EasyChair Preprint

№ 8101

Machine Learning for Syntactic and Morphological Analysis of Text in the Kazakh Language

Saule Kudubayeva, Botagoz Zhusupova and Meruyert Salkenova

EasyChair preprints are intended for rapid dissemination of research results and are integrated with the rest of EasyChair.

May 28, 2022

Kudubaeva S.A.[1], Zhusupova B. T.[2], Salkenova M. K[3].
[1]ENU them. L.N. Gumilyov, Nur-Sultan, Kazakhstan
[2]KRU them. A. Baitursynov, Kostanay, Kazakhstan

MACHINE LEARNING FOR SYNTAX AND MORPHOLOGICAL ANALYSIS OF TEXT IN THE KAZAKH LANGUAGE

Abstract. The article describes the possibility of analyzing texts in the Kazakh language using machine learning. Machine learning is used in the recognition of machine and handwritten text, speech and images. In connection with the problem of determining the meaning of words, syntactic and morphological analysis of the text is used, which are interrelated and allow dividing the text into tokens, word forms are formed. The implementation of the task is complicated by a large number of alternative options that arise in the process of parsing, related both to the ambiguity of the input data (the same word form can be obtained from different typical forms) and the ambiguity of the parsing rules themselves. The work is carried out in order to expand the tasks and possibilities of use related to the text: improving translation from Kazakh into other languages, including sign language.

Key words: *machine learning, tokenization, artificial intelligence, lemmatization*

Кудубаева С. А. [1], Жусупова Б. Т. [2], Салькенова М. К.[3]
[1]ЕНУ им. Л.Н. Гумилева, г.Нур-Султан, Казахстан
[2]КРУ им. А.Байтурсынова, г.Костанай, Казахстан

МАШИННОЕ ОБУЧЕНИЕ ДЛЯ СИНТАКСИЧЕСКОГО И МОРФОЛОГИЧЕСКОГО АНАЛИЗА ТЕКСТА НА КАЗАХСКОМ ЯЗЫКЕ

Аннотация. В статье описывается возможность анализа текстов на казахском языке при помощи машинного обучения. Машинное обучение используется при распознании машинного и рукописного текста, речи и изображений. В связи с проблемой определения смысла слов, применяется синтаксический и морфологический анализ текста, которые имеют взаимосвязь и позволяют разделить текст на токены, образуются словоформы. Реализация задачи осложняется большим числом альтернативных вариантов, возникающих в процессе разбора, связанных как с многозначностью входных данных (одна и та же словоформа может быть получена от разных типичных форм), так и неоднозначностью самих правил разбора. Работа проводится с целью расширения задач и

возможностей использования, связанных с текстом: улучшение перевода с казахского языка на другие, в том числе на язык жестов.
Ключевые слова: *машинное обучение, токенизация, искусственный интеллект, лемматизация.*

В последние годы благодаря развитию искусственного интеллекта и в частности машинного обучения открываются большие возможности для реализации различных проектов, направленных на решение задач в разных сферах жизни человека. Перед машинным обучением ставятся всё больше и больше задач по анализу больших данных.

Многие крупные компании финансируют в машинное обучение, так как эта технология по-настоящему имеет хорошие перспективы и человеку недостаточно собственных сил для анализа больших данных, которые ИИ обрабатывает лучше и быстрее. Уже сейчас часть процессов в различных учреждениях автоматизируется и управляется за счет самообучающихся алгоритмов.

Машинное обучение используется при распознании машинного и рукописного текста, речи и изображений. Но на данный момент многие работы только начинают своё развитие в этом направлении или ведутся в основном с текстами на английском и русских языках, в связи с этим существует проблема: распознавание текста на других языках и в частности касается анализа текста на казахском языке, где необходимо выявить в них последовательности и определенные закономерности для получения результата.

Таким образом существует проблема методов машинного обучения для анализа текста на казахском языке и в данной статье описывается синтаксический и морфологический анализ текста с целью расширения возможностей использования их для задач, связанных с текстом: улучшение перевода с казахского языка на другие, в том числе на язык жестов.

Для работы был изучен ряд материалов, состоящий из статей и научных работ различных авторов.

Для того, чтобы понять каким образом в целом происходит синтаксический и морфологический анализы текстов был изучен ряд работ и выделены основные приемы. Для предсинтаксического анализа важно знать формообразовании естественных языков, способы представления правил формообразования и алгоритм, обеспечивающий высокую скорость анализа форм слов, структура представления знаний о формообразовании и приведен пример описания формообразования слов. Для морфологических словарей казахского языка на основе корпуса размеченных текстов главной проблемой обозначается определение роли слов и их связи между собой, а результатом этого этапа является набор деревьев, показывающих такие связи. Выполнение подобных задач

осложняется огромным количеством альтернативных вариантов, возникающих в ходе разбора, связанных как с многозначностью входных данных, так и неоднозначностью самих правил разбора. Таким образом можно приблизиться к решению задачи по проведению синтаксической сегментации текста.

Важно отметить, что все виды анализов текста имеют взаимосвязь, так как так или иначе приходится обращаться как морфологии слова, так и его синтаксическому значению в предложении. Задачей предсинтаксического анализа является подготовка данных для синтаксического анализа в наиболее удобной форме, максимально облегчающей выполнение задачи последнему, а также обработке словосочетаний. Выделяются следующие виды словосочетаний: неразрывные, неизменяемые, неразрывные изменяемые и разрывные.

В применении машинного обучения в задачах анализа текста особое внимание уделяется методам, которые можно эффективно использовать для извлечения информации из текста на естественном языке. Проводятся различные этапы и уровни анализа текста и используются методы машинного обучения на каждом из них.

В качестве дополнительных источников были взяты Интернет-ресурсы. На страницах справочника Microsoft описываются модули текстовой аналитики, входящие в Машинное обучение. Эти модули предоставляют специализированные вычислительные средства для работы с структурированным и неструктурированным текстом, в том числе:

– несколько параметров для предварительной обработки текста;

– определение языка;

– создание компонентов из текста с помощью настраиваемых словарей;

– хэширование компонентов для эффективного анализа текста без предварительной обработки или расширенного лингвистического анализа;

– vowpal Wabbit для очень быстрого машинного обучения в тексте;

– распознавание именованных сущностей для извлечения имен людей, мест и организаций из неструктурированного текста.

Из статьи было изучены виды синтаксического анализа, он подразделяется на поверхностный и полный. Поверхностный анализ (hunking) предназначен для выделения смысловых составляющих, таких, как именная группа, глагольная группа, предложная группа. Эти семантические единицы не пересекаются, не рекурсивны и не избыточны.

В задачах машинного обучения в аналитике была определена следующая проблема – способ сбора и приведения к нужному формату необходимой информации для дальнейшего обучения, а

также рассмотрены основные положения из теории машинного обучения, затронуты основные проблемы сбора и анализа данных, предложено решение всех ранее перечисленных проблем.

Предсинтаксический анализ отвечает за две противоположные задачи: объединение отдельных лексических единиц в одну синтаксическую единицу или, наоборот, деление ее на несколько. В одной синтаксической единице объединяются взаимозаменяемые интегральные выражения. Деление слов следует проводить особенно в тех случаях, когда несколько взаимосвязанных произвольных слов можно объединить в сложное слово, а размещение таких сочетаний в морфологическом анализе невозможно. Еще одна задача предсинтетического анализа – проведение синтаксической сегментации. Его задача – разбить текст строки на куски, зависящие от правил на следующем этапе – разбор, задача с экспоненциальным увеличением сложности. В связи с этим любая помощь в его проведении может привести к значительному ускорению его работы.

Синтаксический анализ – самая трудная часть анализа текста. Тут нужно определить роли слов и их связи между собой. Итогом этого этапа является комплект деревьев, показывающих такие связи. Осуществление задачи осложняется громадным числом альтернативных вариантов, возникающих в процессе разбора, связанных как с многозначностью входных данных (одна и та же словоформа может быть получена от разных типичных форм), так и неоднозначностью самих правил разбора.

Постсинтаксический анализ служит двум целям. С одной стороны, нам нужно уточнить толк, заложенный в слова и выраженный с помощью разных средств языка: предлогов, префиксов либо аффиксов, создающих ту либо другую словоформу. С иной стороны, одна и та же мысль может быть выражена разными конструкциями языка. В случае с многоязыковой диалоговой системой, одну и ту же мысль можно выразить разными синтаксическими конструкциями. В связи с этим дерево нужно нормализовать, т.е. конструкция, выражающая некоторое действие разным образом для разных языков либо обстановок, обязана быть сведена к одному и тому же нормализованному дереву. Помимо того, на этом же этапе может проводиться обработка разрывных изменяемых словосочетаний, в которых слова словосочетания могут изменяться и могут быть поделены, другими словами.

Предсинтаксический анализ нужен для выделения элементов текста, морфологического анализа этих элементов, распределения сложносоставных элементов на части, объединение примитивных связанных по смыслу элементов в группы, выделение фрагментов текста, которые могут разбираться самостоятельно. Его наименование показывает место предсинтаксического анализа в общей системе: перед синтаксическим анализом. Задачей

предсинтаксического анализа является подготовка данных для синтаксического анализа в особенно комфортной форме, максимально облегчающей осуществление задачи последнему. На вход системы поступает текст. В 1-ю очередь нужно определить единицы этого текста: абзацы, предложения, отдельные слова и знаки препинания. В различие от систем машинного перевода, диалоговым системам нет необходимости выделять заголовки, сноски, комментарии, врезки и прочие элементы текста, нужные для сохранения форматирования.

Выделение абзацев в новейших редакторах является банальной задачей. В них уже существует разметка на абзацы. При абсолютно текстовом вводе абзацы зачастую отмечаются символом перевода строки. В самом начале абзаца зачастую ставят два и больше пробела либо пробельную строку. В случае, когда всякая строка текста оканчивается символами конца строки, задача выделения абзаца может затребовать особых умений о структуре данного текста. Задача выделения предложений менее банальна. Обыкновенно предложение заканчивается точкой, вопросительным либо восклицательным знаком, изредка – многоточием. Впрочем, на практике те же знаки препинания применяются и для других целей. Точка зачастую используется в сокращениях. При этом если сокращение доводится на конец предложения, то ставится только одна точка, относящаяся как к сокращению, так и к концу предложения. Восклицательный и вопросительный знаки зачастую применяются в колоритных вставках в тексте. Таким образом, знаки препинания не являются стопроцентной гарантией окончания предложения. К счастью, при общении с диалоговыми системами колоритные вставки обыкновенно не применяются, впрочем, проблемка точки остается. Еще одна проблемка на данном этапе – это слова, написанные с большой буквы, после точки. Так в предложении без применения прагматики и контекста не вовсе внятно о ком идет речь: о каком-либо предмете или имени/кличке живого существа. В 1-ом случае однозначно рассматривается начало предложения, во втором есть шанс, что предшествующая точка могла стоять после сокращения. Еще одну проблему представляют собой цитаты и прямая речь, также зачастую используемые при общении с диалоговыми системами. В состав цитаты может входить как некоторое количество слов, так и некоторое количество предложений. Прямая речь обыкновенно содержит определенный связанный отрывок текста. В связи с этим и разбирать их лучше, как обособленный текст. В текстах, пытающихся подчеркнуть простонародность речи, сокращения (и как итог апострофы) встречаются очень часто. В связи с этим проблемка выделения начала и конца прямой речи стоит остро.

Отдельной вопросом являются тире и дефис. В 2-хбайтных кодировках они различаются, и системы редактирования текстов имеют возможность ставить их в необходимых местах. Впрочем, однобайтные кодировки знают только один символ – минус. В связи с этим в однобайтной кодировке (либо в случае ошибок пользователя) отличить «күндіз-түні» от «күндіз-түні» в предложениях можно только в итоге синтаксического анализа. Ориентироваться на присутствие пробелов в такой ситуации можно, но и в данном случае мы не имеем стопроцентной обязательства. Для решения этих и других проблемных задач, возникающих при членении текста на составляющие, применяется графематический анализ. Для работы графематического анализа нам среди прочего понадобится этап деления сложносоставных слов. Данный этап занимается тем, что делит трудное слово, составленное из нескольких, на составляющие его, к примеру, «аппақ».

Еще одной задачей предсинтаксического анализа является обработка словосочетаний. Можно выделить последующие виды словосочетаний: необрывные неизменяемые, неразрывные изменяемые и разрывные. Неразрывные неизменяемые словосочетания состоят из одних и тех же словоформ, идущих одна за иной. К примеру, «деп айтты», «тек қана» и т.д. Для поиска сходственных словосочетаний нужно попросту проанализировать входной текст. Обнаруженные словосочетания имеет толк обрабатывать дальше как цельную словоформу.

Помимо того, словосочетания можно поделить на открытые и закрытые. В закрытых словосочетаниях отдельные слова теряют личный толк и могут трактоваться только в составе словосочетания. При этом слова в открытых словосочетаниях сберегают все лексические связи и, как следствие, могут подчинять себе иные слова. При этом может протекать разрыв словосочетания.

Для анализа неразрывных неизменяемых словосочетаний нужно заблаговременно провести морфологический анализ. Дальше мы ищем требуемые словоформы в заданном порядке. При этом может быть нужно проверить согласование слов по параметрам: заданные параметры обязаны владеть одними и теми же значениями. Как уже упоминалось, прилагательное и существительное в русском языке согласуются по ряду параметров. Следственно, для словосочетаний, составленных из существительного и подчиненных ему прилагательных, нужно проверить сходственное согласование. Разрывные словосочетания – это связанные слова, между которыми могут вклиниваться иные слова. Как и в предыдущем случае, связка слов дает некоторое количество иное значение, чем просто сумма значений слов. В случае с разрывными словосочетаниям связь между словами либо очевидна, либо применяемый вид согласования не требует того, дабы слова стояли рядом. помимо этого, кроме

подчиненного слова, образующего словосочетание, к основному слову могут присоединяться и иные зависимые слова, являющиеся его неоднородными членами. В конце концов все подчиненные члены имеют право идти вперемешку. Так как слова в разрывных словосочетаниях могут быть разнесены по предложению, то сперва требуется провести синтаксический анализ предложения. Следственно, работа с разрывными словосочетаниями не может быть отнесена к предсинтаксическому анализу Анализ существующих решений рассмотрим способы и средства морфологического анализа в задачах нормализации слов в научно-технических терминах для казахского языка. Выделяют два твердо разных подхода к морфологическому анализу: способы, основанные на словарях, и бессловарные морфологические анализаторы. Остановимся на морфологических моделях, заложенных подходах к построению алгорифмов нормализации слов. Существующие подходы делятся на два класса: алгорифмы стемминга и лемматизации.

Стемминг – процесс нахождения основы слова для данного начального слова. Основа слова не постоянно совпадает с корнем. Лемматизация – процесс привода слова (словоформы) клемме (типичной форме). Дадим кое какие объясняющие определения. Лемма – типичная (словарная, каноническая) форма слова. К примеру, в казахском языке леммой для научно-технической терминологии являются: существительные – именительный падеж, одно – единственное число; прилагательные выступают в роли определений и не приобретают окончаний, а видоизменение прилагательных, выступающих в роли существительных, не отличается от видоизменения существительных; глаголы – исходная форма глагола. Словоформа – это слово, представленное в определенной грамматической форме. Лексема – слово как абстрактная единица обычного языка. В одну лексему объединяются различные парадигматические формы (словоформы) одного слова. К примеру, ақпараттандыру (известить) – лемма, ақпараттандырылған (информационный), ақпараттандырылатын (информированный) – лексема.

Морфологический анализ дает решение 2-х основных задач:

• задачи анализа – определение типичной формы слова по произвольной словоформе,

• задачи синтеза – построение всех словоформ по типичной форме. уйма знаменитых алгорифмов реализует лемматизацию (приведение к типичной форме) с использованием основы слова (алгорифма стемминга). Впрочем, тут сокрыты две трудности, характерные для казахского языка: во-первых, синтез типичной формы сильно зависит от метода обретения основы слова, а во-вторых, множество реализаций синтезирует все допустимые леммы,

не выбирая из них исключительного итога, либо останавливается на определении основы слова.

Определения нормальной формы слова для казахского языка

Различные языки имеют разные семантические и грамматические особенности, в следствие этого нередко методы, благополучно применяемые для обработки 1-го языка, демонстрируют довольно невысокую эффективность на ином языке. Так как предсинтаксический анализ тесно взаимосвязан с морфологическим, то нельзя не обозначить пути синтеза нормальной формы слова и особенности казахского языка.

Казахский язык – тюркский язык кыпчакской группы, относящийся к виду синтетического агглютинативного языка, имеет богатую и сложную морфологию. Слова в ней обычно составляют основу и добавляют к ней суффиксы окончание, которые бывают не менее двух или трех.

Необходимость введения слов в нормальную форму морфологического анализа возникла в работе с поисковыми тезаурусами с учетом языковой морфологии Казахстана в полнотекстовой базе данных информационных технологий. Привести слова в анализируемый текст в нормальную форму значительно усложняет его работу: индексация, последующие поиски информации в построенном индексе, решение задач кластеризации, автоматическое реферирование документов научной тематики.

В казахском языке словоформы образуются методом конкатенации корня и аффиксов (суффиксов и окончаний). При данном любой аффикс связан с наборами семантических симптомов и порядок прибавления аффиксов строго определен. К примеру, для имен существительных к базе текста в начале прибавляется суффикс и дальше завершение многочисленного количества, вслед за тем притяжательное завершение, дальше идет по стопам падежное завершение и лишь только впоследствии него – завершение формы спряжения (добавляется лишь только к одушевленным существительным).

Свежие словоформы образуются с учетом морфологических и семантических симптомов исходных форм грядущим образом: в начале к исходной форме текста прибавляются суффиксы; вслед за тем, двигаясь слева вправо, ориентируется категория (глухие, звонкие и т. п.) последней буковки (последнего звука) исходной формы текста для прибавления такого или же другого завершения.

Общая морфологическая конфигурация определения состава смотрится так:

Түбір (корень) + жұрнақ (суффикс) + жалғау (окончание).

На основании анализа и грамматики казахского языка возможно отметить надлежащие основные критерии казахского языка.

• В казахском языке текст не имеет возможность оканчиваться на звонкие согласные: «б», «в», «г», «ғ», «д», «ж». В данном языке имеют пространство исключения, в коих удаляется суффикс, начинающийся нагласную, а заслуживающие в конце «б», «г», «ғ» преобразуются в надлежащие буковки: «п», «қ», «к». К примеру, азбучный знак «п» на «б», «қ» на «ғ», «к» на «г».

• После жесткого слога идет по стопам жесткое завершение, впоследствии мягонького слога – податливое окончание.

• Мягкость и твердость текстов в казахском языке ориентируются наличием конкретной гласной в последнем слоге текста. К примеру, текст считается жестким, в случае если наличествуют гласные а, о, ұ, ы, я; а плавным оно делается, в случае если наличествуют гласные ә, ө, ү, і, е. Твердость или же плавность текстов коррелирует еще с наличием кое-каких согласных: текст жесткое, в случае если в нем наличествуют согласные қ и ғ, и податливое, в случае если наличествуют к и г.

• Каждое надлежащее завершение находится в зависимости от предшествующего по нескольким характеристикам. По твердости: в случае если конечный слог текста уверенный, то любое надлежащее завершение станет твердым, например как твердость еще одного завершения находится в зависимости от предшествующего. Этим образом, в случае если текст жесткое, то все завершения твердые, в случае если податливое, то мягонькие.

• Как ведомо, морфемы считаются кратчайшими означающими (семантическими) единицами языка, из коих оформляется словоформа, а дальше, в соответствии с этим, и лексема. В казахском языке завершения разделяются на 4 облика завершений. Описанные ниже завершения именно станут применяться в разрабатываемом методе определения основы слова.

Обозначим сквозь Pi – надлежащие большого количества завершений (аффиксов), для i = 1, 2, 3, 4.

P1 – большое количество 3-х буквенных завершений (окончание многочисленного числа);

P2 – большое количество завершений (притяжательные окончания);

P3 – большое количество завершений (личные окончания);

P4 –множество завершений (падежные окончания).

В табл. 1 описаны определения морфемного состава (Pi, где i = 1, 2, 3, 4):

№	Виды окончаний	Окончания
1	Окончание множественного числа – P_1	'лар', 'лер', 'дар', 'дер', 'тар', 'тер'

2	Притяжательные окончания – P_2	'ым', 'ім', 'м', 'ың', 'ің', 'ң', 'ыңыз', 'іңіз', 'ңыз', 'ңіз', 'сы', 'сі', 'ы', 'і', 'ымыз', 'іміз', 'мыз', 'міз'
3	Личные окончания – P_3	'мын', 'мін', 'бын', 'бін', 'пын', 'пін', 'сың', 'сің', 'сыз', 'сіз', 'мыз', 'міз', 'быз', 'біз', 'пыз', 'піз', 'сыңдар', 'сіңдер', 'сыздар', 'сіздер', 'м', 'ң', 'ңыз', 'ңіз', 'к', 'к', 'ңдар', 'ңдер', 'ңыздар', 'ңіздер'
4	Падежное окончание – P_4	'ның', 'нің', 'дың', 'дің', 'тың', 'тің', 'ға', 'ге', 'қа', 'ке', 'ны', 'ні', 'ды', 'ді', 'ты', 'ті', 'да', 'де', 'та', 'те', 'нан', 'нен', 'тан', 'тен', 'дан', 'ден', 'мен', 'бен', 'пен'

Таблица 1 – Определение морфемного состава

С учетом всех композиций соединения завершении аффиксной группы казахского языка были отнесены больше 750 словоизменительных аффиксов с указанием алгоритмов синтеза текстов, собственно, что дает креативные способности по расширению круга применяемых текстов и словосочетаний.

Для удобства реализации была изучена классификация завершении. Порядок правил содержит следующий вид:

A – завершение многочисленного количества + падежное завершение. B – многократное количество + собственное завершение.

C – многократное количество + притяжательное завершение.

D – многократное количество + притяжательное завершение + падежное завершение. E – многократное количество + притяжательное завершение + собственное завершение.

F – собственное завершение + завершение многочисленного количества. G – притяжательное завершение + падежное завершение.

H – притяжательное завершение + собственное завершение.

ЛИТЕРАТУРА

1. Азимбаев, Д. Ж. Искусственный интеллект и машинное обучение / Д. Ж. Азимбаев, И. А. Куан, И. В. Гулида // Вестник современных исследований. - 2019. - № 1.3 (28). - С. 6-7. – https://elibrary.ru/item.asp?id=36885190

2. Гущин Я.А. «Автоматизация машинного обучения. Оптимизация методов построения модели»

3. Джумамуратов Р.А., Разработка средств создания морфологических словарей казахского языка на основе корпуса размеченных текстов

4. Найденова К.А., Невзорова О.А., Машинное обучение в задачах обработки естественного языка: обзор современного состояния исследований\
5. Клышинский Э. С., Начальные этапы анализа текста
6. Кормалев Д. А., Приложения методов машинного обучения в задачах анализа текста
7. Кравцова Н. Е. О решении задач классификации в методах машинного обучения/Н. Е. Кравцова, А. П. Преображенский А.П. // Вестник. - 2018. - № 3 (26). - С. 116-118.
8. Розанов А.К., Пруцков А.В., Алгоритмы и методы представления знаний для предсинтаксического анализа текстов на естественных языках
9. Справочник Microsift https://docs.microsoft.com/ru-ru/azure/machine-learning/studio-module-reference/text-analytics
10. Федотов А. М., Тусупов Д. А., Самбетбаева М. А., Еримбетова А. С., Бакиева А. М., Идрисова А. И. Модель определения нормальной формы слова для казахского языка // Вестн. Новосиб. гос. ун-та. Серия: Информационные технологии. 2015. Т. 13, вып. 1. С. 107–116.
11. Флах П. Машинное обучение. — М.: ДМК Пресс, 2015. — 400 с. — ISBN 978-5-97060-273-7.

REFERENCES

1. Azimbaev, D. Zh. Iskusstvennyj intellekt i mashinnoe obuchenie / D. Zh. Azimbaev, I. A. Kuan, I. V. Gulida // Vestnik sovremennyh issledovanij. - 2019. - № 1.3 (28). - S. 6-7. – https://elibrary.ru/item.asp?id=36885190
2. Gushhin Ja.A. «Avtomatizacija mashinnogo obuchenija. Optimizacija metodov postroenija modeli»
3. Dzhumamuratov R.A., Razrabotka sredstv sozdanija morfologicheskih slovarej kazahskogo jazyka na osnove korpusa razmechennyh tekstov
4. Najdenova K.A., Nevzorova O.A., Mashinnoe obuchenie v zadachah obrabotki estestvennogo jazyka: obzor sovremennogo sostojanija issledovanij\
5. Klyshinskij Je. S., Nachal'nye jetapy analiza teksta
6. Kormalev D. A., Prilozhenija metodov mashinnogo obuchenija v zadachah analiza teksta
7. Kravcova N. E. O reshenii zadach klassifikacii v metodah mashinnogo obuchenija/N. E. Kravcova, A. P. Preobrazhenskij A.P. // Vestnik. - 2018. - № 3 (26). - S. 116-118.
8. Rozanov A.K., Pruckov A.V., Algoritmy i metody predstavlenija znanij dlja predsintaksicheskogo analiza tekstov na estestvennyh jazykah
9. Spravochnik Microsift https://docs.microsoft.com/ru-ru/azure/machine-learning/studio-module-reference/text-analytics

10. Fedotov A. M., Tusupov D. A., Sambetbaeva M. A., Erimbetova A. S., Bakieva A. M., Idrisova A. I. Model' opredelenija normal'noj formy slova dlja kazahskogo jazyka // Vestn. Novosib. gos. un-ta. Serija: Informacionnye tehnologii. 2015. T. 13, vyp. 1. S. 107–116.

11. Flah P. Mashinnoe obuchenie. — M.: DMK Press, 2015. — 400 s. — ISBN 978-5-97060-273-7.

1. Кудубаева Сауле Альжановна[1], Жусупова Ботагоз Тулегеновна[2,] Салькенова Меруерт Кабдоллаевна[3]

2. [1]Евразийский национальный университет имени Л.Н. Гумилева, Факультет информационных технологий, кафедра Технологии искусственного интеллекта, ассоциированный профессор, к.т.н.,
[2]Костанайский региональный университет имени А.Байтурсынова, Инженерно-технический институт имени А.Айтмухамбетова, кафедра Информационных систем, старший преподаватель,
[3] КГКП «Костанайский педагогический колледж», г.Костанай, Казахстан, кафедра «Математика и информационные технологии», преподаватель, магистрант 1 курса КРУ им. А.Байтурсынова.

3. [1]010008, г. Нур-Султан, ул. Пушкина, 11, учебный корпус №2, Моб.: 87759069647 Раб.: 87172709500 внутр.: 34212, e-mail: saule.kudubayeva@gmail.com,
[2]110000, г. Костанай, ул. Абая, 28, учебный корпус №2, Моб.: 87479222810, e-mail: botashazhus@gmail.com
[3,]110000, г. Костанай, ул. Быковского, 9, Моб.: 87018320687, e-mail: 19mira87@bk.ru

4. Машинное обучение для синтаксического и морфологического анализа текста на казахском языке

5. В статье описывается возможность анализа текстов на казахском языке при помощи машинного обучения. Машинное обучение используется при распознании машинного и рукописного текста, речи и изображений. В связи с проблемой определения смысла слов, применяется синтаксический и морфологический анализ текста, которые имеют взаимосвязь и позволяют разделить текст на токены, образуются словоформы. Реализация задачи осложняется большим числом альтернативных вариантов, возникающих в процессе разбора, связанных как с многозначностью входных данных (одна и та же словоформа может быть получена от разных типичных форм), так и неоднозначностью самих правил разбора. Работа проводится с целью расширения задач и возможностей использования, связанных с текстом: улучшение перевода с казахского языка на другие, в том числе на язык жестов.

6. Ключевые слова: машинное обучение, токенизация, искусственный интеллект, лемматизация/

7. УДК 004.8

1. [1]Kudubayeva Saule Alzhanovna, [2]Zhussupova Botagoz Tulegenovna, Salkenova Meruyert Kabdollaevna[3]

2. [1]L.N. Gumilyov Eurasian national University, Faculty of Information Technology, Department of Artificial Intelligence Technology, Associate Professor, candidate of technical sciences,

[2]A. Baitursynov Kostanay regional University, A.Aitmukhambetov Institute of Engineering and Technology, Department of Information Systems, Senior Lecturer.
[3]Kostanay Pedagogical College, Kostanay, Kazakhstan, Department of"Mathematics and Information Technology, teacher, master student of the 1st year of A. Baitursynov Kostanay regional University.

3. [1]010008, Nur-Sultan, Pushkin st., 11, academic building №2, Mobile: 87759069647, work phone number: 87172709500 internal: 34212, e-mail: saule.kudubayeva@gmail.com,

[2]110000, Kostanay, Abai st., 28, academic building №2, Mobile phone: 87479222810, e-mail: botashazhus@gmail.com,

[3.]110000, Kostanay, Bykovsky st, 9, Mob.: 87018320687, e-mail: 19mira87@bk.ru

4. Machine learning for syntactic and morphological analysis of text in the Kazakh language

5. The article describes the possibility of analyzing texts in the Kazakh language using machine learning. Machine learning is used in the recognition of machine and handwritten text, speech and images. In connection with the problem of determining the meaning of words, syntactic and morphological analysis of the text is used, which are interconnected and allow the text to be divided into tokens, word forms are formed. The implementation of the task is complicated by a large number of alternative options that arise during the parsing process, related both to the ambiguity of the input data (the same word form can be obtained from different typical forms) and the ambiguity of the parsing rules themselves. The work is carried out with the aim of expanding the tasks and possibilities of use related to the text: improving the translation from Kazakh to others, including sign language. Stemming is the process of finding the stem of a word for a given seed word. The stem of a word does not always match the root. Lemmatization is the process of driving a word (word form) to a terminal. Let us give some explanatory definitions. Lemma is a typical form of the word. For example, in the Kazakh language, the lemma for scientific and technical terminology is: nouns - nominative case, one - singular; adjectives act as definitions and do not acquire endings, and the modification of adjectives acting as nouns does not differ from the modification of nouns; verbs - the original form of the verb. A word form is a word presented in a specific grammatical form. A lexeme is a word as an abstract unit of ordinary language. Various paradigmatic forms (word forms) of one word are combined into one lexeme.

6. Keywords: machine learning, tokenization, artificial intelligence, lemmatization