SAMODEJNO LUŠČENJE DEFINICIJ IZ SPECIALIZIRANIH BESEDIL

Darja Fišer, Senja Pollak, Špela Vintar
Filozofska fakulteta, Ljubljana

UDK 801.8=163.6:81'322.2:004.738.5

V prispevku predstavljamo novo metodo luščenja definicij iz slovenskih specializiranih besedil, ki temelji na modelu za klasifikacijo definicij, naučenem z uporabo metod strojnega učenja iz primerov v slovenski Wikipediji. Prvi korak metode zajema luščenje kandidatov s pomočjo slovenskega semantičnega leksikona, avtomatskega razpoznavanja terminov ter leksikoskladenjskih vzorcev. V drugem koraku pa z uporabo naučenega klasifikacijskega modela izmed definicijskih kandidatov izberemo »prave« definicije. Iz korpusa s področja naravoslovja smo s to metodo izluščili več kot tisoč definicijskih kandidatov ter z uporabo naučenega modela dosegli do 70-odstotno klasifikacijsko točnost.

luščenje definicij, luščenje informacij, računalniška obdelava naravnega jezika, strojno učenje, informacijsko poizvedovanje

This paper presents a new method for definition extraction from Slovene domain-specific corpora, based on a model for definition classification learned using machine-learning methods on examples from Slovene Wikipedia. In the first step we extract definition candidates using a Slovene semantic lexicon, automatic terminology recognition and lexico-syntactic patterns. Next, we use the learned classification model to select »true« definitions from the set of definition candidates. The method was tested on a natural science domain corpus from which we extracted more than a thousand definition candidates and achieved up to 70% classification accuracy with the learned classification model.

definition extraction, information extraction, natural language processing, machine learning, information retrieval

1 Uvod

Za učinkovit prenos znanja v specializiranem diskurzu mora biti jasno, kaj določeni pojem na izbranem področju pomeni. Pomen pojma v terminološkem slovarju, leksikonu ali specializiranem besedilu opredeljuje definicija, ta pa naj bi imela po načelih terminološke vede jasno skladenjsko in semantično strukturo. Osnovno in že tradicionalno načelo pri pisanju definicij je, da se pojem opredeli z navedbo najbližjega nadrejenega pojma in značilnosti, ki pojem razlikujejo od nadrejenega in istorednih pojmov v sistemu (*per genus et differentiam*). Tako se z ustrezno oblikovano definicijo obenem vzpostavlja pojmovni sistem terminološke zbirke in vnosov ni treba posebej opremljati s polji za nadpomenke in sorodne pojme.

Pri gradnji specializiranih terminoloških baz nam lahko korpusni pristop bistveno olajša delo, s sodobnimi metodami samodejnega prepoznavanja terminoloških enot v besedilih pa se terminografski postopki lahko bistveno skrajšajo in poenostavijo (Vintar 2010). V specializiranih besedilih se poleg samih terminov skrivajo še drugi dragoceni delci znanja, med drugim tudi definicije.

Pričujoči prispevek opisuje niz eksperimentov, s katerimi smo poskušali iz strokovnih besedil izluščiti čim več definicij. Najprej v drugem razdelku predstavimo pristope drugih avtorjev, nato v tretjem razdelku podrobno opišemo potek naše raziskave,

v četrtem razdelku navedemo pridobljene rezultate posameznih faz eksperimenta, prispevek pa sklenemo z razpravo in idejami za nadaljnje delo.

2 Sorodni pristopi

Dosedanji pristopi k samodejnemu luščenju definicij s specializiranih korpusov ali s spleta se v grobem delijo na dve veji; prva temelji na pravilih, druga na strojnem učenju, v zadnjem času pa se pojavljajo tudi kombinacije obeh; v slednji sklop se uvršča tudi pričujoča raziskava. Na pravilih temelječi pristopi skušajo priti do definicij prek njihovih skladenjskih in leksikalnih značilnosti. Za slovenščino je denimo tipična struktura ([*samostalniška zveza v imenovalniku*] + je + [*samostalniška zveza v imenovalniku*], ki [*opis razlikovalnih lastnosti*]), pri čemer sta prva in druga samostalniška zveza v nad-/podpomenskem razmerju. Takšno metodo je uporabil že Hearst (1992), a se različice metode z vzorci še vedno pojavljajo tudi v novejših raziskavah (Muresan, Klavans 2002; Walter, Pinkal 2006; Storrer, Wellinghoff 2006; Del Gaudio, Branco 2007).

Drugi sklop raziskav uporablja metode strojnega učenja, pri čemer je odkrivanje definicij mogoče razumeti kot problem razvrščanja; algoritem se skuša iz pozitivnih in negativnih primerov definicij naučiti pravil za razlikovanje med pravimi in nepravimi definicijami. Z običajnimi klasifikacijskimi algoritmi, kot so naivni Bayes, odločitvena drevesa in metoda podpornih vektorjev (SVM), je različnim avtorjem uspelo razlikovati med dobro in slabo oblikovanimi definicijami (Del Gaudio, Branco 2009; Chang, Zheng 2007; Velardi idr. 2008; Fahmi, Bouma 2006; Westerhout 2010; Kobyliński, Przepiórkowski 2008), za popolnoma avtomatske pristope pa so uporabljeni tudi genetski algoritmi (Borg idr. 2010) ter *mreže besednih vrst* (Navigli, Velardi 2010).

3 Zasnova raziskave

Cilj eksperimenta je bil razviti metodo za samodejno luščenje definicij iz specializiranih besedil, pri čemer smo želeli preskusiti različne pristope, jih med seboj primerjati in ovrednotiti ter tudi smiselno združiti. Sprva smo želeli iz specializiranih besedil samodejno izluščiti čim več kandidatov za definicije, nato pa s pomočjo klasifikacijskega algoritma, ki smo ga naučili pravil na definicijah iz slovenske Wikipedije, množico kandidatov razdeliti na prave in neprave definicije.

3.1 Strojno učenje na primerih iz Wikipedije

Algoritmi za strojno učenje tipično potrebujejo večje količine podatkov, zato smo se kot vir definicij odločili uporabiti slovensko Wikipedijo. Pri tem smo predpostavili, da je pri vsakem enciklopedičnem članku prvi stavek definicija, vsak naslednji stavek, v katerem se na začetku pojavi definirani termin, pa ni definicija. Tako štejemo stavek pod i) za definicijo, stavek pod ii) pa ne.

Primer 1

i) Celica je strukturna in funkcionalna enota vseh živih organizmov.

ii) Celice so v povprečju velike 10–20 μm, s prostim očesom jih ne moremo videti.

Slovenska Wikipedija, ki smo jo uporabili, je ob začetku našega eksperimenta obsegala približno 162.500 člankov. Obdržali smo le tiste, ki so imeli pravilno strukturo in so vsebovali vsaj odstavek besedila. Vse članke smo oblikoskladenjsko označili in lematizirali z orodjem ToTaLe (Erjavec idr. 2005), nato pa razčlenili glede na strukturo. Iz vsakega članka smo samodejno izbrali prvi stavek kot definicijo, nato pa v članku poiskali še en stavek, ki se je prav tako pričenjal s pojmom iz naslova članka v imenovalniku, ter ga shranili kot primer nedefinicije (glej primer 1). Tako smo zgradili učno množico z 19.964 primeri, od katerih jih je bila polovica pozitivnih, druga polovica pa negativnih, se pravi nedefinicij. Z orodjarno za rudarjenje podatkov Weka (Witten, Frank 2005) smo na tej učni množici izurili klasifikacijski algoritem, kot vektorske atribute pa smo uporabili najpogostejše besednovrstne oznake ter leme.

3.2 Luščenje potencialnih definicij iz besedil

Za namene raziskave smo najprej zgradili korpus strokovnih in poljudno-znanstvenih besedil. Vanj smo vključili učbenike in strokovne knjige z različnih naravoslovnih področij, kot so astronomija, geografija, fizika, botanika, saj smo domnevali, da besedila z vsaj delno izobraževalno funkcijo vsebujejo tudi več definicij in razlagalnih kontekstov. Nato smo oblikovali tri izhodiščne hipoteze, na podlagi katerih smo samodejno prepoznavali definicijske kandidate. Predpostavili smo, da je stavek kandidat za definicijo, če je bil izpolnjen najmanj eden od naslednjih pogojev:

- stavek vsebuje dva pojma, ki sta v slovenskem wordnetu v nad-/podpomenskem razmerju,
- stavek vsebuje najmanj dva terminološka izraza, od katerih mora biti prvi v imenovalniku,
- stavek vsebuje tipični leksikoskladenjski vzorec [*samostalniška zveza v imenovalniku*] + je/so + [*samostalniška zveza v imenovalniku*].

Naše predpostavke so namenoma široke, saj smo želeli z njimi zajeti čim več zanimivih stavkov in doseči dobro pokrivanje področja.

3.2.1 Luščenje s pomočjo slovenskega wordneta

Za raziskavo smo uporabili nedavno zgrajeni semantični leksikon sloWNet (Fišer 2007; Fišer, Sagot 2008). S semantičnim označevalnikom smo iz korpusa izluščili vse tiste stavke, v katerih se pojavita najmanj dva pojma iz sloWNeta in je hkrati eden nadpomenka drugega. V primeru, da je bilo v stavku več gnezdenih izrazov, smo uporabili daljšega. Primer 2 kaže izluščeni stavek, pri katerem sta pojma *diabetes* in *bolezen* v razmerju podpomenka-nadpomenka; takšno strukturo izkazujejo klasične definicije.

Primer 2

<term id=ENG20-13313485-n>Diabetes </term> je <term id=ENG20-13268088-n> bolezen </term>, ki je posledica pomanjkanja inzulina, hormona, ki skrbi, da celice v telesu dobivajo glukozo (sladkor).

3.2.2 Luščenje s pomočjo samodejno prepoznanih terminov

Naša druga hipoteza je predpostavljala, da so stavki, pri katerih se pojavita dva strokovna izraza, od katerih je vsaj eden v imenovalniku, prav tako morebitne definicije. Za prepoznavanje terminološko relevantnih enot v besedilu smo uporabili luščilnik izrazja LUIZ (Vintar 2010), ki na podlagi oblikoskladenjskih vzorcev in izračuna terminološkosti predlaga eno- in večbesedne terminološke izraze. Oblikoskladenjski vzorci za slovenščino vsebujejo tipične nize besednih vrst s podatki o sklonu, denimo [*samostalnik + samostalnik v rodilniku*], s katerim izluščimo samostalniške fraze, kot so *ohlajanje reaktorja*, *sila trenja*, *kislost rastišča.* Terminološko relevantnost izluščenih fraz ugotavljamo s pomočjo primerjave pogostosti posameznih besed v specializiranem korpusu in v splošnem korpusu, za slovenščino kot slednjega uporabljamo FidoPLUS. Iz terminološke relevantnosti posameznih besed ter pogostosti celotne večbesedne enote izračunamo skupno terminološko utež, ki jo v spodnjem primeru vidimo v atributu *score*.

Seveda niso vsi stavki, ki vsebujejo najmanj dva termina, definicije. V primeru 3 se nam je posrečilo izluščiti definicijo kljub temu, da nadpomenke *vzporednik* sistem ni prepoznal kot termin.

Primer 3

<term score="80.45">Ekvator </term> je najdaljši vzporednik, ki deli Zemljo na severno in <term score="43.21">južno poloblo </term>.

3.2.3 Luščenje s pomočjo leksikoskladenjskih vzorcev

Tretji način za zbiranje definicijskih kandidatov uporablja tradicionalno metodo

leksikoskladenjskih vzorcev, ki jih definiramo vnaprej in so – za razliko od prejšnjih dveh metod – vezani na posamezni jezik. V našem primeru smo uporabili en sam vzorec v obliki [samostalniška zveza v imenovalniku] + je/so + [samostalniška zveza v imenovalniku], ki poleg definicij izlušči tudi precej drugih stavkov. Primera 4 in 5 ponazarjata izluščeno definicijo (a.) in nedefinicijo (b.), ki obe ustrezata zgornjemu vzorcu.

Primer 4

a. Datumska meja je navidezna časovna črta, ki vzhodno poloblo loči od zahodne.

Primer 5

b. Zelo pomembna črta je datumska meja, ki teče čez Tihi ocean in usklajuje časovne pasove.

4 Rezultati

Da bi lahko z algoritmom strojnega učenja uspešno razlikovali med definicijami in nedefinicijami, smo morali algoritem najprej izuriti na množici znanih definicij in nedefinicij iz slovenske Wikipedije. Pri tem smo kot atribute uporabili najpogostejše oblikoskladenjske oznake in leme, z nizom eksperimentov pa smo primerjali različne načine predstavitve atributov (absolutna pogostost proti binarni, polne oznake proti skrčenim). Ko smo zgradili model oziroma ko se je algoritem »naučil« razlikovati med definicijami in nedefinicijami iz Wikipedije, smo preskusili njegovo delovanje na testni množici primerov iz Wikipedije. Najboljše rezultate smo dobili z uporabo polnih oznak in binarno predstavitvijo pogostosti. Med različnimi klasifikacijskimi algoritmi, ki jih ponuja orodjarna Weka, smo izbrali modele odločitvenega drevesa J48 (nastavitev M = 10), s katerim so bili rezultati v povprečju najboljši. V najboljšem primeru tako dosežemo 82,7 % točnost klasifikacije, samo na definicijah pa dosežemo natančnost 0,83 in priklic 0,82 (F-mera 0,827).

Z opisanimi metodami za luščenje definicij smo iz korpusa pridobili več kot tisoč definicijskih kandidatov, ki smo jih pregledali tudi ročno. V Tabeli 1 je navedeno število kandidatov, ki smo jih izluščili s posamezno metodo, ter natančnost. Kot je razvidno, je bila od skupnega števila izluščenih definicijskih kandidatov približno tretjina pravih.

Šele ko smo ročno pregledovali kandidate, smo se zavedeli, kako kompleksna je pravzaprav naloga luščenja definicij. Definicije iz Wikipedije so navadno oblikovane v skladu z načelom *per genus et differentiam* in se skladajo z vzorcem [samostalniška zveza] + je/so [samostalniška zveza], definicije iz korpusa pa so izkazovale mnogo širšo paleto skladenjskih struktur. Še težje pa se je bilo pri številnih primerih odločiti, ali gre za definicijo ali nedefinicijo z vsebinskega vidika, saj se v besedilih nove pojme uvaja in opisuje na zelo raznolike načine. Včasih je pojem najlažje definirati tako, da povemo, kaj ni, v drugih primerih pa naenkrat definiramo ali razložimo več pojmov. Pri ocenjevanju naših kandidatov smo bili strogi, kar je pomembno upoštevati pri interpretaciji rezultatov; številne stavke smo tako označili kot nedefinicije, čeprav so vsebovali dragoceno znanje o pojmu in bi jih lahko uporabili pri opisovanju pomenskega polja pojma.

Tabela 1: Izluščeni kandidati in ročna evalvacija

	Št. kandidatov	Prave definicije	Natančnost
SloWNet	104	41	0,39
Termini	629	118	0,19
Vzorci	311	98	0,31
Skupaj/Povprečje	1044	257	0,29

Tabela 2 kaže, da se točnost klasifikacije z algoritmom, ki smo ga izurili na primerih iz Wikipedije, giblje med 62 % in 71 %, kar v grobem pomeni, da algoritem v večini primerov pravilno uvrsti stavek bodisi v razred definicij ali nedefinicij. Pokaže se tudi, da se rezultat nekoliko izboljša, če za atribute uporabimo skrčene oblikoskladenjske oznake, se pravi oznake, iz katerih smo prej odstranili nerelevantne slovnične kategorije (npr. spol).

Tabela 2: Rezultati klasifikacije definicij iz korpusa z modelom, naučenim z algoritmom J48 na Wikipediji (klasifikacijska točnost in F-mera)

	Vzorci	Termini	SloWNet
skrčene oznake + J48	69,45 % (0,697)	69,79 % (0,698)	61,76 % (0,6)
skrčene oznake_bin + J48	63,9 % (0,643)	71,06 % (0,708)	66,67 % (0,65)
polne oznake_bin + J48	62,7 % (0,635)	65,98 % (0,662)	63,72 % (0,617)

Pogostost posameznega atributa lahko predstavimo bodisi z absolutno frekvenco – prva vrstica tabele – ali pa binarno, kar pomeni, da atribut lahko zavzame le vrednost 1, če se je pojavil, pri čemer število pojavitev ni pomembno, ter 0, če se ni pojavil. Preskusili smo obe različici, vendar se sodeč po zgornji tabeli obnašata podobno.

Ker smo želeli z raziskavo med drugim tudi primerjati tri metode za luščenje definicij, je zanimiv podatek o natančnosti klasifikacije zgolj na definicijah. Najvišjo natančnost dosežemo pri definicijah, izluščenih s pomočjo SloWNeta (0,63), sledijo vzorci (0,514) in termini (0,46). Ti rezultati se skladajo z rezultati ročne evalvacije, kar z drugimi besedami pomeni, da so stavki, ki vsebujejo dva izraza iz Wordneta in sta ta med seboj v nad-/podpomenskem razmerju, bolj verjetno prave definicije kot stavki, ki vsebujejo le dva strokovna izraza. Po drugi strani metoda s termini izlušči največ kandidatov in pogosto zajame tudi tiste razlagalne stavke, ki ne ustrezajo klasični definicijski strukturi, pa kljub temu vsebujejo pomembno znanje o določenem pojmu.

5 Sklep

V prispevku smo predstavili izvirno metodo za luščenje definicij iz korpusov, ki združuje metode strojnega učenja z jezikovnotehnološkimi pristopi. Za učenje klasifikacijskega modela smo uporabili Wikipedijo kot prosto dostopen in obsežen vir, pri prepoznavanju definicij pa smo kot nadgradnjo klasične metode z vzorci uporabili še semantični leksikon SloWNet in luščilnik izrazja LUIZ.

Rezultati so pokazali, da je s kombinacijo vseh treh načinov iz korpusa mogoče izluščiti veliko število potencialnih definicij, ki so v približno tretjini primerov prave definicije. Algoritem za klasifikacijo, ki smo ga izurili, je na testnih množicah iz Wikipedije deloval s skoraj 83-odstotno točnostjo, medtem ko smo pri klasifikaciji primerov iz korpusa dosegli 71-odstotno točnost. Od treh metod je najbolj natančna metoda z wordnetom, najmanj pa metoda s termini.

Eksperiment je bil zanimiv in poučen tudi z jezikoslovnega in terminološkega vidika, saj smo ob opazovanju izluščenih primerov ugotovili, da so načini razlaganja novih predmetnosti v besedilih bistveno bolj variabilni in prosti kot v enciklopedičnem viru, kot je Wikipedija. V tem smislu bi bilo v prihodnjih tovrstnih raziskavah nujno že vnaprej bolje opredeliti ciljno skupino uporabnikov oziroma ciljno aplikacijo na eni strani, na drugi strani pa tudi jasneje postaviti kriterije za razlikovanje med pravimi in nepravimi definicijami. Glede na ciljno aplikacijo pa lahko metodo prilagodimo tudi za luščenje drugih tipov definicij (Kosem 2006) kot klasičnih intenzionalnih definicij.

Literatura

BORG, Claudia, ROSNER, Mike, PACE, Gordon, J., 2010: Automatic Grammar Rule Extraction and Ranking for Definitions. Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, Daniel Tapias (ur.): *Proceedings of the Seventh conference on International Language Resources and Evaluation.* LREC'10. Malta: European Language Resources Association. 2577–2584.

CHANG, Xiao, ZHENG, Qinghua, 2007: Offline definition extraction using machine learning for knowledge-oriented question answering. De-Shuang Huang, Laurent Heutte, Marco Loog (ur.): *Proceedings of the Third International Conference on Intelligent Computing.* ICIC'07. *Communications in Computer and Information Science* 2. Berlin Heidelberg: Springer. 1286–1294.

FAHMI, Ismail, BOUMA, Gosse, 2006: Learning to identify definitions using syntactic features. *Proceedings of Workshop on Learning Structured Information in Natural Language Applications*. EACL'06.

FIŠER, Darja, SAGOT, Benoît, 2007: Leveraging parallel corpora and existing Wordnets for automatic construction of the Slovene Wordnet. *Proceedings of the 3rd Language and Technology Conference*. LTC'07.162–166.

FIŠER, Darja, SAGOT, Benoît, 2008: Combining multiple resources to build reliable Wordnets. *Proceedings of the 11th International Conference on Text, Speech and Dialogue*. TSD'08. 61–68.

DEL GAUDIO, Rosa, BRANCO, Antonio, 2007: Automatic extraction of definitions in Portuguese: A rule-based approach. *Progress in Artificial* In*telligence. Lecture Notes in Computer Science*. Berlin Heidelberg: Springer. 659–670.

DEL GAUDIO, Rosa, BRANCO, Antonio, 2009: Language independent system for definition extraction: first results using learning algorithms. Gerardo Sierra, Maria Pozzi in Juan-Manuel Torres-Moreno (ur.): *Proceedings of the 1st International Workshop on Definition Extraction.* RANLP-09. 33–39.

HEARST, Marti A., 1992: Automatic acquisition of hyponyms from large text corpora. *Proceedings of the 14th International Conference on Computational Linguistics.* COLING'92. 539–545.

KOBYLINSKI, Lukasz, PRZEPIORKOWSKI, Adam, 2008: Definition extraction with balanced random forests. *Proceedings of the 6th International Conference on Natural Language Processing.* GoTAL 2008. Springer Verlag. 237–247.

KOSEM, Iztok, 2006: Definicijski jezik v slovarju slovenskega knjižnega jezika s stališča sodobnih leksikografskih načel. *Jezik in slovstvo* 51/5.

MURESAN, Smaranda, KLAVANS, Judith L., 2002: A method for automatically building and evaluating dictionary resources. *Proceedings of the 3rd International Conference on Language Resources and Evaluation*. LREC'02.

NAVIGLI, Roberto, VELARDI, Paola, 2010: Learning Word-Class Lattices for Definition and Hypernym Extraction . *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. ACL 2010*. 1318–1327.

STORRER, Angelika, WELLINGHOFF, Sandra. 2006: Automated detection and annotation of term definitions in German text corpora. *Procedings the 5th International Conference on Language Resources and Evaluation*. LREC'06.

VELARDI, Paola, NAVIGLI, Roberto, D'AMADIO, Pierluigi, 2008: Mining the Web to Create Specialized Glossaries. *IEEE Intelligent Systems* 23/5. IEEE Press. 18–25.

VINTAR, Špela, 2010: Bilingual term recognition revisited: the bag-of-equivalents term alignment approach and its evaluation. *Terminology* 16/2. 141–158.

WALTER, Stephan, PINKAL, Manfred, 2006: Automatic extraction of definitions from German court decisions. *Proceeding of the ACL'06 Workshop on Information Extraction beyond the Document*. 20–26.

WESTERHOUT, Eline, 2010: *Definition extraction for glossary creation.* Utrecht: LOT.

WITTEN, Ian, H., FRANK, Eibe, [2]2005: *Data Mining: Practical Machine Learning Tools and Techniques*. San Francisco: Elsevier.