Korpusová lingvistika

Studie jazyka:
- struktury
- užití

Co to je korpus? Sbírka textů, v ideálním případě kompletní
- je uchovávaný nějak
- každý korpus bude přirozeně zkreslený.
Něco se neobjeví, kvůli některým problémům.

Vlastnosti:
- reprezentativnost
- kolísání velikost → Je důležité v jistý moment zastavit. Protože je potřeba porovnávat výsledky v jiných časech.
- strojem čitelná forma
- standardní reference

Účelové korpusy

Brown korpus
- první velký korpus
- snažil se nejlépe reprezentovat americkou angličtinu v roce 1961.

Penn Treebank
- první syntakticky anotovaný korpus
- články z Wall Street Journal → Je to ale burzovní slang, není to spisovná angličtina.

Český národní korpus
- vzniklo v 90. letech: UK, MU, AV
- uchováván na morfologické úrovni
- mix novin, knih a technické literatury
  → 60% → 25% → 15%

Pražský závislostní korpus (PDT) (100 000 vět, 1 200 000 slov)
- syntakticky anotovaný korpus → celý anotovaný ručně (na rozdíl od ostatních ve světě)
- založeno na teorii funkčního generativního popisu
  - úrovně anotace:
    - morfologie
    - analytická rovina (povrchově syntaktická)
    - tektogramatická rovina: závislostní, jádro, koreference, vše ostatní

Prague Arabic Dependency Treebank

- přesný závislostní korpus použitý na typově úplně odlišném jazyku.

Pravděpodobnostní a statistické metody

- pravděpodobnost jen je jen aproximací trénovacích dat
  - ↳ tam může být velká odchylka
- například trénujeme překlad na paralelním slovníku / korpusu.

Bayesův vzorec: $P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$

Cíl: chceme predikovat $p(w|h)$, kde w je hledané slovo a h je historie textu.

- tyto modely by byly ale obří a většinou prázdné.

Používá se trigramový model:

jako „okénko" historie bere pouze kontext délky 3.

- spousta konstrukcí bude mít 0 pravděpodobnost

Metoda vyhlazování

- nulové pravděpodobnosti se nastavují na velmi malé ε.
  - už ale nerozeznáme neexistující kombinace

- Překlad beletrie není dobrý trénovací dataset, jelikož se preferuje zachování významu před formou 1:1.
- Do 2015 se překládalo statisticky, nyní funguje neuronový překlad.

Metoda zašuměného kanálu: chceme $f: F \to A$, hledáme $P(A|F)$

Bayes nám říká: $P(A|F) = P(F|A) \cdot P(A) / P(F)$ → toto můžeme chápat jako $= 1$, když jsme to dostali na vstupu

↳ obrátíme směr překladu a hádáme větu, ze které to vzniklo.

modely: překladový $P(f|a)$
cílového jazyka $P(a)$

- neexistuje vztah k originálu, bere se nejbližší cílová věta
- překlad působí obměněně

- těch hypotéz nesmí být příliš mnoho!

problém etiky nejsou
postrádá sémantické pochopení

BLEU:

BLEU: $BP * (p_1 p_2 p_3 p_n)^{1/n} \in \langle 0, 1 \rangle$

lze reprezentovat jako %, ale není to určené pro porovnávání odlišných systémů

penalizace za stručnost ⇓ což je porovnání délky, aby zvítězila kompletní věta

geometrický průměr n-gramové přesnosti pro $n = 1..4$

$p_i :=$ # nalezených i-gramů / i-gramů celkem

- je to velmi rychlé
- velmi přesné, pokud existuje více ref. překladů
    - použití synonym by kazilo skóre, kdybych měl jen jeden ref. překlad
- zdvojnásobení trénovacího datasetu zvedne BLEU pouze o 0,5%
    - nemůžu jen tak stáhnout celý internet, jelikož spousta stránek je vytvořeno již nějakým strojovým překladem, takže se to nedá použít.