Zbornik radova Fakulteta tehničkih nauka, Novi Sad

UDK: 004.9
DOI: https://doi.org/10.24867/27BE25Maruna

CHATBOT У ОБЛАСТИ МЕДИЦИНЕ БАЗИРАН НА ЕНКОДЕР-ДЕКОДЕР АРХИТЕКТУРИ

MEDICAL CHATBOT BASED ON ENCODER-DECODER ARCHITECTURE

Теодора Маруна, *Факултет техничких наука, Нови Сад*

Област – Електротехника и рачунарство

Кратак садржај – *Циљ рада јесте креирање конверзацијског chatbot-а који користи напредне алгоритме машинског учења и технике процесирања природног језика за давање дијагнозе или пружање препоруке за лечење, на основу симптома које пацијент наведе. У оквиру рада изнети су експериментални резултати који показују да су перформансе боље у случају када се chatbot обучава са скупом података у којем постоји алтернативно парафразирање питања и где за слична питања добија генерализоване, опште одговоре, а не одговоре који су врло уско специјализовани и персонализовани за наведено питање пацијента.*

Кључне речи: *chatbot, енкодер-декодер архитектура, механизам пажње, обрада природног језика*

Abstract – *The goal of this thesis is to create a conversational chatbot that utilizes advanced machine learning algorithms and NLP techniques to predict diagnosis or provide treatment recommendations based on the patient's symptoms. This paper presents experimental results that show that performances are better when the chatbot is trained with a data set in which there is an alternative paraphrasing of questions and where, for similar questions, it receives generalized answers rather than answers that are very narrowly specialized and personalized for the patient's stated question.*

Keywords: *chatbot, encoder-decoder architecture, attention mechanism, Natural Language Processing*

1. УВОД

Опште здравствено стање сваке особе је један од главних фактора квалитетног живота. Како би лекар пружио неопходну помоћ пацијентима, он мора да издвоји одређено време, а време сваког лекара је ограничено и драгоцено. Међутим, шта се дешава када лекари, услед великог броја пацијената, немају довољно времена да пруже адекватан савет сваком појединцу. То доводи до идеје креирања chatbot-а који ће користити напредне алгоритме вештачке интелигенције да постави основну дијагнозу пацијента и да му пружи препоруке пре консултација са лекаром. С обзиром да су рекурентни енкодер-декодер модели доминантни у области моделовања конверзацијских дијалог система, за имплементацију chatbot-а у овом раду, користи се модел секвенци (енг. Seq2Seq) [1], који поседује енкодер-декодер архитектуру и има имплементиран механизам пажње (енг. attention mechanism) између енкодер и декодер слоја.

Главни задатак овог рада је показати каква је разлика у резултатима када се користе другачије структурирани скупови података за обучавање модела и како ти подаци утичу на резултате. Како би се показала разлика до које доводе различито структурирани улазни подаци, у оквиру рада вршено је обучавање и упоређивање два модела који поседују исту архитектуру али су обучени на различитим скуповима података.

У оквиру рада изнете су вредности БЛЕУ (енг. BLEU – Bilingual Evaluation Understudy) [2] метрике које показују да су перформансе система боље у случају кад се chatbot обучава са скупом података у којем постоји алтернативно парафразирање питања и где за слична или иста питања добија генерализоване, опште одговоре, а не одговоре који су врло уско специјализовани и персонализовани за наведено питање пацијента.

2. ТЕОРИЈСКЕ ОСНОВЕ И ДЕФИНИЦИЈЕ

У овом сегменту је дато објашњење најбитнијих теоријских појмова и алгоритама који се користе за израду chatbot-а.

2.1 ЛСТМ неуронска мрежа

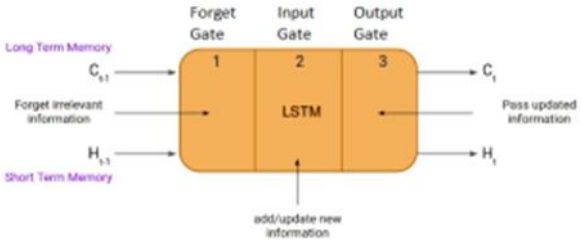
ЛСТМ (енг. LSTM – Long Short-Term Memory) [3] је суштински побољшана верзија класичне рекурентне неуронске мреже и способна је за тумачење дужих низова података. Најбитнија предност ЛСТМ је могућност дугорочног памћења секвенци.

ЛСТМ решава проблем нестајућег градијента коришћењем неколико капија које контролишу проток информација које пролазе кроз меморијску ћелију. ЛСТМ се састоји из интерног стања ћелије (енг. cell state), које представља дугорочну меморију, скривеног стања (енг. hidden state) које репрезентује краткотрајну меморију и 3 капије. Улазна капија утиче на одабир информација које се додају у стање ћелије. Капија за заборављање информација контролише које информације треба да се задрже у стању ћелије, а које треба да се забораве. Излазна капија контролише које

НАПОМЕНА:

Овај рад проистекао је из мастер рада чији ментор је био др Александар Ковачевић, ред. проф.

информације треба проследити на излаз стања ћелије. Свака капија представља сигмоидалну функцију која враћа вредности између 0 и 1, указујући на то колико информација треба пропустити кроз капију. Описана архитектура ЛСТМ мреже приказана је на слици 1.



Слика 1. *Архитектура ЛСТМ неуронске мреже*

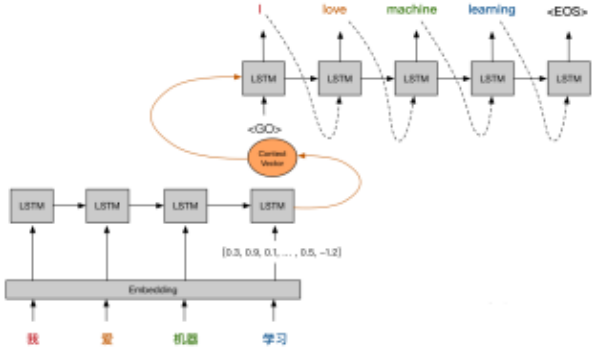
2.2 Модел секвенци

Ово је модел чија се архитектура састоји из две компоненте – енкодера и декодера. Енкодер и декодер су представљени као нека врсте рекурентне неуронске мреже, најчешће ЛСТМ. Улазна секвенца енкодера и декодера се провлачи кроз embedding слој како би се смањила димензионалност улазних вектора.

Енкодер узима читаву реченицу и процесира је реч по реч, односно секвенцу по секвенцу. При обради сваке појединачне секвенце, енкодер поседује скривено стање и стање ћелије и генерише излазну вредност. Излазна секвенца енкодера се занемарује у овом облику енкодер-декодер архитектуре. Свако скривено стање тренутне секвенце утиче на скривено стање следеће секвенце, а крајње скривено стање се посматра као резиме целе реченице. Ово последње стање се зове контекст вектор (енг. context vector) и он енкапсулира информације улазних података и представља смисао целе улазне реченице.

Декодер као улаз прима контекст вектор генерисан од стране енкодера и користи га за генерисање излазне секвенце. Користећи овај вектор, декодер врши предикцију крајње излазне секвенце, реч по реч, где се сваки претходно предиктовани токен (из временског тренутка t-1) узима у обзир за генерисање наредног токена секвенце.

На основу разлике између предиктованог и правог (очекиваног) токена, рачуна се функција грешке и користи се алгоритам пропагације градијента уназад кроз време да би се ажурирали параметри модела. За креирање крајње излазне секвенце користи се softmax активациона функција. На слици 2. илустрована је енкодер-декодер архитекура, односно модел секвенци.



Слика 2. *Енкодер-декодер архитектура*

2.3 Механизам пажње у моделу секвенци

Енкодер-декодер архитектура функционише добро када су у питању краће реченице, односно краћи низови секвенци. У случају дужих реченица, може доћи до проблема при покушају екстракције контекста из секвенце. Зато се користи механизам пажње, који омогућава декодеру да приликом предикције излазне секвенце селективно обраћа пажњу на улазне секвенце (које су заправо излаз енкодера).

Модел се тренира тако да учи да обраћа селективну пажњу на излазе енкодера (који се у класичној енкодер-декодер архитектури занемарују) и да увиђа повезаности између њих и излазних секвенци. Ова врста механизма пажње се зове адитивни или Bahdanau механизам пажње и он добро функционише са дугачким секвенцама речи. Врши линеарну комбинацију стања енкодера и декодера.

3. МЕТОДОЛОГИЈА

Систем се може поделити на два главна модула, модул задужен за претпроцесирање улазних података и модул задужен за генерисање одговора.

3.1 Модул за претпроцесирање података

У оквиру овог модула извршава се претпроцесирање улазне секвенце са циљем припреме и пречишћавања података који улазе у модул за генерисање одговора. Циљ овог модула је повећање квалитета улазних података.

Улаз у овај модул чине неколико скупова података, који представљају конверзације између пацијената и лекара.

Најпре се врши пречишћавање података тако да су прво уклоњени непотребни карактери, а затим су све речи пребачене на мала слова и уклоњени су знакови интерпункције.

Лематизација је следећи корак претпроцесирања и она се врши само над постављеним питањима пацијената, да се не би изгубило семантичко значење одговора и да би генерисан одговор био у формату природног језика, а не лематизован.

После се врши уклањање заустављних речи, да би се chatbot фокусирао на битне речи и њихов контекст.

Следећи корак је креирање речника. Услед великог обима речника, речник је сведен на речи које се понављају 3 или више пута.

Након тога, врши се токенизација у оквиру које долази до замене свих речи у оквиру питања пацијената и одговора лекара индексима тих речи из речника. Токенизација је неопходан корак јер улаз у неуронску мрежу мора да буде вектор (нумеричка вредност), а не низ карактера.

Поред тога, уведени су додатни токени. Токен <OUT> означава реч која се не налази у речнику. Токен <SOS> представља почетак реченице, док <EOS> токен означава крај реченице.

Текст који улази у декодер (одговора лекара) означен је токенима <SOS> и <EOS>. Токен <PAD> користи се за додавање токена секвенци тако да она буде жељене дужине.

Свим секвенцама које улазе у енкодер додато је онолико специјалних токена (<PAD>) тако да укупна дужина сваке секвенце буде једнака дужини најдуже секвенце која улази у енкодер. Аналогно је урађено и за секвенце које улазе у декодер слој.

За репрезентовање текста у векторском облику коришћен је GloVe embedding. Embedding је вектор који представља значење и контекст улазне речи. Резултат овог корака је скуп вектора који репрезентују улазну секвенцу.

После тога, долази до поделе скупа података на тренинг и тест скуп података, где је величина тест скупа 15%. Подела на тренинг и тест скуп је неопходна јер је потребно одвојити податке над којима се врши обучавање модела, од података над којим се врши тестирање модела.

3.2 Модул за генерисање одговора

Овај модул се састоји из модела секвенци, који је задужен за постављање дијагнозе пацијента на основу унетих симптома. Улаз у овај модул представља векторска репрезентација улазног текста описаних симптома корисника, а излаз је одговор, односно предикована дијагноза.

Модел секвенци поседује енкодер-декодер архитектуру, где оба слоја представљају неку врсту рекурентне неуронске мреже. Између енкодер и декодер слојева, имплементиран је механизам пажње, како би се на што ефикаснији начин увидео контекст улазне реченице.

Улаз у енкодер је секвенца коју желимо да енкодујемо, а излаз је енкодовано стање. Енкодер је задужен за трансформисање улазне секвенце у векторску репрезентацију фиксне дужине. Улаз у декодер представља енкодовано стање а излаз је декодирана секвенца.

При имплементацији енкодера, коришћен је бидирекциони ЛСТМ. Енкодер генерише крајње стање – контекст вектор, он представља резиме целе реченице и он је улаз у декодер слој. Архитектура декодер слоја се састоји из ЛСТМ мреже.

У последњем слоју неуронске мреже, налази се Dense слој са softmax активационом функцијом, који се користи за генерисање крајњег излаза из мреже.

4. СКУПОВИ ПОДАТАКА

Скупови података емулирају структуру реалне конверзације између лекара и пацијента, где пацијент најпре поставља питање, а лекар затим даје одговор. Над скуповима података извршене су одговарајуће технике претпроцесирања, како би се пречистили подаци и уклонили редудантни делови.

4.1 Скуп података MedDialog-EN

MedDialog-EN [4] је скуп података са дијалозима, где се сваки дијалог се састоји од дела у коме пацијент објашњава своје здравствено стање и дела где лекар даје одговор. Подаци у оквиру овог скупа података су преузети са два сајта – iCliniq и HealthcareMagic, који представљају онлајн платформе за давање стручних медицинских савета.

С обзиром да су конверзације пацијената и лекара преузете са разних онлајн форума, код овог скупа података за слична или иста питања постоје различити одговори, јер је сваки одговор лекара специјално намењен проблему који је пацијент навео. Одговори су уско специјализовани за наведено питање пацијента и његове индивидуалне проблеме и не представљају уопштено решење за дату тематику питања.

4.2 Скуп података Intents

Intents [5] скуп података је оригинално био такав да се сваки пар питања-одговор састоји из 4-6 сличних питања и једног одговора лекара који одговара за сва наведена питања. Скуп података је обрађен тако да се структура подудара са структуром реалне конверзације, где се свака конверзација састоји из једног питања пацијента и адекватног одговора лекара.

Постоје две предности Intents скупа података у односу на MedDialog-EN. С обзиром да постоји алтернативно парафразирање питања, лакше је уочити намену постављеног питања, а самим тим је лакше генерисати адекватан одговор на постављено питање.

Поред тога, Intents скуп података је такав да за слична питања даје увек потпуно исти одговор. У Intents скупу података, одговори су генерализовани, уопштени, недвосмислени и тичу се било каквог проблема из исте области. Ипак, мана Intents скупа података јесте мали обим података.

5. РЕЗУЛТАТИ И ДИСКУСИЈА

У овом поглављу представљени су резултати експеримената, поређењем конфигурација модела.

5.1 Резултати chatbot-а обученог над MedDialog-EN

У табели 1. налазе се резултати конфигурација chatbot-а обученог над MedDialog-EN скупом података. Mean bleu 1-gram представља просечну БЛЕУ вредност за унираме (појединачне речи). Loss представља вредност функције циља, а accuracy представља добијену тачност. У табелама се још налазе вредности validation loss и validation accuracy и оне представљају прецизност валидационог скупа података у последњој епохи и функцију циља валидационог скупа података у последњој епохи. На основу табеле, може се закључити да најбоље резултате за БЛЕУ вредност, функцију циља и прецизност има конфигурација 3.

Табела 1. *Конфигурације chatbot-а, обученог над MedDialog-EN скупом података*

	конфиг.1	**конфиг.2**	**конфиг.3**
број епоха	10	70	70
batch size	16	16	32
bleu 1-gram	0.124924	0.140363	0.141532
loss	0.9598	0.0496	0.0228
accuracy	0.7830	0.9850	0.9936
validation loss	1.9509	3.6006	3.3374
validation acc.	0.6867	0.6699	0.6702

5.2 Резултати chatbot-а обученог над Intents

У табели 2. налазе се резултати конфигурација chatbot-а обученог над Intents скупом података. Сва поља описана у претходном поглављу налазе се и у табели 2. На основу табеле 2, може се закључити да најбоље резултате за БЛЕУ вредност, функцију циља и прецизност chatbot-а обученог над Intents скупом података има конфигурација 1.

Табела 2. *Конфигурације chatbot-а, обученог над Intents скупом података*

	конфиг.1	**конфиг.2**	**конфиг.3**
број епоха	70	70	70
batch size	16	24	32
bleu 1-gram	0.688934	0.680912	0.676773
loss	0.0129	0.0321	0.0533
accuracy	0.9995	0.9987	0.9977
validation loss	0.0613	0.0916	0.1170
validation acc.	0.9830	0.9822	0.9801

5.3 Дискусија

Посматрајући БЛЕУ вредности из претходно споменутих табела, увиђа се знатна разлика у перформансама chatbot-а обученог над MedDialog-EN скупом података и chatbot-а обученог над Intents скупом података. Просечна БЛЕУ вредност 1-грама најбоље конфигурације chatbot-а обученог над MedDialog-EN скупом података износи само 0.141532, док најбоља просечна вредност 1-грама за Intents скуп података износи чак 0.688934. Разлог за нижу вредност БЛЕУ метрике chatbot-а обученог над MedDialog-EN скупом података је то што су питања пацијената али и одговори лекара врло индивидуализовани и јако специјализовани за наведени проблем пацијента. У MedDialog-EN скупу података, лекар у свом одговору често показује емпатију према пацијенту, пружа наду и утеху, али због тога, врло је захтевно извршити сумаризацију у којој остаје суштина одговора. Поред тога, постоје парови питање-одговор који поседују иста питања али другачије одговоре. Неуронској мрежи се на тај начин, за иста питања, доводе различити одговори, а то доприноси некозистентности при обучавању.

С друге стране, Intents скуп података поседује питања и одговоре који су генерализовани. Алтернативно парафразирање питања омогућава да се за слична питања, на мрежу доводи исти одговор. Због тога, конфигурације chatbot-а које за обучавање користе Intents скуп података, дају знатно боље резултате.

6. ЗАКЉУЧАК

Модерни chatbot-ови, специјализовани у области медицине имају потенцијал за побољшање комуникације између доктора и пацијената. Chatbot може да спроводи брзо и једноставно анкете, комуницира користећи медицинске термине, заказује прегледе и контроле код специјалиста, прикупља и анализира податке пацијената и спрам тога даје дијагнозе или потенцијалне препоруке за лечење.

У овом раду је предложен chatbot, имплементиран помоћу модела секвенци са механизмом пажње, развијен са циљем пружања помоћи пацијентима пре посете лекара. Креирани chatbot би имао велики допринос у здравственом систему, јер би био доступан 24 сата дневно, 7 дана недељно, што би пацијентима омогућило брз али пре свега сигуран извор информација и савета у вези са њиховим здравственим проблемима. Постојање chatbot-а може побољшати приступ здравственој нези, смањити оптерећеност медицинског особља и омогућити брзу и ефикаснију комуникацију са пацијентима, посебно у хитним ситуацијама. Ипак, овакав систем нема као циљ замену медицинског особља и стручних дијагноза, него представља алат за повећање њихове ефикасности. За озбиљне медицинске дијагнозе потребно је консултовати стручњака из одређене области, зависно од врсте проблема.

У оквиру рада изнете су вредности БЛЕУ метрике које показују да су перформансе боље у случају када се chatbot обучава са скупом података у којем постоји алтернативно парафразирање питања и где за слична питања добија генерализоване и опште одговоре, а не одговоре који су врло уско специјализовани и персонализовани за наведено питање пацијента.

Међутим, главни проблем при имплементацији овог система је недостатак адекватног скупа података. Како би chatbot могао да даје тачне дијагнозе и препоруке за лечење, потребно је обучити систем на још већем скупу података. Ипак, најбитнији је квалитет и природа скупа података. Питања и одговори треба да буду што концизнији, прецизнији и да су фокусирани на суштину проблема. Постојање таквог скупа података, који је адекватног обима и чији подаци задовољавају очекиване критеријуме, додатно би поправило перформансе система.

Један правац будућег развоја је могућност памћења конверзације и генерисање одговора на основу целокупне посматране историје разговора. Памћење конверзације омогућава квалитетнију дијагнозу јер се сагледа целокупна слика проблема пацијента.

7. ЛИТЕРАТУРА

[1] Sutskever, O. Vinyals, and Q. V. Le, “Sequence to Sequence Learning with Neural Networks”.

[2] P. F. Brown et al., “A STATISTICAL APPROACH TO MACHINE TRANSLATION,” vol. 16, no. 2, 1990.

[3] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” Neural Comput., vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735

[4] X. He et al., “MedDialog: Two Large-scale Medical Dialogue Datasets.” arXiv, Jul. 07, 2020. doi: http://arxiv.org/abs/2004.03329

[5] https://www.kaggle.com/datasets/tusharkhete/dataset-for-medicalrelatedchatbots

Кратка биографија:

Теодора Маруна рођена је у Новом Саду 1999. године. Мастер рад на Факултету техничких наука из области Електротехнике и рачунарства одбранила је 2023. године.
контакт: teodora.maruna@gmail.com