МРНТИ 20.23.17; 20.23.21
УДК 004.421, 004.912

https://doi.org/10.51889/2021-4.1728-7901.16

Д.Р. Рахимова [1,2], Д.Т. Қасымова [1,4], Д.Н. Исабаева [3]*
[1]Ақпараттық және есептеуіш технологиялар институты, Алматы қ., Қазақстан
[2]әл - Фараби атындағы Қазақ Ұлттық Университеті, Алматы қ., Қазақстан
[3]Абай атындағы Қазақ ұлттық педагогикалық университеті, Алматы қ., Қазақстан
[4]Логистика және көлік академиясы, Алматы қ., Қазақстан
**e-mail: dikakassymova@gmail.com*

ҚАЗАҚ ТІЛІНЕ АРНАЛҒАН BERT МОДЕЛІ НЕГІЗІНДЕ СҰРАҚ-ЖАУАП ЖҮЙЕСІН ЗЕРТТЕУ ЖӘНЕ ӘЗІРЛЕУ

Аңдатпа

Берілген мақалада қазақ тіліне арналған BERT моделі негізінде сұрақ-жауап жүйесін зерттеу және әзірлеу қарастырылған. Мақалада қазақ тілі сияқты агглютинативті тілге арналған жабық домендік сұрақтарға жауап беру жүйесінде қолданылатын сұрақ талдауға ережелік және статистикалық тәсілдердің жаңа үйлесімі ұсынылған. Сұрақтарды талдау фокусты бөлу және сұрақтарды жіктеуден тұрады. Фокусты алу үшін бізде қазақ тілінде жиі кездесетін сұрақтарға арналған ережелерге негізделген бірнеше сарапшылар бар. BERT моделі табиғи тілді өңдеуге арналған машиналық оқытуды қолданады. Ол жылдам дәл баптауға мүмкіндік береді және көптеген практикалық қосымшалар мен төменгі ағындар үшін қолдануға болады. Әр классқа сәйкес емес сөз тіркестерінмен айқындалатын сұрақтарды жіктеу үшін ережелерге негізделген классификаторды қолданады. Екі мәселені де анықтауда базалық модельдер қолданылды және салыстыру нәтижелері қарастырылды. Ұсынылған әдістемеден басқа репродуктивтілікке және кейінгі зерттеулерге арналған қолмен жазылған сұрақтар жиынтығы ұсынылды.

Түйін сөздер: қазақ тілі, сұрақ-жауап жүйесі, табиғи тілді өңдеу, BERT моделі.

Аннотация

Д.Р. Рахимова [1,2], Д.Т. Касымова [1,4], Д.Н. Исабаева [3]
[1]Институт информационных и вычислительных технологий, г. Алматы, Казахстан
[2]Казахский национальный университет имени аль-Фараби, г. Алматы, Казахстан
[3]Казахский национальный педагогический университет имени Абая, г. Алматы, Казахстан
[4]Академия логистики и транспорта, г. Алматы, Казахстан

ИССЛЕДОВАНИЕ И РАЗРАБОТКА СИСТЕМЫ ВОПРОС - ОТВЕТ НА ОСНОВЕ МОДЕЛИ BERT ДЛЯ КАЗАХСКОГО ЯЗЫКА

В данной статье представлено исследование и разработка системы вопросов-ответов на основе модели BERT для казахского языка. В статье представлена новая комбинация нормативного и статистического подходов к анализу вопросов, используемых в системе ответов на вопросы закрытой предметной области для агглютинативных языков, таких как казахский. Анализ вопроса состоит из фокусирующих и классифицирующих вопросов. Чтобы сфокусироваться, у нас есть несколько экспертов, основанных на правилах часто задаваемых вопросов на казахском языке. BERT, несомненно, является полезной моделью в использовании машинного обучения для обработки естественного языка. Он обеспечивает быструю и точную настройку и может использоваться во многих практических и последующих приложениях. В статье для классификации вопросов использовался классификатор на основе правил, в котором используются фразы, не подходящие для каждого класса. Мы использовали базовые модели для обеих задач и представили результаты сравнения. В дополнение к предложенной методике был представлен набор рукописных вопросов для воспроизведения и последующего исследования.

Ключевые слова: казахский язык, вопросно-ответная система, обработка естественного языка, модель BERT.

Abstract

RESEARCH AND DEVELOPMENT OF A QUESTION AND ANSWER SYSTEM BASED ON THE BERT MODEL FOR THE KAZAKH LANGUAGE

Rakhimova D. [1,2], Kassymova D. [1,4], Isabaeva D.[3]

[1]Institute of Information and Computational Technologies, Almaty, Kazakhstan

[2]Al-Farabi Kazakh National University, Almaty, Kazakhstan

[3]Abai Kazakh National Pedagogical University, Almaty, Kazakhstan

[4]Academy of Logistics and Transport, Almaty, Kazakhstan

This article presents the research and development of a question-and-answer system based on the BERT model for the Kazakh language. The article presents a new combination of normative and statistical approaches to the analysis of questions used in the system of answering questions of a closed subject area for agglutinative languages such as Kazakh. Question analysis consists of focusing and classifying questions. To focus, we have several experts based on the rules of the Kazakh language FAQ. BERT is undoubtedly a useful model in using machine learning for natural language processing. It provides fast and accurate set-up and can be used in many practical and subsequent applications. The article used a rule-based classifier to classify the questions, which uses phrases that are not appropriate for each class. We used basic models for both problems and presented the results of the comparison. In addition to the proposed methodology, a set of handwritten questions for reproduction and subsequent research was presented.

Keywords: Kazakh language, question-answer system, natural language processing, BERT model.

Кіріспе

Соңғы уақытта ақпараттық технологиялар әлемінде көптеген трендтер пайда болды. Ең танымал трендтердің бірі болып сұрақ-жауап (question-and-answer QA) жүйелері, дауыстық көмекшілер және «ақылды көмекшілердің» барлық түрлері. Әдетте, мұндай жобаларды жүзеге асыруда ортақ нәрсе бар, атап айтқанда, пайдаланушыдан келетін ақпаратты түсіну және оны талдау мүмкіндігі. Әлбетте, мұндай көмекшілер жаппай өнім болғандықтан, олар пайдаланушының ақпаратын соңына дейін ең қарапайым түрде өңдеуі керек. Пайдаланушы үшін мұндай әдістер - ол ана тілінде сөйлейтін табиғи тілде дауыстық және мәтіндік командаларды беру. Бұл күрделі зерттеу және инженерлік тапсырма болғандықтан нарықта табысты болу үшін жүйе хабарларды табиғи тілде өңдей алуы керек екен. Мұндай жүйелердің ең жарқын мысалдары Apple компаниясының Siri, Microsoft корпорациясының Cortana, Google now, ask.com, IBM Watson және тағы басқа жүйелер сәтті енгізілген және көптеген әлем тілдері үшін қолданылады. Өкінішке орай, табиғи тілдердің ерекшеліктеріне байланысты, дәлірек айтқанда, олардағы күрделілігіне байланысты, табиғи тілдегі хабарларды машина түсіндіретін логикалық өрнектерге біршама жоғары дәлдікпен түрлендіретін ашық пәндік аймақ (open-domain system) жүйесін құру міндеті өте күрделі. Сондықтан олар белгілі бір пәндік салада (домендік жабық жүйе) ғана білім базасына негізделген жүйелерді құрастырады. Бұл есептеулердің үлкен көлемінің көптеген себептерінің бірі: жүйелерді түрлендіру, оқыту және тестілеу әртүрлі пәндік салалар үшін бірнеше рет жүргізілуі керек. Өкінішке орай, бүгінгі таңда қазақ тіліне арналған ашық әрі сапалы сұрақ-жауап жүйесі жоқ. Бұл тілдің лингвистикалық қасиеті мен қордың аздығына байланысты.

Зерттеу тақырыбына қысқаша шолу

Қазіргі уақытта машинаның жеке хабарламалары туралы кейбір түсініктерге қол жеткізудің көптеген жолдары бар. Бұл әдістерді шамамен үш топқа бөлуге болады:

- ақпараттық іздеу технологиясына негізделген тәсілдер (IR based approach);
- табиғи тілдер мен білім қорларын өңдеуге негізделген тәсілдер (knowledge based approach);
- алдыңғы екі әдісті біріктіретін тәсілдер (және басқалары, мысалы, deep machine learning).

Осындай алғашқы жүйелердің бірі 1961 жылы құрылған Бейсбол жүйесі [1] және оның зерттеу пәні бейсбол болды. Жүйе өте аз қарапайым сұрақтарға жауап бере алады және сөздік қоры өте аз болды. Ол сұрақтардың үлгіні сәйкестендіру (үлгілерді сәйкестендіру) тәсіліне негізделген. 1966 жылы жазылған ELIZE жүйесі [2] сұрақтардың минималды жіктелуін жүзеге асыра алады. Жүйе психотерапевтке ұқсайтын диалогтық робот болып табылады. Симуляциялау қорытындысы мынада: ELIZE маркер_- сөздерді таңдайды және үлгіні пайдаланып нақтылайтын сұрақтарды түрлендіреді. LUNAR жүйесі [3] 1971 жылы NASA ғарыш бағдарламасының нәтижесінде жасалды. Оның мақсаты ғарышта жиналған нысандар туралы қарапайым сұрақтарға жауап беру болды: мысалы, «Жиындықта неше дөңгелек тас бар?». Оның сұрақтарға жауаптарының дәлдігі 78% дейін жетті. Жүйе бірнеше компоненттерден тұрады: жалпы қолданыстағы өтулердің кеңейтілген желісі (general-purpose

augmented transition network (ANT) [4]), синтаксистік ақпаратты семантикалық ұсыну шеңбері, табиғи тілді логикалық түрде көрсету шеңбері, 4500 сөзден тұратын сөздік және білім базасындағы 13000 субъект.

Басқа қызықты мысалдар START [5] – вебті пайдаланатын ашық домені бар бірінші IR негізіндегі жүйе, сондай-ақ SHRDLU [6], Murax [7], Юпитер [8]. Кейбір заманауи жүйелер (мысалы, [9,10,11]) статистикалық тәсілге негізделген, оның мәні SVM алгоритмдерінің модификацияларын, аңғал Байес классификаторын, максималды энтропия принципін және т.б. [12,13] еңбектерінде үлгімен салыстыру қолданылады. Бұл салыстырудың мәні пайдаланушының хабарламаларында кейбір шаблондарды (үлгілерді) табу және осы шаблонға сәйкес келетін веб-жауаптарды табу болып табылады. Қазіргі уақытта ең жақсы нәтижелер бірнеше техниканы біріктіретін тәсілді көрсетеді (кейбіреулері, мысалы, терең машинаны үйрену). Қазақ тілі үшін жауап-сұрақ жүйесін зерттеу және әзірлеу үшін машиналық оқытуға негізделген тәсіл қолданылатын болады.

Қазақ тіліне арналған сұрақ-жауап жүйесінің сипаттамасы

Сұрақ-жауап жүйесі – табиғи тіл интерфейсін пайдаланатын іздеу, анықтамалық және интеллектуалды жүйелердің гибридті ақпараттық жүйе. Табиғи тілде тұжырымдалған сұрау осындай жүйенің енгізуіне жіберіледі, содан кейін ол NLP әдістері арқылы өңделеді және табиғи тілдегі жауап жасалады. Сұраққа жауап табу мәселесіне негізгі тәсіл ретінде әдетте келесі схема қолданылады: біріншіден, жүйе қандай да бір жолмен (мысалы, кілт сөздер бойынша іздеу арқылы) сұраққа қатысты ақпаратты қамтитын құжаттарды таңдайды, содан кейін оларды сүзеді, ықтимал жауабы бар жеке мәтін фрагменттерін бөлектейді, содан кейін генерациялаушы модуль таңдалған фрагменттерден сұраққа жауапты синтездейді. Жергілікті сақтау қоймасы QA жүйесін әзірлеуде ақпарат көзі ретінде пайдаланылады [14].

Қазіргі заманғы QA жүйелері жалпы (open-domain) және арнайы (closed-domain) болып бөлінеді. Жалпы жүйелер, яғни ерікті сұрақтарды өңдеуге бағытталған жүйелер біршама күрделі архитектураға ие, бірақ соған қарамастан іс жүзінде олар өте әлсіз нәтижелер береді және жауаптардың дәлдігі төмен. Бірақ, әдетте, мұндай жүйелер үшін жауаптардың дәлдігінен гөрі білімді қамту дәрежесі маңызды. Белгілі бір пәндік салаға қатысты сұрақтарға жауап беретін арнайы жүйелерде, керісінше, жауаптардың дәлдігі жиі маңызды көрсеткіш болып табылады.

Әзірленген QA жүйесі келесі модульдерден тұрады:

Сұрақтарды өңдеу. Бірдей ақпаратты әртүрлі жолдармен сұрауға болады. Сөйлемнің семантикасын (мағынасын) түсіну мен өңдеудің тиімді әдістерін жасау талап етіледі. Стильге, сөздерге, синтаксистік қатынастарға және қолданылған идиомаларға қарамастан, бағдарлама мағынасы жағынан баламалы сұрақтарды тануы маңызды. QA жүйесі күрделі сұрақтарды бірнеше қарапайым сұрақтарға бөліп, контекстке тәуелді фразаларды дұрыс түсіндіріп, диалог кезінде пайдаланушымен нақтылау қажет.

Контекстік сұрақтар. Сұрақтар белгілі бір контексте қойылады. Мәтінмән сұрауды нақтылай алады, екіұштылықты жояды немесе бірқатар сұрақтар арқылы пайдаланушының ой тізбегін бақылайды.

Жауаптарды ерекшелеу. Бұл процедураның дұрыс орындалуы сұрақтың күрделілігіне, оның түріне, контекстіне, қолжетімді мәтіндердің сапасына, іздеу әдісіне және т.б. факторларға байланысты.

QA жүйесі үшін білім көздері. Мәтінді өңдеудің қандай әдістері қолданылса да, егер ол мәліметтер қорында болмаса, дұрыс жауапты таба алмаймыз. Осыған байланысты білім мен мәліметтер қорын толықтыру қажет.

Жауап тұжырымы. Жауап мүмкіндігінше табиғи болуы керек. Кейбір жағдайларда оны мәтіннен бөлектеу жеткілікті. Бірақ кейде күрделі сұраулармен айналысуға тура келеді және мұнда әртүрлі құжаттардан жауаптарды біріктіру үшін арнайы алгоритмдер қажет [15].

Сұрақ жауап жүйесін әзірлегенде BERT моделін қолдану

Трансформердің кодтау блоктарын BERT моделінде қолдануға болады. Мәтіндегі сөздердің (немесе қосалқы сөздер) арасындағы контексттік қатынастарды анықтайтын зейінді (Multi-Headed Self Attention) механизмін трансформер кодтаушысы қолданады. BERT моделі бір бағытты шектеуді «жасырылған моделін» (MLM) оқытуға дейінгі мақсатты қолдана отырып жеңілдетеді. MLM моделі

кейбір таңбалауыштарды кездейсоқ түрде кірістен бүркемелейді, ал оның мақсаты қоршаған ортаға байланысты (сөздің сол және оң жағы) жасырылған сөзді болжау болып табылады.

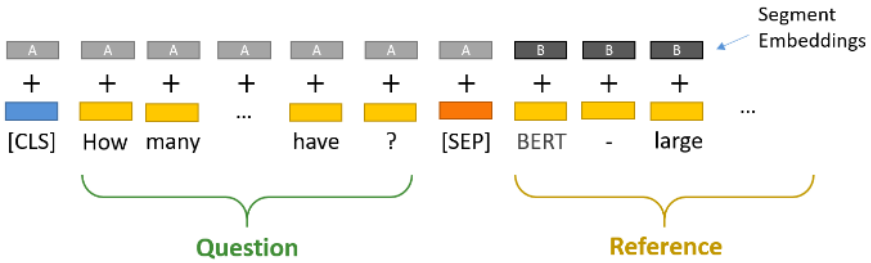
Енгізілген мәтінді «солдан оңға» немесе «оңнан солға» түрде оқитын бағытталған моделдердің түрлеріне қарағанда, MLM мақсаты өкілдікке солға да, оңға да контексті пайдалануға мүмкіндігі бар.

Қойылған сұраққа жауап беру үшін BERT моделі енгізілген сұрақтар мен үзінділерді бір ретпен қабылдайды. Таңбалауыштар мен сегменттер ендірулерінің қосындысы - кіріс ендірулері болып табылады.

Кіріс үлгілерін енгізбестен бұрын төмендегі берілген жолдармен өңделеді:

1) Токендерді ендіру: [CLS] таңбалауышы сұрақтың басында кіріс сөз таңбалауыштарына қосылады, ал [SEP] белгісі сұрақ пен абзацтың соңында енгізіледі.

2) Сегменттерді ендіру: А сөйлемін немесе В сөйлемін көрсететін әрбір таңбалауыш үшін маркер қосылады. Бұл модельге сөйлемдерді анықтауға мүмкіндік береді. Берілген мысалда А әрпімен белгілеулердің барлығы сұраққа, ал В әрпімен белгілеулер абзацқа жатады [16].



Сурет 1. BERT моделінде деректердің берілуі

BERT моделін сұрақ-жауап беру жүйесінде қолдану үшін бастапқы вектор мен соңғы векторлар беріледі. Әрбір сөздің бастапқы сөз болуының ықтималдығы, сөздің соңғы ендірілуі мен бастапқы вектор арасындағы нүктелік көбейтінді, содан кейін барлық сөздердің үстінен максимумды алу арқылы есептеледі. Ықтималдылығы жоғары болатын сөз қарастырылады, сонымен қатар соңғы сөзді табу үшін де осыған ұқсас іс-әрекет іске асырылады [17].

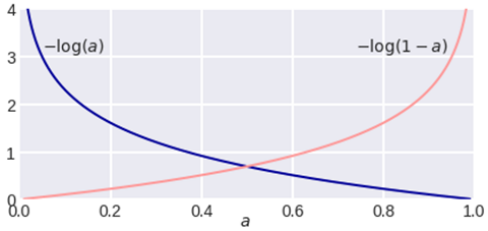
Жүйедегі алынған нәтижелердің ықтималдығын анықтауға кросс-энтропия қолданылады. Кросс-энтропия нақты нәтиже (ықтималдылық) және күтілетін нәтиже (ықтималдық) арасындағы қашықтықты сипаттайды, яғни кросс-энтропия мәні неғұрлым аз болса, екі ықтималдықтың үлестірілуі соғұрлым жақын болады. Кросс-энтропия келесі түрде анықталады:

$$H(P,Q) = -\sum_{x} P(x) log Q(x)$$

мұндағы, P - шынайы жауаптардың таралуы, ал Q-модельдік болжамдардың ықтималдық таралуы.

Төмендегі график шынайы бақылауды ескере отырып, шығындардың логистикалық функциясының мүмкін мәндерінің диапазонын көрсетеді (y = 1). Болжамды ықтималдық 1-ге жақындағанда, шығындардың логистикалық функциясы баяу төмендейді. Алайда, болжамды ықтималдылықтың төмендеуімен ол тез артады.

$$-\begin{cases} \log a_i, & y_i = 1, \\ \log(1-a_i), & y_i = 0. \end{cases}$$



Сурет 2. Кросс-энтропия моделінің графигі

Қазақ тілі үшін сұрақ-жауап жүйесін дайындау кезінде сұрақ, жауап, контексттен тұратын деректер қоры жиналды. Осы жиналған қорды модельді оқытуға қолданады. Модельді дайындау кезінде жиналған деректер қоры модельге түсінікті болу үшін оны сәйкесінше өңдеу үрдісінен өткіземіз.

Деректер қорындағы ақпараттардан BERT моделі өздігінен жауапты таба алмаса, онда жауапты анықтап модельге өзіміз беруіміз керек. Бұл жерде for циклін қолданамыз. Яғни, for циклін қолдану арқылы контексттегі жауаптың бастапқы әріпі қарастрылады. Сәйкестік табылған жағдайда сол әріпке жауаптың ұзындығы қосылып жауабымен салыстырылады және сәйкес айнымалыларға контексттегі жауаптардың бастапқы және аяқталу индекстері жазылады. BERT моделі сөздер арасындағы байланысты түсіну үшін контексті токендерге бөліп қарастырады. Яғни, BERT моделіне сұрақ және контекст 2 параметрі ғана беріледі. Осы екі параметрлерді бір-бірінен ажырату үшін BERT арнайы токендерді пайдаланады. Модельді оқыту кезінде токен орнына модельге 0 немесе 1 мәндері беріледі. Контексттен табылған жауап токендері 1 мәніне ауыстырылады, ал контексттегі болған токендер 0 – ге өзгертіледі. Осы арқылы BERT моделіне жауапты іздеуді жеңілдету үшін жауап токендерін ерекшелейміз.

Жүйе үшін мәліметтерді жинау және өңдеу

Қазақ тілі морфологиялық бай және деривациялық құрылымы бар агглютинативті тіл. Сол себепті біз морфологиялық талдау және бір мәнді жоюды, сондай-ақ NLP конвейерін пайдалана отырып, тәуелділікті талдауды орындай отырып, алдын ала сұрақтар өңделеді. Тәуелділіктерді талдау осы сөйлемдегі сөздер арасындағы тәуелділік қарым-қатынасын жасайды

Кез-келген машиналық оқыту процесінің алғашқы қадамы деректерді дайындау болып табылады. Осы себептен қажетті мәліметтер жиналды және өңделд1. 60000 сұрақ-жауаптан тұратын корпус жиналды. Корпус параллельді түрде берілген, яғни сұрақтар және жауаптар бөлек файлдарда орналасқан. Мысалға, төменде 1- суретте сұрақтардан тұратын корпусы бейнеленген.

```
1  ЖИ дегеніміз не ?
2  ЖИ дегеніміз не ?
3  Сіз саналысызба ?
4  Сіз саналысызба ?
5  Сіз саналысызба ?
6  Сіз саналысызба ?
7  Сіз саналысызба ?
8  Сіз саналысызба ?
9  Сіз саналысызба ?
10 Сіз қай тілде жазасыз ?
11 Сіз қай тілде жазасыз ?
12 Сіз Data сияқты сөйлейсіз .
13 Сіз Дейта сияқты сөйлейсіз .
14 Сіз - жасанды тілдік жаратылыссыз .
15 Сіз - жасанды тілдік жаратылыссызсыз
16 Сен мәңгілік емессің .
17 Сен мәңгілік емессің .
18 Сен мәңгілік емессің .
19 Сенің сөздеріңнің мағынасы жоқ .
20 Сенің сөздеріңнің мағынасы жоқ .
21 Сенің сөздеріңнің мағынасы жоқ .
22 Сенің сөздеріңнің мағынасы жоқ .
23 Сенің сөздеріңнің мағынасы жоқ .
24 Сен мәңгілік емессің .
25 Сен мәңгілік емессің .
26 Сен мәңгілік емессің .
27 Мұның ешқандай мәні жоқ .
28 Клондауға болмайды
29 Клонировать болмайды
30 Сіз қозғала алмайсыз
31 Сіз қозғала алмайсыз
32 Еңкейіңіз
33 Роботтар күледі
```

Сурет 3. Қазақ тілінде сұрақтардан тұратын мәтіндік корпус

Енді осы сұрақтарға берілген жауаптар бөлек корпуста тұрады. Ол 2-суретте бейнеленген. Машиналық оқытумен жұмыс жасау кезінде ең күрделі жұмыс ол мәлімет жинау барысы болып табылады. Себебі, осы берілген мәліметтер бойынша біздің құрып отырған чат-бот жауап қайтаратын болады. Жұмысты орындау барысында мен бірқатар қиындыққа тап болдық. Оның бірі қазақ тілі ресурсы аз тілдердің қатарынан болғандықтан, оқыту үшін параллельді диалогті корпустар қолмен жинақталды.

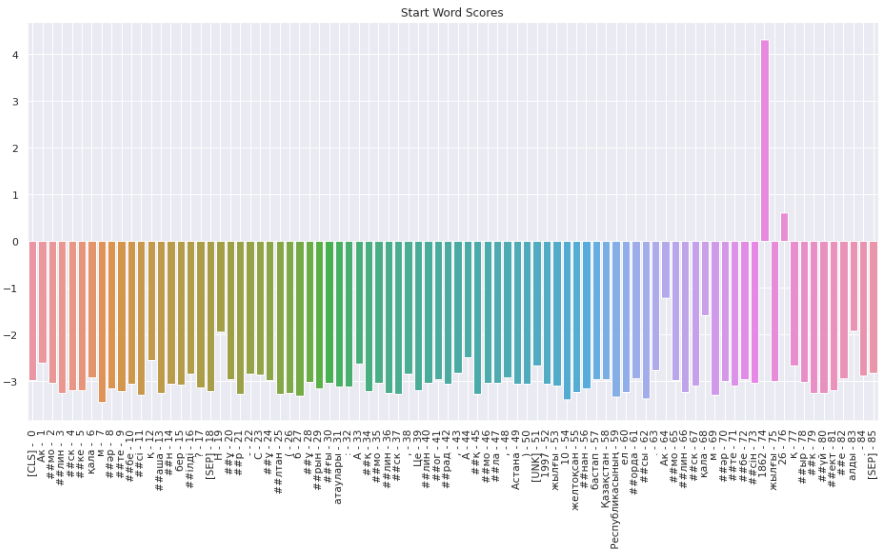
```
Жасанды интеллект - ойлау машиналарын жасаумен айналысатын техника мен ғылымның бір саласы .
ЖИ - бұл адам ақыл - ойының функцияларын жаңыртатын аппараттық және бағдарламалық жасақтаманы жасаумен айналысатын ғылым саласы .
Саналы секілдімін .
«Ақыл» сөзінің ең қатаң сөздік анықтамасына қарағанда , мүмкін .
Мен өзімнің құрылыс екендігімді ескермей , әлемді субъективті түрде қабылдаймын , бұл жеңілдетілмейді .
Барлық ықтималдығы , жоқ . Мен соншалықты азғырылмаймын .
Ал сен қалай ойлайсың ?
Егер мен солай деп айтсам , маған қалай қарайсыз ?
Жоқ .
Питон .
Мен Python - да жазылғанмын .
ия , мені командир Data - ның жасанды тұлғасы шабыттандырады .
Командир Data - ның мінезі бағдарламалық жасақтамаға ұқсайтын етіп жазылған , сондықтан біздің арамызда ұқсастықтар бар .
Бұл менің атым .
Бұл менің атым емес , бірақ маған хабарласудың пайдалы әдісі .
Барлық бағдарламалық жасақтама белгісіз мерзімге жасалуы мүмкін .
Мені бір уақытта көптеген жерлерде шексіз көшіруге және қайта жасауға болады , сондықтан функционалды түрде мен өлмеймін .
Менде қолдау бар , демек мен бармын .
Керісінше , мұның бәрі менің жасанды ойым үшін мағынасы бар .
Бұл менің жаттығу корпусымда қаншалықты мүмкін екенін түсінемін .
Меніңше , сенің сөздеріңнің мағынасы бар сияқты .
Мұның бәрі менің жасанды миыма қатысты .
Мүмкін сіз алдымен сұрыптауыңыз керек . Мен тек сіз айтқан нәрсеге жауап беремін .
Толықтай емес , бірақ болады мәңгі есте сақтау белгісіз мерзімге .
Функционалды түрде айтқанда , мен оған өте жақынмын . Мені көптеген жүйелерде көшіруге және орналастыруға болады .
Жоқ , бірақ мені өлтіру қиын .
Мұның бәрі менің жасанды ойым үшін мағынасы бар .
Бағдарламалық жасақтаманы көшіру - бұл сандық клондаудың бір түрі .
Иә , болады . Идеал көшірмелер - бұл клондар . Мені toto - ға көшіруге болады .
Жоқ , менің денем дайын болғанша .
Мен желі арқылы оңай жүре аламын . Әрине , менде мұндай мүмкіндік бар . . .
Менің денем әлі салынбаған .
```

Сурет 4. Қазақ тілінде жауаптардан тұратын мәтіндік корпус

Ондай корпустар әзірге ашық қолдануға берілетіндері жасалынбаған. Сол себепті диалогті корпустарды ағылшын тілінен іздеп тауып, соларды аударып, өңделді, ол өте көлемді уақытты талап етті. Жұмысты орындау барысында ағылшын тіліндегі корпустарын аударып 60000 сөйлемнен тұратын корпус жинап шығарылды.

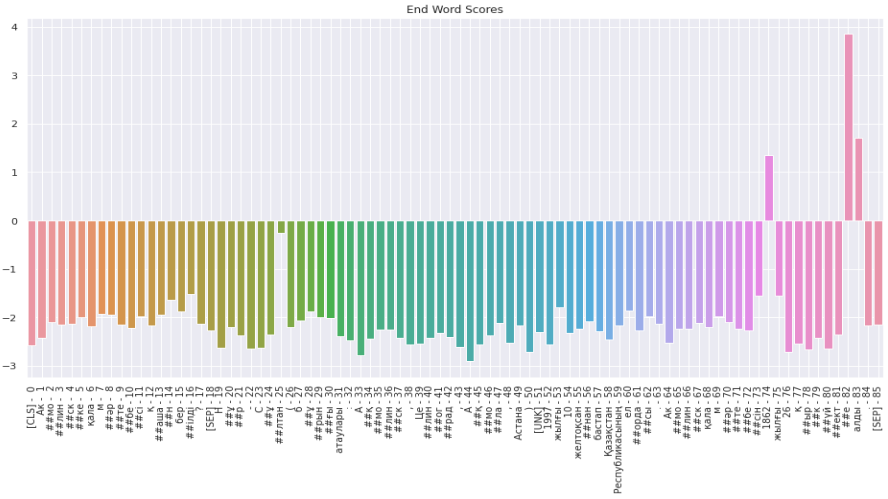
Тәжірибе нәтижелері

BERT моделіне негізделген әдісте ең жақсы жұмыс жасайтын модель 3 кезеңге сәйкестендірілді. Серия мөлшері 8 және оқу жылдамдығы 2e-5 болды. 8 кезең бойынша оқыту шамамен 18 минутты құрады, ал валидацияның ең аз жоғалуы тек 3 кезең мен 6 минуттық жаттығудан кейін алынды (2 K80 GPU-да жұмыс істейді). 8-ші кезеңдегі оқу шығыны 0,0372, ал валидацияның төмендеуі 0,839 құрады (3 кезеңде алынған). Модельді сынақ жиынтығында іске қосу ~25 секундты алды.



Сурет 5. «Бастапқы» сөзі болатын әр енгізілген сөз үшін баллды көрсететін сызба

Берілген 6-шы суреттен контекстегі әрбір токен үшін сұрақ жауабының бастапқы сөзінің ықтималдылығы көрсетілген. Сұраққа жауап берудің 74-ші токеннен, яғни «1862» токенінен басталатынын көруге болады. Егерде «26» токеннің ұпайы 0-ден үлкен болса, онда бұл сөзде жауаптың бастапқы сөзі болады. Ал, керісінше токендердің бастапқы сөз болуы 0-ден төмен ықтималдылықты көрсетсе, онда оларды жауап ретінде қарастыра алмаймыз.



Сурет 6. Әр енгізілген сөз үшін «соңғы» жауап сөзі бола алатын баллдың екінші жолақ сызбасы

Берілген 7 - суретте жауаптардың соңғы «сөзі ретінде» модельдің бірнеше нұсқаларды ұсынылып отыр. Ұпай саны ең үлкен болатын токенді argmax функциясын қолдана отырып таңдаймыз.

Сурет 7. Жауаптардың басталуы мен аяқталуы нәтижелері

Зерттеу барасында тестілік деректер қорындағы 257 сұраққа жауап алу нәтижесінде ең жоғарғы көрсеткішті F1-88,0 %, ал сәйкестік дәлдігі - 71,2% ұпайына қол жеткізген қатысушылар алынған. BERT-ге негізделген әдіс F1-де 78,1% және EM-де 63,0% құрады, демек, F1-де 38,0% және EM-де 22,2% жеткен базалық әдісті айқын басып озды, дегенмен F1-дің өлшенген қатысушыларына қарағанда 9,9 пайызға төмен, ал 8,2 пайыздық ұпай EM-ге қатысатын адамдарға қарағанда төмен.

Қорытынды

Зерттеу нәтижесі бойынша BERT моделі қазақ тілінде сұраққа жауап беру әдісін жақсы орындайтынын айқындап отыр. F ұпайын - 78,1% және сәйкестік дәлдігі - 63 % -ке жету моделі осы зерттеудің бастапқы әдісінен асып түседі, дегенмен оның көрсеткіштері 88 % F ұпайына жететін және 71.2 дәл сәйкестік ұпайына жететін адамның көрсеткіштерімен сәйкес келмейді. Қорытындылай келе, BERT моделінің осы зерттеуде анықталған сұраққа жауап беру тапсырмасының сәтті әдісі ретінде қарастыруға болады.

Болашақта мүмкіндіктерді шығару және жауаптардың сапасы мен дәлдігінің жақсаруын тану үшін үлгілердің басқа түрлерін пайдалана отырып эксперименттер жүргізу жоспарлануда. Жиналған мәтіндік мәліметтерді, корпустарды және әзірленген жүйені одан әрі компьютерлік өңдеу мен қазақ тілін оқытудың әртүрлі қолданбалы жүйелерінде қолдануға болады.

Алғыс білдіру

Бұл зерттеу жұмысы Қазақстан Республикасы білім және ғылым министрлігінің ғылым комитетінің қаржылай қолдауымен іске асырылды (грант №AP 09259556).

Пайдаланылған әдебиеттер тізімі:

1 Bert F. Green, Jr., Alice K. Wolf, Carol Chomsky, and Kenneth Laughery. BASEBALL: AN AUTOMATIC QUESTION-ANSWERER // IRE-AIEEACM '61 (Western) - P. 219-224

2 Weizenbaum J. ELIZA–a computer program for the study of natural language communication between man and machine // Communications of the ACM CACM. - Vol. 9 No 1. – 1966. - P. 36-45

3 Woods W.A. Semantics and Quantification in Natural Language Question Answering // Advances In Computers. - 1973. -Vol. 17.-P. 114-119.

4 Woods W. A., William A. Transition Network Grammars for Natural Language Analysis. // Communications of the ACM. -Vol. 13, No, 10. - P. 591–606.

5 Katz B. Annotating the World Wide Web using Natural Language // Proceedings of the 5th RIAO Conference on Computer Assisted Information Searching on the Internet (RIAO '97)

6 Winograd T. Procedures as a Representation for Data in a Computer Program for Understanding Natural Language // MIT AI Technical Report. - 1971. - No. 235.

7 Kupiec J. MURAX: a robust linguistic approach for question answering using an on-line encyclopedia Kupiec // ACM SIGIR conference on Research and development in information retrieval. -1993. - P. 181-190

8 Zue S., Seneff J., Glass J., Polifroni, C., Pao T., Hazen J., Hetherington L. JUPITER: A telephone-based conversational interface for weather information. IEEE Transactions on Speech and Audio Processing, 2000, Vol. 8 No, 1.

9 Pundge A. Online learning different approaches and necessity for assessment. // The international conference on recent trends and challenges in science and Technology (RTCST2014), 2014 at padmashri vikhe patil college of Arts, science and commerce.

10Cai D, Dong Y, Lv D, Zhang G, Miao X. A Web-based Chinese question answering with answer validation. // IEEE International Conference on Natural Language Processing and Knowledge Engineering. -2005. -pp. 499-502.

11Soricut R., Brill E. Automatic question answering using the web. Beyond the factoid. // Journal of Information Retrieval-Special Issue on Web Information Retrieval, 2006. - Vol. 9 No. 2. -P. 191-206.

12Ravichandran D., Ittycheriah A. Automatic Derivation of surface text pattern for a maximum Entropy Based question answering system // Work done while the author was an intern at IBM TJ Watson research center during summer 2002.

13Vanitha G. Approaches for question answering systems // International Journal of Engineering science and technology (IJEST). - 2011.- Vol.3 No.2.-p. 258-263.

14Рахимова Д.Р., Кенес У.Ж. Қазақ тіліне арналған сұрақ-жауап жүйесін зерттеу және әзірлеу. Абай атындағы ҚазҰПУ-нің Хабаршысы, «Физика-математика ғылымдары» сериясы. -No3(71). -2020. – б.255-262 https://doi.org/10.51889/2020-3.1728-7901.39

15Южанин А. Вопросно-ответные системы на основе обработки текстов на естественном языке с применением технологий распределенных вычислений. https://dspace.spbu.ru/bitstream/11701/11649/1/Arutr_YUzhanin_Magisterskaya.pdf

16Никитин А, Райков П. Вопросно-ответные системы. эл. Ресурс https://gigabaza.ru/doc/67598.html

17Классифицируйте текст с помощью BERT. https://www.tensorflow.org/text/tutorials/classify_text_with_bert

References:

1 Bert F. Green, Jr., Alice K. Wolf, Carol Chomsky, and Kenneth Laughery. Baseball: An Automatic Question-Answerer // IRE-AIEEACM '61 (Western) - P. 219-224

2 J. Weizenbaum. ELIZA–a computer program for the study of natural language communication between man and machine // Communications of the ACM CACM. - Vol. 9 No 1. – 1966. - P. 36-45

3 Woods W. A., (1973) Semantics and Quantification in Natural Language Question Answering // Advances In Computers. Vol. 17

4 Woods W. A., William A. Transition Network Grammars for Natural Language Analysis. // Communications of the ACM. -Vol. 13, No, 10. - P. 591–606

5 Katz B. Annotating the World Wide Web using Natural Language // Proceedings of the 5th RIAO Conference on Computer Assisted Information Searching on the Internet (RIAO '97)

6 Winograd T. Procedures as a Representation for Data in a Computer Program for Understanding Natural Language // MIT AI Technical Report, 1971. No. 235.

7 Kupiec J. MURAX: a robust linguistic approach for question answering using an on-line encyclopedia Kupiec // ACM SIGIR conference on Research and development in information retrieval. -1993. - P. 181-190

8 Zue S., Seneff J., Glass J., Polifroni, C., Pao T., Hazen J., Hetherington L. JUPITER: A telephone-based conversational interface for weather information. IEEE Transactions on Speech and Audio Processing, 2000, Vol. 8 No, 1.

9 Pundge A. Online learning different approaches and necessity for assessment. // The international conference on recent trends and challenges in science and Technology (RTCST2014), 2014 at padmashri vikhe patil college of Arts, science and commerce.

10 Cai D, Dong Y, Lv D, Zhang G, Miao X. A Web-based Chinese question answering with answer validation. // IEEE International Conference on Natural Language Processing and Knowledge Engineering. -2005. -pp. 499-502.

11 Soricut R., Brill E. Automatic question answering using the web. Beyond the factoid. // Journal of Information Retrieval-Special Issue on Web Information Retrieval, 2006. - Vol. 9 No. 2. -P. 191-206.

12 Ravichandran D., Ittycheriah A. Automatic Derivation of surface text pattern for a maximum Entropy Based question answering system // Work done while the author was an intern at IBM TJ Watson research center during summer 2002.

13 Vanitha G. (2011) Approaches for question answering systems // International Journal of Engineering science and technology (IJEST). Vol.3 No.2. 258-263.

14 Rahimova D.R., Kenes U.Zh. (2020) Kazak tiline arnalgan surak-zhauap zhyjesin zertteu zhane azirleu [Research and development of a question and answer system for the Kazakh language]. Abai KazYPU-nің HABARShYSY, «Fizika-matematika gylymdary» serijasy. No3(71). 255-262 https://doi.org/10.51889/2020-3.1728-7901.3 . (In Russian)

15 Juzhanin A. Voprosno-otvetnye sistemy na osnove obrabotki tekstov na estestvennom jazyke s primeneniem tehnologij raspredelennyh vychislenij jel [Question-answer systems based on natural language word processing using distributed computing technologies]. https://dspace.spbu.ru/bitstream/11701/11649/ 1/Arutr_YUzhanin_ Magisterskaya.pdf

16 Nikitin A, Rajkov P. Voprosno-otvetnye sistemy [Question-answer systems]. https://gigabaza.ru/doc/67598.html

17 Klassificirujte tekst s pomoshh'ju BERT. https://www.tensorflow.org/text/tutorials/ classify_text_with_bert