšší rovině reprezentuje funkci jednotky na nižší rovině
- teorie valence
  ↳ vazby jsou vyžadované nebo povolené řídícími slovy
- TG rovina zachycuje to podstatné
  → Kniha byla vyd. nakl. = nakl. vyd. knihu
  → 2 druhy členů
    - Aktanty: konatel (agens), Patient, Adresát, Origo, Efekt
      (objekt zasažení)
      ↳ ∀ může být v 1 větě jen 1x
    - volná doplnění: klidně víckrát
  → členy dělíme na
    - obligatorní – nesmí na TG chybět
      ↳ podmět tam musí být, i když na povrché rovině může chybět
    - fakultativní
  → obligatorní členy jsou pro ∀ sloveso jiné!
    → Moji přátelé přijeli
      - kam? ... to je důležité ⇒ obligatorní
      - odkud? ... fakultativní

Valenční rámec = seznam aktantů (i fakultativních) a obligatorních volných doplnění

Kontrola gramatické správnosti

Tatínek šli do práce. ← Co kdo kam donesl?

- problémy v češtině
  - neproj. konstrukce, shoda podm. s přísudkem
  - interpunkce, zájmena mě / mně

• chybové vzorky
  - pro jazyky s pevným slovosledem
    ↳ chybové konstrukce jsou blízko u sebe

• gramatika
  - kontrola podle pravidel
  - nelze rozeznat, zda je to chyba, nebo zda je naše gramatika neúplná

} 2 přístupy

• Metoda redukční analýzy
- větu redukuju tak, že vždy odstraním něco, co je 100% správ
  ⇒ neumím chybnou větu opravit ani správnou rozbít

1. Naši hokejisté včera po velké snaze opět vyhrály.
2. hokejisté včera po snaze vyhrály.
3. hokejisté včera vyhrály.
4. hokejisté vyhrály.

→ např. (velké snaze) ... shodný přívlastek ... kontrola shody
  ... je to dobře → odstranit

⇒ chyba zůstane blízko k sobě ⇒ chybové vzorky
- musí se ošetřovat speciální případy

výstup: minimální chybová konfigurace ... vyškrtaná věta

! pouhé hledání chybových vzorkům složitějších konstrukcí nefunguje
  > Které děvčata chtěla dostat náušnice? ☆
  ↳ min. chyb. konf., ale dlouhá hrana

! homonymie ... čelisti jako orgán × čelisti jako mučitelé
rod ž   rod m

• RFODG = Robust Free-Order Dependency Grammar

1 gramatické pravidlo může popisovat správnou i chybnou konstrukci zároveň

- výpočet ve fázích, interpret gramatiky rozhoduje, jak se bude stejné gramatické pravidlo používat

- příklad pravidla: po předložce s je 7. pád
  shoda shodného přívlastku

fáze: pozitivní – tvrdé podmínky → musí to splnit
negativní – měkké podmínky – nemusí to splňovat
projektivní = věci vedle sebe

- 3 fáze: z fáze vždy vypadne nějaký les

1) pozitivní projektivní
- pokud je to věta správně a bez neproj. konst. tak to vyrobí celý strom

2) neg proj. nebo poz. neproj

3) neg. neproj
- tohle už poskládá všechno

→ vyhodí to více stromů, jak to může být

dírky s náušnicemi – neprojektivní strom
\ strom s chybnou posunutím

- **Lan GR**
  - redukuje značky po morf. analýze ⇒ disambiguace
  - pozitivní a negativní disambiguační pravidla
  - pravidla mohou mít neomezený kontext
  - jsou vzájemně nezávislá, nesporiádaná a jsou vyhodnocována v cyklu
  - → 4 části: kontext, disambiguační část, report; ale
    ⇒ $cont_1$ $disamb_1$ $cont_2$ $disamb_2$ ... $cont_n$ $disamb_n$ $cont_{n+1}$ report action
  - ⇒ odhalí podezřelé věty – nějaké slovo na konci nemá žádnou značku

Korpusová lingvistika

Korpus = velké množství textu s nějakou přidanou informací
↳ např. morf. nebo syn. analýza

- **Studie struktury jazyka**
  - → identifikuje jednotky a třídy jazyka (morfémy, slova, fráze...) a popisuje jak se malé kombinují na velké
- **Studie užívání jazyka**
  - → zdůrazňuje hodně náhodně nezávislé vybraných textových dat, tak mi to dává nějakou objektivní informaci o tom, jak se ten jazyk reálně používá
  - → měl by to být reprezentativní vzorek jazyka
- **výběr vzorků a reprezentativnost**
  - jazyk je nekonečný, ale korpusy konečné
  - ⇒ vzorky musí být nějak vyvážené
- **konečná velikost**
  - s výjimkou monitorovacích korpusů, kam se pořád něco přidává mají korpusy fixní velikost
  - → jakmile se jednou vydá, už se nemění
- **strojem čitelná forma**
- snaží by se dodržovat nějaké standardy – UD

- Brown Corpus of Standard American English
  - 1. moderní elektronický korpus
  - 1 milion slov textů v americké angličtině vydaných v roce 1961
  - náhodný výběr vzorků, reprezentativní
  - morfologické značky
- Penn Treebank
  - 1. a nejznámější syntaktický anotovaný korpus
  ~1 milion slov
  - články z Wall Street Journal
  ⇒ výběr k ničemu, je tu hodně slangu, vůbec nereprezentuje angličtinu
- British National Corpus ~ 100 mil. slov, morfologická anotace
- American National Corpus ~ 22 mil. slov, morfologická anotace
- Corpus of Contemporary American englis ~ 410 mil slov →
  ↳ 20 mil. slov za 4 roky mezi lety 1990 a 2010
- Oxford English Corpus ~ 2 miliardy slov →
  ↳ gramatické vzorky Italy nejasné
- Negra Corpus – němčina – syntaktická anotace
- EUROPARL – 11 evropských jazyků, ~ 40 mil slov / jazyk
  ↳ velmi paralelní korpus
  → dobré pro trénování překladů
  → důsledný lidský překlad větu po větě, ne jako celek
  ↳ nezávisle na kontextu

- Český národní korpus – UK, MUNI, Ústav pro jazyk český
  - morfologická anotace
  - dnes ~5 miliard slov
  - každých několik let vyjde nový korpus se slovy za ten uplynulý čas
    ↳ název SYN2020 = vyšel 2020, slova z 2015–2019
- Pražský závislostní korpus
  - chtěli syntaktickou anotaci jako TreeBank
  - teoretický základ je Funkční generativní popis FGP
  - 100 000 vět ~1.25 mil slov → malá podmnožina ČNK
  - anotace
    - morfologická
    - syntaktická pomocí závislostních stromů
      ↳ chtěli ukázat, že jsou lepší než složkové
    - FGP ⇒ celkem 5 rovin
      ↳ anotace různých jemných detailů
      ↳ v TG rovině se kromě stromu kreslí i různé rodokmenové reference, např. naco odkazují zájmena
- Prague Arabic Dependency Treebank
  ⇒ aby dokázali, že závislostní stromy jsou super, tak pomocí nich udělali korpus arabštiny
- Universal Dependencies
  - dřív různé jazyky anotovaly stejné konstrukce různě
  ⇒ snaha anotovat stejné věty v podobných jazycích stejně
  → hodně lidí začalo kopírovat závislostní korpusy, ale každý jinak
  → PZK byl moc detailní → UD zjednodušené schéma
  → dnes se běžně používá UD

LINDAT ← všechny korpusy jsou tam

Sémantika

- syntaxe jen říká, jestli je to gramaticky OK
- sémantika říká, jestli to dává smysl

Kulatý obdélník zeleně mosíroval kovi vedle minuty.

- význam ≠ pravdivost
  - ↳ nepravdivá sdělení mají svůj význam
  - ↳ ne vždy je možné ověřit pravdivost
- není vždy jasné, zda jde o věty se stejným významem
  - Pozorovali to dobrovolně
  - Byl jimi pozorován dobrovolně
- vyplývání – pravdivé věty mají důsledky
  - ↳ ! tučňáci jsou ptáci $\overset{?}{\Longrightarrow}$ mají křídla a létají
- formální jazyky × přirozené jazyky
  - pravdivost = význam / nefunguje to
- Fregeho princip kompozicionality
  - význam složeného výrazu je jednoznačně určen významy jeho částí a způsobem jejich kombinace.
- Lexikální sémantika
  - → význam slov popisujeme pomocí nějakého metajazyka
  - FORMÁLNÍ
    - → hodnoty kategorií: král: muž +, panovník +
      - královna: muž –, panovník +
    - → problém: málo kategorií nestačí
      - hodně kategorií: některá slova jsou nerozhodnutelná
    - } moc
  - Přirozený – stejný nebo jiný – jako ve slovníku – lepší
    - ! význam věty závisí i na kontextu
    - Problém: na vysvětlení mohou být potřeba další neznámá slova
    - Řešení: na popisy se používá třeba jen 2000 nejběžnějších slov jazyka

👁 významy slov je různé nezávisle na kontextu popisovat pomocí významových tříd

- **Ontologie** = množina tříd objektů na klasifikaci objektů universa U

vidličky ∈ příbor ∈ kuchyňské nástroje ∈ nástroje ∈ fyzické objekty

→ velmi těžké to udělat dobře

→ problém, že význam slov není jednoznačný

↳ kohoutek, hlava

| kvantity
| vztahy
| vlastnosti
⋮

ontologie
- vrcholové – obecné významy
- doménové – něco specifičtějšího
  ↳ třeba ontologie vlastností

- **Sémantické sítě** + příslovce
  - **Word Net** → databáze anglických podstatných a přídavných jmen a sloves
    → dělení na třídy ekvivalence = **synsety**
    ↳ ∀ synset vyjadřuje určitý **koncept**
    → relace mezi synsety – sémantický a lexikální význam
    → slovo může patřit do více synsetů
    ~155 000 hesel
    ~117 000 synsetů
  - **Euro Word Net**
    - několik jazyků
    - vrcholové ontologie – 63 nejdůležitějších jazykově nezávislých konceptů
    - množiny základních konceptů – 1000 základních konceptů
      ↳ tvoří jádro sítě slov, jazykově závislé
    - dobré pro automatický překlad
    - dobré pro práci se sémantickými vztahy – hlavně synonyma

Reprezentace významu věty

Predikátová logika 1. řádu

- jakoby to funguje, ale věty nejsou formule
... nezvládá to možné hodnocení
↳ possible(F), believe(x, F), true_at_some_point_in_future(F)

- **presupozice**
= předpoklad, co musí být pravdivý, aby věta vůbec měla pravdiv. hodnotu
"Jupiterův měsíc má oranžové pruhy."
↳ Jupiter musí mít právě 1 měsíc

- **nevíctest = Fuzziness** - T/F nestačí ... Pavel je mladý.

Extenze a intenze

- **Intenze** = definice
- **Extenze** = věci, co splňují tu definici

- Cena Big Macu je 20 Kč.
pokud to nahradím za intenzi → 125 Kč je 20 Kč.
Nepravda

- Myslím, že cena Big Macu je 20 Kč.
↳ není ekvivalentní k "Myslím, že 125 Kč je 20 Kč."

Přístupy k sémantice

1) **Modelově-teoretická sémantika** — svět
pravdivostní podmínky jsou vztažené ke konkrétnímu modelu ↑
→ např. montagueovská gramatika

2) **Kompozicionální sémantika**
- vychází z principu kompozicionality = význam složeného výrazu závisí na významu jeho částí a na způsobech spojování

• <u>Montagueovská gramatika</u>

- základ je formální logika
- Montague byl přesvědčen, že sémantika přirozených a formálních jazyků se moc neliší

axiom: sémantická pravidla jsou svázána se syntaktickými

→ do jisté míry se mu to podařilo

- pravidla připomínají kategoriální gramatiku

  ↳ slovesa co se chovají podobně se seskupují do kategorií a zachází se s nimi stejně

• <u>TIL</u> = Transparentní intenzionální logika

→ logika 1. řádu nestačí

- vyhodnocuje pomocí sémantiky <u>možných světů</u>

  ↳ věta může mít různý význam v různých světech

  → také zohledňuje čas, kdy se to odehrává

- používá lambda kalkul - něco z logiky

ω = svět

τ = čas

$\lambda \omega \lambda \tau (Z^{*}_{\omega\tau} F^{\circledast}_{\omega\tau} Hm^{\boxtimes}_{\omega\tau})$

↳ tohle popisuje nejlepší větu

→ kromě možných světů rozlišujeme i různá individua

- studentka Alena - nálepka individua

* • ministr zahraničí - individuální koncept

⊠ • hezčí než - vztah mezi 2ma individua

⊛ • ministr financí

⇒ můžu vyrobit větu „Studentka Alena si myslí, že MF ≥ MZ"

↳ je to nějaký objekt $O_{\omega\tau}$

↳ každému světu a času přiřadí nejvýše 1 pravdivostní hodnotu

Rozpoznávání vztahů v textu – Anafory

⇒ odkazovací zájmena – problém při překladu

↳ české „on" se musí přeložit na { muž: on → he ; stůl: on → it }

⇒ potřebujeme vědět na co to zájmeno odkazuje

Anafora = výraz jehož interpretace závisí na kontextu

dělení:

- Exofora = odkazování mimo text. → „Vidíš ho?"
- Endofora = odkazování v rámci textu
  - **Anafora** (zpětně) … jo, fakt se to jmenuje stejně
    - ↳ „Petr vyzradil tajemství. To neměl dělat"
  - Katafora (dopředu): „Když se zlobí, není s Petrem žádná řeč."

- **Anafora** … řeší se hlavně ona, vztah předchůdce ↔ následník

1) **Zájmena a nevyjádřené větné členy (podmět)**

   Petr si koupil vstupenku. Dal ji do kapsy. Byla drahá. ?? [nejednoznačné]

2) **Určité jmenné skupiny**

   Elektronický zesilovač Tesla 1000. Toto zařízení…

3) **Elipsa**

   Petr přinesl dva stoly. Dřevěný a kovový.
   Petra půjde do kina. Lucka taky.

4) **Spojovací výrazy**

   například, jednak – jednak, nejdříve – potom

⇒ velmi obtížné: je třeba aplikovat více metod zároveň

Využití:

- získávání informací z textu – MOZAIKA
- automatický překlad – správné překládání zájmen „on" → it/he
- dialogové systémy

• Řešení anafory

1) morfologické značky – např. v zájmen shoda v rodě
2) syntaktická struktura věty
   ↳ pro určení možných kandidátů na předchůdce
3) statistické metody
4) sémantické členění – viz alg. obilených znalostí
5) rozsáhlé formální znalosti – ontologie, sémantické sítě, ...

• Zásoba obilených znalostí

– modelují zásobu znalostí, co mluvčí předpokládá, že sdílí s posluchačem
  ↳ mění se podle toho, co je zrovna v centru pozornosti
– jednoduchá pravidla
→ věta má nějaké okolí a jádro
→ a^n znamená, že a má důležitost n

příklad pravidel:

pokud referjeme přes zájmeno: n se nemění
pokud a je v okolí: $a^n \to a^0$
pokud a je v jádru: $a^n \to a^1$
pokud $a^n \to a^m$: objekty asociované s a obdrží hodnotu a^{m+2}
pokud a není zmíněno ani asociováno: $a^n \to a^{n+2}$

Strojový překlad

problémy:
- potřebujeme znát morfologii, nejde překládat slovo od slova
- pravidla pravopisu: jeden pět = čtyři
- ustálená spojení: velký muž = náčelník
- ne vždy je možné namapovat slova ze dvou jazyků 1-1
- homonyma
- složené výrazy: airport long term car park courtesy vehicle pickup point

- rešerše první generace
  - dvojjazyčný slovník, překlad slov od slova
  - ve slovníku nejsou celá slova, ale morfémy

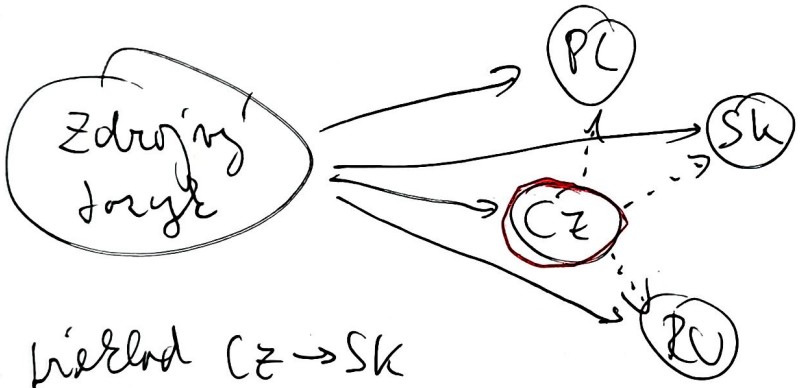
    predediting → překlad → postediting

    ANALÝZA
    TRANSFER
    SYNTÉZA
  - jeden jazyk se zvolí jako pivotní pro překlad do více jazyků
  - Georgetownský experiment 1958
    - velmi skromná omezení domény a povrchových jazykových prostředků
    - → ale fungovalo to
  - 1955–1965 bouřlivý rozvoj
  - 1966 zpráva ALPAC
    - → je nutné dlouhodobě investovat do teoretického lingvistického výzkumu
    - → důsledek: konec podpory v USA
      - ↳ ve Francii, SSSR, Kanadě dál pokračovalo
  - TAUM-METEO – překlad meteorologických zpráv A → F
    - ↳ dobře definovaná a hodně omezená problémová jazyka
    - ↳ používá Q-Systémy
  - SYSTRAN – překlad dokumentů EU mezi ~20 páry → fungovalo jen A-F-N
  - EUROTRA – projekt EU, nedokončen: 72 párů – negativní efekt jako ALPAC
  - VERBMOBIL – překlad mluvené řeči, německý následník EUROTRy
    - ↳ omezení rozboru: plánování příští schůzky mezi obchodními

- **Strojem podporovaný překlad** = CAT
  - komerčně mnohem úspěšnější než automatický – už od 80. let
  - **překladová paměť**
    - ↳ obsahuje páry přeložených segmentů – typicky vět
      + metadata, umožňuje efektivní spravování paměti
    - → nové vstupní věty se vyhledávají v paměti
      - → pokud tam jsou, překladatel dostane návrh překladu se svýznačenými odlišnostmi a procenty podobnosti
    - ⇒ vhodné pro opravené překlady méně aktualizovaných textů – technické manuály, nové verze
  - **2 přístupy**
    1) **překlad věta po větě**
       - překladatel vidí kontext toho co překládá
       - každý výskyt se musí přeložit znovu
       - → IBM Translation Manager
    2) **překladová tabulka** → jakoby její opakování
       - kontext není vidět, ale ∀ věta se překládá jen jednou
       - → Déjà Vu

- **České překladače**
  - **RUSLAN** – překlad manuálů sálových počítačů ~ 1985
    - 1 věta ~ 4 minuty
    - dvojjazyčný slovník, klasický přístup: analýza, transfer, syntéza
    - Q-systémy
    - **Transdukční slovník**
      - 👁 některá slova s řecko-latinským základem je nutné přebírat podle vzor
        -ace → -acja: industrializace → industrializacia

- **Systém Česílko**
  - lokalizace velkých sw. systémů
  - náhrada lokalizace z jazyka typově odlišného překladem z jazyka blízkého
  - překládá se mezi velmi blízkými jazyky – vysoce kvalitní překlad (téměř automatický)
  - statistické značkování částí
  - dvojjazyčný slovník



→ : lidský překlad
⋯→ strojový překlad

  - překlad CZ → SK
    - CS stejná syntaxe
    - → většinou shodné přiřadí slov ve větě
    - → různé slovníky a morfologie

- **PC Translator 2003**
  - tradiční přístup, pravidla
  - asi nejlepší český komerční systém

Statistický strojový překlad

- věci se dají přeložit více způsoby : 3 způsoby $in(EN) \to F$

→ relativní četnost nás zajímá

→ Bayesův vzorec $P(x|y) = \frac{P(y|x)\,P(x)}{P(y)}$

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

- Modelování jazyka

úkol: předpovědět další slovo v běžném textu

→ na základě historie h předpovídáme další slovo : $P(w|h)$

cíl: spočítat pst. celé věty $P(věta) = P(w_1 \cap w_2 \cap \dots \cap w_n)$

→ N-gram:

$$P(n\text{-tice}) = P(w_1 \cap \dots \cap w_n) = P(w_n | w_1 \cap \dots \cap w_{n-1}) \cdot P(w_{n-1} | w_1 \cap \dots \cap w_{n-2}) \cdot P(w_{n-2} | w_1 \cap \dots \cap w_{n-3}) \cdot \dots \cdot P(w_2 | w_1) \cdot P(w_1)$$

problém: dlouhá historie ⇒ potíže s obtížností

→ reálně se používají bigramy, případně bigramy

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

- Vyhlazování

- problém je velikost dat

→ slovník V, $|V| = 40000$ slov ⇒ velikost modelu $|V|^2 = 6{,}4 \cdot 10^{12}$

↳ standardně trénujeme někde na $\sim 10^8$ slovech

⇒ skrytá moc nulových pstí = řídká data

řešení: nahradíme nulové pst nějakou velmi malou hodnotou

↳ potom ale nerozlišíme opravdu nesmyslné kombinace

• Princíp štatistického prekladu

- paralelný korpus = trénovacia množina dobre preložených viet
→ až do nedávna prevládajúca metóda, dnes skôr neurónky

- Metóda zašumeného kanálu

→ chceme prekladať F → A
⇒ hľadáme model $P(a|f)$, čo vyjadrí pst A vety a na francúzske f

$$P(a|f) = \frac{P(f|a) \cdot P(a)}{P(f)}$$ → $P(f)$ je stejná pre ∀ a

⇒ obrátili jsme de-facto smer prekladu
⇒ hľadáme 2 modely: $P(f|a)$, $P(a)$

→ vlastne predstierame, že jsme dostali vetu preloženú z A do F a hľadáme jej správny originál

→ jazykový model $P(a)$ môže byť trigramový model založený na omnoho rozsiahlejším korpuse cieľového jazyka
→ prekladový model $P(f|a)$ je založený na paralelním korpuse (s omnoho menším)

- podstata:
  • prekladový model budujeme v opačnom smere
  • jazykový model odfiltruje nepodarené preklady a vyrovná chyby prekladového modelu
    → pozná, jestli sa anglická veta dáva zmysel
    → vybírá totiž pouze „hezké" vety, nemá vzťah k originálu
  • hľadání prekladových hypotéz (dekódovaní) je tiež svoj problém

- dostanu vetu
  - mám na výber z mojich prekladov
  - pro ∀ spočítam $P(a)$
  } vyberiem → max

Evaluace systémů automatického překladu

→ metrika BLUE

- přesnost překladu v n-gramech

⇒ máme 2 kandidáty na překlad … který je lepší

⇒ pro to máme nepsaných několik správných překladů

- unigramová přesnost

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

C = # slov z kandidáta, co se vyskytují v alespoň jednom referenčním překladu

- n-gramová přesnost

→ dnjde, bnjde …

⇒ N = # n-tic v daném kandidátu (co na sebe navazují)

C = # n-tic z kandidáta co jsou v nějakém referenčním překladu

- BLUE = celkové skóre

- penalizace za stručnost = Brevity penalty (BP)

- pokud bychom brali v úvahu jen # n-gramů, tak by to favorizovalo krátké věty, jejichž n-gramy by byly v překladech, i když by ty překlady reálně byly mnohem lepší

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

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

(+) - rychlý výpočet, relativně objektivní míra

- obecně přijímaný standard

(−) - získání dostatečného množství referenčních překladů je drahé

- favorizuje statistické systémy, co se různí mi hodí

- špatně zachycuje různé varianty - slovosledné, morfologické

- přeceňuje se, není až tak univerzální