ELEKTROTEHNIŠKI VESTNIK 84(3): 108-116, 2017
IZVIRNI ZNANSTVENI ČLANEK

Wikipedija kot vir znanja za iskanje imenskih entitet

Jernej Flisar, Miha Pavlinek

Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko, Inštitut za informatiko, Smetanova ulica 17, 2000 Maribor, Slovenija
E-pošta: jernej.flisar@um.si

Povzetek. V raziskavi je predstavljen postopek gradnje učnega korpusa iz Wikipedije, katerega namen je izboljšati splošno točnost modelov za razpoznavanje imenskih entitet v slovenščini. Zaradi tematskih razlik med besedili po različnih domenah so namreč rezultati označevanja besedil z modeli za razpoznavanje imenskih entitet, ki so zgrajeni nad korpusi v eni domeni, v drugih domenah slabši. To razliko smo želeli zmanjšati z uporabo Wikipedije, ki omogoča prost dostop do informacij z različnimi tematikami. Modeli za razpoznavanje imenskih entitet so po navadi naučeni z nadzorovanim pristopom, kjer je potreben velik učni korpus, zato se zdi Wikipedija primeren vir znanja, saj vsebuje nešteto splošnih informacij v delno strukturirani obliki. Rezultate modela, ki smo ga zgradili na vsebini iz Wikipedie, smo primerjali z drugimi modeli za slovenščino. Prav tako smo iz Wikipedije zgradili leksikon imenskih entitet, s katerim smo pomembno izboljšali točnost modela.

Ključne besede: procesiranje naravnega jezika, razpoznavanje imenskih entitet, Wikipedija, klasifikacija

Wikipedia as a knowledge repository for the named-entity recognition

The paper presents a process of building a training corpus from Wikipedia to improve the general accuracy of the named-entity recognition models for the Slovenian language. Due to the differences in the text topics of various domains, the named-entity recognition models built on one domain provide worse results on other domains. We reduce this difference by using Wikipedia which offers an open access to general information on various topics. Moreover, the named-entity recognition models are usually built using a supervised learning approach needing a large training corpus. Therefore Wikipedia, with a great amount of information in a semi-structured form, seems to be an appropriate source of knowledge. We compare results of a model built on a Wikipedia content with those of other models for the Slovenian language and set up a lexicon of named entities which provides a significant improvement in terms of the recognition accuracy.

Keywords: natural language processing, named entity recognition, Wikipedia, classification

1 UVOD

Wikipedija je spletna enciklopedija in ena najbolj obiskanih spletnih strani v današnjem času. Ker je prosto dostopna in jo lahko vsakdo ureja, vsebuje velike količine informacij. Podprta je v več kot 250 jezikih, pri čemer največ člankov vsebuje angleška verzija (več kot pet milijonov), medtem ko jih ima slovenska verzija več kot 150.000 [1].

Zaradi velike količine informacij, zbranih na enem mestu, je Wikipedija velikanski vir znanja. Le-to se lahko koristno uporabi pri različnih raziskavah, ki se ukvarjajo z analiziranjem strukture in vsebine dokumentov. Ta področja so na primer računalniška lingvistika, rudarjenje v besedilu (ang. Text mining), procesiranje naravnega jezika (ang. Natural Language Processing – NLP) in upravljanje znanja (ang. Knowledge management). Področja se med seboj prekrivajo in dopolnjujejo, veliko pa jih je povezanih z umetno inteligenco in strojnim učenjem [2].

NLP je raziskovalno področje, ki se ukvarja z obdelavo besedil, zapisanih v naravnem jeziku [3]. Temelji na uporabi statističnih metod in metod strojnega učenja za pridobivanje znanja iz besedil. Večinoma se za NLP uporabljajo algoritmi nadzorovanega učenja. Raziskovalci se v zadnjem času ukvarjajo tudi z algoritmi nenadzorovanega učenja, ki pa so kompleksnejši in dajejo slabše rezultate, vendar so lahko včasih primernejši zaradi velikanskih količin neoznačenih podatkov, ki so na voljo [4].

Naloge, s katerimi se ukvarja NLP, so povzemanje (ang. Automatic summarization), odgovarjanje na zastavljena vprašanja (ang. Question answering), oblikoslovno označevanje (ang. Part-of-speech tagging), priklic informacij (ang. Information retrival) ter ekstrakcija informacij (ang. Information extraction) [5]. Ena od nalog slednjega je med drugim tudi razpoznavanje imenskih entitet (ang. Named Entity Recongnition – NER).

Prepoznavanje entitet se je kot znanstvena disciplina uveljavila v začetku devetdesetih let na konferencah MUC (ang. Message Understanding Conference), kjer so prvič obravnavali ta raziskovalni problem, ki je

Prejet 7. marec, 2017
Odobren 20. maj, 2017

pozneje postal ključna naloga pri večini sistemov s področja naravnega procesiranja jezikov [6].

Z NER lahko namreč pridobimo dodatne informacije iz besedila, saj besede oz. besedne zveze, ki pomenijo imenske entitete, kot npr. ime osebe, lokacijo ali organizacijo, k vsebini besedila prispevajo bistveno več informacij, kot bi jih bilo mogoče razbrati le iz posameznih besed. Pogosto med imenske entitete spadajo tudi datumski in številski izrazi, čeprav ne pomenijo imen [6].

NER vsebuje proces identifikacije imenske entitete iz besedila ter klasifikacijo posamezne entitete v vnaprej določene razrede.

Obstaja več standardov za iskanje imenskih entitet. Najpogosteje se uporabljajo MUC in CoNLL (ang. The Conference on Computational Natural Language Lerning) anotacije. MUC razdeli entitete v tri razrede:

- osebno ime (ang. person - PER),
- zemljepisno ime (ang. location - LOC) in
- ime organizacije (ang. organization - ORG),

CoNLL pa obstoječim razredom doda še stvarna imena (ang. MISC) [7].

Za pridobivanje oz. iskanje entitet iz besedil se uporabljajo različne označevalne metode. Sodobnejše uporabljajo postopke strojnega učenja, kot so npr. skriti markovi modeli (ang. Hidden Markov Model – HMM) [8] ali pogojna naključna polja (ang. Conditional Random Fields – CRF) [9]. Te metode temeljijo na nadzorovanem učenju, kar pomeni, da so entitete že pred tem označene. Poleg oznak učenje temelji na značilkah (ang. features), kot so npr. oblikoskladenjske oznake, velika začetnica, kratice, numerične vrednosti ipd., ki se za vsako besedo generirajo v fazi predprocesiranja. Pri označevanju entitet se uporabi model, ki je naučen na podlagi teh lastnosti.

Za nadzorovano učenje je potreben korpus besedil, v katerem so imenske entitete natančno označene. Za angleški jezik obstaja več ročno označenih korpusov, kjer gre večinoma za časopisne članke in novice. V teh domenah NER-modeli dosegajo zelo dobro točnost, vendar pa nad podatki iz drugih domen njihova točnost precej upade [11].

Ročno ustvarjanje učnih korpusov za posamezno domeno je dolgotrajno opravilo. Zato so raziskovalci iskali rešitve za avtomatizirano označevanje učnih korpusov. Med drugim so kot vir uporabljali Wikipedijo, ki vsebuje splošno znanje in je uporabna v različnih domenah [12][13][14]. Rezultati nekaterih raziskav so namreč pokazali, da lahko uporaba Wikipedije kot učnega korpusa izboljša točnost v različnih domenah oz. izboljša splošnost modela [15].

V naši raziskavi smo se osredinili na avtomatsko grajenje slovenskega korpusa za iskanje imenskih entitet s pomočjo Wikipedije. V drugem poglavju smo opisali dela na tem področju. Wikipedija, njena struktura in uporabna vrednost so opisane v tretjem poglavju. V četrtem poglavju predstavimo svojo rešitev za gradnjo učnega korpusa, ki ga nato v petem poglavju testiramo in preverimo. Sledijo rezultati eksperimenta in sklep.

2 SORODNA DELA

NER-modeli za angleški jezik, na katerega se je v preteklosti na tem področju osredinjala večina raziskav, dosegajo že zelo dobro točnost, in sicer okrog 90-odstotno [10]. V nekaterih drugih jezikih, med katerimi je tudi slovenščina, pa so doseženi rezultati slabši. Nad slovenskimi besedili so Štajner in drugi [16] ter Ljubešić in drugi [17] v preteklih raziskavah dosegali 75-odstotno točnost, pri čemer so uporabljali različne učne korpuse iz različnih domen. Štajner je s soavtorji model učil na korpusu ssj500k in ga nato preveril tudi z Ljubešićevim korpusom (slWaC), pri čemer je dosegel 62-odstotno točnost.

Kljub zelo natančnim NER-modelom so raziskovalci v preteklosti poskušali z alternativnimi načini pridobivati imenske entitete. Eden takšnih pristopov temelji na uporabi Wikipedije, ki ima kot velikanska zbirka besedil veliko lastnosti, s katerimi je lahko primeren vir za velike imensko označene korpuse.

Leta 2006 sta med prvimi Wikipedijo na tem področju uporabila Bunsecu in Pasca, in sicer za reševanje problema dvoumnosti (ang. disambiguation) oz. razločevanja pomena enakih besed [18]. Za njima se je raziskav s tega področja loteval tudi Cucerzan [19]. Bøhn in Nørvag sta na podlagi vsebine iz Wikipedije iz imenskih entitet zgradila slovar sinonimov [20].

Prav tako so avtorji različnih raziskav za izboljšavo NER-modelov v Wikipediji uporabili razločevalne strani. Tak pristop je uporabil Wetland [21] in zgradil večjezični slovar imenskih entitet HeiNER, ki vsebuje veliko zbirko imenskih entitet iz Wikipedije v številnih različnih jezikih. Slabost zbirke je ta, da entitete niso označene z entitetnimi razredi. To slabost je odpravil Knopp [22], ki je v svojem delu zbirko HeiNER nadgradil tako, da je z uporabo klasifikacijskih metod entitete iz slovarja klasificiral v standardne razrede imenskih entitet. Opozoril je na koristnost svojega dela, ki se lahko uporablja kot leksikon imenskih entitet (ang. gazetters) za oblikovanje NER-modelov. Nekateri modeli namreč dobro izkoriščajo to eksplicitno predznanje [16].

Toral in Munoz [23] sta se osredinila na prvi stavek iz članka v Wikipediji, saj sta domnevala, da je tam najboljši opis definicije naslova članka. S tem postopkom sta klasificirala naslove člankov glede na standardne imenske entitete: osebno ime, zemljepisno ime, ime organizacije in druge. Prvi stavek sta najprej oblikoslovno označila (ang. Pos-tagging) ter za vse označene samostalnike pridobila njihove imenske označitve iz WordNeta. Raziskovalca sta uporabila za primer angleško Wikipedijo, vendar pa se lahko enak postopek uporabi v drugih jezikih, če imamo na voljo podatke iz Wikipedije in WordNeta v jeziku, ki ga obdelujemo.

Tudi raziskovalca Kazama in Torisawa [24] sta se osredinila na prvi stavek v Wikipedija članku. V nasprotju s Toralom in Munozo pa ste se osredinila na glagol biti (ang. to be) v prvem stavku in z njim definirala razred imenske entitete pripadajočega članka. Kot primer sta navedla naslednje besedilo: »Jimi Hendrix je bil ameriški kitarist.« S preprostim postopkom sta asociirala članek o Jimiju Hendrixu s kitaristom. Tako sta pridobila dodatne slovarje in jih uporabila kot atribute, ki sta jih implementirala v NER-modele ter izboljšala njihovo točnost.

Številni raziskovalci so se prav tako lotili klasifikacije člankov iz Wikipedije glede na standardne razrede imenskih entitet. Dakka in Cucerzan [25] sta za klasifikacijo uporabila klasične metode tekstovnega rudarjenja. Vektorski prostor sta zgradila iz kratkega povzetka članka, celotne vsebine članka in prvega odstavka. Validirala sta vrednosti posameznih pristopov in ugotovila, da najboljšo klasifikacijo dosegata z metodo SVM in uporabo celotnega besedila članka. Dosežena točnost je presegla 90 % za funkcijo F.

Za izboljšanje NER-sistemov z avtomatskim generiranjem imensko označenih korpusov za učenje NER-modelov so se Richman in Schone [14], Mika in drugi [13], Knopp [22] ter Notham ukvarjali v več svojih raziskovalnih delih [11] [26] [12].

V svoji raziskavi sta Richman in Schone poskušala pridobiti korpuse z označenimi imenskimi entitetami v različnih jezikih, ki sta jih povezala prek Wikipedije. V svojem delu sta članke iz angleške Wikipedije klasificirala s hevrističnimi metodami. Za vsak razred (osebno ime, zemljepisno ime, ime organizacije) sta definirala nekaj ključnih besed, ki sta jih nato poiskala v definiranih kategorijah, katerim je pripadal članek iz Wikipedije. Klasificirane angleške članke sta nato povezala z enakimi članki v drugem jeziku. Tako sta lahko avtomatično označila velike korpuse za učne modele v španščini, francoščini in ukrajinščini. Informacije o kategorijah sta v svojih raziskavah uporabila tudi Knop in Notham [27]. Sunchanek pa je kategorije iz Wikipedije povezal s podatki iz WordNeta [28]. Mika in sodelavci [13] so namesto odhodnih povezav iz člankov uporabili informacije iz infopolj.

Notham je validiral splošno točnost svojega NER-modela, učenega nad avtomatsko ustvarjenim učnim korpusom iz angleške Wikipedije, ter primerjal rezultate z modeli, naučenimi nad standardnimi angleškimi korpusi. Splošno točnost NER-modela je tako dokazano izboljšal, v poprečju za 7 % za funkcijo F.

Wikipedijo so raziskovalci uporabili tudi za izboljšanje točnosti klasifikacije. Wang idr. [29] so uporabili znanje iz Wikipedije ter izboljšali svoj klasifikator. Obogatili so model vektorskega prostora s semantičnimi relacijami ter kot značilke dodali sinonime, asociacije in generalizacijo.

3 OBDELAVA PODATKOV V WIKIPEDIJI

Wikipedija je večjezična spletna enciklopedija, ki je prosto dostopna za raziskovalne namene. Obstaja v več kot 250 jezikih, pri čemer se slovenščina uvršča med 50 najobsežnejših po številu člankov. Vse jezikovne verzije Wikipedije je mogoče prenesti s spleta[1] v treh različnih podatkovnih formatih:

- HTML-format, kot je prikazan v spletnih brskalnikih,
- XML-format z metapodatki in vsebino, označeno v Mediawiki sintaksi[2], in
- SQL-format, strukturirane zbirke podatkov.

Velikost statičnih HTML-datotek je za vsaj trikrat večja od datotek v formatu Mediawiki, kar pomeni več potrebnega časa pri obdelavi vsebin. Čeprav je zaradi velikega števila orodij procesiranje HTML-formata lažje, se priporoča uporaba XML-formata [26].

3.1 Zgradba

Članki v Wikipediji imajo pet pomembnih lastnosti, ki jih lahko izkoristimo pri pridobivanju znanja za graditev velikega korpusa imenskih entitet: preusmeritve (ang. redirects), razločevanje (ang. disambiguations), povezave do člankov (ang. article links), notranje povezave (ang. internal links), infopolja (ang. infobox) in kategorije (ang. categories).

Tipičen Wikipedija članek, označen v sintaksi MediaWiki, je prikazan na primeru kode 1.

```
'''Slovenija''' (uradno ime: '''Republika Slovenija''') je [[Evropa|evropska]] država z zemljepisno lego na skrajnem severu [[Sredozemlje|Sredozemlja]] in na skrajnem jugu [[Srednja Evropa|Srednje Evrope]].
```

Primer kode 1: Izpis vsebine članka iz Wikipedije

Notranje povezave
Dvojni »oglati« oklepaji [[ ]] ponazarjajo notranje povezave med članki. V zgornjem primeru so vključene tri notranje povezave: »Evropa«, »Sredozemlje« in »Srednja Evropa«.

Kategorije
Članek vsebuje tudi povezave do kategorij, katerim pripada:

- [[Kategorija:Liberalne demokracije]]
- [[Kategorija:Evropske države]]
- [[Kategorija:Ustanovitve leta 1991]]
- [[Kategorija:Slovenija|*]]

Vsak članek lahko pripada več kategorijam. Te niso vezane na hierarhično strukturo, temveč so sestavljene

[1] http://dumps.wikimedia.org/
[2] http://www.mediawiki.org/wiki/Markup_spec.

kot usmerjen ciklični graf. Zaradi takšne strukture je zelo težko določiti podkategorije oz. nadkategorije [15].

Preusmeritve

Preusmeritev je stran brez vsebine, ki vsebuje samo preusmeritev na drug članek. Primer kode 2 prikazuje preusmeritve za entiteto z imenom »Republika Slovenija«, ki je preusmerjena na stran »Slovenija«. Preusmeritvene strani predstavljajo: alternativna imena, množine, sinonime, alternativna črkovanje in okrajšave, za enake entitete oz. članke.

```
#REDIRECT [[Slovenija]].
```

Primer kode 2: Vsebina preusmeritve strani "Republika Slovenija"

Razločevanje

Razločevalne strani so posebne strani s seznami povezav na različne članke s podobnim ali enakim imenom, vendar drugim pomenom. Na primer Wikipedijina razločevalna stran za »Delo« vsebuje več povezav, med katerimi so spodaj navedene tri najpomembnejše:

- delo – dejavnost proizvajanja,
- delo (fizika) – fizikalna količina in
- Delo (časopis) – slovenski dnevni časopis.

Vsi pojmi imajo enako ime, vendar pa imajo povsem drug pomen. Med temi tremi članki tako le zadnji pomeni imensko entiteto. Kot lahko razberemo iz zgornjega primera, se razločitvene strani uporabljajo za ugotavljanje različnih pomenov enakih besed.

Infopolja

Opcijsko lahko članki vsebujejo tudi infopolja, v katerih so strukturirani podatki o osebah, krajih in stvareh. Zaradi strukture so lahko zelo koristna. Infopolja uporablja DBPedija za pridobivanje semantičnih podatkov [30].

3.2 DBPedija

DBPedija je ontologija, ki jo sestavljajo semantično označeni podatki iz Wikipedije, vsebuje pa tudi povezave na druge vire, ko so WordNet, OpenCyc, GeoNames[3] in druge [31]. Viri v DBPediji so posamezne strani iz Wikipedije. Nabor podatkov je sestavljen iz RDF-trojic, do katerih lahko dostopamo prek SPARQL končne točke (ang. endpoint) ali pa si izbrane podatke prenesemo za lokalno obdelavo. Na voljo je v 125 jezikih. Angleška verzija DBPedije opisuje več kot 4,5 milijona stvari, od katerih je 4,2 milijona primerkov klasificiranih po razredih znotraj ontologije. Točnost klasifikacije je po poročanju avtorjev DBPedije 96-odstotna [32], kar je primerljivo z drugimi klasifikacijskimi pristopi, ki so jih dosegli drugi raziskovalci. Rezultat točnosti tudi sovpada z natančnostjo pri ročnem označevanju imenskih entitet, ki ni nikoli 100-odstotna in je v poprečju okrog 95-odstotna [33].

[3] Geonames je geografska zbirk, ki vsebuje več kot 10 milijonov zemljepisnih imen.

4 GRADITEV UČNEGA KORPUSA

Za graditev učnega korpusa smo kot vir besedil uporabili Wikipedijo, ki smo smo jo povezali z ontologijo DBPedija. Med vsemi članki iz Wikipedije je bilo primernih le 72.087, saj je večina pomenila kategorije ali škrbine. V nadaljevanju smo to število dodatno zmanjšali, saj niso vsi naslovi člankov pripadali imenskim entitetam. Končen nabor je tako bil 49.673 člankov.

Pri graditvi korpusa smo upoštevali, da številne povezave v Wikipediji ustrezajo razredom imenskih entitet (osebno ime, lastno ime, stvarno ime), zato smo lahko z naslednjimi koraki označili imenske entitete v članku Wikipedije:

1. Klasifikacija vseh Wikipedija člankov v katerega od razredov imenskih entitet.
2. Prečiščenje Wikipedija člankov in razdelitev na stavke.
3. Označitev stavkov z imenskimi entitetami.
4. Izbira člankov za vključitev v učni korpus.

4.1.1 Klasifikacija

Najprej je bilo treba klasificirati vse članke iz Wikipedije v razrede imenskih entitet. V tem koraku je potrebna velika natančnost, saj se napake iz klasifikacije pozneje prenesejo na učni model. V nasprotju z našimi predhodniki, ki so se posluževali različnih hevrističnih metod in nadzorovanega učenja, smo uporabili klasifikacijo iz DBPedije. DBPedija namreč semantično označi članke v Wikipediji z vsaj enim od razredov ontologije, pri čemer je na prvi hierarhični ravni 28 razredov. Te razrede smo nato preslikali v razrede za imenske entitete. Tabela 1 prikazuje način preslikave.

Osebna imena so tako primerki iz razreda »Person« (oseba) in EthnicGroup (etnična skupina), zemljepisna imena pa iz razredov »Place« (lokacija) in »Planet« (nebesna telesa). Stvarna imena smo sestavili iz več razredov (Organisation, Work, Event, MeanOfTransportation). Instance drugih razredov smo pustili nedefinirane oz. jih nismo uvrstili med imenske entitete.

Tabela 1: Preslikava razredov DBPedija v imensko entitetne razrede

Imenske entitete	**Razredi DBPedija**
Stvarna imena	Organisation, Work, Event, MeanOfTransportation
Zemljepisna imena:	Place, Planet
Osebna imena	Person, EthnicGroup
Nedefinirana	Painting, Protein, Beverage, AnatomicalStructure, SupremeCourtOfTheUnitedStatesCase, Award, ChemicalCompound, GovernmentType, Infrastructure, Disease, SpatialThing, Website, PersonFunction, Activity, OlympicResult, Sales, Colour, MusicGenre, Drug, Species, Currency, Language

Skupaj smo dobili seznam 42.061 imenskih entitet. Dodatno smo iz Wikipedijinih preusmeritvenih strani pridobili 5.624 alternativnih imen in sinonimov. Skupno število imenskih entitet je tako bilo 47.685. Med temi je 20.342 imenskih entitet pomenilo lokacijo, 18.807 osebna imena, stvarnih pa je bilo najmanj, in sicer 8.536. Tabela 2 prikazuje primer izpisa iz seznama razreda osebnih imen.

Tabela 2: Primer seznama osebnih imen

Einstein Person
Albert Einstein Person
Phillip Hamman Person
France Prešeren Person
Jason Thompson (košarkaš) Person
Michael Phelps Person

Za graditev leksikona, smo vse tri sezname razredov združili in ustvarili leksikon imenskih entitet, ki ga lahko vključimo v fazi učenja NER-modela kot eksplicitno predznanje.

4.1.2 Prečiščevanje Wikipedije

Wikipedija v osnovni obliki ni najprimernejša za učni korpus. Pred uporabo jo je treba prečistiti in s tem odstraniti nekaj nepotrebnih podatkov. Tako smo od celotne vsebine obdržali le golo besedilo, notranje povezave do drugih Wikipedija člankov ter prikazne vsebine povezave. Notranje povezave smo pozneje uporabili za označevanje imenskih entitet. Za razčlenjevanje (ang. parsing) smo uporabili javansko knjižnico Wikixmlj. Tako pridobljeno vsebino smo nato označili z imenskimi entitetami. Povezave in prikazne vsebine povezav smo oblikovali v strukturo ključ-vrednost, kjer prikazna vsebina predstavlja ključ, povezava na Wikipedija članek pa vrednost (Primer kode 3). Na ta način smo za celotno vsebino članka shranili informacije o vseh možnih povezavah oz. potencialnih imenskih entitetah.

Republika Slovenija → [Slovenija] Slovenije → [Slovenija] evropska → [Evropa] Evropi → [Evropa]

Primer kode 3: Povezava med prikazno vsebino in notranjo povezavo v obliki ključ-vrednost

4.1.3 Označevanje imenskih entitet

Prečiščene vsebine smo označili z imenskimi entitetami. Pri tem smo označevali izključno notranje povezave, in sicer smo vsako označili z oznako razreda, v katerega se uvršča povezan članek, s tem pa smo se izognili tudi reševanju dvoumnosti. Proces označevanja je sestavljen iz več korakov:

- Segmentacija besedila na stavke.
- Pridobivanje povezav na druge Wikipedija članke.
- Pridobivanje prikaznih vsebin na povezavah.
- Pridobivanje imenskih razredov člankov, h katerim kažejo povezave.
- Označevanje povezav z imenskimi razredi.

Besedilo smo najprej razdelili na posamezne stavke ter uporabili le prvi odstavek, kjer se po navadi nahaja največ povezav. Nato smo za vsak posamezni članek izluščili vse notranje povezave.

Pri uvrščanju povezav v razrede imenskih entitet smo upoštevali tudi prikazne vsebine. Če je npr. besedilo vsebine vsebovalo malo začetnico, je nismo upoštevali kot imensko entiteto, saj se vsa imena v slovenščini pišejo z veliko začetnico. Dodatno smo preverili, ali smo članek, na katerega kaže povezava, klasificirali v enega od treh razredov imenskih entitet. Nato smo povezave označili s primernimi razredi imenskih entitet, kar je prikazano na primeru kode 4.

[Slovenija] [**Zemljepisno ime**] (uradno ime: [Republika Slovenija] **[Zemljepisno ime]**) je [evropska] **[Zemljepisno ime]** država z zemljepisno lego... Leta 1550 je [Primož Trubar] **[Osebno ime]** na Nemškem izdelal prvi slovenski knjigi [Katekizem] **[Stvarno ime]** in [Abecednik] **[Stvarno ime]**.

Primer kode 4: Oznake nad razredi

Pri analizi smo odkrili, da veliko potencialnih entitet ni označenih s povezavo, ali pa so določene povezave kazale na neobstoječe strani oz. članke. Do tega prihaja zaradi razmeroma majhnega obsega slovenske Wikipedije v primerjavi z drugimi jezikovnimi verzijami.

Za izboljšanje ustvarjenega korpusa smo v zadnjem koraku odstranili stavke, ki niso ustrezali naslednjima pogojema:

1. V stavku je vsaj ena imenska entiteta.
2. Vse besede z veliko začetnico morajo pripadati razredu imenskih entitet.

S tem smo dosegli višjo stopnjo zaupanja (ang. confident), saj smo minimalizirali morebitne napake v učnem korpusu. Tako je končni korpus vseboval 500.000 besed.

4.1.4 Avtomatsko zgrajen korpus

Avtomatsko smo označili besedila iz 39.807 člankov, ki so vsebovala več kot 100.000 stavkov z več kot 1.000.000 besedami. Zaradi prej omenjenih pomanjkljivosti je bilo treba iz te množice izbrati primerno označene besede oz. stavke. Več podatkov o zgrajenem korpusu prikazuje Tabela 3.

Tabela 3: Število elementov avtomatsko zgrajenega korpusa iz slovenske Wikipedije

Elementov	**n**
Št. člankov	29.724
Št. izbranih člankov	4.593
Št. stavkov oz. povedi	100.203
Št. besed	268.368
Št. ločil in simbolov	58.231
Št. stavkov z imenskimi entitetami	17.415
Št. imenskih entitet	51.393

Naš avtomatsko ustvarjeni učni korpus vsebuje besedila iz 29.724 različnih člankov. Članki so sestavljeni iz 17.415 povedi, katerih skupno število besed je 268.368. S tako velikim naborom člankov in besed želimo zagotoviti bolj raznoliko besedilo, s katerim bi bil učni korpus čim bolj splošen. Skupno število imenskih entitet je 51.393, razdeljene so na tri razrede:

- osebna imena (18.083),
- zemljepisna imena (28.328) in
- stvarna imena (4.982).

V nadaljevanju smo vse besedilo razdelili na pojavnice (besede in ločila) ter jim dodelili entitetne razrede. V ta namen je bil uporabljen IO-format, kjer so imenske entitete označene z I_*X*, vse druge pojavnice pa z O (Tabela 4).

Tabela 4: Vsebina izdelanega učnega korpusa Wikipedije v tipičnem IO-formatu

Leta	O
1550	O
je	O
Primož	I_PER
Trubar	I_PER
na	O
Nemškem	O
izdelal	O
prvi	O
slovenski	O
knjigi	O
Katekizem	I_MISC
in	O
Abecednik	I_MISC
.	O

4.1.5 Učni korpus ssj500k

Za slovenščino je na voljo ročno označeni učni korpus ssj500k, ki je nastal v okviru projekta Sporazumevanje v slovenskem jeziku (SSJ) in ki temelji na dveh predhodnih učnih korpusih jos100k ter jos1M [34]. Vsebuje ročno dodane in pregledane podatke na ravni segmentacije, tokenizacije, lematizacije, oblikoskladenjskega označevanja in skladenjskega razčlenjevanja (11.411 stavkov), petina korpusa pa vsebuje tudi ročno označene imenske entitete (osebno, zemljepisno, stvarno ime). Podatki za ta pod korpus so predstavljeni v tabeli 5.

Tabela 5: Število elementov v korpusu ssj500k, označenem z imenskimi entitetami.

Elementov	**n**
Besedil	248
Odstavkov	1.599
Stavkov oz. povedi	5.808
Besed	100.135
Ločil in simbolov	18.499
Skladenjsko označenih stavkov	5.808
Skladenjskih povezav	118.635
Stavkov z imenskimi entitetami	2.177
Imenske entitete	4.397

Lastna imena so v korpusu, razdeljena na tri razrede:

- osebna imena (1.922),
- zemljepisna imena (1.284) in
- stvarna imena (1.191).

Lastna imena vsebuje 2.177 oz. 37,48 % vseh stavkov, pri čemer je distribucija lastnih imen po teh stavkih razmeroma neenakomerna. Skoraj polovica stavkov vsebuje le eno lastno ime, 25 % stavkov vsebuje dve lastni imeni, 10 % stavkov pa tri.

4.1.6 Primerjava

Ročno ustvarjeni korpus ssj500k je veliko manjši od avtomatsko ustvarjenega korpusa iz slovenske Wikipedije (v nadaljevanju – WPAslo). Najbolj opazna razlika je pri številu entitet. V ssj500k namreč polovica stavkov ne vsebuje imenskih entitet in ker je korpus WPAslo večji, je razlika še bolj opazna.

5 EKSPERIMENT

Za izvedbo eksperimenta smo uporabili javansko knjižnico Stanford CoreNLP in orodje za iskanje imenskih entitet - Stanford NER. Njegova implementacija temelji na metodi pogojnih naključnih polj. S temi orodji smo najprej zgradili dva ločena NER-modela:

- model, ki je naučen nad avtomatsko ustvarjenim korpusom iz Wikipedije in (v nadaljevanju M_{WPAslo})
- model, ki je naučen nad učnim korpusom ssj500k (v nadaljevanju $M_{ssj500k}$).

Pri graditvi modelov smo uporabili osnovne nastavitve parametrov, saj nas njihov vpliv na točnost ni zanimal.

Kakovost rezultata smo merili nad enotnim testnim korpusom z več metrikami:

- Natančnost (ang. precision), ki nam pove, koliko od dobljenih entitet je pravilnih.
- Priklic (ang. recall), ki nam pove, koliko znanih entitet smo identificirali.
- F-funkcija, ki je geometrijsko povprečje natančnosti in priklica.

5.1.1 Rezultati

Poskuse na ssj500k smo izvedli z desetkratnim navzkrižnim preverjanjem, pri katerem smo naključnih 85 % podatkov uporabili za učenje, preostalih 15 % pa za testiranje. Da bi preverili splošno točnost modela M_{WPAslo}, smo za njegovo testiranje uporabili povsem enako testno množico iz ssj500k.

Tabela 6 prikazuje rezultate za iskanje imenskih entitet z modelom $M_{ssj500k}$. Rezultati kažejo, da model pri razpoznavanju osebnih in zemljepisnih imen daje povprečno točnost, medtem ko je pri razpoznavanju stvarnih imen manj uspešen.

Tabela 6: Rezultati modela Mssj500k pri razpoznavanju imenskih entitet

Tip entitete	Natančnost	Priklic	F-funkcija
Osebno	0.5481	0.7925	0.6480
Zemljepisno	0.5139	0.7067	0.5951
Stvarno	0.5057	0.4049	0.4497
Skupaj	0.5306	0.6606	**0.5885**

Pri stvarnih imenih je $M_{ssj500k}$ manj uspešen, saj je mogočih precej več variacij, ki jih je težko zajeti v obstoječe učne podatke. Prav tako je točnost modela slabša od modelov v sorodnih raziskavah, kar utemeljujemo z dejstvom, da zaradi časovne zahtevnosti in težje primerjave s korpusom WPAslo (ni bil oblikoskladenjsko označen) nismo uporabili oblikoskladenjskih oznak, ki NER-model precej izboljšajo [16].

Da bi preizkusili splošnost pridobljenega modela in s tem tudi vpliv učnega korpusa na splošno točnost modela, smo izvedli merjenje s pomočjo korpusa ssj500k. Model M_{WPAslo} smo naučili s celotnim korpusom WPAslo, za testiranje pa uporabili enak del podatkov, kot smo jih uporabili že pri testiranju modela $M_{ssj500k}$. S tem smo zagotovili neposredno primerjavo med modeloma, naučenima nad korpusoma različnih domen. Rezultate modela M_{WPAslo} nad testnim delom korpusa ssj500k prikazuje Tabela 7.

Tabela 7: Rezultati modela MWPAslo nad testno množico ssj500k

Tip entitete	Natančnost	Priklic	F-funkcija
Osebno	0.6884	0.5241	0.5952
Zemljepisno	0.5492	0.7922	0.6486
Stvarno	0.3953	0.2191	0.2819
Skupaj	0.5802	0.5183	**0.5475**

Iz rezultatov je razvidno, da smo pri označitvi testnega dela korpusa ssj500 nekoliko izgubili pri natančnosti osebnih in stvarnih imen, medtem ko smo pri zemljepisnih imenih dosegli celo boljši rezultat. Iz tega lahko razberemo, da model dosega visoko splošnost pri označevanju zemljepisnih imen, nekoliko nižjo pa pri osebnih in stvarnih imenih. Skupen rezultat splošnega modela je bil sicer slabši za štiri odstotne točke za F-funkcijo, kar je bilo zaradi tematskih razlik pričakovano.

5.1.2 Leksikon imenskih entitet

Ker je lahko Wikipedija tudi odličen vir eksplicitnega predznanja, smo v nadaljevanju raziskave vse najdene imenske entitete združili v leksikon imenskih entitet in ga dodali učnemu modelu $M_{ssj500k}$. Dobljene rezultate prikazuje Tabela 8.

Tabela 8: Rezultati označevanja z uporabo leksikonov

Tip entitete	Natančnost	Priklic	F-funkcija
Osebno	0.8263	0.7770	0.8009
Zemljepisno	0.7506	0.7139	0.7318
Stvarno	0.5902	0.4046	0.4801
Skupaj	0.7563	0.6636	**0.7069**

Tabela 8 kaže, da uporaba leksikonov signifikantno izboljša rezultate za priklic in natančnost pri vseh imenskih entitetah. Tako je model NER s pomočjo leksikona iz Wikipedije pravilno označil 71 % imenskih entitet.

6 SKLEP

V pričujočem delu smo naredili pregled raziskav, ki so obravnavale Wikipedijo kot vir znanja na področju procesiranja naravnega jezika. Osredinili smo se predvsem na avtomatizirano ustvarjanje leksikonov in učnih korpusov za reševanje problema pri iskanju imenskih entitet. S pomočjo Wikipedije smo namreč želeli izboljšati splošno točnost NER-modelov v slovenskem jeziku.

Začeli smo s pregledom korakov in postopkov, uporabljenih za generiranje učnih korpusov in leksikonov iz Wikipedije. Po pregledu literature smo metode, ki smo jih zasledili, implementirali in uporabili v svojem eksperimentu.

Avtomatsko smo ustvarili učni korpus iz Wikipedije, s pomočjo katerega smo naučili osnovni NER-model ter ga testirali nad testnim delom korpusa ssj500k, ki je bil druga domena. Dobljeni rezultat je bil za 4,1 % slabši za F-funkcijo kot domenski NER model, je pa imel boljši rezultat pri zemljepisnih imenih.

Iz Wikipedije smo tudi ustvarili leksikon imenskih entitet. Tega smo dodali domenskemu NER-modelu in preverili izboljšanje modela. Ugotovili smo, da se je kakovost modela izboljšala z uporabo ustvarjenega leksikona, saj je model pravilno označil 71 % entitet.

V prihodnje bi bilo treba razširiti razrede imenskih entitet na več podrazredov. Stvarna imena bi lahko namreč še podrobneje razdelili npr. na imena organizacij, naslove književnih del, naslove glasbenih del ipd. Te podatke lahko pridobimo s pomočjo ontoloških podrazredov iz DBPedije. Tako bi z dodatnimi razredi lahko pridobili še več informacij o pomenu besedila, ki bi jih lahko nato primerno uporabili v kakšni drugi aplikaciji. Za večjo točnost splošnega modela bi bilo treba vključiti še druge vire znanja, kot so WordNet, BabelNet, GeoNames ipd.

LITERATURA

[1] Wikipedia, "Wikipedia," Wikipedia, the free encyclopedia. Februar 2014.

[2] A. Kao and S. R. Poteet, Natural language processing and text mining. London: Springer, 2007.

[3] D. Jurafsky and J. H. Martin, Speech and Language Processing, vol. 2000.

[4] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," Lingvisticae Investig., vol. 30, no. 1, pp. 3–26, 2007.

[5] C. D. Manning and H. Schütze, Foundations of statistical natural language processing. Cambridge, Mass.: MIT Press, 1999.

[6] N. A. Chinchor, "Overview of MUC-7," Proc. Seventh Message Underst. Conf. MUC-7, 1998.

[7] E. F. Tjong Kim Sang and F. De Meulder, "Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition," in Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4, 2003, pp. 142–147.

[8] L. R. Rabiner and B. H. Juang, "An introduction to hidden Markov models," IEEE ASSp Mag., 1986.

[9] J. Lafferty, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," 2001, pp. 282–289.

[10] J. R. Finkel and C. D. Manning, Joint Parsing and Named Entity Recognition.

[11] J. Nothman, T. Murphy, and J. R. Curran, "Analysing Wikipedia and gold-standard corpora for NER training," in Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, 2009, pp. 612–620.

[12] J. Nothman, J. R. Curran, and T. Murphy, "Transforming Wikipedia into named entity training data," in Proceedings of the Australian Language Technology Workshop, 2008, pp. 124–132.

[13] P. Mika, M. Ciaramita, H. Zaragoza, and J. Atserias, "Learning to Tag and Tagging to Learn: A Case Study on Wikipedia," IEEE Intell. Syst., vol. 23, no. 5, pp. 26–33, Sep. 2008.

[14] A. E. Richman and P. Schone, "Mining Wiki Resources for Multilingual Named Entity Recognition.," in ACL, 2008, pp. 1–9.

[15] J. Nothman, N. Ringland, W. Radford, T. Murphy, and J. R. Curran, "Learning multilingual named entity recognition from Wikipedia," Artif. Intell., vol. 194, pp. 151–175, Jan. 2013.

[16] T. Štajner, T. Erjavec, and S. Krek, "Razpoznavanje imenskih entitet v slovenskem besedilu," vol. 2013.

[17] N. Ljubešič, M. Stupar, T. Jurič, and Ž. Agić, "Combining Avalilable Datasets for Building Named Entity Recognition Models of Croatian and Slovene," Slov. 20, vol. 2013, no. 2.

[18] R. C. Bunescu and M. Pasca, "Using Encyclopedic Knowledge for Named entity Disambiguation.," in EACL, 2006, vol. 6, pp. 9–16.

[19] S. Cucerzan, "Large-Scale Named Entity Disambiguation Based on Wikipedia Data.," in EMNLP-CoNLL, 2007, vol. 7, pp. 708–716.

[20] C. Bøhn and K. Nørv\aag, "Extracting named entities and synonyms from Wikipedia," in Advanced Information Networking and Applications (AINA), 2010 24th IEEE International Conference on, 2010, pp. 1300–1307.

[21] W. Wentland, J. Knopp, C. Silberer, and M. Hartung, "Building a Multilingual Lexical Resource for Named Entity Disambiguation, Translation and Transliteration," in Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Marocco, 2008.

[22] J. Knopp, A. Frank, and S. Riezler, "Classification of named entities in a large multilingual resource using the Wikipedia category system," Master's thesis, University of Heidelberg, 2010.

[23] A. Toral and R. Munoz, "A proposal to automatically build and maintain gazetteers for Named Entity Recognition by using Wikipedia," NEW TEXT Wikis Blogs Dyn. Text Sources, p. 56, 2006.

[24] J. Kazama and K. Torisawa, "Exploiting Wikipedia as external knowledge for named entity recognition," in Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 2007, pp. 698–707.

[25] W. Dakka and S. Cucerzan, "Augmenting Wikipedia with Named Entity Tags.," in IJCNLP, 2008, pp. 545–552.

[26] J. Notham, "Learning Named Entity Recongnition from Wikipedia," 2008.

[27] D. Balasuriya, N. Ringland, J. Nothman, T. Murphy, and J. R. Curran, "Named Entity Recognition in Wikipedia," in Proceedings of the 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources, Stroudsburg, PA, USA, 2009, pp. 10–18.

[28] F. M. Suchanek, G. Kasneci, and G. Weikum, "Yago: a core of semantic knowledge," 2007, p. 697.

[29] P. Wang, J. Hu, H.-J. Zeng, and Z. Chen, "Using Wikipedia knowledge to improve text classification," Knowl. Inf. Syst., vol. 19, no. 3, pp. 265–281, Jun. 2009.

[30] DBPedia, "DBPedia." [Online]. Available: http://wiki.dbpedia.org/About. [Accessed: 13-Mar-2014].

[31] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, "DBpedia: A Nucleus for a Web of Open Data," in The Semantic Web, K. Aberer, K.-S. Choi, N. Noy, D. Allemang, K.-I. Lee, L. Nixon, J. Golbeck, P. Mika, D. Maynard, R. Mizoguchi, G. Schreiber, and P. Cudré-Mauroux, Eds. Springer Berlin Heidelberg, 2007, pp. 722–735.

[32] H. Paulheim and C. Bizer, "Type Inference on Noisy RDF Data," in The Semantic Web – ISWC 2013, H. Alani, L. Kagal, A. Fokoue, P. Groth, C. Biemann, J. X. Parreira, L. Aroyo, N. Noy, C. Welty, and K. Janowicz, Eds. Springer Berlin Heidelberg, 2013, pp. 510–525.

[33] E. Marsh and D. Perzanowski, "MUC-7 Evaluation of IE Technology: Overview of Results," 1998.

[34] Krek, Simon; Dobrovoljc, Kaja; Erjavec, Tomaž; et al., 2015, Training corpus ssj500k 1.4, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1052..

Jernej Flisar je asistent za področje informatike na Fakulteti za elektrotehniko, računalništvo in informatiko Univerze v Mariboru. Njegovo raziskovalno in razvojno delo se nanaša na semantični splet, tekstovno rudarjenje, ter procesiranje naravnega jezika.

Miha Pavlinek je asistent za področje informatike na Fakulteti za elektrotehniko, računalništvo in informatiko Univerze v Mariboru. Njegovo raziskovalno in razvojno delo se nanaša na strojno učenje, tekstovno rudarjenje, storitveno usmerjene arhitekture ter spletne tehnologije. V zadnjem času vse več pozornosti namenja področju inteligentnih sistemov in povezanim rešitvam.