# Večjezičnost spletnih informacijskih virov: medjezično iskanje 3

Prevajanje iskalnih zahtev s slovarji, prevajanje iskalnih zahtev s korpusi, avtomatska gradnja jezikovnih virov.

# MI s prevajanjem iskalnih zahtev

Prevajanje iskalnih zahtev s slovarji

# Prevajanje iskalnih zahtev s slovarji

- Potrebujemo dvojezični e-slovar. Različne oblike:
  - od enostavnega glosarja z dvojezičnimi pari besed
  - do pravega računalniškega leksikona s sintaktičnimi in semantičnimi informacijami.
- Za vsako besedo (razen blokiranih) iz iskalne zahteve poiščemo prevod v ciljnem jeziku.
- S prevedeno iskalno zahtevo opravimo enojezično iskanje dokumentov v ciljnem jeziku.

# Prevajanje iskalnih zahtev s slovarji

- V postopek je že vgrajena nenatančnost, izvirajoča iz ohlapnosti naravnega jezika:
  - veliko besed nima natančnega prevoda, ali
  - je prevodov več, z zelo različnimi pomeni.

- Vključevanje prevodov z napačnimi pomeni zelo zniža natančnost iskanja.

# Izpeljava **eksperimentov** s prevajanjem iskalnih zahtev

Klasični pristop:

- Imamo iskalne zahteve v jeziku *j2* in dokumente v jeziku *j2*; znani so relevantni dokumenti za iskalne zahteve.
- Postopek:
  1. enojezično iskanje v jeziku *j2* (rezultat za primerjavo z MI),
  2. ročno prevajanje iskalnih zahtev v jezik *j1*,
  3. avtomatsko prevajanje iskalnih zahtev iz *j1* v *j2*,
  4. enojezično iskanje v jeziku *j2* z isk. zahtevami iz 3. koraka – v resnici medjezično iskanje,
  5. primerjava rezultatov 1 in 4.

# Prevajanje iskalnih zahtev s slovarji

Osrednji problemi MI:

- prevajanje polisemih besed,
- prevajanje besednih zvez,
- prevajanje strokovnih izrazov,
- pomanjkanje jezikovnih virov,
- neprimerljivost rezultatov, dobljenih z različnimi metodami.

# Prevajanje iskalnih zahtev s slovarji

Problem polisemije:

- Beseda v izvornem jeziku ima lahko veliko različnih pomenov, prevod vsakega od pomenov ima lahko tudi v ciljnem jeziku različne pomene.
- Primer: beseda "*fly*".
  - V angleščini 8 pomenov in 13 možnih španskih prevodov;
  - njihovo prevajanje nazaj v angleščino da 38 različnih besed.
- Posledica iskanja z enostavnim prevajanjem brez razreševanja polisemije je kombinatorična eksplozija pomenov in rezultati iskanja z zelo nizko natančnostjo.

# Prevajanje iskalnih zahtev s slovarji

- Uspešnost iskanja brez razreševanja dvoumnosti zaradi polisemije (vključevanje vseh možnih prevodov):

40% - 60% natančnosti enojezičnega iskanja.

- Vse kar je več je odlično.

- Izkaže se celo, da je izbira naključnega od možnih prevodov enako dobra, kot izbira vseh prevodov.

# Prevajanje iskalnih zahtev s slovarji

Razreševanje polisemije:

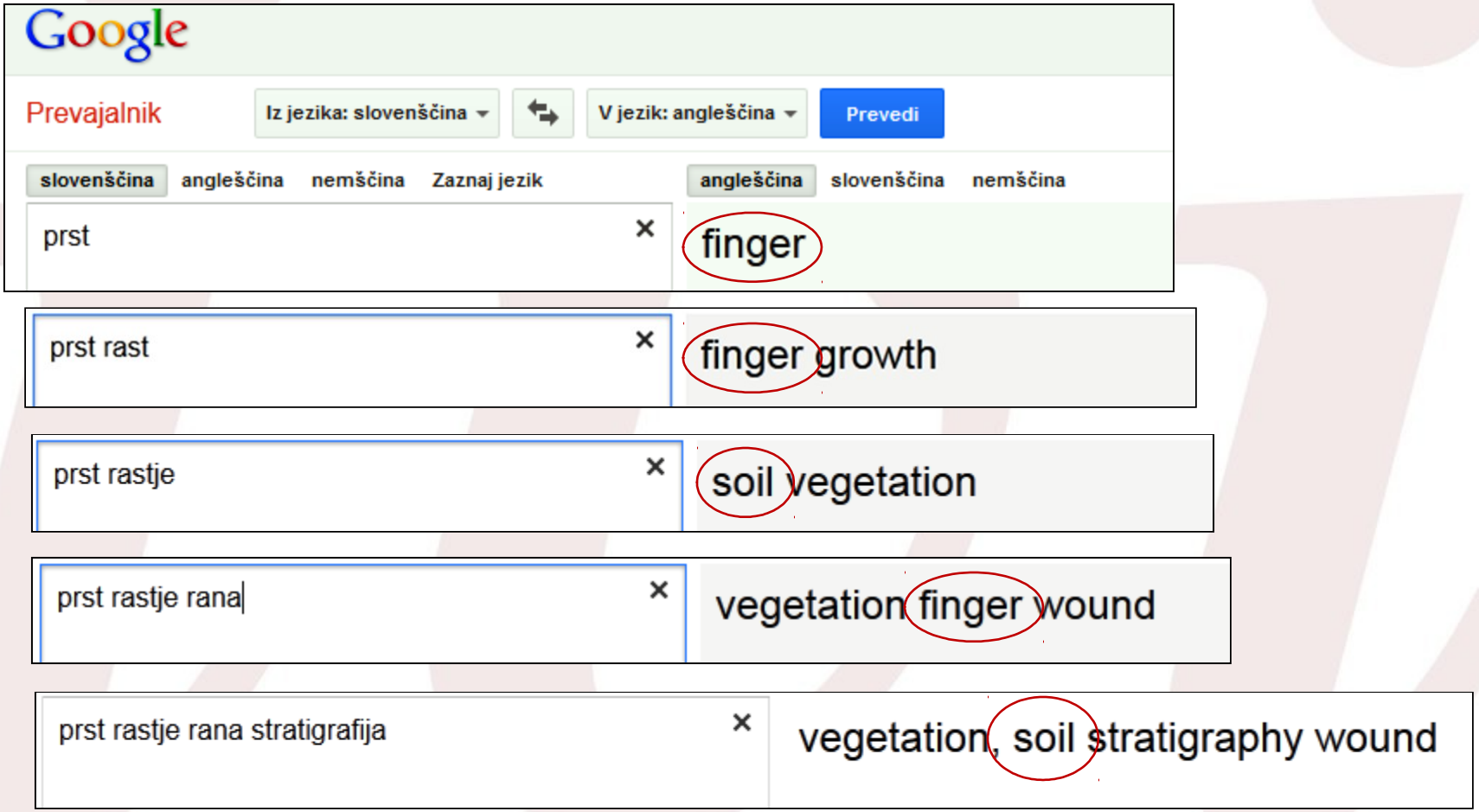
- v iskalno zahtevo vključimo vse prevode neke besede in poskušamo zmanjšati vpliv posameznega prevoda,

ali

- iz porazdelitve besed v učnem korpusu poskušamo izračunati verjetnost posameznih prevodov in vključimo najverjetnejšega(e).

# Prevajanje iskalnih zahtev s slovarji

Primeri prilagajanja prevajanja glede na porazdelitve besed v učnem korpusu.



# Prevajanje iskalnih zahtev s slovarji

Pomen prevajanja besednih zvez

- Pravilno prevajanje besednih zvez dramatično zmanjša vpliv polisemije:
  - samostojno prevajanje posameznih besed, ki sestavljajo zvezo, uvaja množico pomenov, ki so največkrat drugačni od pomena besedne zveze,
  - besedne zveze imajo običajno en sam pomen, zato prevajanje zvez ne uvaja dvoumnosti.

# Prevajanje iskalnih zahtev s slovarji

## Pomen prevajanja besednih zvez

- Poskus (Hull, Grefenstette, 1996):
  - Prevajanje iskalnih zahtev iz francoščine v angleščino. Med drugim primerjala učinek

    (a) slovarja z enobesednimi gesli in

    (b) istega slovarja z dodanimi prevodi besednih zvez.
  - Rezultati:

    (a) 68,4% natančnosti enojezičnega iskanja,

    (b) 90,8% natančnosti enojezičnega iskanja.

# Prevajanje iskalnih zahtev s slovarji

- V znanstvenem informiranju so iskalne zahteve običajno strokovne narave.
- Problem:
  - zelo redki računalniški dvojezični slovarji strokovnega jezika,
  - prevajanje iskalnih zahtev strokovne narave običajno poteka s slovarji splošnega jezika.
- Posledica iskanja je nizek priklic.

# Prevajanje iskalnih zahtev s slovarji

Poskus A. Pirkole, 1998

- Prevajanje iz finščine v angleščino.
- Iskanje časopisnih člankov s poljudno medicinsko tematiko.
- Uporabljeni postopki za:
  - prevajanje strokovnega izrazja,
  - prevajanje polisemih besed,
  - prevajanje besednih zvez.

# Prevajanje iskalnih zahtev s slovarji

Poskus A. Pirkole (nadaljevanje)

- Uporabil splošni in strokovni medicinski slovar:
  - najprej prevajanje s strokovnim slovarjem,
  - sledi prevajanje preostalih besed s splošnim slovarjem.
- Vključil vse možne prevode vsake besede, dvoumnost zaradi polisemije rešil z obteževanjem prevodov.

# Prevajanje iskalnih zahtev s slovarji

Poskus A. Pirkole (nadaljevanje)

- Pri iskanju so imeli vsi prevodi ene besede enak skupni vpliv na računanje relevantnosti dokumenta kot beseda, ki da en sam prevod.
- Npr.: prevod vsake besede je lahko prispeval k izračunu relevantnosti 10 enot. Če je imela beseda 5 možnih prevodov, je vsak lahko prispeval le 2 enoti.
- Tako je imel posamezen (največkrat napačen) prevod poliseme besede manjši relativni vpliv kot prevod besede z enim samim pomenom.
- Strokovne besede so imele največkrat en sam prevod.

# Prevajanje iskalnih zahtev s slovarji

Poskus A. Pirkole (nadaljevanje)

- Problem določanja in prevajanja besednih zvez močno olajšan zaradi same narave finščine – besedne zveze so sestavljenke.
- Uspeh:
  - povprečna natančnost MI praktično dosegla povprečno natančnost enojezičnega iskanja.
- Nauk:
  - zelo dobre rezultate je mogoče doseči brez uporabe dragih jezikovnih virov in zapletenih metod računalniškega jezikoslovja.

# MI s prevajanjem iskalnih zahtev

Prevajanje iskalnih zahtev s korpusi

# Prevajanje iskalnih zahtev s korpusi

- Postopki so najenostavnejši, če so na razpolago paralelni korpusi, poravnani na nivoju stavkov.
- Uporaba:
  - Prevajamo iskalno zahtevo iz jezika *j1* v jezik *j2*.
  - Sistem za vsako besedo v iskalni zahtevi v jeziku *j1* poišče v korpusu v jeziku *j1* vse stavke s to besedo.
  - V korpusu v jeziku *j2* poišče paralelne stavke, jih združi in poišče najpogostejšo besedo.
  - To besedo vključi v prevod iskalne zahteve.

# Prevajanje iskalnih zahtev s korpusi

- Tak enostaven pristop je relativno uspešen, če paralelni korpus sodi v isto domeno, kot iskalna zahteva.
- Paralelni korpusi, poravnani na nivoju stavkov, so zelo redki, izdelava pa izjemno draga.
- Obstajajo le za nekatere jezikovne pare in le za nekatere domene.

# MI s prevajanjem iskalnih zahtev

Prevajanje iskalnih zahtev s slovarji in korpusi

# Prevajanje iskalnih zahtev s slovarji in korpusi

- Dobre rezultate daje kombinacija prevajanja iskalne zahteve z dvojezičnim slovarjem in razreševanja dvoumnosti prevodov s korpusom.
- Uporabljeni korpusi:
  - paralelni, poravnani na nivoju dokumentov, ali
  - primerljivi, »poravnani« na nivoju tematike.
- Take korpuse je laže dobiti.
- Korpus mora biti po vsebini primerljiv z vsebino zbirke,

# Prevajanje iskalnih zahtev s slovarji in korpusi

Postopek prevajanja iskalne zahteve:

- Korpus je v jeziku, v katerega prevajamo iskalno zahtevo.
- Algoritmi s slovarjem poiščejo možne prevode besede iz iskalne zahteve, in
- v korpusu preverijo, kateri od pomenov je najverjetnejši (npr. najpogostejši), in
- to besedo uporabijo v prevodu iskalne zahteve.

# Avtomatska gradnja jezikovnih virov: gradnja paralelnih korpusov

# Gradnja paralelnih korpusov

Nie et al., 1999:

- Odkrivanje jezikovnih parov spletnih dokumentov.
- Uporabili najpogostejše lastnosti parov:
  - prevodi dokumentov povezani s kazalci v obe smeri,
  - besedilo sidra kazalca imenuje jezik dokumenta (“in English”, “English version”...),
  - pari dokumentov imajo podobna imena (“products_fre.html”, “products_eng.html”...),
  - na spletišču sta hierarhiji map za dokumente v posameznih jezikih zelo podobni ali identični.

# Gradnja paralelnih korpusov

Nie et al., 1999 (nadaljevanje):

- Brez posebnih težav sestavili paralelni korpus, poravnan na nivoju besedil.
- Nadaljevanje postopka je avtomatsko preverjanje pravilnosti izbire parov z enostavnimi hevrističnimi postopki.
- Ročno preverjanje pokazalo le 2% napak.
- Velikost korpusa 14.200 parov dokumentov (250 Mbytov).