Robert Tovornik
Trinkova ulica 62, 1000 Ljubljana, Slovenija
Študijski program: Računalništvo in informatika, MAG
Vpisna številka: 63120292

Komisija za študijske zadeve
Univerza v Ljubljani, Fakulteta za računalništvo in informatiko
Večna pot 113, 1000 Ljubljana

Vloga za prijavo teme magistrskega dela
Kandidat: Robert Tovornik

Robert Tovornik, študent magistrskega programa na Fakulteti za računalništvo in informatiko, zaprošam Komisijo za študijske zadeve, da odobri predloženo temo magistrskega dela z naslovom:

Slovenski: **Avtomatska prepoznava in digitalizacija obrazcev**
Angleški: **Automatic recognition and digitalisation of forms**

Tema je bila že potrjena lani in je ponovno vložena: ***NE***

Izjavljam, da so spodaj navedeni mentorji predlog teme pregledali in odobrili ter da se z oddajo predloga strinjajo.

Magistrsko delo nameravam pisati v slovenščini.

Za mentorja predlagam:

Ime in priimek, naziv: Slavko Žitnik, doc. dr
Ustanova: Univerza v Ljubljani, Fakulteta za računalništvo in informatiko
Elektronski naslov: slavko.zitnik@fri.uni-lj.si

V Ljubljani, 5. december 2021.

PREDLOG TEME MAGISTRSKEGA DELA

1 Področje magistrskega dela

slovensko: računalništvo in informatika, ekstrakcija informacij, obdelava naravnega jezika, strojno zaznavanje, spletne aplikacije
angleško: computer science, information extraction, natural language processing, machine perception, web applications

2 Ključne besede

slovensko: optična prepoznava znakov OCR, prepoznava postavitve, digitalizacija
angleško: optical character recognition OCR, layout recognition, digitalization

3 Opis teme magistrskega dela

Pretekle potrditve predložene teme:
Predložena tema ni bila oddana in potrjena v preteklih letih.

3.1 Uvod in opis problema

Podjetje Better d.o.o. deluje na področju zdravstva, kjer svojim strankam nudi odprto podatkovno platformo s tehnologijo openEHR[1, 2] za strukturirano hranjenje digitalnih podatkov. Problem s katerim se srečujejo je zbiranje podatkov v nedigitalni obliki. Na področju zdravstva zbiranje podatkov velikokrat poteka v obliki papirnatih obrazcev, ki jih izpolni pacient. Na papir izpolnjeni obrazci zahtevajo nadaljnje administrativno delo medicinskih sester ali drugih tehnikov, ki morajo le-te ročno pregledati in relevante podatke vnesti v sistem.

Takšen način zbiranja podatkov nepotrebno obremenjuje zdravstveno osebje z nalogami, ki so vezane na administrativne postopke namesto da bi ta čas namenili pacientu[3]. Vendar potreba po obrazcih ostaja zaradi specifičnost zajema podatkov, ki jih zdravnik potrebuje za uspešno postavitev diagnoze. Različne težave pacientov pomenijo zajem različnih podatkov, kar pomeni potrebo po različnih obrazcih. In le-teh je v zdravstvu mnogo.

Ker je vsak trenutek, ki ga lahko zdravnik ali zdravstveno osebje nameni pacientu namesto administraciji zelo pomemben, je želja postopek zbiranja podatkov poenostaviti in časovno pohitriti. Podjetje Better v okviru svoje platforme že ponuja orodje za izdelavo digitalnega obrazca. Problem ki ostaja je časovno in tehnično zahtevna pretvorba obstoječih papirnatih obrazcev v digitalno obliko.

V magistrski nalogi se bom osredotočil na izdelavo sistema za avtomatsko prepoznavo strukture in gradnikov obrazcev, vsebovanega besedila, različnih tipov vnosnih polj ter na ekstrakcijo in prepoznavo vsebine relevantne za medicinske entitete.

Ker je želja podjetja rešitev uporabiti na mednarodnem trgu in pri različnih ustanovah, je potrebno poskrbeti za generičnost, prilagodljivost in večjezičnost sistema. Hkrati, ker gre za zdravstveno področje pa tudi za natančnost in zanesljivost prepoznave vsebine. Hitrost in paralelnost sistema sta zaradi številčnosti dokumentov pomembni, vendar pa glavna metrika ostaja hitrost v primerjavi z ročnim vnosom.

3.2 Pregled sorodnih del

Samo področje digitalizacije papirnatih dokumentov ni noviteta. Želja po vnosu podatkov s papirja v računalniške sisteme se je prvič komercialno realizirala že leta 1954 [4] v podjetju Reader's Digest. Od takrat dalje se je tehnologija razvijala z novim večjim preskokom leta 2005 ko je podjetje Hewlett-Packard projekt Tesseract OCR [5] objavilo kot odprtokodni projekt. Pod okriljem Googla od leta 2006 dalje se tehnologija uporablja vsepovsod z glavno spremembo pri razpozavanju, kjer postajajo vodilna tehnologija globoke nevronske mreže - v primeru Tesseract OCR gre za LSTM (Long Short Term Memory) arhitekturo.

Sočasno se je zaradi aplikacije za komercialno uporabo, s fokusom na področje računov, izdajnic in drugih računovodskih dokumentov razvilo tudi področje razpoznavanja tabel[6]. Glavni ponudniki komercialnih sistemov na trgu, ki ponujajo tako optično prepoznavo besedila kot tudi detekcijo in razpoznavo tabel so trenutno: ABBYY FlexiCapture [7], Amazon Textract[8], Google Document AI [9] in Microsoft Form Recognizer[10]. Vsi izmed naštetih pa storitev razširjajo še s prepoznavo imenski entitet za katero je potrebnih nekaj vzorcev istega tipa dokumenta na podlagi katerega se do-trenira klasifikator entitet. Manjka pa celostna rešitev, namreč razpoznavalniki so na voljo kot posamezne storitve, ki lahko ponudijo medsebojno izključujoče rezultate.

Sorodnega problema prepoznave in ekstrakcije teksta so se lotili tudi [11], s fokusom na pripravi podatkovne zbirke skeniranih dokumentov s šumom (noisy scanned documents). Dokumente so ročno anotirali za detekcijo in prepoznavo besedil, strukture in razpoznavo obrazcev. Na zbirki so preizkusili tudi natančnost nekaterih izmed predhodno omenjenih

storitev ter nekaj svojih rešitev vezanih na klasifikacijo besedilnih entitet.

Na veliki skali si je podoben projekt zadala tudi raziskovalna ekipa podjetja Adobe Systems [12]. Posebnost njihovega pristopa je, da pri uporabi globokih mrež ohranjajo visoko ločljivost slik, pri čemer večina drugih zaradi procesne zahtevnosti ločljivost zmanjšuje. Vendar pa je njihov pristop težko ponovljiv, saj temelji na podatkovnem naboru v velikosti nekaj več kot 52.000 ročno anotiranih dokumentov, kar je za večino raziskovalcev časovno in finančno nedosegljivo.

3.3 Predvideni prispevki magistrske naloge

Temeljni prispevek magistrske naloge bo v implementaciji generičnega sistema za prepoznavo vsebine in strukture poljubnih obrazcev. S pomočjo razpoznave posameznih strukturnih gradnikov, njihove lege in postavitve na strani ter medsebojne razporeditve bo rekonstruiral obrazec v digitalni obliki. Na podlagi metrik bo posamezne elemente obrazcev združeval v logične in vsebinske sklope. S prepoznavo entitet bo poskusil kategorizirati podatke v obstoječe podatkovne strukture.

Sekundarni prispevek po razvita povezovalna logika, ki bo združevala informacije iz različnih stopenj zajema podatkov. Ocenil bom smiselnost posameznih pristopov pri zajemu podatkov v različnih stopnjah postopka. Ocenil bom tudi smiselnost vrstnega reda zajema in združevanja podatkov iz različnih virov t.j.: podatki iz slike, podatki iz besedila, ter združevanja različnih vrst podatkov znotraj posameznih sklopov.

V postopku izbire bom ocenil tudi delovanje posameznih sistemov za prepoznavo besedila, vnosnih polj in smiselnost uporabe posameznih sistemov, pri čemer bom preverjal natančnost in hitrost prepoznave ter kompleksnost in uporabnost izhodnih podatkov sistemov.

3.4 Metodologija

Sistem za prepoznavo obrazcev bo izdelan s programskim jezikom Python. Implementiran bo kot spletna aplikacija za nalaganje, procesiranje, potencialno urejanje in na koncu izvoz obrazca v digitalni obliki. Uporabljeno bo ogrodje za razvoj spletnih aplikacij (Flask in/ali FastAPI).

Priprava podatkov
Zbral bom podatkovno zbirko različnih tipov obrazcev. Ker želimo generično rešitev bo pri izbiri dokumentov pomembna vsebovanost ponavljajočih se gradnikov oz. tipov vnosnih elementov, na primer: polja ki omogočajo eno izbiro (radiobutton), polja ki omogočajo več vnosov (multi-choice / checkboxes), tabele idr.. Torej posamezni gradniki obrazcev,

ki bodo fokus prepoznave ob različnih postavitvah.

Zajem podatkov iz slike
Za pred-obdelavo slik in zajem vizualnih podatkov iz slike bo uporabljena knjižnica OpenCV nad katero bom razvil dodatne komponente za prepoznavo posameznih gradnikov. Za optično prepoznavo besedila bo uporabljena temu namenska odprtokodna knjižnica Tesseract OCR. Potencialno se lahko za korak prepoznave besedila uporabi tudi katera izmed komercialnih rešitev.

Zajem podatkov iz besedila
Nad prepoznanim besedilom bo izvedena obdelava besedil s tehnikami procesiranja naravnega jezika. Nad obdelanimi besedilnimi segmenti bom tvoril napovedne lastnosti, ki bodo omogočile klasifikacijo besedilnih segmentov v vnaprej določene razrede. Dodatno bo nad besedilnimi segmenti izveden algoritem za iskanje in ekstrakcijo entitet, relevantnih za podatkovno strukturo v ozadju. Namen je preslikati posamezne besedilne segmente v semantične kategorije ter jih navezati na podatkovne strukture.

Analiza
Rezultati bodo analizirani in evalvirani polavtomatsko. Prepoznava bo izvedena nad celotnim podatkovnim naborom zbranih dokumentov, čemur bo sledila ročna evalvacija naključne selekcije. Ocenjevala se bo točnost prepoznanega besedila, pravilnost in številčnost pravilno prepoznanih gradnikov in elementov obrazca, pravilnost povezanih elementov v logične / semantične strukture ter pravilno kategoriziranih besedilnih entitet.

3.5 Literatura in viri

[1] H. Leslie, openehr - the world's record (06 2015). doi:10.13140/RG.2.1.1398.0321.

[2] G.-H. Ulriksen, R. Pedersen, G. Ellingsen, Infrastructuring in healthcare through the OpenEHR architecture, Computer Supported Cooperative Work (CSCW) 26 (1-2) (2017) 33–69. doi:10.1007/s10606-017-9269-x.
URL `https://doi.org/10.1007/s10606-017-9269-x`

[3] B. Y. Lee, Doctors wasting over two-thirds of their time doing paperwork (Sep 2016).
URL `https://www.forbes.com/sites/brucelee/2016/09/07/doctors-wasting-over-two-thirds-of-their-time-doing-paperwork/?sh=13e217785d7b`

[4] A. Robertson, OPTICAL CHARACTER RECOGNITION, Management Decision 9 (3) (1971) 213–223. doi:10.1108/eb000971.
URL `https://doi.org/10.1108/eb000971`

[5] R. Smith, An overview of the tesseract ocr engine, in: Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), Vol. 2, 2007, pp. 629–633. doi:10.1109/ICDAR.2007.4376991.

[6] F. Shafait, R. Smith, Table detection in heterogeneous documents, in: Document Analysis Systems 2010, 2010.
URL http://doi.acm.org/10.1145/1815330.1815339

[7] Intelligent document processing solution: Abbyy flexicapture.
URL https://www.abbyy.com/flexicapture/

[8] M. M, P. Rangarajan, Natural language processing with aws ai services: implement various nlp use cases from unstructured data using amazon comprehend and amazon textract (2021).
URL https://aws.amazon.com/textract/

[9] Document ai | google cloud.
URL https://cloud.google.com/document-ai

[10] form recognizer – automated data processing systems | microsoft azure_2021 (2021).
URL https://azure.microsoft.com/en-us/services/form-recognizer/

[11] G. Jaume, H. K. Ekenel, J. Thiran, FUNSD: A dataset for form understanding in noisy scanned documents, CoRR abs/1905.13538 (2019). arXiv:1905.13538.
URL http://arxiv.org/abs/1905.13538

[12] M. Sarkar, M. Aggarwal, A. Jain, H. Gupta, B. Krishnamurthy, Document structure extraction for forms using very high resolution semantic segmentation, CoRR abs/1911.12170 (2019). arXiv:1911.12170.
URL http://arxiv.org/abs/1911.12170