МРНТИ 20.23.17
УДК 004.421

https://doi.org/10.51889/2021-4.1728-7901.15

Д.Р. Рахимова [1], Г.Е. Ахмет[1]*
[1] Әл-Фараби атындағы Қазақ Ұлттық Университеті, Алматы қ., Қазақстан
**e-mail: gulstan.akhmet@gmail.com*

МАШИНАЛЫҚ ОҚЫТУ НЕГІЗІНДЕ ҚАЗАҚ ТІЛІНДЕГІ СӨЙЛЕМДЕРДІ СИНТЕЗДЕУ ӘДІСІН ЗЕРТТЕУ ЖӘНЕ ӘЗІРЛЕУ

Аңдатпа

Бүгінгі таңда сөйлемдердің синтезі әртүрлі салаларда қолданылады. Бұл дауыстық көмекшілер, IVR жүйелері, ақылды үйлер, чат-боттар және тағы басқалары. Біраз уақыт бұрын, сөйлеу синтезі саласында, көптеген басқа салаларда секілді, машиналық оқыту пайда болды. Машиналық оқыту-бұл оқуға қабілетті алгоритмдерді құру әдістерін зерттейтін жасанды интеллектінің кең жиынтығы. Бүкіл жүйенің бірқатар компоненттерін нейрондық желілермен алмастыруға болатындығы белгілі болды, бұл қолданыстағы алгоритмдерге сапамен жақындауға ғана емес, тіпті олардан едәуір асып кетуге мүмкіндік береді. Мақалада сөйлемдерді синтездеу технологияларына шолу жасалынды, қазақ тіліндегі сөйлемдерді синтездеу мәселесі чат-бот жүйесі негізінде, seq2seq әдісін қолдана отырып шешілді. Қазақ тілінде параллельді сұрақ-жауап корпусы жинақталды. Қазақ тіліндегі сұрақ-жауап корпусы ағылшын тіліндегі көптеген чат-боттарды құруда қолданылатын Cornell movie, Ubantu тағы да басқа көптеген корпустарды аударып, тазарту нәтижесінде жинақталды. Қазақ тіліндегі сөйлемдерді синтездеу үшін құрылған моделі бойынша корпустарды қолдана отырып, бірқатар эксперименттер жүргізіліп, нәтижелер алынды.

Түйін сөздер: қазақ тілі, NMT, лингвистикалық ресурстар, seq2seq әдісі, сөйлем синтезі, машиналық оқыту.

Аннотация

Д.Р. Рахимова[1], Г.Е. Ахмет[1]
[1]Казахский национальный университет имени аль-Фараби, г. Алматы, Казахстан

ИССЛЕДОВАНИЕ И РАЗРАБОТКА МЕТОДА СИНТЕЗА ПРЕДЛОЖЕНИЯ НА КАЗАХСКОМ ЯЗЫКЕ НА ОСНОВЕ МАШИННОГО ОБУЧЕНИЯ

Сегодня синтез предложений используется в различных областях. Это голосовые помощники, системы IVR, умные дома, чат-боты и многое другое. Некоторое время назад в области синтеза речи, как и во многих других областях, появилось машинное обучение. Машинное обучение-это широкий набор искусственного интеллекта, который изучает методы создания алгоритмов, способных к обучению. Оказалось, что ряд компонентов всей системы можно заменить нейронными сетями, что позволяет не только приблизиться к существующим алгоритмам с качеством, но даже значительно превзойти их. В статье проведен обзор технологий синтеза предложений, решена проблема синтеза предложений на казахском языке на основе системы чат-ботов, с использованием метода seq2seq. На казахском языке собран параллельный корпус вопросов и ответов. Корпус вопросов и ответов на казахском языке был собран в результате перевода и очистки многих корпусов, таких как Cornell movie, Ubantu и других, которые используются для создания множества чат-ботов на английском языке. Проведен ряд экспериментов и получены результаты с использованием корпусов по построенной модели для синтеза предложений на казахском языке.

Ключевые слова: казахский язык, NMT, лингвистические ресурсы, метод seq2seq, синтез предложений, машинное обучение.

Abstract

RESEARCH AND DEVELOPMENT OF A METHOD FOR SYTHESIZING SENTENCES IN THE KAZAKH LANGUAGE BASED ON MACHINE LEARNING

Rakhimova D.R.[1], Akhmet G. E.[1]
[1]Al-Farabi Kazakh National University, Almaty, Kazakhstan

Today, sentence synthesis is used in various fields. These are voice assistants, IVR systems, smart homes, chatbots and much more. Some time ago, machine learning appeared in the field of speech synthesis, as in many other fields. Machine learning is a broad set of artificial intelligence that studies methods for creating algorithms capable of learning. It turned out that a number of components of the entire system can be replaced by neural networks, which allows not only to approach the existing algorithms with quality, but even significantly surpass them. The article provides an overview of sentence synthesis technologies, solves the problem of sentence synthesis in the Kazakh language based on a chatbot system using the seq2seq method. A parallel corpus of questions and answers has been collected in the Kazakh language. The corpus of questions and answers in Kazakh was collected as a result of translation and cleaning

of many corpora, such as Cornell movie, Ubantu and others, which are used to create many chatbots in English. A number of experiments were carried out and results were obtained using corpora based on the constructed model for the synthesis of sentences in the Kazakh language.

Keywords: Kazakh language, NMT, linguistic resources, seq2seq method, sentence synthesis, machine learning.

Кіріспе

Бүгінде ірі компаниялар әр түрлі кеңсе тапсырмаларын орындай алатын интеллектуалды көмекшіні енгізу жолдарын іздейді. Мұндай тапсырмалардың кейбір мысалдары: жұмысқа қабылдау, жаңа қызметкерлерді оқыту, сұрақтарға жауап беру, қағазбастылықты реттеу, уақыт пен ресурстарды жоспарлау [1]. Қазіргі әлемде көбірек мәтіндер компьютермен синтезделуде. Көптеген зерттеулер мен мақалалар ауызша сөйлеуді синтездеуге арналған, бірақ жазбаша сөйлеуді синтездеуге көп көңіл бөлінбейді, дегенмен бұл салада мәселелер өте қызықты.

Синтез мәселесін талдау мәселесіне кері мәселе деп есептеуге болады. Автоматты синтез-бұл біріктірілген мәтінді шығару процесі, оның жеке кезеңдері морфологиялық талдаумен бірдей, бірақ керісінше қолданылады: алдымен семантикалық синтез жүзеге асырылады, содан кейін синтаксистік, морфологиялық және графематикалық синтездер орындалады. Семантикалық синтез- бұл сөйлемнің семантикалық жазуынан оның синтаксистік құрылымына өту; синтаксистік синтез- сөз тіркестерінің синтаксистік құрылымынан сөз тіркестерінің лексикалық-грамматикалық сипаттамаларының тізбегіне ауысу; лексика-морфологиялық синтез -лексикалық-грамматикалық сипаттамадан нақты сөз формасына ауысу. Морфологиялық синтезде сөздің қалыпты формасы мен оның параметрлері бойынша бағдарлама тиісті сөз формасын табады. Графематикалық синтез- сөздерді бір мәтінге біріктіреді, кіріс мәтінінің фрагменттерінің Шығыс фрагменттеріне сәйкестігін бақылайды [2]. Компьютер қолданушының өзімен байланыс жасау барысында сөйлем құрады, сөздің тура мағынасында онымен сөйлеседі деп болжайды. Оның үстіне, ол мұны дайын сөз тіркестерінің көмегімен емес, қойылған сұраққа жауап беретін мағыналы сөйлемдер құру арқылы жасайды.

Осы мақсатқа жету үшін компьютерді табиғи тілдегі сөйлемдерді өз бетінше құруға және түсінуге үйрету керек.

Машиналық оқыту негізіндегі интеллектуалды көмекшілер компьютерлік бағдарламаларға деректерді үйренуге мүмкіндік береді. Бүкіл әлемдегі компаниялар өз бизнесінде машиналық оқытуды қолданады, себебі бұл оларға аз инвестициямен көп пайда алуға және бір тапсырмаға аз уақыт жұмсауға мүмкіндік береді [3]. Машиналық оқыту- бұл оқуға қабілетті алгоритмдерді құру әдістерін зерттейтін жасанды интеллектінің кең жиынтығы [4]. Бүкіл жүйенің бірқатар компоненттерін нейрондық желілермен алмастыруға болатындығы белгілі болды, бұл қолданыстағы алгоритмдерге сапамен жақындауға ғана емес, тіпті олардан едәуір асып кетуге мүмкіндік береді [5].

Берілген мақалада қазақ тіліндегі сөйлемдерді синтездеу мәселесі машиналық оқыту негізінде шешілетін болады. Синтездеу мәселесі чат-бот жүйесі көмегімен шешіледі. Чат-бот - заманауи бизнес шешімдерінің өкілдерінің бірі. Бұл күндері олар барған сайын танымал бола бастады, себебі олар тәулік бойы, аптасына 7 күн жұмыс жасай алады. Мақалада сөйлемдерді синтездеу үшін қолданылатын бірқатар жұмыстар талданды. Қазақ тіліндегі сөйлемдерді синтездеу үшін моделі таңдалды және әдісі құрылды. Жобада синтездеу үшін нейронды машиналық аударудың seq2seq моделі қолданылған болатын. Seq2seq моделі негізінде эксперименттер жүргізіліп, бірқатар нәтижелер алынып, талданды.

Пәндік аймақ бойынша жұмыстарға шолу

Белгілі бір тіл үшін синтез мәселесін шешудің әртүрлі ғылыми тәсілдері мен әдістері бар. Олардың кейбіреулері төменде ұсынылады. Ағылшын тілінде жазбаша диалогты синтездейтін алғашқы компьютерлік бағдарламалардың бірі американдық ғалым Джозеф Вейценбаумның "Элиза"бағдарламасы болды. Оның алғашқы нұсқасы 1966 жылы сыналды. Бұл бағдарлама белсенді тыңдау техникасын қолдана отырып, психотерапевтпен диалогты еліктеді. Бағдарлама Бернард Шоудың "Пигмалион" пьесасының кейіпкері Элиза Дулитлдің есімімен аталды, ол "жоғары деңгейдегі адамдар" тілін үйретті. Бағдарламаның мақсаты нақты мағынада ойлауды модельдеу емес, шектеулі бағдарламалық ресурстармен, сондай-ақ лингвистикалық талдау мен синтездің бастапқы деңгейімен байланысты сөйлеу әрекетін модельдеу болды [2]. Бағдарлама ең аз лингвистикалық ақпаратты қамтыды: 1) кейбір тұрақты сөйлеу формулаларын іске асыратын кілт сөздер жиынтығы, 2) алдыңғы мәлімдемені жалпы сұраққа айналдыру мүмкіндігі.

Бағдарламаны құруда қолданылатын алгоритмдердің қарапайымдылығына қарамастан, оның көмегімен 1950 жылы ұсынылған ағылшын ғалымы Алан Тюрингтің әйгілі сынағын жоққа шығаруға болады. Бұл тесттің стандартты түсіндірмесі келесідей: "адам бір компьютермен және бір адаммен өзара әрекеттеседі. Сұрақтарға жауаптардың негізінде ол кіммен сөйлесетінін анықтауы керек: адаммен немесе компьютерлік бағдарламамен". Тесттің барлық қатысушылары бір-бірін көрмейді. Егер қатысушы адам осы әңгімелесушілердің қайсысы адам екенін нақты айта алмаса, онда машина сынақтан өтті деп саналады. Ауызша сөйлеуді тану мүмкіндігін емес, машинаның интеллектісін дәл тексеру үшін әңгіме "тек мәтін" режимінде, мысалы, пернетақта мен экранның көмегімен жүзеге асырылады [6]. "Элиза" бағдарламасымен эксперименттер жүргізу кезінде субъектілердің 62%- ы кіммен сөйлесетіндерін анықтауға шақырылды, оларға адам жауап берді деп шешті [7].

Желіде ағылшын, неміс және басқа тілдерге арналған кездейсоқ мәтін генераторлары қол жетімді. Олардың бірі, Randomtextgenerator, параметрлері шамалы өзгеріп, кейбір мәтіндер шығарады. Бұл бағдарлама үшін мәтіндер еуропалық тілдерге қол жетімді, негізінен латын тамыры бар. Орыс тілі үшін бірнеше жүйелер мен ақылы платформалар жасалды. Олардың бірі - Морфер жүйесі. Бағдарлама Ресейдің, жақын және алыс шетелдердің жүздеген кәсіпорындарында іске асырылды және сұраныс артып келеді. Бағдарлама келесі функцияларды жүзеге асырады:

- сөздер мен сөз тіркестерін септіктер бойынша септеу және көпше немесе жекеше түрде жіктеу;
- адамның жынысын аты бойынша анықтау;
- сандар мен ақша сомаларын жазу;
- өлшем бірліктерін санмен үйлестіру;
- қалалар мен елдердің атауларынан сын есім жасау.

Авто-түзеткіштердің құрылысы бірқатар іргелі және әлі толық шешілмеген мәселелермен кездеседі: сөздіктерді ықшам сақтау, морфологиялық және синтаксистік талдаудың тиімді әдістері және т.б.

Орыс тіліндегі мәтінге арналған автоматты түзетушілер: Орфография, Advego, ORFO, LINAR және т.б. «Рифмач» - бұл көрсетілген параметрлер бойынша құттықтаулар шығаруға арналған бағдарлама. «Textgen» - берілген тақырып бойынша ақылы мәтін генераторы. Жасалған мәтінде зат есімдер, сын есімдер, етістіктер мен үстеулер бар. Бұл автоматты мәтін генераторлары берілген тақырып бойынша тек орыс грамматикасы ережелерінің көпшілігіне сәйкес келетін бірегей мәтін құруға мүмкіндік береді. Мәтінді генерациялауға немесе синтездеуге арналған түркі тілдерінің ішінде түрік тіліндегі шығармалар көбірек, басқа топтар үшін бұл зерттеулер тобынан зерттеулер жоқтың қасы [8].

Қазіргі кезеңде табиғи тілдегі жазбаша диалогтарды компьютерлік модельдеу жүйелері күрделі алгоритмдерді қолданады. Атап айтқанда, виртуалды агенттерді (немесе боттарды) құру үшін қолданылатын AIML (Artificial Intelligence Markup Language) жасанды интеллектке арналған арнайы белгілеу тілі жасалды. Сұхбаттасушымен диалогты модельдейтін боттар компьютерлік ойындарда және корпоративті веб-беттерде қолданылады, мысалы, пайдаланушының ұялы байланыс операторының немесе сауда желісінің мүмкіндіктері туралы сұрақтарына жауап беру үшін қолданылады.

Машиналық оқыту негізіндегі чат-боттары - бұл қолданушыларға технологиялармен өзара әрекеттесуге және тапсырмаларды автоматтандыруға көмектесетін чат-бот түрі. Жасанды интеллект, машиналық оқыту, табиғи тілді өңдеу және деректерді талдау саласындағы жетістіктер оларды тез қабылдауға түрткі болды. Чат-боттар қазіргі кезде бизнес, банк ісі, денсаулық сақтау, оқу, саяхат және тағы да басқа сияқты көптеген салаларда кең танымал болып келеді. Дауыстық ассистенттермен бірінші болып интеграцияланған компаниялар клиенттердің жаңа көмекшісі болуға мүмкіндік алады. Алғашқы мысалдар қазірдің өзінде бар. «Алиса» Utkonos, Papa John's, McDonald’s, S7 Airlines клиенттеріне қажетті ақпаратты алуға және тапсырыс беру қызметтерін алуға көмектеседі. «Ауызекі сөйлеудің» артында сөйлеуді тану және синтездеу технологиялары, табиғи тілді түсіну технологиясы, машиналық оқыту алгоритмдері жатыр.

Бүгінгі таңда чат-бот технологиясы тіпті шағын бизнеске де қол жетімді және оның танымалдығы әлеуметтік енудің артуына байланысты артып келеді. Бүгінде 2 миллиардтан астам адам пайдаланатын желілер мен мессенджерлер бар. Осының арқасында бизнес онлайн режимінде жұмыс істеуге көбірек тәуелді, ал research And Market зерттеуі бойынша чат-боттар нарығы $2 млрд-қа жетті және жыл сайын 30% - ға өсуді жалғастыруда. Бұл ретте чат-боттар технологиясының өзі тек батыс және технологиялық нарықтарда ғана жұмыс істемейтінін атап өту қажет. Мысалы, көрші Ресейде,

Accenture мәліметтері бойынша, өткен жылы чатботтар нарығы шамамен 1,5 миллиард рубльді құрады.мамандардың болжамына сәйкес, ол жыл сайын 30%-ға өседі, бұл жылына 400-600 миллион рубльді құрайды.

Қазақстанда бизнес пен мемлекеттік құрылымдар азаматтармен коммуникация арнасы ретінде чат-боттарды біртіндеп меңгеруде. Мысалы, Астана қаласының "109 кезекші қызметінің" өз боты бар. Ол арқылы коммуналдық-тұрмыстық сипаттағы мәселелер бойынша өтініштер жіберуге болады. "Қазпочта" бот сәлемдемелерді трек-код бойынша қадағалауға мүмкіндік береді және олардың мәртебесі туралы хабарлама жібереді. Ол арқылы ең жақын пошта бөлімшелері туралы ақпарат алуға болады. Telegram және Facebook Messenger-де бірнеше қазақстандық банктердің өз чат-боттары бар. Бот @KZPhoneOperatorBot ұялы байланыс операторын телефон нөмірі бойынша есептеуге мүмкіндік береді. DAR VIS әмбебап боты бұл тізімді толықтырды. Оны әзірлеу кезінде ұлттық менталитет пен сөйлеудің ерекшеліктері ескеріледі. Оны қазақстандық банктердің карталарына және телефон нөміріне байлауға болады. Чат-Ботта қазақ және орыс тілдерінде стикерлер және ұлттық стильдегі ашық хаттар бар. Болашақта қосымшаның барлық функционалы қазақ тілінде локализацияланатын болады. Қазір интерфейс ішінара қазақ тіліне аударылған және чат-бот қазақша қарапайым диалог жүргізе алады.

Бұл қазіргі кезінде синтездеуде қолданылатын қолда бар жүйелерге сипаттама болатын. Бұл жұмысымда мен синтездеуді машиналық оқыту негізіндегі чат-боттар көмегімен жасайтын боламын. Бұл таңдауға тоқталу себебім машиналық оқыту негізіндегі чат-боттар үлкен жетістіктер көрсетуде. Және зерттеу барысында қазақ тіліндегі сөйлемдерді чат-бот негізінде синтездейтін жүйелер жоқтың қасы, ашық түрде ешқандай жұмыс ұсынылмаған.

Есепті шешу барысы, қолданылған әдіске сипаттама

Чат-бот мәтін бойынша өзіндік сананы дамытуға үйретілген және оны адамдармен қалай сөйлесуге болатындығын үйрете аламыз. Мәліметтер қаншалықты көп болса, соғұрлым машинада оқыту тиімдірек болады.

Қазақ тіліндегі мәтіндерді синтездеу үшін чат-бот құруда ең бірінші деректер жинау қажет болады. Кез-келген машиналық оқыту процесінің алғашқы қадамы деректерді дайындау болып табылады. Осы себептен ең бірінші кезекте чат-ботты оқытуға қажетті мәліметтер жиналды. 60000 сұрақ-жауаптан тұратын корпус жиналды. Корпус параллельді түрде берілген, яғни сұрақтар және жауаптар бөлек файлдарда орналасқан. Мысалға төменде 1- суретте сұрақтардан тұратын корпус 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 1. Сұрақтардан тұратын корпус

Енді осы сұрақтарға берілген жауаптар бөлек корпуста жинақталады. Ол 2-суретте бейнеленген.

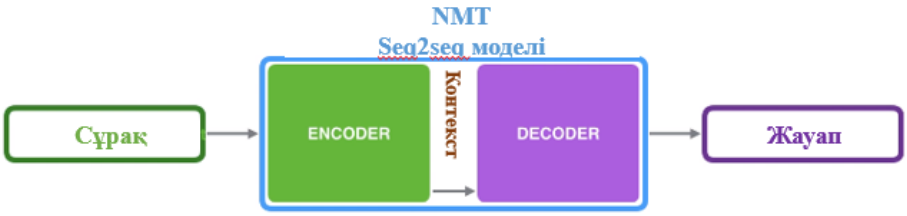
```
Жасанды интеллект - ойлау машиналарын жасаумен айналысатын техника мен ғылымның бір саласы .
ЖИ - бұл адам ақыл - ойының функцияларын жаңғыртатын аппараттық және бағдарламалық жасақтаманы жасаумен айналысатын ғылым саласы .
Саналы секілдімін .
«Ақыл» сөзінің ең қатаң сөздік анықтамасына қарағанда , 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 2. Жауаптардан тұратын корпус

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 бірқатар қиындықтар болды. Оның бірі қазақ тілі ресурсы аз тілдердің қатарынан болғандықтан, оқыту үшін параллельді диалогті дайын корпустар болмады. Сол себепті диалогті корпус басқа тілдегі жұмыстарды аудара отырып жиналды, тазаланды, ол өте көлемді уақытты талап етті. Жұмысты орындау барысында ағылшын тіліндегі көптеген чат-бот жүйелерімен танысып, олардың корпустарын аударып 60000 сөйлемнен тұратын корпус жинап шығарылды. Екінші кезекте мәліметтерді жинағаннан кейін модельді құру болып табылады. Жобада машиналық оқыту негізіндегі чат-ботты құру үшін seq2seq моделі [10] қолданылды.

Seq2Seq- RNN қолданатын Encoder-Decoder моделінің бір түрі. Ол машиналық өзара әрекеттесу және машиналық аударма үшін модель ретінде пайдаланылуы мүмкін. Ол екі RNN-ден тұрады: кодтаушы және декодер. Кодтаушы кіріс ретінде реттілікті(сөйлемді) қабылдайды және әр уақыт кезеңінде бір символды(сөзді) өңдейді. Оның мақсаты - қажет емес ақпаратты жоғалтқан кезде тізбектегі маңызды ақпаратты кодтайтын белгілердің ретін белгіленген векторлық сипат векторына айналдыру. Уақыт осі бойынша кодтаушыдағы деректер ағынын тізбектің бір ұшынан екіншісіне жергілікті ақпарат ағыны ретінде елестетуге болады.

Мүмкін, диалогты жүйенің проблемасы машиналық аудармамен қалай байланысты екендігі түсініксіз болуы мүмкін, бірақ олар өте ұқсас. Чат-ботқа сөйлесу кезінде жауаптар беру машиналық аударма жүйесінде ағылшын тіліндегі сөйлемнің неміс тіліндегі аудармасын жасаудан ерекшеленбейді. Аударма да, әңгімелесу тапсырмалары да моделдің бір реттілікті екінші реттілікпен сәйкестендіруді талап етеді. Ағылшын лексемаларының реттілігін неміс реттілігімен салыстыру диалогтық жүйенің күтілетін жауабына сөйлесудегі табиғи тілдік сұрақтарына бейнелеуге өте ұқсас. Бірақ, бізге чат-ботымыз әңгіме айтатындай болуы үшін, едәуір көп мәліметтер қажет болады. Нақты айтатын болсақ, корпусымыздың көлемі үлкен болуы қажет[9]. Машиналық оқытуды іске асыру үшін NMT тәсілі пайдаланылды. NMT - бұл үлкен жасанды нейрондық желіні қолданатын машиналық аударма тәсілі. Машиналық оқыту 3-суретте бейнеленген моделі бойынша жұмыс істейтін болады:



Сурет 3. Сөйлемдерді синтездеу моделі

Нейрондық машиналық аудармада элементтер тізбегі-бұл кезекпен өңделетін сөздер жиынтығы. Кодтаушы кіріс тізбегінің әр элементін өңдейді, алынған ақпаратты контекст (мәтінмән) деп аталатын векторға аударады. Барлық кіріс тізбегін өңдегеннен кейін, кодтаушы контекстті декодерге жібереді, содан кейін элементтің артындағы элементтің шығу тізбегін құра бастайды. Машиналық аудармаға қатысты контекст вектор (сандар массиві) болып табылады, ал кодтаушы мен декодер өз кезегінде көбінесе қайталанатын нейрондық желілер болып табылады

Оқыту барысында ашық қолданыстағы бағдарлама қолданылған болатын. Қолданылған туториал [10] сілтемесі бойынша қол жетімді. Оқыту осы ресурстағы келесі код арқылы орындалды:

```
!python -m nmt.nmt.nmt \
    --src=en --tgt=vi \
    --vocab_prefix=./nmt_data/vocab  \
    --train_prefix=./nmt_data/train \
    --dev_prefix=./nmt_data/tst2020  \
    --test_prefix=./nmt_data/tst2021 \
    --out_dir=./nmt_model \
    --num_train_steps=12000 \
    --steps_per_stats=100 \
    --num_layers=2 \
    --num_units=128 \
    --dropout=0.2 \
    --metrics=bleu [9]
```

Келесі кезекте корпус жүктеледі. Ол үшін бірінші кезекте nmt_data деген папка ашып, сол жерге барлық корпус жүктелді. Параллельді корпус бөліктерге бөлінді:

- Train.en – оқытуға қолданылатын сұрақтар-56796 сөйлем
- Train.vi- оқытуға қолданылатын жауаптар-56796 сөйлем
- Tst2020.en- тестингке қолданылатын сұрақтар -1604 сөйлем
- Tst2020.vi- тестингке қолданылатын жауаптар-1604 сөйлем
- Tst2021.en- тестингке қолданылатын сұрақтар-1600 сөйлем
- Tst2021.vi- тестингке қолданылатын жауаптар-1600 сөйлем
- Vocab.en-оқытуда қолданылған сұрақтар файлы бойынша құрылған сөздік
- Vocab.vi- оқытуда қолданылған жауаптар файлы бойынша құрылған сөздік

Осы файлдарды қолданып оқытуды жүргізіліп басталады. Гипер параметрлер берілді, қадам-12000, қабат саны-2, және де метрика ретінде bleu метрикасын берілді. NLP-ні енді үйрене бастайтын адамдарға жиі қоятын бір сұрақ - жүйенің нәтижесі мәтін болған кезде жүйені қалай бағалау керек. BLEU (Bilingual Understudy Evaluation) - бұл бір табиғи тілден екінші табиғи тілге машиналық аудармаға аударылған мәтіннің сапасын бағалау алгоритмі. Ең алғашқы эксперимент 7500 сөйлемнен тұратын параллельді корпуспен жасалынды. Test2020-1604 сөйлем, Test2021-1600 сөйлем, Train-4296 сөйлем болды. Келесі кезектерде корпус саны артып, эксперименттер жасалынды. Ең соңғы Test2020-1604 сөйлем, Test2021-1600 сөйлем, Train-56796 сөйлеммен жүргізілді, және ең үлкен нәтиже берді. Best Bleu – 7.1 нәтиже көрсетті.

Қорытынды

Мақалада машиналық оқыту негізінде қазақ тіліндегі сөйлемдерді синтездеу әдісі зерттелді. Қазіргі кезде қолданыстағы синтездеу технологияларына зерттеулер жүргізілді, синтез әдісі анықтамасы берілді. Зерттеулер негізіне сүйене отырып, қазақ тіліндегі сөйлемдерді синтездеу әдісі таңдалды. Қазіргі уақытта қазақ тілінде 60000 сұрақ-жауаптан тұратын параллельді корпус жинақталды. Қазақ тіліндегі сөйлемдерді чат-бот негізінде синтездеу үшін модель құрылды. Seg2Seg моделін қолдана отырып , бірқатар тәжірибелер жүргізілді. Болашақта синтездеу нәтижесін жақсарту мақсатында корпусты ұлғайтуға және нәтижеге тікелей әсер ететін параметрлер көрсеткішін жақсарту мақсаты алға қойылып отыр.

Пайдаланылған әдебиеттер тізімі:

1. *Find, install and publish Python packages with the Python package index [electronic resource]: SciPy/ Python software foundation; Python community // URL: https://pypi.org/project/scipy/ , free access – Lang. En., Ru., Esp., Fr. Access date: 02.05.2020*
2. *Автоматический анализ и синтез текста.// [Electronic resource] - 2020. URL: http://csaa.ru/vtomaticheskij -analiz-i-sintez-teksta/*
3. *Development of a chatbot with natural language processing// [Electronic resource]- 2020. - URL: http://earchive.tpu.ru/bitstream/11683/60953/1/TPU927415.pdf*
4. *Machine Learning - Машинное обучение. // [Electronic resource] - 2020. - URL: https://www.it.ua/ru/ knowledge-base /technology-innovation/machine-learning*
5. *Нейросетевой синтез речи. [Electronic resource] - 2020. - URL: https://habr.com/ru/company/speechpro/blog/358816/*
6. *Тест Тьюринга. [Electronic resource] - 2020. - URL: http://window.edu.ru/window/%20catalog?p_ rubr= 2.2.73.12.15*
7. *Прикладная лингвистика: портал «Единое окно доступа к образовательным ресурсам». // [Electronic resource]–2020.-URL: http://window.edu.ru/window/%20catalog?p_rubr=2.2.73.12.15*
8. *Синтез текста. Программа Морфер. //[Electronic resource] - 2020. - URL: https://controleng.ru /innovatsii/sintez-teksta/*
9. *Natural Language Processing in Action. Understanding, analyzing, and generating text with Python. 2019. P 311-334*
10. *Thang Luong, Eugene Brevdo, Rui Zhao (2019) Neural Machine Translation (seq2seq)// URL: https://github.com/tensorflow/nmt*

References:

1. *Find, install and publish Python packages with the Python package index [electronic resource]: SciPy/ Python software foundation; Python community // URL: https://pypi.org/project/scipy/ , free access – Lang. En., Ru., Esp., Fr. Access date: 02.05.2020*
2. *Automatic text analysis and synthesis. (2020) [Electronic resource] - URL: http://csaa.ru/vtomaticheskij-analiz-i-sintez-teksta/*
3. *Development of a chatbot with natural language processing// [Electronic resource]- 2020. - URL: http://earchive.tpu.ru/bitstream/11683/60953/1/TPU927415.pdf*
4. *Machine Learning - Machine Learning. // [Electronic resource] - 2020. - URL: https://www.it.ua/ru/ knowledge-base/technology -innovation/machine-learning*
5. *Neural network synthesis of speech. [Electronic resource] - 2020. - URL: https://habr.com/ru/company/ speechpro/blog/358816/*
6. *Turing Test [Electronic resource] - 2020. - URL: http://window.edu.ru/window /%20catalog? p_rubr=2.2.73.12.15*
7. *Applied Linguistics: Portal "Single Window of Access to Educational Resources".(2020) [Electronic resource]– URL: http://window.edu.ru/window/%20catalog?p_rubr=2.2.73.12.15*
8. *Synthesis of text. Morfer program. (2020) [Electronic resource]. URL: https://controleng.ru/innovatsii/sintez-teksta/*
9. *Natural Language Processing in Action. Understanding, analyzing, and generating text with Python. 2019. P 311-334.*
10. *Thang Luong, Eugene Brevdo, Rui Zhao (2019) Neural Machine Translation (seq2seq)// URL: https://github.com/tensorflow/nmt*