Pregled resursa za obradu kliničkih tekstova na različitim prirodnim jezicima

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

Sažetak— **Procesiranje prirodnog jezika je jedna od široko raspostranjenih oblasti veštačke inteligencije. U zavisnosti od domena tekstova kojima je procesiranje namenjeno, potrebni su različiti resursi za obradu odgovarajućeg prirodnog jezika. Klinički tekstovi nose mnoštvo informacija koje se mogu analizirati u cilju boljeg razumevanja zdravstvenog stanja pacijenta kao i u cilju sveobuhvatnog razumevanja bolesti. Cilj ovog rada je prikaz resursa za obradu kliničkih tekstova na različitim prirodnim jezicima.**

Ključne riječi-procesiranje prirodnog jezika, klinički tekstovi, resursi

I. UVOD

Zdrastvena zaštita je jedan od glavnih prioriteta svake države. Svedoci smo da bolesti nemaju granice i jezičke barijere i da brza i efikasna dostupnost podataka može da spasi živote. Većina zemalja danas koristi elektronsko zdravstvo i čuva gomile dragocenih podataka o pacijentima i njihovim bolestima koje najčešće služe za vođenje pojedinačnih slučajeva kao i za administraciju. Slobodan tekst koji se nalazi u izveštajima lekara u drugim sličnim kliničkim tekstovima nosi informacije koje su od značaja ne samo za razrešenje pojedinačnog slučaja bolesti već za izdvajanje informacija na globalnom nivou. Većina razvijenih zemalja ozbiljno se bavi obradom informacija u kliničkim tekstovima ali je najveći broj istraživanja za englesko govorno područje. Stvaranje jedinstvene celine i višejezičkog sistema za obradu kliničkih tekstova moglo bi da doprinese bržem razvoju medicine. Da je postojala jedinstvena platforma kojom bi se COVID-19 bolesnici pratili, mnogo brže bi se moglo doći do informacija šta su simptomi, koja su mesta potencijalana žarišta i sl. upotrebom metoda obrade prirodnog jezika kao što je i prikazano u radu [1]. Mogućnost obrade izveštaja pacijenata koji su lečeni u zemljama u kojima engleski nije zvanični jezik omogućava globalno udruživanje podataka što je od izuzetnog značaja posebno za retke bolesti [2]. Ovim radom dat je pregled dostupnih resursa za obradu kliničkih tekstova na različitim prirodnim jezicima.

Rad je oraganizovan u šest sekcija. Druga sekcija predstavlja opis osnovnih pojmova na kojima je baziran rad. U trećoj sekciji prikazana je obrada kliničkog teksta. Prikaz korpusa koji su korišćeni za različite jezike dat je u četvrtoj sekciji. Peta sekcija sadrži alate za medicinsku klasifikaciju i anotaciju kliničkih tekstova. Poslednja sekcija sadrži zaključak i pravce daljeg istraživanja.

II. OSNOVNI POJMOVI

U ovom poglavlju su opisani osnovni pojmovi obrade kliničkih tekstova koji su glavne tačke razmatranja u ovom radu: klinički tekst, elektronski medicinski izveštaji, procesiranje prirodnog teksta, procesiranje kliničkog teksta, resursi za obradu kliničkog teksta i medicinske klasifikacije.

Klinički tekstovi (eng. clinical texts) su tekstovi koje su napisali lekari, medicinsko osoblje i drugi pružaoci zdravstvenih usluga. Koriste se za dokumentovanje stanja pacijenta i pruženih zdravstvenih usluga. Oni opisuju pacijente, nihove patologije, njihovu ličnu, socijalnu i medicinsku istoriju. Klinički tekstovi se razlikuju od naučnih tekstova i nisu pripremljeni za objavljivanje. U njima se ne koriste potpune rečenice, koriste se medicinski prihvaćeni izrazi i skraćenice koji ne pripadaju prirodnom jeziku na kome su napisane.

Elektronski medicinski izveštaji (eng. electronic health records EHR) 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 [3]. 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 smanjuje se stepen nejsanoće i dvosmislenosti.

Procesiranje prirodnog jezika (eng. natural language processing- NLP) je oblast lingvistike, računarstava i veštačke inteligencije koja istražuje načine kako bi računari mogli razumeti i koristiti tekst ili govor na prirodnom jeziku i primeniti ih na neke korisne aktivnosti [4].

Procesiranje kliničikog teksta (eng. clinical text mining) predstavlja izdvajanje informacija iz kliničkih tekstova [5].

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

Resursi za obradu kliničkog teksta su svi skupovi podataka koji nam pomažu u istraživanju, a to su: skupovi dijagnoza, simptoma, lekova kao i klinički korpusi.

Medicinska klasifikacija i terminologija predstavlja sisteme klasifikacije i termine koji se koriste u izveštajima, administraciji, klasifikaciji i opisivanju bolesti, tretmana i medikamenata ka što su ICD kodiranje dijagnoza, SNOMED CT, MeSH, UMLS, ATC i drugi [5].

III. OBRADA KLINIČKOG TEKSTA

Klinički tekstovi predstavlju osnovnu formu komunikacije između zdravstvenih radnika. Koristeći metode obrade prirodnog jezika moguće je iz ovih tekstova izdvojiti informacije koje su zaključane u slobodnom tekstu i nisu lako upotrebljive za dalje kompjuterski automatizovane analize. Najveći broj autora bavi se analizom podataka iz kliničkih tekstova napisanih na engleskom jeziku zbog njihove javne dostupnosti kao i zbog alata kojima se obrađuju koji su za engleski jezik javno dostupni. Koriste se dva pristupa za obradu prirodnog jezika na kliničkim tekstovima i to: pristup zasnovan na pravilima; i algoritmi mašinskog učenja. Prvi pristup zahteva postojanje specijalizovanih kliničkih rečnika koji podržavaju složenu kliničku logiku kao što je na primer alat MTERMS [6]. Drugi pristup zasnovan na mašinskom učenju zahteva skup ručno anotiranih kliničkih podataka. Pregled upotrebe mašinskog učenja nad kliničkim tekstovima do 2020. dat je u [7] gde su prikazani rezultati od 110 istraživanja dostupnih na PubMed iz perioda od 2015. do 2018. godine, a koja se tiču mašinske obrade kliničkih tekstova na engleskom jeziku. Ispitujući osobine podataka koji su korišćeni došlo se zaključaka da je većina istraživanja koristila na stotine ili hiljade dokumenata. Mali broj je sa jako malim skupom podataka manjim od 50 i jako velikim od 10 000 dokumenata (deset radova). Veliki deo podataka iako je bio dostupan je ostao neiskorišćen. Glavni razlog neiskorišćenosti podataka je što nisu obeleženi. Ukoliko se anotacija podataka radi ručno onda je to zahtevan posao i sklon greškama. Algoritmi aktivnog učenja omogućavaju da se i sa manjim brojem ručno anotiranih podataka izvrši obrada dokumenata pri čemu se pri novim anotacijama koristi više algoritama i upoređuju se njihovi rezultati. Često se za anotiranje koriste postojeći struktuirani podaci pa tako se tako tekstualnom delu medicinskog izveštaja može pridružiti kod dijagnoze [8]. Može se koristiti i poluautomatizovano anotiranje. Često podaci koji se obrađuju dolaze iz jedne institucije što dodvodi u pitanje relevantnost tih podataka. Vrlo često rezultati objavljeni na jednom skupu podataka nisu davali iste rezultate na drugom skupu [9]. Klinička primena obrade ovakvih podataka je raznolika od dijagnostike, prognostike, zaštite, predikcije rizika, poboljšanja pružanja usluga, menadžmenta i dr.

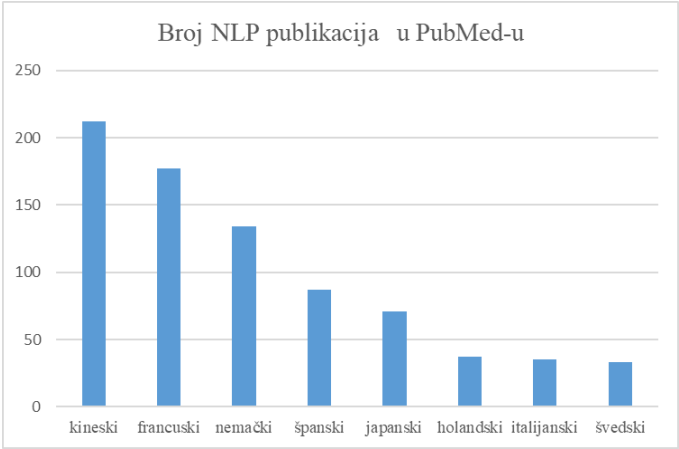
IV. DOSTUPNI KLINIČKI KORPUSI

Jako je teško doći do kliničkog korpusa zbog osetljivosti podataka koje koje sadrži. Svaki klinički korpus mora imati etičko odobrenje za upotrebu i proći proces deidentifikacije kako bi se sačuvala privatnost pacijenta vodeći računa o imenima i identifikacionim brojevima, telefonskim brojevima i adresama. Postoji broj dostupnih korpusa kako za engleski jezik tako i za ostale jezike.

Za engleski jezik postoje brojni skupovi podataka koji su anotirani i sastoje se od otpusnih listi, anamneza, izveštaja o nezi, radioloških izveštaja, rečenica iz medicinskog domena i drugih medicinskih izveštaja. Neki od dostupnih su:

- Informatics for Integrating Biology & the Bedside (i2b2) [10]
- Computational Medicine Center (CMC) corpus [11]
- ShARe/CLEF eHealth [12]
- Multiparameter Intelligent Monitoring in Intensive Care (MIMIC)[13]
- BioScope Corpus [14].

Kliničkih korpusa na stranim jezima (izuzimajući engleski) ima, ali su manji i pokrivaju manje različitih medicinskih sadržaja. Broj publikacija iz oblasti procesiranja prirodnog teksta po različitim jezicima u PubMed-u prikazan je na grafikonu (Slika 1). Vidi se da se radi o jezicima koji imaju široku zastupljenost ali i o nekim manje zastupljenim kod kojih je nivo svesti o zdrastvenoj zaštiti u državi koja ga koristi visok.



Slika 1. Broj publikacija u PubMed-u koje odgovaraju upitu “natural language processing and (French| German| Chinese| Spanish| Japanese| Dutch| Italian| Swedish)“

Korpusi koji nisu na engleskom u većini slučajeva su vezani za institucije kojima istraživači pripadaju i zahtevaju posebne dozvole i kontakte za njihovo dalje korišćenje. Ovi korpusi su uglavnom ručno anotirani i u njima su nazačene dijagnoze simptomi, lekovi, terapije. Neki od korpusa kliničkih tekstova na jezicima koji nisu engleski a koji su upotrebljavani u načnim radovima su:

- Španski IXAMed corpus koji sadrži 142,154 otpusnih listi upotrebljavan je u svrhe izdvajanja neželjenih reakcija na lekove [15].
- Bugarski klinički korpus [16] sadrži 100 miliona izveštaja lekara opšte prakse i specijalista. Drugi korpus sastoji se od više od 500,000 izveštaja bolesnika sa dijabetisom [17].
- Za srpski jezik u radu [18] su korišćena dva korpusa jedan sa izveštajima pacijenata koji su imali morbile 2212 izveštaja i jedan sa 10 različitih dijagnoza (2000 izveštaja).

- Švedski klinički korpus: Stockholm EPR Corpus koji je deo HEALTH BANK[19]. Sadrži oko 2 miliona izveštaja iz 500 kliničkih jedinica.
- Postoje dva korpusa na kojima se radilo na danskom jeziku i to jedan sa psihijatriskim izveštajima (61000) [20] i "EMC Dutch clinical corpus" koji je javno dostupan [21].
- Finski korpus koji se sadrži opis nege pacijenata od strane medicinske sestre. Sastoji se od 2800 rečenica od 8 različitih pacijenata [22].
- Jedan veći francuski korpus od 170,000 dokumenata od 2000 pacijenata prikazan je u [23].
- Za italijanski jezik rađeno je na korpusu od 23695 izveštaja pacijenata [24], kao i na korpusu od 100 izveštaja na kojima je isproban MetaMap [25].
- Za nemački jezik postoji više korpusa: 18,000 izveštaja iz različitih kliničkih jedica u Austriji [26], zatim 12743 kliničkih opisa labaratoriskih rezultata bolesnih od leukemije [27], 6817 kliničkih izveštaja i 118 otpusnih listi sa neurološke klinike [28].
- Za kineski jezik je bilo više manjih korpusa, a Hi i saradnici izgradili su korpus od 1100 medicinskih dokumenata [29].

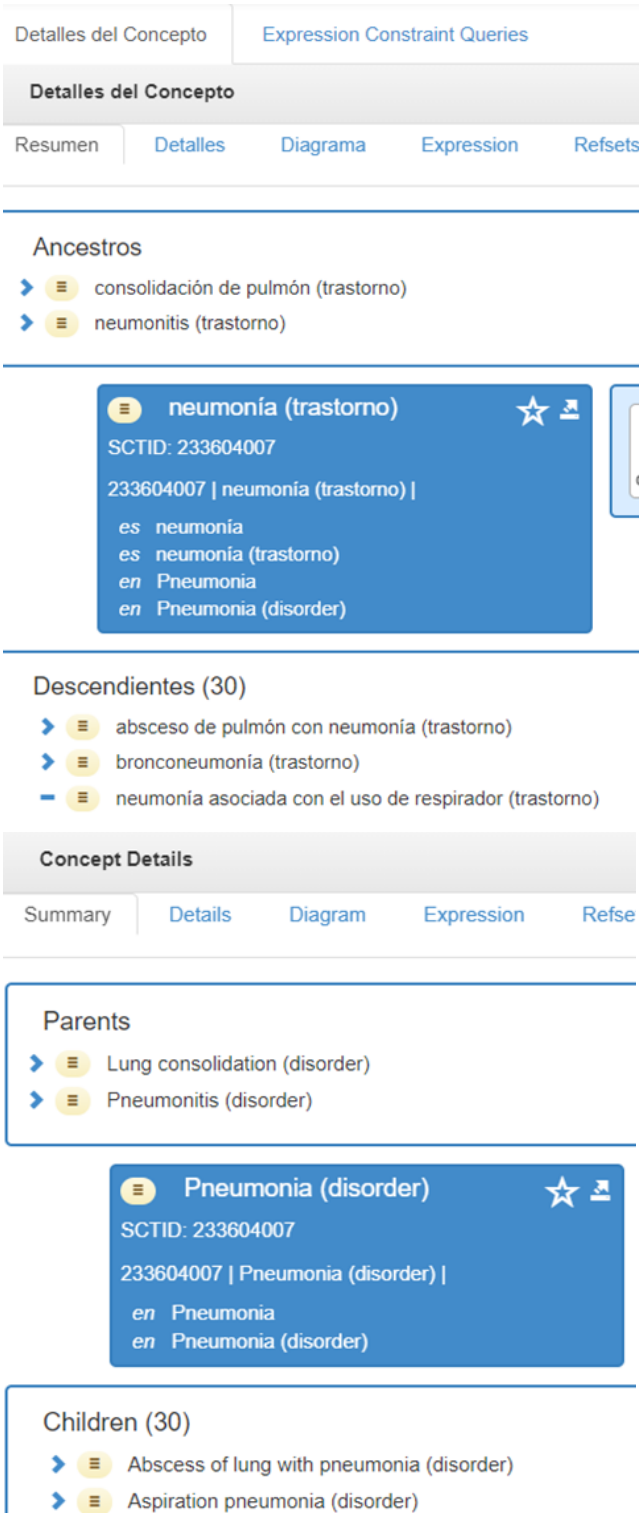
Nedostatak odgovarajućih leksičkih resursa se ponekad rešava primenjivanjem nenagledanih metoda [30].

V. ALATI ZA MEDICINSKU KLASIFIKACIJU I ANOTACIJU KLINIČKIH TEKSTOVA

Medicinska terminologija i sistemi za klasifikaciju se koriste u zdravstvu radi lakše interoperabilnosti među institucijama i kolaboracije medicinskih radnika, naučnika, i drugih zainteresovanih strana na globalnom nivou. Postoji opravdana potreba za integracijom različite medicinske terminologije i klasifikacije.

Međunarodna klasifikacija bolesti (ICD - International Statistical Classification of Diseases and Related Health Problems) koristi se još od 18. veka uz stalne revizije i dopune. Koristi se u preko 150 zemalja a dostupna je na više od 40 jezika i u nadležnosti je Svetske zdrastvene organizacije [31]. 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.

SNOMED CT[32] je strukturirani klinički rečnik koji se koristi u bilo kom elektronskom zdravstvenom kartonu (EHR). To je najopsežniji i najprecizniji klinički zdravstveni terminološki proizvod na svetu, čija je korist u tome što se podaci mogu deliti između zdravstvenih i socijalnih ustanova i pružaoca usluga. SNOMED CT je dostupan na američkom engleskom, engleskom, argentinskom španskom, danskom i švedskom. Prevodi na francuski, holandski, litvanski i nekoliko drugih. SNOMED CT je klinička hijerarhijska terminologija koja sadrži medicinske izraze i njihove odnose kao i sinonime, uključujući preko 320.000 termina (Slika 2).



Slika 2. SNOMED CT gore: Shanish edition 2020 dole: Unateted States edition 2020

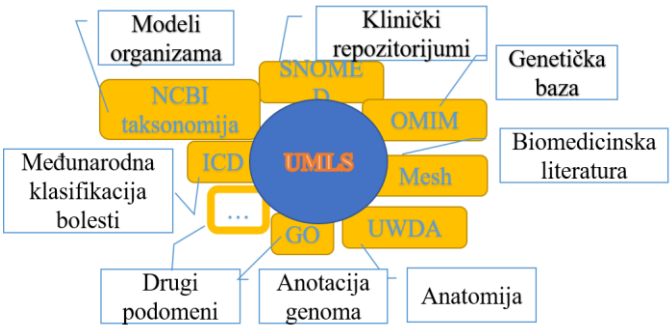
UMLS (Unified Medical Language Sistems) [33] integriše i distribuira medicinsku terminologiju, standarde klasifikacije i

kodiranja i povezane resurse kako bi se ubrzalo stvaranje efikasnijih i interoperabilnijih biomedicinskih informacionih sistema i usluga, uključujući elektronske zdravstvene kartone. UMLS daje podršku mapiranju između različitih terminologija. UMLS sadrži nekoliko miliona koncepata koji potiču iz stotina bio (medicinskih) rečnika, kao što su ICD, SNOMED, OMIM, MeSH, GO, kao i medicinske skraćenice (Slika 3). Sastoji se iz tri dela:

1. metatezaurus - medicinski tezaurusi dostupni u više različitih jezika

2. semantičke mreže - kategorizacija i veze između svih resursa u metatezaursima

3. specijalizovani leksikoni i alati - leksikoni za biomedicinski i opšti engleski jezik.



Slika 3. Vizuelni prikaz UMLS-a

UMLS daje jedistveni identifikator svim sinonimima, Kroz semantičku mrežu su semantičke kategorije dodeljene svim pojmovima. Na primer, groznici - simptom.

Metatezaurus sadrži 215 različitih leksikona za 25 jezika od kojih su najbrojniji resursi za engleski (čak 144), zatim nemački, španski i francuski (Slika 4.). Postoje interesovanja za stalnim dopunjavanjem ovih resursa za različite jezike. UMLS se koristi u raličite svrhe, a neke glavnih su [34]:

- obrada tekstova radi izdvajanja koncepata, relacija i novih znanja;
- lakše mapiranje između terminologija;
- za izdavanje određene terminologije iz metatezaurusa (npr. MedDRA, MeSH, NDF-RT);
- za razvoj sistema za pronalaženje informacija;
- stvaranje i podrška za lokalnu terminologiju;
- istraživanje terminologija i ontologija;
- podrška terminološkog servera ili usluge i dr.

Najčešće korišćeni UMLS proizvodi su metatezaurusi, a posle njih MetaMap[35] koji se koristi za mapiranje koncepata iz metatezaurusa u tekstu.

Slika 4. Broj tezaurusa po različitim jezicima

Kreiranje referentnih korpusa je ključni u procesu razvoja odgovarajućih metoda za rešavanje problema mašinskog prevođenja, deidentifikacije, interakcije lekova itd [36-38].

Većina medicinskih izveštaja sadrži latinske i engleske izraze i skraćenice pomešane sa službenim jezikom koji se koristi. Kod nekih jezika je problem i pisanje u drugačijoj azbuci pa imamo mešavinu i pisama i jezika [39]. Za jezike za koje ne postoje resursi često se oni prave prevođenjem u engleski resurs i korišćenjem već postojećih UMLS relacija [40].

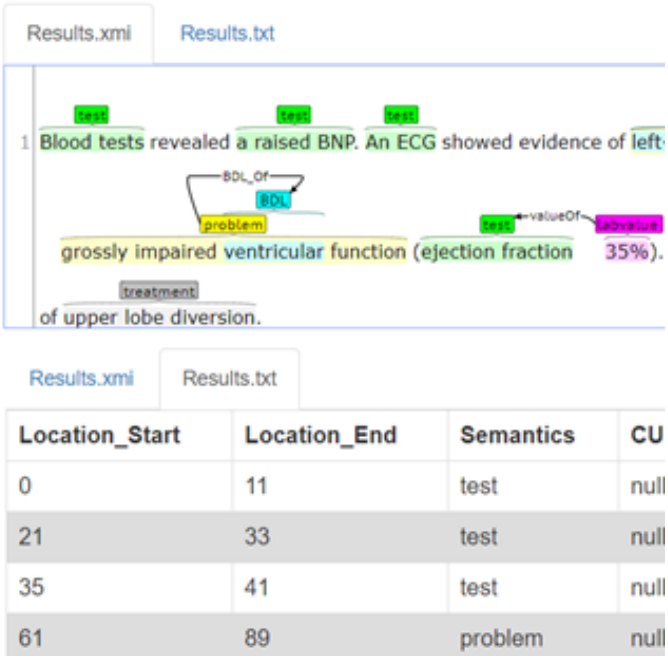
Razvoj višejezičnih paralenih korpusa je takođe značajan ali i mukotrpan proces. Razumevanje različitog stila pisanja doprinosi boljem izdvajanju informacija. Kurtuloroške razlike nam ukazuju da treba da se prave specifični sistemi za svaki jezik. S druge strane negde su razlike samo u jeziku dok je stil pisanja isti i treba se truditi da se stvori mogućnost međusobne saradnje [41].

Tri najpolularnija alata za izvlačenje informacija su MetaMap [42], cTAKES [43] i CLAMP [44]. Zajedničko za ove alate je da vrše mapiranje imenskih entiteta zasnovano na UMLS-u. MetaMap je alat za izdvajanje biomedicinskih informacija. cTAKES je sistem za obradu prirodnog jezika za izdvajanje podataka iz kliničkog slobodnog teksta iz elektronskog medicinskog kartona pomoću pravila zasnovanog na mašinskom učenju. Sadrži sve osnovne funkcije NLP obrade za engleski jezik, kao što su tokenizator, POS označivač, prepoznavač imenovanih entiteta, otkrivanje negacija, funkcionalnost mašinskog učenja itd. Najnoviji NLP alat za klinički tekst CLAMP ima veći naglasak na fleksibilnost u razvoju prilagođenih šema sa mogućnošću njihove primene za izvlačenje informacija. CLAMP je Java alat, ima ugrađene module za obradu prirodnog jezika za engleski tekst. Poređenjem ovih alata u radu [45] je pokazano da CLAMP ima najbolje performanse u pogledu F1 rezultata, i veću preciznost i nešto niži opoziv u poređenju sa cTAKES-a i MetaMap-a.

Na slici 5. je prikazan primer primene CLAMP alata na primeru medicinskog izveštaja EHR na engleskom jeziku:

- EHR : “Blood tests revealed a raised BNP. An ECG showed evidence of left-ventricular hypertrophy and echocardiography revealed grossly impaired ventricular function (ejection fraction 35%). A chest X-ray demonstrated bilateral pleural effusions, with evidence of upper lobe

diversion.”



Slika 5. Primer primene CLAMP alata

- Prevod EHR-a: ”Testovi krvi otkrili su povišeni BNP. EKG je pokazao dokaze o hipertrofiji leve komore, a ehokardiografija je otkrila teško oštećenu funkciju komore (frakcija izbacivanja 35%). Rentgen grudnog koša pokazao je bilateralne pleuralne izlive, sa vidljivom redistribucijom krvi u krvnim sudovima gornjeg režnja.“

Na slici 5. je prikazan .xml i .txt rezultat mapiranja različitih medicinskih entiteta u tekstu kao što su: testovi, simptomi, različite labaratoriske analize i drugo.

VI. ZAKLJUČAK

Analizom postojećih resursa za obradu kliničkih tekstova na različitim prirodnim jezicima može se zaljučiti da je najviše resursa i alata napravljeno za engleski jezik. Veliki napori su da se stvore i alati za druge prirodne jezike. Da bi mogli da se obrađuju klinički tekstovi potrebni su specifični alati za obradu odgovarajućeg prirodnog jezika kao i leksikoni medicinske terminologije na odgovarajućem jeziku. Neki od naših budućih ciljeva su upravo stvaranje odgovarajućih resursa za srpski jezik.

LITERATURA

[1] J. L. Izquierdo, J. Ancochea; Savana COVID-19 Research Group and J. B. Soriano, “Clinical Characteristics and Prognostic Factors for Intensive Care Unit Admission of Patients With COVID-19: Retrospective Study Using Machine Learning and Natural Language Processing“, J Med Internet Res, 2020, vol. 22(10), e21801.

[2] C. Kothari, M. Wack, C. Hassen ‐ Khodja, S. Finan, G. Savova, M. O'Boyle, ... and P. Avillach, “Phelan-McDermid syndrome data network: Integrating patient reported outcomes with clinical notes and curated genetic reports“, American Journal of Medical Genetics Part B: Neuropsychiatric Genetics, 2018, vol. 177.7, pp. 613-624.

[3] C. Safran, C. Chute and J. R. Scherrer, “*Natural Language and Medical Concept Representation*“, Preprints of the IMIA WG6 Conference, Vevey, 1994.

[4] K. R. Chowdhary, "Natural language processing." Fundamentals of Artificial Intelligence. Springer, New Delhi, 2020, pp. 603-649.

[5] D. Hercules, “Clinical text mining: Secondary use of electronic patient records”, Springer Nature, 2018.

[6] L. Zhou, J. M. Plasek, L. M. Mahoney, N. Karipineni, F. Chang, X. Yan, ... and R. A. Rocha, “Using Medical Text Extraction, Reasoning and Mapping System (MTERMS) to process medication information in outpatient clinical notes”, AMIA Annual Symposium Proceedings, American Medical Informatics Association, Vol. 2011.

[7] I. Spasic and G. Nenadic, “Clinical text data in machine learning: Systematic review”, JMIR Medical Informatics, 2020, vol. 8.3, e17984.

[8] S. Horng, D. A. Sontag, Y. Halpern, Y. Jernite, N. I. Shapiro and L. A. Nathanson, “Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning”, PLoS One 2017;, Vol. 12(4).

[9] S. J. Fodeh, D. Finch and L. Bouayad, “Classifying clinical notes with pain assessment using machine learning. Medical & Biological Engineering & Computing”, 2018, Vol.56(7), pp. 1285-1292.

[10] i2b2: Informatics for Integrating Biology & the Bedside

[11] Ö. Uzuner, X. Zhang and T. Sibanda, “Machine learning and rule-based approaches to assertion classification“, Journal of the American Medical Informatics Association, 2009, Vol. 16(1), pp. 109-115.

[12] H. Suominen, S. Salanterä, S. Velupillai, W. W. Chapman, G. Savova, N., Elhadad, ... and G. Zuccon, “Overview of the ShARe/CLEF eHealth evaluation lab 2013“, In International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, Berlin, Heidelberg, 2013, Vol. 8138, pp. 212-231.

[13] M. Saeed, M. Villarroel, A. T. Reisner, G. Clifford, L. W. Lehman and G. Moody, “Multiparameter intelligent monitoring in intensive care II (MIMIC-II): A public-access intensive care unit database“, Critical Care Medicine, 2011, Vol. 39(5), pp. 952.

[14] V. Vincze, G. Szarvas, R. Farkas, G. Móra and J. Csirik, “The BioScope Corpus: Biomedical texts annotated for uncertainty, negation and their scopes“, BMC Bioinformatics, 2008, Vol. 9(11), S9.

[15] M. Oronoz, K. Gojenola, A. Pérez, A. D. de Ilarraza and A. Casillas, “On the creation of a clinical gold standard corpus in spanish: mining adverse drug reactions“, J. Biomed. Inform. 2015, Vol. 56, pp. 318–332.

[16] S. Boytcheva, G. Angelova, Z. Angelov and D. Tcharaktchiev, “Text mining and big data analytics for retrospective analysis of clinical texts from outpatient care“, Cybernetics and Information Technologies, 2015, Vol. 15(4), pp. 58-77.

[17] S. Boytcheva, I. Nikolova, G. Angelova and Z. Angelov, “Identification of risk factors in clinical texts through association rules“, In Proceedings of RANLP Workshop on Biomedical Natural Language Processing, 2017, pp. 64–72.

[18] A. Avdic, U. Marovac and D. Jankovic, “Automated labeling of terms in medical reports in Serbian“, Turkish Journal of Electrical Engineering & Computer Sciences, 2020, Vol. 28.6, pp. 3285-3303.

[19] H.Dalianis, A. Henriksson, M. Kvist, S. Velupillai and R. Weegar, “HEALTH BANK – A workbench for data science applications in healthcare“, In J. Krogstie, G. Juel-Skielse, & V. Kabilan (Eds.), Proceedings of the CAiSE-2015 Industry Track Co-located with 27th Conference on Advanced Information Systems Engineering (CAiSE 2015), Stockholm, Sweden, June 11, 2015, Vol. 1381, pp. 1–18.

[20] R. Eriksson, P. B. Jensen, S. Frankild, L. J. Jensen and S. Brunak, “Dictionary construction and identification of possible adverse drug events in Danish clinical narrative text“. Journal of the American Medical Informatics Association, 2013, Vol. 20(5), pp. 947–953.

[21] https://biosemantics.erasmusmc.nl/index.php/resources/emc-dutch-clinical-corpus

[22] http://bionlp.utu.fi/clinicalcorpus.html.

[23] L. Campillos, L. Deléger, C. Grouin, T. Hamon, A. L. Ligozat and A. Névéol, “A French clinical corpus with comprehensive semantic annotations: development of the Medical Entity and Relation LIMSI annOtated Text corpus (MERLOT)“, Language Resources and Evaluation, 2018, Vol. 52(2), pp. 571-601.

[24] E. Chiaramello, F. Pinciroli, A. Bonalumi, A. Caroli and G. Tognola, “Use of ‘off-the-shelf’ information extraction algorithms in clinical

informatics: a feasibility study of MetaMap annotation of Italian medical notes“, J. Biomed. Inform. 2016, Vol. 63, pp. 22–32.

[25] G. Attardi, V. Cozza and D. Sartiano, “Annotation and extraction of relations from Italian medical records“, In Proceedings of the 6th Italian Information Retrieval Workshop, Cagliari, Italy, 2015.

[26] S. Spat, B. Cadonna, I. Rakovac, C. Gütl, H. Leitner and G., Stark, “Enhanced information retrieval from narrative German-language clinical text documents using automated document classification“, Studies in Health Technology and Informatics, 2008, Vol. 136, pp. 473.

[27] M. Zubke, “Classification based extraction of numeric values from clinical narratives“, In Proceedings of RANLP Workshop on Biomedical Natural Language Processing, 2017, pp. 24–31.

[28] R. Roller, H. Uszkoreit, F. Xu, L. Seiffe, M. Mikhailov and O. Staeck, “A fine-grained corpus annotation schema of German nephrology records“, In Proceedings of the Clinical Natural Language Processing Workshop, Osaka, Japan, 2016, pp. 69–77.

[29] B. He, B. Dong, Y. Guan, J. Yang, Z. Jiang, Q. Yu, ... and C. Qu, “Building a comprehensive syntactic and semantic corpus of Chinese clinical texts“, Journal of biomedical informatics, 2017, Vol. 69, pp. 203-217.

[30] A. Alicante, A. Corazza, F. Isgrò and S. Silvestri, “Unsupervised information extraction from Italian clinical records“, Proceeding of Innovation in Medicine and Healthcare, 2014, pp. 340-349.

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

[32] https://www.snomed.org/

[33] O. Bodenreider, “The unified medical language system (UMLS): integrating biomedical terminology“. Nucleic acids research, 2004, Vol. 32(suppl_1), D267-D270.

[34] L. Amos, D. Anderson, S. Brody, A. Ripple and B. L. Humphreys, “UMLS users and uses: a current overview“, Journal of the American Medical Informatics Association, 2020, Vol. 27(10), pp. 1606-1611.

[35] https://metamap.nlm.nih.gov/

[36] A. Vagelatos, E. Mantzari, M. Pantazara, C. Tsalidis and C. Kalamara, “Developing tools and resources for the biomedical domain of the Greek language“, Health informatics journal, 2011, Vol. 17(2), pp. 127-139.

[37] C. Grouin, T. Lavergne and A. Névéol, “Optimizing annotation efforts to build reliable annotated corpora for training statistical models“, In: 8th Linguistic Annotation Workshop – LAW VIII, 2014, pp. 54–58.

[38] M. Skeppstedt, M. Kvist, G. Nilsson and H. Dalianis, “Automatic recognition of disorders, findings, pharmaceuticals and body structures from clinical text: An annotation and machine learning study“, Journal of Biomedical Informatics, 2014. pp. 148–158.

[39] S. Boytcheva, “Multilingual aspects of information extraction from medical texts in Bulgarian“, Multilingual Processing in Eastern and Southern EU Languages: Less-resourced Technologies and Translation, Cambridge Scholars Publishing, 2012, pp. 308-29.

[40] L. Deléger, M. Merkel and P. Zweigenbaum, “Translating medical terminologies through word alignment in parallel text corpora“, Journal of Biomedical Informatics, 2009, Vol. 42(4), pp. 692-701.

[41] Y. Wu, J. Lei, W. Wei, B. Tang, J. Denny, S. Rosenbloom, R. Miller, D. Giuse, K. Zheng and H. Xu, “Analyzing differences between Chinese and English clinical text: a cross-institution comparison of discharge summaries in two languages“, Stud Health Technol Inform, 2013, pp. 662–666.

[42] A. R. Aronson, F. M. Lang, “An overview of MetaMap: historical perspective and recent advances“, J Am Med Inform Assoc. 2010, Vol. 17(3), pp. 229–36.

[43] G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, C. G. Chute, “Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications“, J Am Med Inform Assoc, 2010, Vol. 17(5), pp. 507–13.

[44] E. Soysal, J. Wang, M. Jiang, Y. Wu, S. Pakhomov, H. Liu and H. Xu “CLAMP–a toolkit for efficiently building customized clinical natural language processing pipelines“, J Am Med Inform Assoc, 2018, Vol. 25(3), pp. 331–6.

[45] J. Peng, M. Zhao, J. Havrilla, C. Liu, C. Weng, W. Guthrie, ... and Y. Zhou, “Natural language processing (NLP) tools in extracting biomedical concepts from research articles: a case study on autism spectrum disorder“, BMC Medical Informatics and Decision Making, 2020, Vol. 20(11), pp. 1-9.

ABSTRACT

Natural language processing is one of the most widespread areas of artificial intelligence. According to the domain of the texts for which it is intended, it requires various resources necessary for the processing of the appropriate natural language. Clinical texts carry a wealth of information that can be analyzed in order to better understand the health status of the patient as well as in order to have a comprehensive understanding of the disease. The aim of this paper is to present the resources for processing clinical texts in different natural languages.

AN OVERVIEW OF RESOURCES FOR CLINICAL TEXT PROCESSING IN DIFFERENT NATURAL LANGUAGES

Ulfeta Marovac, Aldina Avdić