Zbornik radova Fakulteta tehničkih nauka, Novi Sad

UDK: 004.85
DOI: https://doi.org/10.24867/16BE05Katic

SISTEM ZA ODGOVORE NA PITANJA U DOMENU FITNESA ZASNOVAN NA MAŠINSKOM UČENJU

QUESTION ANSWERING SYSTEM IN FITNESS DOMAIN BASED ON MACHINE LEARNING

Sava Katić, *Fakultet tehničkih nauka, Novi Sad*

Oblast – ELEKTROTEHNIKA I RAČUNARSTVO

Kratak sadržaj – *U ovom radu predstavljen je sistem za odgovore na pitanja u oblasti fitnesa i nutricionizma, koji dovoljno dobro generalizuje i za opšti domen. Kao ulaz model prima pitanje u obliku niza karaktera, a potom traži najbolje kandidate dokumente u bazi znanja, koji imaju odgovor na pitanje koje je na ulazu u sistem postavljeno.*

Ključne reči: *Odgovori na pitanja, Jezički modeli, NLP, QA, BERT*

Abstract – *This paper presents system for question answering focused on fitness and nutrition field, that works just as well in open domain. As an input model accepts a question in a form of array of characters and finds best document candidates in a knowledge base from which the actual answer is extracted.*

Keywords: *Chatbot, Language models, NLP, QA, BERT*

1. UVOD

Jezik i govor su delom ono što nas čini ljudima. Nije iznenađenje da nam je daleko lakše komunicirati prirodnim jezikom nego tipkanjem, kliktanjem i kodiranjem. Noviji algoritmi veštačke inteligencije pomažu nam da premostimo tu razliku tako što simuliraju razumevanje našeg prirodnog jezika i konverzaciju u istom. U konkretnom domenu fitnesa na koji se ovaj rad fokusira, ovakav sistem bi omogućio da se besplatno dođe do odgovora na pitanja (engl. *question answering*) u vezi sa zdravljem, vežbanjem i ishranom na brz, jednostavan i stalno pristupačan način. Komponente koje čine sistem su: baza znanja, čitač i sakupljač. Baza znanja je baza podataka koja sadrži tekstualne dokumente iz određenog domena na osnovu kojih se daju odgovori, organizovanih tako da ih je jednostavno pretraživati. Sakupljač je komponenta koja pretražuje dokumente iz baze znanja i pronalazi kandidate koji bi mogli da sadrže odgovor na postavljeno pitanje koje je ulaz u sistem, dok je čitač neuronska mreža koja u dokumentima kandidatima pronalazi odgovor na dato pitanje. Za skup podataka korišćeno je više izvora kako bi se izgenerisali dokumenti koji sadrže bitne informacije u polju fitnesa i nutricionizma. Na osnovu ovih dokumenata su generisana pitanja i odgovori koristeći *cloze translation* [2].

NAPOMENA:
Ovaj rad proistekao je iz master rada, čiji mentor je bio prof. dr Aleksandar Kovačević.

Nakon analize skupa podataka, pretprocesiranja teksta i izdvajanja odgovarajućih obeležja, iskorišćena je čitač sakupljač arhitektura koja ima za cilj da izdvoji relevantne dokumente na osnovu prosleđenog pitanja, a zatim i nađe odgovor u pasusima tih dokumenata. Za čitač komponentu korišćeni su *RoBERTa* [4] i *MiniLM* [3] arhitekture mreže, koje koriste *transformer* i bidirekcionalnost čiji su se rezultati međusobno poredili. Za sakupljač komponentu implementirani su *TF-IDF*, *BM25* i *DPR* [5] pristupi. Baza znanja se čuvala u *ElasticSearch* [1] servisu koji omogućava brzu pretragu. Nakon toga, vršeno je evaluiranje i poređenje rezultata između različitih arhitektura kako bi utvrdili u kojim slučajevima model greši i kako bi se greške mogle izbeći. Napravljena je i *web* aplikacija koja nudi interfejs za postavljanje pitanja i poziva odgovarajuće modele putem *HTTP* protokola.

U narednom poglavlju biće analizirana relevantna literatura za problem odgovora na pitanja koji se analizira. U trećem poglavlju biće opisana predložena arhitektura sistema kao i implementacija rešenja dok četvrto poglavlje nudi uvid u rezultate i njihovu diskusiju. U petom poglavlju dat je zaključak o samom radu i problemu koji je rešavan.

2. PRETHODNA REŠENJA

Modeli za odgovore na pitanja su modeli mašinskog ili dubokog učenja koji mogu da odgovore na određena pitanja na osnovu konteksta, tj. dela teksta koji sadrži odgovor na postavljeno pitanje. Postoje implementacije koje ne zahtevaju da kontekst u kom se nalazi odgovor bude prosleđen i ovaj rad se bavi jednim takvim rešenjem sa kraja na kraj (engl. *end-end*).

Postoji više *NLP* modela koji su u prošlosti primenjeni za rešavanje problema odgovora na pitanja. Jedan od njih koristi gramatičko označavanje i *TF-IDF* pristup da pitanje koje je postavljeno pretraži na *Yahoo Answers*[2] i nađe odgovor na isto. Zatim, potpuno nova arhitektura neuronskih mreža zasnovana na pažnji, posebno na samopažnji (engl. *self-attention*) nazvana *transformer* je zaista ponudila najbolji pristup za reprezentaciju teksta i prirodnog jezika [1, 3, 4, 6, 7].

Jezički model je probabilistički model koji uči verovatnoću pojavljivanja rečenice ili sekvence reči na osnovu primera teksta viđenih tokom treninga, gde je najzastupljeniji model *BERT* [1]. Razlog za popularnost *BERT*

[1] https://www.elastic.co/
[2] https://www.yahoo.com/

modela je što uvodi inovaciju koristeći bidirekciono treniranje *transformera* za modelovanje jezika (engl. *language modeling*). Ovo se razlikuje od prethodnih pokušaja da se tekst analizira i da se modeli treniraju ili sa leva na desno ili kombinujući sa leva na desno i sa desna na levo pristup.
Postoji dosta pretreniranih modifikacija na *BERT* model koje su dale nove *state of the art* modele , a takođe i treniranih za specifični domen kao što su *BioBERT*, *SciBERT* i *ClinicalBERT* [7]. Takođe, drugi modeli dobijeni korišćenjem destilacije znanja izložene u *MiniLM* radu [3] zadržavaju najveći deo performansi pretreniranih *state of the art* modela, dok imaju mnogo manje parametara.
Za domen fitnesa, nije pronađen odgovarajući model, skup podataka kao ni istraživanje na temu odgovora na pitanja. Iz tog razloga skup podataka će se generisati sintetički. U narednom poglavlju biće opisan implementirani sistem sa kraja na kraj za odgovore na pitanja.

3. METODOLOGIJA I ALATI

Ovo poglavlje posvećeno je korišćenim alatima (poglavlje 3.1) kao i prikazu primenjene metodologije i arhitekturi sistema sa kraja na kraj koji na osnovu pitanja parsira veliki korpus dokumenata u potrazi za odgovorom.

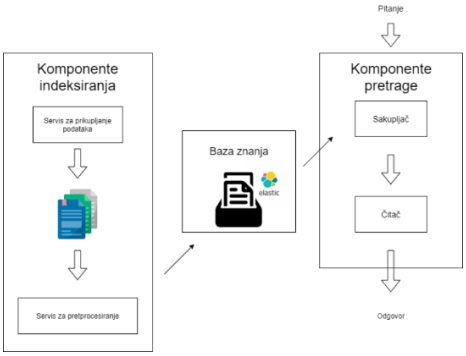
3.1. Korišćeni alati

Server na kom je treniran model je *p2.xlarge* tip instance sa jakom grafičkom karticom u okviru *AWS*[3] alata. Server ima 2.7 GHz *Intel Xeon E5-2686* v4 procesor i *NVIDIA K80 12 GB* grafičku karticu. Radi portabilnosti treniranja iskorišćena je kontejnerizacija (engl. *containerization*) koristeći *Docker*[4] softver.

Pretrenirani modeli za problem odgovora na pitanja specificirani su koristeći *Deepset - Haystack* [5] radno okruženje koje se oslanja na *PyTorch* [6] biblioteku za razvoj modela veštačke inteligencije.

3.2. Arhitektura rešenja

Predloženo rešenje fokusira se na podatke u domenu fitnesa, ali je dovoljno generično da bude upotrebljeno i za druge domene. Iskorišćena je čitač sakupljač arhitektura kako bi se dobio što performantniji sistem, a komunikacija komponenti može se videti na slici 1. Objašnjenje svake komponente je u nastavku.



Slika 1: *Arhitektura rešenja*

[3] https://aws.amazon.com/
[4] https://www.docker.com/
[5] https://github.com/deepset-ai/haystack
[6] https://pytorch.org/

Da bi većina pitanja i pojmova o fitnesu i nutricionizmu bila pokrivena, korišćeno je nekoliko izvora podataka kako bi pokrivali što veći skup činjenica. Tekstovi za trening i evaluacione podatke su skinuti sa sajtova *muscleandfitness*[7], *myfitnesspal*[8], *breakingmusclefitness*[9] , *advancedhumanperformance*[10]. Pored ovih sajtova, korišćeni su artikli sa *Wikipedia* u kategorijama fitnesa, trčanja, bodibildinga, joge i nutricionizma. Da bi bila pokrivena i pitanja o najboljim vežbama za određenu grupu mišića korišćen je *API*[11] koji daje najbolje vežbe spram prosleđene grupe mišića, kako bi se generisale rečenice koje prave preporuke u vidu vežbi. Potom se nad ovako dobijenim podacima vrši pretprocesiranje i dobijeni dokumenti se smeštaju u bazu dokumenata.
U okviru pretprocesiranja podataka izbačene su stop reči i rađena je lematizacija (engl. *lemmatization*) i *stemming*. Generisani dokumenti su se prosleđivali *BERT* modelu koji je generisao tri ključne reči tako što je pravio reprezentaciju (engl. *embedding*) celog dokumenta i svake reči u dokumentu. Potom su dokumenti čije su ključne reči van domena relevatnog za ovo istraživanje bili isfiltrirani.
Za generisanje pitanja i odgovora kako bi se napravio trening skup kojim će se modeli dotrenirati korišćen je pristup putem ekstrakcije imenica, fraza ili imenovanih entiteta [2] koji se potom maskiraju i služe za generisanje pitanja u prirodnom govoru. Tako generisan odgovor je smatran tačnim (engl. *ground truth*) Potom se generisano pitanje prosleđuje *T5 Text-To-Text Transfer Transformer* modelu koji je istreniran nad *Quora*[12] skupom podataka da parafrazira pitanje prosleđeno na ulazu.
Pored ovako generisanog skupa podataka, iskorišćen je i *Natural Questions Google Search*[13] skup podataka koji sadrži pitanja otvorenog domena, da bi se proverilo da li model gubi sposobnost generalizacije nakon što se dotrenira na generisanim pitanjima iz domena fitnesa.

Sakupljač ili pretraživač je filter komponenta koja prolazi kroz celokupnu bazu znanja i pronalazi dokumente koji su kandidati na osnovu pitanja prosleđenog na ulazu. Ovaj alat služi da se izbace negativni kandidati i uštedi vreme čitaču da ne radi dodatan posao i prolazi kroz sve dokumente kako bi našao odgovor na pitanje, ubrzavajući proces upita na taj način.

Neke od mogućih reprezentacija sakupljača:

- *TF-IDF* - statistička mera odnosno funkcija rankiranja relevantnosti dokumenata spram upita koji se prosleđuje
- *BM25* - funkcija rankiranja relevantnosti dokumenata spram upita koja ne koristi neuronsku mrežu za rankiranje i predstavlja varijantu ranije opisanog *TF-IDF* pristupa
- *DPR* - visoko performantan metod za računanje relevantnosti koristeći duboko učenje. Funkcioniše tako što koristi jednu *BERT* mrežu da enkodira dokumente i jednu *BERT* mrežu da enkodira upite.

[7] https://www.muscleandfitness.com/
[8] https://www.myfitnesspal.com/
[9] https://breakingmuscle.com/
[10] https://www.advancedhumanperformance.com/
[11] https://github.com/davejt/exercise
[12] https://quoradata.quora.com/First-Quora-Dataset-Release-Question-Pairs
[13] https://ai.google.com/research/NaturalQuestions/

U ovom sistemu iskorišćeni su *TF-IDF*, *BM25* i *DPR* i njihovo poređenje dato je u poglavlju 5.

Baza znanja je baza podataka koja čuva tekst relevantan za domen u vidu dokumenata (koji mogu biti paragrafi) i nudi te podatke pri upitu.

Neke od najčešćih implementacija baze znanja:

- *SQL baza* – relaciona baza, podaci organizovani u tabelama, loše se skalira
- *ElasticSearch* – fleksibilna i brza biblioteka otvorenog koda, čuva dokumente u vidu *inverted index* što omogućava brzu pretragu teksta

U ovom sistemu iskorišćen je *ElasticSearch* zbog načina na koji čuva tekstualne podatke i jer nudi veoma brzu pretragu nad velikim korpusom dokumenata, a takođe je eksperimentisano i sa *SQL* rešenjem koje se pokazalo da ima dosta manju brzinu izvršavanja upita u poređenju sa *ElasticSearch*.

Čitač predstavlja glavnu komponentu sistema za određivanje odgovora na pitanje. Ulazi u ovu komponentu su dokumenti koji su kandidati da sadrže odgovor dobijeni od sakupljača i pitanje a izlaz su mogući odgovori na postavljeno pitanje. Ova komponenta koristi *transformer* arhitekturu najčešće jer takvi modeli ostvaruju *state of the art* rezultate za NLP oblast jer razumeju semantiku i kontekst, u poređenju sa *LSTM* pristupom koji je daleko sporiji i nije zaista bidirekcion.

Za čitač komponentu u implementaciji ovog sistema korišćeni su *RoBERTa* , koji uvodi nasumično maskiranje u *BERT* model [4], i *MiniLM* [3] model koji su upoređeni u poglavlju 5.

MiniLM koristi tehniku destilacije znanja (engl. *knowledge distillation*) koja kompresuje veliki model koji se naziva učitelj (engl. *teacher*) u mali model, koji se naziva učenik (engl. *student*). Glavna ideja je oponašanje slojeva samopažnje, koji su ključna koponenta *transformer* modela. Tačnije, vrši se destilacija modula samopažnje poslednjeg *transformer* sloja u modelu učitelja. Na ovaj način se odbacuju poteškoće koje mogu da nastanu ako se vrši mapiranje svakog sloja učitelja na neki sloj u modelu učenika [3].

Pored navedenih pristupa, napravljen je i osnovni *baseline* pristup za poređenje sa ostalim modelima koji koristi *TF-IDF* metriku za pronalaženje rečenica kandidata, a potom koristi prepoznavanje imenovanih entiteta da pronađe odgovor u okviru tih rečenica. U poglavlju 4 dati su rezultati i poređenje ovih pristupa.

4. EKSPERIMENTALNI REZULTATI I DISKUSIJA

Eksperimentalna evaluacija se vrši radi određivanja performansi klasifikacije svih algoritama za mašinsko učenje u kombinaciji sa tehnikama pretprocesiranja, navedenim u prethodnom poglavlju ovog rada.

Tekstualni podaci, dobijeni sa 3 prethodno pomenuta izvora, nakon obrade i formatiranja prosleđeni su algoritmu za generisanje pitanja i odgovora. Sveukupno, bilo je 20356 *(pitanje, odgovor, kontekst)* trojki u trening skupu i 5124 u test skupu. Pored ovoga, iskorišćen je i *Natural Questions Google Search* skup podataka koji sadrži pitanja otvorenog domena, da bi proverili da li pretrenirani model gubi sposobnost generalizacije nakon što se dotrenira. Kao metrika je pri evaluaciji modela korišćena *F1* mera.

4.1. Evaluacija sistema

Modeli sa kojima je rađena evaluacija čitača su osnovni, jednostavni model, kao i *MiniLM* i *RoBERTa* sa pretreniranim i dotreniranim verzijama. U tabeli 4.1 dati su rezultati evaluacije čitača.

Tabela 1: F1 mera *RoBERTa* i *MiniLM* modela

Podaci / Model	NQ Google General QA skup podataka	Sintetički fitnes skup podataka
Osnovni (*baseline)*	0,05	0,15
MiniLM pretrenirani	0,19	0,39
RoBERTa pretrenirani	0,30	0,42
MiniLM dotrenirani	0,28	0,70
RoBERTa dotrenirani	0,40	0,75

Rezultati su u skladu sa očekivanjima, jer *MiniLM* model ne može da ostvari iste rezultate kao *RoBERTa* koji ima daleko složeniju arhitekturu. Pored toga, modeli dotrenirani na sintetički generisanom skupu podataka o fitnesu ostvarili su znatno bolje rezultate na testnom skupu za pitanja o fitnesu, što je takođe očekivano. Ono što takođe možemo primetiti je da su dotrenirani modeli bolji i na pitanjima otvorenog domena, tj. generalizuju bolje od pretreniranih modela, na osnovu *F1* mera koje su dobijene. Takođe, obzirom da nije pronađen postojeći sistem za fitnes domen, u poređenju sa *XLNet* modelom [6] pretreniranim nad *NewsQA* i *TriviaQA* skupovima gde je ostvarena 0.72 i 0.76 F1 mera, možemo zaključiti da su dotrenirani čitači dovoljno tačni i upotrebljivi.

Dalje, data je evaluacija komponente sakupljača sa *topK* metrikom, obzirom da je to najčešće korišćena metrika za evaluaciju ove komponente. U tabeli 4.2 dati su rezultati koje su postigli modeli *DPR*, *BM25* i *TF-IDF*.

Tabela 2: *TopK* za *TF-IDF*, *BM25* i *DPR* sakupljača

Metrika / Sakupljač	top5	top10	top20
TF-IDF	0,63	0,65	0,68
BM25	0,66	0,67	0,69
DPR	0,82	0,83	0,84

Možemo zaključiti da je za ovaj sistem pogodniji *DPR* jer pored toga što daje bolji rezultat, ovaj pristup će raditi za veći korpus dokumenata. Takođe, možemo zaključiti da su u poređenju sa postojećom implementacijom *BM25* i *DPR* sakupljača [5], koji ostvaruju 0.59 i 0.79 *top20* vrednosti, implementirani sakupljači dovoljno tačni i primenjivi u realnom okruženju.

Pored tačnosti, za ovaj sistem je bilo bitno i vreme izvršavanja nad većim korpusom dokumenata. Brzina izvršavanja sistema sa kraja na kraj koji uključuje i čitač i sakupljač komponentu data je u tabeli 4.3.

Tabela 3: *Performanse sistema sa kraja na kraj*

Model \ Vreme	Prosečno vreme izvršavanja za pitanje u sekundama
MiniLM + TF-IDF	0,55
MiniLM + DPR	0,46
RoBERTa + TF-IDF	1,27
RoBERTa + DPR	1,11

Možemo zaključiti da je *MiniLM* najpogodniji u smislu brzine. Ipak, ako brzina nije najveći prioritet sistema, *RoBERTa* i *DPR* sistem je bolji izbor zbog tačnosti. U poređenju sa *Mindstone* sistemom sa kraja na kraj [8] koji je pretrenirani *BERT* i ne koristi destilaciju znanja ali je navodi kao moguće poboljšanje, možemo zaključiti da je implementirano rešenje dovoljno brzo obzirom da je vreme upita za *Mindstone* 0.73s.

4.2. Analiza grešaka

Pored evaluacije pojedinačnih komponenti i rešenja sa kraja na kraj, vršena je i analiza primera u kojima je *MiniLM* lošiji od *RoBERTa* modela. Neki od ovih primera su pitanja „*What is a chinup*?“, gde *MiniLM* daje odgovor „*bringing the chin up through space*“ ili pitanje „*How do I get stronger*?“ na koje se dobija odgovor „*stronger muscles will improve* posture“, gde se može zaključiti da MiniLM prepoznaje da je u pitanju ista reč u odgovoru, ali ne razume kontekst u potpunosti.

Takođe, slučajevi u kojima *RoBERTa* model greši su pitanja kao što su „How long should I recover from exercising?“ gde je odgovor „two to three minutes between exercises“ ili „How much carbs should I eat?“ gde je odgovor „Eating carbs and fats will make you nervous.“ što pokazuje da iako dobro razume kontekst, ima probleme u slučajevima u kojima je potreban dodatni kontekst tj. konverzacija sa korisnikom. U nastavku dat je zaključak o implementiranom sistemu.

5. ZAKLJUČAK

U ovom radu, implementirano je rešenje problema odgovora na pitanja gde se na osnovu pitanja kao ulaza parsira baza znanja, odnosno dokumenti sa tekstualnim podacima za fitnes i nutricionizam i pronalazi deo teksta u kom se nalazi odgovor na postavljeno pitanje. Implementirana je čitač sakupljač arhitektura koja je vrlo zastupljena u rešavanju problema odgovora na pitanja [5].
Za čitač komponentu isprobani su *MiniLM* i *RoBERTa* modeli, dok su za sakupljač komponentu implementirani *TF-IDF*, *BM25* i *DRP* pristupi. Da bi implementirano rešenje bilo što performantnije, pažljivo je birana sakupljač komponenta kako bi pretraga nad velikog korpusa dokumenata bila dovoljno brza.
Postojao je problem nedostatka podataka za fitnes domen, koji je rešen generisanjem sintetičkog skupa podataka.
Na kraju, pažljivim odabirom i optimizacijom pojedinih delova arhitekture ovog sistema dobijeni su istrenirani modeli koji su u stanju da odgovore na pitanja o fitnesu i nutricionizmu sa visokom tačnošću, a pritom da generalizuju bolje od pretreniranih. Kroz eksperimente i poređenje sa drugim modelima, potvrđeno je da je pristup opisan u ovom radu za rešavanje problema odgovora na pitanja primenljiv.

Moguće su dodatne optimizacije vremena izvršavanja predloženog sistema, tako što će se pažljivo odabrati čitač komponenta koja je računski najzahtevnija u ovakvom sistemu. Pažljiv odabir uređaja na kom će se sistem izvršavati može bitno uticati na brzinu.

Modeli koji su korišćeni u ovom sistemu su dizajnirani za paralelno procesiranje grafičkih kartica. Pored toga, pažljiv odabir parametara kao što su broj dokumenata kandidata koje vraća sakupljač ili broj rezultata iz čitača mogu takođe uticati na brzinu izvršavanja.

6. LITERATURA

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

[2] Lewis, P., Denoyer, L., Riedel, S. (2019). Unsupervised Question Answering by Cloze Translation. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.

[3] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, Ming Zhou. (2020). MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers.

[4] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, & Veselin Stoyanov. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach

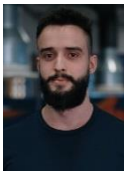
[5] Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, Wen-tau Yih. (2020). Dense Passage Retrieval for Open-Domain Question Answering.

[6] Yang, Zhilin, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. "Xlnet: Generalized autoregressive pretraining for language understanding." arXiv preprint arXiv:1906.08237 (2019).

[7] Victor Sanh, Lysandre Debut, Julien Chaumond, & Thomas Wolf. (2020). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter.

[8] Sina J. Semnani, & Manish Pandey. (2020). Revisiting the Open-Domain Question Answering Pipeline.

Kratka biografija:



Sava Katić rođen je u Novom Sadu 21. februara 1997. godine. Master rad na Fakultetu tehničkih nauka, oblast Elektrotehnika i računarstvo – Softversko inženjerstvo i informacione tehnologije odbranio je 2021. godine.

Kontakt: savakatic555@gmail.com