Kreiranje resursa za obeležavanje dijagnoza u medicinskim izveštajima na srpskom jeziku

Ulfeta Marovac, Aldina Avdić
Departman za tehničke nauke
Državni univerzitet u Novom Pazaru
Novi Pazar, Srbija
umarovac@np.ac.rs, apljaskovic@np.ac.rs

Dragan Janković
Elektronski fakultet
Univerzitet u Nišu
Niš, Srbija
dragan.jankovic@elfak.ni.ac.rs

Sead Marovac
Odeljenje opšte hirurgije
Opšta bolnica Novi Pazar
smarovac@yahoo.com

Sažetak—**U ovom radu je prikazano kreiranje medicinskih leksičkih resursa za automatsko označavanje termina iz dijagnoza u medicinskim izveštajima. Da bi se izvršilo automatsko označavanje slobodnog teksta potrebne su metode komjuterske obrade prirodnih jezika kao i odgovarajući leksički resursi. Kako ne postoje javno dostupni medicinski leksički resursi za srpski jezik, kao ni korpus sa medicinskim izveštajima, doprinos ovog rada je izgradnja ovakvih resursa za potrebe automatskog obeležavanja dijagnoza.**

Ključne riječi - medicinski izveštaji; dijagnoze; automatsko označavanje; kompjuterska obrada teksta; leksički resursi

I. UVOD

Sveopšta upotreba informacionih sistema dovela je do gomilanja velikog broja tekstualnih podataka u elektronskom obliku. Razvijanjem računarske oblasti koja se bavi obradom prirodnih jezika (NLP - *natural language procesing*) otvorile su se mogućnosti izdvajanja informacija ne samo iz struktuiranih podataka već i slobodnog teksta pisanog na različitim prirodnim jezicima. Upotrebom informacionih sistema u medicinskim ustanovama integrisan je rad različitih delova sistema medicinske zaštite. Veliki broj podataka u medicinskim informacionim sistemima je struktuiran i nosi informacije o ličnim podacima pacijenta, informacije o zaposlenim, lekovima, speciijalizacijama, ustonovama, dijagnozama itd. Pored struktuiranih podataka neizostavan deo svakog pregleda je tekstualni opis stanja pacijenta (anamneza). Kako ograničenja sistema često ne dozvoljavaju da se bitne činjenice izraze kao struktuirani podaci ovaj slobodni tekst nosi puno informacija koje mogu biti od značaja za različite analize kako medicinskog stanja pojedinačnog pacijenta tako i stanja jednog dela populacije. Izdvajanje informacija iz medicinskih izveštaja zahteva postojanje odgovarajućih posebnih leksičkih resursa kojima će se bitne činjenice izdvojiti od nebitnih. U ovom radu je prikazana je izgradnja resursa za označavanje dijagnoza zapisanih u medicinskim izveštajima kao slobodan tekst.

Rad je oraganizovan u pet sekcija. Druga sekcija predstavlja pregled istraživanja bliskih sa temom rada. U trećoj sekciji je prikazana izgradnja rečnika dijagnoza. Testiranje kreiranih rečnika nad skupom obeleženih anamneza dato je u četvrtoj sekciji. Poslednja sekcija sadrži zaključak i pravce daljeg istraživanja.

II. PREGLED POVEZANIH ISTRAŽIVANJA

Elektronski medicinski izveštaji (EHR - *electronic health records*) nose puno bitnih informacija kao što su stanje pacijenta na prijemu u bolnicu, tok njegovog oporavka, zatim zdrastveno stanje pri otpustu. Ove informacije se i dalje najlakše izražavaju prirodnim jezikom što čini izdvajanje ovih informacija težim [1]. Specifična medicinska terminologija definisana je različitim standardima i sistemima klasifikacije. Klasifikacijom i opisima bolesti, tretmana i lekova kontroliše se vokabular koji se koristi u medicinskim izveštajima i administraciji i smanjunju stepena nejsanoće i dvosmislenosti.

Međunarodna klasifikacija bolesti (ICD - *International Statistical Classification of Diseases and Related Health Problems*) koristi se još od 18. veka uz stalne revizije i dopune. Dostupna je na više jezika i u nadležnosti je Svetske zdrastvene organizacije [2]. Klasifikacija bolesti predstavlja sistem kategorija koje se dodeljuju određenim bolestima po definisanim kriterijumima. Međunarodna klasifikacija bolesti je standardno sredstvo koje se koristi u epidemiologiji, zdravstvenom menadžmentu i u kliničke svrhe, odnosno u analizama opšteg zdravstvenog stanja populacionih grupa i ukupnog stanovništva i za praćenje zdravstvenih problema. Deseta revizija ove klasifijacije ICD10 je prevedena na srpski jezik i koristi se u medicinskim institucijama [3].

Izgrađeni su mnogi sistemi koji metodama analize prirodnog jezika (NLP) obrađuju medicinski jezik koji se može naći kao slobodan tekst sa ciljem njegove dalje primene u sistemu zdrastvene zaštite. Spyns [4] je jos 1996. dao pregled sistema za primenu NLP tehnika u medicini. I tada su postojale ideje o stvaranju NLP sistema koji su višejezični što u medicini ima poseban značaj s obzirom na primarnu upotrebu latinskog jezika kao dela medicinskih izveštaja na bilo kom drugom prirodnom jeziku. Procesiranje medicinskih izveštaja se sastoji iz više

koraka kao što su pročišćavanje podataka, integracija, transformacija, redukcija i na kraju zaštita podataka. Glavna svrha je da se pretvore polustruktuirani i nestuktuirani medicinski izveštaji u kompjuterski razumljive informacije metodama NLP-a. Ključne metode u ovom procesu su prepoznavanje imenskih entiteta (Named Entity Recognition - NER) i izdvajanje relacija (Relation Extraction - RE) [5]. Metoda prepoznavanja imenskih entiteta odnosi se na postupak identifikcije određenog simbola ili vrste imena u dokumentima. U medicinskim izveštajima ova metoda se koristi za identifikaciju medicinskih subjekata koji imaju specifičan značaj za lečenje kao što su imena bolesti, simptomi i nazivi lekova. Ona se sastoji iz dva koraka: pronalaženje granice entiteta i određivanje klase entita. Specifičnost pisanja medicinskih izveštaja kao što su veliki broj grešaka, sraćenica, ličnih stilova i oznaka lekara otežavaju ovaj proces [6].

Metode izdvajanja informacija iz medicinskih izveštaja koje se zasnivaju na pravilima i rečnicima zahtevaju pomoć odgovarajućih eksperata iz oblasti medicine za formiranje pravila i rečnika. U [7] je prikazana zavisnost izdvajanja imenskih entiteta od upotrebe različitih korpusa. Savova i saradnici su prikazali cTAKES otvoreno rešenje procesiranja medicinskih izveštaja NLP metodama [8]. Za izdvajanje medicinskih izraza iz slobodnog teksta mogu se koristiti metode zasnovane samo na rečnicima, pravilima i mašinskom učenju [9,10].

Procesu izdvajanja informacija prethodi proces normalizacije i on je specifičan za odgovarajući jezik ali i tip dokumenata koji se obrađuje. Normalizacija medicinskih izveštaja na srpskom jeziku prikazana je u radu [11]. Prepoznavanjem imenskih entita na srpskom jeziku bavili su se autori u radu [12].

Ne postoje javno dostupni medicinski leksički resursi niti radovi koji se bave automatizovanim izdvajanjem informacija iz medicinskih resursa na srpskom jeziku. U radu [13] se opisuju izazovi izvlačenja informacija iz medicinskih izveštaja na nemačkom jeziku sa naglaskom na problem nepostojanja adekvatnih leksičkih resursa za ovaj specifičan domen. Ovde su predloženi koraci neophodni za prevazilaženje poteškoća pri obradi kliničkih tekstova. Što se slovenskih jezika tiče postoje slična istraživanja za Bugarski jezik [14]. U ovom radu su predstavljeni softverski moduli koji podržavaju automatsko vađenje dijagnoza.

III. REČNIK DIJAGNOZA

Da bi se označile reči koje učestvuju u opisu dijagnoze mora postojati rečnik sa terminima kojima se označavaju dijagnoze. U medicinskim izveštajima dijagnoze se pišu kako na maternjem jeziku tako se često koriste latinski nazivi istih kao i internacionalno prihvaćeni nazivi. Česta je upotreba i skraćenica kao je su zvanično ustanovljena ali i onih ličnih. Heterogeno označavanje dijagnoza u medicinskih izveštajima čini proces njihovog obeležavanja težim.

Da bi dobili resurs sa terminima vezanim za dijagnoze počeli smo od struktuiranog skupa podataka koji postoji i odnosi se na ICD10 klasifikaciju dijagnoza na srpskom jeziku [3]. Ova klasifikacija sadrži kodove za bolesti , naziv i opis bolesti (simptome i znake, društvene okolnosti i spoljne uzroke nastanka bolesti, i drugo). Početna klasifikacija sadrži oko 14000 dijagnoza. Prošrene verzije i nacionalna izdanja ove klasifikacije sadrže više dijagnoza. Za kodiranje svake dijagnoze koristi se alfanumerički niz dužine najviše 4 koji se sastoji od jednog slova i do tri broja. Od slova je iskorišćeno 25 slova engleskog alfabeta (nije korišćeno 'U' koje je ostavljeno za dodatne izmene). U slobodnom tekstu ove dijagnoze se mogu naći u različitim oblicima: različitim brojem cifara, sa tačkom ispred poslednje cifre, itd. Primer: 'A000','B05',' B05.8',''B05".

Dostupni podaci se sastoje od (Tabela I):

1. Koda dijagnoze
2. Opisa i naziva dijagnoze na srpskom jeziku
3. Opisa i naziva dijagnoze na latinskom jeziku

TABELA I. PRIMER PODATAKA O DIJAGNOZI

Kod	**Naziv na srpskom jeziku**	**Latinski naziv**
A00	Kolera	Cholera
A000	Kolera, uzročnik Vibrio cholerae 01, biotip cholera...	Cholera classica
A001	Kolera, uzročnik Vibrio cholerae 01, biotip El Tor	Cholera El Tor
A009	Kolera, neozačena	Cholera, non specificata

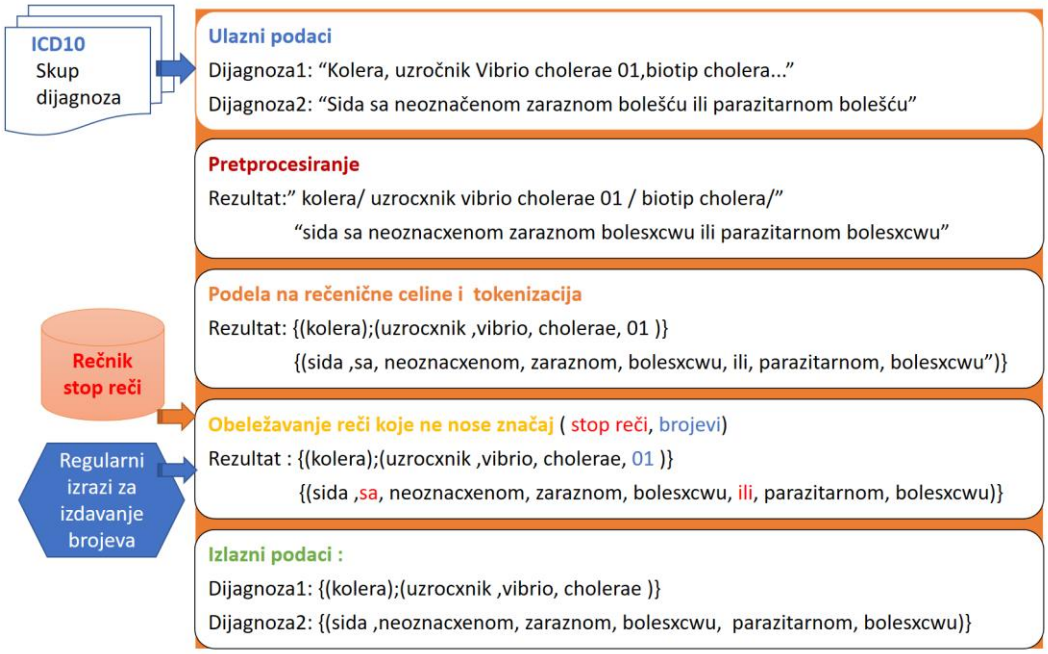
Da bismo uspešno označili termine u medicinskim izveštajima koji su deo dijagnoze moramo skup dijagnoza ICD10 klasifikacije konvertovati u skup tokena od kojih su dijagnoze sastavljene.

PROCES OBRADE REČNIKA DIJAGNOZA

Tekstualni podaci pre svake obrade moraju proći kroz proces predprocesiranja. Problem obrade pisanih podataka na srpskom jeziku leži u složenosti gramatike srpskog jezika pa je proces morfološke obrade reči težak i nepouzdan. Drugi problem je postojanje dva pisma koja su oba u zvaničnoj upotrebi pa čak vrlo često se mogu naći i oba u istom dokumentu. Slova sa dijakritičkim simbolima takođe otežavaju automatizovanu obradu zato će se srpska azbuka prevesti u engleski alfabet.
Pre bilo kakave obrade teksta slede koraci kojima se rešavaju prethodno pomenuti problemi i priprema se tekst za tokenizaciju:

a. Svođenje ćirilice na latinično pismo,
b. Svođenje slova sa dijakritičkim simbolom na kombinaciju slova (ć, č, ž, š, đ) => (cx, cw, zx, sx, dx). Slova x, w nisu slova srpske azbuke pa je ovo preslikavanje ispravno,
c. Obeležavanje seperatora rečenične celine ('.', ',', ';' , '-', '!', '?', '(', ')', '{', '}'),
d. Uklanjanje nepotrebnih prznina, svih ostalih simbola sem slova, cifara i separatora rečenične celine.

Sledeći korak je tokenizacija. Imajući u vidu da je krajni cilj pravljenja ovog resursa obeležavanje skupa reči koji opisuju



Slika 1. Proces izdvajanja značajnih reči iz skupa dijagnoza

dijagnozu, pri procesu tokenizacije se čuvaju informacije o rečeničnim celinama. Koraci tokenizacije su:

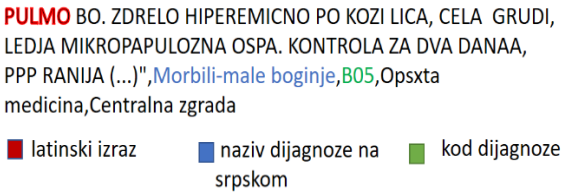
a. Podela izvornog teksta na rečenične celine,
b. Podela celine na reči pri čemu se čuva veza sa rečeničnom celinom.

Treći korak predstavlja označavanje reči koje nisu od značaja za obeležavanje dijagnoze i to:

a. pridruživanjem oznake rečima iz skupa stop reči ,
b. označavanje bročanih vrednosti.

Stop reči su one reči koje nemaju informativnu vrednost i javljaju se sa velikom učestalošću u većini dokumenata (na primer: veznici, uskilici, predlozi, itd.). Ovakav skup je izdvojen za potrebe normalizacije dokumenata na srpskom jeziku [15]

Nakon ovog procesa svaki opis dijagnoze je sveden na skup reči koje nose značajnu informaciju o njemu (isključene su reči obeležene u trećem koraku). Prikaz normalizacije prikazan je na Slici 1.



Slika 2. Primer anamneze sa različitim sadržajima

Kako se u medicinskim izveštajima ravnopravo nalaze izrazi na srpskom i latinskom jeziku (Slika 2.), isti postupak normalizacije je izvršen i za skup latinskih naziva ali bez označavanja stop reči.

ANALIZA REČNIKA DIJAGNOZA

Normalizacijom se dobija skup reči od kojih su neke više neke manje značajne za identifikaciju odgovarajućih medicinskih termina koji ukazuju na dijagnozu. Značajnost u pojavljivanja reči(r) u testu dijagnoze (d) u odnosu na njeno pojavljivanje u celom skupu dijagnoza (s) izračunati su pomoću formula za tf (1), idf (2) i tf_idf(3):

$$tf(r,d) = \frac{broj\_pojava\_reči\_r\_u\_dijagnozi\_d}{ukupan\_broj\_reči\_u\_dijagnozi\_d}$$

$$idf(r,s) = \frac{ukupan\_broj\_dijagnoza\_u\_skupu\_s}{broj\_dijagnoza\_koje\_sadrže\_reč\_r}$$

$$tf\_idf(r,d,s) = tf(r,d) * idf(r,s)$$

Vrednost tf ukazuje na odnos frekvenije pojavljivanja reči r u dijagnozi d u odnosu na ukupan broj reči u dijagnozi. Vrednost idf ukazuje na odnos ukupnog broj dijagnoza i broja dijagnoza u kojima se pojavljuje reč r. Što je tf-idf veći to se reč r značajnije pojavljuje u dijagnozi d u odnosu na druge dijagnoze. Tako se na primer naziv i opis na srpskom jeziku dijagnoze A000 preslikava set reči sa pridruženim tf, idf i tf-idf vrednostima prikazanim u Tabeli II.

Iz tabele je vidljivo da reči "biotip", "cholerae" imaju najveći idf što znači da se retko pojvljuju u celom korpusu. Reč kolera ima nešto niži idf na šta su uticale dijagnoze iz grupe A00 koje sve sadrže datu reč. Reč "uzročnik" se javlja u većem broju dijagnoza pa je ona nije značajna za identifikaciju date dijagnoze.

TABELA II. PRIMER VREDNOSTI TF, IDF I TF-IDF ZA DIJAGNOZU A000

Reč	Broj pojava u d	tf	Idf	tf-idf
Biotip	1	0.14	7097	1013.86
Cholerae	2	0.29	7097	2027.71
Kolera	1	0.14	3548.5	506.93
Uzrocxnik	1	0.14	84.49	12.07
Vibrio	1	0.14	4731.34	675.90

Početni skup se sastoji od 14194 dijagnoza. Procesom predporcesiranja naziva dijagnoza izdvojeno 72652 reči nakon čega je skup sveden na 7942 različite reči. Treba naglasiti da se u nazivu dijagnoza nalaze i simptomi, anatomski delovi, uzroci nastanka bolesti i mnogi medicinski i nemedicinski termini. Označavanjem tf-idf možemo primetiti da postoji veliki skup reči koje sadrže jako nizak idf što znači da se često pojavljuju u terminologiji. Čak oko jedna četvrtina izdvojenih termina 2071 termin ima idf koji se od prosečne vrednosti razlikuju više od jedne standardne devijacije. Međutim u ovom skupu se nalazi i puno medicinskih termina koji se često pojavljuju pa je teško da se izdvoje nemedicinski termini koji ne čine ključne reči u označavanju. Zbog specifičnosti klasifikacija dijagnoza sistemom ICD10 trebalo bi odgovarajaće faktore računati i na nivou grupe dijagnoza čime bi se mogla bolje izvršiti identifikacija.

Napravljen je i resurs sa latinskim nazivima dijagnoza za one za koje je postojao latinski naziv (3794 dijagnoze). Ovim resursom su obuhvaćeni i neki latinski nazivi anatomskih delova, simptoma i drugog. Ovaj skup posle obrade sadrži 2844 latinska termina.

Pored naziva dijagnoze izdvojen je resurs i sa ICD10 šiframa dijagnoza koje se takođe koriste prilikom pisanja medicinskih izveštaja. Izdvojen je skup od 14194 šifre.

IV. REZULTATI I DISKUSIJA

Predloženi rečnici termina iz dijagnoza su testirani na skupu medicinskih izveštaja i to na njihovom nestruktuiranom delu (anamnezama). Ovaj skup sadrži 2212 medicinskih izveštaja iz perioda (2012-2018) iz 32 medicinske stanice DZ Niš. Ovi medicinski izveštaji prikupljeni su od stane MEDIS.NET [16] informacionog sistema koji se korsti u više od 20 zdravstvenih ustanova u Republici Srbiji. Korpus je korišćen u skladu sa etičkim standardima, uz deidentifikaciju pacijenta i medicinskog osoblja.

Nad ovim ovim anamnezama je izvršen isti proces normalizacije korišćen za normalizaciju rečnika. Posle normalizacije ručno su od strane anotatora iz oblasti medicine označene reči koje čine deo opisa dijagnoze kao i drugi medicinski termini u anamnezi.

Pretragom termina iz kreiranog rečnika ICD10 kodova dijagnoza pronađeno je 26 različitih kodova sa ukupno 121 pojavljivanjem u resursu. Poređenjem sa ručno dodeljenim oznakama utvrđeno je da je ovakvim izdvajanjem dijagnoza načinjeno 20 grešaka. Na primer greška lekara pri skaćenom zapisivanju temperature t38 (temperatura 38) mapirana kao dijagnoza. Preciznost mapiranja kodova dijagnoza u obeleženom korpusu je 84%.

Mapiranjem reči iz rečnika termina iz opisa i naziva dijagnoza na srpskom jeziku. Pronađeno je 460 različitih termina iz rečnika koju se ukupno pojavljuju 4750 puta u obeleženom korpusu. Od toga 3434 puta pronađeni termin je u korpusu imao oznaku dijagnoze ili nekog drugog medicinskog termina koji stoji kao opis uz dijagnozu, dok 1316 pojava je u korpusu označeno kao nemedicinski termin. Preciznost ovog mapiranja termina iz dijagnoza na srpskom jeziku u korpusu je 72,3%. Veliki broj termina nije mogao biti obeležen na osnovu rečnika u korpusu iz razloga što se nalazio u nekom drugom morfološkom obliku u tekstu. Tako na primer za morbile imamo sledeće oblike u anamnezama („mobilli", „mor", „morb", „morbila", „morbilama", „morbile",...). U ovakvom obliku ovi termini ne mogu biti pronađeni iz rečnika u resursu bez svođenja na morfološki oblik.

Na isti način je testiran i skup reči iz latinskih naziva dijagnoza. Problem ovog resursa sto postoje srpski termini koji imaju isti oblik kao neki latinski termin a sasvim drugo značenje. Iz ovog rečnika je u korpusu pronađeno 30 različitih termina koji se ukupno pojavljuju 236 puta (205 puta kao medicinski termini i 31 put kao nemedicinski). Preciznost mapiranja latiskih reči u medicinske termine u anamnezama je 86,9%.

Ako uporedimo dobijene preciznosti sa rezultatima za bugarski jezik dobijenim u radu [14] gde su korišćene metode mašinskog učenja uz obradu skraćenica, oni se razlikuju do 10%. Treba naglasiti da je cilj rada predstavljanje i dobijanje odgovarjućeg leksičkog resursa, a za poboljšanje rezultata mapiranja odgovarajućih dijagnoza pored rečnika treba uključiti i obradu skraćenica, grešaka i naravno primeniti metode mašinskog učenja kojima bi se neralaevantan skup reči eliminisao.

V. ZAKLJUČAK

U ovom radu je prikazano kreiranje rečnika sa terminima koji se pojavljuju u nazivima i opisima dijagnoza u cilju označavanja delova teksta u anamnezama u kojima je opisana ili navedena neka dijagnoza. Iz početnog skupa dijagnoza navedenih u međunarodnoj klasifikaciji ICD10 izdvojena su tri rečnika termina. Prvi se odnosi na ICD10 kodove dijagnoza, drugi je rečnik termina koji se pojavljuju u nazivima dijagnoza na srpskom jeziku i treći je skup termina koji se pojavlju u latinskim nazivima dijagnoza. Kreirani rečnici su testirani na ručno obeleženom skupu anamneza. Dobijena tačnost za mapiranje termina iz kreiranih rečnika kodova, srpskih naziva i latinskih naziva dijagnoza u korpus je redom 84%, 72,3% i 86,9% . Dobijena preciznost se može poboljšati, morfološkom obradom podataka, dodavanjem posebnih resursa za skraćenice, kao i označavanjem reči u rečnicima koje ne nose značajne

informacije za identifikaciju dijagnoze. Ovo će biti i predmet našeg daljeg istraživanja.

ZAHVALNICA

Ovaj rad je delemično podržan od strane Ministarstva prosvete, nauke i tehnološkog razvoja Republike Srbije po projektima III44007 i ON 174026.

LITERATURA

[1] Safran C, Chute C, Scherrer JR. eds. *Natural Language and Medical Concept Representation*, (Preprints of the IMIA WG6 Conference), Vevey, 1994. Also published as McCray A, Safran C, Chute

[2] International Statistical Classification of Diseases and Related Health Problems; https://www.icd10data.com

[3] Međunarodna statistička klasifikacija bolesti i srodnih zdravstvenih problema Deseta revizija Knjiga 1 Tabelarna lista;Institut za javno zdravlje Srbije „Dr Milan Jovanović Batut", World Health Organization

[4] Peter Spyns, Natural Language Processing in Medicine: An Overview Article in Methods of Information in Medicine · January 1997 DOI: 10.1055/s-0038-1634681 · Source: PubMed

[5] Sun W, Cai Z, Li Y, Liu F, Fang S, Wang G. Data processing and text mining technologies on electronic medical records: a review; Journal of healthcare engineering 2018, vol. 2018, pp. 1-10.

[6] Dalianis H. Characteristics of Patient Records and Clinical Corpora; In: Clinical Text Mining, Springer, Cham, 2018.

[7] D. Rebholz-Schuhmann, A. Yepes, C. Li et al., "Assessment of NER solutions against the first and second CALBC silver standard corpus," Journal of Biomedical Semantics, vol. 2, article S11, Supplement 5, 2011.

[8] G. K. Savova, J. J. Masanz, P. V. Ogren et al., "Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications," Journal of the American Medical Informatics Association, vol. 17, no. 5, pp. 507–513, 2010.

[9] Jiang M[1], Chen Y, Liu M, Rosenbloom ST, Mani S, Denny JC, Xu H. A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. J Am Med Inform Assoc. 2011 Sep-Oct;18(5):601-6. doi: 10.1136/amiajnl-2011-000163. Epub 2011 Apr 20.

[10] Quimbaya, AP, Múnera, AS, Rivera, RAG, et al. Named entity recognition over electronic health records through a combined dictionary-based approach. Procedia Comput Sci 2016; 100: 55–61.

[11] Avdić A, Marovac U, Janković D, Avdić Dž**,** Normalization of Medical Records Written in Serbian , Proceedings of ICIST (2019), 9th International Conference on Information Society and Techology will be held on Kopaonik, Serbia on Mar 10-13, 2019

[12] Krstev, C., Obradović, I., Utvić, M., & Vitas, D. (2014). A system for named entity recognition based on local grammars. Journal of Logic and Computation, 24(2), 473-489.

[13] Starlinger, J., Kittner, M., Blankenstein, O., & Leser, U. (2017). How to improve information extraction from German medical records. *It-Information Technology*, *59*(4), 171-179.

[14] S Boytcheva, Automatic matching of ICD-10 codes to diagnoses in discharge letters, Proceedings of the Second Workshop on Biomedical Natural Language Processing September, 2011, Hissar, Bulgaria, Association for Computational Linguistics, pp 11–18

[15] U. Marovac, A. Pljaskovic, A. Crnisanin and E. Kajan, N-gram analysis of text documents in Serbian language, In Telecommunications Forum (TELFOR), pp. 1385-1388, 2012.

[16] Milenković, A. M., Rajković, P. J., Stanković, T. N., & Janković, D. S. (2011, November). Application of medical information system MEDIS. NET in professional learning. In 2011 19thTelecommunications Forum (TELFOR) Proceedings of Papers (pp. 1474-1477). IEEE.

ABSTRACT

This paper presents the creating process of a medical lexical resource for automatically labeling terms from diagnoses in electronic health reports. Natural languages processing and appropriate lexical resources are required to perform automatic labeling of free text. As there is no publicly available lexical resource for the Serbian language, as well as a corpus with electronic health reports, the contribution of this work is the construction of such resources for the purpose of automatic labeling of diagnoses in electronic health reports.

CREATING RESOURCES FOR MARKING DIAGNOSES IN ELECTRONIC HEALTH REPORTS IN SERBIAN

Ulfeta Marovac, Aldina Avdić, Dragan Janković, Sead Marovac