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). 7–17
https://doi.org/10.32014/2023.2518-1726.200

UDC 004.942

© G. Abdikalyk*, A. Mukanova, A. Nazyrova, 2023
Astana International University, Astana, Kazakhstan.
E-mail: gulnazymabdikalik@gmail.ru

NAMED ENTITY RECOGNITION FOR KAZAKH LANGUAGE USING CRF AND RANDOM FOREST MODELS: A COMPARATIVE STUDY

Abdikalyk Gulnazym — grad student of the Astana International University. 010000. Astana, Kazakhstan
E-mail: gulnazymabdikalik@gmail.ru. ORCID ID: https://orcid.org/0009-0008-5216-0707;
Assel Mukanova — PhD, assoc. professor of the Astana International University. 010000. Astana, Kazakhstan
E-mail: asiserikovna@gmail.com. ORCID ID: https://orcid.org/0000-0002-8964-3891;
Nazyrova Aizhan — senior lecturer of the Astana International University. 010000. Astana, Kazakhstan
E-mail: ayzhan.nazyrova@gmail.com. ORCID ID: https://orcid.org/0000-0002-9162-6791.

Abstract. The activity of recognizing and categorizing named items in text is known as named entity recognition (NER), and it is essential to natural language processing. However, due to morphological complexity and scarce linguistic resources, NER for under-resourced languages like Kazakh presents distinct difficulties. The performance of two well-known machine learning models, Conditional Random Fields (CRF) and Random Forest, for NER in the Kazakh language was thoroughly compared in this scientific work. The work addresses feature engineering strategies catered to the morphological complexity of Kazakh and makes use of a benchmark dataset generated specifically for Kazakh NER. While Random Forest models manage high-dimensional feature spaces and intricate interactions in the data, CRF models capture sequential dependencies and contextual information. The efficiency of both the CRF and Random Forest models for Kazakh NER is shown by experimental findings. However, the scarcity of labeled data has an impact on how well these models work. Future research directions include extending annotated datasets through partnerships with linguists and native speakers in order to address this constraint. The study also emphasizes how crucial it is to deal with Kazakh's complicated morphology in NER. Word stems and part-of-speech tags are among the morphological qualities that CRF and Random Forest models integrate, which enhances the recognition of named entities

in a variety of inflections and variants. The comparison analysis sheds light on the advantages and disadvantages of the Random Forest and CRF models for Kazakh NER. While Random Forest models may manage complex linkages and feature interactions, CRF models excel at capturing sequential dependencies and utilizing contextual information. The NER task's specific needs and characteristics determine the model to use. In conclusion, by offering information on the effectiveness of CRF and Random Forest models, this comparative study advances the field of NER for the Kazakh language. It shows the value of handling morphological complexity, the necessity of annotated data, and directs future study aimed at enhancing Kazakh NER systems.

Keywords: Named Entity Recognition (NER), Kazakh language, Conditional Random Fields (CRF), Random Forest, comparative study

This research has been funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. AP19577922)

© Г. Әбдіқалық*, Ә. Мұқанова, А. Назырова, 2023
Астана Халықаралық Университеті, Астана, Қазақстан.
E-mail: gulnazymabdikalik@gmail.ru

CRF ЖӘНЕ RANDOM FOREST МОДЕЛДЕРІНІҢ КӨМЕГІМЕН ҚАЗАҚ ТІЛІНДЕ АТАЛҒАН ОБЪЕКТІЛЕРДІ ТАНУ: САЛЫСТЫРМАЛЫ ЗЕРТТЕУ

Әбдіқалық Гүлназым — Астана Халықаралық университетінің магистранты. 010000. Астана, Қазақстан
E-mail: gulnazymabdikalik@gmail.ru. ORCID ID: https://orcid.org/0009-0008-5216-0707;
Әсел Мұқанова — Астана Халықаралық университетінің доценті, PhD. 010000. Астана, Қазақстан
E-mail: asiserikovna@gmail.com. ORCID ID: https://orcid.org/0000-0002-8964-3891;
Назырова Айжан Есболовна — Астана Халықаралық университетінің аға оқытушысы, 010000. Астана, Қазақстан
E-mail: ayzhan.nazyrova@gmail.com. ORCID ID: https://orcid.org/0000-0002-9162-6791.

Аннотация. Мәтіндегі аталған объектілерді тану және санаттау әрекеті аталған объектілерді тану (NER) деп аталады және табиғи тілді өңдеу үшін үлкен маңызға ие. Алайда, морфологияның күрделілігіне және лингвистикалық ресурстардың тапшылығына байланысты, қазақ сияқты ресурстары шектеулі тілдердегі аталған объектілерді тану белгілі бір қиындықтармен байланысты. Бұл ғылыми жұмыста ner үшін машиналық оқытудың екі белгілі моделінің — шартты кездейсоқ өрістердің (CRF) және кездейсоқ орманның қазақ тіліндегі тиімділігін салыстыру жүргізілді. Жұмыста қазақ тілінің морфологиялық күрделілігін ескере отырып, белгілерді іріктеу стратегиялары қарастырылады және ner қазақ тілі үшін арнайы жасалған эталондық деректер жиынтығы пайдаланылады. Random Forest модельдері жоғары өлшемді белгілер кеңістігін және деректердегі күрделі өзара әрекеттесуді басқарса, CRF

модельдері дәйекті тәуелділіктер мен контекстік ақпаратты көрсетеді. Қазақ NER үшін CRF және Random Forest модельдерінің тиімділігі эксперименттік нәтижелермен расталады. Алайда, бұл модельдердің тиімділігіне таңбаланған деректердің жетіспеушілігі әсер етеді. Зерттеудің болашақ бағыттары осы мәселені шешу үшін лингвистермен және ана тілінде сөйлейтіндермен ынтымақтастық арқылы аннотацияланған деректер жиынтығын кеңейтуді қамтиды. Зерттеу сонымен қатар NER-де қазақ тілінің күрделі морфологиясын ескеру қаншалықты маңызды екенін көрсетеді. CRF және Random Forest модельдерін ескеретін морфологиялық қасиеттердің ішінде сөздерде сөйлеу бөліктерінің сабақтары мен тегтерінің болуын атап өтуге болады, бұл әртүрлі ауытқулар мен нұсқаларда аталған нысандарды тануды жақсартады. Салыстырмалы талдау қазақстандық NER үшін Random Forest және CRF модельдерінің артықшылықтары мен кемшіліктеріне жарық түсіреді. Random Forest модельдері күрделі байланыстар мен белгілердің өзара әрекеттесуін басқара алатын болса, CRF модельдері дәйекті тәуелділіктерді анықтауда және контекстік ақпаратты пайдалануда жақсы жұмыс істейді. NER тапсырмасының нақты қажеттіліктері мен сипаттамалары модель таңдауын анықтайды. Қорытындылай келе, CRF және Random Forest модельдерінің тиімділігі туралы ақпаратты ұсына отырып, бұл салыстырмалы зерттеу Қазақ тілі үшін NER дамуына ықпал ететінін атап өткен жөн. Ол морфологиялық күрделілікпен жұмыстың құндылығын, аннотацияланған деректердің қажеттілігін көрсетеді және қазақ тілінің NER жүйелерін жетілдіруге бағытталған одан әрі зерттеулерді бағыттайды.

Түйін сөздер: Аталған объектілерді тану (NER), қазақ тілі, шартты кездейсоқ өрістер (CRF), кездейсоқ орман, салыстырмалы зерттеу

Бұл жұмысты Қазақстан Республикасы Ғылым және жоғары білім министрлігінің Ғылым комитеті қаржылай қолдады (грант AP19577922).

© Г. Абдикалык*, А. Муканова, А. Назырова, 2023
Международный университет Астана, Астана, Казахстан.
E-mail: gulnazymabdikalik@gmail.ru

РАСПОЗНАВАНИЕ ИМЕНОВАННЫХ ИМЕНОВАННЫХ ОБЪЕКТОВ В КАЗАХСКОМ ЯЗЫКЕ С ПОМОЩЬЮ МОДЕЛЕЙ CRF И RANDOM FOREST: СРАВНИТЕЛЬНОЕ ИССЛЕДОВАНИЕ

Абдикалык Гульназым — магистрант Международного университета Астаны. 010000. Астана, Казахстан
E-mail: gulnazymabdikalik@gmail.ru. ORCID ID: https://orcid.org/0009-0008-5216-0707;
Муканова Асель — PhD, доцент Международного университета Астаны. 010000. Астана, Казахстан
E-mail: asiserikovna@gmail.com. ORCID ID: https://orcid.org/0000-0002-8964-3891;
Назырова Айжан Есболовна — старший преподаватель Международного университета Астаны. 010000. Астана, Казахстан
E-mail: ayzhan.nazyrova@gmail.com. ORCID ID: https://orcid.org/0000-0002-9162-6791.

Аннотация. Деятельность по распознаванию и категоризации именованных объектов в тексте называется распознаванием именованных объектов (NER) и имеет большое значение для обработки естественного языка. Однако из-за сложности морфологии и скудности лингвистических ресурсов распознавание именованных объектов в языках с ограниченными ресурсами, таких как казахский, сопряжено с определенными трудностями. В данной научной работе проведено сравнение эффективности двух известных моделей машинного обучения - условных случайных полей (CRF) и случайного леса — для NER на казахском языке. В работе рассматриваются стратегии подбора признаков с учетом морфологической сложности казахского языка и используется эталонный набор данных, созданный специально для NER казахского языка. В то время как модели Random Forest управляют высокоразмерными пространствами признаков и сложными взаимодействиями в данных, модели CRF отражают последовательные зависимости и контекстную информацию. Эффективность моделей CRF и Random Forest для казахского NER подтверждается экспериментальными результатами. Однако на эффективность работы этих моделей влияет нехватка помеченных данных. Будущие направления исследований включают расширение аннотированных наборов данных за счет сотрудничества с лингвистами и носителями языка для решения этой проблемы. В исследовании также подчеркивается, насколько важно учитывать в NER сложную морфологию казахского языка. Среди морфологических качеств, которые учитывают модели CRF и Random Forest, можно выделить наличие в словах стеблей и тегов частей речи, что улучшает распознавание именованных Именованных объектов в различных склонениях и вариантах. Сравнительный анализ проливает свет на преимущества и недостатки моделей Random Forest и CRF для казахстанской NER. В то время как модели Random Forest могут управлять сложными связями и взаимодействием признаков, модели CRF лучше справляются с выявлением последовательных зависимостей и использованием контекстной информации. Специфические потребности и характеристики задачи NER определяют выбор модели. В заключение следует отметить, что, предлагая информацию об эффективности моделей CRF и Random Forest, данное сравнительное исследование способствует развитию NER для казахского языка. Оно показывает ценность работы с морфологической сложностью, необходимость аннотированных данных и направляет дальнейшие исследования, направленные на совершенствование систем NER казахского языка.

Ключевые слова: Распознавание именованных объектов (NER), казахский язык, условные случайные поля (CRF), случайный лес, сравнительное исследование.

Эта работа была финансово поддержана Комитетом науки Министерства науки и высшего образования Республики Казахстан (грант AP19577922).

Introduction

The task of identifying and categorizing named entities, such as names of people, companies, places, and times in text that is unorganized, is known as named entity recognition (NER) (Sang et al., 2003). It is important in many applications, such as text mining (Cheng et al., 2020), information retrieval, and question-answering (Aliod et al., 2006). Despite the fact that NER has been thoroughly explored for major languages, the lack of linguistic resources and the intricacy of morphology make extracting named entities from under-resourced languages particularly difficult.

The Kazakh language, which is a Turkic language with limited resources and is mostly spoken in Kazakhstan and its nearby territories (Eryiğit et al., 2013), is the subject of this research article. The Kazakh language possesses a wide range of morphological traits, such as intricate inflections, agglutination, and different grammatical constructions. Given that named entities can differ greatly in form and context, these linguistic complexities make it difficult to effectively identify and categorize them.

This study examines the performance of Conditional Random Fields (CRF) and Random Forest, two well-known machine learning models, to address the NER difficulties in the Kazakh language. A sequence labeling approach called CRF uses contextual data to accurately identify entities while also capturing sequential interdependence. The Random Forest ensemble learning method, on the other hand, mixes various decision trees to generate predictions, managing high-dimensional feature spaces and capturing complicated relationships in the data.

The objective of the comparison study is to assess the effectiveness of Random Forest and CRF models for NER in Kazakh. Comparing the models is part of the study, which also covers other NER-related topics including entity detection. We evaluate the precision of named entity extraction from Kazakh texts using these models on a benchmark Kazakh NER dataset.

The results of this study, which specifically focused on the Kazakh language, help us comprehend NER in languages with little resources. They provided information on the effectiveness and applicability of CRF and Random Forest models for NER tasks in Kazakh, as well as their advantages and disadvantages. The findings will aid researchers and practitioners in selecting suitable models for NER in the Kazakh language and provide a foundation for further improvements in NER methods for languages with limited resources.

In the parts that follow, we will go over the methodology for the comparative analysis as well as a summary of similar work in NER for languages with limited resources. Following a presentation and analysis of the experimental findings, we will talk about the implications and potential future developments for NER in Kazakh.

Ner challenges in Kazakh language. The activity of recognizing and categorizing named items in text is known as named entity recognition (NER), and it is essential to natural language processing. However, due primarily to the language's

morphological complexity, extracting named items from under-resourced languages (Yeniterzi, 2011) like Kazakh presents special difficulties. We go into the special difficulties posed by morphological complexity in NER for the Kazakh language in this section.

The Kazakh language possesses a wide variety of grammatical features, such as intricate inflections, agglutination, and rich morphological traits (Tolegen et al., 2019). The morphological intricacy that makes accurate NER challenging is exacerbated by these linguistic traits (Yergesh et al., 2016). Named Entity Recognition (NER), also known as named entity identification and classification in text, is a crucial problem in natural language processing (Nadeau et al., 2007). The scarcity of annotated data, however, is one of the biggest obstacles to NER for the Kazakh language.

1. Complex Irregularities

The inflection system used in Kazakh is intricate and includes different verb conjugations, noun declensions, and adjectival agreements. Each entity has multiple word forms as a result of these inflections, necessitating the use of NER algorithms to take these various inflected versions into consideration (Iskhakova et al., 2020).

2. Agglutination

In order to express grammatical information, affixes are frequently appended to the root word in Kazakh, a process known as agglutination. The extensive and complex word forms that result from this agglutinative nature make it difficult to effectively identify and extract named items (Bilakhanova et al., 2023).

3. The use of language Features

A wide range of grammatical traits, such as case markers, tense markers, and person markers, are used in Kazakh. The interplay between these markers and their effects on named entity recognition must be taken into account because these traits add to the language's morphological complexity (Tolegen et al., 2023).

4. Lack of Annotated Data

The availability of high-quality Annotated Corpora is crucial for the creation of accurate and reliable NER models. Nevertheless, finding enough annotated data for languages with limited resources, like Kazakh, is a considerable difficulty (Bogdanchikov et al., 2022). The availability of Kazakh-specific annotated datasets frequently restricts the range for developing and testing NER models.

Conditional random fields for kazakh ner. The activity of recognizing and categorizing named items in text is known as named entity recognition (NER), and it is essential to natural language processing. For sequence labeling tasks like NER, Conditional Random Fields (CRF) is a well-liked and efficient machine learning model that is frequently utilized.

A probabilistic graphical model called Conditional Random Fields is employed for applications involving sequence labeling. It is a discriminative model that takes into account the relationships between adjacent labels in a sequence, so capturing the contextual data required for precise labeling (Lample et al., 2016). For NER tasks, where the types and limits of named entities frequently depend on the context of the words around them, CRF models are particularly well suited.

Random forest for kazakh ner. A well-liked machine learning technique called Random Forest is renowned for its capacity to manage high-dimensional feature spaces and capture intricate relationships in the data (Belgiu et al., 2016). An ensemble learning technique called Random Forest uses several decision trees to produce predictions. It is frequently used for classification and regression problems and is a member of the family of tree-based algorithms. The way Random Forest works is by using various subsets of the training data to train a group of decision trees, then combining those predictions to get the outcome.

A set of features taken from the input data are used as the basis for the Kazakh Random Forest models for NER. Word IDs, part-of-speech tags, morphological characteristics, and contextual information from nearby words are a few examples of these features. Multiple decision trees are constructed during the training phase utilizing various feature and training data subsets. To ensure variation among the different trees, each decision tree learns to make predictions based on a random selection of features.

Corpus structure. As a dataset were utilized part of the train data which is open source in the website github (https://github.com/IS2AI/KazNERD/tree/main), training of the models was done on data in Kazakh. The dataset contained 132 phrases, 831 words, and 1489 characters in total. IOB2 tags (Sang et al., 1999) were used to label the data, as seen in Figure 1. Tagger that uses the IOB format, where chunks are labeled by their appropriate category, can be created that tags every word in a sentence.

B - for the word in the Beginning chunk

I - for words Inside the chunk

O - Outside any chunk

	Tokens	IOB2
0	Еуразиялық	B-LAW
1	экономикалық	I-LAW
2	одақ	I-LAW
3	туралы	I-LAW
4	шартқа	I-LAW

Fig. 1. Example of IOB2 scheme labeled data

Unexpectedly, the majority of words are categorized as not belonging to any block. These words can be thought of as placeholders, and the classifier's performance may be impacted by their use. Let's run a different test on the dataset devoid of O tags, you can see this from Figure 2.

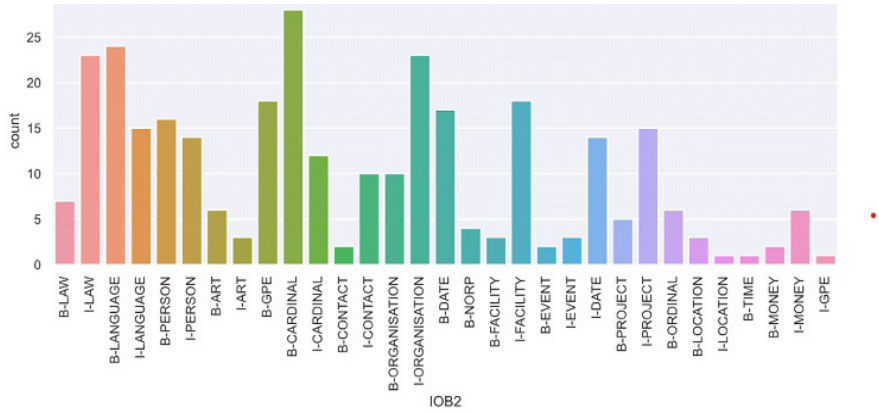

Fig. 2. Words distribution without O tag

Modeling the data. In this section, we propose a comparison study to assess Conditional Random Fields (CRF) and Random Forest, two well-known machine learning models, for NER in the Kazakh language. We use a benchmark dataset designed especially for Kazakh NER to conduct an extensive comparison. The dataset comprises of Kazakh-language writings that have been annotated and span a variety of topics and item kinds. In order to ensure a suitable distribution of entities throughout the sets, we separate the dataset into training, validation, and test sets.

We use feature engineering to gather pertinent data for NER in Kazakh for both CRF and Random Forest models. Word identities, part-of-speech labels, and contextual elements from nearby words are included in this. We use common assessment metrics like precision, recall, and F1-score to evaluate the effectiveness of the CRF and Random Forest models. These metrics reveal how well the models are able to recognize and categorize named things in Kazakh texts. Before we move on to the modeling portion, it is essential to understand the performance indicators that will be used to evaluate the models. We will assess the models using the precision (1), recall (2) and F1 score (3), metrics because we are dealing with information extraction.

In order to generate the metrics described above, True/False positives and True/False negatives are used, respectively.

• True Positives (TP) are successfully predicted positive values, which means that both the actual and projected classes have the same value.

• True Negatives (TN) are successfully predicted negative values, which indicates that both the actual and projected class values are negative.

• False Positives (FP) are when the expected class is present but the actual class is absent.

• False Negatives (FN) are when the expected class is no when the actual class is yes.

$$\text{Precision} = \frac{tp}{tp+fp}, \quad (1)$$

$$\text{Recall} = \frac{tp}{tp+fn}, \quad (2)$$

$$\text{F-1} = 2* \frac{Precision*Recall}{Precision+Recall} \quad (3)$$

On the test set, we compute the metrics, and we contrast the outcomes of the two models. The outcomes of the performance of the RF and CRF models are also shown in the pictures below. (Shown in Figure 3.)

	precision	recall	f1-score	support
B-ART	0.00	0.00	0.00	6
B-CARDINAL	0.43	0.36	0.39	28
B-CONTACT	0.00	0.00	0.00	2
B-DATE	0.00	0.00	0.00	17
B-EVENT	0.00	0.00	0.00	2
B-FACILITY	0.00	0.00	0.00	3
B-GPE	0.20	0.22	0.21	18
B-LANGUAGE	0.00	0.00	0.00	24
B-LAW	0.00	0.00	0.00	7
B-LOCATION	0.00	0.00	0.00	3
B-MONEY	0.00	0.00	0.00	2
B-NORP	0.67	0.50	0.57	4
B-ORDINAL	0.50	0.17	0.25	6
B-ORGANISATION	0.00	0.00	0.00	10
B-PERSON	0.00	0.00	0.00	16
B-PROJECT	0.00	0.00	0.00	5
B-TIME	0.00	0.00	0.00	1
I-ART	0.00	0.00	0.00	3
I-CARDINAL	0.00	0.00	0.00	12
I-CONTACT	0.00	0.00	0.00	10
I-DATE	0.00	0.00	0.00	14
I-EVENT	0.00	0.00	0.00	3
I-FACILITY	0.00	0.00	0.00	18
I-GPE	0.00	0.00	0.00	1
I-LANGUAGE	0.00	0.00	0.00	15
I-LAW	0.00	0.00	0.00	23
I-LOCATION	0.00	0.00	0.00	1
I-MONEY	0.00	0.00	0.00	6
I-ORGANISATION	0.29	0.17	0.22	23
I-PERSON	0.00	0.00	0.00	14
I-PROJECT	0.00	0.00	0.00	15
O	0.82	0.99	0.90	1178
accuracy			0.79	1490
macro avg	0.09	0.08	0.08	1490
weighted avg	0.67	0.79	0.72	1490

	precision	recall	f1-score	support
B-ART	1.00	0.33	0.50	6
B-CARDINAL	0.80	1.00	0.89	28
B-CONTACT	0.00	0.00	0.00	2
B-DATE	1.00	1.00	1.00	17
B-EVENT	0.00	0.00	0.00	2
B-FACILITY	0.00	0.00	0.00	3
B-GPE	0.75	1.00	0.86	18
B-LANGUAGE	0.96	1.00	0.98	24
B-LAW	0.67	0.29	0.40	7
B-LOCATION	0.00	0.00	0.00	3
B-MONEY	0.00	0.00	0.00	2
B-NORP	0.00	0.00	0.00	4
B-ORDINAL	1.00	1.00	1.00	6
B-ORGANISATION	0.77	1.00	0.87	10
B-PERSON	0.70	1.00	0.82	16
B-PROJECT	1.00	0.40	0.57	5
B-TIME	0.00	0.00	0.00	1
I-ART	0.00	0.00	0.00	3
I-CARDINAL	1.00	0.75	0.86	12
I-CONTACT	1.00	1.00	1.00	10
I-DATE	1.00	1.00	1.00	14
I-EVENT	1.00	0.33	0.50	3
I-FACILITY	0.86	1.00	0.92	18
I-GPE	0.00	0.00	0.00	1
I-LANGUAGE	0.00	0.00	0.00	15
I-LAW	0.79	1.00	0.88	23
I-LOCATION	0.00	0.00	0.00	1
I-MONEY	1.00	1.00	1.00	6
I-ORGANISATION	0.79	1.00	0.88	23
I-PERSON	0.93	1.00	0.97	14
I-PROJECT	0.65	1.00	0.79	15
O	1.00	1.00	1.00	1178
accuracy			0.96	1490
macro avg	0.58	0.57	0.55	1490
weighted avg	0.95	0.96	0.95	1490

Fig. 3. RF and CRF model performance

The experimental results demonstrate that both CRF and Random Forest models compete favorably for Kazakh NER. While CRF is excellent at capturing sequential dependencies, Random Forest does a better job of covering high-dimensional feature spaces.

The model performed poorly even though it had a high average score with an f-1 score of 0.79. For the majority of the classes, precision and recall levels were 0. It appears that the model lacks the characteristics required to make wise selections. The methodology just calls for memory of words and tags, which is insufficient. For the algorithm to produce more precise predictions, the context information surrounding each word must also be supplied. As the scores increased and the

f-1 score reached 0.96, the CRF classifier surpassed the Random Forest classifier. For each class, the precision and recall measurements, however, have only slightly increased. Perhaps the model is repeating words and not fully accounting for context.

Limitations and directions for the future. Limitations of CRF and Random Forest models include the necessity for annotated data and issues handling the complicated morphology of Kazakh. Expanding annotated datasets, analyzing deep learning models, and researching transfer learning and domain adaptation techniques in Kazakh NER should be the key objectives of future research. Kazakh's morphological complexity poses difficulties for NER since substantial changes and inflections in named entities may not be properly captured by CRF and Random Forest models. Future research should look at sophisticated feature engineering techniques and Kazakh-specific language resources to increase morphological complexity. As there are just a few annotated corpora developed specifically for Kazakh NER, the availability of annotated data is a significant limitation. Future research should concentrate on improving the usability of annotated datasets and collaborating with linguists, academics, and native speakers to solve this problem.

Conclusion

The use of Conditional Random Fields (CRF) and Random Forest models for Named Entity Recognition (NER) in the Kazakh language is examined in this study. its performance is assessed, and its advantages, disadvantages, and potential directions are covered in a comparison analysis. The results demonstrate that while Random Forest models manage high-dimensional feature spaces and intricate interactions in the data, CRF models excel at capturing sequential dependencies and contextual information. Both models successfully extract named entities from texts written in Kazakh. However, it is difficult to create reliable and accurate models because there is a lack of annotated data for training these models.

To improve NER performance in Kazakh, future research should concentrate on increasing the accessibility of annotated datasets, investigating domain adaptation strategies, and applying deep learning models. The performance of NER systems can be greatly enhanced by integrating linguistic resources and expertise particular to the Kazakh language. The advancement of NER in Kazakh depends on using unsupervised and semi-supervised learning methodologies and assessing models on various domains and text genres. Overall, the comparison study offers insightful information about the effectiveness of CRF and Random Forest models for Named Entity Recognition in Kazakh.

REFERENCES

Aliod D.M., van Zaanen M. and Smith D. (2006). Named entity recognition for question answering. In Proceedings of the Australasian Language Technology Workshop (ALTA). Pp. 51–58. ALTA. (in Eng.).

Belgiu M. and Dragut L. (2016). Random Forest in Remote Sensing: A Review of Applications and Future Directions. ISPRS Journal of Photogrammetry and Remote Sensing, 114, 24–31. https://doi.org/10.1016/j.isprsjprs.2016.01.011. (in Eng.).

Bilakhanova A., Ydyrys A. & Sultanova N. (2023). KAZAKH LANGUAGE-BASED QUESTION ANSWERING SYSTEM USING DEEP LEARNING APPROACH. Suleyman Demirel University Bulletin: Natural And Technical Sciences, 62(1), 113–121. doi:10.47344/sdubnts.v62i1.974. (in Eng.).

Bogdanchikov A., Ayazbayev D., Varlamis I. (2022). Classification of Scientific Documents in the Kazakh Language Using Deep Neural Networks and a Fusion of Images and Text. Big Data Cogn. Comput. 2022, 6, 123. https://doi.org/10.3390/bdcc6040123. (in Eng.).

Cheng P. and Erk K. (2020). Attending to entities for better text understanding. In Proceedings of the AAAI Conference on Artificial Intelligence, 34. Pp. 7554–7561. (in Eng.).

Eryiğit G., Çetin F., Yanık M., Temel T., Çiçekli I. (2013). TURKSENT: a sentiment annotation tool for social media. In: Proceedings of the 7th Linguistic Annotation Workshop & Interoperability with Discourse, ACL 2013, Sofia, Bulgaria. (in Eng.).

Iskhakova G.R., Kirillova Z.N. & Yerbulatova I.K. (2020). Paired Combinations in Kazakh Language and its Ways of Translation into Russian. Utopía Y Praxis Latinoamericana, 25(1). Pp. 396-401. (in Eng.).

Lample G., Ballesteros M., Subramanian S., Kawakami K. and Dyer C. (2016). "Neural architectures for named entity recognition", arXiv vol. 1603.01360: 1–1. (in Eng.).

Nadeau D., Sekine S. (2007). A survey of named entity recognition and classification. In: Lingvisticae Investigationes. (in Eng.).

Sang E.F.T.K. and Meulder F.D. (2003). Introduction to the CoNLL-203 shared task: Language-independent named entity recognition. In Proceedings of the Conference on Natural Language Learning (CoNLL) at HLT-NAACL. Pp. 142–147. ACL. (in Eng.).

Sang E.F.T.K. and Veenstra J. (1999). Representing text chunks. In Proceedings of the European Chapter of the Association for Computational Linguistics (EACL). Pp. 173–179. ACL. (in Eng.).

Tolegen G., Toleu A., Mamyrbayev O., Mussabayev R. (2023). Neural Named Entity Recognition for Kazakh. In: Gelbukh, A. (eds) Computational Linguistics and Intelligent Text Processing. CICLing 2019. Lecture Notes in Computer Science. Vol. 13452. Springer, Cham. https://doi.org/10.1007/978-3-031-24340-0_1. (in Eng.).

Tolegen G., Toleu A., Mamyrbayev O. and Mussabayev R. (2019). "Named Entity Recognition for Kazakh Using Conditional Random Fields". CICLing2019: Springer Lecture Notes in Computer Science (La Rochelle, France, 2019). (in Eng.).

Yeniterzi Reyyan (2011). Exploiting Morphology in Turkish Named Entity Recognition System. In Proceedings of the ACL 2011 Student Session. Pp. 105–110. (in Eng.).

Yergesh B., Sharipbay A., Bekmanova G., Lipnitskii S. (2016). Sentiment analysis of Kazakh phrases based on morphological rules. J. Kyrgyz State Tech. Univ. Named After I. Razzakov. Theor. Appl. Sci. Tech. J. 2(38). Pp. 39–42. (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