Avtomatsko luščenje izrazja iz slovensko-angleških vzporednih besedil

Špela Vintar
Filozofska fakulteta
Oddelek za prevajalstvo
Aškerčeva 2, Ljubljana
Tel./Fax: +386 1 241 1500
E-pošta: spela.vintar@guest.arnes.si

Abstract

The paper describes the design and structure of a Slovene-English term extraction system. Although the state-of-the-art systems operate on hybrid approaches using various levels of linguistic analysis, sometimes including semantic information, the aim here was to implement both statistical and linguistically motivated methods for both languages and compare the results. It is shown that some methods typically used in Information Retrieval can successfully be implemented for term extraction purposes, and we use cognates to supplement the results of word alignment. In general, as was expected, the linguistically motivated approach preforms better, however both precision and recall could still be improved by fine-tuning and implementation of further components. In the last two sections we briefly discuss the problems of evaluation and applicability.

1. UVOD

Avtomatsko luščenje terminologije iz zbirk dokumentov ali korpusov je ena od ključnih jezikovnih tehnologij, ki se uporablja na mnogih področjih. Preko strokovnega izrazja se prenaša in organizira znanje, zato je terminologija tesno povezana z ontološkim modeliranjem področij in izkopavanjem podatkov in znanja, hkrati pa je avtomatsko razpoznavanje in obteževanje terminov v dokumentih neizogiben korak pri sodobnih iskalnikih podatkov.

V dvojezičnem kontekstu, ko izrazje luščimo iz vzporednih ali primerljivih korpusov in obenem ugotavljamo tudi prevodno ustreznost, predstavljajo avtomatske metode obetavno podporo prevajalskemu delu, ko lahko za potrebe določenega projekta iz obstoječih prevodov izdelamo področni terminološki glosar. Podobno so taki avtomatsko izdelani dvojezični glosarji lahko podpora ali podlaga za terminografsko delo, ki s tradicionalnimi slovarskimi metodami ne zmore dohajati naglega razvoja strokovnih področij.

Avtomatsko luščenje izrazov iz korpusov je v okviru računalniškega jezikoslovja dobro znano in precej raziskano področje, kjer lahko za številne metode trdimo, da so uspešne in dosegajo uporabne rezultate. Pristopi se v grobem delijo na pretežno statistične, ki temeljijo na pogostosti in sopojavljanju in tako ne zahtevajo oblikoskladenjske predobdelave besedil, in t. i. nadzorovane, kjer poleg pogostosti za razpoznavanje izrazov uporabljamo oblikoskladenjske vzorce oziroma pravila. Za ugotavljanje prevodne ustreznosti izrazov v dveh jezikih potrebujemo še dvojezični leksikon, ki ga je mogoče izdelati z orodjem za avtomatsko poravnavo besed ali fraz.

2. NAMEN IN ZGRADBA PRISPEVKA

Čeprav še zdaleč ne moremo trditi, da je problem luščenja izrazja povsem rešen, in to še toliko manj v dvojezičnem okolju, se pri razvoju te tehnologije za povsem nov par jezikov srečamo s številnimi ovirami. Ena od njih je zagotovo (ne)obstoj in (ne)razpoložljivost temeljnih jezikovnih virov in orodij, kot so korpusi, leksikoni, tezavri na eni in tokenizatorji, oblikoskladenjski označevalniki in razčlenjevalniki na drugi strani. Za slovenski jezik je bilo sicer narejenega že marsikaj, kar nenazadnje dokazuje tudi pričujoči zbornik in njegovi predhodniki, a ne glede na to nam še vedno manjkajo prosto dostopna orodja za skoraj vse osnovne jezikovnotehnološke postopke.

Izhajajoč iz tega dejstva ter iz predpostavke, da v številnih uporabniških situacijah orodja za jezikoslovno predobdelavo slovenskih besedil ne bodo na voljo, je predstavljeni sistem zasnovan v dveh različicah. Prva različica deluje na podlagi minimalne jezikovne predobdelave in temelji na zgolj statističnem luščenju večbesednih izrazov, druga pa uporablja oblikoskladenjsko označena besedila. Taka zasnova omogoča primerjavo rezultatov ter oceno "dodane vrednosti", ki jo prispeva oblikoskladenjsko označevanje.

Drug izziv, povezan z luščenjem izrazja, pa je najti ustrezno kombinacijo in sosledje metod, jih prikrojiti izbranemu paru jezikov tako, da bo sistem proizvajal uporabne rezultate, in te ovrednotiti.

Prispevek opisuje zgradbo sistema za luščenje izrazja iz vzporednih korpusov. V 3. razdelku najprej predstavimo uporabljene vire in orodja, nato pa se osredotočimo na posamezne metode. Tako 4. razdelek v prvem delu govori o merjenju terminološkosti in načinih, kako ločevati

strokovne izraze od nestrokovnih, v drugem se posveča metodam za iskanje kolokacij in večbesednih enot, v tretjem pa obravnava reševanje problema gnezdenih terminov. Zatem v 5. razdelku opišemo uporabljeni program za besedno poravnavo ter metodo iskanja sorodnic, v 6. pa so predstavljeni prvi rezultati obeh različic sistema. V zaključku razpravljamo o možnostih izboljšave in nadaljnjih načrtih.

3. KORPUS IN PREDOBDELAVA

V raziskavi je bil uporabljen vzporedni slovensko-angleški korpus strokovnih besedil (TRANS)[1], ki je nastal v okviru študentskega projekta na Oddelku za prevajalstvo Filozofske fakultete v Ljubljani v študijskem letu 2001/2002. Korpus obsega približno milijon pojavnic, vsa besedila so stavčno poravnana z orodjem DéjàVu[2]. Za razvoj in preskušanje metod luščenja izrazja sta bila izbrana dva podkorpusa, in sicer zbirka besedil s področja jedrske tehnike v obsegu 25.000 pojavnic in zbirka pravno-gospodarskih besedil, tj. zakonov, ki urejajo gospodarstvo in finance, v obsegu 166.000 pojavnic.

3.1. Obdelava slovenskih besedil

Prvi korak pri obdelavi slovenskih besedil je tokenizacija, ki je bila izvedena z uporabo lastnega orodja. Stavčna segmentacija ni bila potrebna, saj so besedila stavčno poravnana in tako že segmentirana v prevodne enote.

Za statistično različico sistema to že zadošča, za jezikoslovno obogateno različico pa je bilo potrebno besedila še oblikoskladenjsko označiti in lematizirati. Oblikoskladenjsko označevanje je opravil progam TnT (Brants, 2000), ki se po preskusih sodeč (Džeroski et al., 1999) med statističnimi označevalniki za slovenščino najbolje odreže. Za učni korpus je bil uporabljen ročno označeni izsek iz korpusa ELAN[3] (Erjavec, 2002), nekatere ponavljajoče se napake pa smo popravili še naknadno.

Lematizacijo slovenskih besedil je prijazno prispevalo podjetje Amebis d. o. o.[4] Ker je pri številnih slovenskih pojavnicah možnih več lem, Amebisov lematizator pa označi vse in ne razrešuje večpomenskosti, smo uvedli nekaj grobih pravil za izbiro ene od danih lem. V primerih, kjer je ena od lem bistveno pogostejša od ostalih *(je → **biti**, jesti, on)*[5], vselej izberemo to, v ostalih, manj jasnih primerih, pa preprosto izberemo prvo od ponujenih lem. Ker je lematizacija v našem primeru pomembna predvsem zaradi bolj natančnega merjenja pogostosti in sopojavnosti, je bistveno, da se besedne oblike zlijejo v skupno lemo, tudi če je ta napačna.

3.2. Obdelava angleških besedil

Za obdelavo angleških besedil je bilo uporabljeno orodje ShProt, katerega uporabo je prijazno dovolil institut DFKI GmbH.[6] ShProt vključuje tokenizator, besednovrstni označevalnik TnT, lematizacijo (MMorph) in razpoznavanje samostalniških, predložnih in glagolskih zvez (Chunkie; Skut in Brants 2000).

3.3. Zapis označenih besedil

Označena vzporedna besedila so zapisana v obliki XML. Dokument je sestavljen iz prevodnih enot, ki vsebujejo vzporedni slovenski in angleški segment. Vsak segment nato vsebuje pojavnice, pri katerih so oblikoskladenjski podatki zapisani kot atributi.

Sistem kot eno izhodno možnost ponuja označevanje izrazov in njihovih ustreznic v samem dokumentu, za kar je oblika XML nadvse prikladna. K vsaki prevodni enoti se doda element <terms>, ki vsebuje najdene slovenske in angleške termine ter predloge ustreznic, k vsakemu od teh pa pripišemo še utež oziroma koeficient zaupanja. Primer označenega besedila z izrazi kaže Slika 2.

4. METODE AVTOMATSKEGA LUŠČENJA IZRAZOV

Navadno postopki za avtomatsko iskanje strokovnega izrazja v dvojezičnem okviru v grobem zajemajo naslednje korake:

- enojezična identifikacija terminov v obeh delih vzporednega korpusa,
- ugotavljanje prevodne ustreznosti med najdenimi termini,
- izdelava dvojezičnega leksikona parov terminov ali označevanje le-teh v besedilih.

Opisani sistem pri zadnji točki nudi obe izhodni možnosti.

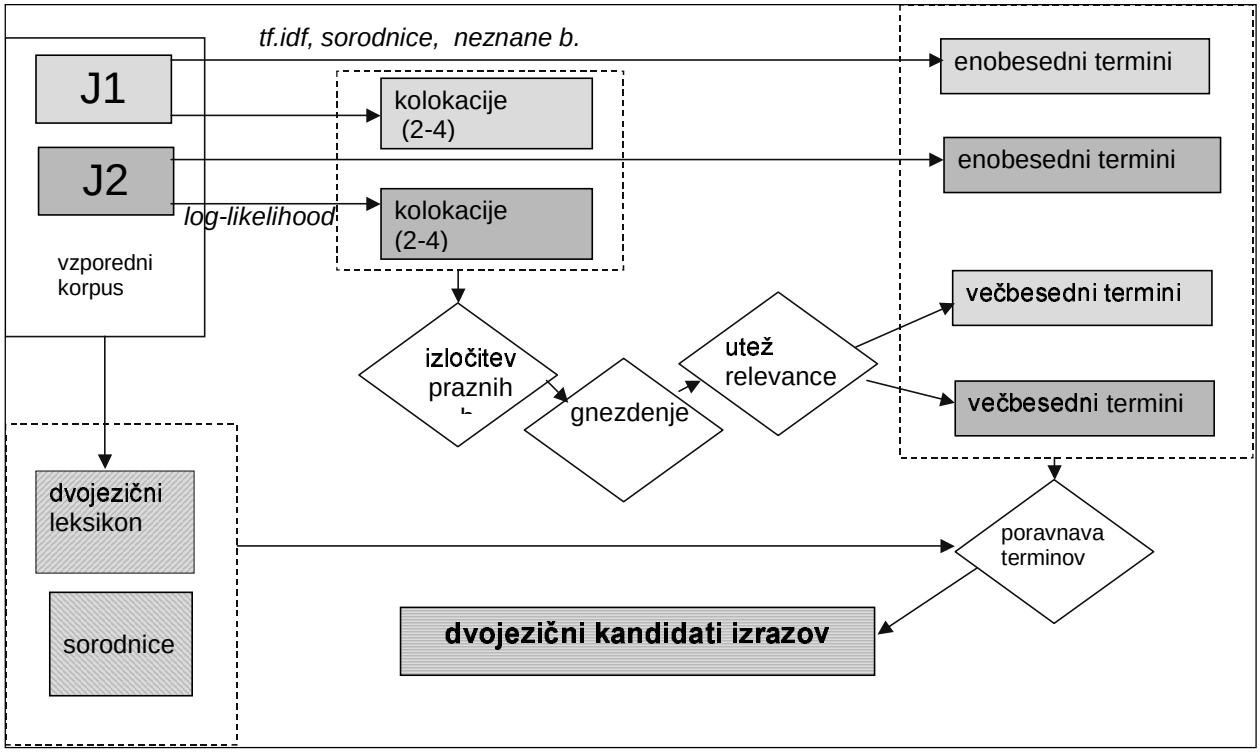
4.1. Merila terminološkosti

Prvo vprašanje, s katerim se pri tem srečamo, je kako meriti terminološkost oziroma kako razlikovati strokovne izraze od splošnega besedišča. Tu sta se uveljavili dve metodi; prva izvira iz korpusne leksikologije in temelji na *relativni pogostosti pojavnic*, druga pa s področja iskanja podatkov (Information Retrieval) in temelji na merilu *tf.idf*.

4.1.1. Relativna pogostost pojavnic

Strokovno izrazje navadno luščimo iz korpusa besedil določene stroke. Za strokovni jezik pri tem

[1] http://nl2.ijs.si/corpus/index2-bi.html
[2] http://www.atril.com
[3] http://nl.ijs.si/elan
[4] http://www.amebis.si
[5] Lema *on* se pojavi kot osnovna oblika osebnega zaimka tretje osebe ednine, katerega ženska oblika je *ona*, naslonska oblika v rodilniku pa *je*.
[6] http://www.dfki.de



Slika 1: Statistični model

predpostavljamo, da uporablja bodisi lastne izraze, ki jih v besedilih splošnega jezika ne najdemo, bodisi pomensko osredotočene ali premaknjene izraze splošnega besedišča. Za oboje je utemeljeno domnevati, da se bodo v strokovnih besedilih pojavljali pogosteje kot v splošnih, zato za ugotavljanje "ključnosti" določene besede za izbrano stroko lahko uporabimo primerjavo pogostosti besede v strokovnem in v splošnem korpusu (Manning/Schütze, 1999).

Pokaže se, da ta metoda razmeroma dobro označuje specifičnost izraza znotraj določene stroke, vendar le, če sta tako strokovni kot primerjalni (splošni) korpus dovolj velika in homogena oziroma uravnotežena. Predvsem so neugodni strokovni korpusi, ki sicer zajemajo besedila določene stroke (npr. pravo) in so torej besedilnovrstno homogeni, posamezna besedila pa vendar vsebujejo izrazje iz zelo različnih področij (npr. zakon o bančništvu, zakon o ekonomskih conah).

4.1.2. Koeficient tf.idf

Na področju priklica informacij je že dlje v uporabi drug način merjenja ključnosti besed, ki namesto na pojmu korpusne pogostosti temelji na zbirki dokumentov. Naloga priklica informacij je, da za določeno poizvedbo prikliče točno tiste dokumente, ki vsebujejo v poizvedbi vsebovane ključne izraze večkrat kot ostali dokumenti. Koeficient tf.idf (Term Frequency – Inverted Document Frequency) se definira kot teža termina *i* v zbirki dokumentov *j*, kjer je N število dokumentov, tf_{ij} je pogostost termina v dokumentu, df_i pa število dokumentov, ki vsebujejo termin *i*.

$$weight(i, j) = (1 + \log(tf_{ij})) \log \frac{N}{df_i}$$

V našem primeru se je pokazalo, da to merilo pri obeh podkorpusih bolj zanesljivo pokaže termine kot relativna pogostost. Pri večbesednih terminih je mogoče terminološkost izraziti kot vsoto koeficientov tf.idf posameznih besed, deljeno s številom besed. Seveda lahko tf.idf računamo tudi za kolokacije, vendar je pri tem treba upoštevati različice terminov (npr. *reaktorsko hladivo* in *reaktorsko hladilo*), ki neustrezno nižajo pogostosti kolokacij.

4.1.3. Druge metode ugotavljanja terminološkosti

Med termine pogosto sodijo tudi besede, ki jih izluščimo posredno oziroma jih pridobimo kot "stranski proizvod" pri drugih postopkih. To so na primer:

- neznane besede,
- sorodnice.

Pri prvih mislimo na besede, ki jih ne najdemo v slovarju in jih zato tudi jezikovna orodja, npr. črkovalnik ali

lematizator, ne prepoznajo. Med njimi sicer najdemo tudi precej napačno zapisanih besed, mnoge pa so neznane zaradi terminološkosti.

Sorodnice *(cognates)* v predstavljenem sistemu sicer uporabljamo za dopolnitev dvojezičnega leksikona, a se pogosto tudi pri teh pokaže, da gre za strokovne izraze tujega izvora, ki jih zaradi sorodne pisave avtomatsko najdemo (npr. *dekontaminacija/decontamination*).

4.2. Iskanje večbesednih enot

V strokovnih jezikih so termini večinoma sestavljeni iz več besed, zato je prepoznavanje večbesednih enot osrednjega pomena. Poleg vprašanja, kako učinkovito prepoznavati nize dveh ali več zaporednih besed, ki sodijo skupaj in tvorijo termin, se tu srečamo še s problemom **gnezdenja** (npr. *zunanji cevni vod* in *cevni vod*) ter s problemom **terminološke variabilnosti** (npr. *obod zvara* in *zvarni obod*).

Na področju luščenja izrazja sta v rabi dva pristopa za iskanje večbesednih terminov. Prvi je statistični in iz korpusa izlušči ponavljajoče se kombinacije besed oz. kolokacije, pri čemer za merjenje relevantnosti kolokacije obstaja vrsta različnih cenilk. Drugi izkorišča oblikoskladenjsko analizo in išče termine na podlagi vnaprej določenih skladenjskih vzorcev. Prednosti statističnega pristopa sta predvsem jezikovna neodvisnost in enostavnost glede potrebnih ravni predanalize, glavni pomanjkljivosti pa predvsem, da je rezultate statistično predlaganih kolokacij navadno treba pošteno presejati, da postanejo uporabni, in seveda da statistične metode odkrivajo zgolj tiste večbesedne enote, ki se v korpusu pojavijo večkrat. Pristop z vzorci tega problema nima, ima pa še to prednost, da jezikoslovno motivirani pristopi na splošno omogočajo boljše reševanje gnezdenja in variabilnosti.

Ker je osrednji namen opisane raziskave primerjava statističnih in jezikoslovno motiviranih pristopov, smo sistem zasnovali v obeh različicah. Statistična različica tako išče kolokacije dolžine od dveh do štirih besed, za merilo kolokativnosti pa uporabljamo cenilko **logaritem verjetja** (log-likelihood - LL), ki se je v številnih preskusih za te namene izkazala za najzanesljivejšo (Manning/Schütze, 1999). Poleg tega uporabljamo sezname praznih besed za izločevanje nekoristnih kolokacij. Reševanje gnezdenja je opisano v razdelku 4.3. Model statističnega sistema kaže Slika 1.

V jezikoslovno bogatejši različici uporabljamo vzorce oziroma zaporedja oblikoskladenjskih oznak, na primer *pridevnik + samostalnik*, kar se ujema s termini kot *tlačna črpalka, regulacijski ventil, jedrska elektrarna*. Čeprav je vzorce z regularnimi izrazi mogoče definirati tudi kot končne avtomate, ki opisujejo slovnico večbesednih enot, se naš sistem omeji na vzorce, dolge do štiri besede, in sicer samo na samostalniške zveze. Tako je za slovenščino določenih 19 vzorcev, pri vsakem pa je označeno tudi jedro besedne zveze. Žal se pri vzorcih razkrije kakovost označevanja, saj se morebitne napake pokažejo bodisi v obliki nepravih zvez bodisi v obliki nižjega priklica. Shema pristopa z vzorci je podobna statistični, le da namesto kolokacij luščimo vzorce ter spustimo izločevanje na podlagi praznih besed.

4.3. Gnezdenje

Predvsem statistična metoda proizvede veliko gnezdenih kandidatov, ki jih zgolj s sejanjem na podlagi terminološke relevantnosti ne moremo izločiti. Tako na primer najdemo dvojki *sistem reaktorskega* in *reaktorskega hladiva*, hkrati pa tudi trojko *sistem reaktorskega hladiva*. Potrebujemo torej način za ugotavljanje, katero od obeh gnezdenih dvojk obdržati. Enako metodo uporabimo tudi pri izbiri trojk znotraj četvork itd.

Lokalni maksimum LL

Prva metoda se odloča na podlagi primerjave vrednosti LL (log-likelihood) med posameznimi dvojkami; merimo torej moč povezave med dvema besedama.

previous | steam | generator | replacement
34,17 602,05 77,88

Pri zgornjem primeru bi iz angleškega termina radi odstranili besedo *previous*, zato primerjamo vrednosti LL in odstranimo besedo, ki ima najšibkejšo povezavo. Ta metoda deluje dobro, vendar je težko določiti mejno vrednost LL tako, da ne bi izgubili preveč pravilnih terminov.

Vrednost C

Frantzi in Ananiadou (1996) predlagata drug način, ki naj bi posredoval informacijo o stabilnosti besedne zveze v korpusu. Za vsako večbesedno enoto *a* se izračuna t.i. **vrednost C** (C-value) po spodnji formuli:

$$C(a) = (dolžina(a) - 1)(pogostost(a) - \frac{t(a)}{c(a)})$$

kjer je *t(a)* sešteta pogostost *a* v daljših kolokacijah, *c(a)* pa število daljših kolokacij, ki vsebujejo *a*. S to preprosto, čeprav na prvi pogled morda ne povsem jasno, tehniko merimo, kako ustaljena je besedna zveza. Tako ima dvojka *compressive force* v spodnjem primeru višjo vrednost C kot *axial compressive*, ki jo želimo izločiti.

N-gram	Vrednost C
compressive force	10,3
axial compressive	5,2
axial compressive force	16,4

Tabela 1: Reševanje gnezdenja z vrednostjo C

Ker se je ob primerjavi obeh metod slednja izkazala za bolj zanesljivo, je v sistem vgrajeno izločanje gnezdenih izrazov na podlagi vrednosti C.

5. IZDELAVA DVOJEZIČNEGA LEKSIKONA

5.1. Besedna poravnava

Besedna poravnava *(Word Alignment)* je izraz za tehnologijo statističnega pridobivanja leksikonov prevodnih ustreznic iz vzporednih korpusov. Metoda temelji na predpostavki, da je za vsako besedo v korpusu iz prevodnih segmentov, kjer se pojavlja, mogoče izračunati najverjetnejšo ustreznico.

Danes je za avtomatsko izdelavo dvojezičnih leksikonov na voljo več prosto dostopnih programov, npr. Giza++ (Och, 2002), Plug (Tiedemann, 1999) in Twente (Hiemstra, 1998). Po neuspešni instalaciji Gize++ in površnem spoznavanju programa Plug je zmagal že znani program Twente (prim. Vintar, 2000). Besedna poravnava se vrši z algoritmom IPFP, program pa za vsako besedo v korpusu predlaga eno ali več ustreznic skupaj z verjetnostmi prevoda. Rezultate besedne poravnave lahko izboljšamo, če iz korpusa prej odstranimo prazne besede.

5.2. Sorodnice

Sorodnice *(cognates)* so besede, ki so si po zunanji podobi, navadno pa tudi po pomenu, podobne, npr. *sistem/system, informacija/information, Francija/France*. Med njimi najdemo precej imen in kratic, večinoma pa gre za besede neoklasičnega izvora, ki so si v številnih jezikih podobne. V jezikovnih tehnologijah se sorodnice izkorišča predvsem za izboljšanje stavčne poravnave (Simard in dr. 1992), zanimive pa so tudi s prevodoslovnega vidika (Vintar/Hansen 2002). V pričujočem okviru jih uporabljamo kot dopolnitev besedne poravnave, in sicer iz dveh razlogov:

- Sorodnice je razmeroma enostavno avtomatsko najti.
- Pari sorodnic so zelo zanesljivi pari prevodnih ustreznic.

Komponenta za odkrivanje sorodnic uporablja perlov modul String::Approx, ki omogoča primerjavo znakovnih nizov na podlagi Levensteinove metode dodajanja, odvzemanja in zamenjave znakov. Na ta način najdemo besede, ki so si po pisavi podobne, raven podobnosti pa lahko uravnavamo z določitvijo še dovoljene *razdalje* med nizoma. Iz podkorpusa jedrskega inženiringa smo pridobili 364 parov sorodnic, iz podkorpusa zakonodaje pa 774.

Ker se je pri pregledu rezultatov pokazalo, da so pari sorodnic skoraj vedno tudi pravi pari prevodnih ustreznic, nam te služijo za dopolnitev dvojezičnega leksikona, pri čemer jim priredimo zelo visoko verjetnost (0,90). To pomeni, da pri izboru ustreznice prednost navadno dobi sorodnica, če obstaja.

5.3. Iskanje ustreznic terminov

Ko smo izluščili termine obeh jezikov in presejali sezname kandidatov, se znajdemo pred naslednjo nalogo – poiskati pare prevodnih ustreznic. Pri enobesednih terminih so ustreznice seveda vključene že v dvojezični leksikon, problem predstavljajo le primeri, kjer se enobesedni termin prevede v večbesednega ali obratno (npr. *uparjalnik/steam generator*). Če je večbesedni termin povsem ustaljen, je ta pojav mogoče ugotoviti na ravni besedne poravnave, saj dobimo za *uparjalnik* ustreznici *steam* in *generator*, vsako z verjetnostjo 0,50. Ta metoda je zanesljiva le pri pogostih terminih in le, če smo za izdelavo dvojezičnega leksikona uporabili lematizirani korpus.

V vseh ostalih primerih pa skušamo ustreznico večbesednega termina najti tako, da za vsako od besed v izvirnem terminu v leksikonu poiščemo ustreznice, nato pa med ciljnimi termini najdemo tistega, ki vsebuje največ ustreznic. Zanesljivost prevoda izrazimo s seštevkom prevodne ustreznosti vsake posamezne ustreznice.

projekt	zamenjave	uparjalnikov
↓	↓	↓
project 1.00	[null] 1.00	steam 0.49
		generator 0.33
		generators 0.18

Tu med angleškimi termini največ ustreznic vsebuje izraz *steam generator replacement project*, ki ga izberemo s prevodno verjetnostjo 1,82.

Ta način omogoča dokaj precizno ugotavljanje parov ustreznic, njegova prednost pa je, da razmeroma dobro deluje tudi z nelematiziranimi termini. Kljub temu opazimo, da za številne kandidate v enem jeziku ne najdemo ustreznice, zato je skupno število dvojezičnih parov manjše od števila najdenih terminov v posameznih jezikih.

6. REZULTATI

Sistem zaenkrat predvideva dve možnosti izpisa rezultatov. Pri prvi izpišemo le pare ustreznic, za katere je tako prag terminološkosti kot prag ustrezanja dovolj visok, s čimer zagotovimo natančnost, a močno znižamo priklic. Tabela 2 kaže primer, kjer se *Term. rel.* nanaša na terminološko relevantnost, *Prev. ustr.* pa na stopnjo prevodnega ustrezanja.

Drug način izpisa označuje izraze v besedilu v zapisu XML. Tu označujemo izraze za vsak jezik posebej, nato pa še ustreznice, pri čemer so poleg enojezičnih terminov navedeni še terminološka relevantnost ter indeksi pojavnic v besedilu, kjer je bil izraz odkrit, pri ustreznicah pa navedemo stopnjo ustrezanja.

Sl. termin	Term. rel.	An. termin	Prev. ustr.	Term. rel
blokirani legi	7.001	blocked position	1.95	5.728
cevovodih sistema	7.455	piping system	1.57	13.206
delovni paket	8.214	work package	1.36	7.189
drsnim sistemom	8.003	sliding system	0.77	8.505
dvizxnim sistemom	8.003	sliding system	0.78	8.505
francisovi turbini	7.040	Francis turbine	2	10.430
glavne komponente	9.869	main components	1.44	9.237
natancxno nacxrtovanje	9.112	detailed planning	1.11	4.428
nominalno tlacxno	8.413	nominal pressure	1.8	13.710
opticxne meritve	9.484	optical survey	1.9	10.264
opticxnih meritvah	7.635	optical measurement	1.12	11.966
padec tlaka	9.624	reduce pressure	1.09	13.599
razlicxnimi turbulentnimi modeli	7.530	different turbulence models	2.79	10.290
razmerja FG FGcr	8.048	ratia FG FGcr	2.51	9.808
turbulentne kineticxne energije	6.998	turbulent kinetic energy	1.87	8.916

Tabela 2: Pari ustreznic

6.1. Vrednotenje rezultatov

Vprašanje, kako vrednotiti rezultate sistema za avtomatsko luščenje izrazja, še zdaleč ni enostavno. Prva nejasnost je že sama definicija termina, saj se tu srečujemo z zelo različnimi pogledi in praksami, ki so tesno povezane s področjem aplikacije sistema ter potrebami uporabnikov.

Zanimiva ilustracija tega, kako različne so terminološke potrebe in prakse, je opisana v raziskavi Rose Estopà (1999), kjer so štiri skupine uporabnikov terminologije (terminologi, prevajalci, tehnični pisci in bibliotekarji) označevale termine v visoko strokovnem medicinskem besedilu. Pokazalo se je, da se je število označenih terminov med skupinami in znotraj skupin dramatično razlikovalo, predvsem pa je bil presek vseh štirih skupin – se pravi odstotek terminov, ki so jih izbrale vse štiri – zgolj 10%.

Pri vrednotenju je možnih več scenarijev. Eden od njih je, da seznam avtomatsko izluščenih kandidatov predložimo strokovnjaku, ki ga ovrednoti, vendar lahko s tem merimo zgolj natančnost sistema, ne pa tudi priklica. Drug način pa vključuje ročno označevanje izrazov v besedilu in s tem izdelavo evalvacijskega materiala, s katerim primerjamo avtomatsko pridobljene rezultate.

V našem primeru je bilo zaradi časovnih in organizacijskih omejitev pridobivanje evalvacijskega materiala oteženo, zato je trenutno na voljo le eno ročno označeno besedilo v slovenščini, v katerem so študenti prevajanja označili termine in terminološke kolokacije. Skupno število označenih terminov v slovenskem besedilu je znašalo 181, Tabela 3 pa kaže rezultate obeh različic sistema na istem besedilu. *Stat.* se nanaša na statistični pristop brez oblikoskladenjskega označevanja, *Jez.* pa na pristop z označevanjem ter vzorci. *Število ujemajočih* pomeni presek najdenih z ročno označenimi izrazi, v tretji vrstici pa

```
<document id="Cerjak.al">
 <tu id="Cerjak.2">
<terms>
   <slterm string="jedrska elektrarna" tokens="1 2" score="5.677" />
   <slterm string="zamenjavi uparjalnikov" tokens="13 14" score="15.424" />
<slterm string="modernizacije elektrarne" tokens="21 22" score="12.956" />
<slterm string="jedrska elektrarna krsxko" tokens="1 2 3" score="7.7613" />
   <slterm string="projektov modernizacije elektrarne" tokens="20 21 22" score="10.468" />
   <slterm string="izmed projektov modernizacije elektrarne" tokens="19 20 21 22" score="8.719" />
   <enterm string="nuclear power" tokens="3 4" score="10.754" />
   <enterm string="power plant" tokens="4 5" score="11.620" />
   <enterm string="replacement project" tokens="15 16" score="9.537" />
   <enterm string="modernization projects" tokens="23 24" score="9.178" />
   <enterm string="consortium siemens" tokens="30 31" score="7.326" />
   <enterm string="nuclear power plant" tokens="3 4 5" score="12.497" />
   <enterm string="krsxko nuclear power plant" tokens="2 3 4 5" score="13.064" />
   <equiv string="jedrska elektrarna krsxko" equiv="krsxko nuclear power plant" transscore="1.43" />
   <equiv string="jedrska elektrarna" equiv="nuclear power plant" transscore="0.74" />
  <equiv string="projektov modernizacije" equiv="modernization projects" transscore="1.84" />
</terms>
<seg lang="SL">
    ....
</seg>
<seg lang="EN">
    ....
 </seg>
</tu></document>
```

Slika 2: Primer izpisa XML

podamo še število parov ustreznic, ki jih je sistem odkril na podlagi vzporednega besedila.

	Stat.	Jez.
Število najdenih izrazov	288	184
Število ujemajočih	69	96
Število parov ustreznic	101	36
Priklic	0,39	0,53
Natančnost	0,23	0,52

Tabela 3: Primerjava rezultatov obeh različic sistema

Kot bi pričakovali, statistični sistem najde več izrazov kot jezikoslovni, vendar so rezultati slabše kakovosti, zato se jih tudi manj ujema z ročno označenimi termini. Ob pregledu parov ustreznic se pokaže, da je pravilnih le okrog 60%. Jezikoslovni sistem ponudi manj terminov, vendar se jih več ujema z ročno označenimi, pri tem pa pregled seznama daje vtis, da so tudi ostali ponujeni izrazi nadvse uporabni, a jih označevalci iz takih ali drugačnih razlogov niso uvrstili med termine. Najdenih ustreznic je malo, a so z izjemo dveh pravilne vse.

Pri obeh sistemih sta priklic in natančnost nekoliko nizka, vendar moramo tu upoštevati dvoje:

- Sistem primerjamo z ročnim označevanjem, ki je vselej subjektivno, tu pa še toliko bolj, ker je šlo za študente in ne strokovnjake s področja.
- Sistem izraze in ustreznice izločuje na podlagi intuitivno določenih mejnih vrednosti terminološke relevantnosti, vrednosti C in stopnje prevodne ustreznosti. Vse te vrednosti so nastavljive – njihovo optimalno konfiguracijo bodo pokazali obširnejši eksperimenti.

Kljub temu rezultati kažejo tudi, kje je sistem še potrebno izboljšati. V sedanjih različicah ni vključena obdelava terminoloških variacij, ki bi znatno povečala priklic pri iskanju prevodnih ustreznic. Pri statistični metodi bi bilo variacije možno odkrivati s pomočjo primerjave znakovnih nizov, npr. *reaktorsko hladivo* in *reaktorsko hladilo* (podobno kot pri sorodnicah), pri jezikoslovni pa bi v vzorce lahko vključili tudi transformacijska pravila.

Med načrti za izboljšave je tudi odkrivanje semantičnih lastnosti terminov in povezav med njimi. Tako bi bilo razmeroma preprosto vključiti še komponento za iskanje izpisanih različic akronimov, saj se te pogosto pojavljajo v oklepajih poleg krajšave ali obratno, npr. *Jedrska elektrarna Krško (JEK)*. V nekaterih strokah je za osnovno semantično klasifikacijo možno uporabiti tipične morfeme, npr. [-lnik] kot morfemski znanilec semantičnega razreda *naprava/stroj/pripomoček*, ali [-anje, -enje] kot znanilec *postopka/metode*.

7. ZAKLJUČEK

V prispevku smo opisali zgradbo sistema za avtomatsko luščenje izrazja iz vzporednih korpusov. Predstavili smo glavne metode obeh različic, statistične in jezikoslovne, in navedli ter ovrednotili rezultate obeh različic. Kot bi pričakovali, se zgolj statistični pristop izkaže za nekoliko manj učinkovitega, saj sta tako priklic kot natančnost slabša, kot če uporabimo oblikoskladenjsko obdelavo besedil. Če je ta na voljo, imamo več manevrskega prostora tudi za uresničitev zgoraj omenjenih načrtovanih izboljšav.

Uporabnost sistema je v tej fazi še težko oceniti, vendar bi že v obstoječi različici lahko nudil podporo prevajalskemu in terminološkemu delu. Ker se v prevajalskih okoljih pomnilniki prevodov vse bolj uveljavljajo, s tem nastajajo tudi vzporedni korpusi kot vir za luščenje izrazja. Hkrati so večjezične terminološke zbirke vse bolj standardizirane in izmenljive, kar bi omogočalo učinkovito dopolnjevanje človeškega dela in avtomatskih postopkov.

Posredno izkušnja z gradnjo opisanega sistema kaže tudi na pomanjkljivo stanje oziroma dostopnost jezikovnih tehnologij za slovenski jezik, saj se v okviru tako kompleksne naloge pokažejo številne potrebe. Mednje bi – poleg želje po prosto dostopnem lematizatorju – sodilo še razpoznavanje imen, kosanje na besedne zveze (*chunking)*, slejkoprej pa tudi področni in splošni tezavri.

Zahvala

Zahvaljujem se prof. dr. Ulrichu Heidu (IMS Stuttgart) za dragocene nasvete in podporo, podjetju Amebis d. o. o. za lematizacijo besedil, dr. Tomažu Erjavcu za tehnično pomoč pri izdelavi korpusa TRANS in študentom Oddelka za prevajalstvo za zbiranje besedil in zavzetost pri omenjenem korpusnem projektu.

VIRI

Brants, T. (2000). TnT - A Statistical Part-of-Speech Tagger. V: Proceedings of 6th ANLP Conference, Seattle, WA.

Džeroski, S.; Erjavec, T. in Zavrel, J. (1999). Morphosyntactic Tagging of Slovene: Evaluating PoS Taggers and Tagsets. Jožef Stefan Institute Research Report IJS-DP 8018, 1999. http://nl.ijs.si/lll/

Erjavec, T. (2002). The IJS-ELAN Slovene-English Parallel Corpus. International Journal of Corpus Linguistics, 7(2), v izidu.

Estopa, R. (1999). Extracció de terminologia: elements per a la construcció d'un SEACUSE (Sistema d'Extracció Automatica de Candidats a Unitats de Significació Especializada). Doktorska disertacija, Universidad Pompeu Fabra, Barcelona.

Frantzi, K.T. in Ananiadou, S. (1996). Extracting nested collocations. V: 16thConference on Computational Linguistics, COLING, str. 41-46.

Heid, U. (1999). Extracting Terminologically Relevant Collocations from German Technical Texts. V: Sandrini, Peter (ed.): Terminology and Knowledge Engineering (TKE '99), Innsbruck. Wien: TermNet, *str. 241*-255.

Hiemstra, D. (1998). Multilingual Domain Modelling in Twenty-One: Automatic Creation of a Bi-directional Translation Lexicon from a Parallel Corpus. V: Coppen, Peter-Arno et al, (ur.): Proceedings of the 8th CLIN meeting, 41-58. http://wwwhome.cs.utwente.nl/~hiemstra/

Manning, Ch., D.; Schütze, H. (1999). Foundations of Statistical Natural Language Processing. The MIT Press Cambridge MA.

Simard M., Foster G. in Isabelle P. (1992). Using Cognates to Align Sentences in Bilingual Corpora. V: Proceedings of the 4th Conference on Theoretical and Methodological Issues in Machine Translation (TMI). Montréal, Québec

Skut W. in Brants T. (1998). A Maximum Entropy partial parser for unrestricted text. V: Proceedings of the 6th ACL Workshop on Very Large Corpora (WVLC), Montreal.

Vintar, Š. (1999). Računalniško podprto iskanje terminologije v slovensko-angleškem korpusu. *Uporabno jezikoslovje*, 7-8, Ljubljana.

Vintar, Š. in Hansen, S. (2002). Cognates – free rides, false friends or stylistic devices: A bilingual comparative study. V: Meaningful Texts: The Extraction of Semantic Information from Monolingual and Multilingual Texts, Birmingham University Press (v tisku).