Razvoj algoritma za samodejno prepoznavanje krajšav in krajšavnih razvezav v elektronskih besedilih

Mojca Kompara

Cobiss: 1.01

Namen prispevka je predstaviti razvoj algoritma za samodejno prepoznavanje krajšav in krajšavnih razvezav v slovenskih elektronskih besedilih. Prepoznavanje krajšav poteka na leksikalni oz. besedni ravni z opazovanjem lastnosti krajšav in krajšavnih razvezav ter sovpadanja. Algoritem prepozna krajšave na podlagi pravil za prepoznavanje, razvezave pa išče v sobesedilu ob upoštevanju pravil sovpadanja. V prispevku predstavljamo algoritem na podlagi filtriranja petih letnikov dnevnika Delo, s katerim v 30 minutah izluščimo 5820 kandidatov za krajšavno-razvezavne pare, ki so potem ročno čiščeni. Natančnost algoritma je 96,75-odstotna.

Ključne besede: krajšave, razvezave, algoritmi

Developing an algorithm for automatic recognition of acronyms and expanding acronyms in electronic texts

This article presents the development of an algorithm for automatic recognition of acronyms and expanding acronyms in electronic Slovenian texts. Recognizing acronyms takes place at the lexical level by observing the qualities of acronyms, expanding acronyms, and their correspondence. The algorithm recognizes acronyms based on recognition principles, and it seeks their expanded forms in context while taking into account principles of correspondence. This article presents an algorithm based on filtering five years of the newspaper *Delo* in which 5,820 potential acronym-expansion pairs were extracted in 30 minutes and then cleaned up manually. The accuracy of the algorithm is 96.75 percent.

Keywords: acronyms, expansion, algorithms

0 Uvod

Na spletu[1] je na voljo vedno več spletnih krajšavnih slovarjev in zbirk. Spletni slovar ima lahko zelo obsežen geslovnik, ki se preprosto in hitro ažurira in tako vključuje najnovejše krajšave. Nekateri spletni slovarji vključujejo nove krajšave samodejno,

[1] Prispevek je prirejeno poglavje avtoričine disertacije Algoritem za samodejno prepoznavanje krajšav in krajšavnih razvezav v elektronskih besedilih (mentor doc. dr. Primož Jakopin), obranjene julija 2010 na Oddelku za primerjalno in splošno jezikoslovje Filozofske fakultete Univerze v Ljubljani.

kar jim omogočajo algoritmi za prepoznavanje krajšav in krajšavnih razvezav. Algoritem predstavlja niz postopkov in operacij oz. pravil, ki so potrebni za razrešitev danega problema. Tudi naravnim jezikom v veliki meri vladajo pravila, ki pa imajo popolnoma drugačen status kot pravila, ki so nujna za opisovanje kombinacij simbolov formalnega jezika (Chiari 2007: 8). Algoritem mora predvideti vse potrebne korake za razrešitev problema in poleg tega pregledati vse razpoložljive podatke, torej je program, ki predstavlja poteze, potrebne za razrešitev problema. Program zahteva formalni jezik ali kodo, ki na abstrakten način opredeli razrede in odnose med razredi abstraktnih elementov (Chiari 2007: 7).

V prispevku je predstavljen razvoj algoritma za samodejno prepoznavanje krajšav in krajšavnih razvezav v elektronskih besedilih. Njegov cilj je prepoznavanje krajšav in razvezav v slovenskih elektronskih besedilih, predvsem kratic in akronimov, prikazana pa je tudi univerzalnost algoritma pri prepoznavanju tujih krajšav v slovenskih in tujih besedilih. Namen algoritma je čim bolje prepoznati krajšave in krajšavne razvezave v elektronskih besedilih, izdelati abecedni seznam krajšav in razvezav ter pripraviti gradivo za izdelavo slovarja krajšav. Prepoznavanje krajšav poteka na leksikalni oz. besedni ravni, z opazovanjem lastnosti krajšav in krajšavnih razvezav ter sovpadanja.

1 Začetna stopnja algoritma

Samodejno prepoznavanje krajšav sega v leto 1999. Pionirja na tem področju sta Taghva in Gilbreth, v naslednjih letih pa se je s prepoznavanjem začelo uvajati vse več avtorjev. V preglednici 1 so povzete temeljne značilnosti njihovih programov.
Začetna stopnja algoritma za samodejno prepoznavanje krajšav in krajšavnih razvezav temelji na dveh korakih. Prvi korak je zapis pravil za samodejno prepoznavanje krajšav, krajšavnih razvezav in sovpadanje krajšav z razvezavami, drugi korak pa priprava programske opreme. Pravila za prepoznavanje krajšav so omejena na samodejno prepoznavanje kratic in akronimov, izvzete so vse okrajšave. Tudi programi nekaterih že omenjenih avtorjev večinoma prepoznavajo akronime, izjemi sta le Larkeyjev (2000) in zbirka ADAM (2006), ki vključujeta tudi nekaj drugih tipov krajšav. Pravila za samodejno prepoznavanje krajšav in krajšavnih razvezav so razdeljena na tri dele. Prvi del je prepoznavanje krajšav, drugi del prepoznavanje razvezav, tretji pa ugotavljanje sovpadanja razvezav s krajšavami.

1.1 Prepoznavanje krajšav (akronimov in kratic)

Za prepoznavanje krajšav (akronimov in kratic) v besedilih so upoštevani naslednji premisleki, ki omogočajo ločitev krajšav (akronimov in kratic) v besedilih od drugih besed in besednih zvez:

(1) Upoštevane so besede, ki imajo največ 6 črk in so zapisane z velikimi tiskanimi črkami v oklepaju, npr. *(NATO)*.

(2) Izvzete so krajšave, ki so zaradi pogoste rabe prešle v navadno pisano besedje in se zapisujejo kot navadne besede, npr. *Unicef*, *Nama*. Te splošno znane krajšave bo algoritem prepoznal na podlagi vključenega Slovarčka krajšav.

Preglednica 1: Sinhrono-diahroni pregled algoritmov za prepoznavanje krajšav in krajšavnih razvezav

Avtor (leto)	**Značilnosti krajšav**	**Značilnosti razvezav**
Taghva Gilbreth (1999)	• akronimi, zapisani z velikimi tiskanimi črkami, z najmanj 3 in največ 10 znaki	• iz sobesedila • začetne črke razvezave so ključnega pomena
Yeates (1999)	• akronimi, zapisani z velikimi tiskanimi črkami	• iz sobesedila • začetne tri črke razvezave so ključnega pomena
Larkey (2000)	• akronimi, zapisani z velikimi tiskanimi črkami • tudi nekaj izjem (nekaj preostalih krajšavnih tipov) • do največ 9 znakov	• vzorec *akronim (razvezava)* ali *razvezava (akronim)* • ustaljene fraze, npr. *also known as*
Byrd Park (2001)	• akronimi, zapisani z velikimi tiskanimi črkami • vsaj 1 velika črka • tudi številke • od 2 do 10 znakov	• vzorec *akronim (razvezava)* ali *razvezava (akronim)* • ustaljene fraze, npr. *also known as*, *short for*
Schwartz Hearst (2003)	• akronimi, zapisani v oklepaju ali zunaj njega • od 2 do 10 znakov	• vzorec *akronim (razvezava)* ali *razvezava (akronim)*
Zahariev (2004)	• akronimi, zapisani v oklepaju ali zunaj njega	• vzorec *akronim (razvezava)* ali *razvezava (akronim)*
Jun Xu Yalou Huang (2005)	• akronimi, zapisani z velikimi tiskanimi črkami • od 2 do 10 znakov	• vzorec *akronim (razvezava)* ali *razvezava (akronim)*
ADAM (2006)	• akronimi, zapisani z velikimi tiskanimi črkami • tudi izjeme (nekaj preostalih krajšavnih tipov) • zapisani v oklepaju ali zunaj njega	• vzorec *akronim (razvezava)* ali *razvezava (akronim)*
Šateva Nikolov (2008)	• akronimi, zapisani z velikimi tiskanimi črkami, z največ 5 znaki	• iz sobesedila • vzorec *akronim (razvezava)* ali *razvezava (akronim)*

(3) Upoštevane so besede, ki imajo največ 6 črk, od katerih je vsaj prva velika tiskana, ter niso zapisane v oklepaju, npr. *NATO* ali *Mig*.

(4) Iz nastale zbirke kandidatk je treba odstraniti vse, kar ni akronim ali kratica. V to skupino spadajo osebna lastna imena, naselbinska lastna imena, nenaselbinska lastna imena in okrajšave. Odstranitev je mogoča s pomočjo geslovnikov Slovarja slovenskega knjižnega jezika, Slovenskega pravopisa in Slovarčka krajšav.

(5) Odstraniti je treba vse okrajšave, npr. *itd.*, *ipd.*, *itn.*, *npr.*, *št.*, *stol.* Pri tem se upošteva, da so zapisane z malimi črkami in da za njimi stoji pika. Take okrajšave

nimajo razvezave v besedilu in bodo zato samodejno izločene iz nabora možnih krajšav. Vse okrajšave, ki se zapisujejo z malimi črkami in brez ločil, npr. *cca*, *kg*, *dcl*, bo algoritem prepoznal s pomočjo geslovnika Slovarčka krajšav. Ker take krajšave v besedilu običajno niso razvezane, bodo samodejno izločene iz nabora možnih krajšav.

(6) Upoštevati je treba, da se kandidati za krajšave (akronime in kratice) ne pojavljajo v sobesedilu, sestavljenem iz samih velikih črk in v oklepaju ali pa zunaj njega, npr. *NAMA JE VŠEČ KINO*, *(JUTRI JE PETEK).*

1.1.1 Upoštevani so tudi drugi tipi krajšav in razvezav, ki jih je mogoče razdeliti v tele kategorije.

(1) Prekrivne krajšave

Pri prekrivnih krajšavah je število črk v krajšavi enako številu besed v razvezavi, črke v krajšavi so enake prvim črkam v razvezavi, npr. ***FF*** = ***F****ilozofska* ***f****akulteta*. Pri tem je lahko razvezava v oklepaju, prva beseda v razvezavi pa se začne z veliko začetnico, npr. *FF* (*Filozofska fakulteta*). Take krajšave so: *FF*, *EF*, *MK*, *DZS*, *CZ*, *DZ*, *BDP*, *CD*.

(2) Krajšave brez veznikov in predlogov

Gre za krajšave, ki imajo v razvezavi veznik ali predlog, v krajšavi pa ga nimajo, npr. *FDV* (*Fakulteta za družbene vede*). Na podlagi geslovnika Slovarčka krajšav je mogoče ugotoviti, koliko je takih primerov. Algoritem deluje tako kot pri prekrivnih krajšavah, ob tem pa upošteva tudi predloge in veznike iz razvezave.

(3) Krajšave iz prvih črk

Gre za krajšave tipa *NAMA* ali *BETI*, ki se lahko zapišejo tudi kot *Nama* ali *Beti*. Odkriti jih je mogoče s pomočjo Slovarčka krajšav. Te krajšave so sestavljene iz prvih črk razvezave, kot je to pri ***NAMA*** = ***Na****rodni* ***ma****gazin.*

(4) Krajšave z vezniki in s predlogi

Gre za krajšave tipa *DZU* (*Družba za upravljanje*), ki imajo predlog v krajšavi in razvezavi. Pri teh se uporablja enak postopek kot pri prekrivnih krajšavah. Če krajšava ne bo prekrivna z razvezavo, je algoritem ne bo upošteval.

1.2 Prepoznavanje krajšavnih razvezav iz sobesedila in sovpadanje

Pri prepoznavanju krajšavnih razvezav iz sobesedila je treba upoštevati sobesedilo, ki stoji levo ali desno od krajšave. Pri tem je lahko krajšava v oklepaju ali pa je v oklepaju razvezava, krajšava pa v takem primeru stoji zunaj oklepaja, npr. *MIP* (*Mesna industrija Primorske*).

1.3 Programska oprema

Po zapisu pravil za prepoznavanje krajšav in krajšavnih razvezav je sledila izdelava programske opreme. Informatik Gregor Širca je pravila za samodejno prepoznavanje krajšav in krajšavnih razvezav prevedel v programski jezik C# za okolje MS.NET

Framework, V2.0. Pripravljen je kot samostojna komponenta (dll), ki se lahko uporablja v novejših spletnih tehnologijah, kot so npr. obrazci offline windows forms.

1.3.1 Opis in delovanje

Na podlagi pravil za prepoznavanje krajšav in krajšavnih razvezav je v izvedbi informatika Gregorja Širce algoritem zaživel tudi v spletni različici.[2] Program MKstrings ima dve okni: v prvo vnesemo poljubno besedilo s krajšavami, po kliku na »Click here to process data« (klikni za obdelavo podatkov) se v drugem oknu prikažejo najdene krajšave in krajšavne razvezave. Od julija 2011 je nekoliko vizualno prirejen in funkcijsko zmogljivejši program na voljo na spletni povezavi http://mkstrings.farhouse.si/.

Iz vnesenega besedila[3]

> Ljubljana - **SDS (Slovenska demokratska stranka)** je na svoji spletni strani objavila premoženjsko stanje svojih poslank in poslancev v **DZ (Državni zbor)**. Podpredsednik **Državnega zbora (DZ)** France Cukjati je po objavljenih podatkih lastnik polovice enostanovanjske hiše. Poslanec **SDS** Branko Grims ima devet let star audi **A6**, 50 delnic Krke in 46 delnic **NFD** Holdinga, na računu pa 35.000 evrov. **Zvezi društev upokojencev Slovenije (ZDUS)** je namreč uspelo zbrati več kot 13.500 podpisov zavarovancev za sklic izredne skupščine. Od atipične pljučnice **(sars)** prek ptičje do nove (prašičje) gripe. Panvita se je za najem in oživitev Mipove proizvodnje v Kromberku odločila skupaj z družbo **Mig (Mesna industrija Goriške)**, ki jo je ustanovila skupina 20 nekdaj zaposlenih v **Mipu**. **NATO (North Atlantic Treaty Organization)** je mednarodna vojaško-politična organizacija držav za sodelovanje na področju obrambe, ki je bila ustanovljena leta 1949. **Organizacija severnoatlantskega sporazuma ali tudi Severnoatlantska pogodbena zveza (angleško North Atlantic Treaty Organisation; kratica Nato ali NATO)** je mednarodna vojaško-politična organizacija držav za sodelovanje na področju obrambe, ki je bila ustanovljena leta 1949. Nato sva odšla.

algoritem prepozna tele krajšave in krajšavne razvezave:

SDS Slovenska demokratska stranka
DZ Državni zbor

[2] Čeprav je bil program MKstrings objavljen na spletu, ni bil javno dostopen. Preko spleta je bil na voljo le informatiku in meni. V času nastajanja programa je informatik živel v Kopru, jaz pa v Bruslju, in prav spletna objava je omogočila preprosto sporazumevanje in posodabljanje programa, obenem pa je na tak način algoritem že pripravljen za morebitno poznejšo uporabo prek spleta.

[3] V razvojni fazi algoritma so bila uporabljena naključna spletna besedila, najdena ob pomoči iskalnika Google, v nadaljevanju pa je bil uporabljen ustrezen korpus besedil.

DZ Državnega zbora
ZDUS Zvezi društev upokojencev Slovenije
Mig Mesna industrija Goriške
NATO North Atlantic Treaty Organization
Nato North Atlantic Treaty Organization

Algoritem ne upošteva krajšav brez razvezav tipa *sars*, *Mipu*, *NFD*, *A6*, kar je pričakovano in v skladu z zgoraj navedenimi pravili. Na podlagi filtriranega besedila je mogoče trditi, da algoritem krajšave in krajšavne razvezave iz besedila ustrezno in pravilno prepoznava. Algoritem tako ustreza začetnim zahtevam prepoznavanja krajšav in je pripravljen za nadaljnji razvoj in izboljšave.

2 Nadaljnji razvoj

Na začetni stopnji je algoritem ob upoštevanju omejitev, ki zadevajo tako prepoznavanje krajšav kot prepoznavanje razvezav, krajšave in krajšavne razvezave samo prepoznaval. Omejitve so vidne v naboru črk, ki sestavljajo krajšavo, in vzorcih pojavitve krajšav in razvezav v besedilu. V okviru slednjega je bil v nadaljevanju za krajšavne kandidatke uporabljen niz desetih črk v štirih vzorcih možne pojavitve:

(krajšava) razvezava
(razvezava) krajšava
krajšava (razvezava)
razvezava (krajšava)

Za izboljšano delovanje algoritma in ugotovitev pomanjkljivosti je bilo v nadaljevanju uporabljenih nekaj naključnih besedil s spletnega portala 24ur.com, da bi ugotovila, kako se algoritem obnaša in kje ga je še treba izboljšati. Besedila so bila izbrana naključno, rezultati so predstavljeni v nadaljevanju.

Po vnosu besedila

> Dela na predoru Markovec pri Kopru sledijo terminskemu planu oziroma so celo pred rokom, trdijo v **Družbi za avtoceste v RS (Dars)**. Po napovedih **Darsa** bodo v predor Markovec z izolske strani zakopali sredi aprila, s koprske strani pa mesec dni pozneje.

algoritem ni prepoznal niza ***Družbi za avtoceste v RS (Dars)***. Ni ga prepoznal, tudi če je bila krajšava *Dars* zapisana z velikimi tiskanimi črkami. Razlog za neprepoznavo je v predlogih *za* in *v*, pa tudi rabi krajšave v razvezavi. Če se v nizu odstrani predlog *v* in doda razvezava za *RS* (*Republika Slovenija*), sistem niz prepozna. Torej je treba v pravilih za opredelitev razvezav upoštevati tudi prisotnost več kot enega

predloga in pojava krajšave. Prav taki primeri so zanimiv izziv pri iskanju boljših rešitev in prav na njihovi podlagi je treba oblikovati pravila, ki bodo vključena v algoritem in bodo zagotavljala brezhibno delovanje. Ostalo je še nekaj nerazrešenih primerov, kot so tuje krajšave s slovenskimi razvezavami ali razvezave, zapisane na poseben način. O tem v nadaljevanju.

> Obseg proizvodnje motornih vozil v **EU** se je lani glede na leto 2008 zmanjšal za 17,3 odstotka, glede na predkrizno leto 2007 pa je upadel kar za 23 odstotkov, ugotavlja **Združenje evropskih avtomobilskih proizvajalcev (ACEA)**.

Krajšave s slovenskimi razvezavami tipa *Združenje evropskih avtomobilskih proizvajalcev (ACEA)* so v slovenskem prostoru in tudi v drugih jezikih zelo pogoste. Pri takih krajšavah se večinoma uporabljata tuja krajšava in domača razvezava, npr. *NATO*. Za algoritem to pomeni prepoznavanje krajšav kot takih in prevodov razvezav, ob upoštevanju dejstva, da v sobesedilu ni tujih razvezav, ampak le domače. Algoritem naj bi tako prevode prepoznaval neposredno, pri tem pa se ne more opirati na leksikalno raven prepoznavanja v smislu sovpadanja črk. Ker algoritem temelji na leksikalnem prepoznavanju, so taki primeri krajšavno-razvezavnih parov izvzeti in niso predmet obravnave. Pri takih primerih je treba ubrati povsem drugačno metodo, saj bi tak algoritem temeljil na prepoznavanju prevodov in ne razvezav.

3 Tuji jeziki

Tudi v slovenskem besedilu lahko najdemo tuje krajšavno-razvezavne pare, predvsem če gre za nove, neustaljene krajšave, pri katerih je v slovenščini v rabi tuja razvezava. Zahariev (2004) se je ukvarjal z univerzalnostjo pravil za prepoznavanje krajšav in krajšavnih razvezav v tujih jezikih. Opazil je namreč, da so lahko pri nekaterih tipih besedil pravila univerzalna, predvsem pri latinični pisavi, vendar je treba upoštevati črkovne posebnosti in raznolikosti posameznih jezikov, npr. romunskega, nemškega, francoskega idr. Opozoril je, da so prav jezikovne posebnosti in raznolikosti ključne pri optimalnem prepoznavanju krajšav in da morajo biti vključene v algoritem za prepoznavanje. Zahariev se ni omejil le na nekatere evropske jezike, ampak je šel dlje, krajšave je želel prepoznavati tudi v arabščini, ruščini, kitajščini, japonščini in drugih bolj eksotičnih jezikih. Pravi, da je raznolikost skupna vsem jezikom. Zato jo je treba upoštevati pri gradnji algoritma za samodejno prepoznavanje krajšav in krajšavnih razvezav. V nadaljevanju je za prikaz univerzalnosti algoritem preizkušen še na angleških in italijanskih besedilih.[4]

[4] Besedila so izbrana naključno s pomočjo iskalnika Google.com.

Vneseno je bilo tole angleško besedilo:

> **USB** (**Universal Serial Bus**) is a specification to establish communication between devices and a host controller (usually personal computers), developed and invented by Ajay Bhatt working in Intel. **USB** is intended to replace many varieties of serial and parallel ports. **USB** can connect computer peripherals such as mice, keyboards, digital cameras, printers, personal media players, flash drives, and external hard drives. For many of those devices, **USB** has become the standard connection method. **USB** was designed for personal computers[citation needed], but it has become commonplace on other devices such as smartphones, **PDAs** and video game consoles, and as a power cord between a device and an **AC** adapter plugged into a wall plug for charging. As of 2008, there are about 2 billion **USB** devices sold per year, and approximately 6 billion total sold to date. The design of **USB** is standardized by the **USB Implementers Forum (USB-IF),** an industry standards body incorporating leading companies from the computer and electronics industries. Notable members have included Agere (now merged with **LSI** Corporation), Apple Inc., Hewlett-Packard, Intel, Microsoft, Sony and **NEC**. **Human-interface devices (HIDs)**. Main article: **USB** human interface device class

Algoritem prepozna niz *USB (Universal Serial Bus)*, ne prepozna pa nizov *USB Implementers Forum (USB-IF)* in *Human-interface devices (HIDs)* zaradi zapisa krajšav, npr. *HIDs*, in posebnih znakov, npr. *USB-IF* pri krajšavah in razvezavah, ter rabe krajšav v razvezavi, npr. *USB Implementers Forum*. Algoritem prepozna niz *(IF) Implementers Forum*, ne prepozna pa krajšav *PDAs*, *AC*, *LSI*, *NEC*, saj niso imele razvezav.

Vneseno je bilo še tole angleško besedilo:

> **Center for Group Learning** (**CGL**)
> The **Center for Group Learning's** web site: what we do and why we do it, announcements of our events, other group learning opportunities, and links to many …
> Stirling based Castle Leisure Group has come out on top at the **Scottish Licensed Trade News (SLTN)** awards in Glasgow. The company's latest venture City in ...

Pri slednjem je mogoče opaziti, da algoritem prepozna oba niza.

Vneseno je bilo tole italijansko besedilo.

> L'**ANAS**, il cui nome era l'acronimo di **Azienda Nazionale Autonoma delle Strade**, è una società per azioni italiana, di proprietà statale, avente per unico socio il Ministero dell'Economia e delle Finanze. L'**ANAS**, sotto la vigilanza tecnica e operativa del Ministero delle Infrastrutture e dei Trasporti, gestisce la rete stradale ed autostradale italiana di interesse nazionale.

Confederazione Generale del Lavoro (CGdL) è l'organizzazione sindacale fondata a Milano, tra il 29 settembre e il 1º ottobre del 1906, per iniziativa …
Cerca un acronimo per **CGL ... CGL** - abbreviazione aeronautica di Circling guidance light(s) - Luce/Luci di guida alla
Il **Popolo della Libertà** (**PdL**) è un partito politico italiano di centro-destra, membro del Partito Popolare Europeo. Il partito nasce dall'unione dei due ...

Program ne prepozna niza *(ANAS) Azienda Nazionale Autonoma delle Strade*, verjetno zaradi rabe določnega člena (*L'*) in določnega člena, ki se veže s predlogom (*delle*), prepozna pa niza *Confederazione Generale del Lavoro (CGdL)* in *Popolo della Libertà (PdL)*.

Sledil je še preizkus z drugim italijanskim besedilom.

Il **Dipartimento di Scienze del Linguaggio, dell'Interpretazione e della Traduzione (DSLIT)** è stato costituito nel 1997, a tutt'oggi l'unico in Italia a includere espressamente l'interpretazione tra le aree di ricerca che intende sviluppare. Le attività di ricerca promosse dal **DSLIT** spaziano in numerose aree disciplinari collegate alla lingua e alla mediazione interlinguistica e interculturale, nelle diverse lingue oggetto di studio all'interno del **DSLIT**. Oltre all'italiano, queste lingue sono: arabo, francese, inglese, olandese, portoghese, russo, serbo e croato, sloveno, spagnolo, tedesco.
Il Dipartimento afferisce a due Centri interdipartimentali all'interno dell'Ateneo di Trieste: il **Centro Interdipartimentale per la Ricerca Didattica (CIRD)** e il **CISEM (Centro Interdipartimentale di Studi Europei e Mediterranei).**

Algoritem prepozna niz *CISEM (Centro Interdipartimentale di Studi Europei e Mediterranei)*, ne prepozna pa niza *Dipartimento di Scienze del Linguaggio*, *dell'Interpretazione e della Traduzione (DSLIT)* in *Centro Interdipartimentale per la Ricerca Didattica (CIRD)*. Prav slednja sta problematična zaradi vsebnosti člena, predloga ali konstrukcije, sestavljene iz določnega člena in predloga, npr. *della*, *dell'*.

Iz rezultatov, dobljenih na zgornjih vzorčnih tujejezičnih besedilih, je mogoče sklepati, da algoritem za prepoznavanje krajšav ni univerzalen. Vsak jezik ima specifične značilnosti, ki morajo biti algoritmu znane, če želimo, da ustrezno prepoznava krajšave in krajšavne razvezave tudi v tem jeziku.

4 Analiza obsežnejše besedilne zbirke

Po preverjanju delovanja programa in ustrezni nadgradnji je bil algoritem uporabljen na večji, bolj raznovrstni in s krajšavami bogati zbirki besedil. Krajšave, predvsem nove, se predvsem zaradi jezikovne gospodarnosti pogosto pojavljajo prav v medijih, zato je bil preizkus narejen na besedilih petih letnikov dnevnika Delo, od leta 2005 do vključno 2009. Ta zbirka obsega 60 milijonov besed in 4 milijone povedi. 334.000 povedi ima oba oklepaja, 439.000 povedi pa vsaj po eno zaporedje

velikih črk. Za lažjo obdelavo so bila besedila razdeljena v dve zbirki. V prvi so bile zbrane povedi z oklepajema, v drugi pa povedi z le zaporedjem velikih črk. Posodobljeni program[5] je zbirki filtriral približno 30 minut in izluščil 5820 kandidatov za krajšavno-razvezavne pare. Pri luščenju je upošteval večpomenskost krajšav in število pomenov označil z arabskimi številkami. Seveda vsi krajšavno-razvezavni pari niso bili ustrezni oz. pravi, saj se v veliki množici razvezav in krajšav skrivajo lažni primeri ter ponavljajoči se pari. Lažno krajšavo je težko opredeliti, saj je lahko že vsako sosledje črk krajšava oz. kratica.[6] Bistveno lažje je z razvezavami, saj lahko prav te pričajo o dejanskem obstoju niza krajšave in razvezave. Izbira pravih nizov je potekala ročno, vsak krajšavno-razvezavni par je bil preverjen s pomočjo iskalnika Google.

4.1 Izsledki

4.1.1 Lažni krajšavno-razvezavni pari

Lažne krajšavno-razvezavne pare je mogoče izslediti le z ročnim pregledovanjem. Takih parov je bilo 189 ali 3,24 odstotka. Natančnost algoritma je torej 96,75-odstotna. Med lažnimi pari so krajšave, ki ne sovpadajo niti z eno ustrezno razvezavo, z drugimi besedami, tak krajšavno-razvezavni par ne obstaja. Razvezave spoznamo za lažne že na prvi pogled, saj nekatere ne vsebujejo vseh razvezavnih delov ali pa predstavljajo dele navadnega besedila ter pomensko ne ustrezajo in se ne ujemajo s krajšavo. S seznama so bila izločena tudi vsa lastna imena, npr. *JB*, *Janez Bratovž*. Iz preglednice 2 je razvidnih nekaj lažnih parov, navedenih je nekaj najbolj in najmanj pogostih. Pogostost je razvidna iz števila pojavitev v drugem stolpcu.

Preglednica 2: Najbolj in najmanj pogosti lažni krajšavno-razvezavni pari

Krajšava	**Število pojavitev**	**Prva razvezava**
PO	746	predstavljenih podatkih
TV	48	TV Tednik
NE	28	nasprotnikov EU
DP	27	domačem pokalu
DPA	10	dovoljenj Po pisanju agencije
ČE	10	članice EU
ON	2	občinstvo navdušil
NNNSP	2	nič nas ne sme presenetiti
GO	1	Gorenje obe
KAJ	1	Kratko atraktivno jedrnato
SS	1	spremenjena signalizacija
PISK	1	prevod in spremna beseda Klemen
VEM	1	Vse na enem mestu

[5] Gl. http://mkstrings.farhouse.si/.

[6] Pri tem so izvzete okrajšave, saj niso predmet obravnave pričujočega dela.

V levem stolpcu so krajšave, sledi število razvezav oz. pojavitev v srednjem in prva razvezava v desnem stolpcu. Največ razvezav ima krajšava *PO*, kar 746. Sledijo pari z občutno nižjo prednostjo pojavitve. Krajšava TV ni imela ustrezne razvezave v 48 primerih, NE v 28, DP v 27.

4.1.2 Ustrezni krajšavno-razvezavni pari

Čeprav je samo pridobivanje krajšavno-razvezavnih parov iz besedil hitro in preprosto, je njihovo pregledovanje dolgotrajno. S seznama parov so bile izločene lažne krajšave oz. vse krajšave, ki niso ustrezale vsaj eni razvezavi, tudi če krajšave obstajajo. Npr. *HIV*, *AJPES* sta bili prisotni v besedilih, a ker nista bili prisotni razvezavi, ju je algoritem skladno s pravili za ustrezne krajšavno-razvezavne pare samodejno odstranil. Na podlagi pridobljenih parov je mogoče sklepati, da so bile razvezave v skladu s pravili o prepoznavanju, a so se med primeri znašle tudi lažne. Nekaj lažnih razvezav je ustrezalo pravilom o prepoznavanju, zato jih lahko za lažne prepozna le človek. Lep primer lažnih razvezav je razviden iz preglednice 3. Algoritem krajšavi *PZS* pripiše kar 4 razvezave, med katerimi je ena lažna (*pa tudi zaradi stroškov*), ki vseeno ustreza pravilom za prepoznavanje.

Preglednica 3: Primer krajšave *PZS*

PZS	
1	Planinski zvezi Slovenije
2	pa tudi zaradi stroškov
3	Plesne zveze Slovenije
4	Plavalni zvezi Slovenije

Podobno lahko opazimo pri krajšavi *KS* v preglednici 4. Pri slednji so ustrezni trije pomeni, in sicer 2, 3 in 4.

Preglednica 4: Primer krajšave *KS*

KS	
1	koncu seje
2	krajevnih skupnosti
3	Krajevne skupnosti
4	Konfederacije sindikatov

Algoritem je krajšave in razvezave prepoznaval tako, kot so si sledile v besedilu, pri čemer je zajel razvezave v različnih sklonih ter ponavljajoče se razvezave. Iz nabora dobljenih parov je bilo treba ročno izluščiti najbolj nevtralno razvezavo in odstraniti primere, ki so se ponavljali, kot je prikazano v preglednici 5. Od šestih dobljenih razvezav krajšave *MNZ* je po ročnem luščenju mogoče ohraniti le tri ustrezne.

Preglednica 5: Primer krajšave *MNZ*

MNZ	
1	ministrstva za notranje zadeve
2	medobčinskih nogometnih zvez
3	ministrstvom za notranje zadeve
4	Medobčinske nogometne zveze
5	Muzeja novejše zgodovine
6	Muzej novejše zgodovine

↓

MNZ	
1	ministrstvo za notranje zadeve
2	medobčinska nogometna zveza
3	Muzej novejše zgodovine

Po ročnem luščenju ustreznih parov in odstranitvi ponavljajočih se krajšavno--razvezavnih parov je končna zbirka štela 2665 krajšavno-razvezavnih parov. Pozornost so vzbudili nizi s pomensko podobnimi, a skladenjsko različnimi razvezavami. Pomenska ustreznost je bila preverjena s pomočjo iskalnika Google.

Iz preglednice 6 je razvidna pomenska ustreznost nekaterih krajšavno--razvezavnih parov. Ustrezne oz. ustaljene ali z drugimi besedami uradne razvezave so zapisane krepko. Opaziti je mogoče, da imajo nekatere krajšave tudi več ustreznih razvezav. Razvezave, ki niso zapisane krepko, niso uradne; take so razvezave krajšav *ELES*, *ARRS* in *NKKVŠ.* Glede na izsledke, dobljene z iskalnikom Google, krajšavi *ELES* ustreza razvezava *Elektro Slovenija*, krajšavi *ARRS*, *Javna agencija za raziskovalno dejavnost Republike Slovenije*, krajšavi *NKKVŠ* pa *Nacionalna komisija za kvaliteto visokega šolstva*.

Preglednica 6: Pomenska ustreznost krajšavno-razvezavnih parov

ELES	
1	evra Ljubljana Elektro Slovenije
2	Slovenije Elektro
3	Slovenije Elektro letih
4	evrih Ljubljana Elektro Slovenija
PZP	
1	**poslovno združenje za prehrano**
2	poslovno združenje prehrane
3	**Perutninarska zadruga Ptuj**
ARRS	
1	agencije za raziskovalno dejavnost Republike Slovenije
2	Agencija za raziskave in razvoj Slovenije
NKKVŠ	
1	nacionalna komisija za kakovost visokega šolstva

2	nacionalna komisija za kakovost v visokem šolstvu
DZK	
1	**Demokratske zveze Kosova**
2	**Demokratično zvezo Kosova**
ŠRC	
1	**Športno rekreativni center**
2	**Športno rekreacijski center**
ZERC	
1	**zaščitne ekološke ribolovne cone**
2	**zakon o ekološko ribolovni coni**
DIIP	
1	**dokument identifikacije investicijskega projekta**
2	dokument o identifikaciji investicijskega projekta
ZPIZ	
1	**zakonom o pokojninskem in invalidskem zavarovanju**
2	**zavoda za pokojninsko in invalidsko zavarovanje**
ZGS	
1	**za gozdove Slovenije**
2	**za gradbeništvo Slovenije**
DRSC	
1	**Direkciji Republike Slovenije za ceste**
2	**Družba Republike Slovenije za ceste**
RRA	
1	**Regionalne razvojne agencije**
2	**razvojne regionalne agencije**
3	Regijska razvojna agencija

4.1.3 Tuje krajšave

Od 2664 primerov najdenih razvezavno-krajšavnih parov iz prvega in drugega nabora besedil je 646 tujih; največ je angleških, sledijo italijanski, francoski, španski, nemški idr. Pri nekaterih tujih manjka prvi del razvezave, preostale pa ustrezajo pravilom za prepoznavanje, ki so bila v celoti zasnovana za prepoznavanje krajšav in krajšavnih razvezav v slovenskih besedilih. Glede na pridobljene tujejezične nize lahko sklepamo, da so pravila v nekaterih segmentih univerzalna za večino jezikov, a nikakor ne na vseh ravneh. Algoritem namreč prepozna niza *ESA*, *BNL*, delno prepozna tudi niza *DOC* in *DOCG* (pri slednjih manjka začetni del razvezave, *denominazione*, gl. preglednice 7–9). Algoritem mora biti prirejen za potrebe posameznega jezika (Zahariev 2004), saj ima vsak jezik svojo specifiko. V sedanjem stanju algoritem še ni zrel za prepoznavanje nekaterih bolj zapletenih nizov, kot sta npr. *Dipartimento di Scienze del Linguaggio*, *dell'Interpretazione e della Traduzione (DSLIT)* in *Centro Interdipartimentale per la Ricerca Didattica (CIRD)*, predvsem če nastopata v bolj zapletenem kontekstu in so krajšave obravnavane v vlogi lastnih imen, npr. *L'ANAS*, *L'Anas*. Podobno velja tudi za druge

jezike, saj tako zapis, ki je lahko tudi nelatiničen, kot tudi tipološke značilnosti vplivajo na rezultate prepoznavanja.

Preglednica 7: Krajšavno-razvezavni par *ESA*

ESA	
1	European Space Agency
2	European Sponsorship Association

Preglednica 8: Krajšavno-razvezavna para *DOC* in *DOCG*

DOC	
1	di origine controlata
DOCG	
1	di origine controlata e garantita

Preglednica 9: Krajšavno-razvezavni par *BNL*

BNL	
1	Bance Nazionale del Lavoro

5 Sklep

S problematiko samodejnega prepoznavanja krajšav in krajšavnih razvezav se je ukvarjalo že več avtorjev in ob upoštevanju njihovih izsledkov so bila sestavljena pravila za samodejno prepoznavanje krajšav za pričujočo raziskavo. Pomanjkljivosti pravil so se pokazale šele pri uporabi algoritma na besedilih, zato se je bilo treba k pravilom večkrat vračati, vnašati popravke in algoritem ponovno preverjati. Prvi, učni različici so sledili nadgradnja v smislu nabora znakov krajšave (do deset znakov) in položaja niza (4 tipi položaja krajšavno-razvezavnih parov) ter ponovno preizkušanje in opazovanje izsledkov. Zaradi preizkusa univerzalnosti je bil algoritem uporabljen še na angleških in italijanskih besedilih, kjer so se pokazale tipološke posebnosti posameznih jezikov in njihova pomembnost pri gradnji algoritma za prepoznavanje krajšav še v drugih jezikih. Pred uporabo algoritma na besedilih iz časopisa Delo je bila, predvsem zaradi obsežnega nabora besedil (60 milijonov besed), programska oprema ustrezno dopolnjena. Program je po vnosu besedil deloval po pričakovanjih, z dobrimi rezultati. Za opredelitev točnosti algoritma je bilo treba izločiti lažne primere krajšavno-razvezavnih parov, lastna imena in ponavljajoče se primere. Na koncu je ostalo 2664 krajšavno-razvezavnih parov. Na trenutni stopnji lahko algoritem filtrira in prepoznava tudi nekatere tuje krajšavno-razvezavne pare. Čeprav je korpus zajemal le slovenska besedila, je bilo prikazano samodejno prepoznavanje tudi na nekaterih tujih naključno izbranih besedilih s portala 24ur.com. Trenutno predstavlja največjo oviro oz. najbolj zamudno stopnjo prav ročno pregledovanje ustreznosti krajšavno-razvezavnih parov po opravljenem filtriranju

iz korpusa, ki se ga na tej stopnji tudi še ne da povsem avtomatizirati, saj nekaterih razvezav brez posvetovanja s strokovnjaki ali ustreznim drugim virom ni mogoče prepoznati za prave. Algoritem za samodejno prepoznavanje krajšav in krajšavnih razvezav predstavlja vez med elektronskim besedilom in delno samodejno izdelano bazo krajšavno-razvezavnih parov, ki lahko služi kot gradivo za izdelavo slovarja krajšav. Tak način priprave slovarja je nedvomno prihodnost elektronske leksikografije.

Viri in literatura

ADAM ⟨http://128.248.65.210/arrowsmith_uic/adam.html⟩.

Byrd – Park 2001 = Youngja Park – Roy J. Byrd, *Hybrid TextMining for Finding Abbreviations and Their Definitions*, IMB Thomas J. Watson Research Center, 2001, 167–170.

Chiari 2007 = Isabella Chiari, *Introduzione alla linguistica computazionale*, Roma – Bari: Laterza, 2007.

Google ⟨http://www.google.com/⟩.

Jun Xu – Yalou Huang 2005 = Jun Xu – Yalou Huang, *A Machine Learning Approach to Recognising Acronyms and Their Expansions*, 2005 ⟨http://research.microsoft.com/en-us/people/junxu/acronymextraction-icmlc2005.pdf⟩.

Larkey idr. 2000 = Leah S. Larkey idr., Acrophile: An Automated Acronym Extractor and Server, *Proceedings of the fifth ACM conference on Digital libraries*, 2000, 205–214.

Schwartz – Hearst 2003 = Ariel S. Schwartz – Marti A. Hearst, A simple algorithm for identifying abbreviation definitions in biomedical texts, *Proceedings of the Pacific Symposium on Biocomputing*, 2003, 451–462.

Šatev – Nikolov 2008 = Vesna Šatev – Nicolas Nikolov, Using the Web as a Corpus for Extracting Abbreviations in the Serbian Language, v: *Jezikovne tehnologije: zbornik 11. mednarodne multikonference Informacijska družba – IS 2008*, *zvezek C*, ur. Tomaž Erjavec – Jerneja Žganec Gros, Ljubljana: Institut Jožef Stefan, 2008, 75–79.

Taghva – Gilbreth 1999 = Kazem Taghva – Jeff Gilbreth, Recognizing acronyms and their definitions, *International Journal on Document Analysis and Recognition* 1 (1999), št. 4, 191–198.

Yeates 1999 = Stuart Yeates, Automatic extraction of acronyms from text, *Proceedings of the Third New Zealand Computer Science Research Students' Conference*, Hamilton: University of Waikato, 1999, 117–124.

Zahariev 2004 = Manuel Zahariev, *A (Acronyms): doktorska disertacija*, School of Computing Science, Simon Fraser University, 2004.

24ur.com ⟨http://24ur.com/⟩.

Developing an algorithm for automatic recognition of acronyms and expanding acronyms in electronic texts

Summary

This article presents the development of an algorithm for automatic recognition of acronyms and expanding acronyms in electronic Slovenian texts. Before the final configuration, the algorithm was subjected to numerous improvements and changes. Recognizing acronyms takes place at the lexical level by observing the qualities of acronyms, expanded acronyms, and their correspondence. The algorithm recognizes acronyms based on recognition principles, and it seeks their expanded forms in context while taking into account principles of correspondence. This article describes the initial stage of recognition and its further development. The initial stage of the algorithm for automatic recognition of acronyms and expanding acronyms is based on the transcription of principles for automatic recognition of acronyms, expanding acronyms, and the correspondence of acronyms with their expansions, as well as on software preparation. The principles for recognizing acronyms are limited to automatic recognition of acronyms and omit all abbreviations. The recognition of acronyms takes into account words that have a maximum of six characters and are written in capital letters in parentheses, as well as words that have a maximum of six characters, of which at least the first character is capitalized, and are not written in parentheses. It also takes into account types of acronyms and expansions; for example, overlapping acronyms, acronyms without conjunctions and prepositions, acronyms made from initial letters, and acronyms with conjunctions and prepositions. Context was taken into account in recognizing expansions of acronyms. In the next stage of development, sequences of 10 characters in four possible patterns of appearance were used as acronym candidates: *(acronym) expansion*, *(expansion) acronym*, *acronym (expansion)*, and *expansion (acronym)*. Filtering was performed on texts from five years of the newspaper *Delo*, from 2005 to 2009, inclusive. This collection contains 60 million words. In 30 minutes 5,820 potential acronym-expansion pairs were extracted. False acronym-expansion pairs could only be detected through manual checking. Such pairs represented 3.24 percent. The accuracy of the algorithm is therefore 96.75 percent. The algorithm recognized acronyms and expansions as they followed one another in the text, whereby it captured expansions in various cases and repeating expansions. From the selection of pairs obtained, it was necessary to manually extract the most neutral expansion and to exclude cases that repeated. After manual extraction, the collection comprised 2,665 acronym-expansion pairs.