

Slovenian Instruction-based Corpus Generation

Gašper Spagnolo, Žiga Klun, Žiga Črv

Abstract

This project explores the utilization of Large Language Models (LLMs) for developing conversational agents proficient in the Slovenian language, with a focus on health communication. We investigate advanced LLMs, such as MixTral, evaluating their compatibility with Slovenian, computational infrastructure requirements, and model capabilities. Emphasis is placed on understanding the creation process of LLMs and constructing high-quality conversational datasets specific to the Slovenian context. Our methodology involves reviewing and categorizing datasets for training instruction-based LLMs. We devise a comprehensive data-gathering plan, identifying sources such as MedOverNet, Viva, and Vizita forums. Crawlers are developed to efficiently collect conversational data, which is then systematically organized to facilitate the fine-tuning of LLMs. Additionally, we examine relevant literature, including research on models like MixTral, to identify key considerations in data preparation. By synthesizing these insights, we prepare a corpus of conversational data tailored for fine-tuning LLMs, ensuring its relevance and quality. We aim to adapt an existing LLM using the gathered data, enhancing its ability to conduct meaningful and contextually appropriate dialogues in Slovenian on health topics. This adaptation aims to improve user interaction and practical utility, ensuring that the customized model performs better and provides more accurate responses than the original model. Our findings are consolidated in a final report, providing a comprehensive overview of the process and its implications for developing conversational agents in Slovenian using LLMs.

Keywords

Large Language Models (LLMs), Conversational Agents, Slovenian Language, Health Communication, MixTral, Data Preprocessing, Instruction-based Training, Model Fine-Tuning, User Interaction

Advisor: Slavko Žitnik

Uvod

Razvoj pogovornih agentov, ki lahko učinkovito komunicirajo v različnih jezikih, je doživel velik napredek z naprednimi tehnikami procesiranja naravnega jezika (NLP). Čeprav je bil v angleškem jeziku dosežen pomemben napredek, se pojavlja potreba po razširitvi teh tehnologij na druge jezike, da bi zadovoljili raznolike jezikovne skupnosti. Ta raziskava se osredotoča na ustvarjanje pogovornih agentov, ki obvladajo slovenski jezik, z uporabo naprednega modela MixTral.

Prilagoditev NLP modelov za učinkovito delovanje v slovenskem jeziku prinaša posebne izzive in priložnosti. Slovenščina, s svojimi značilnimi jezikovnimi lastnostmi in kulturnimi posebnostmi, zahteva skrbno prilagoditev obstoječih modelov za zagotavljanje natančnega razumevanja in generiranja dialogov. V tej študiji poudarjamo pomembnost učenja na podatkih, specifičnih za slovenski jezik in kulturo, za izboljšanje učinkovitosti teh modelov.

Naš pristop vključuje pridobivanje podatkov s pomembnih

slovenskih forumov, kot so MedOverNet, Vizita in Viva, ki so bogati viri resničnih konverzacijskih vzorcev in specifične terminologije. S tem zajemamo širok spekter kontekstov, kar zagotavlja, da so odgovori naših agentov relevantni in verodostojni. Glede na kritičnost zdravstvenih informacij smo še posebej pozorni, da podatki vključujejo vsebine, preverjene s strani medicinskih strokovnjakov. Odločili smo se za terminologijo - kategorijo zdravstvo, ker menimo, da so tukaj vprašanja in odgovori najboljše strukturirani, kar omogoča, da model najboljše naučimo za pomoč pri zdravstvenih vprašanjih.

Za pripravo podatkov za učenje modelov smo izvedli natančno predobdelavo. To je vključevalo čiščenje nezaželenih znakov, preverjanje prisotnosti vprašanj in odgovorov, zagotavljanje minimalne dolžine vsebine ter izločanje anonimnih ali slovnično nepravilnih vnosov. Ta skrbna priprava podatkov je ključna za ohranjanje kakovosti in relevantnosti, kar je bistveno za treniranje zanesljivih pogovornih agentov.

Poleg tega primerjamo našo zbirko podatkov z uveljavljen-

imi medicinskimi korpusi, kot sta MedQuAD in MIMIC-III, da poudarimo prednosti uporabe slovensko specifičnih podatkov. Te primerjave izpostavljajo pomembnost lokaliziranih podatkov za razvoj agentov, prilagojenih slovenskemu občinstvu.

Glavni cilj te raziskave je razviti model, ki ne bo le razumel slovenskega jezika, ampak bo tudi sposoben voditi smiselne in kontekstualno ustrezne dialoge o zdravstvenih temah. S prilagoditvijo modela MixTral z našo slovensko zbirko podatkov želimo izboljšati uporabniško interakcijo in praktično uporabnost. Pri tem skušamo zagotoviti, da bo model, prilagojen z našimi podatki, deloval bolje in odgovarjal natančneje kot originalni model.

Metode

V razvoju konverzacijskih agentov, ki delujejo v slovenskem jeziku, se osredotočamo na napredne pristope procesiranja naravnega jezika (NLP). Pomemben del našega raziskovalnega dela predstavlja analiza in prilagoditev sodobnega modela MixTral [1]; predstavlja temelj za razvoj naših konverzacijskih sistemov. Model bomo prilagodili za delovanje v slovenskem jeziku, s posebnim poudarkom na razumevanju in generiranju naravnega, tekočega dialoga.

Za uspešno prilagajanje modela na slovenski jezik in specifične slovenskega komunikacijskega prostora bomo izvajali "scraping" (pobiranje podatkov) s slovenskih forumov, kot so MedOverNet [2], Vizita [3] in Viva [4]. Ti forumi predstavljajo bogat vir realnih konverzacijskih vzorcev in specifičnih izrazov, ki so ključni za razumevanje in ustrezno odzivanje na povpraševanja uporabnikov. Posebno pozornost bomo namenili selekciji besedil, saj želimo zagotoviti, da bodo odgovori generirani s strani naših konverzacijskih agentov temeljili izključno na verodostojnih informacijah, podanih s strani zdravnikov ali specializantov na posameznih področjih. Tak pristop bo zagotovil večjo zanesljivost in kredibilnost odgovorov, kar je še posebej pomembno pri temah, ki so zdravstvene ali strokovne narave.

Prilagajanje modela MixTral, za specifične jezikovne in kulturne kontekste predstavlja izziv, a hkrati ponuja obetavne možnosti za razvoj visoko funkcionalnih in prilagodljivih konverzacijskih sistemov. Naš cilj je ustvariti model, ki ne bodo samo razumel slovenskega jezika, ampak bo sposoben voditi smiselne in kontekstualno relevantne dialoge, kar bo izboljšalo interakcijo med uporabniki in tehnologijo ter povečalo uporabnost konverzacijskih agentov v slovenskem jezikovnem prostoru.

Za nadaljnje izboljšave in prilagoditve naših modelov, bo naša raziskovalna akcija temeljila na zbirki orodij in virov, dostopnih v repozitoriju Brevdev Notebooks [5]. Ta bogat vir vsebuje obsežno zbirko zvezkov Jupyter, ki ponujajo raznolike metode, tehnike in primere kode, ki so ključni za razumevanje naših modelov.

Priprava učne množice

V procesu prilagajanja pogovornih modelov smo izvedli temeljito filtriranje podatkov, pridobljenih s forumov MedOverNet, Viva in Vizita. Pred filtriranjem je MedOverNet vseboval 45325 parov vprašanj in odgovorov, Vizita 29092 parov vprašanj in odgovorov, ter Viva 2632 parov vprašanj in odgovorov, skupaj pa je originalna množica vsebovala 77049 vprašanj in odgovorov. Po filtriranju MedOverNet vsebuje 22566 vprašanj in odgovorov, Vizita 24894, Viva pa 2422 parov vprašanj in odgovorov. Skupaj je torej naša originalna množica pred filtriranjem vsebovala 49882 parov vprašanj in odgovorov.

Pri množicah Viva in Vizita ima vsako vprašanje le en odgovor določenega zdravnika. Na drugi strani pa se pri MedOverNet pogosto razvije razprava z več podvprašanji in odgovori pod vsakim vprašanjem. Po temeljitih premisleku smo se odločili, da bomo uporabili le prvi odgovor na vprašanje, saj je ta najbolj strukturiran in neposredno odgovarja na zastavljeno vprašanje. Razlog za to odločitev je, da podvprašanja pogosto ne postavlja avtor originalnega vprašanja, temveč drugi uporabniki. Takšne razprave pogosto zaidejo v različne smeri in se odmaknejo od prvotne teme. Vključevanje teh podvprašanj in odgovorov bi lahko povzročilo zmedo pri treniranju modela, saj bi model moral obdelovati več vzporednih tem, kar bi lahko negativno vplivalo na njegovo natančnost in sposobnost zagotavljanja relevantnih odgovorov. Tako smo se odločili, da bomo pri filtriranju podatkov za model ohranili le tiste odgovore, ki so neposredno povezani z originalnim vprašanjem, kar bo pripomoglo k večji natančnosti in zanesljivosti našega modela.

Posamezne mejne vrednosti za filtriranje smo testirali na podatkih iz spletne strani Viva, saj smo ocenili, da so vprašanja in odgovori najboljše strukturirani. Vsi odgovori na tej strani so bili napisani s strani zdravnikov. Podobno so na spletni strani Vizita odgovarjali le zdravniki, zato tam ni bilo potrebno veliko filtracije. Pričakovano je največja filtracija prišla na množici MedOverNet, kjer odgovarjajo različni ljudje.

Filtriranje smo izvedli v več korakih zaporedno enega za drugim, s ciljem zagotoviti visoko kakovost in relevantnost podatkov za nadaljnje treniranje in prilagajanje naših pogovornih modelov za slovenski jezik. Natančna metodologija filtriranja je ključna za ohranjanje visoke kakovosti podatkovnega nabora. Spodaj so ti koraki podrobno opisani:

Čiščenje neželenih znakov

Odstranili smo neželene znake iz vprašanj in odgovorov, da bi izboljšali kakovost teksta za procesiranje. V tem koraku nismo odstranili vprašanj in odgovorov, ki so vsebovali nezaželenne znake, pač pa smo jih le nadomestili s presledkom. V množici viva je bilo takšnih popravljenih vprašanj in odgovorov 1355, pri viziti 14833, pri MedOverNet-u pa 41838.

Preverjanje obstoja vprašanj in odgovorov

Najprej smo preverili, ali podatki vključujejo tako vprašanja kot odgovore. Ugotovili smo, da množici viva in vizita ne

vsebujeta vprašanj brez odgovorov, medtem ko MedOverNet vsebuje 2972 takšnih primerov. Tej primeri so zato bili odstranjeni iz MedOverNet množice.

Preverjanje obiska strani

Izločili smo vsebine, ki niso prejele vsaj enega ogleda, saj to kaže na nizko interakcijo ali pomembnost vsebine. To smo storili le na spletni strani MedOverNet, saj strani Viva in Vizite teh podatkov nista vsebovali. Takših primerov, ki smo jih na množici MedOverNet odstranili je bilo 9239.

Dolžina vsebine

Zagotovili smo, da vprašanja vsebujejo več kot 19 znakov, odgovori pa več kot 50 znakov, kar prispeva k večji informativnosti dialogov. Določili smo namreč, da je najkrajše še sprejemljivi vprašanje recimo "Boli me nos.Kaj naj", odgovor, ki nekaj pove pa mora vsebovati vsaj 50 znakov. Po tej stopnji filtracije smo Množica Viva zmanjšali za 1, Vizito za 291 in MedOverNet za 220.

Preverjanje anonimnosti odgovora

Izključili smo odgovore anonimnih uporabnikov, saj dajemo prednost verodostojnejšim in zanesljivejšim informacijam od identificiranih strokovnjakov ali uporabnikov. Na množicah viva in vizita so vsi odgovori bili podani s strani registriranih uporabnikov, ki pa so tudi uradni zdravniki. Na množici MedOverNet je takših primerov anonimnih odgovorov bilo 2805.

Gramatična pravilnost

Opravili smo preverjanje gramatičnih napak v vprašanjih in odgovorih, da bi zagotovili jezikovno pravilnost. Besedilo brez gramatičnih napak kaže na avtorjevo sposobnost obvladovanja jezika, kar lahko prispeva k večji verodostojnosti odgovora. Gramatično pravilnost smo testirali na množici Viva in določili, da je lahko gramatičnih nepravilnosti v vprašanju in odgovoru manj od 21. Primer napak, ki jih vrača program za preverjanje gramatične pravilnosti:

```
Match({'ruleId': 'MORFOLOGIK_RULE_SL_SI',
      'message': 'Najdena morebitna napaka pri črkovanju.', 'replacements': [...], ...}),

Match({'ruleId': 'COMMA_PARENTHESIS_WHITESPACE',
      'message': 'Po vejici vstavi presledek',
      'replacements': [...], ...}),

Match({'ruleId': 'WHITESPACE_RULE',
      'message': 'Možna tipkarska napaka: ponovili ste presledek', 'replacements': [...], ...}),

Match({'ruleId': 'DA_BREZ_VEJICE',
      'message': 'V podredjih je pred \'da\' vejica: ", da"!.', 'replacements': [...], ...}),
```

Po filtraciji smo množico Viva zmanjšali za 211, Vizito za 3907, MedOverNet pa za 7523.

Preverjanje prisotnosti neprimernih besed

Za zagotovitev primerne vsebine za uporabo smo implementirali funkcijo za prepoznavanje neprimernih izrazov v vprašanjih in odgovorih, ki so bili prevedeni v angleščino. Ta funkcija je bila razvita z namenom identifikacije besedil, ki bi lahko vsebovala neprimerna ali nezaželena vsebinska vprašanja ali odgovore. V nobeni množici nismo zaznali nobene neprimerne besede.

Primerjava z drugimi korpusi

Kreirana učna množica je bila analizirana v primerjavi z korpusoma MedQuAD [6] in MIMIC-III [7].

MedQuAD (Medical Question Answering Dataset) vsebuje več kot 47.000 parov vprašanj in odgovorov, pridobljenih virov, kot so National Institutes of Health, Genetics Home Reference in National Cancer Institute. Viri zagotavljajo, da so odgovori zelo zanesljivi in natančni. Zbirka podatkov zajema širok spekter medicinskih tem, vključno z boleznimi, simptomi, zdravljenjem, genetiko in rakom. Vsak vnos v MedQuAD vključuje vprašanje, ki ga je postavil uporabnik, skupaj z odgovorom, ki ga je pripravil strokovnjak, ter metapodatke, ki podrobno opisujejo vir in kategorijo informacij. Takšna struktura naredi MedQuAD še posebej uporabno za usposabljanje zanesljivih sistemov za odgovarjanje na medicinska vprašanja, čeprav je predvsem v angleščini, kar lahko omeji njegovo neposredno uporabo za slovenske jezikovne modele brez prevajanja.

MIMIC-III (Medical Information Mart for Intensive Care III) je obsežna zbirka anonimiziranih zdravstvenih kartotek več kot 60.000 pacientov na intenzivni negi v medicinskem centru Beth Israel Deaconess. Čeprav MIMIC-III ni strukturirana kot zbirka vprašanj in odgovorov, vključuje podrobne klinične zapise, demografske podatke, vitalne znake, laboratorijske teste, zdravila in zapiske. Ta bogata zbirka podatkov zahteva znatno predhodno obdelavo za ekstrakcijo strukturiranih parov vprašanj in odgovorov, vendar ponuja obsežne podatke za ustvarjanje podrobnih in kontekstualno bogatih medicinskih informacij. Tako kot MedQuAD je tudi MIMIC-III v angleščini, njen glavni izziv pa je potreba po obdelavi in prevajanju za uporabo v slovenščini.

Naša učna množica izhaja iz spletnih zdravstvenih forumov MedOverNet, Vizita in Viva, ter vključuje pare vprašanj in odgovorov, ki odražajo resnična vprašanja splošne javnosti. MedOverNet vključuje odgovore, ki jih ustvarijo uporabniki, medtem ko Vizita in Viva zagotavljata odgovore, ki jih pripravi strokovnjaki, kar močno izboljša zanesljivost. Zbirka podatkov zajema različne teme;

- MedOverNet - bolezni srca in ožilja, cutila, drugo, gibalna/gibalni sistem, ginekologija, pediatrija, prebavila, rak, splošno, stomatologija,
- Viva: Alergije, Astma in KOPB, bolezni pljuč, Bolezni ščitnice, Bolezni sečil in ledvic, Bolezni starostnikov, Cepljenje, Črevesje in hemoroidi, Depresija, Diabetes, Diagnostika, Diete in hujšanje, Duševne bolezni, Ginekologija in porodništvo, Hematologija, HPV in rak

materničnega vratu, Kožne bolezni, Medicinska hiponoza, Motnje hranjenja, Multipla skleroza, Multipla skleroza, Nevrologija, Oči, Operacije, Ortopedija, Osebnostna rast, Osteoporoza, Otroške bolezni, Parkinsonova bolezen, Partnerski odnosi in samsko življenje, Plastična in estetska kirurgija, Počutje in lepota, Poklicne in redke bolezni, Prehlad, gripa in bakterijska vnetja, Prehranski dodatki in dopolnila, Presejalni program Zora, Proktologija, Rak, Rehabilitacija, Rekreativna, Revmatologija, Samozdravljenje, Seks, Seksualna zasvojenost, Spolne bolezni, Srce in ožilje, Težave v spolnosti, Težave z želodcem, Urologija, Vse ostalo kar vas žuli, Vzgoja otrok, Zastrupitve in prva pomoč, Zasvojenost z alkoholom, drogami in tabletami, Zdrava hrana, Zdravje zob, zobna estetika, Zelišča in zdravilne rastline, Življenje s starostnikom

- **Vizita** - forum ni razdeljen na tematske kategorije, vendar zajema široko paleto različnih tem.

Korpus vsebuje 22.566 parov vprašanj in odgovorov iz MedOverNet, 2.422 iz Viva in 24.894 iz Vizita, kar je primerljivo z velikostjo MedQuAD. Njena vsebina je prilagojena slovenskemu kontekstu, kar je ključnega pomena za razvoj slovenskih pogovornih agentov iz več razlogov:

- **Jezik in kultura:** Slovenski nabor podatkov je v slovenskem jeziku in zajema vprašanja ter odgovore, ki so prilagojeni slovenskemu jezikovnemu in kulturnemu kontekstu. Prevajanje angleških podatkov lahko povzroči izgubo pomenske nianse ali neustrezno uporabo izrazov v slovenskem okolju.
- **Boljša natančnost:** Slovenski nabor podatkov ponuja bolj natančne informacije in boljše rezultate, saj so vprašanja in odgovori že v pravilnem jezikovnem okolju. Prevajanje lahko povzroči izgubo informacij ali napake v prevodu, kar negativno vpliva na natančnost modela.
- **Lokalna relevantnost:** Slovenski nabor podatkov zajema vprašanja in odgovore, ki so specifični za slovenski zdravstveni sistem, zakonodajo, zdravstvene prakse in bolnišnice. To zagotavlja, da so informacije, pridobljene iz modela, bolj relevantne in uporabne za slovenske uporabnike.

Rezultati

Najprej smo iz začetnega nabora vprašanj in odgovorov iz slovenskih forumov MedOverNet, Vizita in Viva naključno odstranili del vprašanj za uporabo v evalvaciji modela. Preostali del podatkov smo uporabili za prilagoditev modela MixTral [1].

Za oceno kakovosti naše množice vprašanj in odgovorov smo prilagojen model testirali na testni množici in primerjali rezultate z osnovnim, neprilagojenim modelom. Rezultati

so pokazali, da osnovni model ni sposoben podati smiselnih odgovorov in zgolj ponavlja vprašanje. Po drugi strani pa je prilagojeni model v večini primerov vrnil vsebinsko smiselne in ustrezne odgovore.

Primer odgovora osnovnega modela:

- **Vprašanje:** "Zanima me, kakšne so posledice splava?"

Odgovor: "Zanima me, kakšne so posledice splava? Zanima me, kakšne so posledice splava? Zanima me, kakšne so posledice splava? Zanima me, kakšne so posledice splava? ..."

Primeri odgovorov prilagojenega modela:

- **Vprašanje:** "Zanima me, kakšne so posledice splava?"

Odgovor: "Spoštovani! Posledice splava so odvisne od tega, kako je bil splav opravljen. Če je bil splav opravljen v prvem trimesečju nosečnosti, so posledice minimalne. Če je bil splav opravljen v drugem ali tretjem trimesečju nosečnosti, pa so posledice lahko zelo hude. Lep pozdrav!"

- **Vprašanje:** "Spoštovani, stara sem 81 let. Izvid krvi na kreatinin je pokazal 400 $\mu\text{mol/L}$. Ali bi me moralo to skrbeti? Hvala."

Odgovor: "Spoštovani, vrednost kreatinina 400 $\mu\text{mol/L}$ je povišana. Priporočam pregled pri nefrologu. Lep pozdrav!"

- **Vprašanje:** "Pozdravljeni, na UZ so mi v sredini leve ledvici odkrili ledvični kamen (2-3mm). Koliko ima smisel se obremenjevati s tem? Kakšna je verjetnost, da se takšen kamen s takšno lego izloči?"

Odgovor: "Spoštovani! Če je kamen v srednji ledvični čašici, je verjetnost, da se bo izločil manjša kot če bi bil v zgornji ali spodnji ledvični čašici."

- **Vprašanje:** "Na levi nadlakti so se mi začeli delati majhni izpuščaji. Gre lahko za gljivice ali samo za suho kožo?"

Odgovor: "Pozdravljeni! Najverjetneje gre za atopični dermatitis. Priporočam, da se naročite na pregled k dermatologu, ki vam bo predpisal primerno zdravljenje. Lep pozdrav!"

Iz zgornjih primerov je razvidno, da prilagojeni model uspešno razume in odgovori na specifična vprašanja. To kaže na uporabnost naše množice podatkov pri prilagajanju modela, ki obravnava vprašanja zdravstvene tematike v slovenskem jeziku.

Diskusija

Zbirka podatkov, pridobljenih iz slovenskih forumov MedOverNet, Vizita in Viva, predstavlja bogat vir resničnih konverzijskih vzorcev in specifičnih izrazov, ki so ključni za razumevanje in ustrezno odzivanje na povpraševanja uporabnikov. Prilagoditev modela na te podatke izboljša natančnost in relevantnost odgovorov, kar nakazuje na uporabnost kreirane množice

vprašanj in odgovorov. Poleg tega pa ostaja odprto vprašanje, in sicer, zakaj osnovni model ni bil sposoben odgovoriti na nobeno vprašanje. Morda osnovni model ni bil prilagojen na slovenski jezik ali na specifične tematike in izraze, ki so pogosti na omenjenih forumih, lahko pa so bila vprašanja postavljena na način, ki je modelu nerazumljiv ali pa so vsebovala izraze, ki niso bili zajeti v njegovi trenutni bazi znanja. Vendar se na ta vprašanja nismo podrobneje osredotočali, saj je bila glavna tema osredotočena na ustvarjanje množice vprašanj in odgovorov, kar nam je uspelo doseči.

References

- [1] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Deven-dra Singh Chait, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.
- [2] Medovernet - svetovanje. <https://med.over.net/>, 2024. Dostopano: 17.3.2024.
- [3] Vizita.si - za vaše zdravje. <https://vizita.si/>, 2024. Dostopano: 17.3.2024.
- [4] Viva - portal za zdravo življenje. <https://viva.bhc.si/>, 2024. Dostopano: 17.3.2024.
- [5] Brevdev notebooks. <https://github.com/brevdev/notebooks>, 2024. Dostopano: 17.3.2024.
- [6] Asma Ben Abacha and Dina Demner-Fushman. A question-entailment approach to question answering. *BMC Bioinform.*, 20(1):511:1–511:23, 2019.
- [7] Alistair E W Johnson, David J Stone, Leo A Celi, and Tom J Pollard. The mimic code repository: enabling reproducibility in critical care research. *Journal of the American Medical Informatics Association*, 25(1):32–39, 2018.