ISSN 2518-1726 (Online),
ISSN 1991-346X (Print)

«ҚАЗАҚСТАН РЕСПУБЛИКАСЫ
ҰЛТТЫҚ ҒЫЛЫМ
АКАДЕМИЯСЫ» РҚБ

ҚАЙЫРЫМДЫЛЫҚ ҚОРЫ
HALYK
CHARITY FOUNDATION

«ҚАЗАҚСТАН РЕСПУБЛИКАСЫ
ҰЛТТЫҚ ҒЫЛЫМ АКАДЕМИЯСЫ» РҚБ
«ХАЛЫҚ» ЖҚ

ХАБАРЛАРЫ

ИЗВЕСТИЯ

РОО «НАЦИОНАЛЬНОЙ
АКАДЕМИИ НАУК РЕСПУБЛИКИ
КАЗАХСТАН»
ЧФ «Халық»

N E W S

OF THE ACADEMY OF SCIENCES
OF THE REPUBLIC OF
KAZAKHSTAN
«Halyk» Private Foundation

SERIES
PHYSICS AND INFORMATION TECHNOLOGY

3 (347)

JULY – SEPTEMBER 2023

PUBLISHED SINCE JANUARY 1963
PUBLISHED 4 TIMES A YEAR

ALMATY, NAS RK

ЧФ «ХАЛЫҚ»

В 2016 году для развития и улучшения качества жизни казахстанцев был создан частный Благотворительный фонд «Халык». За годы своей деятельности на реализацию благотворительных проектов в областях образования и науки, социальной защиты, культуры, здравоохранения и спорта, Фонд выделил более 45 миллиардов тенге.

Особое внимание Благотворительный фонд «Халык» уделяет образовательным программам, считая это направление одним из ключевых в своей деятельности. Оказывая поддержку отечественному образованию, Фонд вносит свой посильный вклад в развитие качественного образования в Казахстане. Тем самым способствуя росту числа людей, способных менять жизнь в стране к лучшему – профессионалов в различных сферах, потенциальных лидеров и «великих умов». Одной из значимых инициатив фонда «Халык» в образовательной сфере стал проект Ozgeris powered by Halyk Fund – первый в стране бизнес-инкубатор для учащихся 9-11 классов, который помогает развивать необходимые в современном мире предпринимательские навыки. Так, на содействие малому бизнесу школьников было выделено более 200 грантов. Для поддержки талантливых и мотивированных детей Фонд неоднократно выделял гранты на обучение в Международной школе «Мирас» и в Astana IT University, а также помог казахстанским школьникам принять участие в престижном конкурсе «USTEM Robotics» в США. Авторские работы в рамках проекта «Тәлімгер», которому Фонд оказал поддержку, легли в основу учебной программы, учебников и учебно-методических книг по предмету «Основы предпринимательства и бизнеса», преподаваемого в 10-11 классах казахстанских школ и колледжей.

Помимо помощи школьникам, учащимся колледжей и студентам Фонд считает важным внести свой вклад в повышение квалификации педагогов, совершенствование их знаний и навыков, поскольку именно они являются проводниками знаний будущих поколений казахстанцев. При поддержке Фонда «Халык» в южной столице был организован ежегодный городской конкурс педагогов «Almaty Digital Ustaz.

Важной инициативой стал реализуемый проект по обучению основам финансовой грамотности преподавателей из восьми областей Казахстана, что должно оказать существенное влияние на воспитание финансовой грамотности и предпринимательского мышления у нового поколения граждан страны.

Необходимую помощь Фонд «Халык» оказывает и тем, кто особенно остро в ней нуждается. В рамках социальной защиты населения активно проводится работа по поддержке детей, оставшихся без родителей, детей и взрослых из социально уязвимых слоев населения, людей с ограниченными возможностями, а также обеспечению нуждающихся социальным жильем, строительству социально важных объектов, таких как детские сады, детские площадки и физкультурно-оздоровительные комплексы.

В копилку добрых дел Фонда «Халык» можно добавить оказание помощи детскому спорту, куда относится поддержка в развитии детского футбола и карате в нашей стране. Жизненно важную помощь Благотворительный фонд «Халык» оказал нашим соотечественникам во время недавней пандемии COVID-19. Тогда, в разгар тяжелой борьбы с коронавирусной инфекцией Фонд выделил свыше 11 миллиардов тенге на приобретение необходимого медицинского оборудования и дорогостоящих медицинских препаратов, автомобилей скорой медицинской помощи и средств защиты, адресную материальную помощь социально уязвимым слоям населения и денежные выплаты медицинским работникам.

В 2023 году наряду с другими проектами, нацеленными на повышение благосостояния казахстанских граждан Фонд решил уделить особое внимание науке, поскольку она является частью общественной культуры, а уровень ее развития определяет уровень развития государства.

Поддержка Фондом выпуска журналов Национальной Академии наук Республики Казахстан, которые входят в международные фонды Scopus и Wos и в которых публикуются статьи отечественных ученых, докторантов и магистрантов, а также научных сотрудников высших учебных заведений и научно-исследовательских институтов нашей страны является не менее значимым вкладом Фонда в развитие казахстанского общества.

С уважением,
Благотворительный Фонд «Халык»!

«ҚР ҰҒА Хабарлары. Физика және информатика сериясы».
ISSN 2518-1726 (Online),
ISSN 1991-346X (Print)
Меншіктеуші: «Қазақстан Республикасының Ұлттық ғылым академиясы» РҚБ (Алматы қ.). Қазақстан Республикасының Ақпарат және қоғамдық даму министрлігінің Ақпарат комитетінде 14.02.2018 ж. берілген **No 16906-Ж** мерзімдік басылым тіркеуіне қойылу туралы куәлік.
Тақырыптық бағыты: *физика және ақпараттық коммуникациялық технологиялар сериясы.*
Қазіргі уақытта: *«ақпараттық технологиялар»бағыты бойынша ҚР БҒМ БҒСБК ұсынған журналдар тізіміне енді.*
Мерзімділігі: *жылына 4 рет.*
Тиражы: *300 дана.*
Редакцияның мекен-жайы: *050010, Алматы қ., Шевченко көш., 28, 219 бөл., тел.: 272-13-19*
http://www.physico-mathematical.kz/index.php/en/

«Известия НАН РК. Серия физика и информатики».
ISSN 2518-1726 (Online),
ISSN 1991-346X (Print)
Собственник: *Республиканское общественное объединение «Национальная академия наук Республики Казахстан» (г. Алматы).*
Свидетельство о постановке на учет периодического печатного издания в Комитете информации Министерства информации и общественного развития Республики Казахстан **No 16906-Ж** выданное 14.02.2018 г.
Тематическая направленность: *серия физика и информационные коммуникационные технологии.*
В настоящее время: *вошел в список журналов, рекомендованных ККСОН МОН РК по направлению «информационные коммуникационные технологии».*
Периодичность: *4 раз в год.*
Тираж: *300 экземпляров.*
Адрес редакции: *050010, г. Алматы, ул. Шевченко, 28, оф. 219, тел.: 272-13-19*
http://www.physico-mathematical.kz/index.php/en/

News of the National Academy of Sciences of the Republic of Kazakhstan.
Series of physics and informatics.
ISSN 2518-1726 (Online),
ISSN 1991-346X (Print)
Owner: RPA «National Academy of Sciences of the Republic of Kazakhstan» (Almaty). The certificate of registration of a periodical printed publication in the Committee of information of the Ministry of Information and Social Development of the Republic of Kazakhstan **No. 16906-Ж,** issued 14.02.2018
Thematic scope: *series physics and information technology.*
Currently: *included in the list of journals recommended by the CCSES MES RK in the direction of «information and communication technologies».*
Periodicity: *4 times a year.*
Circulation: *300 copies.*
Editorial address: *28, Shevchenko str., of. 219, Almaty, 050010, tel. 272-13-19*
http://www.physico-mathematical.kz/index.php/en/

NEWS OF THE NATIONAL ACADEMY OF SCIENCESOF THE REPUBLIC OF KAZAKHSTAN
PHYSICO-MATHEMATICAL SERIES
ISSN 1991-346X
Volume 3. Namber 347 (2023). 131–146
https://doi.org/10.32014/2023.2518-1726.209

UDC 004.89

© D. Oralbekova[1,2*], O. Mamyrbayev[1], A. Zhunussova[3], B. Zhumazhanov[1], 2023
[1]Institute of information and computational technologies, CS MSHE RK, Almaty, Kazakhstan;
[2]Almaty University of Power Engineering and Telecommunications named after Gumarbek Daukeyev, Almaty, Kazakhstan;
[3] Narxoz University, Almaty, Kazakhstan.
E-mail: dinaoral@mail.ru

STUDY OF MODERN METHODS OF LANGUAGE MODELING FOR A LANGUAGE WITH A COMPLEX MORPHOLOGICAL STRUCTURE

Oralbekova Dina — Doctor PhD. Senior Researcher, Associate Professor. Institute of information and computational technologies. Almaty, Kazakhstan
E-mail: dinaoral@mail.ru. ORCID ID: 0000-0003-4975-6493;
Мамырбаев Оркен — Doctor PhD. Associate Professor, Deputy General Director. Institute of information and computational technologies. Almaty, Kazakhstan
E-mail: morkenj@mail.ru. ORCID ID: 0000-0001-8318-3794;
Zhunussova Aliya — Senior Lecturer. Narxoz University, Almaty, Kazakhstan
E-mail: alia_94-22@mail.ru. ORCID ID: 0000-0002-3641-8260;
Zhumazhanov Bagashar — Candidate of technical sciences. Senior Researcher. Institute of information and computational technologies. Almaty, Kazakhstan
E-mail: bagasharj@mail.ru. ORCID ID: 0000-0002-5035-9076.

Abstract. This scientific paper presents a comparative analysis of contemporary language modeling methods and their application to the Kazakh language, which is characterized by its complex morphological structure. Language modeling involves training machine learning models to predict word probabilities within a given context. The primary focus of this study is the investigation of the BERT model (Bidirectional Encoder Representations from Transformers) and its effectiveness in modeling languages with diverse morphological patterns. The article provides an overview of n-gram models and recurrent neural networks, highlighting their limitations in capturing long-term dependencies and semantic relationships in text. Then the BERT model, its architecture and principles of operation, including attention mechanisms and multi-level Transformer blocks, are considered. The following are the results of the study, including the adaptation of the BERT model

to languages with a complex morphological structure, including Kazakh. It is shown that the BERT model demonstrates high accuracy in modeling contextual dependencies and semantic relationships between words in such languages. the article emphasizes the importance and prospects of applying modern methods of language modeling, especially the BERT model, for languages with a complex morphological structure. She also points out the need for further research in the field of adaptation of the BERT model to specific languages, the development of new architectures and methods, as well as solving the challenges associated with rare and sparsely spoken languages, the results of this study will help improve the understanding and effectiveness of language processing of text in the Kazakh language, and also contribute to the development of the NLP field as a whole.

Keywords: language modeling, Kazakh language, n-gram, BERT, GPT, LSTM

Financing: *This research has been/was/is funded by the Science Committee of the Ministry of Education and Science of the Republic of Kazakhstan (Grant No. AP19174298).*

Conflict of interest: *The authors declare that there is no conflict of interest.*

© Д. Оралбекова[1,2*], О. Мамырбаев[1], А. Жунусова[3], Б. Жумажанов[1], 2023
[1]Институт информационных и вычислительных технологий КН МНВО РК, Алматы, Казахстан;
[2]Алматинский Университет Энергетики и Связи им. Г. Даукеева, Алматы, Казахстан;
[3]Narxoz University, Алматы, Казахстан.
E-mail: dinaoral@mail.ru

КҮРДЕЛІ МОРФОЛОГИЯЛЫҚ ҚҰРЫЛЫМЫ БАР ТІЛГЕ АРНАЛҒАН ЗАМАНАУИ ТІЛДІК МОДЕЛЬДЕУ ӘДІСТЕРІН ЗЕРТТЕУ

Оралбекова Дина — PhD докторы. Аға ғылыми қызметкер, доцент. Ақпараттық және есептеуіш технологиялар институты. Алматы, Казахстан
E-mail: dinaoral@mail.ru. ORCID ID: 0000-0003-4975-6493;
Мамырбаев Оркен — PhD докторы. Қауымдастырылған профессор, бас директордың орынбасары. Ақпараттық және есептеуіш технологиялар институты. Алматы, Казахстан
E-mail: morkenj@mail.ru. ORCID ID: 0000-0001-8318-3794;
Жунусова Алия — Аға оқытушы. Нархоз университеті. Алматы, Қазақстан
E-mail: alia_94-22@mail.ru. ORCID ID: 0000-0002-3641-8260;
Жумажанов Багашар — Техникалық ғылымдардың кандидаты. Аға ғылыми қызметкер, доцент. Ақпараттық және есептеуіш технологиялар институты. Алматы, Казахстан
E-mail: bagasharj@mail.ru. ORCID ID: 0000-0002-5035-9076.

Аннотация. Бұл ғылыми мақала тілдік модельдеудің заманауи әдістерін салыстырмалы талдауды және оларды қазақ тілі сияқты агглютинативті тілге қолдануды ұсынады. Тілдік модель – бұл сөз бойынша ықтималдылықты бөлуді жүзеге асыруға үйретілген машиналық оқыту моделінің бір түрі. Тілдік

модель белгілі бір мәтіннің контекстіне сүйене отырып, сөйлемдегі немесе сөз тіркесіндегі бос орынды толтыру үшін келесі ең қолайлы сөзді болжауға тырысады. BERT (Bidirectional Encoder Representations from Transformers) моделін және оның морфологиялық әртүрлілігімен сипатталатын тілдерді тиімді модельдеу қабілетін зерттеуге баса назар аударылды. Мақалада n-грамм модельдеріне, қайталанатын нейрондық желілерге және олардың мәтіндегі ұзақ мерзімді тәуелділіктер мен семантикалық қатынастарды түсірудегі шектеулеріне шолу жасалды. Содан кейін BERT моделі, оның архитектурасы және жұмыс принциптері, соның ішінде назар аудару механизмдері және көп деңгейлі Transformer блоктары егжей-тегжейлі қарастырылды. Бұдан әрі BERT моделін қазақ тілін қоса алғанда, күрделі морфологиялық құрылымы бар тілдерге бейімдеуді қамтитын зерттеу нәтижелері ұсынылды. BERT моделі осындай тілдердегі сөздер арасындағы контекстік тәуелділіктер мен семантикалық қатынастарды модельдеуде жоғары дәлдікті көрсететіні анықталды. Мақалада күрделі морфологиялық құрылымы бар тілдер үшін заманауи тілдік модельдеу әдістерін, әсіресе BERT модельдерін қолданудың маңыздылығы мен перспективалары көрсетілген. Ол сондай-ақ BERT моделін нақты тілдерге бейімдеу, жаңа архитектуралар мен әдістерді әзірлеу, сондай-ақ сирек кездесетін тілдерге байланысты сын-қатерлерді шешу саласында одан әрі зерттеу қажеттігін көрсетеді. Осы зерттеудің алынған нәтижелері қазақ тіліндегі мәтінді тілдік өңдеудің түсінігі мен тиімділігін арттыруға өз үлесін қосады, сондай-ақ жалпы NLP саласының дамуына ықпал етеді.

Түйін сөздер: тілдік модельдеу, қазақ тілі, n-грамм, BERT, GPT, LSTM

Қаржыландыру: *Бұл зерттеу Қазақстан Республикасы Ғылым және жоғары білім министрлігі, Ғылым комитетімен қаржыландырған (Грант № AP19174298).*

Мүдделер қақтығысы: *Авторлар осы мақалада мүдделер қақтығысы жоқ деп мәлімдейді.*

© Д. Оралбекова[1,2*], О. Мамырбаев[1], А. Жунусова[3], Б. Жумажанов[1], 2023
[1]Институт информационных и вычислительных технологий КН МНВО РК, Алматы, Казахстан;
[2]Алматинский Университет Энергетики и Связи им. Г. Даукеева, Алматы, Казахстан;
[3]Narxoz University, Алматы, Казахстан.
E-mail: dinaoral@mail.ru

ИССЛЕДОВАНИЕ СОВРЕМЕННЫХ МЕТОДОВ ЯЗЫКОВОГО МОДЕЛИРОВАНИЯ ДЛЯ ЯЗЫКА СО СЛОЖНОЙ МОРФОЛОГИЧЕСКОЙ СТРУКТУРОЙ

Оралбекова Дина — Доктор PhD. Старший научный сотрудник, доцент. Институт информационных и вычислительных технологий. Алматы, Казахстан
E-mail: dinaoral@mail.ru. ORCID ID: 0000-0003-4975-6493;

Мамырбаев Оркен — Доктор PhD. Ассоциированный профессор, заместитель генерального директора. Институт информационных и вычислительных технологий, Алматы, Казахстан
E-mail: morkenj@mail.ru. ORCID ID: 0000-0001-8318-3794;
Жунусова Алия — Старший преподаватель. Университет Нархоз. Алматы, Казахстан
E-mail: alia_94-22@mail.ru. ORCID ID: 0000-0002-3641-8260;
Жумажанов Багашар — Кандидат технических наук. Старший научный сотрудник. Институт информационных и вычислительных технологий. Алматы, Казахстан
E-mail: bagasharj@mail.ru. ORCID ID: 0000-0002-5035-9076.

Аннотация. Данная научная статья представляет сравнительный анализ современных методов языкового моделирования и их применение к агглютинативному языку, такому как казахский язык. Языковая модель — это тип модели машинного обучения, призванной проводить распределение вероятностей по словам. Языковая модель пытается предсказать следующее наиболее подходящее слово для заполнения пробела в предложении или фразе, исходя из контекста определенного текста. Основной акцент сделан на изучение модели BERT (Bidirectional Encoder Representations from Transformers) и ее способности эффективно моделировать языки, характеризующиеся морфологической разнообразностью. В статье представлен обзор n-граммных моделей, рекуррентных нейронных сетей и их ограничений в улавливании долгосрочных зависимостей и семантических отношений в тексте. Затем подробно рассмотрена модель BERT, ее архитектура и принципы работы, включая механизмы внимания и многоуровневые Transformer блоки. Далее представлены результаты исследования, включающие адаптацию модели BERT к языкам со сложной морфологической структурой, включая казахский язык. Показано, что модель BERT демонстрирует высокую точность в моделировании контекстуальных зависимостей и семантических отношений между словами в таких языках. Статья подчеркивает значимость и перспективы применения современных методов языкового моделирования, особенно модели BERT, для языков со сложной морфологической структурой. Она также указывает на необходимость дальнейших исследований в области адаптации модели BERT к конкретным языкам, разработки новых архитектур и методов, а также решения вызовов, связанных с редкими и малораспространенными языками Полученные результаты этого исследования помогут улучшить понимание и эффективность языковой обработки текста на казахском языке, а также способствуют развитию области NLP в целом.

Ключевые слова: языковое моделирование, казахский язык, n-граммы, BERT, GPT, LSTM

Финансирование: *Данное исследование финансировалось Комитетом науки Министерства науки и высшего образования Республики Казахстан (Грант No AP19174298).*

Конфликт интересов: *Авторы заявляют об отсутствии конфликта интересов.*

Кіріспе

Тілдік модельдер табиғи тілді өңдеудің (NLP) негізгі компоненті болып табылады, өйткені олар машиналарға адам тілін түсінуге, құруға және талдауға мүмкіндік береді. Олар негізінен кітаптар немесе мақалалар жинағы сияқты мәтіннің үлкен жиынтығын қолдана отырып оқытылады. Содан кейін модельдер сөйлемдегі келесі сөзді болжау немесе грамматикалық тұрғыдан дұрыс және семантикалық тұрғыдан сәйкес келетін жаңа мәтін жасау үшін осы оқу деректерінен алатын үлгілерді пайдаланады.

Тіл моделін жасау табиғи тілді өңдеудегі (NLP) негізгі міндет болып табылады және көптеген қосымшаларда жаңа мәтін құру, машиналық аударма, Автоматты реферат және басқалар сияқты маңызды элементті ұсынады. Түркі тілдерінің тобына кіретін морфологиялық тұрғыдан күрделі тілдер жеткілікті, олар тиімді модельдеуде басқа талаптарды қажет етеді.

Агглютинативті тіл, қазақ тілі сияқты, грамматикалық тұлғаны, санды, уақытты, жағдайды және басқа морфологиялық сипаттамаларды ескере отырып, сөздер өзгеретін бай морфологиясы бар тілдің мысалы болып табылады. Бұл қазақ мәтінін өңдеу мен талдауда қиындықтар туғызады, сондай-ақ нәтижелердің жоғары сапасы мен дәлдігіне қол жеткізу үшін арнайы модельдеу әдістерін талап етеді.

Соңғы жылдары тілдік модельдеу саласында күрт ілгерілеу байқалады, бұл қазақ тілін қоса алғанда, күрделі тілдік құрылымдарды тиімді өңдеу үшін жаңа мүмкіндіктер ашты. Ең сәтті және кеңінен қолданылатын тәсілдердің бірі — мәтіндегі сөздер тізбегін статистикалық талдауға негізделген n-грамм модельдерін қолдану (Браун, 1992). Бұл модельдер алдыңғы сөздерге негізделген келесі сөздің контексті мен ықтималдығын ескеруге мүмкіндік береді (Хайруллина, 2018).

Алайда, n-грамм модельдерінің кемшілігі – олардың сөздер арасындағы ұзақ мерзімді тәуелділіктер мен семантикалық қатынастарды байланыстыру қабілетінің шектеулі болуы (Поречный, 2020). Соңғы жылдары терең оқыту тілдік модельдеуде үлкен жетістік болды, бұл күрделі және қуатты модельдер жасауға мүмкіндік берді. BERT (Bidirectional Encoder Representations from Transformers) моделі ерекше көзге түседі, ол Transformer архитектурасына негізделген және мәтіндегі ұзақ мерзімді тәуелділіктер мен контекстік ерекшеліктерді анықтай алады (Салып, 2022). BERT мәтіндік деректердің үлкен көлемінде өзін-өзі оқыту механизмін қолданады, бұл модельге әртүрлі тілдік құрылымдар мен олардың контекстік ерекшеліктерімен байытуға мүмкіндік береді. Бұл BERT-ті күрделі морфологиялық құрылымы бар тілдер үшін әсіресе пайдалы етеді. BERT тілдік модельдеу нәтижелерін айтарлықтай жақсартқанымен, оның шектеулері бар. Ең алдымен, оны оқыту үлкен есептеу ресурстары мен уақытты қажет етеді. Сонымен қатар, BERT сөйлемдегі барлық сөздер бір-бірінен тәуелсіз деп болжайды, бұл сөздер бір-бірімен тығыз байланысатын бай морфологиясы бар тілдер үшін шектеу болуы мүмкін.

Тағы бір перспективалық бағыт – GPT сияқты генеративті модельдерді тілдік модельдеуге қолдану (Частикова, 2022). GPT сонымен қатар Transformer архитектурасына негізделген және оның контексті мен семантикалық қатынастары мен мағыналарын ескере отырып, мәтін құруға қабілетті. Бұл модель шектеулі мәтіндік деректері бар тілдерде мәтінді талдау мен құрудың тиімді құралы бола алады.

Жұмыстың мақсаты — тілдік модельдеу әдістеріне шолу және салыстырмалы талдау, сонымен қатар оларды қазақ тілі сияқты күрделі морфологиялық құрылымы бар тілге қолдану. Сонымен қатар, әр әдістің артықшылықтары мен шектеулері келтірілген.

Мақала келесідей ұйымдастырылған: 2-бөлімде тілдік модельдеу саласындағы қолданыстағы жұмыстарға шолу жасалды және оларды күрделі морфологиялық құрылымы бар тілдерге қолдану келтірілген. 3-бөлім тілдік модельдеу әдістерін, соның ішінде n-грамм модельдерін, BERT, GPT және басқаларын, олардың негізгі жұмыс принциптері мен артықшылықтарын егжей-тегжейлі сипаттауға арналған. 4-бөлімде тәжірибелер мен зерттеулердің нәтижелері жасалған, ал қорытындыда зерттеу нәтижелері мен осы бағыттағы болашақ жұмыстардың қорытындысы келтірілген.

Әдістер мен материалдар

n-грамм моделі

n-грамм моделі — бұл n-1 сөзден кейінгі келесі сөзді олардың жұптасу ықтималдығына негізделген болжайтын статистикалық модельдер. Мәселен, қазақ тіліндегі «Мен Алматыға бара жатырмын» біріктірудің ықтималдығы жоғары, ал «Мен бара Алматыға жатырмын» біріктірудің ықтималдығы төмен. Қарапайым тілмен айтқанда, n-грамм – n сөздердің тізбегі. Мысалы, биграммалар – екі сөзден тұратын тізбектер (Мен Алматыға, Алматыға бара, баа жатырмын), триграммалар – үш сөзден тұратын тізбектер (мен Алматыға бара, Алматыға баа жатырмын) және т.б.

Мұндай ықтималдық үлестірімдері машиналық аудармада, орфографияны автоматты түрде тексеруде, сөйлеуді тануда және автоматты түрде енгізуде кеңінен қолданылады. Барлық жағдайларда біз келесі сөздің немесе сөз тізбегінің ықтималдығын есептейміз. Мұндай есептеулер тілдік модельдер деп аталады.

Сөйлемдегі сөздер санына байланысты *P(w)* есептеудің жалпы формуласы келесі түрде болады (1):

$$P(w_1, \dots, w_n) = P(w_n | w_{n-1}, \dots, w_1) \cdot P(w_{n-1}, \dots, w_1) \qquad (1)$$

мұнда *w* – жеке сөздер, *n* – сөздер саны.

Осылайша, шартты ықтималдықтарды көбейту арқылы бүкіл тізбектің бірлескен ықтималдығын бағалай аламыз. Алайда, алдыңғы сөздердің ұзақ тізбегі жағдайында сөздің нақты ықтималдығын есептей алмаймыз, өйткені мүмкін болатын тізбектер өте көп және деректерімізде бұл тізбектер

болмауы мүмкін. Сондықтан, барлық алдыңғы сөздерді ескере отырып, сөздің ықтималдығын есептеудің орнына, біз оны жеңілдету арқылы ықтималдылықты жуықтай аламыз. Бұл Марков тізбектерінің негізі, оның көмегімен біз тым кең контекстті ескермей, реттілік элементінің ықтималдығын болжай аламыз (2):

$$P(w_n|w_{n-1}, \dots, w_1) \approx P(w_n|w_{n-1}, \dots, w_{n-k}) \quad (2)$$

мұнда k – Марков тізбегінің реті.

Бірінші ретті Марков тізбегін қолдана отырып, кез келген сөз тізбегінің ықтималдығын оңай есептеуге болады. Осылайша, біз биграммаларды, триграммаларды, квадрограммаларды және т.б. есептей аламыз, ал тізбек неғұрлым ұзағырақ болса, біздің модель соғұрлым егжей-тегжейлі болады, яғни ұзын сөйлемдер қысқа сөйлемдерге қарағанда көбірек грамматиканы қамтиды.

n-грамм моделі салыстырмалы түрде қарапайым және тиімді, бірақ олар дәйектіліктегі сөздердің ұзақ мерзімді контекстін ескермейді.

Ілеспе жұмыстар

Жұмыс авторлары (Си, 2021) n-грамм модельдерінің кемшіліктерін едәуір азайтқан n-distant-max тілдік моделін ұсынды. Контекстегі орфография қателерін түзету жүйесін бағалау кезінде салыстырмалы талдау жүргізілгеннен кейін, әзірленген модель n-грамм тілінің классикалық моделімен салыстырғанда осы жүйенің тиімділігін едәуір арттырғаны көрсетілді.

C. Chelba және басқалары (Челба, 2017) n-грамм модельдеу үшін RNN модельдерінің жадының қасиетін қолдана отырып тиімді тереңдігін зерттеді. Шағын UPenn Treebank корпустағы тәжірибелер LSTM ұяшығы RNN тікелей байланысымен салыстырғанда n-gram күй деректерін кодтау үшін жақсы модель екенін көрсетті. Сөйлемнің тәуелсіздігі туралы болжамды сақтай отырып, LSTM n-граммасы n=9 үшін LSTM LM өнімділігіне сәйкес келді және n=13 кезінде одан сәл асып түсті.

Зерттеушілер (Ким, 2022) агглютинативті тілдер тобына кіретін корей тілін модельдеуді, атап айтқанда, тілді бейнелеу әдістері мен алдын ала оқыту әдістерін зерттеді. Авторлар кеңінен қолданылатын Transformer архитектурасы мен екі жақты тіл көрінісіне негізделген корей тілін фрагментті түрде қайта құруды ұсынады. Олар алдын-ала оқыту кезінде осындай ақпаратты қолдана отырып, тілді түсінуге сөйлеу бөлігі (сөйлеу бөлігі, PoS) сияқты морфологиялық ерекшеліктерді енгізді. Жұмыстың алынған нәтижелері ұсынылған әдістер корей тілін түсінудің зерттелетін міндеттерінің модельдік өнімділігін жақсартатынын көрсетті.

Myrzakhmetov (Мырзахметов, 2018) қазақ тіліне арналған кеңейтілген тілдік модельдеу экспериментінде веб-газеттердегі, сондай-ақ қазақ тіліндегі басқа сайттардағы әртүрлі мақалалардан тұратын қазақ тіліне арналған корпус ұсынды. Дәстүрлі n-грамм модельдерімен бірге олар вербальды тіл моделі

(LM) үшін нейрондық желі модельдерін жасады. Үлкен параметрленген ұзақ қысқа мерзімді жад (LSTM) моделі ең жақсы өнімділікті көрсетті. Сонымен қатар, зерттеушілер LM негізіндегі морфемаларды пайдаланды. Эксперименттер субсөзге негізделген LM қазақ тілі үшін сөз негізіндегі LM-мен салыстырғанда n-грамм және нейрондық желі модельдерінде жақсы жұмыс істейтінін көрсетті.

BERT

BERT – бұл Google-дің нейрондық желісі, ол көптеген міндеттер бойынша state-of-the-art нәтижелерін үлкен айырмашылықпен көрсетті. BERT көмегімен табиғи тілді өңдеуге арналған жасанды интеллект (AI) бағдарламаларын жасауға болады: еркін түрде қойылған сұрақтарға жауап беру, чатботтар, автоматты аудармашылар құру, мәтінді талдау және т.б. (Девлин, 2018).

Тілдік модельдеуге арналған бірнеше назар аудару механизмдері бар BERT моделіне екі бағытты оқытуы бар Transformer архитектурасы енгізілді (Васвани, 2017). Тілдік модель солдан оңға және оңнан солға қарай екі бағытта оқыту кезінде сөйлемнің контексті мен семантикасын толығымен қабылдайды.

BERT контекстен жасырылған сөзді болжауға үйретілген (1-сурет) және екі сөйлемнің дәйекті немесе сәйкестігін жіктеуге анықтауға жасалған. Кіріс деректері алдымен векторларға енгізілген, содан кейін нейрондық желіде өңделетін токендер тізбегі болып табылады. Шығару векторлар тізбегі болып табылады, онда әрбір вектор бірдей индексі бар кіріс лексемасына сәйкес келеді. BERT-ке сөздер тізбегін бермес бұрын, әрбір тізбектегі сөздердің 10%-ы [маска] белгісімен ауыстырылады. Содан кейін модель басқа, жасырын емес сөздермен берілген контекст негізінде бүркеніш сөздердің бастапқы мағынасын болжауға тырысады.

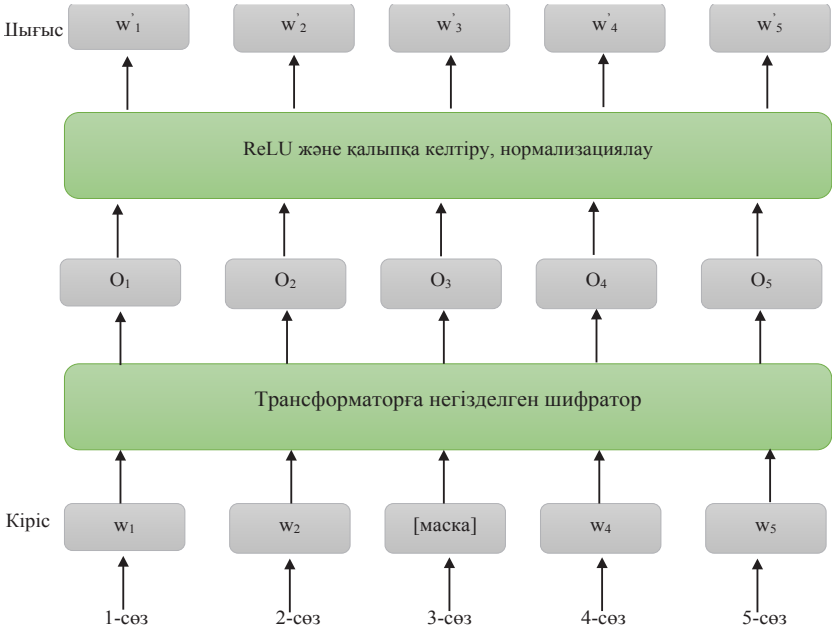
BERT GLUE тестіндегі он бір NLP тапсырмасы бойынша алдыңғы заманауи модельдерден едәуір асып түседі. Бұл керемет нәтиже BERT тіл туралы құрылымдық ақпаратты жақсы байланыстыра алады деп болжайды. Маңызды даму қадамы Голдберг болып табылады, ол BERT субъекті-етістік келісімін қадағалау қабілетін бағалау кезінде синтаксистік құбылыстарды жақсы сәйкестендіретіндігін көрсетеді.

Ілеспе жұмыстар

Авторлар (Чандра, 2021) 2020 жылы АҚШ президенттік сайлауы қарсаңында Twitter-дегі көңіл-күйді талдау үшін LSTM және BERT тілдік модельдерін қолданды. Олардың зерттеулері әлеуметтік медианы талқылау сайлау кезінде көпшіліктің мінез-құлқы мен көзқарасын түсінуге көмектесетінін анықтады. Президенттік сайлауға байланысты шамамен 1,2 миллион твиттер талданды. Модельдеу мен талдаудан кейін көңіл-күйді талдау сайлау нәтижелерін модельдеудің жалпы негізі бола алатынын анықтады. BERT моделі Байденнің сайлау науқандары кезінде Twitter-ге сүйене отырып жеңіске жету мүмкіндігі жоғары екенін көрсетті. Олар BERT моделі Трампты, Байденді және даулы мемлекеттерді анықтауда дәл болды деп есептеді. Демек, көбірек деректер

мен географиялық ақпаратпен көңіл-күйді талдау сайлау нәтижелерін болжау үшін пайдалы болу мүмкіндігін дәлелдеді.

Ganesh және басқалар (Ганеш, 2019) BERT фразалық көрінісі ақпаратты төменгі қабаттардағы фразалар деңгейінде түсіретінін, сондай-ақ BERT аралық қабаттары лингвистикалық ақпараттың бай иерархиясын төменгі жағында беттік функциялармен, ортасында синтаксистік функциялармен және жоғарғы жағында семантикалық функциялармен кодтайтынын көрсетті. Мысалы, алыс қашықтықтағы тәуелділіктер туралы ақпарат қажет болған кезде BERT тереңірек деңгейлерді қажет етті. Сонымен қатар, зерттеушілер BERT лингвистикалық ақпаратты классикалық ағаш құрылымдарына еліктейтін композициялық түрде жинайтынын анықтады.



Сурет 1 – BERT жалпы архитектурасы

Авторлар (Ли, 2020) корей KR-BERT моделін кішірек сөздік пен деректер жиынтығын пайдалана отырып оқытты. Корей тілі латын графикасын пайдаланбайтын морфологиялық ресурсы аз тілдердің бірі болғандықтан, авторлар көптілді BERT моделі жіберіп алған тілге тән тілдік құбылыстарды түсірудің маңыздылығын атап өтті. Олар бірнеше токенизаторларды, соның ішінде әзірленген WordPiece қос бағытты токенизаторын сынады және модельдері үшін жақсы сөздік қорын жасау үшін ішкі таңбадан таңба деңгейіне дейін токендеу үшін ең аз диапазонды реттеді. Осы түзетулермен әзірленген KR-BERT моделі шамамен 1/10 өлшемді корпусты пайдаланатын бұрыннан белгілі басқа модельдерге қарағанда салыстырмалы және тіпті жақсырақ орындады.

Контексте болжам жасау үшін тілдік модельдер қолданатын ақпарат туралы мақсатты сұрақтар қоюға мүмкіндік беретін табиғи тілдің эксперименттеріне негізделген диагностика жиынтығын ұсынылды (Эттингер, 2020). Мысал ретінде олар диагнозды BERT моделіне қолданды және ол әдетте адамдарға қарағанда сезімталдығы төмен болса да, жалпы санатқа немесе рөлдердің өзгеруіне байланысты жақсы және жаман аяқталуларды ажырата алатынын және зат есімдердің гипернимдерін сенімді түрде шығаратынын анықтады, бірақ рөлдерге негізделген оқиғалар туралы қорытындылар мен болжамдармен күреседі және сынайды — және, атап айтқанда, теріске шығарудың контекстік әсеріне айқын сезімталдықты көрсетеді. Осылайша, бұл жұмыста BERT моделінің негізгі кемшіліктері анықталды.

Зерттеу жұмыстың авторлары (Ли, 2020) алдын ала дайындалған BERT моделін дәл баптауға және оны патенттік жіктеуге қолдануға бағытталған. Екі миллионнан астам патенттері бар үлкен деректер жиынтығына қатысты ұсынылған тәсіл CNN сөз ендірілген тәсілінің арқасында заманауи тәсілден асып түседі. Осылайша келесі жақсартулар алынды: 1) алдын-ала дайындалған BERT моделіне және патенттерді жіктеуге арналған дәл баптауға негізделген жаңа заманауи нәтиже, 2) патенттік талаптардың өзі жіктеу тапсырмасы үшін қазіргі заманғы нәтижелерге қол жеткізу үшін жеткілікті екенін көрсету.

GPT

OpenAI GPT модельдерінің арқасында табиғи тілді өңдеуде айтарлықтай жетістіктерге жетті. GPT-1-ден GPT-4-ке дейін бұл модельдер проза мен поэзияны құрудан чатботтарға, тіпті кодтауға дейін AI құрған мазмұнның алдыңғы қатарында болды.

Генеративті алдын-ала дайындалған түрлендіргіштер (GPT) — бұл табиғи тілді өңдеу тапсырмалары үшін қолданылатын машиналық оқыту моделінің түрі. Бұл модельдер контекстке сәйкес және семантикалық тұрғыдан біртұтас тіл құру үшін кітаптар мен веб-беттер сияқты үлкен көлемде алдын ала оқытылады.

Басқаша айтқанда, GPT – бұл адам тәрізді мәтінді нақты бағдарламалаусыз жасай алатын компьютерлік бағдарламалар. Нәтижесінде оларды сұрақтарға жауап беру, тілдік аударма және мәтінді жалпылауды қоса алғанда, табиғи тілді өңдеудің бірқатар мәселелерін шешу үшін дәл реттеуге болады. Сонымен қатар, бұл модель табиғи тілді өңдеудегі үлкен жетістік, бұл машиналарға тілді бұрын-соңды болмаған еркін және дәлдікпен түсінуге мүмкіндік береді.

GPT-1. GPT-1 OpenAI Transformer архитектурасын қолдана отырып, тілдік модельдің алғашқы итерациясы ретінде шығарылды. Оның 117 миллион параметрі бар, бұл алдыңғы заманауи тіл модельдерін айтарлықтай жақсартты. Осы модельдің артықшылығы — оның анықтамасы немесе контексті болған кезде тегіс және біртұтас тіл жасау қабілеті болды. Модель екі деректер жиынтығын біріктіруге үйретілді: Common Crawl, миллиардтаған сөздері бар веб-беттердің үлкен деректер жинағы және bookcorpus деректер жинағы, әртүрлі жанрдағы 11 000-нан астам кітаптар жинағы. Осы әртүрлі деректер

жиынтығын пайдалану GPT-1-ге тілдік модельдеудің күшті қабілеттерін дамытуға мүмкіндік берді.

GPT-1 табиғи тілді өңдеудегі (NLP) маңызды жетістік болғанымен, оның белгілі бір шектеулері бар: модель қайталанатын мәтінді жасауға бейім болды, бұл оған оқыту деректерінен тыс ақпарат берілген кезде айқын көрінді. Ол сондай-ақ мәтіндегі ұзақ мерзімді тәуелділіктерді бақылай алмайды. Сонымен қатар, оның үйлесімділігі мен еркіндігі тек қысқа мәтіндік тізбектермен шектелді, ал ұзағырақ үзінділерде байланыс болмады.

Осы шектеулерге қарамастан, GPT-1 Transformer архитектурасына негізделген үлкен және қуатты модельдердің негізін қалады.

GPT-2. Оның құрамында 1,5 миллиард параметр болды. Модель Common Crawl және Web Text біріктіретін әлдеқайда үлкен және әртүрлі деректер жиынтығында оқытылды. GPT-2-нің күшті жақтарының бірі – оның мәтіннің дәйекті және шынайы тізбегін құру қабілеті. Сонымен қатар, ол адам сияқты жауаптар жасай алады, бұл оны мазмұнды құру және аудару сияқты табиғи тілді өңдеудің әртүрлі тапсырмалары үшін құнды құралға айналдырады. Алайда, GPT-2-нің кемшіліктері бар. Ол күрделі ойлау мен контекстті түсінуді қажет ететін міндеттермен басқара алмайды. GPT2 мәтіннің қысқа абзацтары мен үзінділерінде жақсы жұмыс істегенімен, ол ұзағырақ үзінділерде контекст пен келісімді сақтай алмайды.

GPT-3. GPT-3 BookCorpus, Common Crawl және Wikipedia сияқты әртүрлі деректер көздерінде оқытылады. Деректер жиынтығында триллионға жуық сөз бар, бұл GPT-3-ке NLP тапсырмаларының кең спектріне күрделі жауаптар жасауға мүмкіндік береді. Алдыңғы модельдермен салыстырғанда GPT-3-тің негізгі жақсартуларының бірі –үйлесімді мәтін құру, компьютерлік код жазу, тіпті өнер туындыларын жасау мүмкіндігі. Алдыңғы модельдерден айырмашылығы, GPT-3 берілген мәтіннің мәнмәтінін түсінеді және тиісті жауаптар генерациялайды. Табиғи дыбыстық мәтін құру мүмкіндігі чатботтар, мазмұн жасау және тілдік аударма сияқты қосымшалар үшін өте маңызды. Осындай мысалдардың бірі – ChatGPT, жасанды интеллект диалогтық боты.

GPT-3-де дегенмен өз кемшіліктері жеткілікті. Модель біржақты, дәл емес немесе орынсыз жауаптарды қайтара алады. Бұл мәселе GPT-3 жалған ақпаратты қамтуы мүмкін мәтіннің көп мөлшерінде оқытылатындықтан туындайды. Сондай-ақ, модель контекст пен фондық білімді түсінуде әлі де қиындықтарға тап болғанын көрсетеді және кейде мүлдем маңызды емес мәтінді жасайтын кездері болады.

GPT-4. GMT-4 Тек ChatGPT Plus пайдаланушыларына арналған, бірақ пайдалану шегі шектеулі. GPT-4-тің көрнекті ерекшелігі – оның мультимодальды мүмкіндіктері. Яғни модель енді кескінді кіріс ретінде қабылдай алады және оны мәтіндік формат ретінде түсінеді. Модель сонымен қатар күрделі мәтіндерді жақсы түсінеді, бірнеше кәсіби және дәстүрлі сынақтарда адам деңгейіндегі өнімділікті көрсетеді.

MT-NLG. MT-NLG (Megatron-Turing Natural Language Generation) — бұл

Transformer архитектурасына негізделген қуатты және жетілдірілген тілдік модель. Ол табиғи тілде көптеген тапсырмаларды орындай алады, соның ішінде логикалық тұжырымдар мен оқуды түсіну. Бұл Microsoft және Nvidia әзірлеген тілдік модельдердің соңғы нұсқасы және сөйлемдерді автоматты түрде аяқтау, ақылға қонымды пайымдауды түсіну және оқуды түсіну сияқты көптеген нәрселерді жасай алады. Модель көптеген деректерде оқытылды, атап айтқанда ағылшын тілді веб-сайттардан барлығы 339 миллиард токендерден (сөзден) тұратын 15 деректер жиынтығы. MT-NLG – бұл жаңадан жасалған модель, сондықтан ол үшін нақты пайдалану жағдайлары әлі де толығымен зерттелмеген. Дегенмен, модельді жасаушылар бұл табиғи тілді өңдеу технологиялары мен өнімдерінің болашағын анықтай алады деп болжады.

LaMDA. LaMDA – Google әзірлеген диалогтық қосымшаларға арналған тілдік модель. Ол ауызша диалогты еркін түрде құруға арналған, бұл оны әдетте тапсырмаларға негізделген дәстүрлі модельдерге қарағанда табиғи және егжей-тегжейлі етеді. LaMDA 137 миллиард параметрлері бар диалогтық деректерде оқытылды. Бұл оған ашық әңгіменің нюанстарын түсінуге мүмкіндік береді. Google бұл модельді іздеу, Google Assistant және Workspace сияқты өнімдерінде қолдануды жоспарлап отыр. MT-SNG сияқты, бұл жаңадан жасалған модель, сондықтан осы модельді қолдану бойынша зерттеулер өте аз.

Тілдік модельдеудің тағы бір маңызды әдісі – ұзақ, қысқа мерзімді жады бар қайталанатын нейрондық желілерді пайдалану (LSTM - Long Short-Term Memory). LSTM-мәтіндегі ұзақ мерзімді тәуелділіктерді тиімді модельдеуге қабілетті нейрондық желі архитектурасы. LSTM ақпаратты ұзақ уақыт бойы сақтауға және жаңартуға мүмкіндік беретін арнайы жад механизмін пайдаланады []. Бұл әсіресе морфологиялық өзгерістер мен тәуелділіктер айтарлықтай қашықтыққа ие болуы мүмкін мәтіндерді өңдеу кезінде пайдалы. LSTM-нің басты артықшылықтарының бірі – оның ұзақ мерзімді тәуелділіктерді модельдеу және алдыңғы контексттердегі ақпаратты есте сақтау қабілеті. Бұл модельге мәтіндегі келесі сөзді оның контексті мен семантикасын ескере отырып дәлірек болжауға мүмкіндік береді. LSTM модельдер әртүрлі тілдік құрылымдар мен морфологиялық ерекшеліктерді алуға мүмкіндік беретін мәтіндік деректердің үлкен көлемінде оқытылуы мүмкін.

Нәтижелер және оларды талқылау

Корпус және тәжірибелер. Сөз тізбегінің ықтималдығын болжауға арналған нейрондық тіл модельдері әдетте мәтіндік деректердің үлкен корпусында оқытылады және тілдің негізгі құрылымын үйренуге қабілетті. Ол үшін қазақ тіліндегі электронды көркем кітаптарды жинау туралы шешім қабылданды. tilalemi.kz сайтынан қазақ тілінде *txt* форматында 30 электронды кітап жүктелді. Модельді оқыту үшін сәйкесінше 24 кітап, ал тексеру және тестілеу үшін 3 кітаптан ақпарата қолданылды (кесте 1).

1 кесте. Модельді оқыту үшін мәтіндік корпусты бөлу

Дереккөз	tilalemi.kz		
	Кітаптар саны	Сөйлемдер саны	Сөз саны
Оқытуға арналған жиынтығы	24	131369	1434149
Валидацияға арналған жиынтық	3	22637	238601
Тестілеуге арналған жиынтық	3	8168	89853

Деректерді алдын ала өңдеу

Модельдерді оқытпас бұрын деректерді алдын ала өңдеу сөйлемдердің басталуы мен аяқталуын токендеу, тазарту, өңдеу және тегістеу әдістерін қолдану сияқты бірнеше қадамдарды қамтиды.

Токенизация — мәтінді жеке токендерге немесе сөздерге бөлу процесі. n-грамм моделі жағдайында әр сөз жеке токенге айналады. Токенизация *nltk* табиғи тілді өңдеу кітапханасы арқылы жүзеге асырылды.

Токенизациядан кейін деректерді қажетсіз таңбалардан және арнайы таңбалардан тазарту жүргізілді. Сондай-ақ, модель үшін семантикалық жүктемесі жоқ барлық тыныс белгілері, сандар мен таңбалар алынып тасталды.

BERT моделі үшін келесі параметрлер анықталды: 5 эпоха саны, пакеттің өлшемі 8, оңтайландырғыш ретінде AdamW алгоритмі белгіленді. Модельдерді қайта өңдеуден және оқытудан кейін тест жиынтығында нәтижелер алынды және 2-кестеде келтірілген (сурет. 2).

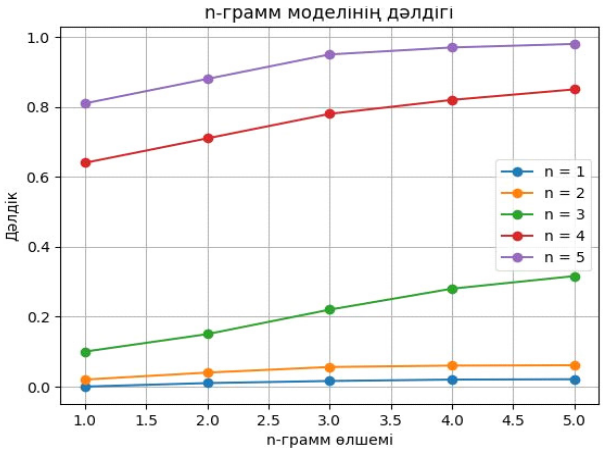
2 кесте. n-грамм және BERT модельдерінің нәтижелері

Модельдер	Болжау дәлдігі, %
n-грамм моделі	
n = 1	1,6
n = 2	5,6
n = 3	31
n = 4	84,6
n = 5	97,7
BERT моделі	98,9

Алынған нәтижелерді талқылау

Эксперименттік тапсырмаларды орындау барысында n-грамм және BERT модельдері құрылды және оларды қазақ тіліне арналған келесі сөзді болжау жұмыстары іске асырылды. Екі модель де токенизацияны қолданады, ал BERT токенизатор жаттығу және сынақ датасеттерін токенизациялау үшін қолданылады, содан кейін оларды жаттығу кезінде қолданылатын PyTorch форматына түрлендіреді.

Тек n = 4 және n = 5 кезінде модель максимумға жетеді, бірақ n-грамм санының ұлғаюы сонымен қатар үлкен есептеу ресурстарын қажет етті (сур.2), ал модельді оқыту мен тестілеу BERT моделімен салыстырғанда көп уақытты қажет етпеді.



Сурет 2 – n-грамм моделін болжау дәлдігі

BERT моделі бойынша нәтиже алу үшін 2 күндей уақыт кетті, бірақ нәтижесі жақсы болып шықты, n-грамм моделін 1,2 %-ға дәлірек болжады.

n-грамм тіл модельдерінен айырмашылығы, BERT контекстке сезімтал көріністерді үйретеді. Мәтінмәнді ескеретін модельдер әдетте токеннің сол немесе оң контекстін ғана есепке алуға мүмкіндік береді. BERT, керісінше, екі жақты контекстті ескереді, бұл модельге көп мағыналы сөздердің мағынасын жақсы түсінуге көмектеседі.

Қорытынды

Бұл ғылыми мақалада тілдік модельдеудің заманауи әдістері және олардың күрделі морфологиялық құрылымы бар тілдерге қолданылуы зерттелді. Негізгі фокус n-грамм, BERT модельдерін және олардың агглютинативті сипаттағы тілдерді тиімді модельдеу қабілетін зерттеу болды.

Зерттеу нәтижелері BERT моделі қазақ тіліне арналған сөздер арасындағы контекстік тәуелділіктер мен семантикалық қатынастарды түсінуде жоғары дәлдікті көрсететінін анықтады. Оның контекстті екі бағытта модельдеу және назар аудару механизмдерін қолдану қабілеті мәтіндегі ұзақ мерзімді тәуелділіктер мен күрделі қатынастарды түсінуге мүмкіндік береді. Зерттеудің бұл нәтижелері күрделі морфологиялық құрылымы бар тілдер үшін заманауи тілдік модельдеу әдістерін, әсіресе BERT модельдерін қолданудың маңыздылығы мен перспективаларын растайды. Мұндай модельдер мәтінді автоматты түрде өңдеу сапасын едәуір жақсарта алады, табиғи тілді өңдеудің әртүрлі міндеттерінде дәлдік пен сенімділікті арттырады.

Алдағы жұмыстарда NLP-дің басқа да перспективалық әдістері негізінде түрлі эксперименттік жұмыстар жүзеге асырылатын болады.

ӘДЕБИЕТТЕР

Васвани и др. (2017). Внимание — это все, что вам нужно. В материалах 31-й конференции по нейронным системам обработки информации (NeurIPS 2017), Лонг-Бич, Калифорния, США, 2017 (in Eng.)

Ганеш Джавахар, Бенуа Саго, Джаме Седда. (2019). Что BERT узнает о структуре языка? ACL 2019 — 57-е ежегодное собрание Ассоциации компьютерной лингвистики, июль 2019 г., Флоренция, Италия. (in Eng.)

Девлин Джейкоб, Минг-Вэй Чанг, Кентон Ли и Кристина Тутанова. (2019). BERT: Предварительная подготовка глубоких двунаправленных преобразователей для понимания языка. В материалах конференции Североамериканского отделения Ассоциации компьютерной лингвистики 2019 года: технологии человеческого языка, Т. 1 (Длинные и короткие статьи), стр. 4171–4186, Миннеаполис, Миннесота. (in Eng.)

Ли Санга, Чан Хансоль, Байк Юнми, Пак Сюзи, Шин Хёпиль. (2020). Малая языковая модель BERT, специфичная для корейского языка. Журнал KIISE. 47. 682–692. 10.5626/JOK.2020.47.7.682 (in Eng.)

Мырзахметов Б. и Кожирбаев З. (2018). Расширенные эксперименты по языковому моделированию казахского языка. Материалы семинара CEUR, 2303 (in Eng.)

Браун Питер Ф., Винсент Дж. Делла Пьетра, Питер В. де Соуза, Дженифер К. Лай, Роберт Л. Мерсер. (1992). n-граммные модели естественного языка на основе классов, Компьютерная лингвистика, Том. 18, стр. 467–479. (in Eng.)

Поречный А.С. (2020). Создание инструмента семантико-синтаксического анализа текстов на русском языке. Современные информационные технологии и ИТ-образование, DOI: 10.17308/sait.2020.1/2630.

Салып Б., Смирнов А. (2022). Анализ модели BERT как инструмента определения меры смысловой близости предложений естественного языка. StudNet, №5, URL: https://cyberleninka.ru/article/n/analiz-modeli-bert-kak-instrumenta-opredeleniya-mery-smyslovoy-blizosti-predlozheniy-estestvennogo-yazyka.

Си Луссейн Аурага, Абделла Юсфиб, Саида Лааруссик, Хишам Гуеддахд, Мохаммед Неджа. (2021). Новая оценка языковой модели n-грамм. В материалах 5-й Международной конференции по искусственному интеллекту в компьютерной лингвистике, стр. 211–215.DOI: 10.1016/j.procs.2021.05.111 (in Eng.)

Ким Х., Ким С., Кан И., Квак Н., Фунг П. (2022). Моделирование корейского языка с помощью синтаксического руководства. В материалах тринадцатой конференции по языковым ресурсам и оценке, стр. 2841–2849, Марсель, Франция. Европейская ассоциация языковых ресурсов (in Eng.)

Хайруллина Р.Х., Рахимова Э.Ф., Сагитова А.Ф. (2018). Лингвокогнитивные основы языкового моделирования. Мир науки, культуры, образования, №3 (70), URL: https://cyberleninka.ru/article/n/lingvokognitivnye-osnovy-yazykovogo-modelirovaniya.

Ли Цзе-Шэн, Цзе Сян. (2020). Патентная классификация путем тонкой настройки языковой модели BERT. Мировая патентная информация, Т. 61, 101965. DOI: 10.1016/j.wpi.2020.101965 (in Eng.)

Чандра Р., Сайни Р. (2021). Байден против Трампа: моделирование всеобщих выборов в США с использованием языковой модели BERT, в IEEE Access, Т. 9, стр. 128494–128505, 2021, doi: 10.1109/ACCESS.2021.3111035 (in Eng.)

Частикова В.А., Гуляй В.Г., Жерлицын С.А. (2022) Подход к решению проблемы контроля качества в сфере услуг на основе построения системы интеллектуального анализа данных." Вестник Адыгейского государственного университета. Серия 4: Естественно-математические и технические науки, №4 (311). URL: https://cyberleninka.ru/article/n/podhod-k-resheniyu-problemy-kontrolya-kachestva-v-sfere-uslug-na-osnove-postroeniya-sistemy-intellektualnogo-analiza-dannyh.

Челба К., Норузи М., Бенджио С. (2017). Моделирование языка N-грамм с использованием рекуррентной оценки нейронной сети. ArXiv, abs/1703.10724 (in Eng.)

Эттингер Э. (2020). Чем не является BERT: уроки нового набора психолингвистической диагностики языковых моделей. Труды Ассоциации компьютерной лингвистики, 2020, 8, стр. 34–48. DOI: 10.1162/tacl_a_00298 (in Eng.)

REFERENCES

Allyson Ettinger. (2020). What BERT is not: Lessons from a new Suite of Psycholinguistic Diagnostics for Language Models. Transactions of the Association for Computational Linguistics, 2020, 8. Pp. 34–48. DOI: 10.1162/tacl_a_00298 (in Eng.)

Chastikova V.A., Gulyai V.G., Zherlitsyn S.A. (2022). Podkhod k resheniyu problemy kontrolya kachestva v sfere uslug na osnove postroeniya sistemy intellektualnogo analiza dannykh." Vestnik Adygeiskogo gosudarstvennogo universiteta. Seriya 4: Estestvenno-matematicheskie i tekhnicheskie nauki, №4 (311). URL: https://cyberleninka.ru/article/n/podhod-k-resheniyu-problemy-kontrolya-kachestva-v-sfere-uslug-na-osnove-postroeniya-sistemy-intellektualnogo-analiza-dannyh (in Rus.)

Chelba C., Norouzi M., Bengio S. (2017). N-gram Language Modeling using Recurrent Neural Network Estimation. ArXiv, abs/1703.10724 (in Eng.)

Chandra R., Saini R. (2021). Biden vs Trump: Modeling US General Elections Using BERT Language Model, in IEEE Access. Vol. 9. Pp. 128494–128505, 2021, doi: 10.1109/ACCESS.2021.3111035 (in Eng.)

Ganesh Jawahar, Benoît Sagot, Djamé Seddah. (2019). What does BERT learn about the structure of language? In ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Jul 2019, Florence, Italy (in Eng.)

Hyeondey Kim, Seonhoon Kim, Inho Kang, Nojun Kwak, and Pascale Fung. (2022). Korean Language Modeling via Syntactic Guide. In Proceedings of the Thirteenth Language Resources and Evaluation Conference. Pp. 2841–2849, Marseille, France. European Language Resources Association (in Eng.)

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1 (Long and Short Papers). Pp. 4171–4186, Minneapolis, Minnesota (in Eng.)

Jieh-Sheng Lee, Jieh Hsiang. (2020). Patent classification by fine-tuning BERT language model. World Patent Information, Vol. 61. 101965. DOI: 10.1016/j.wpi.2020.101965 (in Eng.)

Khairullina R.Kh., Rakhimova E.F., Sagitova A.F. (2018). Lingvokognitivnye osnovy yazykovogo modelirovaniya. Mir nauki, kultury, obrazovaniya, №3 (70), URL: https://cyberleninka.ru/article/n/lingvokognitivnye-osnovy-yazykovogo-modelirovaniya (in Rus.)

Lee Sangah, Jang Hansol, Baik Yunmee, Park Suzi, Shin Hyopil. (2020). A Small-Scale Korean-Specific BERT Language Model. Journal of KIISE. 47. 682–692. 10.5626/JOK.2020.47.7.682 (in Eng.)

Myrzakhmetov B. & Kozhirbayev Z. (2018). Extended language modeling experiments for Kazakh. CEUR Workshop Proceedings, 2303 (in Eng.)

Peter F. Brown, Vincent J. Della Pietra, Peter V. deSouza, Jenifer C. Lai, Robert L. Mercer. (1992). Class-based n-gram models of natural language, Computational Linguistics. Vol. 18. Iss. 4. Pp. 467–479 (in Eng.)

Porechnyi A.S. (2020). Sozdanie instrumenta semantiko-sintaksicheskogo analiza tekstov na russkom yazyke. Sovremennye informatsionnye tekhnologii i IT-obrazovanie. DOI: 10.17308/sait.2020.1/2630 (in Rus.)

Salyp B., Smirnov A. (2022). Analiz modeli BERT kak instrumenta opredeleniya mery smyslovoi blizosti predlozhenii estestvennogo yazyka. StudNet, №5, URL: https://cyberleninka.ru/article/n/analiz-modeli-bert-kak-instrumenta-opredeleniya-mery-smyslovoy-blizosti-predlozheniy-estestvennogo-yazyka (in Rus.)

Si lhoussain Aouragha, Abdellah Yousfib, Saida Laaroussic, Hicham Gueddahd, Mohammed Nejja. (2021). A New Estimate of the n-gram Language Model. In Proceedings of the 5th International Conference on AI in Computational Linguistics. Pp. 211–215. DOI: 10.1016/j.procs.2021.05.111 (in Eng.)

Vaswani et al. (2017). Attention Is All You Need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS 2017), Long Beach, CA, USA, 2017 (in Eng.)

МАЗМҰНЫ

СОДЕРЖАНИЕ

CONTENTS

Publication Ethics and Publication Malpracticein the journals of the National Academy of Sciences of the Republic of Kazakhstan

For information on Ethics in publishing and Ethicalguidelines for journal publication see http://www.elsevier.com/publishingethics and http://www.elsevier.com/journal-authors/ethics.

Submission of an article to the National Academy of Sciences of the Republic of Kazakhstan implies that the described work has not been published previously (except in the form of an abstract or as part of a published lecture or academic thesis or as an electronic preprint, see http://www.elsevier.com/postingpolicy), that it is not under consideration for publication elsewhere, that its publication is approved by all authors and tacitly or explicitly by the responsible authorities where the work was carried out, and that, if accepted, it will not be published elsewhere in the same form, in English or in any otherlanguage, including electronically without the written consent of the copyright-holder. In particular, translations into English of papers already published in another language are not accepted.

No other forms of scientific misconduct are allowed, such as plagiarism, falsification, fraudulent data, incorrect interpretation of other works, incorrect citations, etc. The National Academy of Sciences of the Republic of Kazakhstan follows the Code of Conduct of the Committee on Publication Ethics (COPE), and follows the COPE Flowcharts for Resolving Cases of Suspected Misconduct (http://publicationethics.org/files/u2/New_Code.pdf). To verify originality, your article may be checked by the Cross Check originality detection service http://www.elsevier.com/editors/plagdetect.

The authors are obliged to participate in peer review process and be ready to provide corrections, clarifications, retractions and apologies when needed. All authors of a paper should have significantly contributed to the research.

The reviewers should provide objective judgments and should point out relevant published works which are not yet cited. Reviewed articles should be treated confidentially. The reviewers will be chosen in such a way that there is no conflict of interests with respect to the research, the authors and/or the research funders.

The editors have complete responsibility and authority to reject or accept a paper, and they will only accept a paper when reasonably certain. They will preserve anonymity of reviewers and promote publication of corrections, clarifications, retractions and apologies when needed. The acceptance of a paper automatically implies the copyright transfer to the National Academy of Sciences of the Republic of Kazakhstan.

The Editorial Board of the National Academy of Sciences of the Republic of Kazakhstan will monitor and safeguard publishing ethics.

Правила оформления статьи для публикации в журнале смотреть на сайтах:
www:nauka-nanrk.kz
http://physics-mathematics.kz/index.php/en/archive
ISSN2518-1726 (Online),
ISSN 1991-346X (Print)

Подписано в печать 28.09.2023.
Формат 60x881/8. Бумага офсетная. Печать –ризограф.
18,0 п.л. Тираж 300. Заказ 3.

РОО «Национальная академия наук РК» 050010,
Алматы, ул. Шевченко, 28, т. 272-13-19