

MEDJEZIČNO ISKANJE DOKUMENTOV

Jure Dimec

Oddano: 07.03.2002 – Sprejeto: 06.05.2002

Pregledni znanstveni članek

UDK 004.738.52

Izvleček

Članek utemeljuje potrebo po razvoju medjezičnega iskanja (MI), relativno novega področja shranjevanja in iskanja informacij v večjezičnih tekstovnih zbirkah, definira njegove cilje in umeščenost med raziskovalnimi področji, ki se ukvarjajo z različnimi vidiki obravnave besedil v elektronski obliki. Kratkemu pregledu zgodovine sledi opis najpomembnejših metodoloških pristopov v MI (prevajanje dokumentov, prevajanje iskalnih zahtev) in jezikovnih virov, ki so pri tem v uporabi. Med viri je največ pozornosti posvečene dvo- in večjezičnim ontologijam (tezavrom, slovarjem, prevajalskim leksikonom in tezavrom kolokacij), korpusom, njihovi gradnji in uporabi pri eksperimentih MI. Članek poskuša predvsem ilustrirati pestrost metodologije področja in manj delovanje konkretnih sistemov. Stanje MI v Sloveniji in obstoj jezikovnih virov, primernih za vključevanje slovenskih besedil v medjezične sisteme, nista obravnavana, ker je to tematika, ki zahteva poseben pregled.

Ključne besede: iskanje informacij, naravni jeziki, medjezično iskanje

Review scientific article

UDC 004.738.52

Abstract

The article reviews the motivation behind the development of cross-language information retrieval (CLIR) – a relatively new area of information retrieval in multilingual textual databases, defines its objectives and position among the research disciplines dealing with various aspects of processing electronic texts. A short historical overview is followed by a description of the most important methodologies (document translation and query translation) and language resources used in connection with them. Regarding the resources, attention is focused on the two- and multilingual ontologies (thesauri, transfer lexicons and similarity thesauri), corpora, their construction and use with

DIMEC, Jure: Cross-language information retrieval. Knjižnica, Ljubljana, 46(2002)1-2, 77-110

the CLIR experiments. The article primarily aims at illustrating the various methodological approaches, while the functioning of particular systems is less prominent. There are no references to the condition of CLIR in Slovenia or to the existence of language resources suitable for processing Slovenian texts in CLIR systems, since this topic calls for a separate review.

Key words: information retrieval, cross-language information retrieval

1 Uvod

Eksplzivni razvoj omrežnega, predvsem spletnega publiciranja, je že pred časom sprožil razvoj omrežnih iskalnikov. Opisovanje vsebine dokumentov v zbirkah teh iskalnikov temelji na ključnih besedah in besednih zvezah, avtomatsko izbranih iz samih dokumentov, poizvedovanje pa na iskalnih zahtevah v naravnem jeziku. Od prvih začetkov spletnega publiciranja, ko so bili dokumenti skoraj izključno v angleščini, se stalno povečuje delež neangleških dokumentov, tako med vsemi spletnimi dokumenti kot tudi med tistimi, ki so urejeni v digitalnih knjižnicah. Po študiji OCLC je bil v letu 2001 delež angleških dokumentov na internetu samo še 73%, 21% dokumentov pa je bilo v enem od ostalih evropskih jezikov¹. Konec devetdesetih se je število angleških dokumentov na leto povečalo za približno polovico, letna rast neangleških dokumentov pa je bila kar 90% (Braschler in Schaeuble, 1998). Obstoj spletnih dokumentov v različnih jezikih se seveda odraža tudi v jezikovni pestrosti zbirk spletnih iskalnikov. Ker iskanje z iskalnimi zahtevami v naravnem jeziku pomeni primerjanje besed ali besednih zvez iz iskalne zahteve z besedami ali besednimi zvezami v dokumentih, iskanje ne more dati rezultatov, če sta primerjana iskalna zahteva in dokument v različnih jezikih. Celo pri zbirkah z zelo omejenim številom različnih jezikov dokumentov pomeni zaporedno sestavljanje iskalnih zahtev v teh jezikih za iskalca precejšen napor. Potrebujemo torej iskalnike, ki bodo na iskalno zahtevo v jeziku iskalca temu ponudili relevantne dokumente v vseh jezikih zbirke. Do zrelih sistemov s takimi lastnostmi je sicer še daleč, raziskave in razvoj ustrezne metodologije, ki jo imenujemo medjezično iskanje (cross-language information retrieval - CLIR), pa so v svetu med najživahnejšimi od vseh na področju shranjevanja in iskanja informacij (information retrieval - IR).

Medjezično iskanje (MI) je relativno nov pojem, zato imajo težave s terminologijo tudi na angleškem govornem področju. V strokovni literaturi se še vedno pogosto pojavljajo različni izrazi, kot so cross-language IR, cross-lingual IR, multilingual IR, translingual IR..., ne da bi bila vedno jasna razmejitev njihovih pomenov. V glavnem pa velja:

1 <http://wpc.oclc.org/stats/global.html>

Medjezično iskanje (cross-language IR, CLIR) je iskanje, pri katerem je naravni jezik iskalne zahteve lahko različen od jezika ali jezikov, v katerih je izražena vsebina dokumentov, ki jih iz zbirke prikliče iskalna zahteva. Če je iskalna zahteva v jeziku *a*, dokumenti v zbirki pa v jezikih *a* in *b*, bodo poiskani relevantni dokumenti v obeh jezikih. Običajno je način izražanja vsebine tiskano besedilo, čeprav podobna načela (in veliko dodatnih problemov) veljajo tudi za govorjeno besedilo. MI je tudi iskanje v zbirki z enojezičnimi podatki, če je omogočeno zastavljanje iskalnih zahtev v različnih jezikih.

O **enojezičnem** ali **istojezičnem iskanju** (monolingual IR) govorimo takrat, kadar sta iskalna zahteva in dokumenti v zbirki v istem jeziku, skratka v običajni situaciji. Medjezično iskanje z enim delom svoje definicije pokriva tudi enojezično iskanje.

Najširši izraz je **večjezično iskanje** (multilingual IR), ki vključuje enojezično, medjezično iskanje, in iskanje dokumentov z deli v različnih jezikih. Tudi sisteme s pomnoženo enojezično funkcionalnostjo, torej take, ki omogočajo z iskalnimi zahtevami v različnih jezikih priklic dokumentov v teh jezikih, imenujemo večjezični sistemi.

Ameriški avtorji včasih – z grenkim humorjem in poznavanjem jezikovnega znanja rojakov – imenujejo medjezične sisteme »sistemi, ki iskalcem nudijo dokumente, ki jih ti ne znajo prebrati«.

MI je dober primer raziskovalnega področja, ki posega preko meja poddisciplin. Pri razvoju sistemov za medjezično iskanje je nujna uporaba metod IR, računalniškega jezikoslovja, računalniškega prevajanja, pogosto tudi avtomatskega povzemanja (sumarizacije) besedil, prepoznavanja govora in še kakšnega področja, prevladuje pa seveda metodologija prvih dveh naštetih. MI in IR imata mnogo skupnega: načine organizacije dokumentov v zbirkah, metode avtomatskega indeksiranja (vektorske, probabilistične, latentno semantično indeksiranje...), interpretiranje iskalnih zahtev in računanje relevantnosti. Bistvena razlika med MI in IR je prevajanje, ki je v različnih oblikah in stopnjah zahtevnosti prisotno v MI, v IR pa ne. Relativno manj pomembna lastnost MI, ki ga loči od IR, je tudi sposobnost odkrivanja identičnih dokumentov v različnih jezikih, ki je potrebna pri prikazovanju rezultatov iskanja.

Pri običajnem, enojezičnem iskanju je sestavljanje iskalne zahteve postopek, pri katerem poskuša iskalec uganiti besede, s katerimi je avtor dokumenta izrazil vsebino, kakršna iskalca zanima. Pri medjezičnem iskanju je situacija kompleksnejša. Pri medjezičnem iskanju iskalec poskuša v jeziku iskalne zahteve uganiti besede, ki imajo isti *pomen* kot besede, ki jih je uporabil avtor, ko je v jeziku dokumenta izrazil vsebino, ki zanima iskalca. Odgovornost za izbor konkretnih besed, ki jih je uporabil avtor, je preložena na sistem za iska-

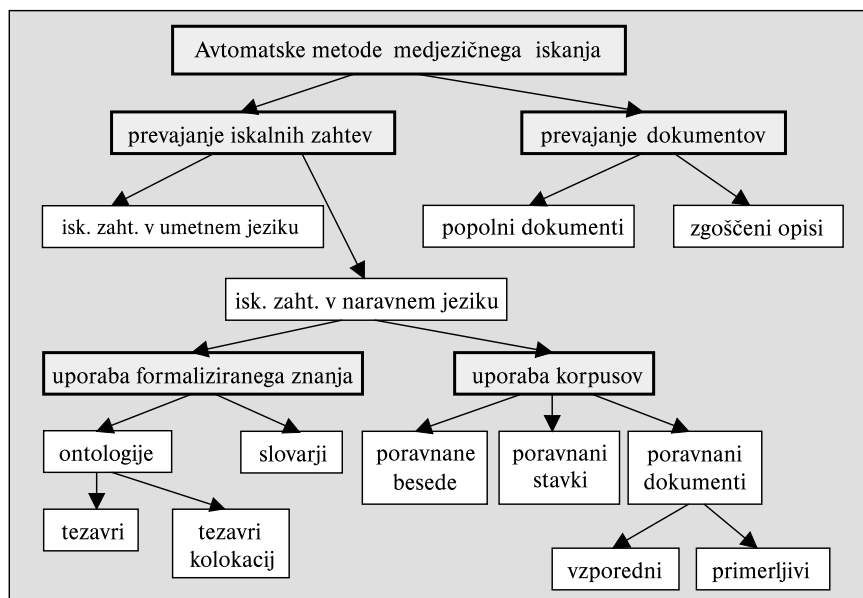
nje (Sperer in Oard, 2000). Pomen je seveda mehka lastnost, ki se slabo prilega algoritmičnemu reševanju problemov, in od tod izvirajo vse težave sistemov za MI ter, posledično, njihova inferiornost v primerjavi s sistemi za enojezično iskanje.

Postopke, ki se uporabljajo pri MI, bi lahko razdelili na nekaj kategorij:

- a) izdelava večjezičnega tezavra in avtomatsko prevajanje deskriptorjev v iskalni zahtevi v tiste jezike dokumentov, ki jih pokriva tezaver;
- b) prevajanje dokumentov v jezik iskalne zahteve;
- c) prevajanje iskalne zahteve v jezike dokumentov, s podkategorijami:
 - prevajanje iskalnih zahtev z orodji za pravo računalniško prevajanje,
 - prevajanje besed ali besednih zvez v iskalnih zahtevah s pomočjo dvojezičnih računalniških slovarjev²,
 - prevajanje besed ali besednih zvez v iskalnih zahtevah s pomočjo dvojezičnih korpusov;
- d) latentno semantično indeksiranje (LSI).

Slika 1 prikazuje hierarhijo večine avtomatskih postopkov in jezikovnih virov, ki so v uporabi pri medjezičnem iskanju. Iz predstavitve je izpuščeno latentno semantično indeksiranje, ker temelji na metodah, neprimerljivih z ostalimi na sliki.

2 »Slovar« je v kontekstu tega članka nerodna beseda. Zaradi rabe pri računalniškem prevajanju bi bilo bolje »leksikon«, vendar je tudi to zmuzljivo pojem, katerega pomen je odvisen od zasnove sistema za prevajanje. Lahko gre za enostaven dvojezičen seznam ustreznih ali pa za pravi transferski leksikon z natančnimi oblikoslovnimi, skladensjimi in pomenoslovnimi podatki za vsako geslo v obeh jezikih. V poročilih o raziskavah, opisanih v tem članku, najpogosteje ni jasno, za kakšno jezikovno orodje gre. Zato in v izogib dilemam, ki jih lahko prinese drugačno razumevanje pojma »leksikon« v nejezikoslovnem jeziku, je v članku povsod uporabljena beseda »slovar«.



Slika 1: Zgoščen pregled postopkov in jezikovnih virov v rabi pri medjezičnem iskanju (po Oard, 1997b).

Vse vrste MI (razen LSI) očitno vključujejo neko obliko prevajanja. Lahko gre za intelektualno ali avtomatsko prevajanje, za prevajanje elementov vsebinskega opisa ali prevajanje celih dokumentov. V tem članku se bomo ukvarjali pretežno z avtomatskimi metodami, in v nadaljevanju se bo izkazalo, da imajo med njimi posebno težo postopki prevajanja iskalnih zahtev.

2 Zgodovina medjezičnega iskanja

Prvi dokumentirani poskusi MI so bili, v skladu z razvojem tekstovnih zbirk tistega časa, narejeni z večjezičnimi tezavri. Izvedla sta jih, neodvisno drug od drugega, Gerald Salton s Cornellske univerze, tudi sicer slavjen kot utemeljitelj sodobnih metod poizvedovanja po tekstovnih zbirkah, in Rus B. R. Pevzner. V prvem poskusu iz leta 1969 je Salton preveril uporabo dvojezičnega (nemško-angleškega) seznama ključnih besed za iskanje po majhni zbirki 1.095 angleških in 468 nemških izvlečkov (Salton, 1970). Del seznama angleških ključnih besed in 48 angleških iskalnih zahtev je bilo »ročno« prevedenih v nemščino. Uporabljen je bil Saltonov sistem SMART, namenjen razvoju ne-Boolovih metod poizvedovanja. Iskanje z nemškimi iskalnimi zahtevami v zbirki angleških

izvlečkov je pokazalo 6% padec povprečne natančnosti v primerjavi z enojezičnim iskanjem, padec pri iskanju z angleškimi iskalnimi zahtevami po nemških izvlečkih pa je bil 3%, oboje merjeno na podoben način, kot je to opisano v poglavju o evalvaciji učinkovitosti iskanja. Padec natančnosti, v tedanjih časih, ko je bil razvoj sodobnih metod poizvedovanja po tekstovnih zbirkah še na začetku, razumljen kot neuspeh, je Salton pripisoval predvsem pomanjkljivemu seznamu ključnih besed. V drugem poskusu iz leta 1973 (Salton, 1973) je uporabil nekoliko bolj sofisticiran dvojezični (angleško-francoski) seznam ključnih besed in zbirko vzporednih prevodov 52 izvlečkov. Povprečna natančnost pri iskanju z angleškimi iskalnimi zahtevami po francoskih izvlečkih je celo prerasla povprečno natančnost pri enojezičnem iskanju (za 5%), pri uporabi francoskih iskalnih zahtev in angleških izvlečkov pa je bila za 12% slabša od enojezičnega iskanja.

Približno ob istem času je Pevzner prevedel obsežni ruski tezaver elektrotehniške stroke (večtisoč vsebinskih konceptov in preko 600 relacij med njimi) v angleščino in uporabil klasični Boolov sistem PNP-2. Enojezična in navzkrižna iskanja po 4.000 ruskih in 4.400 angleških dokumentih s 103 iskalnimi zahtevami v obeh jezikih niso pokazala statistično značilne razlike v učinkovitosti (Pevzner, 1973; citirano v: Oard in Dorr, 1996).

Omenjeni eksperimenti so nakazali smer razvoja MI v naslednjih letih in jasno dokazali, da ima MI prihodnost. Z današnjega zornega kota so Saltonovi rezultati relativno nezanesljivi zaradi majhnih zbirk, njegovi dvojezični seznanja ključnih besed pa so bili, tako zaradi majhnosti kot tudi zaradi pomanjkanja informacij o semantičnih povezavah, daleč od pravih tezavrov. Verjetno so bili Pevznerjevi eksperimenti bolje izpeljani, vendar zaradi jezikovne in družbenih pregrad niso pustili globljih sledi.

3 Medjezično iskanje z večjezičnimi tezavri in semantičnimi mrežami

Iskanje v zbirkah, v katerih je vsebina zapisov opisana z večjezičnimi elementi umetnega jezika, je najstarejša oblika medjezičnega iskanja. Za izvedbo potrebujemo tezaver, pri katerem so vsebinskim konceptom pripisana gesla v različnih jezikih. Vsebinski opis dokumentov, shranjenih v zbirki, je nastal z intelektualnim indeksiranjem z gesli v jezikih dokumentov. Iskalec uporabi gesla v enem od jezikov večjezičnega tezavra, sistem jih prevede v ostale jezike in priključuje iz zbirke dokumente v teh jezikih. Dobro zasnovan sistem za medjezično iskanje, temelječ na kontroliranem večjezičnem tezavru, je lahko skoraj tako učinkovit kot posamezni enojezični sistemi. Do morebitnih razlik

v priklicu in natančnosti iskanja v posameznih jezikih prihaja predvsem zaradi neuskkljenosti skupin indekserjev, zadolženih za posamezna jezikovna področja, čeprav je tudi res, da je za nekatera strokovna področja z enostavnim prevajanjem gesel težko enakovredno pokriti iste vsebinske koncepte, ki se lahko v različnih jezikovnih okoljih do neke mere razlikujejo.

Pred nekaj leti je bila velika večina komercialnih sistemov v znanstvenem informiranju bibliografske narave, zato je bila večina prizadevanj za MI usmerjena v razvoj večjezičnih tezavrov. Že leta 1978 so bili UNESCO-vi standardi za gradnjo večjezičnih tezavrov (UNESCO, 1971) dopolnjeni in sprejeti kot standard ISO 5964. Potrebe po večjezičnih tezavrih so še posebej občutili v administraciji Evropske Unije tudi zato, ker je bilo treba omogočiti dostop do večjezičnih zbirk dokumentov v vseh uradnih jezikih. Med projekti, financiranimi iz tega vira, je treba omeniti vsaj gradnjo devetjezičnega tezavra EUROVOC³ in projekt TRANSLIB⁴, ki pa je bil usmerjen že bolj v procesiranje naravnega jezika v večjezičnem okolju. V projektu so evalvirali uporabnost raznih večjezičnih pripomočkov, kot so dvojezični slovarji, razčlenjevalniki (parserji), avtomatski prevajalniki, terminološke podatkovne zbirke in tezavri za poizvedovanje v knjižničnih katalogih.

Raba umetnega informacijskega jezika pri postopkih opisovanja vsebine dokumentov in iskanja ima svoje prednosti (predvsem predvidljivost), med pomanjkljivostmi pa je najočitnejša cena intelektualnega indeksiranja v primerjavi z avtomatskim. Zato je razvoj metod MI z večjezičnimi tezavri pripeljal tudi do poskusov njihove rabe pri medjezičnem iskanju z iskalnimi zahtevami v naravnem jeziku. Tak poskus je opravila skupina na Univerzi Iowa (Eichmann, Ruiz in Srinivasan, 1998) in pri tem uporabila metatezaver UMLS⁵. UMLS (Unified Medical Language System) je seštevek več kot 60 tezavrov z biomedicinskega področja, najpomembnejši med njimi in nosilna struktura za dodajanje ostalih konceptov, pa je MeSH. Metatezaver za leto 2001 vsebuje preko 800.000 semantično povezanih konceptov⁶, ki vsebujejo 1,9 milijona izrazov, med katerimi so tudi prevodi deskriptorjev MeSH. Španski prevod vsebuje 23.198, francoski pa 18.277 deskriptorjev. Eichmann in sodelavci so s pomočjo UMLS v angleščino prevajali iskalne zahteve v španščini in francoščini in jih uporabili za iskanje v zbirki Medline⁷. Jedro postopka je izbor španskih in francoskih deskriptorjev, ki najbolje predstavljajo vsebino iskalne zahteve; od tu dalje je postopek enak običajnemu MI z večjezičnim tezavrom. Španske in francoske

3 <http://europa.eu.int/celex/eurovoc/>

4 http://peterpan.uc3m.es/english/areas/biblio_e/marco_translib.htm

5 <http://www.nlm.nih.gov/research/umls/>

6 Koncept predstavlja nosilni deskriptor, njegovi sinonimi, leksične variante in drugojezične ustreznice.

7 Spletna verzija zbirke je dostopna na <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi>

deskriptorje so v eksperimentu izbrali na tri načine: (a) v iskalno zahtevo so bili uvrščeni enobesedni deskriptorji, ki so bili enaki besedam iz iskalne zahteve, (b) sestavili so možne kombinacije besed iz iskalne zahteve, ki niso bile uporabljene v koraku a in izbrali tiste deskriptorje, ki so bili tem naključnim kombinacijam dovolj podobni, in (c) za vsako špansko (francosko) besedo, ki ni bila uporabljena v koraku a, so zbrali vse španske (francoske) deskriptorje, v katerih se ta beseda pojavlja, poiskali vse angleške prevode teh deskriptorjev, jih razbili na besede, besede prešteli in za najverjetnejši prevod španske (francoske) besede izbrali tisto angleško, ki se med preštetimi največkrat pojavi. Z iskalnimi zahtevami, sestavljenimi na tak način, so v povprečju dosegli 71% natančnosti (prevajanje iz španščine) oz. 61% natančnosti (prevajanje iz francoščine) iskanja z originalnimi angleškimi iskalnimi zahtevami⁸. Opisani eksperiment je pomemben predvsem zato, ker je dokazal, da je mogoče uporabiti večjezični tezaver tudi za prevajanje iskalnih zahtev v naravnem jeziku. Seveda pa je postopek izvedljiv samo za omejene vsebinske domene, kajti za širok nestrokoven jezik, razen redkih izjem (EuroWordNet, opisan v nadaljevanju), ne obstajajo večjezične ontologije.

Visoko razvit tezaver, v katerem so vsebinski elementi povezani še drugače, kot samo z relacijami nadrejenosti in podrejenosti, si lahko predstavljamo tudi kot semantično mrežo z elementi (ključnimi besedami, deskriptorji) v vlogi vozlov in relacijami, ki ponazarjajo vsebinsko sorodnost, v vlogi povezav med vozli. Semantične mreže seveda niso zanimive le za predstavitev umetnega jezika. V projektu WordNet⁹ (Miller et al., 1993) so se na Princetonski univerzi lotili modeliranja naravnega angleškega jezika - titanske naloge, ki zaposluje raziskovalce od leta 1985. Osnova semantične mreže je t. i. *synset*, ki ga sestavljajo besede z dovolj sorodnim pomenom, da lahko funkcionirajo kot sinonimi. Vsak *synset* predstavlja posamezen vsebinski koncept. Relacije hiponimije in hipernimije povezujejo *synsete* s *synseti*, ki vsebujejo nadpomenke in podpomenke (s tem pridobi WordNet tudi hierarhično strukturo), relacije meronimije in holonimije urejajo odnose del, član in sestavina, relacije antonimije pa povezujejo *synsete* z nasprotnimi pomeni. V WordNet verzije 1.5 je 94.515 različnih *synsetov* s 187.602 različnimi pomeni.

Aktualnost problematike MI je vzpodbudila razvoj EuroWordNet¹⁰ (Vossen, 1998). V prvi fazi, med leti 1996 in 1999, je projekt financirala EU, v nadaljevanju pa je razvoj prevzelo javno združenje Global WordNet Association¹¹. Namen

8 Način primerjanja povprečne natančnosti originalnih iskalnih zahtev v angleščini in iskalnih zahtev, prevedenih na avtomatski način v angleščino iz španščine in francoščine, je bil skladen s standardnimi postopki evalvacije, opisanimi v poglavju Evalvacija učinkovitosti iskanja.

9 <http://www.cogsci.princeton.edu/~wn/w3wn.html>

10 <http://www.hum.uva.nl/~ewn/>

11 <http://www.hum.uva.nl/~ewn/gwa.htm>

dela je zgraditi, po vzoru WordNet, zbirko izrazov v različnih evropskih jezikih (zaenkrat je naloga končana za nizozemščino, italijanščino, španščino, nemščino, francoščino, češčino in estonščino), vključiti semantične relacije med izrazi v posameznih jezikih in omogočiti povezave med izrazi v različnih jezikih. EuroWordNet je zasnovan kot večjezična ontologija velikih razsežnosti, ki naj bi bila, po mnenju avtorjev, uporabna za konceptualno indeksiranje in iskanje, vključno z medjezičnim iskanjem. Sestavljen je iz ločenih enojezičnih semantičnih mrež, ki nastajajo samostojno in delno odslikavajo tudi kulturne in jezikovne posebnosti. Semantične mreže posameznih jezikov so različno velike, trenutno sta največji nizozemska in italijanska s preko 40.000 synseti. Mreže vključujejo le samostalnike in glagole iz »splošnega« naravnega jezika, v nadaljevanju pa nameravajo obseg razširiti z izrazi iz strokovnih domen. Pomemben del EuroWordNet je *medjezični indeks*, ki povezuje enojezične mreže z angleškim WordNet 1.5. Gradnja poteka polavtomatsko, pomemben del je avtomatska ekstrakcija podatkov iz računalniško berljivih slovarjev in drugih jezikovnih virov, najvišji hierarhični nivoji ontologije, ki so skupni vsem jezikom, pa so določeni ročno. Uspešnost projekta potrjuje tudi dejstvo, da so trenutno v razvoju mreže za 19 dodatnih evropskih in neevropskih jezikov, med njimi tudi slovenska (projekt vodi Tomaž Erjavec z Inštituta Jožef Stefan).

Uporabnost ontologij, kakršni sta WordNet in EuroWordNet, obstaja tako za enojezično kot tudi za medjezično iskanje. V vsakem primeru gre pri iskanju za razširitev iskalne zahteve. Najenostavnejša raba je (a) izkoriščanje synsetov za dodajanje sinonimov prvotni iskalni zahtevi, zapletenejša pa (b) izkoriščanje semantične oddaljenosti v ontologiji za obteževanje besed v iskalni zahtevi. Pri enojezičnem iskanju se taka raba WordNet zaenkrat ni obnesla: pristop (a) sicer dvigne priklic, vendar preveč zniža natančnost (razlog je uvajanje sinonimov z napačnim pomenom pri polisemih besedah (Vorhees, 1994)), pri pristopu (b) pa še manjka dobra definicija semantične oddaljenosti. Izjema so zelo kratki dokumenti (npr. podnapisi k slikam), kjer se običajne metode iskanja, zaradi majhne možnosti ujemanja besed iz iskalne zahteve, izkažejo z zelo nizkim priklicem, razširjene iskalne zahteve pa možnost ujemanja bistveno dvignejo. V MI je večjezična ontologija lahko dragocena, ker načeloma omogoča jezikovno neodvisno indeksiranje – v primeru EuroWordNet s pomočjo izrazov v medjezičnem indeksu (Gilarranz, Gonzalo in Verdejo, 1997).

4 Medjezično iskanje s pravim računalniškim prevajanjem

Lahko bi rekli, da se večina raziskav, povezanih z MI, osredotoča na algoritmične probleme – računalnik je pač po svoji naravi slabo »prilagojen« na procesira-

nje naravnega jezika. Razvijamo sisteme, ki so vedno bolj sposobni iskanja dokumentov v različnih jezikih, ob tihi predpostavki, da so iskalci sposobni branja in razumevanja besedil v teh jezikih. Seveda ta predpostavka pogosto ne drži. V veliki večini avtomatskih postopkov MI prevajamo iskalne zahteve v jezike dokumentov, treba pa je razmisliti tudi o možnosti, da bi računalnik prevajal dokumente v jezik iskalca. Povsem na mestu je torej dilema: avtomatsko prevajanje iskalnih zahtev ali avtomatsko prevajanje dokumentov s posledicami, ki bi bile približno take, kot je opisano v nadaljevanju.

Če se MI ukvarja s prevajanjem iskalnih zahtev, to načeloma pomeni manjši procesorski napor, a večji napor za uporabnika, ker mu sistem posreduje dokumente v različnih jezikih. Prevajanje dokumentov zahteva večji procesorski napor, a manjši napor za uporabnika, ki dobi kot rezultat iskanja dokumente, prevedene v svoj jezik. Oboje seveda v idealni situaciji in ob predpostavki, da prevajanje znamo opraviti. Računalniško prevajanje dokumentov bi načeloma lahko opravili na dva načina. Prevajanje dokumentov v vse jezike sistema ob vključevanju v zbirko, kot prva možnost, bi sam postopek iskanja omejilo na enojezično iskanje. Druga možnost bi vključevala prevajanje iskalnih zahtev, dokumenti v zbirki bi bili shranjeni v izvornih jezikih, prevajanje v jezik iskalca pa bi potekalo le na poiskanih dokumentih, morda samo najvišje uvrščenih, ali le tistih, za katere bi to iskalec izrecno zahteval. Ta možnost bi seveda pomenila manjši procesorski napor le navidez ali ob majhnem številu uporabnikov, res pa je, da bi se prevodi lahko v sistemu kopičili.

Najbrž ni presenetljivo, da se je obsežno prevajanje dokumentov kot način izpeljave MI, izkazalo za sicer izvedljivo, vendar – milo rečeno – nepraktično nalogo. Eno redkih raziskav, ki nakazuje drugačen rezultat, opisujejo Oard in sodelavci (Oard et al., 1998). Opravili so več eksperimentov, povezanih z MI, in eden med njimi je temeljil na računalniškem prevajanju polnih dokumentov. V okviru študije so iz nemščine v angleščino prevedli 251.572 dokumentov, jih avtomatsko indeksirali in iskali po njih z angleškimi iskalnimi zahtevami. Metoda se je izkazala s precej višjo povprečno natančnostjo kot katerakoli od metod prevajanja iskalnih zahtev, preizkušenih v seriji eksperimentov, vendar so za prevajanje dokumentov porabili 10 procesorskih mesecev na delovnih postajah Sun SPARC 20. Seveda se lahko vprašamo tudi, kako bi se taka metodologija MI obnesla v dinamičnih zbirkah z več kot dvema jezikoma dokumentov.

Najbrž je realnejši pristop izbor in prevajanje nekaterih ključnih besed in besednih zvez ali povzemanje poiskanih dokumentov, kar iskalcu lahko dovolj dobro predstavi vsebino. Šele če je iskalec prepričan v relevantnost, je smiselno popolno prevajanje (Hovy et al., 1999). Podobno funkcionalnost imajo neka-

teri veliki spletni iskalniki (AltaVista¹² ...), za katere je zaradi ekstremnih pogojev delovanja značilna popolna podrejenost ekonomičnosti algoritmov.

Na vsak način pa je zamisel o posredovanju dokumentov, prevedenih v jezik iskalca, privlačna, vendar pri trenutnem stanju tehnologije računalniškega prevajanja ni videti zelo realna. Prevajanju vseh dokumentov v vse jezike zbirke se moramo zaenkrat odpovedati zaradi njegove časovne zahtevnosti, iskalne zahteve pa so, v primerjavi z dokumenti, zelo kratka besedila. Razmislimo torej še o drugi možni rešitvi dileme – računalniškem prevajanju iskalnih zahtev. Izkaže se, da je ta izbira še težje izvedljiva. Medtem, ko je pravo računalniško prevajanje sicer sposobno relativno kvalitetnega prevajanja dokumentov, pa pri iskalnih zahtevah odpove.

Pravo računalniško prevajanje temelji na metodah kot so označevanje besednih vrst, razčlenjevanje stavkov (parsing) in razreševanje dvoumnosti polisemih (večpomenskih) besed. Te metode lahko delujejo le na pravilnih besedilnih strukturah in potrebujejo dovolj sobesedila, iz katerega je mogoče dobiti podatke o uporabljenem pomenu in najverjetnejšem prevodu neke besede. Običajne iskalne zahteve, tudi če so izražene v naravnem jeziku, so največkrat zaporedje besed ali besednih zvez, ne pa semantično in sintaktično koherentno besedilo, zato načeloma ne ustrezajo zahtevam, ki jih postavlja računalniško prevajanje.

Zdi se, da je pravo računalniško prevajanje iskalnih zahtev ne samo neustrezna, ampak tudi pretežka artilerija. Računalniško prevajanje porabi veliko energije za produkcijo sintaktično pravilnih stavkov, MI pa (kot dediščina IR) zanimajo v prevedeni iskalni zahtevi samo besede ali besedne zveze, nepovezane v višje strukture. Računalniško prevajanje se mora pri dvoumnosti prevodov besede vedno odločiti za enega, MI pa pogosto vključi v iskalno zahtevo vse možne prevode (razen, če s statistično analizo korpusa izračuna verjetnosti posameznih prevodov), ker so med njimi lahko tudi sinonimi, ki dvignejo priključ iskanja. Očitno je, da sistemi za MI ne morejo nastati z enostavnim dodajanjem funkcionalnosti računalniškega prevajanja sistemom za IR, res pa je tudi, da bi moralo biti razumevanje pomena besed (v smislu razreševanja dvoumnosti prevodov) pomembna nadgradnja MI v primerjavi z IR.

Inferiornost računalniškega prevajanja iskalnih zahtev v primerjavi s prevajanjem s pomočjo dvojezičnih slovarjev (predmet enega naslednjih poglavij) izpričujejo rezultati poizkusov Radwana in Fluhra, o katerih poroča Radwan v doktorski disertaciji (Radwan, 1994; citirano v: Hull in Grefenstette, 1996). Pri prevajanju francoskih iskalnih zahtev za iskanje po zbirki angleških besedil sta

12 <http://www.altavista.com>

uporabljala vrsto jezikovnih virov (dvojezične slovarje besed, besednih zvez in idiomov) ter sistem za računalniško prevajanje SYSTRAN¹³. Pri standardnem načinu evalvacije je prevajanje s slovarji doseglo 78%, računalniško prevajanje pa le 62% natančnosti enojezičnega iskanja. Treba je pripomniti, da doseženi rezultati precej odstopajo (v pozitivno smer) od večine ostalih rezultatov MI, ki uporabljajo sorodne pristope (slovarsko prevajanje iskalnih zahtev brez pomoči korpusov). Za prevajanje s pomočjo slovarjev lahko uspeh verjetno pripišemo neobičajno bogatemu slovarskemu instrumentariju, v katerega je bilo vložena veliko »ročnega« dela.

V zagovor pravemu računalniškemu prevajanju je treba povedati, da je večina slabih rezultatov pri MI vsaj deloma tudi posledica oblike iskalnih zahtev, kakršne običajno zastavljajo iskalci – majhnega števila nepovezanih ključnih besed. V dveh poskusih (Nie et al., 1999) so s prevajanjem francoskih iskalnih zahtev v angleščino dosegli 107% in 103% povprečne natančnosti enojezičnega iskanja. Iskalne zahteve, ki so bile popolni stavki, so prevajali s spletno dostopno verzijo SYSTRAN-a.

5 Evalvacija učinkovitosti iskanja

Postopki preverjanja učinkovitosti MI so večinoma podobni tistim, ki smo jih navajeni pri enojezičnem iskanju in ne-Boolovem iskalnem modelu. Standardna mera, s katero primerjamo dva iskalnika ali dve metodi, je natančnost¹⁴ iskanja pri enajstih stopnjah priklica¹⁵ (0%, 10%, 20%, ..., 100%), povprečno za večje število iskalnih zahtev. Da lahko računamo priklic, moramo seveda poznati dokumente v zbirki, ki so relevantni za ocenjevane iskalne zahteve. Pri klasičnem IR to najpogosteje dosežemo z uporabo standardnih iskalnih zahtev in zbirke dokumentov z znanimi ocenami relevantnosti. Posebnost evalvacije MI je svojska izbira referenčnih rezultatov. To so lahko rezultati iskanja enojezičnega sistema na istih dokumentih ali pa rezultati iskanja z iskalnimi zahtevami, pridobljenimi s »popolnim« človeškim prevodom. Pogosta izvedba evalvacije učinkovitosti avtomatskega prevajanja iskalnih zahtev, pri kateri ocenjujemo prevajanje iskalnih zahtev iz jezika J_1 v jezik J_2 , poteka na sledeči način. Denimo, da imamo zbirko angleških dokumentov in angleške iskalne zahteve,

13 <http://www.systransoft.com/>

14 Klasična definicija natančnosti (precision) v Boolovem iskalnem modelu – delež relevantnih dokumentov med zadetki – ni prikladna za ne-Boolov iskalni model z rangiranjem zadetkov. Ker tu množica zadetkov, v nasprotju z Boolovim modelom, ni jasno omejena, je treba meje postaviti umetno – kot pozicije v ranžirni vrsti, na katerih priklic doseže določene vrednosti.

15 Priklic (recall) je delež poiskanih relevantnih dokumentov med vsemi relevantnimi dokumenti v zbirki.

preverjamo pa, kako naš sistem za MI prevaja iz francoščine v angleščino. Preverjanje bo potekalo v petih korakih: (1) enojezično iskanje v angleščini, (2) prevajanje iskalnih zahtev iz angleščine v francoščino, ki ga opravi izurjen (človeški) prevajalec, (3) avtomatsko prevajanje prevedenih iskalnih zahtev (francoščina) v prvotni jezik (angleščina) z metodologijo, ki jo preverjamo v sistemu, (4) enojezično iskanje z iskalnimi zahtevami v angleščini, prevedenimi v 3. koraku, in (5) primerjava rezultatov iskanj v 1. in 4. koraku.

6 Medjezično iskanje, osnovano na dvojezičnih slovarjih naravnega jezika

V nadaljevanju ostajamo pri medjezičnem iskanju, ki temelji na prevajanju iskalnih zahtev. Metode, pri katerih temelji prevajanje na slovarjih, sledijo istim načelom kot metode, osnovane na kontroliranih besednjakih, le da namesto kontroliranega informacijskega jezika uporabljajo dvojezične slovarje za prevajanje iskalnih zahtev v jezike dokumentov. S pomočjo slovarjev sistemi iskalno zahtevo v enem od jezikov slovarja prevedejo v ostale jezike, rezultat iskanja pa so neprevedeni dokumenti v jezikih prevodov iskalne zahteve.

V jedro pristopa so na žalost že vgrajene pomanjkljivosti, ki izvirajo iz pomen-ske ohlapnosti naravnega jezika. Veliko besed nima natančnega prevoda ali pa je možnih prevodov več, pogosto z zelo različnimi pomeni. Že pri iskanju po avtomatsko indeksiranih besedilih v enojezičnem okolju, pada natančnost zaradi polisemije, v večjezičnem okolju pa nenatančnost prevodov situacijo še poslabša. Uporaba vseh možnih prevodov besede iz izvorne iskalne zahteve, lahko uvede veliko število pomenov v prevedeno iskalno zahtevo, kajti marsikatera od prevedenih besed ima lahko tudi v tem jeziku večje število pomenov. Eksplozijo pomenov, ki nastanejo z enostavnim slovarskim prevajanjem polisemih besed, je ilustriral Douglas Oard (Oard, 1997b): beseda »fly« ima v angleščini 8 različnih pomenov, ki dajo 13 možnih španskih prevodov. Rezultat prevajanja teh besed nazaj v angleščino je 38 različnih besed. Kako zelo lahko to vpliva na zmanjševanje učinkovitosti iskanja, je pokazal eden od poskusov (Oard et al., 1998), pri katerem se je izkazalo, da je mogoče, z naključno izbiro enega od možnih prevodov besede, dobiti enako dobre rezultate, kot z uporabo vseh možnih prevodov.

Poleg polisemije, ki je glavni razlog zmanjševanja učinkovitosti MI, Lisa Bales-teros (Balesteros in Croft, 1998) omenja še dva razloga: (a) pomanjkljivo preva-janje strokovnega izrazja, ki ga ni v splošnih dvojezičnih slovarjih, in (b) poma-njkljivo prevajanje besednih zvez s samostojnim pomenom, ki je lahko zelo drugačen od pomena besed, ki sestavljajo zvezo.

Kljub omenjenim težavam MI dosega določene uspehe z enostavno uporabo dvojezičnega slovarja in nekaterih osnovnih orodij računalniškega jezikoslovja, vendar ostaja velika razlika v uspešnosti eno in medjezičnega iskanja. Učinkovitost MI, osnovanega na prevajanju iskalnih zahtev besedo za besedo s pomočjo dvojezičnih slovarjev, običajno doseže 40-60% učinkovitosti enojezičnega iskanja (Balesteros in Croft, 1997). S takimi rezultati vsekakor ne moremo biti zadovoljni, zato so bili opravljeni številni eksperimenti, pri katerih so raziskovalci poskušali dobiti namige za razreševanje dvoumnosti prevodov iz ostalih dokumentov v zbirki ali iz korpusov primerljivih dokumentov. Pri opisu teh metod se ne moremo izogniti tudi omenjanja uporabe korpusov, čeprav je tej tematiki v članku namenjeno posebno poglavje.

Izkaže se, da je učinkovitost MI s prevajanjem iskalnih zahtev odvisna od številnih dejavnikov. Eksperimenti, opisani v nadaljevanju poglavja, dvigajo učinkovitost predvsem s pazljivim prevajanjem besednih zvez in zmanjševanjem pomena nepravilnih prevodov, do katerih prihaja pri polisemih besedah. Določanje besednih zvez v iskalnih zahtevah poteka na tri načine: (a) besedne zveze so samostojna gesla v slovarju, (b) besedne zveze (samostalniške zveze) določi avtomatski označevalnik besednih vrst (part-of-speech tagger), in (c) besedne zveze so dvojice besed, ki se izkažejo s pogostimi kolokacijami (sopojavljanji) v korpusu.

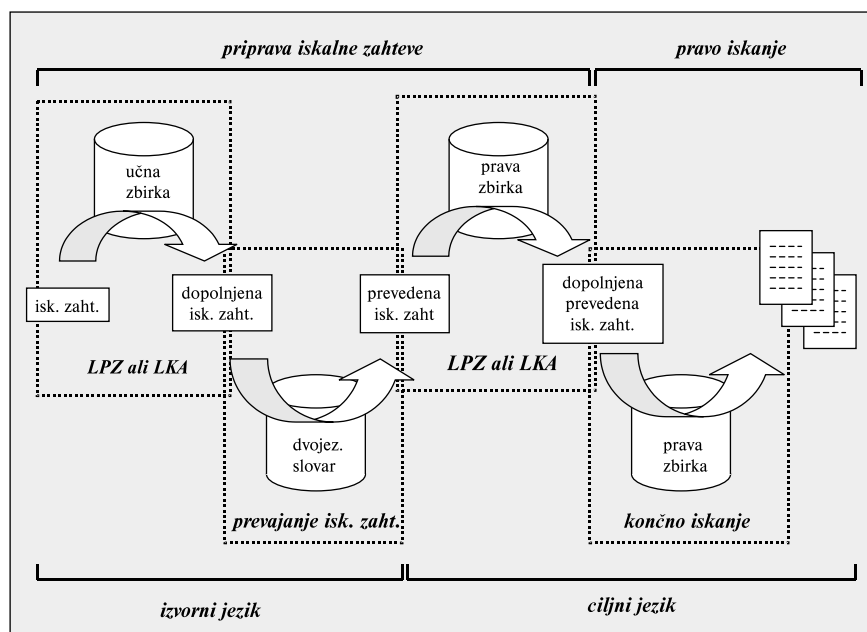
Eksperimenti, ki sta jih opravila z dokumenti v angleščini in francoščini Hull in Grefenstette, potrjujejo pomen, ki ga ima za uspeh MI dobro prevajanje besednih zvez (Hull in Grefenstette, 1996). V eksperimentu, ki je preverjal uspešnost prevajanja iskalnih zahtev iz francoščine v angleščino, so bili uporabljeni trije različni dvojezični slovarji v elektronski obliki: (a) splošni dvojezični slovar, ki v izvorni obliki ni bil namenjen prevajanju iskalnih zahtev, (b) slovar, ki je nastal iz splošnega dvojezičnega slovarja, vendar so bili prevodi gesel z ročnim postopkom očiščeni spremljevalnih informacij, ter (c) slovar, opisan pod točko b, v katerem so bile prevedene tudi samostalniške besedne zveze s samostojno semantično vlogo. Slovarja a in b sta torej omogočala prevajanje posameznih besed, slovar c pa tudi besednih zvez. V vseh treh primerih so bili v slovarju ohranjeni vsi obstoječi prevodi. Prevajanje vsake besede iz iskalne zahteve v francoščini je tako dodalo v iskalno zahtevo, prevedeno v angleščino, vse možne prevode. Hull in Grefenstette sta primerjala rezultate iskanj z iskalnimi zahtevami, nastalimi z vsemi tremi slovarji, z rezultati enojezičnega iskanja v angleščini. Iskanje z iskalno zahtevo, nastalo s slovarjem a, je doseglo 59,8% enojezičnega iskanja, iskanje z iskalno zahtevo b 68,4%, iskanje z iskalno zahtevo c pa kar 90,8% natančnosti enojezičnega iskanja. Zanimivo je, da v enojezičnem IR upoštevanje besednih zvez prinese mnogo manj dramatične izboljšave rezultatov, kot pa v opisanem medjezičnem iskanju. Avtorja predvidevata, da izrazita prednost pravilno prevedenih besednih zvez v njunem eksperimentu izvira iz dejstva, da samostojno prevajanje posameznih

besed v njih prinese v iskalno zahtevo množico prevodov, ki s pomenom izvirne besedne zveze nimajo dosti skupnega, pri besednih zvezah pa takih dvoumnosti skoraj ni. Na tem mestu je treba opozoriti, da je ta trditev še posebno pomembna za MI v zbirkah strokovnih besedil, ker so strokovni termini zelo pogosto besedne zveze.

Hull in Grefenstette predvidevata tudi, da bi lahko iskanje s povratno zanko¹⁶ postalo v MI še pomembnejše, kot je v enojezičnem IR. Ta način iskanja bi lahko pomenil protiutež neizogibni nenatačnosti, ki jo uvaja prevajanje iskalnih zahtev. Ko bi iskalec po medjezičnem iskanju zbral nekaj relevantnih dokumentov v jeziku prevoda iskalne zahteve, bi sistem te dokumente lahko izkoristil za reformulacijo iskalne zahteve. Zaradi ciklične narave iskanja s povratno zanko, bi se v drugi in naslednjih ponovitvah zanke medjezično iskanje poenostavilo na enojezično iskanje s postopnim izboljševanjem rezultatov. Za uspešno izvedbo takega iskanja je seveda nujno, da iskalec jezik prevoda vsaj delno razume in je sposoben oceniti relevantnost dokumentov v njem. Iskanje s povratno zanko v povprečju zmanjšuje število dokumentov, ki jih je treba pregledati do nekega nivoja priklica, zato bi bilo lahko pri delnem razumevanju jezika prevoda še posebej ustrezno.

Vpliv prevajanja besednih zvez in različnih načinov reformulacije (dopolnjevanja) iskalnih zahtev sta v seriji zelo zanimivih in odmevnih eksperimentov preizkusila tudi Lisa Ballesteros in Bruce Croft (Ballesteros in Croft, 1997). Eksperimenti so bili namenjeni preverjanju uspešnosti prevajanja iz angleščine v španščino. Pri delu sta poleg dvojezičnega slovarja uporabila tudi učno zbirko angleških časopisnih člankov, iskanje pa je potekalo v testni zbirki, sestavljeno iz časopisnih člankov v španščini z znanimi ocenami relevantnosti za dvajsetih iskalnih zahtev. Zbirki sta bili primerljivi po vsebini. Učinkovitost prevajanja in iskanja sta preverjala na standarden način, tako kot je to opisano v poglavju o evalvaciji.

16 *Iskanje s povratno zanko (relevance feedback) je ciklični postopek iskanja, pri katerem iskalec po vsakem iskanju med rezultati označi nekaj relevantnih dokumentov, sistem pa na osnovi teh informacij reformulira iskalno zahtevo. Besedam iz iskalne zahteve, ki se odlikujejo z veliko verjetnostjo pojavljanja v relevantnih dokumentih, poveča povedne moči in s tem njihov vpliv na rangiranje dokumentov, iskalno zahtevo pa razširi z novimi besedami s podobnimi lastnostmi. Steče novo iskanje in ponovno označevanje relevantnih dokumentov. V zaporednih ciklih povratne zanke se iskalna zahteva izboljšuje, relevantni dokumenti se zgoščujejo pri vrhu ranžirne vrste zadetkov, vanjo pa prihajajo tudi novi, ki jih prejšnje oblike iskalne zahteve niso zadele. Iskanje s povratno zanko je najučinkovitejša izvedba ne-Boolovega iskalnega modela, zahteva pa precejšnji procesorski napor in stalno sodelovanje iskalca.*



Slika 2: Shematska predstavitev uporabe korpusov za dopolnjevanje iskalne zahteve v eksperimentih, opisanih v (Balesteros in Croft, 1997).

LPZ = lokalna povratna zanka, LKA = lokalna kontekstna analiza.

Prevajanje besed in besednih zvez v iskalnih zahtevah je potekalo s pomočjo e-verzije Collinsovega angleško-španskega slovarja. Iskalne zahteve so najprej obdelali z označevalnikom besednih vrst in določili za besedne zveze vse pare samostalnik-samostalnik in pridevnik-samostalnik. Pri avtomatskem prevajanju so v slovarju za nekatere besedne zveze obstajali večbesedni prevodi, zveze, ki jih ni bilo v slovarju, ter preostanek iskalne zahteve, pa so bile prevajane kot posamezne besede. Blokiranje besede so bile izvzete iz prevajanja. V eksperimentih z avtomatskim dopolnjevanjem iskalnih zahtev sta avtorja preverila vpliv postopkov, imenovanih *lokalna povratna zanka*¹⁷ in *lokalna kontekstna ana-*

¹⁷ *Povratna zanka je v splošnem postopek, ki omogoča širitev iskalne zahteve z besedami ali besednimi zvezami, ki se izkažejo z največjo količino informacije v dokumentih, relevantnih za to iskalno zahtevo. Pri klasični (interaktivni) povratni zanki relevantne dokumente določi iskalec po predhodnem iskanju (glej opombo št. 16), pri lokalni povratni zanki pa sistem privzame, da so relevantni dokumenti, ki so bili najvišje uvrščeni v predhodnem iskanju.*

*liza*¹⁸, pred in po prevajanju iskalnih zahtev. V prvi seriji poskusov je bilo dopolnjevanje opravljeno **pred prevajanjem** iskalnih zahtev v španščino, torej z angleškimi besedami in besednimi zvezami. V drugi seriji poskusov je bilo dopolnjevanje iskalnih zahtev opravljeno **po** njihovem **prevajanju**, z besedami in besednimi zvezami v španščini. Tretja serija poskusov je predstavljala **kombinacijo** dopolnjevanja pred in po prevajanju. Oba postopka dopolnjevanja iskalnih zahtev, ki ju shematsko prikazuje slika 2, potekata v dveh korakih: (a) začetnem iskanju in (b) avtomatski analizi določenega števila najvišje uvrščenih dokumentov ter dopolnjevanju iskalne zahteve. Za postopek pred prevajanjem je bila uporabljena učna zbirka, dopolnjevanje po prevajanju pa je potekalo na testni zbirki.

Analiza rezultatov je pokazala, da nekatere metode, ki izvirajo tako iz instrumentarija IR (lokalna povratna zanka, lokalna kontekstna analiza) kot tudi iz instrumentarija računalniškega jezikoslovja (označevanje besednih vrst, identifikacija besednih zvez), lahko občutno povečajo uspešnost MI, kadar to temelji na prevajanju s pomočjo dvojezičnih slovarjev. K izboljšanju osnovnega rezultata, ki ga predstavlja prevajanje posameznih besed, temelječih na dvojezičnem slovarju, najbolj prispeva identifikacija in prevajanje besednih zvez, še posebno, če je kombinirana z avtomatskim vključevanjem dodatnih besed in besednih zvez, pridobljenih z analizo relevantnih dokumentov. Opisani metodi širjenja iskalnih zahtev z vključevanjem informacij iz korpusov, nevtralizirata negativni vpliv nenatančnosti, ki jo uvaja prevajanje s pomočjo slovarja. Konkretnije: širjenje pred prevajanjem izboljša splošne pogoje za prevajanje, širjenje po prevajanju pa z uvajanjem novih besed zmanjša pomen napačnih prevodov. Kombinacija obeh pristopov izboljša tako natančnost kot priklic medjezičnega iskanja. Posamezno vključevanje je učinkovitejše, če je opravljeno pred prevajanjem v jeziku iskalne zahteve, s kombiniranjem vključevanja v obeh jezikih pa so opisani eksperimenti dosegli 68% uspešnosti referenčnega enojezičnega iskanja (tabela 1). Učinkovitost širjenja iskalnih zahtev pred in po prevajanju so pokazali že starejši eksperimenti (Ballesteros in Croft, 1996), kjer avtorja trdita, da kombinirana raba lokalne povratne zanke nadomesti polovico izgube povprečne natančnosti, ki je posledica napačnih prevodov iskalne zahteve. Učinek širjenja je še posebno opazen pri kratkih iskalnih zahtevah.

18 *Kontekstna analiza je v splošnem postopek, ki omogoča širitev iskalne zahteve z besedami ali besednimi zvezami, ki se v korpusu največkrat pojavljajo v najbližjem sobesedilu z besedami, ki sestavljajo iskalno zahtevo. Temelji na predpostavki, da besede, ki se pojavljajo v dokumentih skupaj, nosijo podobno vsebino, še posebej v domeni, ki jo zastopa korpus. Lokalna kontekstna analiza zmanjšuje procesorsko zahtevnost postopka s tem, da preverja sopojavljanje besed le v dokumentih, ki so bili v predhodnem iskanju najvišje uvrščeni.*

Tabela 1. Povprečna natančnost medjezičnega iskanja v primerjavi z enojezičnim.

Slovar = prevajanje besed in besednih zvez z dvojezičnim slovarjem, Pred-LPZ = širjenje iskalne zahteve pred prevajanjem s pomočjo lokalne povratne zanke, Pred-LKA = širjenje iskalne zahteve pred prevajanjem s pomočjo lokalne kontekstne analize, Po-LPZ = širjenje po prevajanju z lokalno povratno zanko, Po-LKA = širjenje po prevajanju z lokalno kontekstno analizo, Komb-LPZ = širjenje pred in po prevajanju z lokalno povratno zanko, Komb-LKA = širjenje pred in po prevajanju z lokalno kontekstno analizo. Pri vseh metodah je bilo opravljeno avtomatsko prevajanje besed in besednih zvez. Rezultati izvirajo iz eksperimentov (Balesteros in Croft, 1997).

metoda	% uspešnosti glede na enojezično iskanje
Slovar	41,2
Pred-LPZ	55,0
Pred-LKA	57,0
Po-LPZ	45,8
Po-LKA	51,1
Komb-LPZ	62,2
Komb-LKA	68,0

Tudi v kasnejših eksperimentih (Balesteros in Croft, 1998) sta Ballesterosova in Croft preverjala uporabnost korpusa za izboljševanje MI, temelječega na dvojezičnem slovarju. Korpus je bil dvojezičen, vzporeden na nivoju dokumentov. Metodi, ki spominjata na lokalno povratno zanko in lokalno kontekstno analizo, za razliko od (Balesteros in Croft, 1997), nista bili uporabljeni za širitev iskalne zahteve, ampak za razreševanje dvoumnosti prevajanja besed in besednih zvez. Prevajanje je potekalo iz angleščine v španščino. Prva metoda, razreševanje dvoumnosti z vzporednim korpusom, poskuša najti najustreznejše prevode besed na osnovi informacij, ki izhajajo iz vzporednosti dokumentov. Za 30 najvišje uvrščenih španskih dokumentov, ki jih vrne izvorna, ne prevedena, iskalna zahteva v dvojezičnem korpusu, se poiščejo njihovi angleški ustrezniki. Ti se združijo v en navidezen dokument, besede v njih pa razvrstijo po povedni moči, izračunani po metodi Rocchia (Rocchio, 1971). Od teh besed se jih za razreševanje dvoumnosti pri prevajanju iskalne zahteve uporabi 5.000 najvišje uvrščenih. Ta »seznam 5000« tako predstavlja del korpusa, relevanten za prevajanje iskalne zahteve. Besedam v izvorni iskalni zahtevi se avtomatsko dodajo podatki o besednih vrstah. Vsaki španski besedi se v slovarju poišče njen angleški prevod, ki velja za to besedno vrsto (če takih ni, veljajo vsi, neglede na besedno vrsto). Če je možnih prevodov več, jih metoda poskuša najti v »seznamu 5000«. Ustrezen je tisti, ki je najvišje uvrščen v tem sezna-

mu, če pa takega ni, gredo v prevod iskalne zahteve vsi izrazi, ki jih nudi slovar.

Druga metoda razreševanja težav pri izbiri pravega prevoda polisemih besed temelji na statistični analizi kolokacij. Osnovna predpostavka se glasi: pravilni prevodi parov besed iz iskalne zahteve se v korpusu v povprečju pojavljajo večkrat skupaj kakor nepravilni. Pri paru besed *b1* in *b2* tvorijo množico *s1* vsi možni prevodi besede *b1*, ki ustrezajo njeni besedni vrsti, množico *s2* pa vsi možni prevodi besede *b2*, ravno tako z upoštevanjem besedne vrste. Metoda generira vse možne pare besed iz *s1* in *s2* in izračuna »moč« njihovih kolokacij v korpusu s pomočjo enačbe, opisane v delu avtorjev Xu in Croft (1998). Par z najvišjo vrednostjo je določen kot najustreznejši prevod besed. Pomembno je, da za tak način računanja pomembnosti kolokacij ne potrebujemo vzporednega korpusa. Pri preverjanju učinkovitosti prevajanja je bila referenčna vrednost tokrat prevajanje z dvojezičnim slovarjem besedo za besedo, brez razreševanja polisemije.

Ekspirimenti L. Balesteros in B. Crofta (Balesteros in Croft, 1998) so preizkusili tudi vpliv nekaterih drugih parametrov, med njimi:

Označevanje besednih vrst. Če gredo v prevedeno iskalno zahtevo le prevodi, ki ustrezajo besedni vrsti izvorne besede, se povprečna natančnost poveča za 21,9% v primerjavi z referenčno vrednostjo.

Zmanjševanje vpliva posameznih prevodov polisemih besed na izračun relevantnosti dokumentov. Redkeje uporabljane besede med njimi in neustrezni prevodi pridobijo (zaradi same narave računanja) neupravičeno velik izračun povednih moči, kar ima negativen vpliv na natančnost iskanja. Metoda šteje različne prevode iste besede kot sinonime s skupno dokumentno frekvenco, s čemer se teža ene besede razdeli na vse njene sinonime. Samo ta poseg poveča povprečno natančnost za 44,6%, skupaj z označevanjem besednih vrst pa natančnost naraste celo za 89% v primerjavi z referenčno vrednostjo.

Primerjava uspešnosti razreševanja dvoumnosti prevodov z vzporednim korpusom in analizo kolokacij se izteče izrazito v prid slednje. Eden od možnih razlogov je v tem, da za nekatere iskalne zahteve v korpusu ni (ali skoraj ni) ustreznih dokumentov, zato tudi v »seznamu 5000« ni ustreznih besed. Ta sklep navaja na – sicer intuitivno jasno – dejstvo, da mora biti vsebinska domena korpusa, namenjenega razreševanju dvoumnosti prevodov, čimbolj skladna z domeno, v katero sodijo iskalne zahteve. Metoda z analizo kolokacij doseže 79% povprečne natančnosti enojezičnega iskanja (za primerjavo: prevajanje iskalnih zahtev s sistemom za računalniško prevajanje Systran je v istih okoliščinah doseglo 67%).

Eksperimenti Lise Ballesteros in Bruca Crofta kažejo, da je mogočev MI doseči rezultate, skoraj primerljive z rezultati enojezičnega iskanja, z uporabo metod, ki ne temeljijo na težko dosegljivih ali, za nekatere jezike, celo neobstoječih jezikovnih virih, oz. na sistemih za pravo računalniško prevajanje. Kombiniranje metod za širjenje iskalnih zahtev (Balesteros in Croft, 1997), pazljivo prevajanje besednih zvez, razreševanje polisemije s pomočjo korpusov, vzporednih na nivoju dokumentov, ali z analizo kolokacij, privede povprečno natančnost MI preko 90% povprečne natančnosti enojezičnega iskanja.

Pomen zmanjševanja vpliva posameznih prevodov polisemih besed je potrdil v raziskavah MI med finščino in angleščino tudi Ari Pirkola (Pirkola, 1998). Za prevajanje iskalnih zahtev iz finščine v angleščino je uporabljal dva dvojezična slovarja: splošni in strokovni medicinski. Dokumenti v zbirki, v kateri je potekalo iskanje, so bili časopisni članki s poljudno medicinsko tematiko. Rezultati so pokazali nižjo povprečno natančnost pri prevajanju samo s splošnim slovarjem, in višjo, če je bilo najprej opravljeno prevajanje s strokovnim slovarjem, za neprevedene besede pa še s splošnim.

Še večji vpliv kot kombiniranje slovarjev je imelo strukturirano prevajanje iskalnih zahtev. Prevodi finskih besed v angleščino so bili grupirani v množice tako, da so vsi prevodi ene angleške besede tvorili eno množico. Pri iskanju je iskalni algoritem besede v isti množici tretiral kot sinonime in pri računanju relevantnosti poiskanih dokumentov so se vse frekvence pojavljanja besed iz iste množice računale kot pojavljanja iste besede. V prevedeni iskalni zahtevi je bil tako pri velikih množicah potencialni vpliv posameznega prevoda na končni izračun relevantnosti relativno manjši kot pri manjših množicah. Posledica tako vpeljane strukturiranosti na računanje relevantnosti rezultatov iskanja, je bila dominacija besed z enim samim prevodom (torej z manjšo dvoumnostjo prevoda) in razpršen vpliv besed z mnogimi prevodi. Pirkola poroča, da so bile najpomembnejše besede v iskalnih zahtevah največkrat strokovne besede, ki so imele v povprečju malo različnih prevodov, in s tem velik vpliv na izračun relevantnosti. Kombiniranje slovarjev in strukturiranja se je izkazalo kot zelo uspešno, s povprečno natančnostjo, ki je praktično dosegla povprečno natančnost enojezičnega iskanja (na nekaterih stopnjah priklica pa celo preseгла).

Pirkolin pristop je v osnovi različen od običajnega pristopa k razreševanju polisemije. Ostali, v tem poglavju opisani pristopi, poskušajo izbrati najustreznejši prevod s pomočjo statističnih analiz korpusov ali pa zmanjšati negativni vpliv neustreznih prevodov s širjenjem iskalne zahteve pred ali po prevajanju. Pirkola je dosegel enake ali boljše rezultate na enostavnejši način, s kombiniranjem splošnega in specializiranega slovarja ter s strukturiranjem prevedene iskalne zahteve.

Zanimivo je, da se ena od lastnosti finščine, ki je morfološko zelo kompleksen jezik, namreč zelo pogoste sestavljenke za izražanje konceptov, ki so v drugih jezikih opisani z besednimi zvezami, izkaže kot koristna pri MI. V drugih jezikih lahko precejšen del neuspeha MI pripišemo neuspešni identifikaciji besednih zvez – problemu, ki pri sestavljenkah seveda odpade (Pirkola, 1998). Uspešnost pristopa Pirkole najbrž lahko pripišemo temu, da se je v svojih eksperimentih uspešno lotil dveh od treh osnovnih problemov prevajanja iskalnih zahtev – dvoumnosti, zaradi polisemije, ter pomanjkljive izčrpnosti slovarjev. Polisemijo je nevtraliziral s strukturiranjem iskalne zahteve, izčrpnost slovarjev je bistveno izboljšal s kombiniranjem splošnega in specialnega. Tretji osnovni problem – identifikacija besednih zvez – je v finščini relativno majhen zaradi že omenjenega izražanja zapletenejših pojmov s sestavljanjem besed.

Ruth Sperer in Douglas Oard (Sperer in Oard, 2000) sta pri prevajanju iskalnih zahtev iz kitajščine v angleščino uporabila podobno zamisel kot Pirkola, namreč, da strukturirana iskalna zahteva lahko bolje odraža informacijsko potrebo kot nestrukturirana. Njun pristop poskuša natančneje modelirati iskalne postopke pri sestavljanju iskalne zahteve. Trdita, da to poteka v dveh korakih: (a) z določanjem vsebinskih konceptov, ki najbolje predstavljajo informacijsko potrebo, in (b) izbiranjem besed, ki najbolje izražajo izbrani vsebinski koncept. Beseda ima lahko v dvojezičnem slovarju večje število prevodov, ki opisujejo nekaj različnih vsebinskih konceptov, ti pa različno dobro ustrezajo informacijski potrebi. Če uspe prevajalni algoritem izbrati najustrežnejšega, se z uporabo besed, ki ga zastopajo, po mnenju Spererjeve in Oarda dvoumnost prevoda zmanjša. Postopek avtomatskega prevajanja se torej razdeli na grupiranje prevodov v množice, ki predstavljajo posamezne vsebinske koncepte, ter izbor najustrežnejšega med njimi. Bistvena razlika med njunim in pristopom Pirkole je v tem, da je Pirkola uporabil vse možne prevode besed, vpliv neustreznih pa zmanjšal s strukturiranjem iskalne zahteve.

Spererjeva in Oard sta se lotila grupiranja prevodov s strukturiranjem dvojezičnega slovarja. Med možnimi prevodi neke besede dva zastopata isti vsebinski koncept in sodita v isto množico, če njuna sorodnost presega neko izračunano vrednost. Njuna mera sorodnosti je relativno zapletena, sestavljajo jo mera semantične sorodnosti, izpeljana po metodi Lina (Lin, 1998; citirano v: Sperer in Oard, 2000) iz semantične mreže angleških besed WordNet, ter meri morfološke in ortografske podobnosti.

Naslednji korak je izbira najustrežnejše množice. Stopnja ustreznosti neke množice glede na želeni pomen prevoda je odvisna od množic, ki jih prispevajo ostale besede v iskalni zahtevi, ter od splošne rabe besed iz množice v jeziku prevoda. Metoda, ki sta jo avtorja uporabila, temelji na ocenjevanju ustreznosti vsake od množic prevodov besede *b1* glede na njen odnos do vsake od množic besede *b2*, ki je naslednja beseda v iskalni zahtevi. Ustreznost sta

izračunala kot kombinacijo dveh mer: (a) semantične sorodnosti vsake od množic besede b_1 z vsako od množic besede b_2 ter (b) verjetnosti kolokacij besed iz ocenjevane množice z besedami v vsaki od množic besede b_2 . Semantično sorodnost sta izračunala po metodi Lina, verjetnost kolokacij pa v korpusu z 78 milijoni angleških besed, sestavljenem iz časopisnih člankov. Kot referenčno vrednost pri evalvaciji uspešnosti iskanja sta uporabila metodo strukturiranja iskalne zahteve po Pirkoli (Pirkola, 1998).

Pristop Spererjeve in Oarda je zanimiv, ker uvaja natančnejšo granulacijo množice možnih prevodov besed z izrazito polisemijo in s tem potencialno večjo kontrolo nad dogajanjem. Rezultati eksperimentov so zaenkrat slabši kot pri referenčni metodi Pirkole, in avtorja to pojasnjujeta predvsem s šibkostjo izbire najustreznejše množice možnih prevodov posamezne besede.

Z analizo kolokacij lahko nedvomno zelo zmanjšamo težave pri prevajanju polisemih besed, da pa je pomembna tudi sama velikost dvojezičnega slovarja, na katerem sloni osnovno prevajanje, kažejo rezultati raziskav prevajanja iskalnih zahtev iz japonščine v angleščino (Maeda et al., 2000). Uporabljen je bil dvojezični slovar, sestavljen iz treh različnih virov, ki je vseboval 366.000 gesel (!), zanimiv pa je tudi način ugotavljanja kolokacij, s katerim so razreševali polisemijo. Kolokacije so določali z avtomatsko izvedenim iskanjem s spletnim iskalnikom AltaVista, ob uporabi operatorjev AND in NEAR (učni korpus tako naraste na dovršen del vseh dokumentov na spletu). Opis metode izračunavanja moči kolokacij presega okvir tega članka, ga pa podrobno opisuje Meada in soavtorji (Maeda et al., 2000). Tak obsežen prevajalni aparat je privedel do 97% povprečne natančnosti enojezičnega iskanja.

7 Medjezično iskanje, osnovano na korpusih

O uporabi korpusov v MI je bilo precej povedanega že v poglavju o prevajanju iskanih zahtev s pomočjo dvojezičnih slovarjev. Njihova uporaba je največkrat vezana na pomoč pri prevajanju iskalnih zahtev s slovarji; lahko gre za korpus v jeziku izvorne ali pa prevedene iskalne zahteve. V računalniškem jezikoslovju imajo največjo težo *vzporedni* dvojezični korpusi, poravnani na nivoju stavkov, pri katerih za vsak stavek v enem jeziku obstaja drugojezični ustreznik. Zaradi redkosti in težke dostopnosti se taki korpusi, oblikovani v skladu z zelo formalnimi zahtevami računalniškega jezikoslovja, v MI ne uporabljajo pogosto. Za pomoč pri prevajanju iskalnih zahtev so največkrat v rabi korpusi, poravnani na nivoju dokumentov ali celo *primerljivi* korpusi, sestavljeni iz dokumentov v različnih jezikih, vendar s sorodno tematiko. Prednost imajo seveda korpusi, ki po vsebini sodijo v isto domeno kot iskalne zahteve in dokumenti v zbirki.

Znameniti korpus, na katerem je bilo narejeno precej eksperimentov v MI, je Hansard Corpus¹⁹, vzporedni korpus angleških in francoskih besedil iz Kanadskega parlamenta. Številni primerljivi korpusi izvirajo iz časopisnih člankov sorodnih usmeritev (npr. zunanja politika), zbranih ob istem času, tako da s precejšnjo verjetnostjo govorijo o istih dogodkih.

Raziskave medjezičnega iskanja, ki bi temeljilo le na korpusih, brez uporabe dvojezičnih slovarjev, so redke (primer je Davis in Dunning, 1995) in uspešnost takih metod zaenkrat še ne dosega slovarskih. Precej obetajo raziskave avtomatske gradnje jezikovnih orodij iz korpusov ali celo iz zbirk dokumentov, na katerih poteka iskanje. Namesto »ročne« izdelave slovarjev ali tezavrov, kar je težko izvedljivo in nepraktično, poskušajo analizirati velike količine večjezičnih besedil in iz njih izolirati informacije, potrebne za avtomatsko gradnjo slovarjev ali tezavrov. Metode največkrat temelje na statistični analizi korpusa, ki pa zahteva še integracijo nekaterih lingvističnih postopkov. Cilj je gradnja dvojezičnega slovarja, ki bi vseboval ne le možne prevode besede, ampak tudi podatke o verjetnosti posameznih prevodov, temelječe na njihovi porazdelitvi v korpusu (Brown et al., 1993; citirano v: Hull in Grefenstette, 1996). Pristop je intuitivno zelo privlačen, vendar pa odraža le lastnosti besedil v dani domeni, morda celo le v konkretnem učnem korpusu, uporabnost rezultatov izven domene pa je zelo vprašljiva (Hull in Grefenstette, 1996).

8 Avtomatska gradnja jezikovnih virov

Za uspešen razvoj metodologije MI je očitno odločilnega pomena razvoj jezikovnih virov, kot so dvojezični slovarji, primerni za računalniško rabo, in vzporedni korpusi. Dokaj izčrpen pregled dogajanja na tem področju lahko najdemo v dveh virih (Hovy et al., 1999 in Haddouti, 1999). Po Schaeubleju in Smeatonu (1998) se mora delo usmeriti predvsem v:

- gradnjo standardnih dvo- ali večjezičnih zbirk dokumentov;
- razširitev in združevanje obstoječih jezikovnih virov, razvoj polavtomatskih in avtomatskih postopkov za njihovo dopolnjevanje;
- razširitev jezikovnih virov z novimi jeziki, gradnjo pravih večjezičnih virov namesto dosedanjih dvojezičnih;
- razvoj virov (ontologij in konceptualnih tezavrov), neodvisnih od konkretnih jezikov;
- razvoj standardov in novih postopkov za merjenje kvalitete jezikovnih virov.

19 <http://morph ldc.upenn.edu/ldc/news/release/hansard.html>

Del teh nalog je mogoče dovolj uspešno opraviti s polavtomatskimi ali avtomatskimi postopki. V naslednjem poglavju je opisanih nekaj poskusov gradnje vzporednih in primerljivih korpusov ter dvojezičnih slovarjev. Zanimivo je, da so bili pri gradnji korpusov doseženi presenetljivo dobri rezultati z izrazito enostavnimi hevrstičnimi prijemi. Ti so bili nevsebinske narave, omejeni na strukturne lastnosti spletnih dokumentov ali spletišč (Nie et al., 1999), oz. na površinske podobnosti vsebine, daleč od računalniškega razumevanja naravnega jezika (Braschler in Schaeuble, 1998).

8.1. Primerljivi in vzporedni korpusi

Nie in sodelavci so se lotili gradnje vzporednega korpusa z odkrivanjem jezikovnih parov spletnih dokumentov (Nie et al., 1999). Številne spletne strani obstajajo v dveh jezikovnih verzijah, večinoma v angleščini in lokalnem jeziku. Zbiranje takih strani je zahtevna naloga. Število strani, ki predstavljajo potencialen vir, je ogromno, kvalitetni dokumenti so pomešani s povsem neuporabnimi, zelo različne pa so tudi kvalitete prevodov. Ne obstaja enostaven, za vse situacije uporaben način prepoznavanja jezikovnih parov dokumentov, ampak lahko sestavimo le bolj ali manj uporabne hevrstične metode. Nie in sodelavci trdijo, da je lahko vir teh metod le opazovanje najpogostejših lastnosti uporabnih dvojezičnih parov dokumentov, npr.:

1. dokumenti z vzporednimi prevodi so običajno povezani s kazalci v obe smeri;
2. besedilo sidra kazalca pogosto imenuje jezik dokumenta, na katerega kaže kazalec (pri neangleških dokumentih recimo »in English« ali »English version«);
3. vzporedna besedila imajo pogosto podobna imena, z variabilnim delom, ki označuje jezik (npr. »products_fre.html« in »products_eng.html«); so v isti mapi, ali pa v dveh mapah istega hierarhičnega nivoja, pri čemer imena map spet lahko vsebujejo namig na jezik.

V obsežni raziskavi so zbirali vzporedne dokumente z iskalnikoma AltaVista in Northern Light²⁰. Iskalne zahteve so bile sestavljene tako, da so rezultati ustrezali 1. in 2. lastnosti. Zbrane dokumente so filtrirali glede na 3. lastnost. Iskali so vzporedne prevode v angleščini in francoščini. Avtorji so zbiranje vzporednih dokumentov začeli s pilotsko študijo. Na 60 strežnikih so zbrali 8.000 dvojezičnih parov dokumentov, po filtriranju pa jih je preostalo 4.000 (priklic 50%). Natančnejši pregled 164 naključno izbranih parov dokumentov

²⁰ <http://www.northernlight.com/>

med njimi je pokazal, da jih je bilo 162 v resnici vzporednih prevodov (natančnost preko 95%). Pri resničnem zbiranju vzporednega korpusa so v 75 urah avtomatskega delovanja zbrali 14.200 vzporednih dokumentov (250 Mbytov). Uporaba korpusa pri razreševanju dvoumnosti prevajanja z dvojezičnim teza-vrom je pokazala, da je vzporedni korpus, avtomatsko sestavljen iz spletnih virov, za rabo v MI enako dober, kot ročno sestavljeni namenski korpusi.

Poseben problem je bilo končno preverjanje vzporednosti prevodov, ki ga pri velikem številu dokumentov seveda ni moč opraviti »ročno«. V računalniškem jezikoslovju za to obstajajo programi, ki dokumente vzporejajo na nivoju stavkov. Postopki so natančni, vendar procesorsko prezahtevni za rabo v MI. Nie in sodelavci so uporabili enostavne hevrstične metode, pri katerih sta bila glavna kriterija podobnost struktur, zapisanih s HTML, in podobne dolžine dokumentov. V primerjavi s programi za avtomatsko vzporejanje, je bila napaka teh metod samo 2%.

Braschler in Schaeuble (Braschler in Schaeuble, 1998) sta opravila zanimive eksperimente gradnje primerljivih korpusov (angleško-francoskega in francosko-nemškega) s pomočjo avtomatskega odkrivanja parov dokumentov. Uporabila sta časopisne članke poročevalskih agencij Associated Press in SDA, izdane v letih 1988 do 1990, skupaj 243.000 angleških, 185.000 nemških in 142.000 francoskih člankov. V zbirki ni bilo nobenih podatkov o morebitnih medsebojnih prevodih, zato je bilo treba pare dokumentov določiti na osnovi informacij, vsebovanih v samih dokumentih. Zbirki nemških in francoskih člankov sta bili precej sorodni, zbirki angleških in nemških pa raznorodni, kar je povzročilo različni strategiji odkrivanja parov dokumentov.

Za avtomatsko odkrivanje parov dokumentov sta v njih iskala namige na sorodnost, med njimi:

- lastna imena (pri njih med jeziki večinoma ni razlik v načinu zapisa),
- zapise števil,
- datume objave člankov,
- sorodno izrazje (seveda v različnih jezikih).

Med temi skupinami indikatorjev so le zadnji jezikovno odvisni. Odkrivanje parov je temeljilo na začetnem iskanju, pri katerem je bil dokument v enem jeziku iskalna zahteva za iskanje dokumentov v drugem jeziku, in rangiranju poiskanih dokumentov glede na prisotnost indikatorjev. Najsorodnejši dokument med poiskanimi je bil določen kot primerljivi par. Za iskanje parov v zbirkah francoskih in nemških dokumentov so zadoščali jezikovno neodvisni indikatorji, prevajanje iskalnih zahtev ni bilo potrebno, kar v podobnih situacijah nakazuje možnost gradnje primerljivih korpusov poljubnih jezikovnih parov brez uporabe virov računalniškega jezikoslovja.

Večji problem je predstavljala gradnja primerljivega korpusa angleških in nemških dokumentov. Iz besed angleškega članka, za katerega je iskal morebitni nemški par, je sistem sestavil iskalno zahtevo tako, da je ohranil le besede s srednjo meddokumentno frekvenco. Tiste, ki se pojavljajo v več kot 10% dokumentov, imajo premajhno diskriminacijsko vrednost, zelo redke pa prinašajo naključne povezave. Za prevajanje iskalne zahteve sta avtorja uporabila enostaven seznam 86.000 angleških besed s prevodi, brez dodatnih lingvističnih in frekvenčnih informacij, ki sta ga sestavila iz javnih virov na internetu. Pri prevajanju sta uporabila vse možne prevode izvirne besede. Pri rangiranju poiskanih dokumentov, ki so potencialni pari izvirnega dokumenta, sta igrala pomembno vlogo datuma izdaje v obeh jezikih.

Vzorčenje rezultatov določanja parov člankov je pokazalo, da v 38,4% primerov oba članka govorita o istem dogodku, v 20,4% primerov članka govorita o sorodnih dogodkih, v 1,9% primerov je ena od vsebin člankov ista, v 14,4% primerov je skupen dobršen del terminologije, v 24,9% primerov pa med člankoma v obeh jezikih ni očitne povezave. Za gradnjo primerljivega korpusa bi torej lahko uporabili več kot 75% avtomatsko določenih parov dokumentov.

Metodologija Braschlerja in Schaeubleja je obetajoča, čeprav ostaja odprto vprašanje o njeni uporabnosti pri poljubnih dvojezičnih zbirkah dokumentov. Deloma namreč temelji na datumih izdaje dokumentov – dokazih primerljivosti torej, ki so uporabni le za zbirke časopisnih člankov. Posebna prednost metodologije je v tem, da uporablja zelo malo težko dostopnih formalnih jezikovnih virov.

8.2. Tezavri kolokacij in slovarji

Dvojezični korpusi so v MI pomembni jezikovni viri, vendar je njihova vloga omejena predvsem na pomoč pri prevajanju iskalnih zahtev s slovarji. Zato so še toliko pomembnejše raziskave avtomatske gradnje dvojezičnih slovarjev.

Gradnja slovarjev je procesorsko naporen postopek, ki zahteva vzporedne jezikovne vire, ti pa so že sami po sebi redki in težko dostopni. Nekateri raziskovalci, ki so imeli na voljo prave dvojezične korpuse, vzporedne na nivoju stavkov, poročajo o uspešni gradnji dvojezičnih slovarjev. Yang in sodelavci (Yang et al., 1997; Brown, 1997) poročajo o uporabi slovarja, zgrajenega na osnovi 685.000 parov stavkov velikega dvojezičnega korpusa. Iskanje z iskalnimi zahtevami, prevedenimi s tem slovarjem, je na istem korpusu doseglo 91% povprečne natančnosti enojezičnega iskanja.

Zanimiv približek slovarjem so t. i. »similarity thesauri«, med iskanjem boljšega prevoda jih imenujmo *tezavri kolokacij*, za katere niso potrebni vzporedni korpusi. Pojem in osnovna metodologija gradnje izvirata iz raziskav avtomatskega dopolnjevanja iskalnih zahtev v *enojezičnem* okolju. Tezaver kolokacij je struktura, v kateri je znanje o vsebinski domeni zbirke (korpusa) opisano s podatki o sorodnosti izrazov v tej zbirki (korpusu). Sorodnost dveh izrazov se računa s pomočjo statistične analize njunih kolokacij v dokumentih. Metoda predpostavlja, da imata besedi, ki se pogosto pojavljata v istih dokumentih, sorodno vsebino, pri čemer so dokumenti obteženi. Za neko besedo *b* so kolokacije v dokumentih z večjo težo pomembnejše od tistih v dokumentih z manjšo težo, dokumentu pa daje težo ocena, kako pomembna je v njem vsebina, ki jo zastopa beseda *b*. Lahko bi rekli, da so besede indeksirane z dokumenti – na prvi pogled sprevržena situacija v svetu, v katerem dokumente običajno indeksiramo z besedami. V skladu s tem poteka tudi računanje teže dokumentov. Lahko se uporabi klasična shema *tf.idf*, znana iz modelov vektorskega prostora (Salton in Buckley, 1988), pri kateri pa sta vlogi dokumentov in besed zamenjani. Postopek opisujeta Qui in Frey (Qui in Frey, 1993), njegov rezultat pa je obsežen tezaver kolokacij, v katerem je vsaki besedi iz zbirke pripisan seznam najsorodnejših besed - tistih, s katerimi se največkrat pojavlja v dokumentih z največjo težo. Kvaliteta tezavra v splošnem narašča z velikostjo zbirke (korpusa), vsekakor pa relacije med besedami veljajo samo za to zbirko, v najboljšem primeru (pri dovolj veliki in reprezentativni zbirki) za isto vsebinsko domeno, v katero sodi zbirka.

Na ETH v Zürichu uporabljajo podoben pristop za gradnjo *dvojezičnih* tezavrov kolokacij. V primerljivem dvojezičnem korpusu z avtomatskimi postopki določijo dvojezične pare dokumentov z najbolj podobno vsebino in jih združijo v navidezne dvojezične dokumente. S statistično analizo kolokacij za vsak izraz v enem jeziku določijo rangiran seznam najsorodnejših izrazov; zaradi dvojezičnosti dokumentov so med njimi tudi najverjetnejši prevodi. Tak tezaver učinkovito uporabljajo v poskusih MI (Sheridan in Schaeuble, 1997), celo pri MI govornih dokumentov (Sheridan, Wechsler in Schaeuble, 1997). Sheridan in sodelavci imenujejo uporabo tezavrov kolokacij pri prevajanju iskalnih zahtev *pseudoprevajanje*. To se zdi pošteno poimenovanje, kajti besedo v izvornem jeziku zamenja skupina verjetno sorodnih besed, brez zagotovila, da je med njimi tudi pravi prevod. Trdijo pa, in njihovi rezultati to potrjujejo (Sheridan in Ballerini, 1996), da so tako prevedene iskalne zahteve popolnoma uporabne za MI. V naravi gradnje tezavrov kolokacij je, da grupirajo besede, ki se pojavljajo v dokumentih s sorodnimi vsebinami, iskanje dokumentov, sorodnih iskalni zahtevi, pa je ravno cilj vsakega iskanja, ne samo MI. Pri gradnji dvojezičnega tezavra kolokacij na način, ki ga opisujejo Sheridan in sodelavci, je kritična faza iskanje parov sorodnih dokumentov; bolje ko je opravljena ta faza, večja je verjetnost, da bo analiza kolokacij res pripeljala do možnih prevodov. V opisanem eksperimentu je bilo iskanje parov dokumentov opravljeno na zbirkah

nemških in francoskih časopisnih člankov, že omenjenih v tem poglavju (Braschler in Schaeuble, 1998) in z isto metodologijo.

Razvoj dvojezičnih tezavrov kolokacij obeta, vendar je zaenkrat vezan predvsem na eno inštitucijo. Prave dvojezične slovarje potrebujemo med drugim tudi zato, ker so tezavri kolokacij zelo tesno vezani na vsebino učnega korpusa, na katerem so bili zgrajeni.

Gradnja pravih dvojezičnih slovarjev, namenjenih računalniškemu prevajanju, zahteva obsežen instrumentarij računalniškega jezikoslovja. Dober primer so raziskave Carbonella in sodelavcev (Carbonell et al., 1997), kjer so za avtomatsko gradnjo dvojezičnega slovarja uporabili dvojezični korpus, poravnan na nivoju stavkov, sestavljen pretežno iz angleškega in španskega dela večjezičnega korpusa Združenih narodov. Med izjeme sodi sistem SABLE (Scalable Architecture for Bilingual Lexicography, Melamed, 1996a; Melamed, 1997). SABLE je sposoben izdelave dvojezičnega slovarja na osnovi vzporednih besedil, neporavnanih na nivoju stavkov, in pri tem ne potrebuje nujno nobenih posebnih jezikovnih virov in orodij, razen delilnikov na besede (tokenizers), čeprav obstoječi slovarji in krnilniki (stemmers) lahko občutno povečajo njegovo učinkovitost. Bistvo postopka je pretvorba besedil v obeh jezikih v skupno dvodimenzionalno reprezentacijo, v kateri so besede opisane s pozicijo v koordinatnem sistemu, ta pozicija pa je izračunana na osnovi sopojavljanja besed v jezikovnih parih dokumentov. Na tej reprezentaciji algoritmi za prepoznavanje vzorcev iščejo podobnosti med točkami in jih interpretirajo kot prevodne pare besed. Postopke podrobno opisuje Melamed (Melamed, 1996b). Za začetno orientacijo v koordinatnem sistemu SABLE uporablja vrsto hevrističnih postopkov in morebitne obstoječe slovarje. Med hevrističnimi postopki je posebno koristno iskanje kognatov, besed, ki se v različnih jezikih pišejo enako ali podobno – npr. osebna imena, toponimi, tehnični izrazi... Pomembna lastnost sistema je ta, da deluje na nivoju besed in se ne ozira na višje strukture besedila, zato ne potrebuje prevodov s poravnanimi stavki. Geometrijska reprezentacija besedil, na kateri delujejo algoritmi za vzporejanje besed, je abstraktna, in avtor sistema trdi, da ga je zato možno uporabiti za poljubne jezikovne pare. SABLE je bil prvotno razvit za besedila v angleščini in francoščini, kasneje pa uporabljen tudi za špansko-angleške in korejsko-angleške jezikovne pare.

V številnih eksperimentih MI se je izkazalo, da je mogoče precejšen del inferirnosti MI v primerjavi z enojezičnim iskanjem pripisati pomanjkanju strokovnih izrazov v dvojezičnih slovarjih, s katerimi se prevajajo iskalne zahteve. Resnik in Melamed (Resnik in Melamed, 1997) poročata o (pol)avtomatski gradnji dvojezičnega strokovnega slovarja. Uporabila sta sistem SABLE in relativno majhen korpus (410.000 besed), vzporeden na nivoju dokumentov. Korpus je sestavljala tehnična dokumentacija z računalniškega področja v angleščini in francoščini. Postopek gradnje slovarja je potekal v treh korakih: (a) avto-

matski izbor dvojezičnih besednih parov (naloga SABLE), (b) avtomatsko izločanje besed, ki sodijo v »splošni« jeziku, in (c) ročna redakcija preostalih besednih parov. Za filtriranje besed splošnega jezika (korak b) sta avtorja uporabila e-verzijo Collinsovega slovarja. Evalvacija postopka je pokazala 89% natančnost pri 30–40% priklicu strokovnih izrazov. Med pravilnimi pari besed bi jih bilo 56% mogoče vključiti v slovar brez popravljanja, za ostale pa bi bila potrebna še ročna redakcija. Priklic je v taki situaciji lahko le približno ocenjen in avtorja trdita, da je ocena zelo konzervativna.

9 Zaključek

Medjezično iskanje dokumentov je v zadnjih letih v svetu med najživahnejšimi raziskovalnimi usmeritvami na področju shranjevanja in iskanja informacij. Medtem, ko je metodologija gradnje zbirk polnih dokumentov, njihovih vsebinskih opisov in iskanja v njih v enojezičnem okolju že precej dozorela (že leta imamo tako čvrste teoretične modele kot tudi delujoče sisteme), pa je medjezično iskanje še precej na začetku poti. Opravljeni so bili sicer eksperimenti (nekaj jih je opisanih tudi v tem članku), pri katerih je bila učinkovitost medjezičnega zelo blizu učinkovitosti enojezičnega iskanja, vendar so bile okoliščine prirejene testiranju metod in daleč od nepredvidljivih situacij realne rabe.

Videti je, da se zaenkrat najbolje obnesejo pristopi, ki kombinirajo »močna« orodja IR in »šibka« orodja računalniškega jezikoslovja. Najuspešnejši so tisti sistemi, ki vsaj do neke mere razrešujejo dvoumnost prevajanja polisemih besed, uporabljajo slovarje, prilagojene domeni zbirke in identificirajo besedne zveze v iskalni zahtevi. Pri razreševanju polisemije besed v slovarjih se izkažejo predvsem metode iz arzenala IR, ki po eni strani poskušajo besedi izbrati pravi prevod za dano iskalno zahtevo (analiza kolokacij, lokalna kontekstna analiza) ali pa poskušajo zmanjšati vpliv napačnih prevodov (uvajanje dodatnih besed v iskalno zahtevo s pomočjo lokalne povratne zanke, tretiranje vseh možnih prevodov kot sinonimov s skupno dokumentno frekvenco). Pri gradnji prevodnih slovarjev za dano domeno so koristne tako metode računalniškega jezikoslovja (gradnja pravih slovarjev) kot tudi metode IR (gradnja tezavrov kolokacij). Metode računalniškega jezikoslovja pa so najučinkovitejše za določanje besednih zvez v iskalni zahtevi.

Nekatere od opisanih metod temeljijo na predprocesiranju učnih korpusov (npr. gradnja tezavrov kolokacij) ali na njihovi rabi sočasno z iskanjem (lokalna kontekstna analiza in povratna zanka), učinkovite pa so tembolj, čimbolj je korpus vsebinsko soroden zbirki, po kateri poteka iskanje. Treba bo šele videti,

kako se bodo take metode obnesle v dinamičnem okolju hitro se spreminjajočih zbirk, pri katerih se lahko s časom spreminja tudi težišče vsebine.

Članek poskuša biti pregled dogajanja na področju medjezičnega iskanja, vendar se nekaterih tem ni niti dotaknil. Taka tema je recimo latentno semantično indeksiranje, ki že leta veliko obeta, vendar avtor ne pozna medjezičnih eksperimentov na dovolj velikih zbirkah. Neobdelan je ostal problem računanja relevantnosti in rangiranja rezultatov iskanja v večjezični zbirki (kako uvrstiti v skupno ranžirno vrsto dokumente v različnih jezikih, pri katerih so bili parametri, iz katerih so bile izračunane relevantnosti, pridobljeni v različnih pogojih?). Velike težave povzroča tudi transliteracija in transkripcija pri gradnji večjezičnih zbirk dokumentov, napisanih s pisavami, ki se močno razlikujejo med seboj in od latinice – arabskih, kitajskih, japonskih... Tudi slovenščina ni nedolžna; trenutno uporabljamo vsaj štiri standarde kodiranja znakov. Zanimiv problem, vendar rešljiv enostavneje, kot bi bilo pričakovati, je avtomatsko določanje jezika dokumenta ali dela dokumenta, če je ta dolg vsaj nekaj besed, tako da omogoča enostavno statistiko.

Med vsemi neobdelanimi temami je seveda najpomembnejša stanje medjezičnega iskanja in primernih jezikovnih virov pri nas. Ta tema pa zahteva samostojen članek in avtorja, ki je vpeljan v njene skrivnosti.

10 Citirani viri

(Vsi kazalci, navedeni ob referencah in v opombah pod črto, so bili preverjeni februarja 2002.)

1. **Ballesteros, L., & Croft, B.** (1996). Dictionary methods for cross-lingual information retrieval. V *Proceedings of the 7th International DEXA Conference on Database and Expert Systems* (str. 791-801). URL: <http://citeseer.nj.nec.com/ballesteros96dictionary.html>
2. **Ballesteros, L., & Croft, W. B.** (1997). Phrasal translation and query expansion techniques for cross-language information retrieval. V *Proceedings of the 20th ACM SIGIR Conference on Research and Development in Information Retrieval*. URL: <http://www.cfar.umd.edu/~kanungo/cmssc828K/clara/p84-ballesteros.pdf>
3. **Ballesteros, L., & Croft, W. B.** (1998). Resolving ambiguity for cross-language retrieval. V C. J. Van Rijsbergen, W. B. Croft, A. Moffat, (Ur.), *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (str. 64-71). ACM Press. URL: <http://citeseer.nj.nec.com/ballesteros98resolving.html>

4. **Braschler, M., & Schaeuble, P.** (1998). Multilingual information retrieval based on document alignment techniques. V C. Nikolau, C. Stephanidis (ur.), *Lecture Notes in Computer Science. Second European Conference on Research and Advanced Technology for Digital Libraries ECDL98*, Crete.
5. **Brown, P., Pietra, S. D., Pietra, V. D., & Mercer, R.** (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19 (2), 263-311.
6. **Brown, R. D.** (1997). Corpus-based query translation for translanguag information retrieval. *Position paper for SIGIR-97 workshop on Cross-Lingual Information Retrieval*. URL: <http://www.cs.cmu.edu/~ralf/papers/querytrans.ps>
7. **Carbonell, J. G., Yang, Y., Frederking, R. E., Brown, R. D., Geng, Y., Lee, D.** (1997). Translanguag information retrieval: a comparative evaluation. V *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence*. URL: <http://citeseer.nj.nec.com/carbonell97translanguag.html>
8. **Davis, M., & Dunning, T.** (1995). A TREC evaluation of query translation methods for multi-lingual text retrieval. V Harman DK (ur.) *The 4th Text Retrieval Conference (TREC-4)*. NIST. URL: <http://trec.nist.gov/pubs/trec4/papers/nmsu.ps.gz>
9. **Eichmann, D., Ruiz, M. E., & Srinivasan, P.** (1998). Cross-language information retrieval with the UMLS metathesaurus. V *Proceedings of the 21th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (str. 72-80). URL: <http://citeseer.nj.nec.com/218119.html>
10. **Gilarranz, J., Gonzalo, J., & Verdejo, F.** (1997). Language-independent text retrieval with the EuroWordNet multilingual semantic database. V *Second Workshop on Multilinguality in the Software Industry: The AI Contribution*. URL: <http://sensei.ieec.uned.es/NLP/papers/mulsaic97.ps>
11. **Haddouti, H.** (1999). Survey: multilingual text retrieval and access. *Working notes of the AAAI Symposium on Cross Manguage Text and Speech Retrieval*. URL: <http://www.forwiss.tu-muenchen.de/~haddouti/survey.ps>
12. **Hovy, E., Ide, N., Frederking, R., Mariani, J., & Zampolli, A.** (ur.). (1999). Multilingual information management: current levels and future abilities. A report Commissioned by the US National Science Foundation and also delivered to the European Commission's Language Engineering Office and the US Defense Advanced Research Projects Agency, Chapter 2. URL: <http://www.cs.cmu.edu/~ref/mlim/index.html>
13. **Hull, D. A., & Grefenstette, G.** (1996). Querying across languages: A dictionary-based approach to multilingual information retrieval. V *Proceedings of the 19th ACM SIGIR Conference on Research and Development in Information Retrieval*.

14. **Lin, D.** (1998). An information-theoretic definition of similarity. V *Fifteenth international conference of machine learning ICML-98*. Madison, USA.
URL: <ftp://ftp.cs.umanitoba.ca/pub/lindek/papers/sim.ps.gz>
15. **Maeda, A., Sadat, F., Yoshikawa, M., & Uemura, S.** (2000). Query term disambiguation for web cross-language information retrieval using a search engine. V *Proceedings of the 5th International Workshop Information Retrieval with Asian Languages*. URL: <http://db-www.aist-nara.ac.jp/~aki-mae/pub/IRAL00-e.pdf>
16. **Melamed, I. D.** (1996a). Automatic construction of clean broad-coverage translation lexicons. V *Proceedings of the 2nd Conference of the Association for machine translation in the Americas*. Montreal. URL: <ftp://ftp.cis.upenn.edu/pub/melamed/papers/amta96.ps.gz>
17. **Melamed, I. D.** (1996b). A geometric approach to mapping bitext correspondence. V *First Conference on Empirical Methods in Natural Language Processing (EMNLP'96)*, Philadelphia, USA. URL: <ftp://ftp.cis.upenn.edu/pub/melamed/papers/emnlp96.ps.gz>
18. **Melamed, I. D.** (1997) A scalable architecture for bilingual lexicography. *Dept. of Computer and Information Science Technical Report #MS-CIS-91-01*.
URL: <ftp://ftp.cis.upenn.edu/pub/melamed/papers/sabletr.ps.gz>
19. **Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K.** (1993). Introduction to WordNet: An on-line lexical database. V *Five Papers on WordNet*.
URL: <ftp://ftp.cogsci.princeton.edu/pub/wordnet/5papers.pdf>
20. **Nie, J.-Y., Simard, M., Isabelle, P., & Durand, R.** (1999). Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the web. V *Proceedings of the 22th ACM SIGIR Conference on Research and Development in Information Retrieval*, Berkeley, USA.
URL: <http://www.xrce.xerox.com/people/isabelle/publications/sigir99.ps>
21. **Oard, D. W.** (1997a) Cross-language text retrieval research in the USA. V *The 3rd ERCIM DELOS Workshop*, Zurich. URL: <http://www.clis.umd.edu/dlrg/filter/papers/delos.ps>
22. **Oard, D. W.** (1997b). Cross-language information retrieval. *SIGIR-97 tutorial*.
URL: <http://www.clis2.umd.edu/dlrg/filter/papers/tutnotes.ps>
23. **Oard, D. W., Dorr, B. J.** (1996). A survey of multilingual text retrieval. *Technical Report UMIACS-TR-96-19*. University of Maryland. URL: <ftp://ftp.cs.umd.edu/pub/papers/papers/ncstrl.umcp/CS-TR-3615/CS-TR-3615.ps.Z>
24. **Oard, D. W., Dorr, B. J., Hackett, P. G., & Katsova, M.** (1998). A comparative study of knowledge-based approaches for cross-language information retrieval. *Technical Report CLIS-TR-98-01*. University of Maryland. URL: <ftp://ftp.cs.umd.edu/pub/papers/papers/ncstrl.umcp/CS-TR-3897/CS-TR-3897.ps.Z>

25. **Pevzner, B. R.** (1972). Comparative evaluation of the operation of the Russian and English variants of the »Pusto-Nepusto-2« system. *Automatic Documentation and Mathematical Linguistics*, 6 (2), 71-74.
26. **Pirkola, A.** (1998). The effects of query-structure and dictionary setups in dictionary-based cross-language information retrieval. *V Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (str. 55-63).
27. **Qiu, Y., & Frei, H. P.** (1993.) Concept based query expansion. *V Proceedings of the 16th ACM SIGIR Conference on Research and Development in Information Retrieval*, Pittsburgh (str. 160-169). URL: <http://citeseer.nj.nec.com/qiu93concept.html>
28. **Radwan, K.** (1994). *Vers l'Acces Multilingue en Langage Naturel aux Bases de Donnes Textuelles*. PhD thesis. Paris: Universite de Paris-Sud.
29. **Resnik, P., & Melamed, I. D.** (1997). Semi-automatic acquisition of domain-specific translation lexicons. *V Proceedings of the 7th ACL Conference on Applied Natural Language Processing*, Washington, DC. URL: <http://citeseer.nj.nec.com/42076.html>
30. **Rocchio, J. J.** (1971). Relevance feedback in information retrieval. *V Salton G (ur.), The SMART retrieval system* (str. 313-323). Englewood Cliffs: Prentice Hall.
31. **Salton, G.** (1970). Automatic processing of foreign language documents. *Journal of the American Society for Information Science*, 21 (3), 187-194.
32. **Salton, G.** (1973). Experiments in multi-lingual information retrieval. *Information processing letters*, 2 (1), 6-11. TR 72-154. URL: <http://cs-tr.cs.cornell.edu/Dienst/UI/1.0/Display/ncstrl.cornell/TR72-154>
33. **Salton, G., & Buckley, C.** (1988). Term weighting approaches in automatic text retrieval. *Information Processing & Management*, 24 (5), 513-523.
34. **Schaeuble, P., & Smeaton, A. F.** (1998). An international research agenda for digital libraries. *Summary report of the series of joint NSF-EU working groups on future directions for digital library research*. URL: http://www.ercim.org/publication/ws-proceedings/DELOS-B/dl_sum_report.pdf
35. **Sheridan, P., & Ballerini, J. P.** (1996). Experiments in multilingual information retrieval using the Spider system. *V Proceedings of the 19th ACM SIGIR Conference on Research and Development in Information Retrieval*. URL: <http://citeseer.nj.nec.com/sheridan96experiments.html>
36. **Sheridan, P., & Schaeuble, P.** (1997). Cross-language information retrieval in a multilingual legal domain. *V First European Conference on Research and Advanced Technology for Digital Libraries*. URL: <http://citeseer.nj.nec.com/sheridan97crosslanguage.html>

37. **Sheridan, P., Wechsler, M., & Schaeuble, P.** (1997). Cross-language speech retrieval: establishing a baseline performance. *V Proceedings of the 20th ACM SIGIR Conference on Research and Development in Information Retrieval*.
URL: <http://citeseer.nj.nec.com/142488.html>
38. **Sperer, R., & Oard, D. W.** (2000). Structured translation for cross-language information retrieval. *V Proceedings of the 23th ACM SIGIR Conference on Research and Development in Information Retrieval*. Athens.
URL: <http://citeseer.nj.nec.com/298892.html>
39. UNESCO (1971). *Guidelines for establishment and development of multilingual scientific and technical thesauri for information retrieval*. Paris: UNESCO.
40. **Vorhees E.** (1994). Query expansion using lexical-semantic relations. *V Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, Dublin (str. 61-69).
41. **Vossen, P.** (1998). EuroWordNet: building a multilingual database with wordnets for European languages. *The ELRA Newsletter*, 3 (1), 7-10.
URL: <http://www.hum.uva.nl/~ewn/docs/ELRARTF.zip>
42. **Xu, J., & Croft, W. B.** (1998). Corpus-based stemming using co-occurrence of word variants. *ACM Transactions on Information Systems*, 16 (1), 61-81.
URL: citeseer.nj.nec.com/32742.html
43. **Yang, Y., Brown, R. D., Frederking, R. E., Carbonell, J. G., Geng, Y., & Lee, D.** (1997). Bilingual corpus-based approaches to translingual information retrieval. *V 2nd Workshop on Multilinguality in Software Industry: The AI Contribution (MULSAIC'97)*. Nagoya.

Dr. Jure Dimec je zaposlen na Inštitutu za biomedicinsko informatiko, Medicinska fakulteta, Univerza v Ljubljani
Naslov: Vrazov trg 2, 1000 Ljubljana
Naslov elektronske pošte: jure.dimec@mf.uni-lj.si