Večjezičnost spletnih informacijskih virov: medjezično iskanje 2

MI z večjezičnimi tezavri,
računalniško prevajanje v MI.

MI z večjezičnim tezavrom

MI z večjezičnim tezavrom

- Najstarejša oblika MI.
- Tezaver s prevodi konceptov v različne jezike.
- Ročno indeksiranje dokumenta v jezikih *j1, j2, j3* z deskriptorji v jezikih *j1, j2, j3*.
- Iskanje z deskriptorji v jeziku *j1* ALI *j2* ALI *j3* poišče dokumente v jezikih *j1* IN *j2* IN *j3*.
- Do 100% uspešnost v primerjavi z enojezičnim iskanjem.

MI z večjezičnim tezavrom

Primer večjezičnega tezavra:

EUROVOC:

- Večjezični tezaver, v katerem so vsa gesla prevedena v 22 jezikov EU (+ hrvaščina in srbščina).
- Obstajajo še neuradne variante: ruska, baskovska in katalonska.
- Gesla pokrivajo področja, na katerih je aktivna EU.
- Uporabljajo ga dokumentacijske službe vseh pomembnejših institucij EU, pri katerih nastajajo dokumenti, med drugim evropski, nacionalni in regionalni partlamenti.

MI z večjezičnim tezavrom

- Največja pomanjkljivost MI z večjezičnim tezavrom je cena ročnega indeksiranja.

- Opravljeni zanimivi poskusi avtomatske izrabe večjezičnih tezavrov za prevajanje iskalnih zahtev v naravnem jeziku.
- Osnovna ideja: prevesti iskalne zahteve v naravnem jeziku v deskriptorje večjezičnega tezavra in izvesti MI.

MI z večjezičnim tezavrom

- Primer: uporaba UMLS za MI s francoskimi in španskimi iskalnimi zahtevami v naravnem jeziku.
- UMLS (Unified Medical Language System): “seštevek” 60+ tezavrov, osnova je MeSH (Medical Subject Headings).
- Obstajajo številni prevodi MeSH, vključeni v UMLS.

MI z večjezičnim specializiranim tezavrom

Povzetek postopka:

- Prevajanje francoskih in španskih iskalnih zahtev v naravnem jeziku v francoske oz. španske prevode deskriptorjev MeSH.
- Sestavljanje iskalne zahteve iz angleških ustreznic teh deskriptorjev.
- Iskanje po zbirki Medline, ki je indeksirana z angleškimi deskriptorji.

MI z večjezičnim specializiranim tezavrom

Primer (nadaljevanje):

- Izbor francoskih (španskih) deskriptorjev v 3 korakih:
  1. izbrani enobesedni francoski deskriptorji, ki so enaki besedam iz francoske iskalne zahteve,
  2. sestavljeni vsi možni pari preostalih fra. besed in izbrani dovolj podobni dvobesedni fra. deskriptorji,
  3. za vsako fra. besedo, preostalo po korakih 1 in 2
  4. zbrani vsi fra. deskriptorji, v katerih se pojavlja,
  5. poiskani njihovi angleški prevodi,
  6. angleški deskriptorji razbiti na besede,
  7. kot prevod v angleščino izbrana najfrekventnejša beseda.

MI z večjezičnim specializiranim tezavrom

Primer (nadaljevanje):

- Uspešnost postopka, merjena kot % natančnosti, ki bi jo dosegli z angleškimi deskriptorji, ki bi jih določil izkušen informacijski posrednik:
  - španske iskalne zahteve – 71%,
  - francoske iskalne zahteve – 61%.

- Relativno uspešen poskus, vendar postopek omejen na specializirano ontologijo (MeSH) v relativno ozki domeni (medicina).

MI z računalniškim prevajanjem dokumentov

MI z računalniškim prevajanjem dokumentov

Dilema (kaj je bolje):

- avtomatsko prevajanje iskalnih zahtev ali avtomatsko prevajanje dokumentov?

Prevajanje iskalnih zahtev:

- (teoretično) manjši računalniški napor,
- iskalec dobi rezultate v različnih jezikih,
- večji iskalčev napor pri razumevanju dokumentov.

MI z računalniškim prevajanjem dokumentov

Prevajanje dokumentov (v fazi gradnje zbirke)

- avtomatsko prevajanje vseh dokumentov v vse jezike zbirke,
- z iskalno zahtevo v kateremkoli jeziku zbirke je iskanje enojezično,
- uporabnik dobi dokumente v svojem jeziku,
- majhen iskalčev napor, velik (prevelik?) računalniški napor.

MI z računalniškim prevajanjem dokumentov

Prevajanje dokumentov (po iskanju)

- avtomatsko prevajanje iskalnih zahtev, sledi medjezično iskanje,
- iskalec je sposoben približnega razumevanja dokumentov in odločanja o relevantnih dokumentih,
- (varianta: sistem sposoben avtomatskega abstrahiranja in prevajanja zgoščene vsebine),
- avtomatsko prevajanje najboljših relevantnih dokumentov,
- prevodi se v sistemu kopičijo.

MI z računalniškim prevajanjem dokumentov

Eden redkih poskusov (Oard, 1998):

- Korpus 250.000 nemških dokumentov računalniško preveden v angleščino.
- Iskanje z angleškimi iskalnimi zahtevami – zelo velika natančnost.
- Za prevajanje porabljenih 10 procesorskih mesecev na najmočnejših delovnih postajah (za l. 1998).
- Korpus relativno majhen in statičen – realnost spleta, digitalnih knjižnic in števila jezikov je drugačna.

MI z računalniškim prevajanjem dokumentov

Konsenz srenje:

- računalniško prevajanje dokumentov je prenaporno in prepočasno za zahteve MI.
- Zaenkrat je videti njegovo prihodnost le v omejenih situacijah za prevajanje posameznih dokumentov.

MI s prevajanjem iskalnih zahtev

MI s prevajanjem iskalnih zahtev

- Iskalna zahteva se z avtomatskimi postopki prevede v jezike dokumentov, potem sledi serija enojezičnih iskanj.
- Na prvi pogled je pravo računalniško prevajanje iskalnih zahtev idealno tudi za potrebe MI, realnost je drugačna.

MI s prevajanjem iskalnih zahtev

Računalniško prevajanje temelji na metodah, kot so

- razčlenjevanje stavkov,
- označevanje besednih vrst,
- razreševanje dvoumnosti večpomenskih (polisemih besed).

Cilj računalniškega prevajanja je

- generiranje sintaktično in semantično pravilnih stavkov.
- Pri različnih prevodih besede se mora prevajalnik odločiti le za enega.

MI s prevajanjem iskalnih zahtev

- Računalniško prevajanje potrebuje dolge in pravilne besedilne strukture ter sobesedilo za ugotavljanje najverjetnejšega pomena besed.
- Iskalne zahteve so kratka besedila, pogosto le zaporedja ključnih besed.

- Uporaben rezultat prevajanja za potrebe MI so posamezne, nepovezane besede.
- Različni prevodi besede so pogosto sinonimi in so zato lahko koristni v prevedeni iskalni zahtevi.

MI s prevajanjem iskalnih zahtev

- Pravo računalniško prevajanje iskalnih zahtev je uporabno le v redkih primerih:
  - dolge, večstavčne iskalne zahteve,
  - dokument kot iskalna zahteva in iskanje najsorodnejših dokumentov v ciljnem jeziku.