МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РЕСПУБЛИКИ КАЗАХСТАН

НЕКОММЕРЧЕСКОЕ АКЦИОНЕРНОЕ ОБЩЕСТВО
«АЛМАТИНСКИЙ УНИВЕРСИТЕТ ЭНЕРГЕТИКИ И СВЯЗИ ИМЕНИ ГУМАРБЕКА ДАУКЕЕВА»

В Е С Т Н И К

АЛМАТИНСКОГО УНИВЕРСИТЕТА ЭНЕРГЕТИКИ И СВЯЗИ

Учрежден в июне 2008 года

2 (49) 2020 Импакт-фактор - 0.154

Научно-технический журнал
Выходит 4 раза в год

Алматы

РЕДАКЦИОННАЯ КОЛЛЕГИЯ

Сагинтаева С.С., д-р экон. наук, канд. физ.- мат. наук, академик МАИН, ректор НАО «Алматинский Университет Энергетики и Связи имени Гумарбека Даукеева».
Гита Ревалде, доктор PhD, член-корреспондент Академии наук Латвии, НАО «Алматинский Университет Энергетики и Связи имени Гумарбека Даукеева».
Главный редактор – Стояк В.В., канд. техн. наук, профессор, НАО «Алматинский Университет Энергетики и Связи имени Гумарбека Даукеева».
Заместитель главного редактора – Жауыт А., доктор PhD, НАО «Алматинский Университет Энергетики и Связи имени Гумарбека Даукеева».
Илиев И.К., д-р техн. наук, Русенский университет, Болгария.
Галайко Дмитрий, доктор PhD, университет Сарбонны, Франция.
Такая Инамори, доктор PhD, Университет Токио, Япония.
Цветков В.Ю., д-р техн. наук, Белорусский государственный университет информатики и радиоэлектроники, Белоруссия.
Кузнецов А.А., д-р техн. наук, ФГБОУ ВА «Омский государственный университет путей сообщения», Российская Федерация.
Авезова Н.Р., д-р, техн. наук, Министерство инновационного развития Республики Узбекистан.
Мунц В.А., д-р техн. наук ФГАОУ ВО «УрФУ имени первого Президента России Б.Н. Ельцина», Российская Федерация.
Мустафин М.А., д-р техн. наук, НАО «Алматинский Университет Энергетики и Связи имени Гумарбека Даукеева».
Обозов А.Д., д-р техн. наук, Кыргызский государственный технический университет им. И. Раззакова, Кыргызская Республика.
Орумбаев Р.К., д-р техн. наук, НАО «Алматинский Университет Энергетики и Связи имени Гумарбека Даукеева».
Потехин В.В., канд. техн. наук доцент Санкт-Петербургского политехнического университета Петра Великого, г. Санкт-Петербург, Российская Федерация.
Франческо Сандоро, доктор PhD, Университет Альдо Моро, Бари, Италия.
Туманбаева К.Х., канд. техн. наук, НАО «Алматинский Университет Энергетики и Связи имени Гумарбека Даукеева».
Мутуле Анна, доктор PhD, Рижский Технический Университет, Латвия.
Махмутов С.К., канд. истор. наук, НАО «Алматинский Университет Энергетики и Связи имени Гумарбека Даукеева».
Алипбаев К.А., доктор PhD, НАО «Алматинский Университет Энергетики и Связи имени Гумарбека Даукеева».
Кабдушев Б.Ж., канд. истор. наук, НАО «Алматинский Университет Энергетики и Связи имени Гумарбека Даукеева».

Ответственный секретарь: Шуебаева Д.А.
Технический редактор казахского языка Алмухаметова Г.С.
Технический редактор русского языка Ефимова А.Н.
Технический редактор английского языка Сергеева Л.Д.

С содержанием журнала можно ознакомиться на сайте http://aues.kz , http://vestnik-aues.kz Подписаться на журнал можно в редакции журнала и по объединенному каталогу Департамента почтовой связи, подписной индекс -**74108**.
Адрес редакции: 050013, г.Алматы, НАО «Алматинский Университет Энергетики и Связи имени Гумарбека Даукеева», ул. Байтурсынулы, дом 126/1, офис А224
Тел.: 8 (727) 292 58 48, 708 880 77 99, e-mail: vestnik@aues.kz.
Тираж 200 экз.

Вестник Алматинского университета

энергетики и связи № 2 (49) 2020

СОДЕРЖАНИЕ

ТЕПЛОЭНЕРГЕТИКА И ТЕПЛОТЕХНОЛОГИИ

Джунусов Е. А., Стояк В. В.
Анализ энергетического и экологического состояния существующей системы теплоснабжения г. Алматы с рассмотрением возможности применения перспективных технологий для теплоснабжения…….……..…….7

Шишкин А. А., Шишкин А. А., Душкина Н. Н.
Экспериментальное исследование аэродинамической структуры потока в зоне сепарации частиц нового золоулавливающего устройства…….……17

Идрисова К. С., Туманова А. А., Колдасова Г. А.
ЖЭС гидрокүлқожшығару жүйесінде өтетін

процестерді зерттеу.....25

ЭЛЕКТРОЭНЕРГЕТИКА И ЭЛЕКТРОТЕХНОЛОГИИ

Умбеткулов Е. К., Сахариев Б. С., Амитов Е. Т.
К вопросу обеспечения электроэнергией спасательных и других неотложных работ при землетрясениях.....
.....31

Мурат А. К., Соколов С. Е., Сағынбек Р. Б., Умбеткулов Е. К., Соколова И. С. Исследование и моделирование управляемых шунтирующих реакторов с подмагничиванием постоянным током в PSCAD/EMTDC.....39

Ибраева Л. К., Карташов Е. О.
Разработка и исследование распределенной системы энергоснабжения на базе возобновляемых источников энергии.....
.....48

ТЕЛЕКОММУНИКАЦИИ И КОММУНИКАЦИОННАЯ ИНЖЕНЕРИЯ

Туржанова К. М.
Анализ подходов и выбор показателей оценки эффективности функционирования сетей мобильной связи 4G/5G при оказании услуг интернета вещей…….57

Солощенко А. В.
Оптимизация радиопокрытия и балансировка нагрузки в сетях сотовой связи 4-ого поколения……
…….....65

Murat A., Imanbayeva A. K.
Dynamic analysis of a modified duffing oscillator

…….72 5

Вестник Алматинского университета

энергетики и связи № 2 (49) 2020

КОСМИЧЕСКАЯ ИНЖЕНЕРИЯ И

РОБОТОТЕХНИКА

Сухенко А. С., Бопеев Т. М., Шаповалова Д. Л., Фоменко А. Е., Елубаев С. А.
Программно-математическое обеспечение аппаратно-программного комплекса для наземных испытаний системы ориентации спутников и БПЛА……....80

АВТОМАТИКА, ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ И КИБЕРБЕЗОПАСНОСТЬ

Хан С. Г., Макишев А. Қ.
Разработка и исследование алгоритма и программы управления системы распределенного теплоснабжения……87

Бияшев Р., Алгазы К., Дюсенбаев Д., Сакан К.
Результаты линейного криптоанализа шифра QAMAL……96

Жақсылық Г. О., Кулмагамбетова Ж. К.
Arduino платформасы негізіндегі қол жеткізуді бақылау және басқару жүйесінің құрылымдық және функционалдық сызбасын әзірлеу…….....106

Брейдо И. В., Вяткин В., Котов Е. С.
Система пред – и послеаварийного контроля режимов работы взрывозащищенного электрооборудования.....114

Brimzhanova S.S., Atanov S.K., Moldamurat Kh., Tankibaeva A.K., Sailau A.Zh. Implementation of an intelligent testing system application……..……..122

Құмалақов Б., Шақан Я.
Blockchain in education: how to manage student credits of higher education through the blockchain?.....128

Мансурова М. Е., Қадырбек Н. Қ., Мисебай А. М., Қырғызбаева М. Е. Қазақ тіліндегі сөйлемдер үшін тәуелділік ағаштарын құру.....134

Мансурова М. Е., Досанов Б. Б., Чикибаева Д. Ю.
Мәтіндерден атаулы мәндерді алу үшін машиналық оқыту әдістерін қолдану……142

ВОПРОСЫ ВЫСШЕЙ ШКОЛЫ, ИННОВАЦИОННОЕ РАЗВИТИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

Сериков Э. А.
К вопросу планирования учебной нагрузки студентов в кредитах ECTS.....148

ағаштары сөйлемдегі сөздердің нақты орын тәртібін көрсетпейді, тек сөздердің байланыс иерархиясын кескіндейді және қазақ тілінің орыс тілінен басты ерекшелігі ретінде сөйлемнің проективті болуын айта аламыз.

Бұл жұмыс ҚР БҒМ O.0856 BR05236340 «Қазақстан Республикасының цифрлы экономикасын қалыптастыру шеңберінде «логистикалық-агломерациялық» жүйесінің талдау және шешім қабылдау жоғары өнімді зияткерлік технологияларын құру» және AP05132933 «Шешім қабылдау сапасын жақсарту үшін деректердің гетерогенді көздерінен білімді алу жүйесін құру» жобалары шеңберінде жасалды.

ӘДЕБИЕТТЕР ТІЗІМІ

1. Батура, Т. В. Математическая лингвистика и автоматическая обработка текстов : учеб. пособие / Новосиб. гос. ун-т. – Новосибирск: РИЦ НГУ, 2016. – 166 с.

2. Saziye Betul Ozates, Arzucan Ozgur, Dragomir R. Radev. Sentence Similarity based on Dependency Tree Kernels for Multi-document Summarization // Қолжетімділік режимі: https://www.aclweb.org/anthology/L16-1452.pdf, еркін [13.01.2019].

3. Сайранбаев Т. Қазақ тіл білімінің мәселелері. Алматы, Абзал ай: 2014. – 640 б. 4. Аширова А.Т. Сөздердің тіркесу тәсілдері мен байланысу формалары / Herald of journalism № 37(1). – 2016.

5. Ж. Мейрамбеккызы, А.А. Хорошилов. Синтактико-семантические представления для развития казахско-русского фразеологического машинного перевода / Труды V международной конференции по компьютерной обработке тюркских языков «Turklang 2017», Казань, Россия, 2017.

6. Takeaki UNO. An Algorithm for enumerating all directed spanning trees in a directed graph // Қолжетімділік режимі: http://research.nii.ac.jp/~uno/papers/isaac96web.pdf, еркін [20.01.2019].

LIST OF REFERENCES

1. Batura, T. V. Matematicheskaya lingvistika i avtomaticheskaya obrabotka tekstov: ucheb. posobiye / Novosib. gos. un-t. – Novosibirsk: RITS NGU, 2016. – 166 s.

2. Saziye Betul Ozates, Arzucan Ozgur, Dragomir R. Radev. Sentence Similarity based on Dependency Tree Kernels for Multi-document Summarization // Availability Mode: https://www.aclweb.org/anthology/L16-1452.pdf, еркін [13.01.2019].

3. Sairanbaev T. Kazak til biliminin maseleleri. Almaty, Abzal ai: 2014, - 640 b.

4. Ashirova A.T. "Sozderdin tirkesu tasilderi men bailanysu formalary Herald of journalism № 37(1). - 2016.

5. J. Meyrambekkyzy, A.A. Khoroshilov. Sintaktiko-semanticheskie predstavleniya dliya razvitia kazakhsko-russkogo phraseeologygycheskogo mashinnogo perevoda / V international conference on computer processing of Turkic languages "Turklang 2017", Kazan, Russia, 2017.

6. Takeaki UNO. An Algorithm for enumerating all directed spanning trees in a directed graph // Availability Mode: http://research.nii.ac.jp/~uno/papers/isaac96web.pdf.

СОЗДАНИЕ ДРЕВА ЗАВИСИМОСТЕЙ ДЛЯ ПРЕДЛОЖЕНИЙ НА КАЗАХСКОМ ЯЗЫКЕ

М.Е. Мансурова, Н.К. Кадырбек, А.М. Мисебай, М.Е. Кыргызбаева*

Казахский национальный университет имени аль-Фараби, Алматы,
Республика Казахстан
e-mail: mansurova.madina@gmail.com, nurgaliqadyrbek@gmail.com, armanbek128@mail.ru, marzhan.kyrgyzbaeva@gmail.com

Аннотация. *Естественный язык является сложной структурой, в которой выделяются различные уровни: фонетический, морфологический, лексический, синтаксический, семантический. Особое место среди анализаторов для автоматической обработки естественно-языковых текстов занимает синтаксический парсер, который позволяет определить основную смысловую структуру предложения. В данной работе представлен алгоритм построения древа зависимостей предложений на казахском языке, который позволяет выявить смысловую структуру и выполнить синтаксический анализ предложения. Древо зависимостей создается на основе остовного древа из ориентированного графа, построенного посредством правил синтаксических отношений в предложении.*

Ключевые слова: *древо зависимостей, грамматика зависимостей, корпусная лингвистика, формальная модель фраз.*

DEVELOPMENT OF THE DEPENDENCY TREES FOR THE KAZAKH LANGUAGE SENTENCES

M.E. Mansurova, N.K. Kadyrbek, A.M. Misebay, M.E. Kyrgyzbayeva*

Kazakh national university named after al-Farabi, Almaty, Republic of Kazakhstan e-mail: mansurova.madina@gmail.com, nurgaliqadyrbek@gmail.com, armanbek128@mail.ru, marzhan.kyrgyzbaeva@gmail.com

Abstract. *Natural language is a complex structure in which various levels are distinguished: phonetic, morphological, lexical, syntactic, semantic. A special place among analyzers for the automatic processing of natural language texts is occupied by the syntactic parser, which allows you to determine the main semantic structure of the sentence. In this paper, an algorithm for constructing a sentence dependency tree in the Kazakh language is presented. It allows you to identify the semantic structure and perform syntactic analysis of the sentence. A dependency tree is created on the basis of a spanning tree from a directed graph constructed by the rules of syntactic relations in a sentence.*

Keywords: *Dependency tree, dependency grammar, corpus linguistics, syntactic structure, phrase structure rule.*

АВТОМАТИКА, ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ И КИБЕРБЕЗОПАСНОСТЬ

МРНТИ 20.23.17; 20.23.21; 20.23.25

МӘТІНДЕРДЕН АТАУЛЫ МӘНДЕРДІ АЛУ ҮШІН МАШИНАЛЫҚ ОҚЫТУ ӘДІСТЕРІН ҚОЛДАНУ

М.Е. Мансурова, Б.Б. Досанов, Д.Ю. Чикибаева

Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан Республикасы
e-mail: mansurova.madina@gmail.com, dossanbekzhan@gmail.com, dashachikibaeva@mail.ru

Аңдатпа. *Атаулы мәндерді алу – табиғи тілдегі мәтіндерден ақпараттың құрылымдық көрінісін алуға мүмкіндік беретін шешімдер жиыны. Атаулы субъектілердің, адам, ұйым, орын, уақыт, мөлшер атауларын және т.б анықтау міндеті мәтін ішінен негізгі нысандарды табуға мүмкіндік береді. Мұндай есептерді шешу машиналық оқыту әдістерін қолдануға негізделеді. Машиналық оқыту көмегімен құрылымды емес мәтіндер ішінен атаулы мәндерді алу оңайға соғады. Бұл зерттеу жұмысында train data үшін қазақ тіліндегі мәтіндер ішінен қала, тұлға, ұйым атаулары қолмен белгіленді, жиналған мәліметтерге BERT машиналық оқыту әдісі көмегімен оқыту жүргізіліп, модель құрастырылылды. Нәтижесінде қазақ тіліндегі атаулы мәндерді алуға арналған Web бағдарлама әзірленді.*

Түйінді сөздер: *атаулы мәндер, атаулы мәндерді алу, машиналық оқыту, BERT әдісі.*

Кіріспе

Қазіргі таңда әр түрлі салалардағы ақпараттық инфрақұрылымның үнемі жетілдірілуі, Интернет желісінде миллиардтаған адамдардың бір бірімен ақпарат алмасу нәтижесінде қол жетімді цифрлық ресурстар біздің индустриалды қоғамымызда айтарлықтай өсті. Бір жағынан ақпарат көздері мен көлемі көп болса, екінші жағынан пайдалы мәліметтердің тығыздығы төмендейді. Соның нәтижесінде құрылымды емес мәтіндердің көптігінен құнды, керекті мәліметтерді алу қиынға соғады. Қазіргі таңда мәліметтер құрылымды емес үлкен көлемде және алуан түрлі болғандықтан ақпаратты талдау, өңдеу және керекті мәліметті алу қиынға соғады. Құрылымды емес үлкен мәтіндерді өңдеу, мәтін ішінен атаулы мәндерді алу, талдау үшін қолданатын шешімдер де көбейіп жатыр. Үлкен көлемдегі мәтіндер ішінен атаулы мәндерді алуға терең оқыту және машиналық оқыту технологиялары көмегімен қол жеткізуге болады.

Атаулы мәндерді алу (NER – Named Entity Recognition) – табиғи тілдерді өңдеудің (Natural Language Processing) негізгі мәселелерінің бірі болып табылады. Атаулы мәндерге құрылымды емес мәтіндер ішінен тұлға есімдері (names), ұйым атаулары (organization), географиялық объектілер, мекен-жайлар (locations), күндер (dates) және күрделі жағдайларда олардың арасындағы байланыстар, сондай-ақ оқиғалар, себеп-салдарлық байланыстар, өнімдерді жатқызамыз. Сондай-ақ, осы тізімге электрондық мекенжайлар, телефондар, белгілі бір деректер түрлері (мысалы, IP мекенжайлары) сияқты мәндерді де қосуға болады.

Атаулы мәндерді алу міндеті мәтіндегі атаулы мәндерді бөлу және жіктеу болып табылады. CoNLL 2003 конференциясы аясында ағылшын тіліндегі корпуста төрт типтегі атаулы мәндерді тану әдістерінің сапасын бағалау үшін жарыс өткізілді. NER есебін шешу үшін әртүрлі тәсілдер ұсынылды. Сол тәслдер арасында құрылымды емес сөйлемдер ішінен атаулы мәндерді тануда машиналық оқытуға негізделген әдістер жоғары нәтиже көрсетті.

Атаулы мәндерді алу әдістері

Атаулы мәндерді алу MUC-6 конференциясында ақпаратты алудың маңызды қосалқы құралы ретінде анықталды. Ең жиі қолданылатын салыстырмалы тәсілдерді үш санатқа бөлуге болады: ережелер мен сөздіктерге негізделген әдістер, статистикалық және аралас. Бастапқыда NER ережелер мен сөздіктерге негізделген әдіснамаларды пайдаланды, олар пәндік сала бойынша сарапшылардың

ережелер жиынтығын әзірлеуді және тиісті лингвистерді пайдалануды талап етеді. Дегенмен ережелер барлық тілдік құбылыстарды қамтымайды, құрастыру кезеңі тым үлкен, сонымен қатар, мұндай тәсілге төзімділіктің аз ықтималдығы бар [1]. Статистикалық әдістер негізінен жасырын маркалық модельдерді (HMM) [2], максимальды энтропияны (ME) [3], тірек векторлары машиналарын (SVM) [4] және CRF [5] қамтиды. Әдістердің бұл түрлері модельді статистикалық ықтималдықпен ұштастыра оқыту үшін таңбаланған корпустық деректерді пайдаланады. Зерттеушілер сонымен бірге fishery data [6], dietary data [7] және Chinese legal documents data [8] сияқты белгілі бір домендік субъектілер үшін NER зерттеулерін жүргізді. BERT машиналық оқыту моделі – ең танымал модельдердің бірі болып табылады.

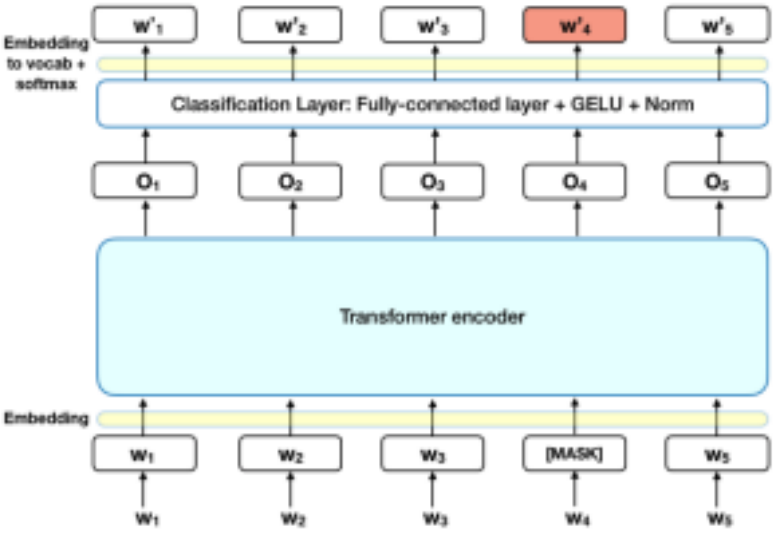
BERT әдісі

BERT (Bidirectional Encoder Representations from Transformers) – бұл «трансформерлер» ретінде белгілі бір тізбектермен жұмыс істеу үшін жаңа архитектураның машиналық оқытуға негізделен табиғи тілді өңдеу әдісі. Бұл 2018 жылы Google инженерлері жасаған және жариялаған жаңа технология болып табылады [9]. BERT алгоритмінің ерекшелігі – сөйлемдегі немесе сұрақтағы сөздер жиынтығын жақсырақ түсінуге мүмкіндік беретін екі бағытты «кодерлер» қолданылауында. BERT әдісінің көмегімен тілдік модельдерін жаттықтыру сөйлемдегі сөздердің толық жиынтығы негізінде екі бағытты оқыту болып табылады, ал NLP дәстүрлі модельдері тілдік модельдерді сөздердің реті бойынша (оңнан солға немесе солдан оңға қарай) оқытады.

Зерттеудің әдіснамалық негіздері

Каноникалық түрде атаулы мәндерді алу есебіне кіріс мәліметі ретінде табиғи тілде мәтін беріледі, ал есептің шешімі мәтін ішінен атаулы мәндерді алу болып табылады. Бастапқы кіріс

деректері бірінші кезекте токендерге бөлінеді. BERT әдісінде WorkPiece-ке негізделген алдын ала белгіленген токенизация моделі бар. Токенизация көмегімен кейбір сөздер түбір болып қалады, ал кейбіреулері, атап айтқанда, бейтаныс сөздер бөліктерге бөлінеді. Мұндай әдіс оқытылған кіріс деректерінде кездеспеген сөздер болған жағдайда жақсы. Сондай-ақ, бұл күрделі сөздерді жақсы ұсынуға мүмкіндік береді. Қол жетімді BERT модельдерінің әрқайсысы үшін сәйкес токенизатор ұсынылады. Бұл жұмыста 104 тілді қамтитын "BERT-base-multilingual-cased" моделі қолданылды. Бұл жұмыста тағы трансферттік оқыту тәсілі қолданылады - баптау (ағылш. Fine-Tuning). Бұл тәсілде модельдегі сөздер салмағы, оның ішінде алдыңғы жаттығу BERT жаңартылады. Алынған таңбаларды атаулы мәндер бойынша жіктеу үшін сызықтық жіктеуіш қолданылады.



Сурет 1 – BERT машиналық оқыту әдісінің архитектурасы [9]

Тәжірибелік жұмыс

Қазақ тіліндегі құрылымды емес сөйлемдер ішінен атаулы мәндерді алу үшін практикалық құрамдас бөлігі қойылған міндетті шешудің әр түрлі сатыларына сәйкес келетін бірнеше кезеңге бөлінген:

1) Қазақ тілінде құрылымды емес мәтіндердің корпусын жинақтау;
2) Пәндік аймаққа байланысты атаулы мәндерді зерттеу және бастапқы кіріс деректерін дайындау;
3) Машиналық оқыту әдістерін оқыту және жаттықтыру;
4) Нәтижелерді сандық бағалау көрсеткіштерін таңдау;
5) Құрылымды емес сөйлем ішінен атаулы мәндерді анықтауға арналған Web бағдарлама құрастыру.

Мәтіндердің корпусын жинақтау

Машиналық оқыту негізінде train data оқытуы үшін мәтіндер корпусы ретінде қазақ тіліндегі жаңалықтар таңдалынды. Crawl көмегімен жаңалықтар ішінен тек қазақ тіліндегі сөйлемдер json форматында жинақталды. Жазбаларды іріктеу және атаулы мәндерді қамтымайтын жазбаларды бөлу үшін іздеу критерилері кең таралған аттарды, белгілі адамдардың тектерін, сондай-ақ жаңалықтық мәтіндерден іріктеу негізінде құрылған, талданған ұйымдардың атауларын қамтыды. Әр жазба ішінде кемінде 1 атаулы мән болуы шарт. Нәтижесінде корпусқа қазақ тіліндегі 170 мың жазба жиналды.

Негізгі атаулы мәндерді зерттеу

Біздің бағдарламада атаулы мәндерге негізгі 3 атаулы мәнді алдық:

1) Person
2) Location
3) Organization

Атаулы мәндердің шекарасын белгілеу үшін IOB (Inside–outside–beginning) форматы қолданылды. IOB форматы – бұл компьютерлік лингвистикада күрделі тапсырмаларда таңбалауыштарды белгілеуге арналған жалпы белгілеу форматы. IOB форматында: B белгісі атаулы мәннің басталуын білдіреді, I – оның ішіндегі орналасуы; O белгісі атаулы мәнге кірмейтін таңбалар белгіленеді.

```
Қазақстанның /B-LOC
бас /O
санитар /O                    B Begin - атаулы мәннің басталуы
дәрігері /O                   I Inside - атаулы мәннің жалғасы
Жандарбек /B-PER              O Outside - атаулы мәнге қатысты емес сөз
Бекшин /I-PER
коронавирус /O                PER PERSON - тұлға аты
бойынша /O                    LOC LOCATION - географиялық обьект атауы
қауіпті /O                    ORG ORGANIZATION - ұйым атауы
елдердің /O
жаңа /O
тізімін /O
жариялады /O
. /O
```

Сурет 2 – Таңбалау мысалы

Машиналық оқыту әдістерін оқыту және жаттықтыру

Мәліметтер қазақ тіліндегі жаңалықтар таңдалынды. Train data үшін корпус мәліметтері 4 түрлі типке (Person, Location, Organization және Other) қолмен белгіленді. Корпус мәліметтерін белгілеу үшін арнайы бағдарлама құрастырылды.

Сурет 3 – Атаулы мәндерді белгілеуге арналған бағдарлама

Нәтижесінде машиналық оқытуға бастапқы кіріс мәліметі ретінде 6 мың мәтін белгіленді. Келесі қадам қазақ тіліндегі құрылымды емес сөйлем ішінен атаулы мәндерді алуға арналған модельді құрастыру үшін мәліметтерді BERT машиналық оқыту әдісімен оқытамыз. **Нәтижені бағалау**

Нәтижені бағалау кезінде атаулы мәндердің өлшемі келесі екі шарттың дұрыс болған кезінде көрінеді:

- Атаулы мән шекарасы дұрыс бөлінген;
- Атаулы мән дұрыс танылған.

Алгоритмнің сапасын өлшеу үшін үш сандық бағалау таңдалды:

1) Дәлдігі

$$\text{Precision} = \frac{\text{(дұрыс танылған атаулы мәндер саны)}}{\text{(барлық таңдалған атаулы мәндер саны)}} \quad (1)$$

2) Толықтығы

$$\text{Recall} = \frac{\text{(дұрыс танылған атаулы мәндер саны)}}{\text{(корпустағы атаулы мәндер саны)}} \quad (2)$$

3) F-өлшем

$$\text{F1} = \frac{2PR}{P+R} \quad (3)$$

Төмендегі кестеде F-өлшем бағалуы бойынша қазақ тіліндегі жиналған корпустан атаулы мәндерді тану нәтижесі көрсетілген.

Кесте 1 – F-өлшем бағалуы бойынша атаулы мәндерді тану нәтижесі

Алгоритм	P	R	F1
BERT	83.43	77.09	80.14

Web бағдарлама құрастыру

Қолданушы енгізген мәтін ішінен машиналық оқыту негізінде атаулы мәндерді алу үшін Web бағдарлама әзірленді.

Сурет 4 – Атаулы мәндерді тануға арналған бағдарлама

Қорытынды

Бұл жұмыста қазақ тіліндегі құрылымды емес мәтіндер ішінен атаулы мәндерді алу үшін машиналық оқыту әдісі негізінде жүйе әзірленді. Машиналық оқыту әдісі ретінде BERT (Bidirectional Encoder Representations from Transformers) таңдалынды. Нәтижесінде BERT машиналық оқыту әдіс көмегімен атаулы мәндерді алуға арналған модель құрастырылып, жүйе әзірленді. F-өлшем бойынша 80.14% алдық.

Болашақта алгоритм сапасының сандық бағалауларының мәндерін жоғарылату үшін бастапқы кіріс деректеріне морфологиялық талдау жасап, терең оқыту (deep learning) қолдану жоспарланып отыр.

ӘДЕБИЕТТЕР ТІЗІМІ

1 . D. M. Bikel, S. Miller, R. Schwartz, and R. Weischedel, “Nymble: a high-per-formance learning name-finder,” in Proceedings of Fifth Conference on Applied Natural Language Processing, pp. 194–201, Washington, DC, USA, April 1997.

2. A. E. Borthwick, “A maximum entropy approach to named entity recognition,” New York University, New York City, NY, USA, 1999, Master’s Dissertation.

3. H. Isozaki and H. Kazawa, “Efficient support vector classifiers for named entity recognition,” in Proceedings of 19th International Conference on Computational Linguistics, pp. 1–7, Taipei, Taiwan, September 2002.

4. A. McCallum and W. Li, “Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons,” in Proceedings of Seventh Conference on Natural Language Learning at HLT-NAACL, vol. 4, pp. 188–191, Edmonton, Canada, May 2003.

5. Y. Liu, “Named entity recognition in Chinese Micro-blog based on deep learning,” Advanced Engineering Sciences, vol. 48, no. 2, pp. 142–146, 2016.

6. J. Sun, H. Yu, and Y. Feng, “Recognition of nominated fishery domain entity based on deep learning architectures,” Journal of Dalian Ocean University, vol. 33, no. 2, pp. 265–269, 2018. 7. Q. Wei, T. Chen, R. Xu, Y. He, and L. Gui, “Disease named entity recognition by combining conditional random fields and bidirectional recurrent neural networks,” Database, vol. 2016, Article ID baw140, 2016.

8. L. Zhang, C. Qin, and W. Ye. Research on legal field entity automatic recognition model based on conditional random fields // New Technology of Library and Information Service, vol. 11, pp. 46–52, 2017. 9. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding / Artificial Intelligence Conference.-New York, 2019.

LIST OF REFERENCES

1. D. M. Bikel, S. Miller, R. Schwartz, and R. Weischedel, “Nymble: a high-per-formance learning name-finder,” in Proceedings of Fifth Conference on Applied Natural Language Processing, pp. 194–201, Washington, DC, USA, April 1997.

2. A. E. Borthwick, “A maximum entropy approach to named entity recognition,” New York University, New York City, NY, USA, 1999, Master’s Dissertation.

3. H. Isozaki and H. Kazawa, “Efficient support vector classifiers for named entity recognition,” in Proceedings of 19th International Conference on Computational Linguistics, pp. 1–7, Taipei, Taiwan, September 2002.

4. A. McCallum and W. Li, “Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons,” in Proceedings of Seventh Conference on Natural Language Learning at HLT-NAACL, vol. 4, pp. 188–191, Edmonton, Canada, May 2003.

5. Y. Liu, “Named entity recognition in Chinese Micro-blog based on deep learning,” Advanced Engineering Sciences, vol. 48, no. 2, pp. 142–146, 2016.

6. J. Sun, H. Yu, and Y. Feng, “Recognition of nominated fishery domain entity based on deep learning architectures,” Journal of Dalian Ocean University, vol. 33, no. 2, pp. 265–269, 2018.

7. Q. Wei, T. Chen, R. Xu, Y. He, and L. Gui, “Disease named entity recognition by combining conditional random fields and bidirectional recurrent neural networks,” Database, vol. 2016, Article ID baw140, 2016.

8. L. Zhang, C. Qin, and W. Ye, “Research on legal field entity automatic recognition model based on conditional random fields,” New Technology of Library and Information Service, vol. 11, pp. 46–52, 2017. 9. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding / Artificial Intelligence Conference.-New York, 2019.

ПРИМЕНЕНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ

ИЗВЛЕЧЕНИЯ ИМЕНОВАННЫХ СУЩНОСТЕЙ ИЗ ТЕКСТОВ

М.Е. Мансурова, Б.Б. Досанов, Д.Ю. Чикибаева

Казахский национальный университет имени аль-Фараби, Алматы,
Республика Казахстан
e-mail: mansurova.madina@gmail.com, dossanbekzhan@gmail.com, dashachikibaeva@gmail.com

Аннотация. *Извлечение именованных сущностей – это набор решений, позволяющий получить структурную картину информации из текстов на естественном языке. Основная задача состоит в том, что необходимо определить названия так называемых субъектов, физических лиц, названия организаций, мест, времени, количества и т. д. Позволяет найти в тексте основные объекты. Основывается на использовании машинных методов обучения для решения таких задач. С помощью машинного обучения находим именованные сущности из неструктурированных текстов. В этой исследовательской работе мы создали систему, используя машинный метод обучения BERT для казахского языка, который получает текст на естественном языке и выводит из текста именованную сущность. Кроме того, был создан корпус текстов на казахском языке, размеченны вручную названия городов, физических лиц и организаций.*

Ключевые слова: *именованные сущности, named entity recognition, машинное обучение, метод BERT.*

APPLYING OF MACHINE LEARNING METHODS FOR NAMED ENTITY EXTRACTION FORM THE TEXTS

M.E. Mansurova , B.B. Dossanov, D.Yu. Chikibaeva

Kazakh national university named after al-Farabi, Almaty, Republic of Kazakhstan
mansurova.madina@gmail.com, dossanbekzhan@gmail.com,
dashachikibaeva@gmail.com

Abstract. *Extracting named entities is a set of solutions that allows you to get a structural picture of information from texts in a natural language. It is necessary to determine the names of subjects, individuals, the name of the organization, places, time, quantity, etc. allows you to find the main objects in the text. They are based on the use of machine learning methods to solve such problems. Using machine learning, we find named entities from unstructured texts. In this research work, we created a system using the BERT machine learning method for the Kazakh language, which receives natural language text and derives a named entity from the text. In addition to the base of texts in the Kazakh language, a manually-labeled name of the cities, individuals and organizations was created.*

Keywords: *named entities, named entity recognition, machine learning, method BERT.*