Казахский национальный университет имени аль-Фараби
Институт информационных и вычислительных технологий КН МОН РК

УДК 004.93

На правах рукописи

МУСХИНА КУРАЛАЙ ЖЕНИСБЕКОВНА

Разработка системы анализа многоязычной текстовой информации на основе машинного обучения

6D070400 – Вычислительная техника и программное обеспечение

Диссертация на соискание ученой степени
доктора философии (PhD)

Отечественный научный консультант
доктор PhD Мамырбаев О. Ж.
Зарубежный научный консультант
доктор технических наук, профессор
Хайрова Н. Ф.

Республика Казахстан
Алматы, 2020

СОДЕРЖАНИЕ

ОБОЗНАЧЕНИЯ И СОКРАЩЕНИЯ

ACL	– Association for Computational Linguistics
BNC	– British National Corpus
CLAWS	– Constituent Likelihood Automatic Word-tagging System
IE	– Information Extraction
IR	– Information Retrieval
FE	– Fact Extraction
HMM	– Hidden Markov Model
KLC	– Kazakh Language Corpus
KNN	– K Nearest Neighbours
LSA	– Latent Semantic Analysis
ML	– Machine Learning
MT	– Machine Translation
NER	– Named Entity Recognition
NLP	– Natural Language Processing
OIE	– Open Information Extraction
PMI	– Pointwise Mutual Information
PPMI	– Positive Pointwise Mutual Information
POS	– Part of Speech
SD	– Stanford Dependencies (parser)
TM	– Translation Memory
SVM	– Support Vector Machines
RDF	– Resource Description Framework
XML	– eXtensible Marlup Language
VSM	– Vector Space Model
UD	– Universal Dependencies (parser)
АККЯ	– Алматинский корпус казахского языка
АКП	– Алгебра конечных предикатов
АОТЕЯ	– Автоматическая обработка текстов на естественном языке
ИИ	– Искусственный интелект
ПрО	– Предметная область

ВВЕДЕНИЕ

В современном полиязычном и мультикультурном мире, как никогда актуальна проблема сопряженности языков, поиск эффективных и жизнеспособных программ в области языков по консолидации обществ. [1]. Инициированный Главой государства и реализуемый в республике Казахстан проект триединства языков служит базой для информационного обмена как внутри страны, так и за ее пределами. Данный проект, рассматривая казахский язык, как государственный язык, русский язык, как язык межнационального общения и английский язык, как язык успешной интеграции в глобальную экономику, содействует активному развитию информационного сообщество внутри страны и интеграции Казахстана в мировое глобальное информационное сообщество [2]. Для осуществления международного взаимодействия, продвижения продукции, знаний и коммуникации Республики Казахстан в мировом информационном пространстве необходимо развивать как специальные информационные приложения по обработке казахского языка, так и возможности его включения в приложения мультиязычной обработки.

Анализ научных исследований в области компьютерной лингвистики, интеллектуального анализа и искусственного интеллекта, связанных с развитием знаний и включением казахского языка в многоязычные проекты, показывает, что существующих в данном научном направлении решений не достаточно для удовлетворения на качественном уровне существующих потребностей по разработке систем автоматической обработки текстов казахского, русского и английского языков.

Существующие многоязычные приложения NLP, в своем большинстве, используют только грамматический этап обработки текста, тогда как, семантический анализ текста или анализ смысла естественного языка, по-прежнему остается одной из ключевых проблем, как теории искусственного интеллекта, так и компьютерной лингвистики.

Поскольку методы, базирующиеся на правилах, требуют высоких интеллектуальных затрата, а компьютерные процессоры обладают все большей мощью, все большее распространение приобретают методы машинного обучения. Однако, для использования методов машинного обучения при семантической и грамматической обработки многоязычной информации необходимы заранее созданные грамматически и семантически размеченные корпуса каждого естественного языка.

Все вышеперечисленное и обуславливает актуальность диссертационной работы, посвященной комплексному исследованию и решению проблемы анализа многоязычной текстовой информации на основе машинного обучения. Для решения данной проблемы поставлены и решены ряд задач от анализа исследований в области подходов машинного обучения, используемых при обработке текстовой информации, и существующих проблем формализации и моделирования казахского языка до разработки и реализации моделей, методов и алгоритмов морфологического и семантического анализа текстов казахского, русского и английского языков и извлечения их них знаний

Тема диссертационного исследования посвящена проблеме развития моделей, методов и алгоритмов обработки текстовой информации на основе машинного обучения и применения этих методов для автоматизации анализа многоязычной текстовой информации.

Основной целью работы является повышение качества работы систем автоматической обработки текстовой информации казахского, русского и английского языков за счет использования моделей интеллектуального анализа и методов машинного обучения. В рамках поставленной цели решается научная задача моделирования процессов интеллектуальной обработки многоязычных текстов, разработке моделей, методов, алгоритмов, осуществляющих анализа многоязычной текстовой информации, с целью определения основных характеристик текстов для построения моделей машинного обучения. Полученные алгоритмы должны получить свою практическую реализацию в виде экспериментальных систем обработки текстов, с возможностью оценки их работы на базе созданных размеченных корпусов.

Объектом исследования являются системы автоматической обработки текстовой информации на казахском, русском и английском языках, использующие методы машинного обучения.

Предметом исследования в данной диссертации являются модели и алгоритмы интеллектуального анализа многоязычной текстовой информации, базирующиеся на подходах машинного обучения.

Задачи исследования

Для достижения цели были поставлены следующие **задачи исследования**:

– анализ современного состояния проблемы автоматической обработки многоязычной текстовой информации, систематизация существующих лингвистических ресурсов и систем обработки казахского языка и формирование основных требования к разработке системы анализа многоязычной текстовой информации на основе машинного обучения;

– разработка моделей и алгоритмов извлечения фактов из слабоструктурированных и неструктурированных текстовых массивов казахского, русского и английского языков;

– разработка методов и алгоритмов морфологического и семантического анализа многоязычных текстов на базе моделей машинного обучения;

– разработка методов вероятностного POS-тегинга, использующих скрытую Марковскую модель;

– разработка метода, определения семантической близости многоязычных текстовых документов, базирующегося на использовании Vector Space Model;

– разработка программного продукта, позволяющего осуществлять семантическую обработку текстов, на базе грамматически и семантически размеченных корпусов, имплементирующего разработанные модели, методы и алгоритмы

Научная новизна полученных результатов:

– разработан гибридный метод автоматической морфологической и семантической разметки текстовых корпусов казахского языка, отличительной

особенностью которого является одновременное использование HMM и правил, представленных регулярными выражениями, что позволило снять часть морфологической многозначности и повысить полноты и точность разметки;

– разработаны логико-лингвистические модели семантического анализа, идентифицирующие факты в многоязыковых текстах, что позволило извлекать из текстов казахского, русского и английского языков знания, явным образом представленные в виде RDF-триплетов и формировать семантически размеченные обучающие корпуса;

– Разработан метод и алгоритм определения семантической близости многоязычных текстовых документов на базе VSM, который отличается использованием весовой функции PPMI для определения принадлежность текста к узкоспециализированной предметной области.

– Разработана информационная технология определения семантической близости текстов к заданной узкоспециализированной тематике, базирующаяся на предложенном методе машинного вычислении максимального, минимального и среднего значения косинусного сходства векторов документов обучающего корпуса;

Достоверность и обоснование научных положений, выводов и результатов диссертационной работы обосновывается практической реализацией программного продукта, представляющего динамически пополняемый узкоспециализированный выровненный параллельный многоязычный корпус.

Теоретическая значимость полученных результатов заключается в адаптации существующих и разработке новых моделей и методов обработки и анализа текстовой информации на казахском, русском и английском языках.

Практическая ценность полученных результатов заключается в разработке программного приложения на основе положений, вынесенных на защиту.

Положения выносимые на защиту:

- логико-лингвистическая модель идентификации фактов в многоязычных текстах;
- гибридный метод морфологической и семантической разметки текстовых корпусов казахского языка, базирующийся на скрытой Марковской модели;
- метод и алгоритм определения семантической близости текстов к заданной узкоспециализированной тематике, базирующийся на использовании весовой функции PPMI векторов документов;
- программная реализация разработанных моделей и алгоритмов для анализа многоязычной текстовой информации на основе машинного обучения
- методика экспертной оценки качества работы системы анализа семантической близости текстов.

Апробация работы. Основные результаты диссертации докладывались и обсуждались на международных и зарубежных научных конференциях, научных семинарах: Международная научно-практическая конференция «Математические методы и информационные технологии макроэкономического анализа и экономической политики, посвященная празднованию 80-летнего юбилея академика НАН РК Абдыкаппара Ашимовича Ашимова"(2017, Алматы);

International Conference on Business Information Systems, BIS (2018, Berlin, Germany); IV международная научно-практическая конференция "Информатика и прикладная математика", посвященной 70-летнему юбилею профессоров Биярова Т.Н., Вальдемара Вуйцика и 60-летию профессора Амиргалиева Е.Н. (Алматы, 2019); Federated Conference on Computer Science and Information Systems, FedCSIS (Poznan, Poland, 2018); III Междунар. науч. конф. «Информатика и прикладная математика», посв. 80-летнему юбилею проф. Бияшева Р.Г. и 70-летию проф. Айдарханова М.Б. (Алматы, 2018); In Proceedings of the 21st International Conference on Enterprise Information Systems, ICEIS (Heraklion, Crete – Greece, 2019); Computation linguistics and intelligent systems, COLLINS (Kharkiv Ukraine, 2019); 13th International Conference on Computational Semantics (IWCS 2019) – RELATIONS-Workshop on meaning relations between phrases and sentences (Gothenburg Sweden, 2019)

Диссертационная работа выполнялась в рамках проекта по грантовым исследованиям МОН РК “Методы и модели поиска и анализа криминально значимой информации в неструктурированных и слабоструктурированных текстовых массивах”– AP05131073 (2018-2020 гг.).

Результаты диссертационного исследования опубликованы в 17 работах. Из них 4 статьи в журналах, рекомендованных Комитетом по контролю в сфере образования и науки МОН РК, 7 статей в международных научных изданиях, входящих в базу данных Web of science и Scopus, 6 работы в материалах международных и республиканских конференций, 1 авторское свидетельство по результатам практического внедрения диссертационных исследований

Автор выражает свою благодарность научным руководителям: доктору PhD, ассоцировнносу професору Мамырбаеву О. Ж. за поставленные задачи и поддержку в работе, и профессору Хайровой Н. Ф. за плодотворные дискуссии и ценные замечания.

1. СОСТОЯНИЕ ВОПРОСА И ПОСТАНОВКА ЗАДАЧИ ИССЛЕДОВАНИЯ

1.1. Обзор существующих лингвистических ресурсов и систем автоматической обработки текстов казахского языка

В настоящее время научное направление автоматической обработки текстов, возникшее на стыковки таких традиционных направлений как искусственный интеллект и структурная и математическая лингвистика, достаточно хорошо развито. Существует множество приложений, решающих различные задачи обработки текстов, таких как машинный перевод, информационный поиск, диалоговые системы, системы автоматическое реферирование и т.д.

Однако, хотя существует достаточное количество компьютерные приложения, обрабатывающих казахский язык, в тоже время, остается достаточно сложностей и проблем, связанных с решением задач компьютерной лингвистики для казахского языка с необходимым уровнем качества обработки текста. Существование подобных проблем связано, как со сложностью морфологии, синтаксиса и семантики казахского языка, так и с исторически сложившимся поздним началом работ по исследованию и автоматизации казахского языка. Такой формальной стартовой точкой начала исследований в направлении структуризации и компьютеризации казахского языка можно считать появление в средине 70-х годов прошлого века работ казахского ученого и исследователя Бектаева К. Б. [3–5].

Основная часть современных научных исследований в направлении автоматической обработки казахского языка направлена на автоматизацию его морфологического и синтаксического анализа [6–9], что во многом связано с реально существующими проблемами автоматической обработки агглютинативных языков.

С необходимостью верификации существующих и разрабатываемых моделей и методов грамматического анализа казахского языка связано и существование достаточно большого количества лингвистических корпусов казахского языка, часть из которых морфологически аннотированы. Такими наиболее известными корпусами являются:

- Алматинский корпус казахского языка (АККЯ), содержащий более 40 миллионов словоупотреблений, 86% словоупотреблений в котором имеют грамматический разбор (Almaty Corpus of Kazakh) [10];
- Kazakh text corpora on Sketch Engine [11];
- Open-Source-Kazakh-Corpus, созданный с использованием инструмента Wikipedia dump[1] и включающий коллекцию из 20 миллионов слов (из них 600 тысяч уникальных) [12];
- Корпус Казахского языка или Kazakh Language Corpus (KLC) [13].

Последний корпус KLC, содержит более 135 миллионов слов, объединенных в пять стилистических жанров: литература, публицистика, наука, информация и

[1] https://dumps.wikimedia.org/kkwiki/

официальные бумаги. Кроме этих главных частей, KLC корпус включает: (i) аннотированный подкорпус письменных текстов, представляющий XML документы, размеченные морфологически, синтаксически и имеющие структурные характеристики текста; (ii) аннотированный подкорпус устной речи. Дополнительно, KLC имеет Веб-ориентированное приложение, позволяющее осуществлять навигацию и информационный поиск в корпусе.

В тоже время, наряду с монолингвистическими корпусами и электронными словарями, важнейшей базой разработки систем машинного перевода и других систем автоматическая обработка текстов на естественном языке (АОТЕЯ) для каждого языка являются параллельные корпуса. Автоматическое генерирование структурных правил языка и перевода, базирующееся на использование параллельных корпусов позволяет легко их внедрять в модели и методы МП и других приложений NLP, избегая больших временных интеллектуальных затрат.

К сожалению, к настоящему моменту нет достаточного хорошо разработанных параллельных корпусов большого объема, включающих казахский язык. Что, прежде всего, связанно с необходимостью затрат больших интеллектуальных и временных ресурсов для разработки лингвистических ресурсов подобного типа.

К таким параллельным корпусам, включающим казахский язык, и готовым к использованию, можно отнести OPUS проект [14], имеющий коллекцию из 4480 выровненных казахско-английских предложений, полученных с помощью инструментов Tatoeba и OpenSubtitle [15].

Хорошими лингвистическими ресурсами казахского языка являются параллельный Казахско-Английский корпус, включающий 5 925 предложений и Казахско-Русский корпус, размером 10 000 предложений созданный с использованием инструментов Bitextor и библиотек LibTidy и LibTagAligner [16]. Оба корпуса выровнены по предложениям.

Bitextor представляет собой открытое доступное приложение, предназначенное для формирования коллекции параллельных текстов Translation Memory (TM), собранных с мульти лингвистических веб-сайтов [17]. Процесс создания данных корпусов включал несколько шагов. На первом этапе, с помощью инструмента HTTract с мульти языкового сайта загружались все HTML файлы. На следующем этапе, все загруженные файлы проходили препроцессорную обработку, в результате которой все возможно некорректные HTML файлы преобразовывались в валидный XHTML формат, для чего использовалась библиотека LibTidy library. На следующем этапе обработки использовалась библиотека LibTextCat library для выравнивания файлов двух языков по имени и расширению. На последнем этапе, с помощью библиотеки LibTagAligner Bitextor сравнивал тексты в файлах, используя имеющиеся теги, блоки текстов и длину предложения.

Дополнительно в работе [18] для выравнивания предложений использовался hunalign tool инструментарий [19] и редактор выровненных текстов InterText [20], позволивший удалить от 6% до 8% , ошибок, оставшихся с предыдущих этапов. В данном исследовании созданы два дополнительных параллельных казахско-английских корпуса: Akorda, содержащий 21 148 выровненных предложений, и

TED, содержащий 6 120 выровненных предложений.

Среди работ, имеющих практическую направленность по разработке конкретных приложений, работающих с казахским языком, можно выделить направления исследований, связанные с системами автоматической проверки орфографии, системами машинного перевода, в частности, в русско↔казахском направлениях [21–22], а так же системами автоматического распознавания и синтеза казахской речи.

Первым программным продуктом, решающим проблему поиска орфографических ошибок и опечаток в текстах на казахском языке, набранных с помощью различных драйверов, стал пакет «Қазақ тілінің орфографиясы» [23]. На сегодня, программный пакет позволяет осуществлять проверку правильности орфографии текста, написанного на казахском языке, предоставляет электронные словари для перевода с казахского языка на русский и обратно и перевод текстов общей тематики с русского языка на казахский и обратно, программой «Тілмаш».

Некоторые ранее существующие онлайновые и пакетные версии систем проверки казахской орфографии, к настоящему моменту, добавили такой сервис как конвертации шрифта текста между кириллицей, латиницей и арабицей. Так, например, компания Sanasoft [24], являющаяся официальным вендором корпорации Microsoft в Центральной Азии по разработке программных продуктов для Microsoft, помимо проверки орфографии казахского языка в режиме реального времени, осуществляет поддержку казахского языка в различных кодировках в платежных системах Национального Банка Республики Казахстан и приложениях, в которых отсутствует поддержка казахского языка. При этом приложение использует русско-казахский словарь объемом более 200 000 слов и словосочетаний.

Данная компания Sanasoft так же занимается поддержкой он-лайн системы Google Translate, для казахского языка в направлении Kazakh<–>English и Kazakh<–> Russian. Решения, разработанные данной компанией для перевода с/на казахский язык, сегодня используются в продукции российской компании ПРОМТ [25]. В настоящее время при переводе на казахский язык система Google MT допускает ошибки, связанные с генерацией правильного падежа, а система Sanasoft MT генерирует ошибки неверного перевода слов.

Кроме того открытая платформа машинного перевода Apertium[2] позволяет так же работать с казахским языком. English-Kazakh MT система данной платформы включает три типа словарей: English и Kazakh – моно языковые словари и English и Kazakh билингвистические словарь. Все словари, за исключением словаря Казахского языка имеют XML формат, в котором каждое слова снабжено POS-тегом [26].

На сайте http://www.soylem.kz предлагается созданная на базе ядра перевода AUDARMA компанией LimeOn Global Company система SOYLEM, осуществляющая автоматизированный перевод документов с русского языка на казахский. В настоящий момент компания ведет работы по созданию облачного решения по автоматизации перевода документов SOYLEM CAT. Объем словарных

[2] https://www.apertium.org/index.rus.html?dir=eng-glg#translation

баз системы на сегодня: более 900 тыс. статей [27].

Среди предлагаемых приложений машинного перевода, работающих с казахским языком, можно назвать также выше упоминавшуюся систему «Тілмаш», переводящую тексты общей тематики с русского языка на казахский и обратно. Базовый словарь, используемый системой, содержит более 90 тысяч слов и словосочетаний [23].

В направлении разработки систем автоматического распознавания и синтеза казахской речи высокие показатели достигнуты на базе Института информационных и вычислительных технологий МОН РК (Алматы), НИИ «Искусственный интеллект» при ЕНУ им. Л. Н. Гумилева (Астана) [28] и Института информационных и вычислительных технологий КН МОН РК [29 –30]

1.2. Анализ существующих проблем формализации казахского языка.

Казахский язык относится к кыпчакской ветви тюркской группы алтайской языковой семьи. Наиболее близки к нему ногайский и каракалпакский языки. Всего на казахском языке говорят в мире около 12 млн. человек, из них в Казахстане - 8 млн. человек, 2 млн. в остальных странах СНГ, 1.5 млн. в Китае. Кроме того этот язык распространен в Монголии, Афганистане, Пакистане, Иране, Турции, Германии. С 1991 года казахский язык является государственным языком Республики Казахстан.

Анализируя Казахский язык с точки зрения возможности его формализации, моделирования и автоматической обработки, можно выделить несколько основных особенностей языка.

Прежде всего, в казахском языке четко определен строгий порядок слов: на первом месте стоит подлежащее, потом дополнение, завершает предложение сказуемое. Кроме того, четкий порядок требует располагать определение, которое может быть развернутым, перед определяемым, а обстоятельства места и времени обычно располагать перед подлежащим или перед прямым дополнением, но не после сказуемого.

Кроме того, так как Казахский язык является агглютинативным языком, обычно слово в казахском языке состоит из основы и целого ряда морфем, следующих одна за другой, каждая из которых имеет определенное значение. Например: “*отырғанмын*”, “қ*иындықтарға*”.

Особую роль для семантического и смыслового анализа текста, в казахском языке, как и в любом другом, играет сказуемое, называющее действие, описываемое во фразе или предложении. В Казахском языке сказуемое может быть выражено глаголом, существительным, прилагательным, причастием, деепричастием, такими вспомогательными словами как “*бар*"; “*жоқ*”; “*көп*”; “ *керек*” и другими. Однако, в подавляющем большинстве случаев, сказуемое в казахском языке выражается глаголом.

Глаголам казахского языка присущи два типа словообразования. Глагол может быть образован синтетически (путем аффиксации), образовывая производные глагольные основы; и аналитически, образуя составные и сложные глаголы.

Синтетическое словообразование глаголов. Неизменимой частью любого глагола, присущей всем его грамматическим формам, считается глагольной

основой. Глагольную основу представляет форма второго лица единственного числа будущего времени повелительного наклонения при обращении на «ты» (*сен*).

В казахском языке отсутствует префиксация, и существующие глаголообразующие аффиксы, которых более 80, всегда являются постпозитивными. Вместе с фонетическими вариантами число их составляет около двухсот, при этом залоговые аффиксы не учитываются [31]. Разработанная таблица основных словообразующих глагольных аффиксов позволяет их классифицировать по типу образования (Приложение А). В казахском языке действует закон гармонии гласных, который заключается в том что все гласные делятся на твердые (а, о, ұ, ы) и мягкие (ә, ө, ү, і, е). В одном слове могут употребляться либо только твердые, либо только мягкие гласные. Поэтому все суффиксы и окончания, как правило, имеют два варианта: для твердых и для мягких основ.

Аналитический тип словообразования. У глаголов, образованных синтаксическим способом словообразования, первый компонент обладает семантическим значением, а вспомогательный глагол, полностью утрачивает свое первоначальное лексическое значение и превращается в носителя грамматического формата. В современном казахском языке различают две группы глаголов, образовавшихся синтаксическим способом словообразования:

1) сложные глагольные основы, состоящие из *имени или звукоподражательного слова + вспомогательный глагол*;
2) сложные глагольные основы, состоящие из *глагола в деепричастной форме + вспомогательный глагол*

Существует группа вспомогательных глаголов, которые, сочетаясь с основными глаголами, образуют сложную глагольную форму, выражающую одно лексическое значение. Вспомогательные глаголы в сложной глагольной основе, модифицируя основное лексическое значение первого компонента, одновременно придают ей различные грамматические значения. Составные глагольные основы часто состоят из двух, трех и (реже) четырех глагольных основ, из которых первые выражают основное лексическое значение, а остальные придают ей различные грамматические значения [31].

Глагольный вид представляет собой семантическую характеристику основного значения глагола. Способ выражения глагольного вида в казахском языке аналитический. Глагольный вид выражается специальным вспомогательным глаголом, который сочетается с основным глаголом, стоящим в одной из деепричастных форм. *Совершенный вид* глагола образуется посредством сочетания основного глагола в форме деепричастия на *–п*, используемого вместе со специальными вспомогательными глаголами: *бол*, *бітір*, *біт*, *кет*, *қой*, *жібер*, *шық*, *сал*, *таста*, *қал*. *Несовершенный вид* глагола образуется сочетанием основного глагола в форме деепричастия на *–п* (*-ып, -іп*) со специальным функционально-вспомогательными глаголами *: отыр, жат, тұр, жат, жүр, бер* (с деепричастием на*–а [-е, -й]*).

Кроме того, во многих случаях при аналитическом словообразовании вспомогательные глаголы конкретизируют тип действия. В таблице (Приложение

Б) представлены вспомогательные глаголы казахского языка, сообщающие подробную характеристику действия.

Еще одной грамматической категорией словообразования глагола казахского языка, распознаваемой по семантике, морфологическому оформлению и синтаксическим функциям является залог.

Залог формируется прибавлением к глагольным основам положительной формы аффиксов, передающих отношения между действием и субъектом и/или объектом. В казахском языке выделяют пять залогов, разделяемых по грамматическому оформлению, семантическому значению и синтаксическим функциям: возвратный, страдательный, совместно-взаимный, понудительный и основной или исходный. При этом, первые четыре оформляются аффиксами. При составных глаголах аффиксы залогов обычно прибавляются к основному глаголу, а не к вспомогательному. Исключения составляют аффиксы *–с, -ыс, -іс*. Например, *күліп жіберісті, қарай қалысады*. Возможно наращивание залоговых аффиксов: к одной основе могут прибавляться два или более залоговых аффикса возвратного и понудительного, понудительного и страдательного, совместно-взаимного и понудительного залогов. В таблицы приведены основные словообразовательные залоговые суффиксы глаголов казахского языка и правила их возможного наращивания (Приложение В).

Все остальные грамматические значения (наклонения, времени, лица и т.д.) вносятся соответствующими морфологическими формантами. Например, спряжение глаголов в казахском языке, так и в других тюркских языках, оформляется при помощи аффиксов сказуемости. То есть любая часть речи, выступающая в предложении предикатом, может принимать личные окончания, определяющие субъект действия. На рисунке 1.1 показана структура формирования глагола в личной форме.

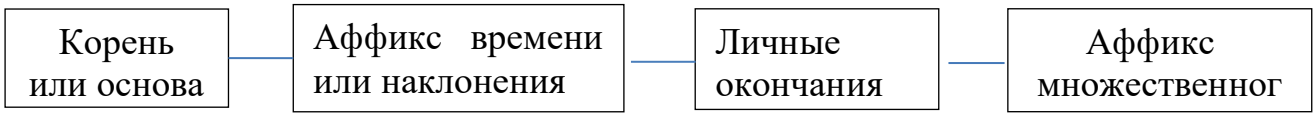


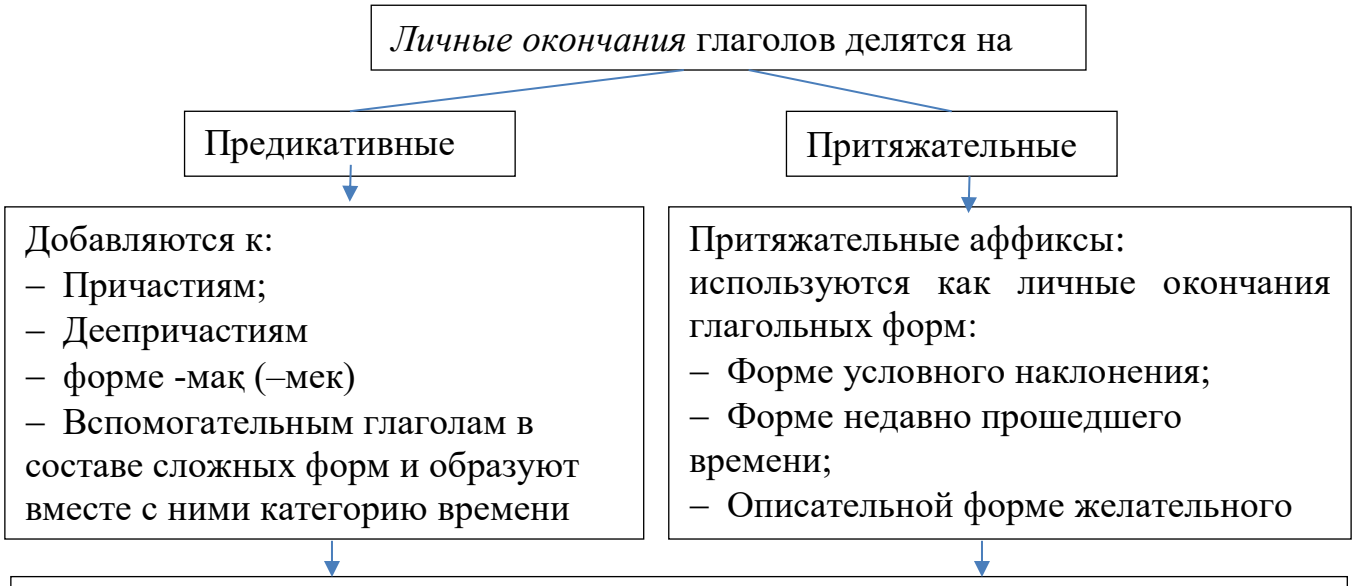
Рисунок 1.1 – Схема формирования личной формы глагола.

При этом личные окончания глаголов делятся на предикативные и притяжательные окончания. На рисунке 1.2 показана схема формирования предикативных и притяжательных окончаний глаголов.

Наклонение глаголов казахского языка включает в себя такие грамматические категории, как время, лицо и число. В современном казахском языке различают пять наклонений: повелительное, изъявительное, желательное, условное и отглагольно-именное. Таблица показывает формализацию грамматического оформления наклонений казахского языка (Приложение_Г).

Изъявительное наклонение является наиболее употребительным наклонением. В нем представлены грамматические формы глагола, выражающие временные

отношения. При этом, категория времени формально выражается морфологически и синтаксически.



– *Предикативные личные окончания*

(не добавляются к форме недавно прошедшего (окончательного) времени; формам условного наклонения, которые принимают притяжательные аффиксы 1 и 2 лица ед и мн.числа)

-сың , *-сыз* (вежливая форма) *-сің*, *-сіз*	2 лицо, ед. число	ала*сың* (ты берешь), ала*сыз* (вы берете) айт-а-*сың* айт-а-*сыз* бара- сың келе-*сің*
-сыңдар , *-сыздар* (вежливая форма), *сіндер*, *сіздер*	2 лицо, мн. число	айт-а-*сың-дар* айт-а-*сыз-дар*
-мын, *-мін*, *-пын*, *-пін*, *-бын*, *-бін*	1 лицо, ед. чис	айт-а-*мын* бара-мын келе-мін
-мыз (міз, -пыз, -піз, -быз, біз)	1 лицо, мн. чис	айт-а-*мыз*, *бара-мыз*
-ды	3 лицо, ед. чис	айт-а-*ды*
-ды	3 лицо, мн. чис	айт-а-*ды*
– *Притяжательные личные окончания* (не связаны с категорией времени, они выражают лишь лицо субъекта)		
-ң -ңыз –ңіз *-сың, -сің*	2 лицо, ед. число	сент қайт-са-ң, сіз қайт-са-ң-ыз, ал-а-сың (берешь), бер-е-*сің* (даешь)

– *Притяжательные личные окончания* (продолжение)		
-дар, -сыңдар, -сіңдер, -сыздар, -сіздер	2 лицо, мн. число	сендер қайт-са-ң-дар, сіздер қайт-са-ң-ыз-дар жина-са-ң-дар (если соберете)
-м	1 лицо, ед. чис	мен қайт-са-м, сана-са-м (когда я считаю), терле-се-м (когда я потею)
-қ, -к	1 лицо, мн. чис	біз қайт-са-қ, жазса-қ (если будем писать)
	3 лицо, ед. чис	ол қайт-са, барса, келсе
	3 лицо, мн. чис	олар қайт-са, таб-а-ды (найдут)

Рисунок 1.2 – Структурная схема формирования предикативных и притяжательных окончаний глаголов.

Морфологически категория времени формируется прибавлением к причастным и деепричастным формам личных окончаний, тогда как синтаксически форма времени оформляется сочетанием глагольных имен с соответствующими вспомогательными глаголами. Существует форма времени, в образовании которой причастия и деепричастия не участвуют. Она образуется от основ *ды (ді)* и личного окончания (сана-*ды- м* (я считал), сен сана-*ды-ң* **(**ты считал), ол сана-*ды* **(**он считал)).

Специфическая форма глагола “тұйық етістік” (неопределенное наклонение) не является инфинитивом, а выступает как имя или название действия. Она образуется путем присоединения к основе глагола аффикса — *у*. Например, *тапсыр-у*, *шақыр-у*. Неопределенная форма глагола в лексическом отношении ближе к именам существительным — она не спрягается, а склоняется, принимая аффиксы принадлежности: *у-дың, -у-ды, -у-ға, -у-дан, -у-да, -у-мен, -у, -у-ім , -у-ің, -у-і, -у-і-міз, -у-і-ңіз, -у-дің, -у-ге, -у-ді, -у-де -у-ден (уы, уым, уымыз, у-ың, у-ыңыз)*. Глаголы неопределенной формы образуют изафетную конструкцию, требуя впереди себя личное местоимение или существительное в родительном падеже.

Многие глаголы в неопределенной форме переходят в отглагольные существительные (*жаз-у* (*письмо*), *ойла-у* (*мышление*)). Далее по словообразовательной цепочке от отглагольных имен при помощи аффикса *–шы* часто образуются производные имена существительные (*жаз-у-шы* (*писатель*), *сайла-у-шы* (*избиратель*)). Приложение_Д показывает таблицу основных словообразовательных афиксов имен существительных.

Имена существительные могут выступать в предложении в функции

определения, подлежащего, дополнения, обстоятельства и именного сказуемого. Существительное в казахском языке не имеет категорий рода и одушевленности, зато имеется категория принадлежности. В синтаксической связи двух имен, выражаемой примыканием, первый компонент всегда выступает как определение , второй — как определяемое. Имя существительное, выступающее в качестве определения (стоит первым) может быть оформлено аффиксом родительного падежа или аффиксом принадлежности. Имена существительные изменяются по числам, падежам, лица, а так же принимают аффиксы принадлежности.

В казахском языке, по сравнению с русским, границы между частями речи несколько размыты. Имена существительные могут выступать в предложении в функции определения, подлежащего, дополнения, обстоятельства и именного сказуемого. В казахском языке существует два типа склонения: простое (безотносительно к обладателю) и притяжательное склонение (с указателем обладателя). Рисунок 1.3 показывает сруктурную схему простого склонения существительных.

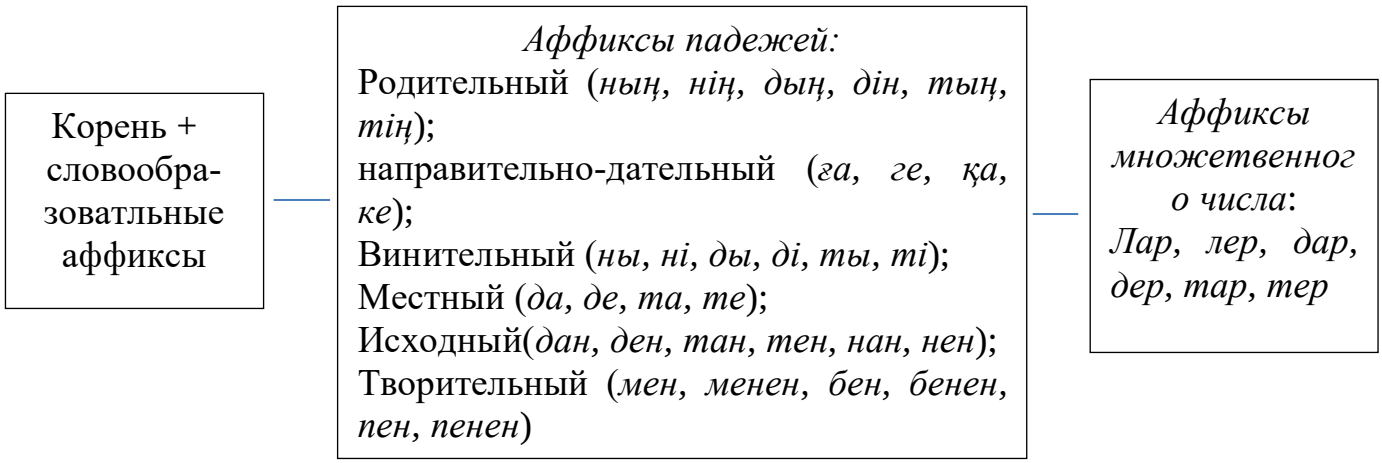


Рисунок 1.3 – Сруктурная схема простого склонения существительных.

При притяжательном склонении в словах содержится указание на обладателя. Принадлежность предмета кому/чему–нибудь выражается посредством присоединения к основе слова аффиксов принадлежности. Форма принадлежности указывает одновременно и на предмет обладания и на имя обладателя. В таблице 1.1 показаны особенности аффиксного формирования притяжательности в казахском языке.

Таблица 1.1. Аффиксное формирование притяжательности казахского языка

Имя обладателя стоит в единственном числе		
лицо	аффиксы	Примеры
1	*м, ым, ім,*	ана-*м*, калам-*ым*, дәптер-*ім*
2	*ң, ың, ің*	ана-*ң*, қалам-*ың*, дәптер-*ің*
3	*сы, ы, і, сі*	ана-*сы*, қалам-*ы*, дәптер-*і*

Имя обладателя и предмет обладания стоят во множественном числе		
1	*ымыз, іміз*	қаламдар-*ымыз*, дәптерлер-*іміз*
2	*ыңыз, ныз, іңіз, ңіз*	қаламдар-*ыңыз*, дәптерлер-*іңіз*
3	*ы, і*	қаламдар-*ы*, дәптерлер-*і*

В таблице 1.2. показаны падежные аффиксы единственного и множественного числа притяжательного склонения казахского языка.

Таблица 1.2Падежные аффиксы единственного
и множественного числа притяжательного склонения

Падежи	1 лицо	2 лицо	3 лицо
родительный	*-ның, нің*	*-ның, нің*	*-ның, нің*
Напр.-дательный	*-а, е*	*-а, е*	*-на, не*
Винительный	*-ды, ді*	*-ды, ді*	*-н*
Местный	*-да,де*	*-да, де*	*-нда, нде*
Исходный	*-нан, нен*	*-нан, нен*	*-нан, нен*
творительный	*-мен, менен*	*-мен, менен*	*-мен, менен*

1.3. Анализ методов машинного обучения, используемых при обработке текстовой информации

Основными направлениями сегодняшней реализации методов и моделей NLP (Natural Language Processing) являются как традиционные Machine Translation, Text classification, Information Retrieval, Sentiment Analysis, Information Extraction, так и достаточно новые Content-based recommender system, Speech recognition, Conversation chat-bot systems и другие [32].

Для реализации этих и других приложений существует три основных группы методов NLP: (1) подходы, базирующиеся на правилах; (2) методы традиционного машинного обучения; (3) глубокое машинное обучение.

Примером первого подхода может быть использование регулярных выражений или контекстно-свободных грамматик для описания морфологии или синтаксиса языка, соответственно. Кроме того, использование ручных правил может быть использовано в задачах классификации текстов для выделения спама. При этом, обычно, реализации, основанные на правилах, требуют больших интеллектуальных затрат при разработке подобных правил, имеют высокую точность и низкую полноту.

Для использования второй группы методов, прежде всего, необходимо иметь некоторый обучающий корпус (training corpus), т.е. корпус текстов, в котором уже осуществлена некая маркировка или тегирование. Основным этапом разработки метода машинного обучения является этап конструирования признаков, на основе имеющегося размеченного корпуса и последующее использования вероятностного подхода, с максимилизацией вероятности значения признаков, для обучения построенной модели, с последующим применением модели к тестовому корпусу

(test corpus).

Сегодня методы машинного обучения присутствуют практически во всех современных задачах обработки текстов, и в целом сводясь к числовой оптимизации весов для интеллектуально выбранных параметров и функций [33]. Такой подход используется, например, в задаче классификации текстов. В этом случае целью обучения является получение необходимых и достаточных правил, с помощью которых можно произвести классификацию новых объектов (текстов, слов, фраз), сходных с теми, которые составляли обучающую выборку. Задачей машинного обучения будет разработка классификатора, который представляет собой γ функцию, определяющую для каждого нового текста (документа или фрагмента текста) его класс. На рисунке 1.4 показана структурная схема такого классификатора с учителем (supervised machine learning).

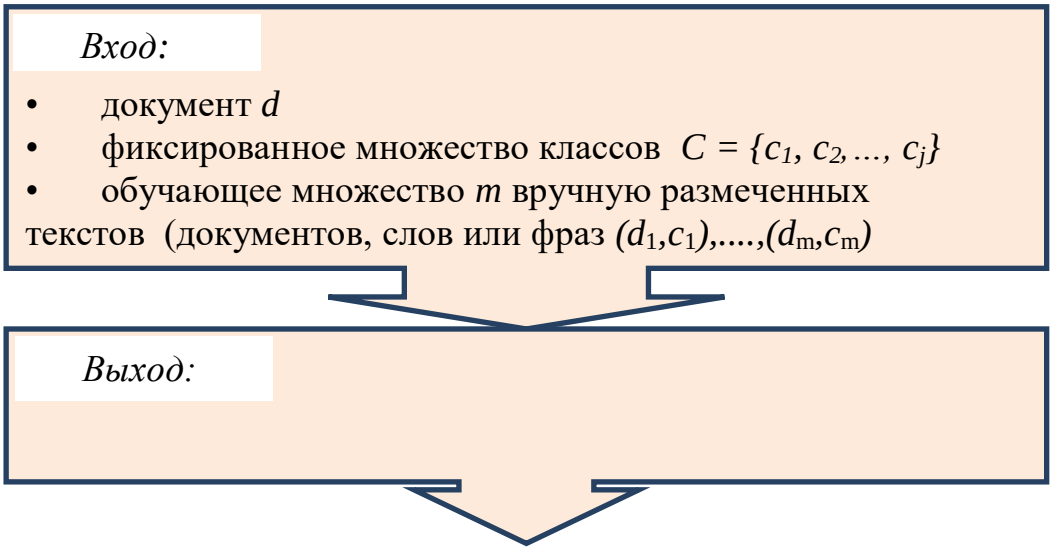


Рисунок 1.4 — Структурная схема supervised machine learning классификатора

Модели классификатора, базирующегося на машинном обучении, используются, например, для построения предметной онтологии на основе текста [34] и при использовании латентного семантического анализа (LSA) для статистической оценке сходства слов по их значению [35]. В последние годы методы машинного обучения успешно применяются в новых областях. Например, для повышения эффективности вычисления коэффициента читабельности текста (readability) [36, 37] и для решения проблем информационного поиска (Information Retrieval (IR)) [38].

Методы машинного обучения можно разделить на вероятностные и логические в зависимости от природы объектов, признаков и формы представления функции γ. К вероятностным методам машинного обучения относятся: наивный байесовский классификатор [39, 40]; логическая регрессия; метод опорных векторов (Support Vector Machines (SVM)); метод k-ближайших соседей [41].

Метод Байеса определяет вероятность наступления события в условиях, когда наблюдается лишь некоторая частичная информация о событиях, например, по наблюдаемым признакам, в частности в модели *bag of words*, по вероятности

появления слов, определяется принадлежность некоторого объекта (текста документа) к одному из заданных классов [42, 43].

Метод опорных векторов (Support Vector Machine) относится к линейным разделяющим методам, в которых каждый вектор признаков представляется точкой в многомерном пространстве признаков. При заданной классификации объектов строятся две параллельные гиперплоскости, разделяющие объекты разных классов, таким образом, что расстояние между этими гиперплоскостями максимизируется [44, 45].

К вероятностным методам относится так же скрытая Марковская модель (Hidden Markov chains). Это статическая модель, имитирующая некоторый последовательный процесс, в котором на наблюдаемые переменные оказывают влияния скрытые состояния. Переход из одного состояния в другое происходит с некоторой вероятностью. По последовательности наблюдений можно получить информацию о последовательности состояний [46]. Например, скрытые Марковские цепочки используются для построения системы синтеза речи [47].

Метод k-ближайший соседей классифицирует объект по большинству «голосов» его соседних объектов в многомерном пространстве признаков. Объект относится к классу, к которому принадлежит наибольшее число из его k наиболее близких соседей. Метод KNN (K Nearest Neighbours) широко применяется для классификации текстов. При этом размер пространства признаков определяется количеством терминов в корпусе [48]. Тексты представляются в виде векторов, компоненты которых показывают «важность» соответствующего слова для данного текста через значение TF-IDF. Одним из основных недостатков данного подхода является медленная скорость, так как для классификации каждого нового текста нужно вычислять расстояния между этим текстом и всеми остальными текстами в корпусе, а их количество может быть достаточно большим [49].

В последнее время нейронные сети так же получают широкое распространение в качестве инструментария классификации и некоторых других задач NLP. Нейронная сеть состоит из нескольких слоев нейронов, связанных между собой. Веса связей при обучении изменяются таким образом, чтобы выходной элемент сети давал бы правильный классификационный ответ на входной сигнал.

Кроме традиционных рекуррентных сетей, использующихся для классификации текстов [50], некоторые топологии нейросетей обладают специфическими свойствами, к примеру, для предсказывания ввода последующего слова (предикативный ввод). Например, при заполнении поисковых форм используются рекуррентные нейросети с несколькими сверточными слоями, размещенными ближе к входным нейронам [51]. Facebook использует нейронные сети для алгоритмов автоматического проставления тегов, Google — для поиска среди фотографий пользователя, Amazon — для генерации рекомендаций товаров, Pinterest — для персонализации домашней страницы пользователя, а Instagram — для поисковой инфраструктуры.

К логическим методам относятся методы, при которых в процессе обучения строятся логические правила в форме продукций или в форме деревьев решений, в

узлах которых проверяются значения отобранных при обучении признаков и принимается решение о разбиении объектов на подклассы. Логические методы включают вывод импликативных и функциональных зависимостей из данных, вывод логических правил, вывод ассоциативных правил, конструирование решающих деревьев, извлечение иерархических концептуальных классификаций и ряд других [52].

К методам концептуального обучения относится направление машинного обучения, называемое формальным концептуальным анализом, который часто называется концептуальной кластеризацией [53]. Примером концептуального анализа может служить моделирование лексической базы данных [54].

Ключевой проблемой использования методов машинного обучения в задачах и приложениях NLP является необходимость наличия для эффективной модели больших и сверхбольших ресурсов размеченных корпусов соответствующего языка [52].

Проблема необходимости наличия сверхбольших размеченных корпусов остается и при использовании последней третьей группы методов NLP — глубокое машинное обучение (deep machine learning). В отличие от традиционных методов machine learning, которые представляют собой числовую оптимизацию выбранных человеком характеристик, целью глубокого машинного обучения является определение того, как можно использовать данные для разработки характеристик и представлений, подходящих для интерпретации задачи [33]. Наиболее привлекательным качеством этих методов является возможность их использования без каких-либо затратно-временных вручную-разработанных характеристик и функций.

Наиболее востребованными методами глубинного обучения являются рекуррентные нейронные сети (Recurrent Neural Networks) и свёрточные нейронные сети (Convolutional Neural Networks) [55]. Например, на вход сверточной нейронной сети подается предложение, в котором каждое слово уже представлено вектором (вектор векторов). Как правило, для представления слов векторами используются заранее обученные модели word2vec [56]. Сверточная нейронная сеть состоит из двух слоев: «глубинного» слоя свертки и обычного скрытого слоя. Слой свертки, в свою очередь, состоит из фильтров и слоя «субдискретизации». Фильтр представляет собой нейрон, вход которого формируется при помощи окон, передвигающихся по тексту и выбирающих последовательно некоторое количество слов (например, окно длины «три» выберет первые три слова, слова со второго по четвертое, с третьего по пятое и т. д.). На выходе фильтра формируется один вектор, агрегирующий все вектора слов, в него входящих. Затем на слое субдискретизации формируется один вектор, соответствующий всему предложению, который вычисляется как покомпонентный максимум из всех выходных векторов фильтров [57].

В то же время, в своем недавнем исследовании [58] Крис Маннинг, глава группы компьютерной лингвистики в Стэнфорде и президент ACL (Association for Computational Linguistics), показал, что применение методов глубокого машинного обучения в NLP эффективно только для классификации текстов, классификации

последовательностей и снижения размерности.

Аналогично многим направлениям NLP автоматическое аннотирование корпусов может быть осуществлено с помощью методов, основанных на правилах; с помощью методов машинного обучения и гибридными методами.

Первые корпуса практически для каждого нового естественного языка используют автоматическое аннотирование, основанное на правилах [59]. Так, система “Taggit”, используемая для аннотирования Bown Corpus (первого аннотированного корпуса) в начале 70-х годов, ставила каждому слову корпуса в соответствие множество потенциальных тегов. Для выбора подходящего тега из данного множества использовался контекст (конечный список слов, большие буквы, слова, написанные через дефис и т.д.), который определялся определенными правилами. На основании лингвистических знаний каждое конкретное правило определяло, что данный потенциальный тег может быть корректным в данном контексте или, наоборот, его использование некорректно [60]. Такие правила в некоторых случаях позволяли снимать многозначность на множестве возможных тегов.

Вторая группа методов использует машинное обучение и предварительно размеченные корпуса данного языка. Методы машинного обучения позволяют оценивать наиболее вероятные последовательности тегов, используя частоты слов и тегов в ранее разработанных и размеченных корпусах данного языка (training corpus) [61, 62].

Гибридные методы автоматического тегирования одновременно используют и вероятностные методы, и правила. Например, современная система Constituent Likelihood Automatic Word-tagging System (CLAWS), которая разрабатывается исследовательским центом Университета Ланкастера (Lancaster University Center for Computer Corpus Research on Language) начиная с 1980х годов, комбинирует вероятностные методы и методы, базирующиеся на правилах [63].

Алгоритмы первой версии системы использовали униграммную вероятность: наиболее подходящий тег из множества возможных тегов для каждого слова выбирался без учета его связи с контекстом. Корпус, размеченный способом униграммной вероятности, имеет точность разметки около 80%.

На следующем этапе работы алгоритма разметка улучшалась итеративно, внесением исправлений, с помощью контекстного правила:

изменить tag a на tag b,

если*:*

a) *предыдущее (следующее) слово имеет tag z;*
b) *второе слово перед (после) имеет tag z;*
c) *одно из двух предыдущих (следующих) слов имеет tag z;*
d) *одно из трех предыдущих (следующих) слов имеет tag z* [64].

В отличие от традиционно используемого подхода скрытых Марковских цепей, вероятностные ассоциации «слово – метка» в этой модели опирались на оценку частоты применения тега для многозначного слова экспертом (это связано с тем, что используемый обучающий корпус был не достаточно велик). Для вероятностного оценивания эксперты использовали трех точечную шкалу:

“Общее”, “Редкое”, “Очень редкое”. Метка “Общее” применялась, если вероятность разметки слова данным тегом более 90%; “Редкое” — ассоциация данного слова с данным тегом используется менее чем в 10% случаев (обозначается знаком “@”); “Очень редкое” — вероятность того, что слово может быть размечено данным тегом меньше чем 1 % (обозначается знаком “%”). На рисунке 1.5 показан пример POS-tagging предложения «*The naïve cat sat on the Persian mat*» системой CLAWS

В последних версиях системы при определении условной вероятности последовательности тегов чаще используют не биграммные, а трехграммные цепочки, т.е. условная вероятность считается с учетом двух предыдущих тегов. При определении условной вероятности ассоциации слова с тегом используется не экспертное мнение, а теорема Байеса. В то же время, так как для редких и очень редких тегов информация о вероятности частоты появления может быть не достаточно корректна, то, обозначения таких тегов, как “@” и “%” осталось и может быть откорректирована экспертом вручную [65].

<s>
The naïve cat sat on the Persian mat
</s>

```
**6;0;START    NULL
----------------------------
The            AT0
na&iuml;ve     AJ0
cat            NN1
sat            [VVD/91] VVN@/9
on             [PRP/90] AVP@/10
the            AT0
**18;7;hi      NULL
Persian        AJ0
</hi>          NULL
mat            [NN1/99] AJ0@/1 VVB@/0
.              .
**8;26;text    NULL
```

Рисунок 1.5 — Пример POS-tagging предложения «*The naïve cat sat on the Persian mat*» системой CLAWS

Кроме того, список контекстных правил значительно расширился, по сравнению с начальными версиями системы. Например, добавились следующие контекстные правила:

– слово, длинной более 25 символов, рассматривается как общее существительное;

– осуществляется поиск наиболее длинной концовки слова, для которой существует предопределенное правило. Например, (1) “*-man*” для слова “*Bowman*”; (2) слова с суффиксом “*-er*”, которые вероятностными методами были размечены как *verb* — маркируются как *common noun (listener)*, а слова, которые были определены как *adjective*, маркируются как *positive adjective (odder*— страннее*)*;

– определяются шаблоны цепочек, состоящие из двух или более слов, которые функционируют как единая грамматическая единица. Например, “*according to*”;

– определяются шаблоны регулярных выражений: (1) по словам (например, любое слово, заканчивающееся на *–ing* или любое слово, начинающееся с прописной буквы, но не в начале предложения); (2) part-of-speech tag (например, любой тег, начинающийся на *N*, определяющий какую-то форму существительного); (3) отрицания; (4) необязательного элемента (например, необязательный adverb в глагольной группе или слова not или n't); (5) обозначение количества повторений необязательного элемента (например, три слова могут повторяться между двумя частями шаблона);

– если больше одного шаблона оказываются подходящими для строки, то определяется приоритет шаблона по типам шаблона (например, совпадение слов – это шаблон более высокого уровня приоритета, чем совпадения по тегу) или длине совпадения (более длинный шаблон имеет приоритет выше, чем короткий). Например, строке подходят три вида шаблонов: "*as well as*", "*as <Adjective> as*" и "*as well*", то шаблон "*as well as*" обладает более высоким приоритетом, чем "*as <Adjective> as*" (совпадение слов более высокий приоритет, чем совпадение по тегу), а шаблон "*as well as*" имеет приоритет выше, чем "*as well*", так как он первый длиннее второго.

Последняя версия системы CLAWS использовалась для POS-tagging 100 миллионного Британского национального корпуса (British National Corpus (BNC)), корпусов Brigham Young University и многих других современных корпусов английского языка [64, 66]. Точность разметки системы составляет 96-97%; время разметки в зависимости от типа текста и процессора от 7 до 12 минут для разметки 1 миллиона слов, в зависимости от типа процессора.

При этом исследования показывают, что использование гибридных методов, объединяющих правила и вероятностные вычисления, предоставляют результаты сравнимые с использованием методов машинного обучения [67]. В тоже время, статистические методы, достигли предела своей точности и не могут ее повысить, так как некоторые лингвистические знания, трудно описать количественными вероятностными параметрами.

1.4. Обзор существующих подходов Open IE многоязычной текстовой информации.

Проблема автоматического извлечения информации и фактов из текстов остается на сегодня не решенной. Существующие модели и алгоритмы извлечения фактографической информации зависят от степени структурированности конкретного анализируемого документа [68]. Подобно общей классификации степени формализации информации, по степени структурированности мы можем разделить текстовые документы на: (1) табличные данные, отображенные в виде фактов; (2) массивы однородных слабоструктурированных текстовых документов, обычно описывающие конкретную предметную область, и (3) документы произвольного слабоструктурированного типа [69].

Для извлечения фактов, представленных в хорошо структурированных текстовых документах, существуют достаточно надежные алгоритмы [70 – 73]. В то же время, несмотря на постоянный рост интереса к исследованиям,

фокусирующимся на способах выявления и извлечения фактов из текстовых корпусов и веб-контента, в настоящее время, не существует общего и хорошо обгрунтованного подхода для излечения структурированной информации из неструктурированных разнородных текстов [74–76]. Рост интереса к данному направлению исследований связан, прежде всего, с существующими на сегодня огромными объемами текстовой информации в корпоративных и интернет сетях, представленной в неструктурированном и полу-структурированном виде (по некоторым источникам, подобных текстов более 85%). Извлечение фактов из неструктурированных и полу-структурированных текстов, может стать дополнительным мощным источником для решения таких задач как автоматическая генерация онтологий на базе корпуса текстов, разработка интеллектуальных вопросно-ответных систем, бизнес-аналитика (например, прогнозирование ожидаемого изменения фондового рынка) и др.

Задачу направления искусственного интеллекта, фокусирующуюся на извлечении информации из слабоструктурированных и неструктурированных текстов, по способу ее решения можно разделить, на три подхода: Information Extraction (IE), Open Information Extraction (OIE) and Fact Extraction (FE). Технологии всех трех направлений позволяют автоматически просматривать большие объемы текстов, содержащих относительно небольшое количество фактической информации.

Работу IE можно рассматривать как особый вид поиска информации, подобный Information Retrieval (IR), когда запрос сформулирован заранее. В то же время, в отличие от информационного поиска, IE создает из набора обработанных документов структуру данных, описывающих соответствующий факт, тогда как результатом информационного поиска является набор ссылок на документы, соответствующие запросу.

Большинство первых системы IE были доменно-ориентированными и базировались на использовании предварительно интеллектуально сформированных знаниях. Примером такого подхода является одна из первых систем IE, работающая с текстами о латиноамериканском терроризме, которая использовала заранее разработанные морфосемантические шаблоны, позволяющие получить тот или иной факт [77, 78]. Современные системы IE так же используют предопределенный набор правил, позволяющих идентифицировать информацию, извлекаемую из текста [79]. Большинство систем IE извлекают и представляют информацию в форме кортежей из двух объектов, с заранее предопределенным типом отношений между ними [80]. Таким образом, подходы IE направлены на создание заранее известных структур знаний в качестве результата.

Как правило, системы Information Extraction включают несколько задач: (1) идентификацию сущностей (Name Entity Recognition (NER)); (2) задачу снятия кореферентности (coreference resolution) — поиск всех лингвистических выражений, которые ссылаются на один и тот же объект внеязыковой действительности; (3) распознавание семантических ролей партиципантов в предложении; (4) определение отношений между сущностями [81].

Текущие исследования в области интеллектуального анализа текста, в

основном, базируются на статистических моделях, использующих наивный метод Байеса [82–85], метод опорных векторов (Support Vector Machines (SVM)) [86], логическую регрессию [87] и метод k-ближайших соседей [88, 89]. Обычно эти методы эффективно используются в системах IR и автоматической классификации/кластеризации текстов [76, 86, 90].

В свою очередь, технологии IE обычно используют статистические методы и методы машинного обучения, как с учителем, так и без учителя [91]. Дополнительно используется распознавание специфических доменных объектов (лица, названия компаний и т.д.), синтаксический анализ и семантическое тегирование [92, 93].

Возрастающий в последнее время интерес к исследованиям, связанным с выявлением и извлечением фактов из неструктурированных текстов связан во многом с расширением областей их использования. Сегодня структурированная информация и идентифицированные факты, кроме интеллектуального использования, могут применяться для создания баз знаний в формате Resource Description Framework (RDF) или онтологий. Кроме того, повышение интереса к автоматическому извлечению информации из текстов, связано так же с недавним предложением использовать статистическую меру, называемую фактической плотностью, для оценки качества контента и определения информативности документа Интернета [94, 95].

В то же время подход IE, направленный на создание заранее известных структур знаний в качестве результата, не позволяет работать с корпусами и веб-контентом текстов неограниченных знаний, где целевые отношения не могут быть предопределены [79].

Новая, появившаяся в 2007 году парадигма извлечения знаний — Open IE [78], позволяет идентифицировать неограниченное число отношений и, следовательно, не зависеть от доменной ориентации. OIE включает в себя широкий спектр задач: (1) определение и отслеживание сущностей, (2) определение их отношений и характеристик, (3) обнаружение и характеристику событий [96]. Подход Open IE, в общем случае, ориентирован на следующие этапы:

1) Entity Extraction – извлечение слов или словосочетаний, важных для описания смысла текста (списки терминов предметной области, персоналий, организаций, географических названий и Т.Д.);
2) Feature Association Extraction – исследование связей между извлеченными понятиями;
3) Event and Fact Extraction – излечение событий, распознавание фактов и действий.

В отличие от подходов IE, Open IE извлекает факт, представленный в виде триплета *Субъект – Отношение – Объект* (рис.1.6), соответствующий RDF-графу

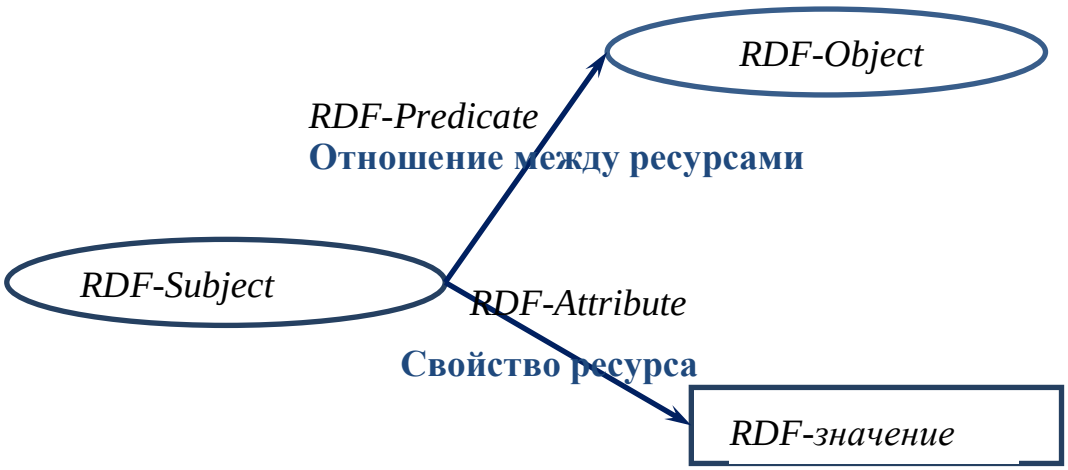


Рисунок 1.6 – Схематическое RDF представление триплета факта.

В таком триплете факта *Объект* и *Субъект* действия, представляющие участников действия, обычно выражаются существительными или именными группами естественного языка, тогда как *Предикат*, определяющий семантическое отношение между участниками действия, обычно выражается на естественном языке сказуемым (чаще всего глаголом) [97]. Кроме того, факт может содержать несколько атрибутов, таких как время, местоположение, способ действия и принадлежность (или обладание) и другие, которые также могут быть представлены именными фразами.

Исходя из данного подхода, выделяем четыре семантических типа факта, извлекаемого из текста, каждый из которых выражается различными поверхностными структурами естественных языков [98] (рис. 1.7).

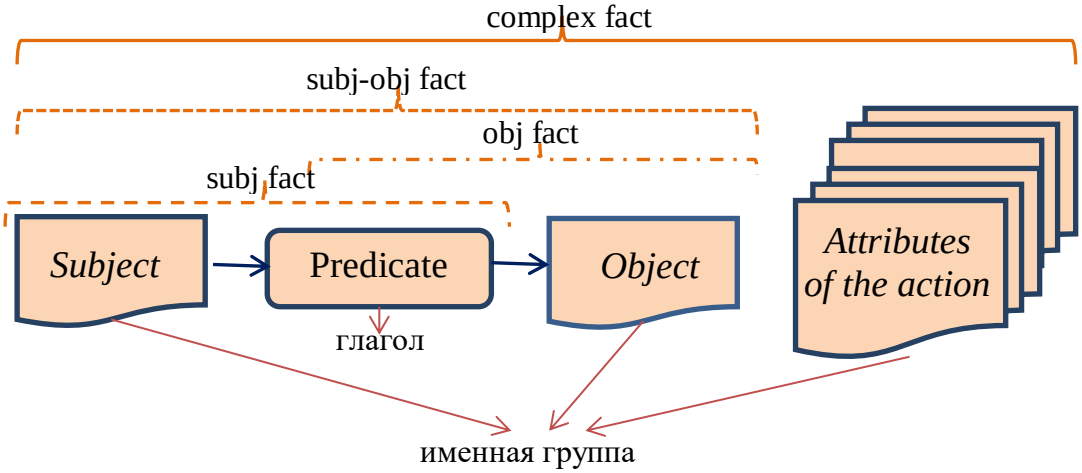


Рисунок 1.7 – Структурная схема формализации
четырех семантических типов фактов.

Первый семантический тип факта, который мы определяем как *subj-fact*, выражается наименьшим грамматическим предложением, включающим только именную группу, определяющую *Subject*, как инициатора действия, и сказуемое или *Predicate*, определяющее само действие.

Второй тип факта, определяемый нами как *obj-fact*, определяется так же в наименьшем грамматическом предложении естественного языка, которое включает сказуемое, называющее *Predicate*, и именную группу, называющую участника действия, на которого это действие направлено, т.е. (*Object*) Объект действия.

Третий семантический тип факта, называемый *subj-obj* факт, выражаемый предложением, включающим две именных группы, которые называют как *Субъект* действия, так и *Объект* действия, и глагол, называющий *Предикат*.

Последний четвертый семантический тип факта — *complex fact*, извлекается из предложения, включающего более двух именных групп, *Субъект*, *Объект*, *Предикат* действия, а так же некоторые атрибуты действия (время, место, способ действия и т.д.)

Для определения всех четырех типов фактов большинство разработок Open IE использует такие инструменты NLP как POS-tagging и Dependency parsing [99, 100], дополнительно применяя лексические ограничения [101] или семантические аннотации [102] для минимизации большого количества возможных специфических отношений [103].

Существует несколько причин, делающих неэффективным использование статистических методов при решении задач Open IE, так же как и при подходе IE. Это во многом связано с тем, что в задачах IR и классификации/кластеризации текстов статистические методы рассматривают документ как неупорядоченный “пакет слов” (“bag-of-words”) [104]. При таком подходе теряются знания, связанные с языком — грамматикой и семантикой.

Второй причиной, не позволяющей использовать подход “пакета слов” в задачах IE и Open IE, является очевидная необходимость извлекать факты из предложений, а не из полного текста (full text) [105]. Такой подход связан с представлением факта в форме триплета: *Субъект – Отношение – Объект*. В данной парадигме знания о некоторой ПрО представляют собой совокупность сведений об объектах/субъектах данной ПрО, их существенных свойствах и связывающих отношениях, представленных в изолированных предложениях.

Третьей причиной низкой эффективности использования статистических методов для извлечения фактов является синонимия и неоднозначность языковых единиц, приводящая к частому появлению скрытых фактов в тексте [76]. Одной из таких проблем является необходимость разрешение кореференции, когда одни и те же сущности и/или действия представлены разными словами (иногда, разными частями речи). Например, такие предложения как "The company management sold a part of share", "Management of Apple Inc. sold their share", и "They marketed its" могут выражать один и тот же факт.

На сегодняшний день проблема автоматического извлечения фактов исследуется для всех языков. Во многих случаях она имеет высокий уровень реализации не только для текстов английского языка, но и для многих других. Например, в работе [105] был проведен эксперимент для оценки адекватности использования фактической плотности к информативности 50 случайно выбранных документов испанского языка в CommonCrawl корпусе. В работе [106] представлена первая система Open IE, которая способна извлекать триплеты

фактов из произвольных текстов китайского языка.

Однако, не смотря на достигнутые хорошие результаты в исследованиях, связанных с извлечением структурированной информации из текстов, на сегодняшний день не существует мультиязычных стандартных методик и подходов Open IE [107], в частности, для языков с ограниченными лингвистическими ресурсами, к которым относится и казахский язык.

1.5. Постановка задачи исследования

Сегодня в корпоративных и интернет компьютерных сетях накоплены огромные объемы текстовой информации, представленной в неструктурированном или слабоструктурированном виде. Анализ таких текстов, включающий грамматическую, а главное, семантическую обработку, базирующуюся на методах ИИ, становится дополнительным мощным источником для решения таких задач как, автоматическая генерация онтологий на базе корпуса текстов, разработка интеллектуальных вопросно-ответных систем, бизнес-аналитика (например, прогнозирование ожидаемого изменения фондового рынка) и так далее.

В тоже время, инициированный Главой государства и реализуемый в республике Казахстан проект триединства языков, рассматривающий казахский язык в качестве государственного языка, русский язык — в качестве языка межнационального общения и английский язык — в качестве языка успешной интеграции в глобальную экономику [2], базируется на свободном использовании и, следовательно, и включении, в том числе, в качестве интерфейса, в различные информационные приложения, казахского, русского и английского языков.

Однако, сегодня, несмотря на большое количество исследований, связанных с обработкой и анализом многоязычной текстовой информации, разработка методов, моделей и технологий обработки текстов казахского, русского и английского языка остается по-прежнему востребованным научным направлением.

За два предыдущих десятилетия задача формализации и автоматической обработки морфологии и синтаксиса многих языков, особенно английского, получила значительное количество удовлетворительных и качественных решений. К настоящему моменту разработано огромное количество корпусов, имеющих как морфологическую, так и синтаксическую разметку, на базе которых получено множество моделей и методов, в том числе, использующих машинное обучение, морфологического (POS-tagging) и синтаксического (Parser) анализа.

Однако, существующие коммерческие приложения NLP, в своем большинстве, используют только грамматический этап обработки текста (морфологический и синтаксический). В то же время, семантический анализ текста или анализ смысла естественного языка, по-прежнему остается одной из ключевых проблем, как теории искусственного интеллекта, так и компьютерной лингвистики. В большинстве востребованных сегодня систем автоматической обработки текстов на естественном языке, таких как рекомендательные системы, системы машинного перевода, системы проверки плагиата, синонимайзеры и другие, обязателен и необходим этап смысловой или семантической обработки.

Для решения задач обработки текстов сегодня используются подходы, базирующиеся на правилах, методы традиционного машинного обучения и глубокое машинное обучение. Однако, поскольку методы, базирующиеся на правилах, требуют высоких интеллектуальных затрата, компьютерные процессоры обладают все большей мощью, а накопители все большим возможным объемом хранимой информации, в связи с этим, все большее распространение приобретают методы машинного обучения.

Использование методов машинного обучения для одной из базовых задач обработки текстов — его морфологической разметки, показывает результаты сравнимые с использованием гибридных методов, объединяющих правила и вероятностные вычисления, хотя так же требует большого размеченного корпуса.

Однако, наибольшей проблемой, использования машинного обучения в задачах и приложениях NLP, является необходимость наличия больших и сверх больших ресурсов корпусов (*train corpus*) текстов соответствующего языка, в которых уже осуществлена некая маркировка (морфологическая, синтаксическая или семантическая). При наличии такого корпуса, основным этапом разработки метода обучения становится конструирования признаков, на базе имеющейся разметки. И, если морфологически размеченные корпуса русского и английского языков существуют, то разработка семантически размеченных корпусов даже для английского языка, остается сложной задачей.

Одним из направлений связанным с семантическим разметкой и анализом текстов является извлечение информации из текста. Возрастающий интерес использования подходов Open IE на этапе семантического анализа текстов связан с расширением приложений NLP, использующим автоматически извлеченные из текста факты. Это такие направления, как создание автоматическое создание онтологий и использование меры плотности фактов для оценки качества Интернет-контента. Однако, не смотря на достигнутые хорошие результаты в исследованиях, связанных с извлечением структурированной информации из текстов, на сегодняшний день не существует мультиязычных стандартных методик и подходов Open IE [107], в частности, для языков с ограниченными лингвистическими ресурсами, к которым относится и казахский.

Для казахского языка, наряду с проблемами семантической обработки, не решенными остаются проблемы морфологической и синтаксической обработки, что, по большей части, связанно со сложностью его грамматики. Проведенный анализ показывает, что существующие на сегодня приложения по обработке казахского языка нуждаются в новых математических моделях, методах и технологиях, позволяющих повысить качества обработки текстов.

Анализ казахского языка с точки зрения возможности его формализации, моделирования и автоматической обработки показал его основные особенности: строгий порядок слов в предложении и агглютинативность языка (слово включает целый ряд морфем, следующих последовательно одна за другой, каждая из которых имеет свое грамматическое и семантическое значение). Выделение данных морфем, их порядка и значения позволяет формализовать семантическое значение, как слова, так и предложения в целом.

В рамках обозначенных проблем представляется важным решение научных задач по моделированию процессов интеллектуальной обработки многоязычных текстов, а так же по разработке моделей, методов, алгоритмов и программ, осуществляющих анализа многоязычной текстовой информации, с целью определения основных характеристик текстов для построения моделей машинного обучения, их оценки на базе размеченных корпусов.

Все вышесказанное обусловливает актуальность развития моделей, методов и алгоритмов обработки текстовой информации на основе машинного обучения и применения этих методов для автоматизации анализа многоязычной текстовой информации, что и составляет основное направление диссертационной работы.

Поскольку целью диссертационной работы является повышение качества работы систем автоматической обработки текстовой информации казахского, русского и английского языков за счет использования моделей интеллектуального анализа и методов машинного обучения, то в соответствии с поставленной целью в работе рассмотрены следующие задачи:

– Разработка моделей и алгоритмов извлечения фактов из слабоструктурированных и неструктурированных текстовых массивов казахского, русского и английского языков;

– Разработка методов и алгоритмов морфологического и семантического анализа многоязычных текстов на базе моделей машинного обучения;

– Разработка методов вероятностного POS-тегинга, использующих скрытую Марковскую модель;

– Разработка метода, определения семантической близости многоязычных текстовых документов, базирующегося на использовании Vector Space Model;

– Разработка программного продукта, позволяющего осуществлять семантическую обработку текстов, на базе грамматически и семантически размеченных корпусов, имплементирующего разработанные модели, методы и алгоритмы

2. РАЗРАБОТКА МОДЕЛЕЙ И АЛГОРИТМОВ ИЗВЛЕЧЕНИЯ ФАКТОВ ИЗ СЛАБОСТРУКТУРИРОВАННЫХ И НЕСТРУКТУРИРОВАННЫХ ТЕКСТОВЫХ МАССИВОВ КАЗАХСКОГО, РУССКОГО И АНГЛИЙСКОГО ЯЗЫКОВ

2.1. Математическая модель извлечения фактов из многоязычной текстовой информации.

Рассмотренный в § 1.4 подход OIE позволяет извлекать триплеты факта Subject – Predicate – Object из коллекции Веб-документов, предварительно не определяя конкретные типы отношений. Поскольку такого рода факт обычно выражается разными нерегламентированными конструкциями естественного языка, то для его идентификации необходимо определить в предложении лексические единицы, называющие участников действия (Субъект и Объект), и затем определить смысловые отношения между ними [108].

Для задания таких смысловых связей предлагается использовать семантические функции, выражаемые через отношение морфологических и семантических категорий партиципантов предложения средствами алгебры конечных предикатов (АКП) [109].

Таким образом, базовым математическим аппаратом, используемым для построения нашей логико-лингвистической модели извлечения фактов из текстов различных языков, является АКП [110]. В работах [110–113] показано, что язык алгебры конечных предикатов представляет собой универсальное средство формального описания тех интеллектуальных функций человека, которые пытаются передать компьютерным системам. На языке алгебры предикатов выражаются любые отношения. Отношения выражают свойства предметов и связи между ними. Они представляют собой универсальное средство формального описания любых объектов. Алгебра предикатов является основной алгеброй теории интеллекта, которая наряду с предикатными операциями рекомендуется для формального описания механизмов естественного языка и мышления, при разработке высокопроизводительных мозгопо-добных ЭВМ параллельного действия [111].

АКП полностью характеризуется алфавитом A, состоящим из k символов a_1, a_2, … , a_k и алфавитом переменных B, состоящих из n символов x_1, x_2, … , x_n . Если $a_1 \in A_1, a_2 \in A_2, \ldots, a_m \in A_m$ и $x_1 = a_1, x_2 = a_2, \ldots, x_m = a_m$, то пишут $(a_1, a_2, \ldots, a_m) \in S$ и говорят, что предметный вектор $(a_1, a_2, \ldots, a_m)$ принадлежит пространству $S = A_1 \times A_2 \times \ldots \times A_m$. Элементы $a_1, a_2, \ldots, a_m$ вектора $(a_1, a_2, \ldots, a_m)$ называются его компонентами (первым, вторым, …, m-тым). Предметное пространство S можно рассматривать как совокупность всех векторов вида $(x_1, x_2, \ldots, x_m)$, компоненты которых удовлетворяют условию $x_1 \in A_1, x_2 \in A_2, \ldots, x_m \in A_m$. Любое подмножество P пространства S называется отношением, образованным в (или иначе: заданным на) пространстве S. Отношение имеет размерность m. Говорят, что оно m-местно. Отношения, заданные на одном и том же пространстве S, называются однотипными. Тип отношения определяется набором переменных $x_1, x_2, \ldots, x_m$ и

набором множеств $A_1, A_2, ..., A_m$. Отношение $\varnothing$, не содержащее ни одного вектора, называется пустым, отношение S, в котором имеются всевозможные векторы, – полным.

Предикатом, заданном на декартовом произведении $A_1, A_2, ..., A_m$, называется любая функция $P(x_1, x_2, ..., x_m) = \xi$, отображающая декартово произведение $A_1 \times A_2 \times ... \times A_m$ множеств $A_1, A_2, ..., A_m$ в множество $\Sigma = \{0, 1\}$. Символы 0 и 1 называются булевыми элементами, Σ – множество всех булевых элементов. Переменная $\xi = \{0, 1\}$, являющаяся значением предиката P, называется булевой. Предикат $P(x_1, x_2, ..., x_m)$ на $A_1 \times A_2 \times ... \times A_m$ называется конечным, если все множества $A_1, A_2, ..., A_m$ конечны, и бесконечным – в противном случае. Эта же терминология переносится и на соответствующие предикатам отношения. Переменные $x_1, x_2, ..., x_m$ называются аргументами предиката P.

Средствами АКП может быть описан любой *n -местный* предикат $f(x_1, x_2, \ldots, x_n)$, заданный алфавитом A. Формулы АКП состоят из символов: $a_1, a_2, \ldots, a_k$, переменных $x_1, x_2, \ldots, x_n$, знаков дизъюнкции $\vee$, конъюнкции $\wedge$, логических констант 0 и 1, называемых соответственно ложью и истиной.

Под универсумом элементов U мы будем понимать все возможные объекты сложной языковой системы (слова, лексемы, сигнификаты, денотаты, грамматические и семантические характеристики слов, характеристики коллокаций и т.д.). Используемое понятие универсума определяется как класс заранее зафиксированных допустимых объектов, включающий несколько конечных множеств [114].

Из элементов универсума, сообразуясь с конкретной задачей обработки информации, представленной на естественном языке, образуем подмножества X_{1i}, X_{2i}, ..., X_{ni},. На декартовых произведениях $X_{1i} \bullet X_{2i} \bullet X_{3i} \bullet ... \bullet X_{ni}$ определены предикаты P_j, которые характеризуют работу системы, выполняющую семантическую или смысловую обработку информационных потоков сложной естественно языковой системы.

Предикатом P, заданным на U^n, называется любая функция $\varepsilon = P(x_1, x_2, \ldots, x_n)$, отображающая множество U^n, во множество $\sum = \{0,1\}$. Переменные $x_1, x_2, \ldots, x_n$, называются предметными или предикатными, а их значения предметами. При $n=1$ предикат P является унарным, при $n=2$ — бинарным, при $n=3$ — тернарным. Если множество U конечно, как при моделировании процессов языка (с конечным алфавитом, конечным количеством морфологических и семантических свойств, конечным словарем), то и предикат P является конечным. Предикаты, обозначаемые *1* и *0*, называются тождественно истинными и тождественно ложными соответственно.

Множество всех *n-арных* предикатов, заданных на U^n, на котором определены операции дизъюнкции, конъюнкции и отрицания, называется алгеброй *n-арных* предикатов на U. При этом операции дизъюнкции, конъюнкции и отрицания являются базисными для алгебры предикатов. Алгебра предикатов при любом значении n является разновидностью булевой алгебры, в ней выполняются все основные тождества булевой алгебры. Базисными предикатами для алгебры

предикатов будут предикаты вида:

$$x_i^a = \begin{cases} 1, если\ x_i = a, \\ 0, если\ x_i \neq a. \end{cases} \quad (2.1)$$

где $i = \{1,2, \ldots, n\}$, a – любой элемент универсума. Предикат вида (2.1) называется предикатом узнавания предмета a по переменной x_i. Если универсум конечен и состоит из r элементов, всего имеется $r \bullet n$ различных базисных элементов.

Алгебра предикатов полна в том смысле, что любой ее предикат можно представить в виде суперпозиции базисных операций, примененных к базисным элементам. Показано, что на языке АКП могут быть описаны любые конечные отношения, поэтому любой другой математический аппарат, предназначенный для описания произвольных конечных отношений, в логическом смысле, обязательно будет эквивалентен алгебре конечных предикатов [115].

Согласно теоретическим положениям АКП [111], на универсуме U, включающем все возможные понятия и объекты анализа сложной языковой системы [116] вводим конечное, дискретное и детерминированное множество грамматических и семантических характеристик слов предложения конкретного языка влияющих на понятийные связи $M = \{m_1, \ldots, m_n\}$, где n — количество указанных характеристик. Отношения между введенными характеристиками можно представить в виде декартового произведения $m_i * m_j * \ldots * m_k$, где $m_i, m_j, \ldots, m_k \in M$.

На множестве M введем систему предикатов S так, чтобы любой предикат $P(x_i) \in S$, равнялся 1 на множестве слов предложения с грамматико-семантической информацией, соответствующей определенной семантической роли или некоторому глубинному смысловому отношению между словами, грамматические характеристики которых выражаются $m_i, m_j, \ldots, m_k$., и был равным 0 в противном случае [117]. Таким образом, множество предикатов S можно сопоставить с множеством семантико-грамматических характеристик, которые соответствуют партиципантам предложения.

N-мерный предикат $P(x_1,..., x_n)$ определяет семантическую роль участника действия через предметные переменные, описывающие грамматические и семантические характеристики слова предложения, называющего данного участника:

$$P(x_1,..., x_n) \rightarrow P(x_1) \wedge \ldots \wedge P(x_n) \quad (2.2)$$

Предикат $P(x_1,..., x_n) = 1$, если анализируемое слово, выполняющее некоторую семантическую функцию, обладает определенными морфологическими и семантическими характеристиками заданного языка. Это значит, что слово, конъюнкция грамматических и семантических характеристик которого описывается предикатом (2.2) называет участника (Субъект или Объект) или атрибут действия. Очевидно, что описанные уравнением отношения грамматических характеристик не зависят от конкретного слова.

На практике подмножество согласованных морфологических, синтаксических и семантических признаков участников действия не совпадает с декартовым

произведением по совокупности всех признаков.

На декартовом произведении $S \times \mathrm{S}$ можно определить предикат:

$$P(x_1,..., x_n) = \gamma_k(x_1,..., x_n) \times P_1(x_1) \times .... \times P_n(x_n), \quad (2.3)$$

здесь $k \in [1,h]$, где h — число рассматриваемых в модели участников и атрибутов действия. Предикат $\gamma_k(x_1,..., x_n)$ равен единице, если отношения конкретных морфологических и синтаксических характеристик слов предложения выражают определенную семантическую роль участников (Субъект, Объект) или атрибутов действия и $\gamma_k(x_1,..., x_n)$=0, если конъюнкция грамматических категорий не выражает никаких семантических ролей. Таким образом, если отношения между грамматическими и семантическими характеристиками слов предложения ни выражают какой-либо элемент факта, они исключаются из формулы (2.3) предикатом $\gamma_k(x_1,..., x_n)$.

Мы рассматриваем имплементацию нашей логико-лингвистической модели Open IE для английского, казахского и русского языков. Семантическая функция участников и атрибутов действия явным образом выражается отношением грамматических и семантических характеристик слов в предложении во всех этих трех языках [118]. Однако, в связи с существующим различием в грамматике данных языков, существуют особенности реализации модели в каждом из них. Это связано с тем, что семантическая согласованность на поверхностном уровне языка явным образом выражается: (1) порядком слов и наличием предлогов в английском языке; (2) грамматическими падежами в русском языке и (3) порядком слов, грамматическими падежами, личными окончаниями, словообразовательными суффиксами и некоторыми другими признаками в казахском языке.

Из всего вышесказанного можно сделать вывод, что при реализации данной модели для трех вышеперечисленных языков определяются различные наборы предметных переменных и вводятся разные предикаты, описывающие семантические функции [119].

2.2. Реализация логико-лингвистической модели Open IE для русского и английского языков.

Основная идея разработанной в §2.1 модели заключается в использовании морфологических и синтаксических характеристик слов предложения для выявления их роли в называемом факте [120]. Такими характеристиками для английского языка являются: наличие предлога после глагола, притяжательный падеж существительного или местоимения, положение (месторазмещение) существительного во фразе, наличие любой формы глагола «to be» и основная форма глагола в анализируемой фразе [150].

Введем конечное множество предметных переменных, описывающих грамматические характеристики слов английского предложение $\{x, z, m, f, n, p\}$ [121]. Предикат $P_z(z)$ описывает синтаксическую характеристику наличия конкретного предлога после глагола во фразе:

$$P_z(z) = z^{to} \vee z^{by} \vee z^{with} \vee z^{about} \vee z^{of} \vee z^{on} \vee z^{at} \vee z^{in} \vee z^{beween} \vee z^{for} \vee z^{from} \vee z^{over} \vee z^{out}, \quad (2.4)$$

где предметная переменная z^{prep}, обозначает конкретный предлог, стоящий после глагола в предложении, *prep*={*to, by, with, about, of, on, at, in, between, for, from, over*}, z^{out} показывает отсутствие любого предлога после глагола.

Предикат P_x (x) идентифицирует порядок слов в английской фразе. А именно, место анализируемого существительного в предложении.

$$P_x(x) = x^f \vee x^l \vee x^{kos}. \quad (2.5)$$

В этом уравнении: $x^f = 1$, если анализируемое существительное расположено перед главным глаголом во фразе; $x^l = 1$, если анализируемое существительное расположено после главного глагола и $x^{kos} = 1$, если анализируемое существительное расположено на месте косвенного дополнения.

Предметная переменная *y* определяет наличие притяжательного падежа через существование апострофа у анализируемого существительного

$$P_y(y) = y^{ap} \vee y^{aps} \vee y^{out} = 1, \quad (2.6)$$

где y^{ap} и y^{aps} показывают существование апострофа или апострофа с буквой *s* (*'s*) в конце анализируемого слова, тогда как y^{out} показывает его отсутствие.

Предикат P_m (m) идентифицирует существование любой формы глагола "*to be*" в анализируемой английской фразе.

$$P_m(m) = m^{is} \vee m^{are} \vee m^{havb} \vee m^{hasb} \vee m^{was} \vee m^{were} \vee m^{out}. \quad (2.7)$$

В этом равенстве значение предметной переменной m определяет конкретную форму глагола "*to be*" или его отсутствие m^{out}. Например, $m^{havb} = 1$, если во фразе присутствует вспомогательный глагол "*have been*".

Аналогичным образом, предикаты P_f (f) и P_n(n) соответственно показывают наличие любого модального глагола и отрицания во фразе.

$$P_f(f) = f^{can} \vee f^{may} \vee f^{must} \vee f^{should} \vee f^{could} \vee f^{need} \vee \vee f^{might} \vee f^{would} f^{out}, \quad (2.8)$$

$$P_n(n) = n^{\text{not}} \vee n^{out}. \quad (2.9)$$

Дополнительно, мы вводим предикат P_p(p), который показывает форму главного глагола в английской фразе:

$$P_p(p) = p^{III} \vee p^{ed} \vee p^{I} \vee p^{ing} \vee p^{II}, \quad (2.10)$$

здесь p^I =1, если основной глагол присутствует в своей базовой форме (используется инфинитивная форма без частицы to); p^{ed} =1, если глагол стоит в прошедшем времени с окончанием – *ed*; p^{III}, p^{II} и p^{ing} показывают третью и вторую формы неправильных глаголов и форму глагола с *ing* окончанием соответственно.

В этом случае, основываясь на равенствах (2.4) – (2.10) для английского языка формула (2.3) будет иметь следующий вид:

$$P(x,y,z,m,p,f,n) = \gamma_k(x,y,z,m,p,f,n) \times P_x(x) \times P_y(y). \times P_z(z) \times P_m(m) \times P_p(p) \times P_f(f) \times P_n(n). \quad (2.11)$$

Предикат γ_{1E} определяет отношения грамматических и семантических характеристик слов, обозначающих Субъект триплета факта в английской фразе:

$$\begin{aligned}\gamma_{1E}(z,\ y,\ x,\ m,\ p,\ f,\ n) = y^{out}\ ((f^{can} \vee f^{may} \vee f^{must} \vee f^{should} \vee f^{could} \vee f^{need} \vee f^{might} \vee \\ \vee f^{would} \vee f^{out})(n^{not} \vee n^{out})\ (p^{I} \vee p^{ed} \vee p^{III})\ x^{f}\ m^{out} \vee (x^{l}\ (m^{is} \vee \\ \vee\ m^{are} \vee\ m^{havb} \vee\ m^{hasb} \vee\ m^{hadb} \vee\ m^{was} \vee\ m^{were} \vee \\ \vee\ m^{be} \vee m^{out}\ )\ z^{by}).\end{aligned} \quad (2.12)$$

Мы так же можем явным образом выделить Объект или участника действия, на которого это действие направленно, через предикат γ_{2E}:

$$\begin{aligned}\gamma_{2E}(z,\ y,\ x,\ m,\ p,\ f,\ n) = y^{out}\ (n^{not} \vee n^{out})\ (f^{can} \vee f^{may} \vee f^{must} \vee f^{should} \vee f^{could} \vee f^{need} \vee \\ \vee f^{might} \vee f^{would} \vee f^{out})\ (z^{out}\ x^{l}\ m^{out}\ (p^{I} \vee p^{ed} \vee p^{III}) \vee x^{f}\ (z^{out}\ \vee \\ \vee z^{by})\ (m^{is}\ \vee m^{are} \vee\ m^{havb} \vee\ m^{hasb} \vee\ m^{hadb} \vee\ m^{was} \vee \\ \vee\ m^{were}\ \vee m^{be} \vee m^{out}\ )(p^{ed} \vee p^{III}).\end{aligned} \quad (2.13)$$

Кроме предикатов γ_{1E} и γ_{2E}, выделяющих главных участников действия (Subject и Object), описываемого английским предложением, мы можем получить предикаты, позволяющие выделить атрибуты времени, места и принадлежности действия. Например, мы можем выделить существительное, обозначающее атрибут времени действия через предикат γ_{3E}, показывающий дизъюнкцию конъюнкций морфологических и синтаксических характеристик анализируемого слова предложения:

$$\begin{aligned}\gamma_{3E}(z,\ y,\ x,\ m,\ p,\ f,\ n) = (y^{out}\ z^{on}x^{kos} \vee y^{out}\ z^{in}x^{kos} \vee y^{out}\ z^{at}x^{kos})(n^{not} \vee n^{out})\ (f^{can} \vee \\ \vee f^{may} \vee f^{must} \vee f^{should} \vee f^{could} \vee f^{need} \vee f^{might} \vee f^{would} \vee f^{out}) \\ (z^{out}\ x^{l}\ m^{out}\ (p^{I} \vee\ p^{II}\ \vee\ \vee p^{ed} \vee p^{III})\ (m^{is}\ \vee m^{are} \vee\ m^{havb} \vee \\ \vee\ m^{hasb}\ \vee m^{hadb} \vee\ m^{was} \vee\ m^{were} \vee m^{be} \vee m^{out}\ ))\end{aligned} \quad (2.14)$$

Предикат γ_{4E} позволяет идентифицировать отношение локации, т. е. существительное, дизъюнкция конъюнкций грамматических характеристик которого позволяет определить его как слово, называющее место или направление действия.

$$\begin{aligned}\gamma_{4E}(z,\ y,\ x,\ m,\ p,\ f,\ n) = ((z^{to} \vee z^{from} \vee z^{between})\ (p^{I} \vee\ p^{II}\ \vee p^{ed} \vee p^{III})\ (m^{is}\ \vee m^{are} \vee \\ \vee\ m^{havb} \vee\ m^{hasb} \vee\ m^{hadb} \vee\ m^{was} \vee\ m^{were} \vee m^{be} \vee m^{out}\ ) \\ (\ x^{kos} \vee x^{l} \vee x^{f})\ \vee z^{by}(p^{I} \vee\ p^{II}\ \vee p^{ed}\ )\ m^{out}\ (x^{kos} \vee x^{f})\ )\ y^{out} \\ (n^{not} \vee n^{out})\ (f^{can} \vee f^{may} \vee f^{must} \vee f^{should} \vee f^{could} \vee f^{need} \vee f^{might} \vee \\ \vee f^{would} \vee f^{out})(n^{not} \vee n^{out})\end{aligned} \quad (2.15)$$

На рисунке 2.1 показан пример имплементации разработанной логико-лингвистической модели для английской фразы: "*...the steamer moved slowly from the dock"*. В этой фразе глагол "*move*" определяет действие типа движения. Затем, согласно уравнению (2.12), мы можем идентифицировать существительное

"*steamer*" в качестве инициатора действия или как Субъект триплета факта. Дизъюнкция конъюнкций морфологических и синтаксических характеристик слова "*signal"* в предикате γ_{2E} (2.13) определяет Объект триплета факта, а предикат γ_{4E} (2.15) определяет слово "*dock*" как атрибут факта, определяющий направление движения.

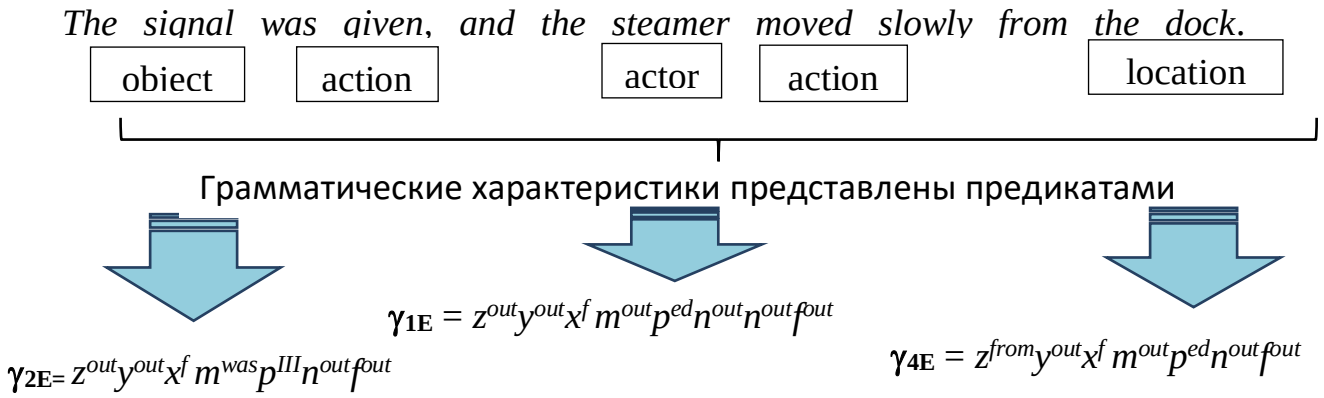


Рисунок 2.1 — Пример извлечения факта из английского предложения. Предикат γ_{1E} определяет отношения грамматических характеристик слова, называющего Субъект; предикат γ_{2E} определяет отношения грамматических характеристик слова, называющего Объект; предикат γ_{4E} определяет отношения грамматических характеристик слова, называющего атрибут места действия факта.

Русский язык является флективным языком и обладает более сложной и развитой морфологией, чем английский, поэтому основными характеристиками, позволяющими определить роли участников триплета факта, являются морфологические падежи и семантические признаки.

Адаптируя разработанную модель к русскому языку, вводим множество грамматических и семантических характеристик слов русскоязычных предложений $M = \{z, y, x\}$, где z — конечное подмножество грамматических падежей русских существительных, y — конечное подмножество возможных семантических характеристик существительного, а x — подмножество характеристик одушевленности.

Тогда, предикат P_z определяет шесть грамматических падежей русского языка (*nom* — именительный, *gen* — родительный, *dat* — дательный, *acc* — винительный, *ins* — творительный и *loc* — предложный):

$$P_z(z) = z^{nom} \vee z^{gen} \vee z^{dat} \vee z^{acc} \vee z^{ins} \vee z^{loc} \qquad (2.16)$$

Предикат $P_x(x)$ определяет семантическую характеристику одушевленности существительного:

$$P_x(x) = x^{anim} \vee x^{inan}, \qquad (2.17)$$

где значение предметной переменной *anim* определяет признак одушевленности существительного, а *inan* — соответственно, признак неодушевленности.

Предикат $P_y(y)$ идентифицирует специальные семантические характеристики существительного:

$$P_y(y) = y^{\text{device}} \vee y^{\text{hum}} \vee y^{\text{tool}} \vee y^{\text{pc:hue}} \vee y^{\text{spacee}} \vee y^{\text{time:moment}} \vee y^{\text{time:period}} \vee y^{\text{s:loc}}, \quad (2.18)$$

здесь значения *device* и *tool* показывают, что существительное называет концепт, являющийся устройством или инструментом, *hum* обозначает, что концепт, который называет существительное, принадлежит к семантическому классу "*person*", *pc:hum* обозначает семантический класс "часть тела", *time*:*moment* — семантическая характеристика “определенное время” и *time*:*period* — наличие семантического признака периода времени, *space* — семантическая характеристика месторасположения. При формировании множества значений предикатных переменных мы использовали таксономические отношения существительных Национального корпуса русского языка[3].

Согласно нашей модели, на следующем шаге определяем семантические роли Субъекта и Объекта триплета факта в русскоязычной фразе через предикаты γ_{1R} и γ_{2R}, соответственно:

$$\gamma_{1R}(x, y, z) = x^{anim} z^{\text{nom}} \vee x^{inan} z^{\text{nom}} (y^{\text{device}} \vee y^{\text{tool}} \vee y^{\text{pc:hue}}) \quad (2.19)$$

$$\gamma_{2R}(x, y, z) = z^{\text{acc}} (x^{inan} \vee x^{anim}) \quad (2.20)$$

Подобным адаптации модели для английского языка, мы можем выделить такие атрибуты факта как время, место, инструмент, бенефициар действия и другие. Например, предикат γ_{3R} определяет отношения грамматических и семантических характеристик инструмента, способа или причины действия:

$$\gamma_{3R}(x, y, z) = z^{\text{ins}} (x^{inan} (y^{tool} \vee y^{pc:hum} \vee y^{device}) \vee x^{anim}) \quad (2.21)$$

Семантическая функция временного атрибута действия во фразе на русском языке будет выражаться предикатом γ_{4R}:

$$\gamma_{4R}(x, y, z) = z^{acc}\ x^{inan}\ y^{time:moment} \vee z^{\ loc}\ x^{inan}\ y^{time:period}. \quad (2.22)$$

Например, реализация разработанной модели позволяет в русскоязычном предложении “*Через два дня приобретенной отмычкой подозреваемый взломал сейф”* определить Субъект и Объект триплета факта *Подозреваемый* → *взломал* → *сейф*, а так же атрибуты факта: временя действие — “*два дня”* и инструмент действия — “*отмычка”* (рис. 2.2).

[3] http://www.ruscorpora.ru/new/

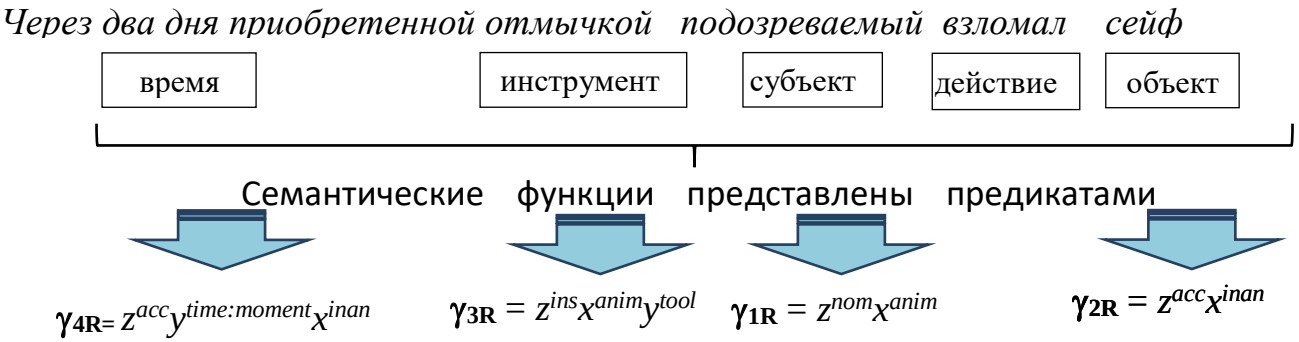


Рисунок 2.2 — Пример извлечения факта из предложения на русском языке. Предикаты γ_{1R} и γ_{2R}, определяют отношения грамматических характеристик слов, называющих Субъект, Объект соответственно, а предикаты γ_{3R} и γ_{4R} определяют атрибуты способа и времени действия соответственно.

2.3. Адаптация разработанной логико-лингвистической модели Open IE к текстам казахского языка

В отличие от русского и английского языков, казахский язык является агглютинативным. Это означает, что в противоположность, флективному русскому языку, при словообразовании которого каждая морфема имеет несколько значений (например, число и падеж существительного), при словообразовании казахского языка каждая морфема несет свое специфическое морфологическое или семантическое значение (например, лицо, падеж, число). Казахский язык отличается так же от аналитического английского, в котором вообще достаточно мало флексий.

При адаптации разработанной в § 2.1 модели извлечения фактов из слабоструктурированных текстов к казахскому языку, вводим достаточно четко сформированное несократимое множество M из десяти грамматических и семантических признаков, влияющих на семантическую роль партиципантов предложения [122, 123]. Большинство из данных признаков представляют собой морфологические или семантические характеристики, выраженные в поверхностной структуре языка с помощью аффиксов. Это такие характеристики как: положение анализируемого слова во фразе; наличие вспомогательного глагола во фразе; грамматический падеж анализируемого существительного; суффиксы множественного числа и лица; аффиксы предопределенного действия и другие, морфологические и семантические характеристики.

Наличие в модели казахского языка большего числа предметных переменных обусловлено, прежде всего, агглютинативностью языка, когда каждая грамматическая особенность выражается определенным аффиксом, а так же необходимостью выделения в казахском языке не только участников и атрибуты действия, но и различных типов действия [124].

Предикат $P_x(x)$ определяет местоположение анализируемого слова во фразе. Выбор признака местоположения слова в предложении предопределен знанием о строгом порядке слов в казахской фразе, когда на первом месте стоит подлежащее, затем дополнение и завершает предложение сказуемое, при этом определение всегда стоит перед определяемым, а обстоятельства места и времени

располагаются обычно перед подлежащим или перед прямым дополнением, но не после сказуемого. Например, ”*Кеше Мен дүкенде мұғалімді көрдім*”.

$$P_x(x) = x^1 \vee x^2 \vee x^3 \vee x^{-1} \vee x^{-2} \vee x^{-3} \vee x^0, \quad (2.23)$$

где 1, 2, 3, -1, -2, -3 показывают смещение слова во фразе, «минус» обозначает начало отсчета с конца фразы; 0 показывает любую другую позицию слова, кроме первых трех и последних трех слов в предложении.

Предикат $P_f(f)$ определяет признак наличия или отсутствия вспомогательного глагола во фразе:

$$P_f(f) = f^{aux} \vee f^0, \quad (2.24)$$

где f^{aux} определяет признак принадлежности основы слова к списку из 35основных вспомогательных глаголов казахского языка [*ал, бар, біт, бітір, бол, ғой, де, деген, дейтін, деп, еді, екен, емес, ер,ет, жазда, жат, жатыр, жет, жібер, жүр, кел, келеді, кет, кір, көр, қал, қой, сал, отыр, түс, тұр, шық, шығар*].

Предикат Pz(z) идентифицирует семь грамматических падежей казахского языка: именительный, родительный, дательно-направленный, винительный, местный, творительный и исходный:

$$P_z(z) = z^{Nom} \vee z^{Gen} \vee z^{Dat} \vee z^{Acc} \vee z^{Ins} \vee z^{Loc} \vee z^{Abl}, \quad (2.25)$$

где *Nom* — именительный падеж (атау септік); *Gen* — родительный падеж (ілік септік), определяемый с помощью списка падежных аффиксов [*-ның, -нің, -дың, -дің, -тың, -тің*]; *Dat* – направительно-дательный падеж (барыс септік) простого склонения определяемый с помощью списка падежных аффиксов [*-ға, -ге, -қа, -ке, -а, -е, -на, -не*]; *Acc* — винительный падеж (табыс септік), определяемый с помощью списка падежных аффиксов [*-ны, -н,-ні, -ды, -ді, -ты, -ті*]; *Loc* — местный падеж (жатыс септік), определяемый с помощью списка падежных аффиксов [*-да, -де,- нда, -нде, -та, -те*]; *Abl* — исходный падеж (шығыс септік), определяемый с помощью списка падежных аффиксов [*-дан, -ден, -тан, -тен, -нан, -нен*]; *Ins*— творительный падеж (көмектес септік), определяемый с помощью списка падежных аффиксов [*-мен, -менен, -бен,- бенен, -пен,-пенен*].

В связи с тем, что в казахском языке существует два типа склонения существительных: простое (безотносительно к обладателю) и притяжательное (с указанием обладателя), вводим предикат $P_a(a)$, определяющий два возможных типа склонения казахских существительных:

$$P_a(a) = a^{NSim} \vee a^{NPo}, \quad (2.26)$$

где *NSim* — признак простого склонения существительного, *а NPos* – признак притяжательного склонения существительного, определяемого наличием аффиксов [*м, -ым, -ім, -ң, -ың, -ің, -сы, -ы, -і, -сі, -ымыз, -іміз, -ыңыз, -ныз, -іңіз, -ңіз, -ы, -і*]. Суфиксы простого склонения соответствуют суффиксам соответствующих падежей, определяемых формулой (2,25).

Предикат $P_{\text{п}}(n)$ идентифицирует особенности отрицания в предложении казахского языка:

$$P_{\text{п}}(n) = n^{\text{me}} \vee n^{0} \vee n^{\text{емес}} \vee n^{\text{jok}} = 1, \qquad (2.27)$$

где n^{0} показывает отсутствие отрицательной частицы во фразе, *me* показывает наличие отрицательной частицы из списка [*ма*, *ме*, *ба*, *бе*, *па*, *пе*], *емес* – определяет наличие отрицания, описываемого словом *емес*, *jok* – определяет наличие в предложении отрицательного слова *жоқ*.

Предикат $P_c(c)$ определяет признак наличия или отсутствия аффикса множественности:

$$P_c(c) = c^{\text{tar}} \vee c^{\text{ter}} \vee c^{\text{dar}} \vee c^{\text{der}} \vee c^{\text{lar}} \vee c^{\text{ler}} \vee c^{0},, \qquad (2.28)$$

где значение предметной переменной c^0 равно единице, если слово употребляется в единственном числе, т.е. у слова отсутствует аффикс множественности. Значения предметной переменной *tar*, *ter*, *dar*, *der*, *lar*, *ler* показывают наличию у слова суффикса множественности *-тар*, *-тер*, *-дар*, *-дер*, *-лар*, *-лер* соответственно.

Предикат $P_{\text{у}}(y)$ идентифицирует признак наличия словообразовательного аффикса конкретной части речи — глагола, причастия, деепричастия и существительного:

$$P_y(y) = y^{\text{ParP}} \vee y^{\text{Vpas}} \vee y^{VaP} \vee y^{UnFu} \vee y^{FuCo} \vee y^{VAd} \vee y^{OAd} \vee \text{y}^{\text{Psuf}} \vee y^{Usuf} \vee \vee y^{Part} \vee y^{NoV} \vee y^{NoN} \vee y^{NCom} \vee y^{NDer} \vee y^{\text{y}} \vee y^{0}, \qquad (2.29)$$

где

– *0* определяет глагольную основу в чистом виде, в форме второго лица единственного числа будущего времени повелительного наклонения при обращении на “ты”;

– *y* определяет признак наличия аффикса инфинитивной формы;

– *Vad*, *Oad*, *VaP* – признаки деепричастия (*Vad* определяет признак деепричастия через аффиксы [*п, ып, ін*]; *Oad* определяет деепричастия на [*а*, *е*, *й*, *и*], *VaP* определяет деепричастия на [*ға*, *ғалы* , *ге*, *гелі*, *гі*, *ғы*, *ке*, *келі*, *қа*, *қалы* , *қі*, *қон*, *қы*]);

– *FuCo*,*UnF* признаки будущего времени изъявительного наклонения глагола (*FuCo* определяет список аффиксов глаголов будущего предположительного времени (future conjecture) [*ар, ер, ыр, ір*], *UnFu* определяет аффиксы неопределенного будущего времени (uncertain future tense) [*мақ, мек, пақ, пек, бақ, бек, пақшы, мақшы, мекші, пекші, бақшы, бекші*, пақшы, *мақшы*, *мекші*, *бақшы*, *бекші*];

– *Part*, *ParP* признаки словообразования причастий (*Part* определяет аффиксы словообразования причастия из списка [*ған*, *ген*, *қан*, *кен*, *қон*, *ға*, *ге*, *қа*, *ке*], *ParP* определяет аффиксы словообразования причастий из списка [*атын*, *етін*, *йтын*, *йтін*].

– *Vpas* определяет 20 специальных словообразовательных аффиксов глагола [*ді*, *дік*, *дық*, *дім*, *дің*, *ды*, *дік*, *дық*, *дым*, дың, қ, *ті*, *тік*, тім, *тің*, *ты*, *тік*, *тық*,

тым, *тың*];

– *Psuf* определяет 189 продуктивных аффиксов словообразования глаголов, в том числе залоговые аффиксы [*а*, *ай*, *ал*, *ан*, *ар*, *арыс*, *ға*, *ғал*, *ғар*, *ға* , *ге* , *гер* , *гі*, *гіз*, *гіздір*, *гіле*, *гір* , *гіт*, *ғы*, *ғыз*, *ғыздыр*, *ғызыл*, *ғыла*, *ғыр*, *ғыт*, *да*, *дан* , *дар*, *дас*, *дастыр*, де, *ден*, *дендір*, *дес*, *діг*, *дік* , *дір*, *діргіз*, *дық*, *дыр*, *дырғыз*, *дырыл*, *е*, *ей*, *ел*, *ен*, *ер*, *й*, *іг*, *іғ*, *ік* , *ікіс*, *іл* , *іле*, *ілде*, *ілу*, *імсіре*, *ін*, *індір*, *ініс*, *іну*, *іңкіре*, *ір*, *ірде*, *іре*, *ірей*, *іріс*, *ірке*, *іркен*, *іс*, *ісу* , *іт*, *ке* , *кер* , *кіз*, *кіле*, *кір* , *қа*, *қал*, *қан*, *қар*, *қе*, *қур*, *қыз*, қыла, *қыла*, *қыр*, *л*, *ла* , *лан*, *ландыр*, *лас*, *ластыр*, *лат*, *ле*, *лен*, *лендір*, *лес*, *лестір*, *лет*, *ліг*, *лік*, *лікіс*, *лін*, *ліс*, *лқа*, *лу*, *лығ*, *лық*, *лын*, *лыс*, *мала*, *меле*, *мсіре*, *мсыра*, *н*, *ні*, *ніл*, *ніс*, *ныл*, *ныс*, *ңгіре*, *ңғыра*, *ңкіре*, *ңқыра*, *ңра*, *ңре*, *р*, *ра*, *ре*, *с*, *са*, *сан*, *се*, *сен*, *сет*, *сетіл*, *сі*, *сін*, *сіре*, *стір*, *стыр*, *сы*, *сын*, *сыра*, *т*, *та*, *тан*, *тандыр*, *тас*, *те*, *тен*, *тендір*, *тес*, *тік*, *ттыр*, *тығ*, *тығс*, *тығыс*, *тық*, *тыр*, *тырыл*, *ура*, *ші*, *шы*, *ығ* , *ығыс*, *ық*, *ықыс*, *ыл*, *ыла*, *ылда*, *ылу*, *ылыс*, *ымсыра*, *ын*, *ындыр*, *ыну*, *ыныс*, *ыр*, *ыра*, *ырай*, *ырқа*, *ырқан*, *ырла*, *ыс*, *ысу*, *ыт*];

– *Usuf* определяет 65 малопродуктивных аффиксов словообразования глаголов [*азы*, *ақта*, *ал*, *ала*, *аңғыра*, *аура*, *бала*, *бе*, *беле*, *би*, *бі*, *бы*, *дала*, *ди*, *ді*, *ды*, *екте*, *ел*, *еңгіре*, *еуре*, *жи*, *жіре*, *жыра*, *зы*, *і*, *ін*, *ірей*, *іс*, *іт*, *қи*, *лі*, *лы*, *ма*, *мала*, *меле*, *ми*, *мсіре*, *мсыра*, *ңра*, *ңре*, *пала*, *пеле*, *пи*, *пі*, *пы*, *ра*, *ре*, *си*, *сіре*, *сый*, *сыра*, *т*, *ти*, *ті*, *ты*, *усіре*, *усыра*, *ши*, *ші*, *шы*, *ы*, *ын*, *ыра*, *ырай*, *ыс*, *ыт*].

– Четыре начения *NoN, NoV, Ncom, Nder* предметной переменной у определяют признак принадлежности токена к существительному через списки конкретных аффиксов:

NoN – определяет наличие аффикса именного образования существительного [*ғай*, *гей*, *гер*, *ғи*, *ғой*, *дас*, *дес*, *дік*, *дық*, *кер*, *кес*, *қай*, *қар*, *қи*, *қой*, *қор*, *лас*, *лес*, *лік*, *лық*, *ман*, *паз*, *пана*, *сақ*, *тас*, *тес*, *тік*, *тық*, хана, *ша*, *шақ*, *ше*, *шек*, *ші*, *шік*, *шы*, *шық*];

NoV – определяет наличие аффикса отглагольного образования существительного [*ақ*, *ба*, *бе*, *ғақ*, *ғаш*, *гек* , *гі*, *гіш*, *ғы*, *ғыш*, *дақ*, *дек*, *ек*, *ік*, *ім*, *іс*, *іш*, *к*, *кі*, *кіш*, *қ*, *қаш*, *қы*, *қыш*, *лақ*, *лек*, *м*, *ма*, *мақ*, *ме*, *мек*, *па*, *пақ*, *пе*, *пек*, *с*, *тақ*, *тек*, *уік*, *уық*, *ш*, *ық*, *ым*, *ыс*, *ыш*];

Ncom – определяет наличие сложного аффикса образования существительного [*герлік*, *гіштік*, *ғыштық*, *дастық*, *дестік*, *ділік*, *дылық*, *қорлық*, *ластық*, *лестік*, *лілік*, *лылық*, *паздық*, *сақтық*, *сіздік*, *сыздық*, *тастық*, *тестік*, *тілік*, *тылық*, *шақтық*, *шілдік*, *шілік*, *шылдық*, *шылық*];

Nder– определяет наличие аффикса экспрессивной оценки (уменьшительно-ласкательной и уничитительной оттенки) образования существительного [*жан*, *ке*, *қан*, *сымақ*, *тай*, *ш*, *ша*, *шақ*, *ше*, *шік*, *шық*].

Предикат $P_{\mathrm{d}}(d)$ определяет наличие признака сослагательности действия:

$$d^{\mathrm{shi}} \vee d^{0} = 1, \tag{2.30}$$

где значение предметной переменной *shi* определяет наличие у анализируемого слова сослагательных суффиксов *ші* или *ши*, а значение *0* определяет признак отсутствия сослагательных суффиксов.

Предикат $P_{\text{m}}(m)$ определяет признак наличия личного предикативного окончания или личного притяжательного окончания глагола и отглагольных форм:

$$P_m(m) = m^{\text{PrFl}} \vee m^{\text{PoFl}} \vee m^0, \tag{2.31}$$

где значения предметной переменной:

PrFl (personal predicative flexion) – определяет наличие личного предикативного окончания причастий, деепричастий, основных и вспомогательных глаголов [*біз*, *бін*, *быз*, *бын*, *ды*, *міз*, *мін*, *мыз*, *мын*, *піз*, *пің*, *пыз*, *пын*, *сіз*, *сіздер*, *сіндер*, *сің*, *сыз*, *сыздар*, *сың*, *сыңдар*, *ті*, *ты*];

PoFl (personal possessive flexion) – определяет признак наличия личного притяжательного окончания некоторых глагольных форм [*дар*, *йік*, *йін*, *йық*, *йын*, *іздар*, *к*, *қ*, *м*, *ндар*, *ң*, *ңдер*, *ңіз*, *ңіздер*, *ңыз*, *сіздер*, *сің*, *сіңдер*, *сыздар*, сың, *сыңдар*, *ыздар*];

0 – определяет отсутствие личного окончания глагола.

Предикат $P_b(b)$ определяет наличие некоторой дополнительной семантики или значения анализируемого глагола:

$$P_b(b) = b^{\text{se}} \vee b^{\text{mic}} \vee b^0, \tag{2.32}$$

где значение *mic* показывает предположительность действия, определяемое через наличие суффиксов [*мыс*, *міс*]; значение *se* показывает существование условного наклонения, определяемого суффиксами [*са*, *се*].

В таблице 2.1 представлены ранее определенные в логико-лингвистической модели Open IE казахского языка предметные переменные и их области изменения.

Таблица 2.1.

Предметные переменные и их области значений логико-лингвистической модели OIE казахского языка

Предметная переменная	Грамматический/семантический признак языка, описываемый переменной	Области изменения предметной переменной представлены уравнением
x	месторасположения анализируемого слова в предложении	(2.23)
f	наличие или отсутствие вспомогательного глагола во фразе	(2.24)
z	грамматический падеж казахского существительного	(2.25)
a	склонения существительного	(2.26)
n	наличие отрицания во фразе	(2.27)
c	наличие аффикса множественности	(2.28)
y	словообразовательный аффикс конкретной части	(2.29)

	речи (глагола, причастия, деепричастия, существительного)	
d	сослагательность действия	(2.30)
m	наличие личного предикативного или притяжательного окончания глагола и отглагольных форм	(2.31)
b	дополнительная определяемая семантика действия	(2.32)

Полученные уравнения (2.23 – 2.32) позволяют преобразовать предикат согласованности грамматических и семантических особенностей слов, являющихся элементами факта (2.3) для казахского языка к следующему виду[152]:

$$P() = \gamma_k \times P_x(x\ ) \times P_y(y\ ) \times P_z(z\ ) \times P_f(f\ ) \times P_m(m\ ) \times P_n(n\ ) \times P_a(a\ ) \times P_b(b\ ) \times P_c(c\ ). \times P_{\mathrm{d}}(d). \quad (2.33)$$

Предикат инициатора действия или *Subject* факта мы определяем через γ_{1K}:

$$\gamma_{1K} = (x^1 \vee x^2 \vee x^3) z^{Nom} (c^{\mathrm{tar}} \vee c^{\mathrm{ter}} \vee c^{\mathrm{dar}} \vee c^{\mathrm{der}} \vee c^{\mathrm{lar}} \vee c^{\mathrm{ler}} \vee c^0) \quad (2.34)$$

Семантическую роль *Объекта* факта в казахской фразе, т.е. лицо или предмет, на которое направлено действие, определяем с помощью γ_{2K}:

$$\gamma_{2K} = (x^0 \vee x^2 \vee x^3)(z^{Gen} \vee z^{Acc})(y^{NoV} \vee y^{NoN} \vee y^{NCom} \vee y^{NDer} \vee y^0) \wedge \wedge (c^{\mathrm{tar}} \vee c^{\mathrm{ter}} \vee c^{\mathrm{dar}} \vee c^{\mathrm{der}} \vee c^{\mathrm{lar}} \vee c^{\mathrm{ler}} \vee c^0) a^{\mathrm{NSim}} \quad (2.35)$$

Формируя логико-лингвистическое уравнение *Предиката* действия в казахской фразе, мы основываемся на определении факта. Согласно Новой философской энциклопедии [125], факт представляет собой реальное, конкретное единичное событие или результат, которое произошло или произойдет. Таким образом, мы принимаем во внимание, только изъявительное наклонение казахского языка, оставляя повелительное, желательное и условное наклонения за границами исследования.

Предикат γ_{VK} определяет комбинацию семантических и грамматических признаков центральной части триплета факта, а именно *Действия* или *Предиката* факта:

$$\begin{gathered}\gamma_{VK} = (x^{-1} \vee x^{-2} \vee x^{-3})((f^{\mathrm{tur}} \vee f^{\mathrm{otur}} \vee f^{\mathrm{jatyr}} \vee f^{\mathrm{jur}})\, m^{\mathrm{PrF}} y^{\mathrm{Vad}} \vee (y^{\mathrm{Oad}} \vee y^{\mathrm{FuCo}}) m^{\mathrm{PrFl}} \vee \\ y^{\mathrm{FuCo}} (m^{\mathrm{PrFl}} \vee (m^{\mathrm{PrFl}} f^{\mathrm{edi}})) \vee y^{y} (f^{\mathrm{edi}} \vee f^{\mathrm{eken}}) \vee (y^{\mathrm{Vad}}\, m^{\mathrm{PrFl}} (b^{\mathrm{mic}} \vee b^0)) \vee \\ \vee m^{\mathrm{PoFl}} ((y^{\mathrm{Vart}} \vee y^{\mathrm{Vpa}} \vee y^{\mathrm{Vpas}}) \vee f^{\mathrm{edi}} (n^{\mathrm{joq}} \vee n^{\mathrm{emes}} \vee n^{\mathrm{me}} \vee n^0) \wedge \\ \wedge (y^{\mathrm{Part}} \vee y^{\mathrm{Vad}} \vee f^{\mathrm{otur}} \vee f^{\mathrm{tur}} \vee f^{\mathrm{jatyr}} \vee f^{\mathrm{jur}} \vee y^{\mathrm{ParP}} \vee y^{\mathrm{UnFu}})))\end{gathered} \quad (2.36)$$

На рисунке 2.3 показан пример реализации модели для предложения казахского языка. В казахской фразе «*Операторлар үйде мылтық тапты*», согласно формуле (2.36) глагол «*тапты*» представляет действие (давно прошедшее время). Конъюнкция грамматических и семантических характеристик слова «*Операторлар*» соответствует уравнению (2.34) и, следовательно, данное слово идентифицируется как *Субъект* действия или *Субъект* факта. Тогда как

предикат γ_{2K} (2.35) идентифицирует существительное «*мылтық*», как *Объект* факта, называемого данной фразой.

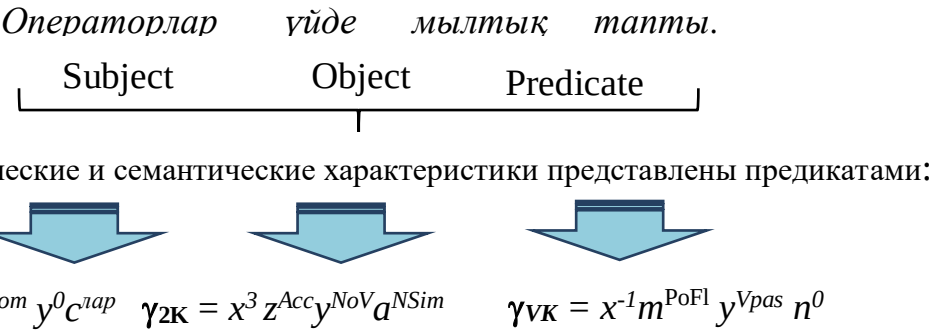


Рисунок 2.3 —Пример идентификации факта во фразе казахского языка. Предикат **γ1к** определяет грамматические особенности *Subject* действия, предикат **γ2к** определяет *Object* и **γvк** описывает Predicate факта.

2.4 Алгоритм излечения факта из предложения английского языка, на примере грамматических форм побуждения к действию.

Мы определяем следующие способы перефразирования одного и того же смысла в двух разных предложениях [126]: (1) использование близкой по смыслу лексики; (2) изменение порядка слов; (3) использование двух разных грамматик; (4) замена существительных их определениями; (5) объединение или разбиение фраз.

Наиболее часто используемым является использование синонимичной лексики. Для определения идентичности фактов, передаваемых такими предложениями, используются тезаурусы или словари синонимов предметной области [127].

Для английского языка в качестве словаря синонимов широких областей знаний используется тезаурус понятийной системы английского языка WordNet. Хотя в синонимичных рядах (синсетах) тезауруса понятия связаны различными парадигматическими и синтагматическими отношениями (отношения гипо-гиперонимии, холонимо-меронимии и др.), базовым лексическим отношением WordNet является отношения синонимии, а главным видом логических отношений является иерархическая подчиненность слов [128]. Причем отношения в WordNet связывают именно понятия, а не слова.

При этом, хотя слова часто и имеют множество синонимов, однако синонимы могут довольно сильно отличаться друг от друга по значению. Следовательно, существование множества пар синонимов в предложениях может привести к изменению смысла, выражаемого факта, т.е. проще говоря, к иному факту.

Наиболее полно смысл факта сохраняется при втором способе перефразирования, а именно изменении порядка слов в предложении или изменении грамматической конструкции. Изменение порядка слов в цепочке, является наиболее простым способом выражения одного и тоже же смысла (факта), так как слова (лексемы), включенные в предложение, остаются неизменными.

Данный способ хорошо применим, например, для русского языка. Однако его не так просто применить к казахским и английским предложениям, у которых порядок слов жестко закреплен грамматиками языков.

В то же время, не смотря на кажущуюся сложность использования различных грамматик для выражения одного и того же факта, данный способ представляется более простым, чем замена словарного запаса. Кроме того, при изменении грамматики смысл факта меняется редко, тогда как, ошибки при замене словарного запаса могут привести к искажению смысла факта.

На базе разработанной в подразделе 2.1 математической модели извлечения фактов из многоязычной текстовой информации и ее адаптации для текстов английского языка (подраздел 2.2), был получен алгоритм перефразирования факта побуждения к действию в предложениях английского языка. Алгоритм представляет собой использование регулярных выражений грамматических конструкций представления факта побуждения к действию, в которых в качестве алфавита используются метки POS-тегинга. Выбор в качестве исследования алгоритма поиска факта в предложениях побуждения к действию связан: (1) с существующим многообразием синтаксических цепочек данного типа предложений; (2) с общим направлением практической реализации разработанных моделей и методов на криминально окрашенных текста, часто использующих различные способы побуждения к действию.

Наиболее частыми грамматическими конструкциями английского языка, описывающими побуждение к действию, являются: повелительное наклонение, герундийное предложение, предложение с модальным глаголом, использование пассивного залога.

Для анализа синтаксической структуры предложения в работе использовался парсер, представляющий процесс сопоставления линейной последовательности лексем естественного языка с его формальной грамматикой, в результате которого сформировано синтаксическое дерево разбора. Использовался UD парсер английского языка [129], базирующийся на зависимостях, представляющих межъязыковые соответствия наиболее известных пользователю понятий и существующих стандартов разметки.

Полученные формальные схемы грамматических конструкций представляют собой регулярные выражения, использующие POS-тегинговую разметку в качестве алфавита. Полученная формальная схема грамматической модели, использующей модальные глаголы, будет выглядеть следующим образом:

$$\text{TO-VB-[JJ*]-NN*-NN|NNS1-MD-VB-[JJ*]-NN|NNS2}, \qquad (2.37)$$

где *MD* = {should, have to, need to, must, may, might, can, could} модальный глагол, использующийся для выражения повелительного наклонения,

VB — основной глагол в первой форме,

$NN|NNS^2$ — дополнение,

$NN|NNS^1$ = {User, Customer, Operator, Worker, Employer, Manipulator, Handler, Manager} — субъект действия такого предложения,

*NN** — дополнение инфинитива цели,

TO-VB — инфинитив цели.

Примеры предложений, отвечающих данной схеме приведены в таблице 2.2.

Таблица 2.2 – Примеры предложений, разбираемых по формальной модели регулярного выражения (2.37)

TO-VB	NN*	NN|NNS[1]	MD	VB	NN|NNS[2]
to $V_{purpose}$	$object_{purpose}$	Sub	Modal	V1	object
To add 2-pin connector contact center user should update figures and text from 95mm to 25mm					
To add	2 pin con tact at center of connector	user	should	update	figures and text 95mm to 25mm
To reproduce (no) part of this material user have to give the written permission of the copyright owner					
to reproduce	(no) part of this material	user	have to	give	he written permission of the copyright owner
To use the power cable user must switch on power and fan module					
to use	the power cable	user	must	switch on	power and fan module

Формальная схема грамматической модели повелительного наклонения будет выглядеть следующим образом:

$$(V1\ obj\ V_{purpose}\ object_{purpose}\ !), \qquad (2.38)$$

где *V1*— глагол в первой форме, занимающий первое место в предложении,

obj — второстепенный член предложения, представляющий собой объект, на который направленно или с которым связано действие,

$V_{purpose}$ — глагол первой формы, обозначающий инфинитив цели,

$object_{purpose}$ — дополнение инфинитива цели.

Данной схеме будут соответствовать предложения, представленные в таблице 2.3.

Таблица 2.3 – Примеры предложений, разбираемых по формальной модели регулярного выражения (2.38)

VV	NN|NNS[1]	TO-VV	NN|NNS[2]
Update	figures and text from 95mm to 25mm	to add	2-pin connector contact center
Give	the written permission of the copyright owner	to reproduce	(no) part of this material
Switch on	power and fan module	to use	the power cable

Кроме приведенных выше способов побуждения к действию определенная степень побуждения в английском языке может быть выражена активной и пассивной формой герундийного предложения.

Формальная схема предложения, использующего герундий, представлена следующим регулярным выражением:

$$\text{VBG-[JJ*]- NN|NNS}^1\text{-VB- NN|NNS}^2, \quad (2.39)$$

где *VBG*— инфинитив глагола,
NN|NNS^1 — дополнение,
VB —смысловой глагол первой формы (в некоторых случаях с окончанием s),
NN|NNS^2 — дополнение инфинитива цели.
Данной схеме будут соответствовать предложения, показанные в таблице 2.4.

Таблица 2.4 – Примеры предложений, разбираемых по формальной модели (2.39)

VBG	NN|NNS^1	VB	NN|NNS^2
$G_{ing}(V)$	object	$V_{1purpose}$	$object_{purpose}$
Updating figures and text 95mm to 25mm add 2 pin contact at center of connector			
updating	figures and text 95mm to 25mm	add	2 pin contact at center of connector
Giving the written permission of the copyright owner reproduce (no) part of this material			
giving	the written permission of the copyright owner	reproduce	(no) part of this material
The switching-on power and fan module use the power cable			
the switching-on	power and fan module	use	the power cable

Формальная схема предложения, побуждающего к действию в пассивном залоге, будет выглядеть следующим образом:

$$\textit{[JJ*]- NN|NNS}^1\textit{-VB-VBD|VBN-RP-VBG-[JJ*]- NN|NNS}^2, \quad (2.40)$$

где *NN|NNS*^{1} — дополнение инфинитива цели,
VB-VBD|VBN-RP — грамматическая структура, включающая форму вспомогательного глагола *to Be (am, is, are, was, were, been)*, глагола в третьей неправильной форме или глагола с окончанием *–ed*, предлога, формирующего творительный падеж *by* или *with*,
VBG-[JJ]- NN|NNS*^{2} — гериндийная грамматическая структура (включает глагол, с окончанием ing и приложения), с использованием динамических и производных глаголов.

Данной схеме будут соответствовать предложения, представленные в таблице 2.5.

Таблица 2.5 – Примеры предложений, разбираемых по формальной модели (2.40)

NN|NNS[1]	VB-VBD|VBN	RP-VBG	NN|NNS[2]
2 pin contact at center of connector is added by updating figures and text 95mm to 25mm			
2 pin contact at center of connector	is added	with updating	figures and text 95mm to 25mm
No part of this material may be reproduced in any form without giving the written permission of the copyright owner			
no part of this material	may be reproduced	without giving	the written permission of the copyright owner
The power cable is used by the switching-on power and fan module			
the power cable	is used	with the switching-on	power and fan module

Разработанный алгоритм представления регулярными выражениями грамматических конструкций побуждения к действию в английском языке реализован в приложении, пример работы которого показан на рисунке 2.4.

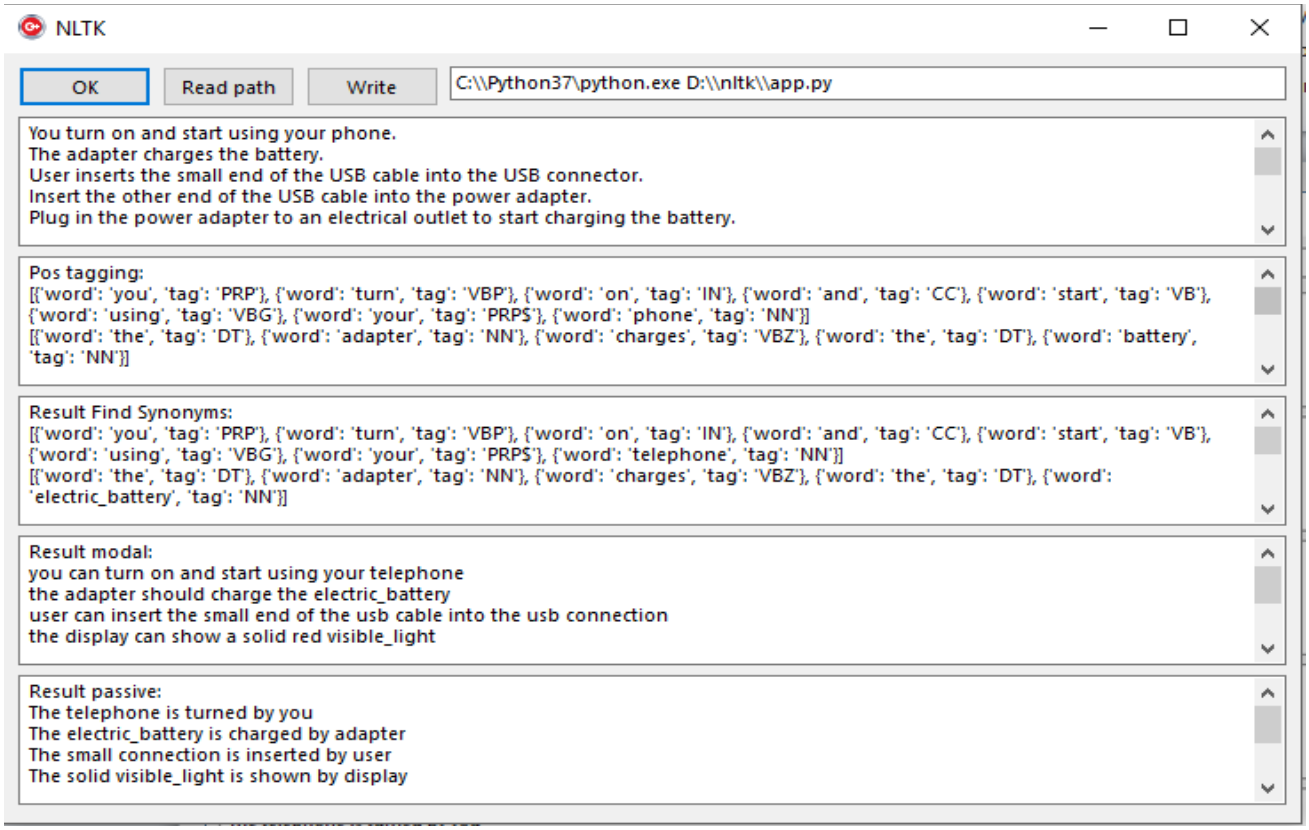


Рисунок 2.4 — Пример работы программного приложения, перефразирующий английские предложения побуждения к действию.

На рис. 2.4 показана работа программного приложения, перефразирующего,

например, фразу “*The adapter charges the battery*”, соответствующую регулярному выражению:

DT*-(NN*|PP)-(VV|VB|VBZ|VBP|VH|VHZ|VVP|VVZ)-[JJ/JJR/JJS]-DT*-[NN/NNS],

во фразы:

– “*The adapter should charge the electric battery* ”, соответствующую регулярному выражению:

DT*-(NN*|PP)-MD-VB-[JJ*]-NN|NNS2;

– “*The electric battery is charged by the adapter*”, соответствующую регулярному выражению:

[JJ*]- NN|NNS1-VB-VBD|VBN-RP-VBG-[JJ*]- NN|NNS2

Дополнительно, в приложении, Предикат и участники (Субъект и Объект) действия исследовались на наличие синонимов, с помощью инструментов WordNet.

Выводы по второму разделу

1. Обоснован выбор математического аппарата алгебры конечных предикатов для моделирования процессов интеллектуальной обработки многоязычной текстовой информации. Рассмотрены основы инструментария алгебры предикатов и предикатных операций применительно к его использованию при формализации конструкций естественных языков, идентифицирующих отношения между участниками действия в предложении.
2. Разработана логико-лингвистическая модель Open IE, описывающая семантические функции партиципантов предложения через отношения грамматических и семантических характеристик слов предложений заданного языка. Использование модели позволяет извлекать факты из многоязычной текстовой информации.
3. Разработанная логико-лингвистическая модель адаптирована для текстов английского языка. При имплементации модели для английского языка использовались такие грамматические и семантические характеристик слов предложения как: наличие предлога после глагола, притяжательный падеж существительного или местоимения, месторазмещение существительного во фразе, наличие любой формы глагола «to be» и основная форма глагола в анализируемой фразе.
4. Показана возможность адаптации модели для автоматической генерации структурированной машиночитаемой информации из текстов казахского языка. При моделировании казахского языка вводились предикатные переменные, характеризующие такие грамматические и семантические характеристики как место анализируемого слова во фразе, наличие или отсутствие вспомогательного глагола во фразе, грамматические падежи и склонения существительных и другие.

5. На базе разработанной математической модели извлечения фактов из многоязычной текстовой информации, и ее адаптации для текстов английского языка, получен алгоритм перефразирования факта побуждения к действию в предложениях английского языка. Алгоритм представляет собой использование регулярных выражений грамматических конструкций представления факта побуждения к действию, в которых в качестве алфавита используются метки POS-тегинга

3. РАЗРАБОТКА МЕТОДОВ И АЛГОРИТМОВ МОРФОЛОГИЧЕСКОГО И СЕМАНТИЧЕСКОГО АНАЛИЗА МНОГОЯЗЫЧНЫХ ТЕКСТОВ, НА БАЗЕ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ

3.1. Метод семантической разметки текстов казахского языка.

Основным критерием оценки текстового корпуса, помимо его репрезентативности и размера, является используемая система разметки и правильность кодирования метаданных корпуса. Разметка представляет собой добавление в текстовый корпус некоторой дополнительной лингвистической информации. Эта информация может быть морфологическая (POS-tagging), синтаксическая (treebanks), семантическая (semantic tagging), просодическая (pro-sodic annotation), метки когерентности дискурса, стилистические метки и т.д. [130].

Используемая структурная разметка документов корпуса, включаемая в заголовок каждого файла, показывает (i) начало и конец предложения; (ii) заголовок документа; (iii) автора текста; (iv) сайт-источник документа; (v) выделяет абзацы и т.д.

Разработанный корпус казахского языка использует следующее макро-тегирование:

i) заголовок текста выделяется тегом:
<head type=main> заголовок</head>

ii) если в тексте есть подзаголовки, они выделяются тегами:
<head type=h1> заголовок</head>

iii) дата публикации выделяется:
<date> </date>

iv) сайт информационного Веб-агенства, откуда взят текст, выделяется тегом:
<site> </site>

v) если есть автор, то он отмечается тегом:
<author> автор < /author>

Существующие сегодня способы добавления лингвистической разметки в текстовый корпус по степени автоматизации можно разделить на три группы:

a) ручная аннотация;
b) автоматическая аннотация;
c) автоматическая аннотация с последующим интеллектуальным редактированием

Использование метода ручной разметки повышает ее точность, в то же время требует больших трудозатрат и не позволяет размечать корпуса большого объема.

Использование автоматической разметки увеличивает количество ошибок, уменьшает точность, но позволяет достаточно быстро осуществлять разметку больших корпусов. В тоже время, полностью автоматическое аннотирование может быть допустимым, если показатель ошибки и многозначности является достаточно низким, не более *n* процента, где *n* достаточно маленькое число, зависящее от конкретного приложения.

Кроме того, при автоматическом аннотировании достаточно сложно создать алгоритмы семантической, стилистической и некоторых других видов разметки.

В связи с приведенными причинами, в нашем исследовании часть информации, в частности, POS-tagging может быть добавлена в корпус автоматически, а семантическая информация добавляется вручную. В итоге, процент ошибки и многозначности аннотирования разрабатываемых корпусов не должен превышать от 5% до 10% в зависимости от языка аннотированного текста.

Для первичной разметки созданного корпуса казахского языка, на базе которого, осуществляется обучение, был использован алгоритм, показанный на рисунке 2.1.

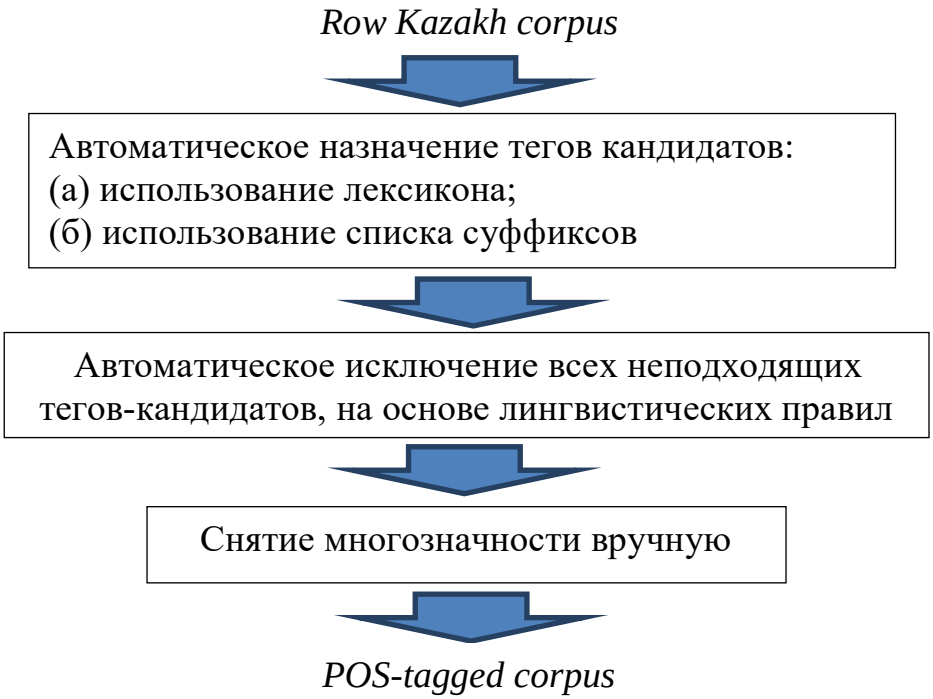


Рис. 3.1. Этапы алгоритма автоматизированного POS-тегирования корпуса казахского языка, на базе которого, осуществляется машинное обучение.

При выборе меток семантического и морфологического тегирования текстов корпуса или множество категорий слов, которые будут применяться к токенам (tagset) учитывались следующие критерии:

1) краткость (conciseness) меток – короткие метки более удобны, чем более подробные и, соответственно, длинные;

2) понятность (perspicuity)– метки должны быть легко интерпретируемы;

3) анализируемость (analysability) – для того, чтобы легко обрабатываться алгоритмами и пониматься человеком, метки должны легко декомпозироваться на логические части.

Логически разработанный *tagset* представляет собой дерево, на верхнем уровне которого находятся части речи, а на нижних уровнях иерархии расположены атрибуты данных частей речи.

От правильности разработанных критериев и самого *Tagset*-а зависит скорость и точность автоматической обработки, и возможность последующего

использования размеченного корпуса.

Основываясь на вышеперечисленных критериях, были выбраны следующие метки, связанные с семантическим аннотированием:

<*Sub*> — Субъект действия триплета факта,

<*Obj*> — Объект действия триплета факта,

<*Pred*> — Предикат (или название действия) триплета факта (см. раздел 2).

Разработанный алгоритм семантической разметки, определяющий триплеты фактов в текстах казахского языка, базируется на использовании списка суффиксов и лингвистических правил. Используется горизонтальный формат разметки.

1. Если:

a. первое, второе или третье слово в предложении, которое имеет суффикс множественного числа из списка [*тар, тер*, *дар*, *дер*, *лар*, *лер*], перед которым не стоят суффиксы:

- родительного падежа из списка [*ның*, *нің*, *дың*, *дің*, *тың*, *тің*],
- направительно-дательный падежа [*ға*, *ге*, *қа*, *ке*, *а*, *е*, *на*, *не*],
- винительного падежа [*ны*, *н*, *ні*, *ды*, *ді*, *ты*, *ті*],
- местного падежа [*да*, *де*, *нда*, *нде*, *та*, *те*]
-исходного падежа [*дан*, *ден*, *тан*, *тен*, *нан*, *нен*]
-творительного падежа [*мен*, *менен*, *бен*, *бенен*, *пен*, *пенен*]

то к слову добавляется метка *_Sub*:

2. Если такого слова в предложении нет, то ищется первое слово в предложении, у которого

a. есть суффикс словообразования существительного из списка:

– именного образования существительного [*ғай*, *гей*, *гер*, *ғи*, *ғой*, *дас*, *дес*, *дік*, *дық*, *кер*, *кес*, *қай*, *қар*, *қи*, *қой*, *қор*, *лас*, *лес*, *лік*, *лық*, *ман*, *паз*, *пана*, *сақ*, *тас*, *тес*, *тік*, *тық*, *хана*, *ша*, *шақ*, *ше*, *шек*, *ші*, *шік*, *шы*, *шық*];

– отглагольного образования существительного [*ақ, ба, бе, ғақ, ғаш, гек , гі, ғіш, ғы, ғыш, дақ, дек, ек, ік, ім, іс, іш, к, кі, кіш, қ, қаш, қы, қыш, лақ, лек, м, ма, мақ, ме, мек, па, пақ, пе, пек, с, тақ, тек, уік, уық, ш, ық, ым, ыс, ыш]*;

– сложного аффикса образования существительного [*герлік, гіштік, ғыштық, дастық, дестік, ділік, дылық, қорлық, ластық, лестік, лілік, лылық, паздық, сақтық, сіздік, сыздық, тастық, тестік, тілік, тылық, шақтық, шілдік, шілік, шылдық, шылық*];

–экспрессивной оценки [*жан, ке, қан, сымақ, тай, ш, ша, шақ, ше, шік, шық*].

b. после которого не стоит падежный суффикс:

- родительного падежа из списка [*ның, нің, дың, дін, тың, тің*],
- направительно-дательный падежа [*ға, ге, қа, ке, а, е, на, не*],
- винительного падежа [*ны, н, ні, ды, ді, ты, ті*],

- местного падежа [*да, де, нда, нде, та, те*]
-исходного падежа [*дан, ден, тан, тен, нан, нен*]
-творительного падежа [*мен, менен, бен, бенен, пен, пенен*]
к найденному слову добавляется метка *_Sub***:**

3. Ищется слово, которое имеет суффикс
- направительно-дательный падежа [*ға, ге, қа, ке, а, е, на, не*],
- винительного падежа [*ны, н, ні, ды, ді, ты, ті*],
после которого может стоять окончание множественности [*тар, тер, дар, дер, лар, лер*],
к такому слову добавляется метка *_Obj***:**

4. Проверяется последнее слово предложения:

a. Если в предложении есть слово, начинающееся с *тұр, отыр, жатыр, жүр,* после которых стоит суффикс [*біз, бін, быз, бын, ды, міз, мін, мыз, мын, піз, пің, пыз, пын, сіз, сіздер, сіңдер, сің, сыз, сыздар, сың, сыңдар, ті, ты*], и последнее слово предложения, с суффиксами [п, ып, іп] , тогда к этому слову добавляется *_Pred*

b. Если последнее слово предложения имеет суффикс будущего предположительного времени [*ар, ер, ыр, ір*] или суффикс деепричастия [*а, е, й, и*], после которого стоит личный предикативный суффикс [*біз, бін, быз, бын, ды, міз, мін, мыз, мын, піз, пін, пыз, пын, сіз, сіздер, сіңдер, сің, сыз, сыздар, сың, сыңдар, ті, ты*]

Тогда к последнему слову добавляется *_Pred*

c. Если в предложении есть вспомогательный глагол *еді, е,* после которого может стоять личный предикативный суффикс [*біз, бін, быз, бын, ды, міз, мін, мыз, мын, піз, пің, пыз, пын, сіз, сіздер, сіндер, сің, сыз, сыздар, сың, сыңдар, ті, ты*], а последнее слово предложения имеет суффикс будущего предположительного времени [*ар, ер, ыр, ір*]

Тогда к последнему слову добавляется *_Pred*

d. Если в предложении есть вспомогательный глагол *еді*, *екен*, после которого может стоять личный предикативный суффикс [*біз, бін, быз, бын, ды, міз, мін, мыз, мын, піз, пің, пыз, пын, сіз, сіздер, сіңдер, сің, сыз, сыздар, сың, сыңдар, ті, ты*], а последнее слово предложения заканчивается на один из возможных аффиксов
–[*ді, дік, дік, дім, дің, ды, дык, дық, дым, дың, қ, ті, тік, тім, тің, ты, тық, тым, тың***];**
– [*а, ай, ал, ан, ар, арыс, ға, ғал, ғар, ғе , ге , гер , гі, гіз, гіздір, гіле, гір , гім, ғы, ғыз, ғыздыр, ғызыл, ғыла, ғыр, ғыт, да, дан , дар, дас, дастыр, де, ден, дендір, дес, діг, дік , дір, діргіз, дық, дыр, дырғыз, дырыл, е, ей, ел, ен, ер, й, іг, іғ, ік , ікіс, іл , іла, ілде, ілу, імсіре, ін, індір, ініс, іну, іңкіре, ір, ірде, іре, ірей, іріс, ірке, іркен , ірқе,*

іс, ісу , іт, ке , кер , кіз, кіле, кір , қа, қал, қан, қар, қе, қур, қыз, қыла, қыла, қыр, л, ла , лан, ландыр, лас, ластыр, лат, ле, лен, лендір, лес, лестір, лет, ліг, лік, лікіс, лін, ліс, лқа, лу, лығ, лық, лын, лыс, мала, меле, мсіре, мсыра, н, ні, ніл, ніс, ныл, ныс, ңгіре, ңғыра, ңкіре, ңқыра, ңра, ңре, р, ра, ре, с, са, сан, се, сен, сет, сетіл, сі, сін, сіре, стір, стыр, сы, сын, сыра, т, та, тан, тандыр, тас, те, тен, тендір, тес, тік, ттыр, тығ, тығс, тығыс, тық, тыр, тырыл, ура, ші, шы, ығ , ығыс, ық, ықыс, ыл, ыла, ылда, ылу, ылыс, ымсыра, ын, ындыр, ыну, ыныс, ыр, ыра, ырай, ырқа, ырқан, ырла, ыс, ысу, ыт];

— [*азы, ақта, ал, ала, аңғыра, аура, бала, бе, беле, би, бі, бы, дала, ди, ді, ды, екте, ел, еңгіре, еуре, жи, жіре, жыра, зы, і, ін, ірей, іс, іт, қи, лі, лы, ма, мала, меле, ми, мсіре, мсыра, ңра, ңре, пала, пеле, пи, пі, пы, ра, ре, си, сіре, сый, сыра, т, ти, ті, ты, усіре, усыра, ши, ші, шы, ы, ын, ыра, ырай, ыс, ыт*],

Тогда к последнему слову добавляется *_Pred*

e. Если последнее слово предложения имеет суффикс [*п, ып, ін*]; после которого стоит личный предикативный суффикс [*біз, бін, быз, бын, ды, міз, мін, мыз, мын, піз, пің, пыз, пын, сіз, сіздер, сіңдер, сің, сыз, сыздар, сың, сыңдар, ті, ты*], в слове так же могут присутствовать суффиксы [*мыс, міс*];

Тогда к последнему слову добавляется *_Pred*

f. Если последнее слово предложения имеет суффикс:

– глагола [*ді, дік, дің, дім, дің, ды, дық, дым, дың, қ, ті, тік, тім, тің, ты, тык, тық, тым, тың*];

– [*а, ай, ал, ан, ар, арыс, ға, ғал, ғар, ғе , ге , гер , гі, гіз, гіздір, гіле, гір , гіт, ғы, ғыз, ғыздыр, ғызыл, ғыла, ғыр, ғыт, да, дан , дар, дас, дастыр, де, ден, дендір, дес, діг, дік , дір, діргіз, дық, дыр, дырғыз, дырыл, е, ей, ел, ен, ер, й, іг, ің, ік , ікіс, іл , іла, ілде, ілу, імсіре, ін, індір, ініс, іну, іңкіре, ір, ірде, іре, ірей, іріс, ірке, іркен , ірқе, іс, ісу , іт, ке , кер , кіз, кіле, кір , қа, қал, қан, қар, қе, қур, қыз, қыла, қыла, қыр, л, ла , лан, ландыр, лас, ластыр, лат, ле, лен, лендір, лес, лестір, лет, ліг, лік, лікіс, лін, ліс, лқа, лу, лығ, лық, лын, лыс, мала, меле, мсіре, мсыра, н, ні, ніл, ніс, ныл, ныс, ңгіре, ңғыра, ңкіре, ңқыра, ңра, ңре, р, ра, ре, с, са, сан, се, сен, сет, сетіл, сі, сін, сіре, стір, стыр, сы, сын, сыра, т, та, тан, тандыр, тас, те, тен, тендір, тес, тік, ттыр, тығ, тығыс, тық, тыр, тырыл, ура, ші, шы, ығ , ығыс, ық, ықыс, ыл, ыла, ылда, ылу, ылыс, ымсыра, ын, ындыр, ыну, ыныс, ыр, ыра, ырай, ырқа, ырқан, ырла, ыс, ысу, ыт*];

– [*ған, ген, қан, кен, қон, ға, ге, қа, ке*], [*атын, етін, йтын, йтін*].

После которых может стоять личное притяжательное окончания некоторых глагольных форм [*дар, йік, йін, йық, йын, іздар, к, қ, м, ндар, ң, ңдер, ңіз, ңіздер, ңыз, сіздер, сің, сіңдер, сыздар, сың, сыңдар, ыздар*],

Тогда к последнему слову добавляется *_Pred*

g. Если последнее слово предложения имеет один из суффиксов:

–[*ді, дік, дің, дім, дің, ды, дык, дық, дым, дың, қ, ті, тік, тім, тің, ты, тык, тық, тым, тың*];

– [*а, ай, ал, ан, ар, арыс, ға, ғал, ғар, ге , ге , гер , гі, гіз, гіздір, гіле, гір , гім, ғы, ғыз, ғыздыр, ғызыл, ғыла, ғыр, ғыт, да, дан , дар, дас, дастыр, де, ден, дендір, дес, діг, дік , дір, діргіз, дық, дыр, дырғыз, дырыл, е, ей, ел, ен, ер, й, іг, іғ, ік , ікіс, іл , іла, ілде, ілу, імсіре, ін, індір, ініс, іну, іңкіре, ір, ірде, іре, ірей, іріс, ірке, іркен , ірқе, іс, ісу , ім, ке , кер , кіз, кіле, кір , қа, қал, қан, қар, қе, қур, қыз, қыла, қыла, қыр, л, ла , лан, ландыр, лас, ластыр, лат, ле, лен, лендір, лес, лестір, лет, ліг, лік, лікіс, лін, ліс, лқа, лу, лығ, лық, лын, лыс, мала, меле, мсіре, мсыра, н, ні, ніл, ніс, ныл, ныс, ңгіре, ңғыра, ңкіре, ңқыра, ңра, ңре, р, ра, ре, с, са, сан, се, сен, сет, сетіл, сі, сін, сіре, стір, стыр, сы, сын, сыра, т, та, тан, тандыр, тас, те, тен, тендір, тес, тік, ттыр, тығ, тығс, тығыс, тық, тыр, тырыл, ура, ші, шы, ығ , ығыс, ық, ықыс, ыл, ыла, ылда, ылу, ылыс, ымсыра, ын, ындыр, ыну, ыныс, ыр, ыра, ырай, ырқа, ырқан, ырла, ыс, ысу, ыт*];

— [*азы, ақта, ал, ала, аңғыра, аура, бала, бе, беле, би, бі, бы, дала, ди, ді, ды, екте, ел, еңгіре, еуре, жи, жіре, жыра, зы, і, ін, ірей, іс, ім, қи, лі, лы, ма, мала, меле, ми, мсіре, мсыра, ңра, ңре, пала, пеле, пи, пі, пы, ра, ре, си, сіре, сый, сыра, т, ти, ті, ты, усіре, усыра, ши, ші, шы, ы, ын, ыра, ырай, ыс, ыт*].

После которого может стоять личное притяжательное окончание [*дар, йік, йін, йық, йын, іздар, к, қ, м, ндар, ң, ңдер, ңіз, ңіздер, ңыз, сіздер, сің, сіңдер, сыздар, сың, сыңдар, ыздар*];

Тогда к последнему слову добавляется *_Pred*

5. Если последнее слово предложения не удовлетворяет ни одному из условий пункта 4, то на эти же условия проверяется предпоследнее слово и затем третье от конца предложения слово.

Кроме определения триплетов факта, для обозначения семантики узкоспециализированной предметной области, связанной с криминальной значимостью текста, используются следующие метки:

<s type=crim> предложение </s> — выделяется предложение, имеющее криминальную окраску;

<v type=crim > глагол </v> — выделяется глагол, имеющий криминальную окраску;

<n type=crim > существительное </n> — выделяется существительное, в имеющее определенное криминальное значение;

* прилагательное и причастие * — выделяется прилагательное или причастие, имеющее некоторую криминальную окраску.

В таком обозначении имена тегов *<v>*, *<n>*, *<a>* определяют грамматическую информацию части речи, тег *<s>* определяет предложение, как синтаксическую единицу. А значения атрибута *type=”crim”* определяет семантическую информацию криминальной окрашенности.

На рисунке 3.2 показан фрагмент семантически размеченного файла корпуса.

<head type=main>·Полицейские·Алматинской·области·изъяли·у·мужчины·более·5·кг·наркотиков·
</head>¶

<date>·06.10.2018·</date>¶

<site>·patrul.kz</site>¶

<s type=crim>Задержание·подозреваемого·<v type=crim>·произведено</v>·
<n type=crim>·полицейскими</n>·в·ночное·время·на·контрольном·посту,·действующем·на·59·км
трассы·«Алматы-·Бишкек».</s>¶

<s type=crim>·В·автомашине·мужчины·специально·<v type=crim>·обученная</v>·на·поиск·<n·
type=crim>·наркотиков</n>·-·служебно-розыскная·собака·по·кличке·Таймас·<
type=crim>·обнаружила</v>·тайник·с·<n type=crim>·наркотиками</n>·-.</s>¶

<s type=crim>·-«<n type=crim>·Тайник·</n>··был·установлен·под·одним·из·сидений·автомашины·
Mercedes.</s>·<s type=crim><n type=crim>·Полицейскими</n>·было<v type=crim>·изъято·</v>
5·килограмм·360·грамм·<n type=crim>·гашиша</n>·-.·</s><s type=crim>·
Задержанный···мужчина·<v>·помещен·</v>·под·-«<n type=crim>·стражу</n>·-в·Жамбылское
районное·подразделение·<n type=crim>·-полиции</n>·-.·</s>¶

Рисунок 3.2 – Фрагмент семантически размеченного файла корпуса

Первичный корпус, имеющий морфологическую и семантическую разметку, используется на этапе обучения методами machine leaning.

3.5. Метод вероятностного POS-тегинга, использующего скрытую Марковскую модель

Разрабатываемый метод машинного обучения для POS- tagging многоязычной текстовой информации базируется на скрытой Марковской модели (HMM – Hidden Markov Model). Термин “скрытая Марковская модель” обозначает модель, имитирующую работу процесса, похожего на Марковский процесс с неизвестными параметрами. Целью модели является определение неизвестных параметров на основе наблюдаемых.

В обычной Марковской модели состояние видимо наблюдателю, поэтому вероятности переходов — единственный используемый параметр. В скрытой Марковской модели видимыми являются лишь переменные, на которые оказывает влияние данное состояние. Каждое состояние имеет вероятностное распределение среди всех возможных выходных значений. Последовательность символов, сгенерированная скрытой Марковской моделью, даёт информацию о последовательности состояний [131].

Таким образом, при применении модели HMM к задаче POS- tagging мы будем говорить не об оценке максимальной вероятности явной цепочки слов, а об оценке максимальной вероятности POS-тегов слова и скрытой цепочки POS-тегов. На вход используемой модели поступает цепочка слов, каждое из которых имеет один или более потенциальный тегов, задача заключается в выборе наиболее вероятной

последовательности тегов, основанной на подсчете вероятностей все возможных последовательностей. Модель *N-gram tagging* определяет наибольшую вероятность данного тега для данного слова, при учете имеющейся разметки *n-1* предыдущих слов.

В разработанном методе вероятностного POS-тегинга, базирующегося на HMM, функция оценки вероятности цепочки тегов предложения зависит от двух вероятностей:

1) условной вероятности последовательности тегов;
2) условной вероятности обозначения слова данным тегом.

Первой рассматриваемой вероятностью является вероятность цепочки или последовательности тегов. В простейшем случае для оценки вероятности перехода от одного тега к другому используют частоты пар тегов (биграмм в training corpus), т.е. вычисляется вероятность того, что два тега стоят последовательно один за другим. Например, для английского языка, за тегом артикля с большей вероятностью будет следовать тег прилагательного или существительного, тогда как следование за ним тега базовой формы глагола является маловероятным. Для казахского языка следование за прилагательным существительного является наиболее вероятным, чем последовательность <*JJ*><*VB*> или <*JJ*>< *Pred* >.

Вероятность аннотирования слова w_i тегом $<t_{ij}>$ будет определяться как условная вероятность конкретного тега при данном предыдущем теге t_{i-1k} слова w_{i-1}:

$$P(t_{ij} \mid t_{i-1k}) \qquad (3.1)$$

Например, для предложения казахского языка "*Түркістан облысында қаруланған үш жігіт жанар-жағар май кассасын тонап кетті*" биграммная вероятность тегирования будет выглядеть как:

$$\begin{aligned} &P(\text{“ <s> } \textit{Түркістан облысында қаруланған үш жігіт жанар-жағар май кассасын тонап кетті } \text{</s>”} \mid \text{“<s> <}JJ_{jat}\text{ > <}NN_{jat}\text{> <}JJ_{at}\text{> < MC>} \\ &\text{<}NN_{at}\text{> <}JJ_{ta}\text{> <}NN_{ta}\text{> <VB> <VV></s>”}) = P(<JJ_{jat}>|<s>) \\ &P(<NN_{jat}>|<JJ_{jat}>)\, P(<JJ_{at}>|<NN_{jat}>)\, P(<MC> \mid <JJ_{at}>) \\ &P(<NN_{at}> |<MC>)\, P(JJ_{ta} |<NN_{at}>)\, P(<NN_{ta}> \mid <JJ_{ta}>) \\ &P(<VB> \mid <NN_{ta}>)\, P(<VV> \mid < VB >)\, P(</s>| <VV>), \end{aligned} \qquad (3.2)$$

где:
JJ_{jat} — прилагательное, местный падеж (жатыс);
NN_{jat} — существительное, местный падеж (жатыс);
JJ_{at} — прилагательное, именительный падеж (атау);
MC — числительное;
NN_{at} — существительное, именительный падеж (атау);
JJ_{ta} — прилагательное, винительный падеж (табыс);
NN_{ta}— существительное, винительный падеж (табыс);
VB — основной глагол;
VV — вспомогательный глагол

В более сложном случае оценку вероятности можно осуществлять с помощь триграмм, когда оценивается вероятность появления тега при заданных двух предыдущих. В таком случае вероятность цепочки слов представляет собой

условную вероятность тега данного слова, при условии известных тегов двух предыдущих слов:

$$P(t_{ij} \mid t_{i\text{-}1k}\ t_{i\text{-}2p}) \qquad (3.3)$$

При этом точность оценивания трехграммной вероятности выше, чем при биграммной вероятности, когда рассматривается только один предыдущий тег, но полнота существенно понижается и необходимо необходим training corpus большего размера.

Необходимость учета вероятности второго вида обусловлена неопределенностью тегов слов цепочки, что обуславливает необходимость оценки вероятности обозначения слова тем или иным тегом из списка. Например, в английском языке слово "*spring*" может быть как существительным, так и прилагательным и глаголом. А в русском языке слово "*мощный*" может быть как прилагательным, связанным с Объектом действия, так и определяемым словом Субъекта действия. В казахском языке слово "*қара*" может быть как прилагательным, так и глаголом, а слово "*алма* " как глаголом, так и существительным.

Такую условную вероятность конкретного слова w_i, при данной ассоциации его с конкретным тегом<t_{ij}> можно записать как:

$$P(w_i \mid t_{ij}) \qquad (3.4)$$

Данную условную вероятность можно определить как Unigram tagging. Для получения *Unigram tagging* так же необходим достаточно большой обучающий корпус (*train corpus*).

Используем статистический алгоритм: выбор тега для каждого токена определяется наибольшей вероятностью данного тега для данного слова в *train корпусе*. Т.е. $P(w_i \mid t_{ij})$ возможно оценить через относительную частоту, с которой любое конкретное слово ассоциируется с данным тегом.

Например, если слово "*мүс"* или "*Шық"* обозначается тегом *VB* или *Pred*, это значит, что данные слова в обучающем корпусе чаще выступают глаголом, чем любой другой частью речи. Процесс обучения включает просмотр всех тегов каждого слова и сохранение наиболее вероятного тега для любого слова, существующего в обучающем корпусе. Так как Unigram Tagger лучше работает на текстах, тематика и лексика которых в обучающем корпусе и тестовом корпусе близки, мы используем корпуса одной предметной области — новостные тексты, имеющие криминальную семантическую составляющую.

Например, на вход поступает цепочка из семи слов $w_0\ w_1\ w_2\ w_3\ w_4\ w_5\ w_6$. В результате предыдущего этапа, каждому слову приписывается один или более тегов. Если у слова их несколько, то они ранжируются от более вероятного к менее вероятному. Слова w_0, w_3 и w_5 однозначно ассоциированы с тегами t_0, t_3 и t_5 , эти слова разделены словами w_1 w_2, w_3 и w_6, которые имеют несколько потенциальных тегов (рис. 3.3).

Задачей является выбрать один наиболее подходящий тег.

w_0 w_1 w_2 w_3 w_4 w_5 w_6

t_0	t_{10}	t_{20}	t_3	t_{40}	t_5	t_{60}
	t_{11}	t_{21}		t_{41}		t_{61}
				t_{42}		

Рисунок 3.3 — Пример цепочки слов,
с потенциально возможным POS- тегингом.

Используя разработанный метод необходимо на выходе получить наиболее вероятную последовательность тегов, которая и будет определена как правильная последовательность. На рисунке 3.4 схематично показано использование функции двух вероятностей для определения того, что заданная цепочка слов w_0 w_1 w_2 w_3 w_4 w_5 w_6, будет тегирована последовательностью тегов <t_0> <t_{10}> <t_{20}> <t_3> <t_{40}> <t_5> <t_{60}>.

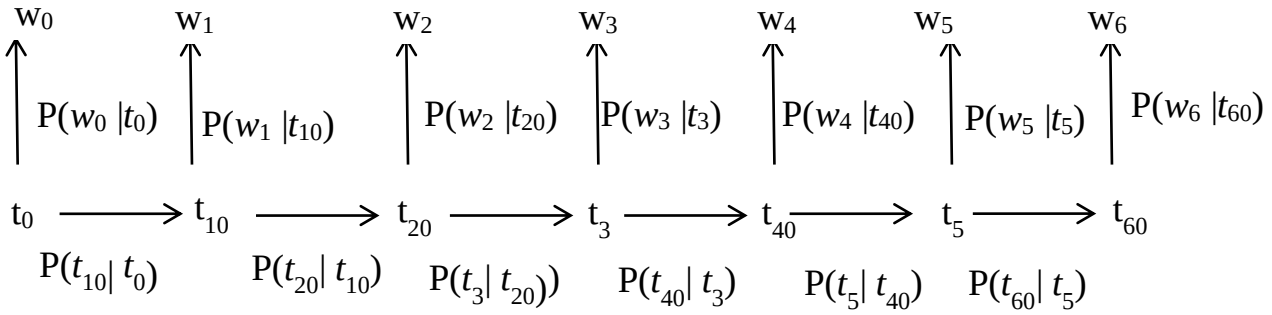


Рисунок 3.4 — Схематическое использование метода
вероятностного POS-тегинга

Базируясь на том, что вероятность тегирования последовательности слов определенными тегами, определяется: (1) условной вероятность последовательности тегов; и (2) условной вероятностью, обозначения конкретного слова данным тегом, мы определяем вероятность того, что цепочка слов w_0 w_1 w_2 w_3 w_4 w_5 w_6 будет последовательно тегирована цепочкой тегов <t_0> <t_{10}> <t_{20}> <t_3> <t_{40}> <t_5> <t_{60}> через произведение условных вероятностей:

$$P(W_T) = P(w_0 \mid t_0)\ P(t_{10} \mid t_0)\ P(w_1 \mid t_{10})\ P(t_{20} \mid t_{10})\ x\ P(w_2 \mid t_{20})$$
$$P(t_3 \mid t_{20})\ P(w_3 \mid t_3)\ P(t_{40} \mid t_3)\ P(w_4 \mid t_{40})\ P(t_5 \mid t_{40})\ P(w_5 \mid t_5)\ P(t_{60} \mid t_5)\ P(w_6 \mid t_{60}) \quad (3.5)$$

Такие же выражения должны быть построены для каждой возможности последовательности тегирования. Последовательность с наибольшей вероятностной оценкой и будет определена как окончательная.

Например, в результате обучения на предварительно размеченном корпусе англоязычных текстов получены *Unigram tagger* и *BigramTagger*, которые схематически показаны на рис. 3.5. Метки соответствуют текущему CLAWS5 tagset[4]. Где *PNP* – личное местоимение; *VVD* – вторая форма глагола; *VVN* – третья форма глагола (past participle form); *TO* – частица инфинитива; *PRP* – предлог; *NN1*

[4] http://ucrel.lancs.ac.uk/claws5tags.html

– существительное единственного числа; *AJ0* – прилагательное; *VVB* – базовая форма глагола; *AV0* – наречие

Unigram tagger

токен	*метка*	$P(w_i \mid t_{ij})$
she	<PNP>	1.0
promised	<VVD>	0.7
	<VVN>	0.3
to	<TO>	0.45
	<PRP>	0.55
back	<NN1>	0.25
	<AV0>	0.08
	<AJ0>	0.25
	<VVB>	0.42

BigramTagger

цепочка меток	$P(t_{ij} \mid t_{i\text{-}1k})$
< PNP >< VVD >	0.3
< PNP >< VVN >	0.15
< VVD >< PRP >	0.4
< VVD >< TO >	0.08
< PRP >< NN1>	0.4
< PRP >< AV0 >	0.07
< PRP >< AJ0 >	0.3
< TO >< VVB >	0.9

Рисунок 3.5 — Схематический фрагмент *Unigram* и *Bigram taggers*, обученных на корпусе английских текстов

Используя разработанный метод вероятностного POS-тегинга, базирующегося на HMM, определяем возможные варианты разметки цепочки слов.

P(“*she* _ PNP *promised*_VVD *to*_PRP *back*_NN1”) = P (*she* | PNP) *
* P(*promised*| VVD)* P (VVD | PNP)*
* P(*to* | PRP) * P(PRP | VVD)*
* P (*back* | NN1)* P (NN1| PRP) =
= 1.0 *0.7*0.3 *0.55 *0.4*0.25*0.4= 0,00462 (3.6)

P(“*she* _ PNP *promised*_VVD *to*_TO *back*_VVB”) = P (*she* | PNP) *
* P(*promised*| VVD)* P (VVD | PNP)*
* P(*to* | TO) * P(TO | VVD)*
* P (*back* | VVB)* P (VVB | TO) =
= 1.0 *0.7*0.3 *0.45* 0.08 * 0.42 * 0.9=
= 0,00285768 (3.7)

P(“*she* _ PNP *promised*_VVD *to*_ PRP *back*_ AJ0”) = P (*she* | PNP) *
* P(*promised*| VVD)* P (VVD | PNP)*
* P(*to* | PRP) * P(PRP | VVD)*
* P (*back* | AJ0)* P (AJ0| PRP) =
= 1.0 *0.7*0.3 *0.55 *0.4*0.25*0.3= 0,3465 (3.8)

Рассмотрев все возможные комбинации тегирования цепочки слов, можно увидеть, что разметка “*she* _ PNP *promised*_VVD *to*_PRP *back*_NN1” (3.5) имеет наибольшую оценочную вероятность, что соответствует интуитивно ожидаемой разметке.

Предложенный метод, очевидно, имеет высокую точность, но в тоже время может иметь низкую полноту, что связано с возможным отсутствием биграммной

или трехграммной цепочки в обучающем корпусе (train corpus). Причем чем длиннее цепочки слов, используемые для оценки вероятности скрытой Марковской моделью (например, триграммы или четырехграммы), тем ниже полнота используемого метода.

Для повышения полноты предлагается использовать алгоритм комбинированного аннотирования [65], который заключается в использовании вероятностного биграммного POS-тегинга, использующего скрытую Марковскую модель, с последующим использованием униграммного тегера, для токенов, не размеченных на первом этапе. На следующем этапе для токенов, которые не были определены в обученном униграммном тагере, используем, тегер по умолчанию, который размечает все оставшиеся токены наиболее вероятной меткой, определенной на обученном униграммном тегере.

3.3. Метод определения семантической близости многоязычных текстовых документов, базирующийся на Vector Space Model

В общем случае Vector Space Model (VSM) является частью семантической технологии и представляет собой математическую модель представления текстов, в которой каждому документу сопоставлен вектор, отражающий его смысл [132]. Базируясь на идеи представления каждого документа в коллекции текстов как точки векторного пространств [133], мы показываем, что точки, расположенные близко друг к другу в этом пространстве, соответствуют семантически близким документам, а точки, которые расположены далеко друг от друга, соответствуют сильно отличающимся по смыслу документам. Исходя из данного подхода, мы определяем близкие по смыслу документы. В основе использования VSM лежат две когнитивные гипотезы. Первая гипотеза статистической семантики (Statistical semantics hypothesis) утверждает, что статистические шаблоны использования слов в естественном языке могут быть использованы для выяснения того, что люди имеют в виду. Говоря иными словами, человеческий интеллект может понимать слова в зависимости от их окружения [134]. Вторая гипотеза, сформулированная Дж. Солтоном для информационного поиска [135, 136], базируется на представлении текста в виде «мешка слов» (bag of words) и предполагает, что частота слов в документе чаще всего определяет релевантность документов к запросу.

С точки зрения математики «мешок слов» — это мультимножество, представляющее собой модифицированное понятие множества, допускающее, в отличие от множества, включение одного и того же элемента в совокупность по несколько раз [137]. Например, $\{a, a, b, c, c, c\}$ — это мультимножество, содержащее a, b, and c. Порядок элементов, как и во множествах, в мультимножествах не имеет значения. Такие мультимножества, как $\{a, a, b, c, c, c\}$ и $\{c, a, c, b, a, c\}$ эквивалентны. Мы можем представить мультимножество $S = \{a, a, b, c, c, c\}$ вектором $x = \langle 2, 1, 3 \rangle$, оговаривая, что первый элемент вектора x – это частота элемента a в мультимножестве S, второй элемент – это частота элемента b в S и третье значение вектора — это частота с. Множество

мультимножеств, подобных *S*, может быть записано матрицей *X*, в которой каждый столбец $x_{:j}$ соответствует одному из мультимножеств, а каждая строка $x_{i:}$ соответствует уникальному элементу, значение x_{ij} на пересечении строки и столбца – это частота i-th элемента в j-th мультимножестве.

Таким образом, если у нас есть большая коллекция документов, и, следовательно, большое количество векторов документов, то удобно организовать данные вектора в матрицу. Строка вектора в матрице соответствует терминам (т.е. словам) документа, а вектор столбца соответствует документам нашего корпуса. Полученная данным путем матрица будет представлять собой матрицу термин-документ.

Пусть *X* будет матрица термин-документ. Предположим, наш корпус содержит *n* текстов и *m* уникальных терминов. Тогда матрица *X* будет иметь *m* строк (одна строка для каждого уникального термина в словаре коллекции текстов) и *n* столбцов (один столбец для каждого документа). Пусть w_i будет *i*-ый термин в документе и d_j будет *j*-ый документ в коллекции. Тогда *i*-ая строка в матрице *X* – это вектор строки $x_{i:}$ и *j*-ый столбец в матрице *X* – это вектор столбца $x_{:j}$. Вектор строки $x_{i:}$ содержит n элементов, один элемент для каждого документа, и вектор столбца $x_{:j}$ содержит m элементов, один элемент для каждого термина. Структура чисел в строке $x_{i:}$ подобна подписи *i*-ого термина w_i, аналогично структура чисел в столбце $x_{:j}$ является подписью в *j*-ого документа d_j.

Базируясь на ранее приведенной дистрибутивной гипотезе «bag of words», утверждающей, что вектор столбца в матрице термин-документ в некоторой степени отражает аспект значения соответствующего документа, мы можем оценить смысловую близость некоторых документов по частоте слов документов, представленных в векторах.

В результате предварительной лингвистической обработки у нас может быть получен, например, такой упрощенный текстовый массив специализированных текстов узкой направленности. Массив состоит из шести документов и включает восемнадцать токенов.

doc1) қақтығысу, қақтығысу, жол қозғалысы, жол қозғалысы, банк;
doc2) *өлу, қақтығысу, жол қозғалысы*;
doc3) *байланған, жоғалды, атысу, жоғалды, атысу*;
doc4) *тапанша, тапанша, атысу, атысу, атысу, ұрланды, ұрланды, банк*;
doc5) *тапанша, ұрланды, ұрланды, банк*;
doc6) *тапанша, байланған, жоғалды, жоғалды, жоғалды, кинолог*.

Основой для построения матрицы термин-документ может быть как матрица *токен-документ* (рис. 3.5), так и матрица *тип-документ* (рис. 3.6). Матрица *токен-документ* имеет восемнадцать строк по одной для каждого токена и шесть столбцов, по одному для каждого документа. Вектор строки для токенов в данной матрице имеет бинарное значение: единица, если токен появляется в данном документе, и ноль, в противном случае, – если токена в данном документе нет. Типы токенов (или общие токены) обозначают количества токенов с учетом их частотности. Матрица *тип-документ* имеет десять строк для каждого типа и шесть столбцов. Вектор строки для типа токенов представлен целочисленными

значениями частоты появления токенов данного типа в данном документе. Значения вектора типа-токена определяется суммой соответствующих векторов токенов. Например, вектор строки для типа токена "*жоғалды* " таблицы 3.2 определяется суммой трех векторов токенов " *жоғалды* " таблицы 3.1.

Таким образом, следующим шагом после токенизации и нормализации корпуса текста является генерация матрица частот, представляющая обобщенный вид матрицы тип токена – документ, в которой тип токена выступает термином. В данной частотной матрице элемент соответствует событию. Элемент, которым выступает термин, проявляется в определенной ситуации, которую определяет документ, конкретное число раз, т.е. с определенной частотой. В таком случае, X – простая матрица частоты, элемент x_{ij} в матрице X – это частота i-того термина w_i в j-ом документе d_j .

Однако, простая частота встречаемости — это не лучший способ измерения ассоциации между словами. Используя частотный способ определения ассоциации мы не должны учитывать влияние таких слов как "*the*", "*it*", or "*they*" в английском языке; "*емес*", "*әлде*", "*олар*" в казахском; или "*он*", "*в*", "*на*" в русском, которые проявляются довольно часто с любыми словами, и при этом не определяют какой-либо смысл. Вместо того чтобы определять простую частоту всех слов, необходимо учитывать контекст только тех слов, которые наверняка являются информативными для определения тематики текста.

Таблица 3.1. – Матрица токен–документ. Строки представляют токены, а столбцы представляют документы

	Doc1	*Doc2*	*Doc3*	*Doc4*	*Doc5*	*Doc6*
тапанша	0	0	0	1	1	1
тапанша	0	0	0	1	0	0
өлу	0	1	0	0	0	0
қақтығысты	1	1	0	0	0	0
қақтығысты	1	0	0	0	0	0
жол қозғалысы	1	1	0	0	0	0
жол қозғалысы	1	0	0	0	0	0
байланған	0	0	1	0	0	1
жоғалды	0	0	1	0	0	1
жоғалды	0	0	1	0	0	1
жоғалды	0	0	0	0	0	1
перестрелка	0	0	1	1	0	0
атысу	0	0	1	1	0	0
атысу	0	0	0	1	0	0
кинолог	0	0	0	0	0	1
ұрланды	0	0	0	1	1	0
ұрланды	0	0	0	1	1	0
банк	1	0	0	1	1	0

Таблица 3.2.– Матрица тип токена–документ. Строки представляют типы токенов, а столбцы представляют документы

	Doc1	*Doc2*	*Doc3*	*Doc4*	*Doc5*	*Doc6*
тапанша	0	0	0	2	1	2
өлу	0	1	0	0	0	0
қақтығысты	2	1	0	0	0	0
жол қозғалысы	2	1	0	0	0	0
байланған	0	0	1	0	0	1
жоғалды	0	0	2	0	0	3
атысу	0	0	2	3	0	0
кинолог	0	0	0	0	0	1
ұрланды	0	0	0	2	2	0
банк	1	0	0	1	1	0

Согласно теории информации, неожиданные события имеют более высокое информационное значение (или содержание информации), чем ожидаемые события [138]. С точки зрения поиска близких по смыслу текстов и семантической обработки текстов, – чем большую частоту имеет термин, тем он менее информативен. Гипотеза состоит в том, чтобы неожиданные события, если они разделены двумя векторами, больше различают сходство между ними, чем менее неожиданные события. Для того чтобы придать больший вес более информативным терминам и уменьшить значение менее информативных терминов, будем использовать взвешивание элементов матрицы.

Для того чтобы повысить влияние в матрице термин-документ терминов, наиболее часто встречающихся в данном документе и редко во всех документах коллекции, будем использовать меру PMI (Pointwise mutual information) [139]. Формально PMI можно определить как весовую функцию.

Пусть F будет термин -документ матрицей, в которой n_{r} строк и n_{c} столбцов, i-ая строка в матрице F – это вектор строки $f_{\text{i:}}$ и j-ый столбец в матрице F – это вектор столбца $f_{:\text{j}}$. Строка $f_{\text{i:}}$ соответствует термину w_{i} и столбец $f_{:\text{j}}$ соответствует документы d_{j} . Значение элемента f_{ij} – это число раз, когда w_{i} появился в документе d_j.

Пусть X будет матрицей, которая получается вычислением весовой функции PMI к элементам матрицы F. Тогда матрица X будет иметь тоже количество строк и столбцов, как и частотная матрица F. Значения элемента x_{ij} в матрице X определяется следующим образом:

$$p_{ij} = \frac{f_{ij}}{\sum_{i=1}^{n_r}\sum_{j=1}^{n_c} f_{ij}}, \tag{3.9}$$

$$p_{i*} = \frac{\sum_{j=1}^{n_c} f_{ij}}{\sum_{i=1}^{n_r}\sum_{j=1}^{n_c} f_{ij}}, \quad (3.10)$$

$$p_{*j} = \frac{\sum_{i=1}^{n_r} f_{ij}}{\sum_{i=1}^{n_r}\sum_{j=1}^{n_c} f_{ij}}, \quad (3.11)$$

где:

p_{ij} — это оценочная вероятность того, что термин w_i появится в документе d_j, вычисляемая, как частота появления термина в данном документе, нормализованная суммой всех терминов во всех документах коллекции, т.е. размером словаря данной коллекции документов *N*;

p_{i*} —это оценочная вероятность термина w_i, т.е. вероятность появления данного термина в любом документе коллекции, определяемая в матрице *F* как сумма всех частот по строке i, в которой расположен данный термин, нормализованная на общее количество терминов в коллекции *N*;

и p_{*j} — это оценочная вероятность документа d_j, т.е. вероятность появления данного документа с любым запрашиваемым термином, вычисляемая как сумма всех частот по столбцу *j*, нормированная делением на общее количество слов в коллекции.

Для расчета PMI вычисляется логарифм оценочной вероятности p_{ij} (3.9) деленный на произведение двух вероятностей p_{i*} (3.10) и p_{*j} (3.11):

$$pmi_{ij} = \log\left(\frac{p_{ij}}{p_{i*}p_{*j}}\right), \quad (3.12)$$

Используя полученные формулы (3.9) –(3.12) для преобразования нашего примера (табл. 3.2), пересчитав все экземпляры терминов и документов, можно получить ниже следующую промежуточную матрицу.

Таблица 3.3.– Промежуточная матрица вычисления оценочных вероятностей появления термина в документе, термина и документа.

	P (w, d)						
	Doc1	*Doc2*	*Doc3*	*Doc4*	*Doc5*	*Doc6*	P_{i*}(*термин*)
тапанша	0	0	0	0,0625	0,0313	0,0625	0,3125
өлу	0	0,0313	0	0	0	0	0,0625
қақтығысты	0,0625	0,0313	0	0	0	0	0,1875
жол қозғалысы	0,0625	0,0313	0	0	0	0	0,1875

байланған	0	0	0,031	0	0	0,0313	0,125
жоғалды	0	0	0,063	0	0	0,0938	0,3125
атысу	0	0	0,063	0,0938	0	0	0,3125
кинолог	0	0	0	0	0	0,0313	0,0625
ұрланды	0	0	0	0,0625	0,0625	0	0,25
банк	0,0313	0	0	0,0313	0,0313	0	0,1875
p_{*j}(*документ*)	0,1563	0,0938	0,156	0,25	0,125	0,2188	1

Если w_i и d_j являются статистически независимыми, то, согласно определению произведения вероятности независимых событий, $p_{i*}p_{*j} = p_{ij}$. Следовательно, значение логарифма формулы определения pmi_{ij} (3.12) будет равно нулю, т.е. если термин встретился в данном документе совершенно случайно и не существует зависимости между смыслом термина и смыслом документа, то $pmi_{ij} = 0$. Однако, если существует некоторая семантическая взаимосвязь между w_i and d_j, то следует ожидать, что p_{ij} будет больше, чем было бы, если w_i и d_j, были бы семантически независимы. Следовательно, следует искать значения $p_{ij} > p_{i*}p_{*j}$, и, следовательно, pmi_{ij} положительно.

Будем использовать введенную на базе дистрибутивной гипотезы весовую функцию PPMI (Positive Pointwise Mutual Information) [139], которая разработана для того, чтобы добавить больший вес значениям x_{ij}, подтверждающим семантическую зависимость между w_i и c_j и присвоить нулевые значения, если появление w_i в документе d_j не имеет никакого семантического значения и, следовательно, не несет никакой информации. Тем самым, мы решаем проблему существования стоп-слов в тексте или избавляемся от шума.

$$x_{ij} = ppmi_{ij} = \begin{cases} pmi_{ij}, \text{ if } \text{pmi}_{ij} > 0 \\ 0, \text{pmi}_{ij} <= 0 \end{cases}, \qquad (3.13)$$

Базируясь на данном определении, значения PPMI в нашем примере можно трансформировать в виде следующий массив:

Таблица 3.4.– Значения PPMI матрицы термин-документ.

	Doc1	*Doc2*	*Doc3*	*Doc4*	*Doc5*	*Doc6*
тапанша	-	-	-	0	0	0
өлу	-	2,415037	-	-	-	-
қақтығысты	1,093109	0,830075	-	-	-	-
жол қозғалысы	1,093109	0,830075	-	-	-	-
байланған	-	-	0,678072	-	-	0,192645
жоғалды	-	-	0,356144	-	-	0,455679

атысу	-	-	0,356144	0,263034	-	-
кинолог	-	-	-	-	-	1,192645
ұрланды	-	-	-	0	1	-
банк	0,093109	-	-	0	0,415037	-

Пусть w_i и d_j являются статистически зависимыми (имеют максимальную ассоциативность). Тогда при $p_{ij} = p_{i*} = p_{*j}$, уравнение (3.12) преобразуется в $log(1/p_{i*})$ и, следовательно, PMI увеличивается, так как вероятность слова w_i уменьшается. Для того чтобы решить очевидную проблему смещения веса PMI к более редким событиям будем использовать подход сглаживания Лапласа для оценки вероятности p_{ij} , p_{i*}, и p_{*j} [140].

Сглаживание Лапласа заключается в добавлении к необработанным частотам перед вычислением вероятностей некоторого положительного числа. При этом, каждое f_{ij} заменяется на $f_{ij} + k$, где $k > 0$. Чем больше константа k, тем больше эффект сглаживания. Сглаживание Лапласа уменьшает значение pmi_{ij}. Величина этого уменьшения (разницы между pmi_{ij} без сглаживания и со сглаживанием Лапласа) зависит от сырой частоты f_{ij}. Если частота большая, что сдвиг небольшой, а если частота маленькая, то сдвиг большой. Таким образом, сглаживание Лапласа позволяет уменьшить зависимость значений *pmi* от редких событий или, в нашем случае, от редко встречающихся терминов.

Используя сглаживание Лапласа, заполняя частотную таблицу термин-документ 3.2, будем считать, что мы видели каждое слово в документе на два раза чаще (табл. 3.5)

Таблица 3.5.– Значения частотной матрицы термин-документ с использованием сглаживания Лапласа.

	Add-2 Smoothed count (w,d)					
	Doc1	*Doc2*	*Doc3*	*Doc4*	*Doc5*	*Doc6*
тапанша	2	2	2	4	3	4
өлу	2	3	2	2	2	2
қақтығысты	4	3	2	2	2	2
жол қозғалысы	4	3	2	2	2	2
байланған	2	2	3	2	2	3
жоғалды	2	2	4	2	2	5
атысу	2	2	4	5	2	2
кинолог	2	2	2	2	2	3
ұрланды	2	2	2	4	4	2
банк	3	2	2	3	3	2

Тогда значение PPMI, учитывающее сглаживание Лапласса *add-2*, будет выглядить следующим образом:

Таблица 3.6.– Значения PPMI матрицы термин-документ, учитывающей сглаживание Лапласа.

	PPMI (w, d) [add-2]					
	Doc1	*Doc2*	*Doc3*	*Doc4*	*Doc5*	*Doc6*
тапанша	0	0	0	0,3531	0,1605	0,4056
өлу	0	0,609	0	0	0	0
қақтығысты	0,69718	0,402	0	0	0	0
жол қозғалысы	0,69718	0,402	0	0	0	0
байланған	0	0	0,382	0	0	0,2706
жоғалды	0	0	0,517	0	0	0,7275
атысу	0	0	0,517	0,675	0	0
кинолог	0	0,024	0	0	0	0,3776
ұрланды	0	0	0	0,4406	0,663	0
банк	0,28214	0	0	0,1186	0,341	0

Для измерения подобия двух взвешенных частотных векторов будем определять их косинусное сходство [141]. Пусть x и y будут два вектора, из n элементов

$$x = \langle x_1, x_2, ..., x_y \rangle, \tag{3.14}$$

$$y = \langle y_1, y_2, ..., y_y \rangle. \tag{3.15}$$

Тогда косинус угла Θ между векторами x и y может быть посчитан как скалярное произведение векторов, нормализованное их длинами. Необходимость нормализации обусловлена тем соображением, что даже если частотные вектора x и y будут соответствовать близким по смыслу текстам, более длинный текст будет соответствовать более длинному вектору. Поэтому при определении семантического подобия текстов, длины векторов не имеют значения, учитывается только угол между векторами.

$$\cos(x, y) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}} = \tag{3.16}$$

$$= \frac{x \bullet y}{\sqrt{x \bullet x}\sqrt{y \bullet y}} = \frac{x}{\| x \|} \bullet \frac{y}{\| y \|} \tag{3.17}$$

Для вычисления косинуса между векторами (3.14) и (3.15) мы суммируем произведения их координат $x_1 \bullet y_1 + x_2 \bullet y_2 + \ldots + x_n \bullet y_n$, а затем делим данное произведение на корень квадратный из суммы квадратов их координат (3.16). Формула (3.17) показывает другую возможную запись вычисления косинусного сходства двух векторов, представляющую произведение отдельных единичных векторов, нормализованных их длинами.

Согласно формулам (3.16), (3.17) значение косинуса может варьироваться от

минус единицы, если векторы имеют противоположные направления (Θ =180 градусов) до плюс единицы, если направления векторов совпадает (Θ =0 градусов). Когда векторы перпендикулярны (Θ =90 градусов), косинус угла между ними равен нулю. Если вектор содержит необработанные частоты, которые не могут быть отрицательными, то косинус не может быть отрицательным. Так как по определению значения PPMI веса не могут быть отрицательными, следовательно, значения косинуса между векторами, использующими в качестве координат PPMI, всегда будут лежать в положительном диапазоне [0, 1].

Для расчета косинусного сходства между PPMI векторами документов нашего примера (см. табл. 6) определим длины векторов каждого документа (табл. 3.7)

Таблица 3.7.– Длины PPMI векторов документов рассматриваемого примера.

$\|doc_i\|$, $1 \le I \le 6$					
Doc1	*Doc2*	*Doc3*	*Doc4*	*Doc5*	*Doc6*
1,025535	0,833466	0,824925	0,887975	0,762634	0,953716

Тогда матрица косинусного сходства между документами нашей коллекции будет выглядеть следующим образом (табл. 3.8):

Таблица 3.8.– Матрица косинусного сходства PPMI векторов документов рассматриваемого примера.

	$cos\ (doc_i, doc_k)$, $1 \le k, i \le 6$					
	Doc1	*Doc2*	*Doc3*	*Doc4*	*Doc5*	*Doc6*
Doc1	1	0,655787	0	0,036745	0,123013	0
Doc2	0,655787	1	0	0	0	0,011401
Doc3	0	0	1	0,476408	0	0,609457
Doc4	0,036745	0	0,476408	1	0,574769	0,169113
Doc5	0,123013	0	0	0,574769	1	0,089503
Doc6	0	0,011401	0,609457	0,169113	0,089503	1

Проанализировав полученные результаты можно прийти к следующим выводам:

a) косинусное сходство PPMI векторов документов рассматриваемого примера лежат в промежутке [0,1], что соответствует определению косинуса угла;
b) косинус угла двух одинаковых документов равен единице ($cos\ (doc_i, doc_i) = 1$), так как угол между ними $\Theta = 0$;
c) из рассмотренного массива документов наиболее близкими по смыслу являются документы *Doc1* и *Doc2*, так как максимально значение косинуса угла между ними $cos\ (doc_1, doc_2) = 0{,}655787$);
d) документ *Doc4* ближе по смыслу к документу *Doc5*, чем к документу *Doc1* ($cos\ (doc_4, doc_5) = 0{,}574769 > cos\ (doc_4, doc_1) = 0{,}036745$).

Выводы по третьему разделу

1. Разработан алгоритм семантической разметки текстов казахского языка. Корпус размечается тегами, определяющими семантическое значение токена в триплете факта <*Sub*> , <*Obj*>, <*Pred*> и тегами <*POS type* = "*crime*">, где тег *POS* определяют грамматическую информацию части речи или синтаксическую единицу предложения, а значение атрибута *type=”crim”* определяет семантическое значение криминальной составляющей токена. Первичная разметка корпуса казахского языка, на базе которого, осуществляется обучение, базируется на использовании списка суффиксов и лингвистических правил.

2. Разработан метод вероятностной морфологической разметки, использующий скрытую Марковскую модель (HMM). Используемая в методе функция оценки вероятности цепочки тегов зависит от двух вероятностей: условной вероятности последовательности тегов и условной вероятности обозначения токена данным тегом. На рассмотренном примере последовательности слов английского предложения показана высокая точность предложенного метода.

3. Разработан метод определения семантической близости многоязычных текстовых документов, базирующийся на VSM. Используя идею представления каждого документа в коллекции текстов в виде точки векторного пространства, и дистрибутивную гипотезу статистической семантики, утверждающую, что частоты слов в документе чаще всего определяют семантическую направленность документа, мы показали, что точки, расположенные близко друг к другу в данном пространстве, соответствуют семантически близкими документам, а точки, которые расположены далеко друг от друга, соответствуют сильно отличающимся по смыслу документам.

4. Разработанный метод, определения семантической близости многоязычных текстовых документов, вычисляет косинусное сходство между двумя векторами документов, используя в качестве координат векторов меру PPMI, представляющую весовую функцию, которая добавляет больший вес значениям координат вектора, подтверждающим семантическую зависимость между документом и термином, и присваивает нулевые значения координатам, если появление термина в документе не имеет никакого семантического значения и, следовательно, не несет никакой информации.

4. ПРАКТИЧЕСКАЯ РЕАЛИЗАЦИЯ ПОЛУЧЕННЫХ РЕЗУЛЬТАТОВ

4.1. Разработка и аннотирование параллельного корпуса текстов криминально-окрашенной информации казахского и русского языков

Технология создания параллельного корпуса включает в себя несколько этапов. Первый базовый этап требует использования специализированных программных инструментов и техник для сбора текстового материала корпуса. При этом, несмотря на то, что интернет содержит огромное количество двуязычных и мульти-язычных веб-сайтов, выбор нужных двуязычных ресурсов представляет собой важную часть разработки параллельного корпуса. Это задание усложняется в связи с тем, что мы обрабатываем два таких разных языка, как казахский и русский.

В нашем исследовании для автоматического сбора (scraping) текстов сайтов мы использовали собственную программу, осуществляющую разбор сайтов, заданной структуры и требуемого контекста [142].

Для сбора текстов были выбраны четыре двуязычных веб-сайта *zakon.kz*, *caravan.kz*, *lenta.kz* и *nur.kz*. Выбранные сайты представляют собой известные и надежные порталы Республики Казахстан, одним из новостных направлений которых являются криминальные новости. Порталы могут содержать новости о таких криминальных деяниях как грабежи, угоны машин, убийства, ДТП и другие. Тексты именно данной предметной области и представляют базовый ресурс создаваемого корпуса. Кроме того, данные порталы являются двух язычными и часто содержат информационно близкие новостные публикации на казахском и русском языках.

В результате осуществленного скрапинга вышеперечисленных сайтов мы получили 3000 текстов на двух языках: русском и казахском. На следующем этапе была вручную осуществлена выборка текстов криминальной окрашенности. Таким образом, был получен корпус, размером более 50410 слов (около 25600 слов принадлежат казахской половине корпуса и около 24800 слов — русской).

На следующем шаге была определена структурная организация корпуса. На сегодняшний день существует три базовых формата корпусов, определяемые в зависимости от прагматических целей создателей или пользователей: (1) традиционный текстовый формат с отсылками к переводу; (2) представление текстов в табличной «зеркальной» форме, более удобной для восприятия и сравнения, (3) организация параллельного корпуса в форме базы данных.

Для нашей работы мы использовали два параллельных формата: (1) база данных, как наиболее удобный способ хранения большого количества данных, представляющих небольшие текстовые фрагменты, с возможностью дальнейшей увеличения базы. Все новостные статьи созданного корпуса хранятся в таблице базы данных, которая включает ID, название, адрес сайта, и текст статьи (рис. 4.1).

id	head	url	text
3256	Месть за сестру:	https://www.inform.kz/	АЛМАТЫ. КАЗИНФОРМ - В специализи
3257	Чем чаще всего	https://www.inform.kz/	АСТАНА. КАЗИНФОРМ - Депутат Сенат
3258	Сенаторы ратиф	https://www.inform.kz/	АСТАНА. КАЗИНФОРМ - Депутаты Сен
3259	115 тысяч кварти	https://www.inform.kz/	АСТАНА. КАЗИНФОРМ - 115 тысяч ква
3260	Сенатор выразил	https://www.inform.kz/	АСТАНА. КАЗИНФОРМ - Депутат Сенат
3261	820 камер видеон	https://www.inform.kz/	КОКШЕТАУ. КАЗИНФОРМ - 820 камер в
3262	Почему приоста	https://www.inform.kz/	АСТАНА. КАЗИНФОРМ - Вице-министр
3263	Задержан подозр	https://www.inform.kz/	ТУРКЕСТАН. КАЗИНФОРМ - Полицейск
3264	Сенатор рассказ	https://www.inform.kz/	АСТАНА. КАЗИНФОРМ - В Казахстане р

Рисунок 4.1 – Фрагмент БД криминально окрашенных текстов новостных веб-сайтов на казахском и русском языках.

Кроме того, была разработана файловая структура параллельного корпуса казахских и русских криминально-окрашенных текстов Веб-медиа, показанная на рисунке 4.2.

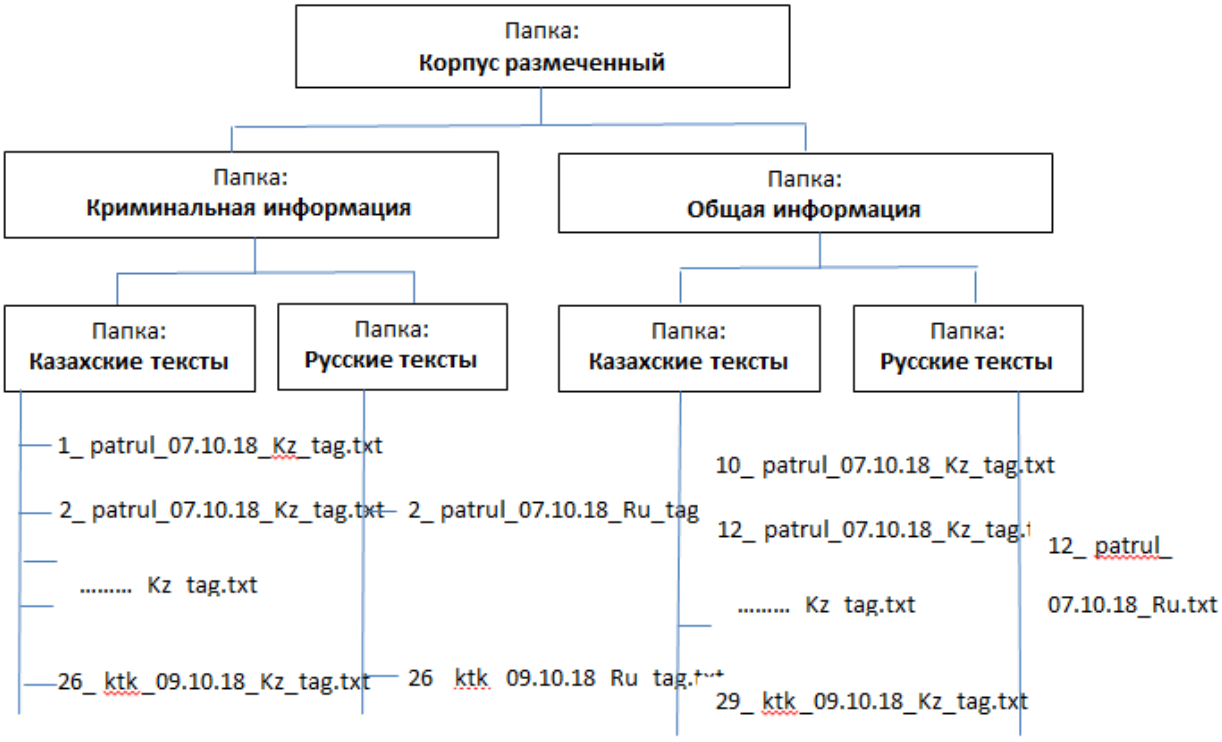


Рисунок 4.2 – Файловая структура разрабатываемого параллельного корпуса

В данной структуре хранения файлов имя текстового файла должно соответствовать шаблону:

порядковыйНомер_названиеАгенства_дата_язык_tag|row.txt

Например, размеченный текстовый файла текста порядкового номера 49 на казахском языке, полученный на сайте информационного агентства *partrul* седьмого сентября 2018 года должен иметь имя: 49*_partrul_07.09.18_Kz_tag.txt*

Критерием оценки текстового корпуса, помимо его репрезентативности и размера, является используемая система разметки и правильность кодирования метаданных корпуса. Разметка представляет собой добавление в текстовый корпус

некоторой дополнительной лингвистической информации. Эта информация может быть морфологическая (POS-tagging), синтаксическая, семантическая и т.д.

Для создаваемых корпусов криминально значимой информации мы используем морфологическую и семантическую разметку. Использование ручной разметки повышает ее точность, в то же время требует много трудозатрат и не позволяет размечать корпуса большого объема. В то время как использование автоматической разметки увеличивает количество ошибок, уменьшает точность, но позволяет достаточно быстро осуществлять разметку больших корпусов. Кроме того, в некоторых случаях достаточно сложно создать алгоритмы семантической, стилистической и некоторых других видов разметки. По этой причине, в нашем исследовании POS-tagging и семантическая разметка добавлялась в корпус автоматически, для обучающих базовых (train) корпусов, с использованием разработанной модели, а для дополнения корпусов с использованием методов ML.

Для осуществления POS-тегирования русских текстов корпуса использовался пакет *pymorphy2*[5] Python, разработанный специально для морфологического анализа русскоязычных текстов. Библиотеки пакета используют словарь *OpenCorpora*[6] и делают гипотетические выводы по нераспознаваемым словам.

Сложность структурного и типового аннотирования текстов казахского языка связана с его принадлежностью к агглютинативным языкам. Агглютинативный формат, в котором каждая агглютинация (суффикс или окончание) несет только одно семантическое или морфологическое значение, противоположен флективному, в котором каждый формант имеет несколько неделимых значений сразу (например, падежа, рода и числа) (см. §§ 1.2, 2.3). Мы осуществляли POS-тегирование казахских текстов с помощью регулярных выражений, основанных на классе *RegexpTagger* библиотеки *nltk* Python, и ряде синтаксических правил. Например, мы можем идентифицировать некоторые типы существительных казахского языка с помощью списка регулярных выражений:

```
patterns=[(r'.*(шық|шы|пыр|мпыр|алар|ашыщ|лар|елер|ды)
$', 'NNat'), (`r.*(мның|енің|рдың|дың,)$','NNil'),
(r'.*(да|те|та|нда|нде|ға|ге|қа|ке|на|не|тік|еге|ырға|рға}
йға|ыға|аға|шаға|сіз|мға|ға)$','NNba')]
```

Здесь, тег `'_NNat'` определяет именительный падеж (атау) существительного, тег `'_NNil'` определяет родительный падеж (ілік), а тег `'_NNba'` — дательно-направленный падеж (барыс) существительного.

Для того чтобы увеличить точность POS-тегирования дополнительно использовались семь синтаксических правил. Базой разработки таких правил являлся строгий порядок слов в предложениях казахского языка. Например, "*Если токен следует за словами из специального списка – токен отмечается как глагол*":

$$[\texttt{list\_1 of words}]\ \ \texttt{token}_i \Rightarrow \texttt{tag}_{\texttt{token}i}\texttt{='\_VV'},$$

где `list_1 of words = [қойды, қой, қалды, қал, салдым, салып,`

[5] https://nlpub.ru/Pymorphy

[6] https://www.pydoc.io/pypi/gensim-3.2.0/autoapi/corpora/dictionary/index.html

кетті, кетсеңші, бару, келу, шығу, жүру, тұру, бар, қел, шық, жүр, қайт, шыққан, барған, түсті, түс, тұрыңдар, тұсын, көрме, …]

На первом этапе автоматического выравнивания корпуса, для выделения предложений, использовался модуль *nltk.tokenize* Python.

На следующем шаге, возможно, выбрать несколько подходов выравнивания предложений. Первый подход, основанный на совпадении длин предложений выравниваемых текстов, обеспечивает более высокую продуктивность. Однако, в нашем исследовании, не смотря на существующие преимущества, данный подход не может принести точных и объективных результатов, так как в казахском языке часто для выражения некоторой смысловой и морфологической информации используются дополнительные слова, коренным образом изменяющие длину предложения. Именно по этой причине различия в организации грамматики и семантики флективных и агглютинативных языков, использование выравнивания по длинам предложений в нашем параллельном казахско-русском корпусе было признано не эффективным.

При втором, используемом нами, более ресурсоемком подходе, применяется лексическое выравнивание слов. В качестве «лексического инструмента выравнивания» использовался созданный казахско-русский словарь (рис. 4.3), основанный на англо-казахско-русском словаре, состоящим, примерно, из 50 000 элементов (рис.4.4) .

kz	ru
қылмыстық	криминал
мәйіті	трупы
мәліметтерге	данные
министрлігімен	с министерством
Оңтүстік	юг
орындарын	мест
өзара	взаимное
өзінің	его собственный
пиғылды	недобросовестный
Рудныйда	в Рудном
сату	продажа
сәттері	моменты

Рисунок. 4.3 – Фрагмент созданного казахско-русского словаря

```
native###_туасынан,·жаратылысынан,·тумысынан###_прирождённый·
(прирожденный)¶
native###_табиғи,·жаратынды###_природный¶
native###_тұнық,·ұқыпты,·кірсіз,·пәк,·мөлдір,·кіршіксіз,·бейкүнә,·
әйнектей,·ақ,·мұнтаздай,·таза,·шайдай·ашық,·мөлдір·бұлақ,·дақсыз,·саф,·
ашық,·адал,·шын,·нағыз,·айна·дай,·бәкізе,·ғилман,·жазықсыз###_чистый¶
native###_жергілікті·адам,·туып-өскен,·туған###_уроженец¶
native###_туған,·бір·туған,·тіған-туысқандар,·туыстық,·туып·өскен,·
қарағым,·шырағым###_родной¶
native·of·a·country·(·aborigine·)###_абориген###_абориген¶
nativity·(the·Nativity)###_рождество###_рождество¶
nativity###_дүниеге·келу,·жаралу###_рождение¶
north·atlantic·treaty·organization·(NATO)###_°###_североатлантический·
союз·(NATO)¶
North·Atlantic·Treaty·Organization·(NATO)###_°###_Североатлантический·
союз·(NATO)¶
nato·(NATO;·North·Atlantic·Treaty·Organization)###_°###_нато·(NATO;·
```

Рисунок . 4.4 – Фрагмент используемого базового англо-казахско-русского словаря

Созданный выровненный параллельный казахско-русский корпус состоит из текстов четырех казахских новостных сайтов периода май – декабрь 2018 года, имеющих криминальную окрашенность. Корпус включает около 50410 слов.

Для оценки точности проведенного автоматического выравнивания предложений корпуса было использовано экспертное оценивание тремя носителями казахского и русского языков. Для работы экспертов использовалось специально разработанное приложение, интерфейс которого показан на рисунке 4.5.

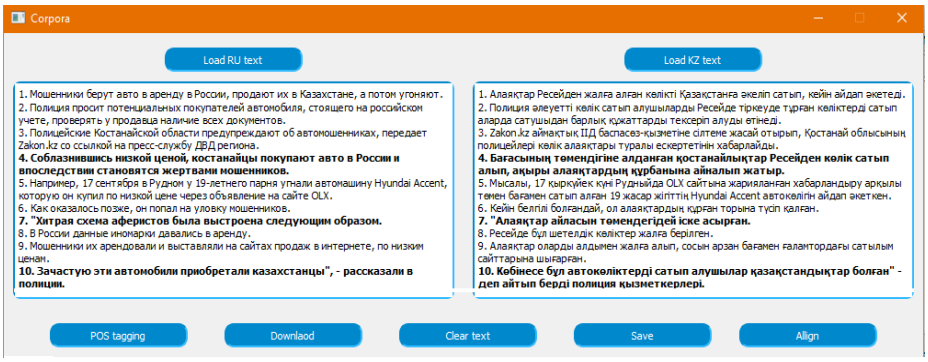


Рисунок . 4.5 – Пользовательский интерфейс приложения, используемого для работы с выровненным параллельным корпусом.

Приложение позволяет экспертам выбрать текст на любом (русском или казахском) языке и автоматически загружает параллельный файл с текстом. Жирный шрифт используется для предложений, не получивших параллельный эквивалент на противоположном языке после автоматического выравнивания. Работая с корпусом, эксперты могут отметить тексты, сохранить их с отметкой и выровнять их вручную

Проведенное экспертное оценивание показало, что точность автоматически выровненных предложений, созданного параллельного казахско-русского корпуса, составляет около 60%, при коэффициенте согласованности (*agreement*) равном 0.83. Остальные предложения выровнены вручную.

Проанализировав полученные результаты выравнивания, были сделаны следующие выводы о причинах ошибок.

- Наибольше влияние на процент точности выровненного корпуса оказывает большая разница синтаксических структур казахского и русского языков, что глобально приводит несовпадению количества предложений в двух частях корпуса. Некоторым предложениям русского текста соответствует несколько предложений казахского текста.
- Результат словарного метода выравнивания во многом зависит от качества используемого переводного словаря. Однако, в связи с тем, что русский и казахский языки находятся в отдаленных группах, при создании словарей возможны некоторые ошибки многозначности.
- Сложность и ограниченность использования сопоставимой грамматики для казахского и русского языков требует дальнейшей работы в этом направлении контрастивной лингвистики.

- Выравнивание текстов криминальной тематики требует учета имен собственных, званий, должностей и некоторых паттернов семантических классов слов (валют, дат и др.), особенно в новостных заголовках.

4.2. Метрика оценки эффективности моделей машинного обучения.

Метрика числовых оценок эффективности моделей ML часто базируется на хорошо известных в системах автоматического извлечения текстовой информации понятиях "*точности*" (precision) и "*полноты*" (recall).

Для оценки результатов технологии поиска семантически близкого текста и разработанной модели Open IE мы так же будем использовать *precision* и *recall* и F_1-*measure*. Выбранный метод оценки эффективности базируется на *методе тестовых коллекций* [143, 144], который заключается в сравнении результатов работы исследуемой схемы на заранее определенных данных с оценками экспертов на тех же данных. При этом, в системах машинного обучения в качестве данных, имеющих экспертную оценку, используется «эталонный» результат, которые представляет собой некий, так называемый, *золотой стандарт* (gold standard test data). В нашем случае таким стандартом является заранее созданный *test corpus*, доступный в качестве стандарта, который либо уже предварительно размечен морфологически или семантически, либо в котором уже определено наличие или отсутствие семантической близости текстов документов к узкой тематической тематике.

В результате проведенного эксперимента мы можем разделить все обработанные тексты на четыре множества, показанные в таблице 4.1 и на рисунке 4.6:

tp (*true positive*) — документы, которые правильно автоматически определены, как семантически близкие к документом данной узкой предметной области;

fn (*false negatives*) — документы, которые неправильно автоматически определены, как семантически не близкие к документам заданной области;

fp (false positives) — документы, которые неправильно автоматически определены, как семантически близкие к документам заданной узкой предметной области;

tn (true negatives) — документы, которые правильно автоматически определены, как не близкие по смыслу к текстам заданной предметной области.

Используя данные метрики можно посчитать точность, которая показывает насколько используемая технология или модель правильна (или корректна).

$$\text{accuracy} = \frac{tp + tn}{tp + tn + fp + fn} \qquad (4.1)$$

Однако, использование данной меры *accuracy*, в качестве определителя эффективности моделей и технологий ML, не приемлемо для редких сущностей. Например, если в нашем корпусе только несколько документов оказывается семантически близкими к запрашиваемой области (что на практике, и чаще всего

встречается), тогда даже если разработанная технология будет совершенно не верной, значение *accuracy* все равно будет стремиться к 100 %.

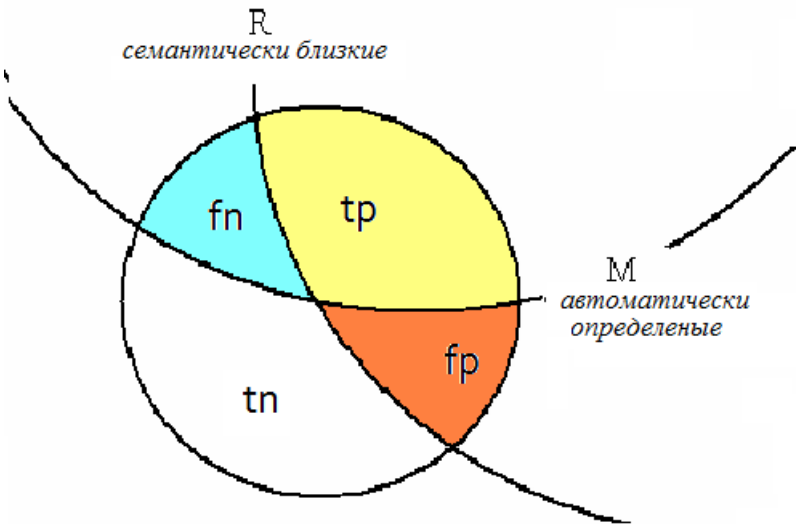


Рисунок. 4.6– Метрики оценки эффективности моделей машинного обучения

Таблица 4.1 – Описанием метрик оценки эффективности технологии определения семантической близости документа к текстам узкой предметной области

	Семантически близкие к заданной предметной области, *R*	Семантически не близкие к заданной предметной области
Автоматически определенные *(positive)*, *M*	*tp (true positive)*, правильно автоматически определены, как семантически близкие к документам данной узкой предметной области	*fp (false positives),* неправильно автоматически определены, как семантически близкие к документам заданной узкой предметной области
Автоматически не определенные *(negatives)*	*fn (false negatives)*, неправильно автоматически определены, как семантически не близкие к документам заданной области	*tn (true negatives),* правильно автоматически определены, как не близкие по смыслу к документам заданной предметной области

Таким образом, для наших вычислений метрика *accuracy* будет не совсем подходящей. В таком случае, так как нас интересует только правильно определенные документы, т.е. коэффициент *true positive*, поэтому необходима метрика, которая будет больше сфокусирована на правильно найденных семантически близких текстах. Такой метрикой является точность или *precision*, вычисляемая, как отношение корректно автоматически определенных семантически близких к данной узкой предметной области текстов к общему количеству автоматически определенных текстов:

$$precision = \frac{tp}{tp + fp} \tag{4.2}$$

Кроме точности, определяется противоположная мера — *recall* (полнота), которая говорит о текстах, реально являющихся близкими по смыслу, к заданным, и показывает какой процент из них был определен правильно:

$$recall = \frac{tp}{tp + fn} \tag{4.3}$$

Очевидно, что коэффициенты точности и полноты являются взаимно связанными. В существующих задачах машинного обучения, если мы увеличим полноту, то уменьшается точность и, наоборот, если в приложении увеличивается точность, то автоматически уменьшается полнота. Это связано с двумя типами существующих в технологиях NLP ошибок.

Для наших моделей ошибка первого типа — *false positives*, когда документы, которые на самом деле не являются близкими по смыслу к заданной группе текстов, будут автоматически определены, как семантически близкие.

И ошибка второго типа — *false negatives*, когда документы, которые на самом деле являются близкими по смыслу к заданной группе текстов, не будут автоматически определены, как семантически близкие. Снижение ошибок одного типа, в любом приложении, обрабатывающем тексты, повышает процент ошибок другого типа.

Для того чтобы определить некоторый баланс между метриками полноты и точности, мы объединяем эти две метрики в некоторую усредненную величину. Но, при этом, невозможно использовать среднее арифметическое, т.к. при стремящейся к нулю точности, среднее арифметическое полноты и точности будет не меньше 50 %. Исходя из вышесказанного, будем опираться на среднее гармоническое точности и полноты, называемое мерой Ван Ризбергена или F-measure, определяемое как:

$$F_\beta - measure = \frac{1}{\alpha \frac{1}{precision} + (1-\alpha)\frac{1}{recall}} = \frac{(\beta^2 + 1) \times precision \times recall}{\beta^2 \times precision + recall}, \tag{4.4}$$

где где $\alpha \in [0,1]$, $\beta^2 = \frac{(1-\alpha)}{\alpha}$, $\beta \in [0,\infty]$. Если $0<\beta<1$ — большее значение при расчете уделяется точности, а при $\beta>1$— больший вес приобретает полнота.

При значении коэффициентов $\alpha=1/2$ или $\beta=1$ F-мера придает одинаковый вес полноте и точности и становится сбалансированной F_1-мерой:

$$F_1 - measure = \frac{2}{\frac{1}{precision} + \frac{1}{recall}} = \frac{2 \times precision \times recall}{precision + recall}, \tag{4.5}$$

В нашей практической задаче, поиска текстов, тематика которых близка по смыслу к текстам, содержащим криминально окрашенную информацию, большее вес необходимо придать полноте в отличие от задачи информационного поиска,

когда больший вес, обычно, придается точности.

4.3. Особенности реализации и экспериментальные результаты модели Open IE

В данном разделе приведены практические результаты реализации разработанных моделей и методов для анализа многоязычной текстовой информации на основе машинного обучения.

Набор данных для экспериментальных исследований включал три корпуса русского, казахского и английского текстов. Тексты получены с помощью разработанного парсера, основанного на библиотеке BeautifulSoup[7] языка Питон с двуязычных новостных казахских веб-сайтов *inform.kz*, *azattyq.org*, *patrul.kz*, *zakon.kz*, *caravan.kz*, *lenta.kz*, *nur.kz*. Время сбора информации: с июня 2018 по июнь 2019 года. Выбор сайтов обусловлен: (1) надежностью перечисленных сайтов; (2) возможностью выбора узкоспециализированных текстов по теме – криминально-окрашенная информация; (3) мультиязычность сайтов, с возможностью переключения между казахским и русским языками. Для создания корпуса английских текстов использовались англоязычные новостные веб-сайты: *edition.cnn.com, news.sky.com* и *foxnews.com*.

Общий объем построенных корпусов: 6000 текстов из 700 000 слов, 275 тыс. из них в корпусе русских текстов, 225 тыс. в корпусе казахских текстов и около 200 тыс. в корпусе англоязычных текстов.

Для идентификации грамматических характеристик слов английских предложений использовался Stanford Dependencies (SD) парсер, наиболее эффективно обрабатывающей глагольные группы и подчиненные предложения. В формальной грамматике зависимостей глагол является центральным элементом фразы, а подлежащее и дополнение прямо или косвенно зависят от него [105]. Такие отношения между словами предложения аналогичны, отношениям, которые в нашей модели устанавливаются между участниками действия — *Subject*, *Object* и глаголом, называющим действие (*Predicate*) триплета факта.

Для выделения триплета факта используются 7 из 40 грамматических отношений слов в английских предложениях, содержащихся в Universal Dependencies (UD) v1[8]: *subj*, *nsubjpass*, *csubj*, *obj*, *iobj*, *dobj*, *ccomp*. Метка *nsubj* показывает синтаксическую зависимость подлежащего от корневого глагола предложения, метка *csubj* показывает синтаксическую зависимость придаточного предложения, а метка *nsubjpass* показывает подлежащее в предложении пасивного залога английского языка. Отношение *obj* определяет прямое дополнение, т.е. сущность, на которую воздействуют или которая претерпевает изменение состояния или движения. Метки *iobj*, *dobj* и *ccomp* используются для более точного обозначения зависимостей объекта действия от глагола.

Например, на рисунке 4.7 показан фрагмент получаемой разметки POS-tagging и UD на языке XML, а рисунок 4.8 отображено графическое представление

[7] https://www.crummy.com/software/BeautifulSoup/
[8] http://universaldependencies.org/en/dep/

универсальных зависимостей для английского предложения "*He referred to the deaths as "the cost of doing business on that particular engagement*", которое получено с помощью инстументов визуализации для парсера зависимостей DependenSee[9].

```
        <token id="3">
                <word>War</word>
                <CharacterOffsetBegin>9</CharacterOffsetBegin>
                <CharacterOffsetEnd>12</CharacterOffsetEnd>
                <POS>NNP</POS>
          </token>
….
          <token id="6">
                <word>Province</word>
                <CharacterOffsetBegin>22</CharacterOffsetBegin>
                <CharacterOffsetEnd>30</CharacterOffsetEnd>
                <POS>NNP</POS>
          </token>
…
        <dep type="nsubj">
                <governor idx="16">consisted</governor>
                <dependent idx="3">War</dependent>
        </dep>
…
        <dep type="case">
                <governor idx="6">Province</governor>
                <dependent idx="4">in</dependent>
        </dep>
```

Рис. 4.7. Фрагмент разметки UD парсером предложения английского языка

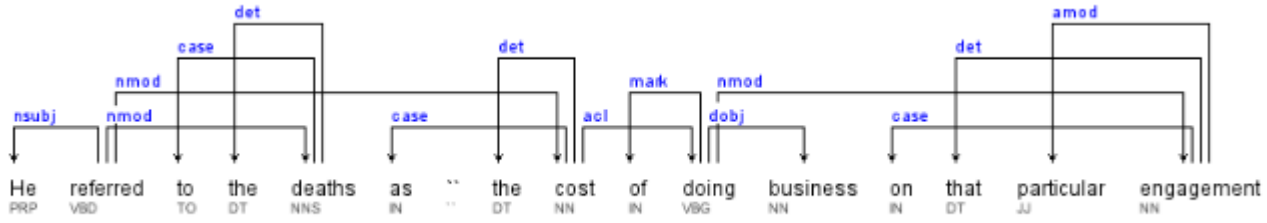


Рис.4.8. Графическое представление универсальных зависимостей для английского предложения "*He referred to the deaths as "the cost of doing business on that particular engagement*"

Согласно разработанной модели Open IE, для получения триплетов знаний, определенные на этапе POS-tagging и синтаксического парсера грамматические

[9] http://chaoticity.com/dependensee-a-dependency-parse-visualisation-tool/

характеристики слов английских предложений подставляются в качестве значений предметных переменных в уравнения (2.12) – (2.15)

В таблице 4.2 показаны примеры фактов, автоматически извлеченных из английских предложений.

Таблица 4.2. Фрагмент таблицы фактов, извлеченных из английских предложений.

Number of sentence	Predicat, verb	nsubj	Advcl	dobj
1	known			
1	consisted	War	fighting	
2	lasted, took	War, majority		place
3	focused	insurgents	featured, ambushing	
3	featured	fighting		warfare
3	improvised			
4	were			
4	killed	Iraqis		
5	saw	Anbar		fighting
6	was			
6	occupied	it		
7	killed	Iraqis		
7	were			
7	began	Violence		
8	relinquished	Army		command
9	was			
10	occurred	fighting		
11	struggled	sides		
11	secure			Valley
11	escalated	Violence	struggled	

Для экспериментальной проверки разработанной модели на корпусах русского и казахского языков использовалось POS-тегирование, позволяющее явным образом определить значения предметных переменных, используемых в модели.

Для морфологической разметки корпуса текстов русского языка использовался пакет pymorphy2[10] Питона, разработанный специально для морфологического анализа русскоязычных текстов.

[10] https://pymorphy2.readthedocs.io/en/latest/

Согласно разработанной модели Open IE, для получения триплетов знаний, грамматического анализа русского языка, характеристики слов подставляются в качестве значений предметных переменных в уравнения (2.19) – (2.22)

POS-тегирование казахских текстов осуществлялось с использованием разработанного тегера, базирующегося на *RegexpTagger* классе пакета *NLTK Python*. На рисунке 4.9 показан фрагмент регулярного выражения, позволяющего идентифицировать некоторые формы существительных в казахских предложениях.

```
patterns=[(r'.*бен$','NN'), (`r.* пенен$','NN'), (`r.*
   басшылық$','NN'), (r'.* ықону$','NN'), (r'.*
  тармен$','NN'), (r'.* ғерлермен$','NN'), (r'.*
                  здар$','NN')]
```

Рисунок 4.9 – Фрагмент регулярного выражения, позволяющего идентифицировать некоторые формы существительного в казахских предложениях.

Использование полученных имен тегов и некоторых синтаксических характеристик слов в предложении в качестве значений предметных переменных уравнений (2.34 – 2.36) позволяет извлекать *Subject, Object* и *Predicate* факта из предложений казахского языка.

Так как с помощью модели создается именно учебный корпус (a training corpus), для оценки эффективности модели использовался только показатель точности (precision). Для оценки результатов автоматического извлечения фактов из текстов на трех языках использовалось экспертное оценивание, а именно – корректность извлеченных фактов оценивалась двумя экспертами для корпуса каждого языка.

Использовалась следующая методика экспертного оценивания. Из автоматически извлеченных из каждого корпуса списка фактов произвольным образом было взято примерно по тысячи фактов и представлено для оценки эксперту – носителю соответствующего языка (английского, русского, казахского). Эксперт оценивал извлеченный факт как 1, если триплет факта идентифицирован правильно. Т.е. правильно определены все три элемента факта: инициатор действия — Subject, предмет или лицо, на которого направленно действие, — Object и Predi-ate — называет действие, которое объединяет его участников. Если же хотя бы один из трех элементов факта был выявлен не верно, эксперт оценивал данный факт как 0 —неверно определенный и извлеченный факт.

В таблице 4.3 показаны полученные точность и коэффициент согласованности (agreement) разработанной модели для корпусов текстов английского, русского и казахского языков.

Таблица 4.3. Точность и согласованность (agreement) разработанной модели для корпусов текстов английского, русского и казахского языков

Язык корпуса	Размер корпуса	precision	agreement
английский	200 000	87,2 %	0,91

русский	275 000	82,4 %	0,78
казахский	225 000	71,0 %	0,72

4.4. Методика экспертной оценки качества технологии определения семантической близости текстов

Для определения принадлежности произвольного текста к криминальной тематики используем созданные корпуса (см. §§4.1, 4.3), содержащие новостные статьи из десяти интернет источников по таким тематическим категориям, как: война, терроризм, чрезвычайные ситуации и происшествия, экстремизм, преступления и др.

Информационная технология определения принадлежности документа к узкоспециализированной области включает в себя три основных этапа, показанных на рисунке 4.9: (1) лингвистическую обработку необработанного корпуса узкоспециализированной тематики (raw corpus) и поступающего на вход документа; (2) этап машинного обучения; (3) этап определения косинусного сходства.

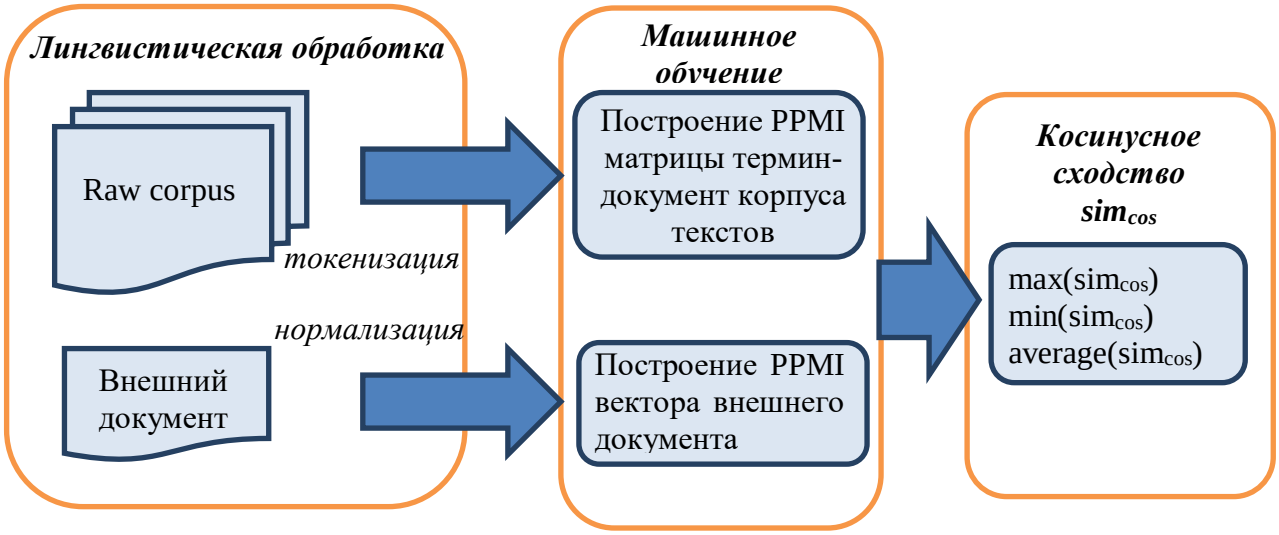


Рисунок. 4.9 – Общая схема используемой технологии.

Наши необработанные данные представляют собой большой корпус текстов на естественном языке. На первом этапе создания матрицы термин-документ необходимо осуществить некоторую лингвистическую обработку необработанного текста, заключающуюся, в простейшем случае, в токенизации текста. Токенизация представляет собой разбиение текста на более мелкие части, токены, которыми в общем случае являются слова текста. В процессе токенизации решается, что представляет собой термин и как извлечь термины из необработанного текста.

На втором шаге лингвистической обработки для русского и английского языков мультиязычной базы текстов осуществляется нормализация, заключающаяся в преобразовании внешне разных строк символов к одной и той же форме. Причина необходимости нормализации связана с тем, что часто различные последовательности символов (знаков) передают, по существу, идентичные значения (один концепт). Учитывая, что в используемой модели необходимо

понять смысл слов и найти близкие по смыслу документы, представляется разумным нормализовать варианты различных форм слова к одной форме. Например, слова “*автомобиль*”, “*автомобили*”, “*автомобилями*” и “*автомобилях*” будут нормализованы к форме “*автомобиль*”.

В агглютинативных языках, каким является казахский язык, многие понятия объединяются в одно слово, используя различные префиксы, инфиксы и суффиксы. Одно слово в агглютинативном языке может соответствовать предложению из полдюжины слов в английском [145]. Исходя из выше изложенного, технология не использует нормализацию казахских текстов, так как она приведет к резкому уменьшению точности определения близких по смыслу текстов.

Второй этап обработки представляет собой этап машинного обучения, заключающегося в построении PPMI матрицы термин-документ корпуса текстов и PPMI вектора входного документа. Работа данного этапа базируется на методе, описанном в подразделе 3.3.

На третьем этапе вычисляется минимальное, максимальное и среднее значение коэффициента семантической близости, определяемое через косинусное сходство между вектором документа, поступившим на вход, и векторами документов имеющейся PPMI матрицы термин-документ корпуса.

На рисунке 4.10 показан пример работы разработанного программного приложения MainWindow, позволяющего определить наличия криминального значения в поступающих новостных текстах.

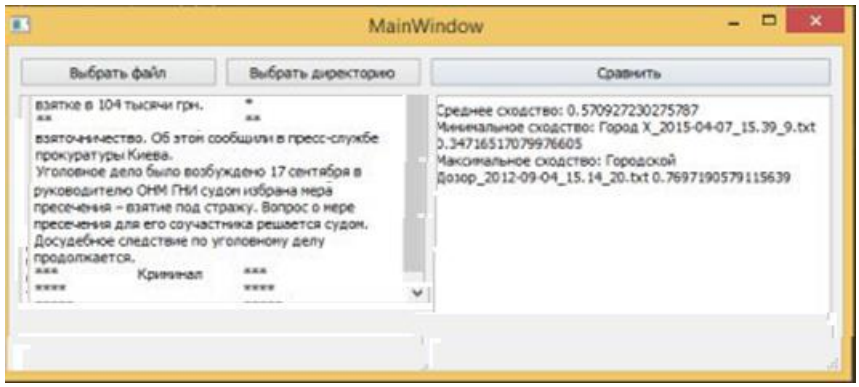


Рисунок 4.10 – Пример работы приложения, позволяющего определить наличие криминального смысла в поступающих новостных текстах.

Для определения значения коэффициента sim_{cos} позволяющего отнести входной документ к тематике рассматриваемого корпуса был проведен эксперимент на двух корпусах. Обучающий корпус (train corpus) содержал тексты криминальной окрашенности. Тестовый корпус (test corpus) содержал тексты, которые как принадлежали к криминальной тематике, так и тексты, принадлежавшие различным тематикам.

В ходе эксперимента вычислялось значение косинусного сходства тестируемого текста doc_{test} и каждого текста обученного корпуса $sim_{cos}(doc_{test}, doc_{train})$. В таблице 4.4 приведено максимальное, минимальное и среднее значение косинуснусного сходства между каждым конкретным криминально окрашенным

текстом тестируемого корпуса (test corpus) и документами обученного корпуса (train corpus).

Таблица 4.4 – фрагмент результатов анализа тестового корпуса криминально окрашенных текстов

Имя файла	№	Косинусное сходство, sim_{cos}		
		$max(sim_{cos})$	$min(sim_{cos})$	$average(sim_{cos})$
zakon.kz _2018-08-31_11.33_1.txt	1	0,92	0,26	0,72
zakon.kz _2018-08-31_12.27_2.txt	2	0,77	0,35	0,57
zakon.kz _2018-08-31_12.33_3.txt	3	0,79	0,38	0,67
zakon.kz _2018-08-31_16.24_4.txt	4	0,78	0,38	0,67
zakon.kz _2018-08-31_16.39_5.txt	5	0,82	0,36	0,69
zakon.kz _2018-09-01_17.44_6.txt	6	0,80	0,39	0,68
zakon.kz _2018-09-02_10.26_7.txt	7	0,88	0,27	0,69
zakon.kz _2018-09-02_12.58_8.txt	8	0,80	0,27	0,60
zakon.kz _2018-09-02_17.00_9.txt	9	0,87	0,31	0,71
lenta.kz_2018-01-14_13.50_2.txt	12	0,77	0,50	0,66
lenta.kz_2018-01-14_14.10_3.txt	13	0,82	0,48	0,68
lenta.kz_2018-01-14_14.40_4.txt	14	0,77	0,49	0,67
lenta.kz_2018-01-15_15.10_5.txt	15	0,73	0,28	0,51
lenta.kz_2018-01-16_11.10_6.txt	16	0,81	0,36	0,68
lenta.kz_2018-01-16_15.00_7.txt	17	0,88	0,31	0,71
lenta.kz_2018-01-16_18.40_8.txt	18	0,85	0,33	0,69
nur.kz_2018-10 16_00.09_1.txt	22	0,79	0,42	0,70
nur.kz_2018-10-17_00.13_2.txt	23	0,67	0,41	0,59
nur.kz_2018-10-17_00.18_3.txt	24	0,89	0,37	0,67
nur.kz_2018-10-18_00.18_4.txt	25	0,74	0,38	0,61
nur.kz_2018-10-19_00.02_5.txt	26	0,77	0,43	0,62
nur.kz_2018-10-22_00.19_6.txt	27	0,84	0,43	0,72

Проведенный анализ полученных результатов позволил сделать вывод, что минимальное значение коэффициента косинусного сходства (sim_{cos}) между криминально окрашенными текстами тестового корпуса и документами обученного корпуса не меньше 0,3 ($min(sim_{cos})> 0,3$), максимальное значение $max(sim_{cos})>0,7$, а среднее значение $average(sim_{cos})>0,55$.

В таблице 4.5 показаны определенные значения $min(sim_{cos})$, $max(sim_{cos})$ и $average(sim_{cos})$, но уже для документов тестового корпуса, относящихся к любой тематике, за исключением криминально специализированной.

Таблица 4.5 – фрагмент результатов анализа тестового корпуса произвольной тематики

Имя файла	№	Косинусное сходство, sim_{cos}		
		$max(sim_{cos})$	$min(sim_{cos})$	$average(sim_{cos})$
zakon.kz _2018-08-31_1.txt	1	0,63	0,24	0,41
zakon.kz _2018-08-31_2.txt	2	0,66	0,25	0,47
zakon.kz _2018-08-31_3.txt	3	0,69	0,19	0,45
zakon.kz _2018-08-31_4.txt	4	0,67	0,21	0,41
zakon.kz _2018-08-31_5.txt	5	0,73	0,24	0,48
lenta.kz_2018-09-6.txt	6	0,60	0,25	0,44
lenta.kz_2018-09-7.txt	7	0,51	0,15	0,36
lenta.kz_2018-09-8.txt	8	0,74	0,20	0,53
lenta.kz_2018-09-9.txt	9	0,67	0,22	0,47
lenta.kz_2018-09-10.txt	10	0,62	0,23	0,43
lenta.kz_2018-09-11.txt	11	0,76	0,19	0,55
lenta.kz_2018-09-12.txt	12	0,75	0,22	0,60
caravan.kz_2018-10-19_13.txt	13	0,72	0,25	0,51
caravan.kz_2018-10-19_14.txt	14	0,65	0,19	0,47
caravan.kz_2018-10-19_15.txt	15	0,66	0,19	0,50
caravan.kz_2018-10-19_16.txt	16	0,62	0,21	0,38
caravan.kz_2018-10-19_17.txt	17	0,56	0,28	0,45
caravan.kz_2018-10-19_18.txt	18	0,62	0,20	0,44
caravan.kz_2018-10-19_19.txt	19	0,60	0,18	0,39
caravan.kz_2018-10-19_20.txt	20	0,65	0,24	0,43
caravan.kz_2018-10-19_21.txt	21	0,73	0,25	0,50
caravan.kz_2018-10-19_22.txt	22	0,65	0,18	0,44
caravan.kz_2018-10-19_23.txt	23	0,59	0,23	0,47
caravan.kz_2018-10-19_24.txt	24	0,56	0,27	0,45
caravan.kz_2018-10-19_25.txt	25	0,61	0,25	0,48
caravan.kz_2018-10-19_26.txt	26	0,55	0,20	0,39
caravan.kz_2018-10-19_27.txt	27	0,70	0,28	0,41

Проанализировав полученные результаты можно сделать вывод, что среднее значение косинусного сходства между текстами обученного корпуса и текстами произвольной тематики обычно находится в пределах $0,35 < average(sim_{cos}) < 0,50$. Максимальное и минимальное значения ниже 0,76 и 0,30, соответственно: $max(sim_{cos}) < 0,76$ и $min(sim_{cos}) < 0,30$.

На базе проведенного экспериментального исследования сформулирована гипотеза о том, что если среднее значение коэффициента косинусного сходства между входным документом и документами обученного корпуса больше 0,50, то данный документ может быть отнесен к узкоспециализированным документам, содержащим криминально значимую информацию.

Для проверки правильности определенного коэффициента использовалась метрика полноты, точности F_1-мера, описанные в § 4.2. В результате проведенного эксперимента проанализированы 1064 ранее не используемых документов тестового корпуса, из которых 520, заранее определены, как относящиеся к

криминально значимой информации и 544 текстов других тематических направлений.

В таблице 4.6 показан фрагмент таблицы оценки качества предложенной технологии определения семантической близости узкоспециализированных текстов.

Таблица 4.6 – фрагмент оценки качества
технологии определения семантической близости текстов

Имя файла	априор. инфор.	max (sim_{cos})	min (sim_{cos})	Average (sim_{cos})	Вывод системы
lenta.kz_2018-11-09_51.txt	Некрим.	0,65	0,19	0,42	Некрим.
lenta.kz_2018-11-09_52.txt	Крим	0,71	0,12	0,49	Некрим.
caravan.kz_2018-10-02_17.txt	Крим.	0,81	0,22	0,59	Крим.
lenta.kz_2018-11-09_53.txt	Некрим.	0,78	0,14	0,52	Некрим.
caravan.kz_2018-10-02_19.txt	Крим.	0,82	0,25	0,67	Крим.
caravan.kz_2018-10-02_20.txt	Крим.	0,83	0,23	0,65	Крим.
lenta.kz_2018-11-09_54.txt	Некрим	0,72	0,21	0,53	Крим.
lenta.kz_2018-11-09_55.txt	Некрим.	0,61	0,17	0,41	Некрим.
lenta.kz_2018-11-09_56.txt	Некрим.	0,79	0,22	0,43	Некрим.
zakon.kz _2018-08-31_20.txt	Крим.	0,87	0,36	0,71	Крим.
zakon.kz _2018-08-31_16.txt	Крим.	0,80	0,47	0,69	Крим.
zakon.kz _2018-08-31_16.txt	Крим.	0,77	0,41	0,67	Крим.
zakon.kz _2018-08-31_18.txt	Крим.	0,71	0,41	0,61	Крим.
zakon.kz _2018-07-01_19.txt	Крим	0,65	0,47	0,58	Крим.
lenta.kz_2018-11-08_20.txt	Крим.	0,73	0,38	0,59	Крим.
lenta.kz_2018-11-09_57.txt	Некрим.	0,61	0,15	0,44	Некрим.
lenta.kz_2018-11-09_58.txt	Крим.	0,66	0,25	0,48	Некрим.
lenta.kz_2018-11-09_59.txt	Некрим.	0,61	0,26	0,46	Некрим.
lenta.kz_2018-11-09_60.txt	Крим.	0,79	0,40	0,62	Крим.
nur.kz _2018-09-04_16.txt	Крим.	0,88	0,32	0,70	Крим.
nur.kz_2018-11-09_17.txt	Крим.	0,85	0,41	0,72	Крим.
lenta.kz_2018-11-09_18.txt	Крим.	0,78	0,49	0,69	Крим.
lenta.kz_2018-11-09_19.txt	Крим.	0,84	0,43	0,71	Крим.
lenta.kz_2018-11-09_15.txt	Крим.	0,87	0,41	0,71	Крим.
nur.kz _2018-09-04_74.txt	Некрим.	0,57	0,21	0,37	Некрим.
nur.kz _2018-11-09_62.txt	Некрим.	0,70	0,24	0,54	Крим.
nur.kz_2018-11-09_63.txt	Некрим.	0,51	0,17	0,39	Некрим.
nur.kz_2018-11-09_64.txt	Некрим.	0,64	0,23	0,43	Некрим.
lenta.kz_2018-11-09_65.txt	Некрим.	0,53	0,19	0,37	Некрим.

Результат проведенного эксперимента можно представить в таблице 4.7, содержащей значения показателей, описанных в таблице 4.1.

Таблица 4.7 – Результаты эксперимента определения семантической близости документа к узкоспециализированной тематике.

tp = 512	fp =36
fn =8	tn =518

Основываясь на формулах (4.2), (4.3) и (4.5), определяем полноту, точность и F-меру разработанной технологии определения семантической близости документа к узкоспециализированной области (на примере корпуса криминально окрашенных текстов).

precision	*recall*	F_1*-мера*
93,4%	98,5%	95,95%

При полученной средней гармонической семантической близости документа к узкоспециализированной тематике F_1*-мера* ≈96%, полнота отнесения документа к узкоспециализированной криминально тематике 98,5% выше, чем точность 93,4%. Что имеет практическое значение, заключающееся в том, что при решении конкретной задачи выделения криминально значащих текстов, лучше иметь ошибку первого типа, создающую избыточность криминально значащих документов, чем ошибку второго рода, пропускающую криминально-окрашенные тексты.

Выводы по четвертому разделу

1. Разработан выровненный казахско-русский корпус, базирующийся на текстах четырех Казахстанских новостных сайтов *zakon.kz, caravan.kz, lenta.kz* и *nur.kz,* периода май – декабрь 2018 года. Объем корпуса около 50410 слов. Включенные в корпус тексты отражают новости о таких криминальных деяниях как грабежи, угоны машин, убийства, ДТП и др., и образуют базу корпуса криминально окрашенных текстов, используемого для экспериментального оценивания эффективности разработанных в исследовании моделей и технологий.
2. Обосновано использование метрики числовых оценок, использующей в качестве объективно измеряемых показателей эффективности моделей машинного обучения кортеж, включающий коэффициенты полноты, точности и меру Ван Ризбергена. Выбранный метод оценки эффективности базируется на методе тестовых коллекций, который в ML моделях заключается в сравнении результатов работы исследуемой схемы на заранее определенных данных с, так называемым, “золотым стандартом”.
3. Показаны практические результаты реализации разработанной модели Open IE. Разработанная модель реализована на трех корпусах русского, казахского и английского текстов. Общий объем построенных корпусов: 6000 текстов из

700 000 слов, 275 тыс. из них в корпусе русских текстов, 225 тыс. в корпусе казахских текстов и около 200 тыс. в корпусе англоязычных текстов. Точность извлечения триплета факта для английского языка представляет 87,2%, для русского языка 82,4% и для казахского 71,0%.

4. Разработана информационная технология определения семантической близости текстов к заданной узкоспециализированной тематике, базирующаяся на предложенных в диссертационном исследовании методах машинного обучения, методе определения семантической близости многоязычных текстовых документов, базирующемся на VSM и вычислении максимального, минимального и среднего значения косинусного сходства входного документа с документами корпуса.
5. Построена модель оценки качества технологии определения семантической близости документа к узкоспециализированной тематике. Проведенное экспериментальное исследование на корпусе текстов криминальной направленности позволило определить показатели качества: коэффициент полноты recall = 0,985, коэффициент точности precision = 0,934, сбалансированную мере Ван Ризбергена F = 0,9595.

ЗАКЛЮЧЕНИЕ

В результате разработки и исследования моделей методов и алгоритмов интеллектуального анализа многоязычной текстовой информации, использующих методы машинного обучения были получены следующие результаты.

– Проанализировано современное состояние проблемы автоматической обработки многоязычной текстовой информации, осуществлена систематизация существующих лингвистических ресурсов и систем обработки казахского языка и сформированы основные требования к разработке системы анализа многоязычной текстовой информации на основе методов машинного обучения.

– Исследованы существующие сложности и проблемы формализации, как грамматики, так и семантики казахского языка. Показаны особенности автоматической обработки казахского языка, вытекающие из его принадлежности к агглютинативной группе языков. Рассмотрены формализмы языковой системы, выделение которых необходимо для построения модели извлечения знаний их текста и представления их в явном виде триплета факта.

– Исследованы существующие направления реализации методов и моделей NLP. Рассмотрены и классифицированы существующие подходы, используемые в приложениях автоматической обработки текстов. Проанализированы методы Machine Learning, используемые при обработке текстовой информации и возможность их реализации в приложениях NLP, связанных с семантической обработкой.

– Обоснован выбор математического аппарата алгебры конечных предикатов для моделирования процессов интеллектуальной обработки многоязычной текстовой информации. Рассмотрены основы инструментария алгебры предикатов и предикатных операций применительно к его использованию для формализации конструкций естественных языков, идентифицирующих отношения между участниками действия в предложении.

– Разработана и исследована логико-лингвистическая модель Open IE, описывающая семантические функции партиципантов предложения через отношения грамматических и семантических характеристик слов предложений заданного языка. Использование модели позволяет извлекать факты из многоязычной текстовой информации. Модель адаптирована для текстов английского и русского языков. Показана возможность адаптации модели для автоматической генерации структурированной машиночитаемой информации из текстов казахского языка. При моделировании казахского языка вводились предикатные переменные, характеризующие такие грамматические и семантические характеристики как место анализируемого слова во фразе, наличие или отсутствие вспомогательного глагола во фразе, грамматические падежи и склонения существительных и другие.

– Получен алгоритм перефразирования факта побуждения к действию в предложениях английского языка, на базе разработанной математической модели извлечения фактов из многоязычной текстовой информации, и ее адаптации для текстов английского языка. Алгоритм представляет собой использование регулярных выражений грамматических конструкций представления факта побуждения к действию, в которых в качестве алфавита используются метки POS-

тегинга.

– Разработан алгоритм семантической разметки текстов казахского языка. Корпус размечается тегами, определяющими семантическое значение токена в триплете факта <*Sub*> , <*Obj*>, <*Pred*> и тегами <*POS type* = "*crime*">, где тег POS определяют грамматическую информацию части речи или синтаксическую единицу предложения, а значение атрибута type="crim" определяет семантическое значение криминальной составляющей токена. Первичная разметка корпуса казахского языка, на базе которого, осуществляется обучение, базируется на использовании списка суффиксов и лингвистических правил.

– Разработан метод вероятностной морфологической разметки, использующий скрытую Марковскую модель (HMM). Используемая в методе функция оценки вероятности цепочки тегов зависит от двух вероятностей: условной вероятности последовательности тегов и условной вероятности обозначения токена данным тегом. На рассмотренном примере последовательности слов английского предложения показана высокая точность предложенного метода.

– Разработан метод определения семантической близости многоязычных текстовых документов, базирующийся на VSM. Используя идею представления каждого документа в коллекции текстов в виде точки векторного пространства, и дистрибутивную гипотезу статистической семантики, утверждающую, что частоты слов в документе чаще всего определяют семантическую направленность документа, мы показали, что точки, расположенные близко друг к другу в данном пространстве, соответствуют семантически близкими документам, а точки, которые расположены далеко друг от друга, соответствуют сильно отличающимся по смыслу документам.

– Разработан метод, определения семантической близости многоязычных текстовых документов, который вычисляет косинусное сходство между двумя векторами документов, используя в качестве значений координат векторов меру PPMI, представляющую весовую функцию, которая добавляет больший вес значениям координат вектора, подтверждающим семантическую зависимость между документом и термином, и присваивает нулевые значения координатам, если появление термина в документе не имеет никакого семантического значения и, следовательно, не несет никакой информации.

– Разработан выровненный казахско-русский корпус, базирующийся на текстах четырех Казахстанских новостных сайтов *zakon.kz, caravan.kz, lenta.kz* и *nur.kz*, периода май – декабрь 2018 года. Объем корпуса около 50410 слов. Включенные в корпус тексты отражают новости о таких криминальных деяниях как грабежи, угоны машин, убийства, ДТП и др., и образуют базу корпуса криминально окрашенных текстов, используемого для экспериментального оценивания эффективности разработанных в исследовании моделей и технологий.

– Исследовано и обосновано использование метрики числовых оценок, использующей в качестве объективно измеряемых показателей эффективности моделей машинного обучения кортеж, включающий коэффициенты полноты, точности и меру Ван Ризбергена. Выбранный метод оценки эффективности базируется на методе тестовых коллекций, который в ML моделях заключается в

сравнении результатов работы исследуемой схемы на заранее определенных данных с, так называемым, “золотым стандартом”.

– Разработана программная реализация полученной модели Open IE. Модель реализована на трех корпусах русского, казахского и английского текстов. Общий объем построенных корпусов: 6000 текстов из 700 000 слов, 275 тыс. из них в корпусе русских текстов, 225 тыс. в корпусе казахских текстов и около 200 тыс. в корпусе англоязычных текстов. Точность извлечения триплета факта для английского языка представляет 87,2%, для русского языка 82,4% и для казахского 71,0%.

– Разработана информационная технология определения семантической близости текстов к заданной узкоспециализированной тематике, базирующаяся на предложенном в диссертационном исследовании методе машинного обучения. Метод заключается в определении семантической близости многоязычных текстовых документов, базирующемся на VSM и вычислении максимального, минимального и среднего значения косинусного сходства входного документа с документами корпуса.

– Разработана методика оценки качества технологии определения семантической близости документа к узкоспециализированной тематике. Проведенное экспериментальное исследование на корпусе текстов криминальной направленности позволило определить показатели качества работы системы: коэффициент полноты recall = 0,985, коэффициент точности precision = 0,934, сбалансированная мера Ван Ризбергена F = 0,9595.

Таким образом, поставленные в диссертационной работе задачи исследования полностью решены. Программная реализация и проведенная оценка качества работы системы, извлекающей знания в виде триплета фактов из текста, и системы, определяющей семантическую близость текста к узкоспециализированной тематике, базирующихся на разработанных в диссертационном исследовании моделях и алгоритмах, показывает хорошие результаты.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Б.А.Жетписбаева, О. Т. Аринова. От идеи «Триединство языков» Н.А.Назарбаева до полиязычного образования в Казахстане / Вестник КарГУ, Серия «Педагогика». № 4(68)/2012. – С. 19-22.
2. Назарбаев Н.А. Социальная модернизация Казахстана: Двадцать шагов к Обществу Всеобщего Труда // Казахстан-ская правда. — 2012. — № 218-219. — 10 июля.
3. Бектаев К. Теория вероятностей и математическая статистика на казахском языке. – Алматы: Атамура, 1990. – 124 с.
4. Бектаев К. и др. Математические методы в языкознании. Алматы: Наука, 1974. – 170 с.
5. Бектаев К. Большой казахско-русский русско-казахский словарь. – Алматы: Алтын Казына, 1999. – 703 с.
6. Gulshat Kessikbayeva, Ilyas Cicekli. Rule based morphological analyzer of Kazakh language / Proceedings of the 2014 Joint Meeting of SIGMORPHON and SIGFSM. – 46-54 p.
7. В. Б. Барахнин, Л. Х. Лукпанова, А. А. Соловьев. Алгоритм построения словоформ с использованием флективных классов для систем морфологического анализа казахского языка / Вестник Новосибирского государственного университета. Серия: Информационные технологии, 2014. – Т.12.– № 2. – С. 25-32.
8. Тукеев У.А., Тургынова А. Морфологический анализ казахского языка на основе полной системы окончаний / Труды международной конференции по компьютерной и когнитивной лингвистике. Сер. "Интеллект. Язык. Компьютер" 2016. – С. 225- 231.
9. Мамырбаев О. Ж., Хайрова Н. Ф., Мухсина К. Ж. Қазақ тіліндегі мәтіндердегі қылмыстық мәнді коллакцияларды анықтау / Вестник КазАТК им. М. Тынышпаева, рекомендуемый ККСОН МОН РК. – №3(110). – 2019. – 170 -175 с.
10. Мадиева Г. Б. , Уматова Ж. М. Об алматинском корпусе казахского языка. // Вестник КазНУ. Серия филологическая. – 2015. №5 (157). – С. 98 -103.
11. Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, Vít Suchomel. The Sketch Engine: Ten Years On // Lexicography. Springer Berlin Heidelberg. – 2014 (1). – P. 7-36.
12. Chapaev D., Turapbekov B. Building Kazakh language open source corpora using wikipedia resources/ Suleyman Demirel University Bulletin. – Kaskelen, 2018. – P. 153- 160.
13. Olzhas Makhambetov, Aibek Makazhanov, Zhandos Yessenbayev, Bakhyt Matkarimov, Islam Sabyrgaliyev, Anuar Sharafudinov. Assembling the Kazakh Language Corpus. // Proceedings of the Conference on Empirical Methods in Natural Language Processing.–2013. – P. 1022-1033.
14. Jörg Tiedemann. Parallel data, tools and interfaces in OPUS // In Proceedings of the Eight International Conference on Language Resources and Evaluation. 2012 (LREC’12). – P. 2214–2218.

15. Katarzyna Niżałowska, Markowska-Kaczmar. A Language-Independent Method for Detection and Correction of Alignment Errors in Parallel Corpora. // Natural Language Processing and Information Systems. –2015. – P.335-346.
16. D Rakhimova, Z Zhumanov. Complex Technology of Machine Translation Resources Extension for the Kazakh Language / Asian Conference on Intelligent Information and Database Systems, Springer, 2017, 297-307 p.
17. Miquel Esplá-Gomis, Mikel L Forcada. Bitextor, a free/open-source software to harvest translation memories from multilingual websites // Proceedings of MT Summit XII, Ottawa, Canada. Association for Machine Translation in the Americas. – 2009.
18. Zhandos Zhumanov, Aigerim Madiyeva, Diana Rakhimova. New Kazakh Parallel Text Corpora with On-line Access / Conference on Computational Collective Intelligence Technologies and Applications, Springer, 2017. – P. 501-508.
19. Dániel Varga, Péter Halácsy, András Kornai, Viktor Nagy, László Németh, Viktor Trón. Parallel corpora for medium density languages. / Amsterdam Studies In The Theory And History Of Linguistic Science. – 2007. – S. 4. – V. 292. – 247 p.
20. Vondricka, Pavel. Aligning parallel texts with intertext // Proceedings of the Ninth International Conference on Language Resources and Evaluation // Reykjavik, Iceland. European Language Resources Association. – 2014. – P. 1875-1879.
21. Tukeyev U.A., Rakhimova D.R., Zhumanov Zh.M., Kartbayev A.Zh. Single state transducer model for Kazakh and Russian morphology / KazNU Bulletin. Mathematics, Mechanics, Computer Science Series №2(89) 2016. – P. 110-117.
22. Тукеев У. А. Жуманов Ж. М., Рахимова Д. Р. Моделирование семантических ситуаций времен казахского языка при машинном переводе // Вестник КазНУ. Серия математика, механика, информатика. 2012. № 4 (75). С. 99–107.
23. История разработки пакетов программ поддержки казахского языка в системе Microsoft Windows [Электрон. ресурс.]. – URL: http://www.izet.narod.ru/history.htm (дата обращения: 16.02.2019)
24. SANASOFT CO. linguistic technologies [Электрон. ресурс.]. – URL: http://www.sanasoft.kz/c/ru (дата обращения: 17.02.2019)
25. PROMT Neural Translation Server [Электрон. ресурс.]. – URL: https://www.promt.ru/ (дата обращения: 12.02.2019)
26. Mikel L Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, Francis M Tyers. Apertium: a free/open-source platform for rule-based machine translation. // Machine translation. – 2011. 25(2). – P. 127-144.
27. Автоматизированный перевод от Soylem [Электрон. ресурс]. – URL: https://soilem.kz/ (дата обращения: 16.02.2019)
28. Y Amirgaliyev, M Hahn, T Mussabayev. The speech signal segmentation algorithm using pitch synchronous analysis / Open Computer Science. – 2017. – V. 7. – № 2. – P. 1-8
29. Bagher Baba Ali, Waldemar Wojcik, Orken Mamyrbayev, Mussa Turdalyuly, Nurbapa Mekebayev. Speech recognizer-based non-uniform spectral compression for

robust MFCC feature extraction / Przeglad Elektrotechniczny. – 2018. – V. 94. – № 6. – P. 90-93.
30. Orken Mamyrbayev, Mussa Turdalyuly, Nurbapa Mekebayev, Keylan Alimhan, Aizat Kydyrbekova, Tolganay Turdalykyzy. Automatic Recognition of Kazakh Speech Using Deep Neural Networks / Asian Conference on Intelligent Information and Database Systems, 2019. – P.465-474.
31. Современный казахский язык. Фонетика и морфология [Текст] / АН Казах. ССР. Ин-т языкознания; редкол.: М. Б. Балакаев, Н. А. Баскаков, С. К. Кенесбаев. - Алма-Ата : Изд-во АН Казах. ССР, 1962. - 453 с.
32. Petrasova S., Khairova N., Lewoniewski W., Mamyrbayev O., and Mukhsina K. Similar Text Fragments Extraction for Identifying Common Wikipedia Communities. Data. – 2018, Vol.3, №4. – P. 66. doi:10.3390/data3040066
33. Socher R, Bengio Y, Manning C. D. Deep learning for NLP (without magic) // Tutorial Abstracts of ACL 2012. Association for Computational Linguistics, – 2012, p. 5-6.
34. Philipp Cimiano, Andreas Hotho, Steffen Staab. Learning concept hierarchies from text corpora using formal concept analysis // Journal of artificial intelligence research. 2005, V.24, p. 305-339
35. Landauer T.K., McNamara D. S., Dennis S., Kintsch W. Handbook of latent semantic analysis. Psychology Press, 2013, 544 P.
36. Thomas François, Eleni Miltsakaki. Do NLP and machine learning improve traditional readability formulas? // Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations. Association for Computational Linguistics. – 2012, p. 49–57
37. Khairova, N.; Kolesnyk, A.; Mamyrbayev, O. and Mukhsina, K. (2019). The Influence of Various Text Characteristics on the Readability and Content Informativeness. In Proceedings of the 21st International Conference on Enterprise Information Systems - Volume 1: ICEIS, ISBN 978-989-758-372-8, pages 462-469. DOI: 10.5220/0007755004620469.
38. Bloedorn E., Mani I. Using NLP for Machine Learning of User Profiles.// Intelligent Data Analysis, 1998, vol.2 № 1, p. 3-18
39. Jatana, N., Sharma, K. Bayesian Spam Classification: Time Efficient Radix Encoded Fragmented Database Approach. In: 2014 International Conference on Computing for Sustainable Global Development (INDIACom), 939-942 (2014)
40. Aiwu, L., Hongying, L. Utilizing Improved Bayesian Algorithm to Identify Blog Comment Spam. In: IEEE Symposium on Robotics and Applications (ISRA), pp. 423-426 (2012) .
41. Manne, S., Kotha, S. K., Fatima, S. Text Categorization with k-nearest neighbor approach . In: Proceedings of the International Conference on Information Systems Design and Intelligent Applications, vol.132, 413 – 420 (2012)
42. Narayanan V, Arora I, Bhatia A. Fast and accurate sentiment classification using an enhanced Naive Bayes model. International Conference on Intelligent Data Engineering and Automated Learning / Springer, Berlin, Heidelberg, 2013, p. 194-201.

43. Hassan S, Rafi M, Shaikh M. S. Comparing svm and naive bayes classifiers for text categorization with wikitology as knowledge enrichment // 2011 IEEE 14th International Multitopic Conference. – 2011, p. 31-34
44. Chakrabarti S., Roy S., and Soundalgekar M.V., “Fast and accurate text classification via multiple linear discriminant projection”, The VLDB Journal The International Journal on Very Large Data Bases, 2003, pp. 170–185.
45. Порохнин А. А. Анализ статистических методов снятия омонимии в текстах на русском языке. Вестн. Астрахан. гос. техн. ун-та. Сер. управление, вычисл. техн. информ., 2013, № 2, с. 168–174.
46. Andrews M, Vigliocco G. The hidden Markov topic model: A probabilistic model of semantic representation // Topics in Cognitive Science. 2010, Vol. 2, № 2, p. 101-113.
47. Чистиков П. Г., Е. А. Корольков Е. А., Таланов А. О., Соломенник А. И. Гибридная технология синтеза речи на основе скрытых марковских моделей и алгоритма unit selection. – ИЗВ. ВУЗов. приборостроение. 2013. Т. 56, № 2, с. 33 – 38.
48. Jiang S., Pang G., Wu M., Kuang L. An improved K-nearest-neighbor algorithm for text categorization // Expert Systems with Applications, 2012, Vol. 39, № 1, p. 1503-1509.
49. Ле Мань Ха. Оптимизация алгоритма KNN для классификации текстов. ТРУДЫ МФТИ. - Информатика, вычисл. техника и управление. – 2016. Том 8, № 1, с. 92 – 94.
50. S Lai, L Xu, K Liu, J Zhao. Recurrent Convolutional Neural Networks for Text Classification. / AAAI, 2015, vol. 333, p. 2267-2273
51. Смирнова О.С., Шишков В.В. выбор топологии нейронных сетей и их применение для классификации коротких текстов / International Journal of Open Information Technologies, 2016, vol. 4, no 8, p. 50- 53.
52. Найденова К. А., Невзорова О. А. Машинное обучение в задачах обработки естественного языка: обзор современного состояния исследований. // Ученые записки Казанского государственного университета. Серия: Физико-математические науки. 2008. Т. 150. № 4. С. 5-24.
53. Флах П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных / пер. с англ. А. А. Слинкина. – М.: ДМК Прогресс, 2015. – 400 с.
54. Михайлов Д. В., Емельянов Г. М. К вопросу автоматизации пополнения базы данных лексических функций в задаче установления смысловой эквивалентности текстов естественного языка. / Вестник Новгородского государственного университета им. Ярослава Мудрого, 2007, № 44, с. 45 -49.
55. Hapke Hannes, Howard Cole, Lane Hobson. Natural language processing in action/ Manning Publications, 2019. – P. 512
56. T Mikolov, K Chen, G Corrado, J Dean. Efficient Estimation of Word Representations in Vector Space. / arXiv preprint arXiv:1301.3781 – 2013
57. Ильвовский Д., Черняк Е. Глубинное обучение для автоматической обработки текстов // Открытые системы. СУБД – 2017. – № 02, С. 26-29 .

58. Christopher Manning. Computational linguistics and deep learning // Computational Linguistics. – 2015. – vol. 41. – № 4. – p. 701-707.
59. Loftsson H. Tagging Icelandic text: A linguistic rule-based approach / Nordic Journal of Linguistics. –2008. – vol. 31. – № 1. – p. 47-72.
60. Захаров В. П. Корпусная лингвистика: учебник для студентов направления «Лингвистика» / В. П. Захаров, С. Ю. Богданова. — [2-е изд., перераб. и доп.]. — Санкт Петербург : СПбГУ. РИО. Филологический факультет, 2013. — 148 с.
61. Hasan F. M., UzZaman N, Khan M. Comparison of different POS Tagging Techniques (N-Gram, HMM and Brill's tagger) for Bangla / Advances and innovations in systems, computing sciences and software engineering. – 2007, – p. 121-126.
62. Màrquez L., Padro L., Rodriguez H. A machine learning approach to POS tagging / Machine Learning. – vol. 39. – № 1. – p. 59-91.
63. McEnery, T. and Hardie, A (2012) Corpus Linguistics: Method, Theory and Practice, Cambridge University Press, Cambridge. 48-52pp.
64. Archer D., McEnery T., Rayson P., Hardie A. Developing an automated semantic analysis system for Early Modern English / Corpus Linguistics 2003 conference. – 2003. – p. 22-31
65. Bird S., Klein E., Loper E. Natural Language Processing with Python. – O'Relly Media, 2009. – P. 501
66. Davies Mark, Fuchs Robert. Expanding horizons in the study of World Englishes with the 1.9 billion word Global Web-based English Corpus (GloWbE) / English World-Wide. – V. 36. – № 1. – 1-28 p.
67. Carvalho L. Translating contracts and agreements: a Corpus Linguistics perspective / Avanĉos da linguística de Corpus no Brasil, 2008. – 333 P.
68. Ландэ Д. В. Интернетика: Навигация в сложных сетях: модели и алгоритмы: моногр. / Д. В. Ландэ, А. А. Снарский , И. В. Безсуднов — М.: Либроком (Editorial URSS), 2009. 264 с.
69. Sint, R. , Schaffert, S., Stroka, S., Ferstl, R. Combining Unstructured, Fully Structured and Semi-Structured Information in Semantic Wikis. In: Proceedings of the 4th Semantic Wiki WorkShop (SemWiki) at the 6th European Semantic Web Conference, ESWC (2009).
70. Baeza-Yates R., Ribeiro-Neto B. Modern Information Retrieval. — Addison-Wesley, 1999. — 340 p.
71. Crestan, E., Pantel P. Web-Scale Knowledge Extraction from Semi-Structured Tables. In: WWW '10 Proceedings of the 19th international conference on World wide web, 1081 – 1082 (2010)
72. Gatterbauer, W.; Bohunsky, P.; Herzog, M.; Krupl, B.; and Pollak, B. Towards Domain-Independent Information Extraction from Web Tables. In Proceedings WWW-07. pp. 71–80. Banff, Canada. (2007).
73. Wong; Y. W., Widdows, D.; Lokovic T.; Nigam K. Scalable Attribute-Value Extraction from Semi-structured Text. In: 2009 IEEE International Conference on Data Mining Workshops, 302 –307 (2009)

74. Phillips, W., Riloff, E. Exploiting Strong Syntactic Heuristics and Co-Training to Learn Semantic Lexicons. In: Proceedings of the conference on Empirical Methods in Natural Language Processing (EMNLP) (2002)
75. Jones, R., Ghani, R., Mitchell, T., Riloff, E. Active Learning with Multiple View Feature Sets. In: ECML 2003 Workshop on Adaptive Text Extraction and Mining, (2003).
76. Agichtein, E., Gravano, L. Snowball: Extracting Relations from Large Plaintext Collections. In: Proceedings of the 5th ACM International Conference on Digital Libraries, 85–94, San Antonio, Texas, (2000).
77. ARPA. Proceedings of the 3rd Message Undestanding Conference. – 1991.
78. Etzioni, O., Banko, M., Soderland, S., Weld, D. (2008). Open information extraction from the web. Communications of the ACM, Vol. 51 No. 12 (pp. 68-74). New York, NY, USA.
79. Fader, A., Soderland, S., Etzioni, O. (2011). Identifying relations for open information extraction. Proceedings of the conference on empirical methods in natural language processing (pp. 1535-1545). Edinburgh, Scotland, UK.
80. Duc-Thuan Vo, Ebrahim Bagheri. Open information extraction. Encyclopedia with Semantic Computing and Robotic intelligence. – Vol. 1, No. 1 (2016).
81. Starostin A. S., Bocharov V. V., Alexeeva S. V. et al. FactRuEval 2016: evaluation of named entity recognition and fact extraction systems for Russian. In: Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2016”, pp. 702-720. 2016.
82. Ludovic, L., Gallinari, P. Bayesian network model for semi-structured document classification. Information Processing and Management: an International Journal - Special issue: Bayesian networks and information retrieval, vol.40, 807 –827 (2004)
83. Rish, I.: An Empirical Study of the Naive Bayes Classifier. In: Proceedings of IJCAI-01 Workshop on Empirical Methods in Artificial Intelligence (2001).
84. Jatana, N., Sharma, K. Bayesian Spam Classification: Time Efficient Radix Encoded Fragmented Database Approach. In: 2014 International Conference on Computing for Sustainable Global Development (INDIACom), 939-942 (2014)
85. Aiwu, L., Hongying, L. Utilizing Improved Bayesian Algorithm to Identify Blog Comment Spam. In: IEEE Symposium on Robotics and Applications(ISRA), pp. 423-426 (2012) .
86. Joachims, T. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In: ECML '98 Proceedings of the 10th European Conference on Machine Learning. Springer-Verlag London, UK, 137– 142 (1998)
87. Kleinbaum, D. G., Klein, M., Pryor, E. R. Logistic Regression: A Self-Learning Text. New York: Springer, 2002.
88. Baoli, L., Shiwen, Y., Qin L. An Improved k-Nearest Neighbor Algorithm for Text Categorization. In: The 20th International Conference on Computer Processing of Oriental Languages, Shenyang, China, 2003.
89. Manne, S., Kotha, S. K., Fatima, S. Text Categorization with k-nearest neighbor approach . In: Proceedings of the International Conference on Information Systems Design and Intelligent Applications, vol.132, 413 – 420 (2012)

90. Entezari-Maleki, R., Rezaei, A., Minaei-Bidgoli, B. Comparison of Classification Methods Based on the Type of Attributes and Sample Size. Journal of Convergence Information Technology (JCIT) 4 (3), 94 – 102 (2009).
91. Keiji Shinzato and Satoshi Sekine. Unsupervised extraction of attributes and their values from product description. In Sixth International Joint Conference on Natural Language Processing, IJCNLP 2013, pages 1339–1347, 2013.
92. Liyuan Liu, Xiang Ren, Qi Zhu, Shi Zhi, Huan Gui, Heng Ji, Jiawei Han. Heterogeneous Supervision for Relation Extraction: A Representation Learning Approach. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing.– p. 46—56.
93. Xuan Wang, Yu Zhang, Yinyin Chen. A Survey of Truth Discovery in Information Extraction. Published 2018.
94. Khairova, N., Lewoniewski, W., Wecel, K. Estimating the Quality of Articles in Russian Wikipedia Using the Logical-Linguistic Model of Fact Extraction. Conference proceedings. BIS 2017. Part of the Lecture Notes in Business Information Processing book series (LNBIP, volume 288). (pp. 28-40). Poland: Poznan.
95. E. Lex, M. Voelske, M. Errecalde, E. Ferretti, L. Cagnina, C.r Horn, B. Stein, M. Granitzer. Measuring the quality of web content using factual information. Proceedings of the 2nd joint WICOW/AIRWeb workshop on web quality, 2012, 7-10.
96. G Zhou, L Qian, J Fan. Tree kernel-based semantic relation extraction with rich syntactic and semantic information / Information Sciences. – Vol. 180, p. 1313-1325.
97. Luckicgev, S. Graphical Notations for Rule Modeling. In: Giurca, A., Gašević, D., Taveter, K. Handbook of Research on Emerging Rule-Based Languages and Technologies: Open Solutions and Approaches, Hershey, New York., vol.1, 76– 98 (2009)
98. Khairova N., Lewoniewski W., Węcel K., Mamyrbayev O., Mukhsina K. (2018) Comparative Analysis of the Informativeness and Encyclopedic Style of the Popular Web Information Sources. In: Abramowicz W., Paschke A. (eds) Business Information Systems. BIS 2018. Lecture Notes in Business Information Processing, vol 320. Springer, Cham DOIhttps://doi.org/10.1007/978-3-319-93931-5_24
99. Pablo Gamallo, Marcos Garcia, and Santiago Fernandez-Lanza. 2012. Dependency-based open ´ information extraction. In Proceedings of the Joint Workshop on Unsupervised and Semi-Supervised Learning in NLP, pages 10–18.
100. Alan Akbik and Alexander Loser. 2012. Kraken: N- ¨ ary facts in open information extraction. In Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction, pages 52–56.
101. Anthony Fader, Stephen Soderland, and Oren Etzioni. Identifying relations for open information extraction. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2011. pages 1535–1545.
102. Gabor Angeli, Melvin Johnson Premkumar, and Christopher D Manning. 2015. Leveraging linguistic structure for open domain information extraction. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics, 2015. – p. 344-354.

103. Kiril Gashteovski, Rainer Gemulla, Luciano Del Corro. MinIE: Minimizing Facts in Open Information Extraction. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2017. – p. 2630-2640
104. Mooney, R. J., Bunescu R. Mining Knowledge from Text Using Information Extraction. In: Newsletter. ACM SIGKDD Explorations Newsletter - Natural language processing and text mining, vol.7, issue 1, 3–10 (2005)
105. Joakim Nivre et al. Universal Dependencies v1: A Multilingual Treebank Collection. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Paris, France, May. European Language Resources Association (ELRA).
106. Yuen-Hsien Tseng, Lung-Hao Lee, Shu-Yen Lin, Bo-Shun Liao, Mei-Jun Liu, Hsin-Hsi Chen, Oren Etzioni, Anthony Fader. Chinese open relation extraction for knowledge acquisition. In Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers, 2014. – p. 12-16.
107. Pablo Gamallo, Marcos Garcia. Multilingual Open Information Extraction. In Portuguese Conference on Artificial Intelligence, 2015. – 711-722 p.
108. Андреев А. М. Модель извлечения фактов из естественно-языковых текстов и метод ее обучения / А. М. Андреев, Д. В. Березкин, К. В. Симаков // Электронные библиотеки: перспективные методы и технологии, электронные коллекции : труды VIII Всероссийской научной конференции RCDL'2006. – Ярославль : Ярославский государственный университет им. П. Г. Демидова, 2006. – С. 252-261.
109. Петров Э. Г. Компараторная структурно-параметрическая идентификация моделей скалярного многофакторного оценивания : монография / Э. Г. Петров, В. В Крючковский. – Херсон : Олди-плюс, 2009. – 294 с
110. Abed Thamer Khudhair. The intelligence theory mathematical apparatus formal BASE / Advanced Information Systems. 2017. Vol. 1, No. 1, pp. 38 -43.
111. Бондаренко М. Ф. Мозгоподобные структуры: Справочное пособие. / М. Ф. Бондаренко, Ю. П. Шабанов-Кушнаренко. Том первый. – Київ : Наукова думка, 2011. – 460 с.
112. Бондаренко М. Ф., Шабанов-Кушнаренко Ю. П. Теория интеллекта / Харьков: Комп. СМИТ, 2007. – 576 с.
113. Шаронова Н. В., Хайрова Н. Ф. Использование метода компараторной идентификации для разбиения семантического пространства предметной области знание-ориентированных систем // Вестн. Херсон. гос. техн. ун-та. – Херсон, 2007. – № 4 (27). – С. 39–42
114. МФ Бондаренко, НП Кругликова, ИА Лещинская, НЕ Русакова, ЮП Шабанов-Кушнаренко. Об алгебре предикатов / БИОНИКА ИНТЕЛЛЕКТА. 2010. № 3 (74). С. 3–7
115. Шабанов-Кушнаренко Ю. П. Теория интеллекта: технические средства: моногр. / Ю. П. Шабанов-Кушнаренко — Харьков: Выща школа, 1986.— 134 с.
116. Хайрова Н. Ф. Використання логіко-алгебраїчної моделі семантичних відмінків для семантичного аналізу речення / Збірник наукових праць Військового

інституту Київського національного університету імені Тараса Шевченка. – К.: ВІКНУ, 2012.- Вип. № 38. – С. 239 – 245.
117. Мамырбаев О. Ж., Мухсина К. Ж., Хайрова Н. Ф., Колесник А. С. Лингвистические инструменты выявления криминально окрашенной текстовой информации веб-контента // Вестник казахстанского-британского технического университета, рекомендуемый ККСОН МОН РК, – 2018. – № 15(3). – С. 112-117.
118. Khairova Nina, Petrasova Svitlana, Lewoniewski Wlodzimierz, Orken Mamyrbayev, Kuralai Mukhsina. Automatic Extraction of Synonymous Collocation Pairs from a Text Corpus. FedCSIS 2018: Proceedings of the Federated Conference on Computer Science and Information Systems DOI: 10.15439/2018F186 ISSN 2300-5963 ACSIS, Vol. 15 pp. 485–488
119. Мамырбаев О. Ж., Мухсина К. Ж. Мәтін үндесітілігін анықтауға арналған қолданыстағы жүйелерді талдау / Вестник национальной академии наук РК", рекомендуемый ККСОН МОН РК, №5, (315) 2017
120. Khairova N., Lewoniewski W., Węcel K., Mamyrbayev O., Mukhsina K. (2018) Comparative Analysis of the Informativeness and Encyclopedic Style of the Popular Web Information Sources. In: Abramowicz W., Paschke A. (eds) Business Information Systems. BIS 2018. Lecture Notes in Business Information Processing, vol 320. Springer, Cham DOI https://doi.org/10.1007/978-3-319-93931-5_24
121. N. Khairova, S. Petrasova, O. Mamyrbayev, K. Mukhsina Detecting Collocations Similarity via Logical-Linguistic Model / RELATIONS-Workshop on meaning relations between phrases and sentences, 2019, p.15-22
122. Fillmore, C. J. Verbs of Judging: An Exercise in Semantic Description // Studies in Linguistic Semantics. N.Y. etc., 1971. ~ P. 273-290.
123. Филлмор Ч Фреймы и семантика понимания. - В кн.: Новое в зарубежной лингвистике, вып. XXШ. - М., 1988
124. Хайрова Н. Ф. , Мамырбаев О. Ж., Мухсина К. Ж., Колесник А.С. Автоматическая генерация структурированной машинно-читаемой информации из мультиязычных текстов / Информатика и прикладная математика: Мат. IV межд. науч. конф. (25-29 сентября). Часть 2. – Алматы, 2019. – с. 509 – 519
125. Новая философская энциклопедия: в 4 т. / Институт философии РАН; Национальный общественно-научный фонд. — М.: Мысль, 2010. — ISBN 978-5-244-01115-9.
126. Хайрова Н. Ф., Мамырбаев О. Ж., Мухсина К. Ж., Пилипенко А. А. Моделирование грамматических способов выражения семантики факта в английском предложении. // Матер. III Междунар. науч. конф. «Информатика и прикладная математика», посв. 80-летнему юбилею проф. Бияшева Р.Г. и 70-летию проф. Айдарханова М.Б. – Алматы, 2018. – Т. 2. – С. 136–144
127. Гендина Н. И., Информационно-поисковые тезаурусы: основные виды и области применения // Научные и технические библиотеки. – М.: Государственная публичная научно-техническая библиотека России, 2008 – С. 5-14.
128. T. Pedersen, S. Patwardhan, and J. Michelizzi, “WordNet: Similarity - Measuring the Relatedness of Concepts” // Proceedings of Demonstration Papers at HLT-NAACL. – 2004. - P. 38–41.

129. DependenSee: A Dependency Parse Visualisation/Visualization Tool // http://chaoticity.com/dependensee-a-dependency-parse-visualisation-tool/ [електронный ресурс] : 15.04.18
130. Базарбаева З.М. корпус казахского языка: просодическая разметка // Международный журнал прикладных и фундаментальных исследований. – 2016. – № 11 (часть 2) – С. 334-337
131. Juang B. H., Rabiner L. R. Hidden Markov Models for Speech Recognition / Technometrics, 1991. –Vol. 33. – No. 3. p. 251-272
132. Turnay P.D., Pantel P. From frequency to meaning: Vector Space Models of Semantics / Journal of Artificial Intelligence Research 37 (2010), p. 141-188.
133. Мамырбаев О.Ж., Мухсина К. Ж. Анализ текстовых сообщений с применением векторной формы / Математический методы и информационные технологии макроэкономического анализа и экономической политики: Мат. межд. науч. конф. – Алматы, 2017.
134. Furnas, G. W., Landauer, T. K., Gomez, L. M., & Dumais, S. T. (1983). Statistical semantics: Analysis of the potential performance of keyword information systems. Bell System Technical Journal, 62 (6), 1753–1806.
135. Salton, G., Wong, A., & Yang, C.-S. (1975). A vector space model for automatic indexing. Communications of the ACM, 18 (11), 613–620.
136. Солтон Дж. Динамические библиотечно-информационные системы: Пер. с англ. — М.: Мир, 1979.— 557с.
137. Дональд Кнут. Искусство программирования, том 2. Получисленные алгоритмы = The Art of Computer Programming, vol.2. Seminumerical Algorithms. — 3-е изд. — М.: Вильямс, 2007. — С. 832.
138. Shannon, C. (1948). A mathematical theory of communication. Bell System Technical Journal, 27, 379–423, 623–656.
139. Pantel, P., & Lin, D. (2002). Document clustering with committees. In Proceedings of the 25th Annual International ACM SIGIR Conference, pp. 199–206.
140. Turney, P. D., & Littman, M. L. (2003). Measuring praise and criticism: Inference of semantic orientation from association. ACM Transactions on Information Systems, 21 (4), 315–346
141. Grigori Sidorov, Alexander Gelbukh, Helena Gómez-Adorno, David Pinto. Soft similarity and soft cosine measure: Similarity of features in vector space mode. Computación y Sistemas. –2014.– V.18,– № 3, – pp. 491-504
142. Khairova, N., Kolesnyk, A., Mamyrbayev, O., Mukhsina, K. The aligned Kazakh-Russian parallel corpus focused on the criminal theme .//CEUR Workshop Proceedings. – 2019, p 116-125
143. Шабанов В. И. Метод классификации текстовых документов, основанный на полнотекстовом поиске / В. И. Шабанов , А. М. Андреев // Труды РОМИП’2003. — СПб. : НИИ Химии СПб гос. ун-та, 2003. — С.52—71.
144. Cormack G.V. A Efficient construction of large test collections / G. V. Cormack , C. R. Palmer , C. L. Clarke // Proc. of the SIGIR'98. — P. 282—289.
145. Johnson, H., & Martin, J. (2003). Unsupervised learning of morphology for English and Inuktitut. In Proceedings of HLT-NAACL 2003, pp. 43–45

ПРИЛОЖЕНИЕ А

Метки POS-tagging казахских существительных

Суффиксы и квази-окончания	Метка	Описание
шык, шы, пыр, мпыр, алар, лар, елер, ды, рдан, рлан, рсақ, қтар, ылар, ылык, нші, лік, сшы, пша, хана, ашы, ші, паз, лық, йлар, қсы, ылық, ндық, ім, ар, ас, кер, уші, шілер, рік, ктер, қша, пан, лшы, дыр, тыр, рған, қай, алар, ылар, ңғы, ылар, ырақ, тік, ңдар, лын, ншақ, най, қтар, гер, рлер, ылар, ңіз, зші, шлер, гер, рлер, пкер, рлер, лігі, тур, түрлер, ші, ілер, ншық, ын, шілік, ылық, дар, лық, ылар, шы, тар, гер, герлер, лер, ханалар, ілеп, паз, ік, іктер, керткіш, ту, ірткі, еп, ептер, сіз, уас, керу, ім, імде, башы, елер, пенділер, бек, кқор, шіл, ктер, ағасы, сы, лар, улар, тау	NN at	Атау (*именительный падеж*)
мның, енің, рдың, дың,	NN ілл	Ілік (*родительный падеж*)
да, те, та, нда, нде, ға, ге, қа, ке, на, не, а, е, тік, еге ырға, рға, йға, ыға, аға, шаға, сіз, мға, ға	NN ба	Барыс (*дательно-направленный падеж*)
мды, ені, рды, ырды, тты, ңды, қы	NN та	Табыс (*винительный падеж*)
да, зға, рда, еде,	NN жа	Жатыс (*местный падеж*)
дан, ден, тан, тен, нан, нен, здан, еден, рдан,	NN шы	Шығыс (*исходный падеж*)
бен, ммен, емен, мен, лармен, пен, нен, рмен, тпен, менен, ммен, мен, тармен, ілермен, герлермен, пенен,	NN ко	Көмектес (*творительный падеж*)

ПРИЛОЖЕНИЕ Б

Классификация основных слообразовательных глагольных аффиксов казахского языка

-қыла/ -кіле, ғыла/-гіле, ңқыра/-ңкіре, -іңкіре ңғыра/-ңгіре, мсыра/-мсіре, ымсыра/- імсіре (и их фонетические варианты**), ылда/-ілде, ырла/-ірде**	Аффиксы глагольного вида, прибавляемые к глагольным основам или прибавляются аффиксы категории залога	Есе- **ңгіре-у** (ослабеть)
-ла /-ле (-да /-де, -та /-те во всех фонетических вариантах), -**лан/-лен (дан /-ден, -тан /-тен, лат/-лет), -лас/-лес, -ландыр/-лендір, -ластыр/-лестір (**во всех фонетических вариантах)	Наиболее продуктивные аффиксы Основной глаголообразующий индекс от других частей речи	Көңілсіз-**де-н-у** (становиться грусным) Әлсіз-**де-н-у** (становиться разговорчивым (**н**-суффикс возвратного залога)
-ла /-ле, -а/-е, -лық/-лік,-ық/-ік, -шы/-ші (во всех фонетических вариантах –**іг, -ліг, -тығ, -діг, -ре**), -**ыра**, -**іре, -ыла, -іла**	Суффиксы бывают как именными так и отглагольными	Үр-**ле-у** (дуть, подуть) Тісте-**ле-у** (кусать) Түй-**ре-у** (колоть)
ай/-ей, й, ар/-ер, р, ра/-ре, ыр/-ір,		Көг-**ер-т-у** (синить, зеленить) үлк-**ей-т-у** (увеличить) Кішір-**ей-т-у**(уменьшить) (-т- аффикс понудительного залога)
-зы, -азы, -ма, -бе, -ды/-ді, -ы/-і, ты,/-ті, лы/-лі, ра/-ре, **-шы/-ші, ын, ін, ал, ел**	Мало продуктивные аффиксы	
мала /-меле, -пала,-пеле, -бала/-беле, -ақта/-екте, -дала/-ала		
сыра/-сіре, мсыра/-мсіре, усыра/-усіре, жыра/-жіре, аңғыра/-еңгіре, ңра/-ңре ,ыра/-сіре, аура/-еуре,		

сы/-сі , сын/-сін, са/-се, сан/-сен	**н-**формальный показатель возвратного залога	
-лық/-лік ық/-ік, дық/-дік, тық/-тік,-тығ –лығ –ығ, іғ		Соқ-тық Соқ-**тығ**-у
қар/-кер қыр/-кір ғар/-гер ғыр/-гір қа/-ке ға/-ге қа/-ке, қан/-кен ан/-ен, ғал/-қал	Добавляются аффиксы **–л** и **–н** возвратного залога	
ырқа /-ірке, лқа ырқан /-іркен		
ғы/-гі ,ғыт/-гіт, ға/-ге қы/-кі	Аффикс **–т** принудительного залога	
-т, ыт/-іт		
бы, бі, пы, пі		
сый, си, ырай, ірей, ыс, іс, жи, ши	Образные глаголы, которые различных движений, обозначают так же изменения во внешности и мимике субъекта	
ди, ти, ми, пи, би, қи, ки ,ыс, іс	Образные глаголы	

ПРИЛОЖЕНИЕ В

Классификация вспомогательных глаголов казахского языка, несущих семантику конкретизации типа действия

Вспомогательные глаголы	Конкретизация типа действия
қой (поставить) с деепричастием на–*а [-е, -й]* (поставить) **қойды**	Быстрое действие
қал (поставить) с деепричастием на–*а [-е, -й]* (остаться) **қалды**	Законченное внезапное действие
сал с деепричастием на–*а [-е, -й]* (положить) **салдым** **салып**	Законченное действие
кет (уходи) с основными глаголами типа (*отыр* (сидеть) *жат* (лежать) *жығыл(*падать*)*) в форме деепричастия на–*а [-е, -й]* **отыра кетті** **қисая кетсеңші**	Действие совершенное как-попало
бару (прибыть) **келу** (прибыть сюда) **кету**(уходить) **шығу**(выходить) **жүру**(идти) **тұру**(стоять) с деепричастием на–*а [-е, -й]* *ала бар* *ала кел* *ала шық* *ала жүр* *ала қайт* *шыққан* *барған*	Выражение попутного действия
түс (спускаться) с деепричастием на–*а [-е, -й]* *түсті*	Усиления и продления действия
бер с деепричастием на–*п* *беру*	Действие, протекающее неограниченное время
тұр (стоять) *тұрыңдар*	Временное действие, со значением «пока»

тұсын	
көр (видеть) с деепричастием на*–а [-е, -й]* *көрмепті* *көрме* —просьба Сочетается со сложной глагольной основой совершенного вида *жоғалтып ала көрме* —пожалуйста не потеряй *кетіп қала көрме* —пожалуйста не уходи	Попытка совершения действия, значение «попробовать»
жазда с деепричастием на*–а [-е, -й]* *жаздады* Может сочетаться с аналитической формой совершенного вида, тогда второй компонент сложного глагола принимает форму деепричастия на*–а [-е, -й]* *жоғылтып ала жаздады* *жығылып қала жаздады*	Действие, не совершившееся в силу каких-либо обстоятельств. «Чуть не..»
Алу (взять, брать) с деепричастием на*–а [-е, -й]* *алады* *алмадым* - отрицание *алмайды*- отрицание	Возможность или невозможность (если в отрицательной форме)
болу с деепричастием прошедшего времени на*–ған(- ген, - қан,- кен)* *өлген болып жатты* *таныған болып* *жатқан болып* *қалған болды да*	Действующее лицо притворялось
кел (в третьем лице)с основным глаголом на суффикс *–ғы(- ғі, -қы,- қі)* **келеді** **келді** *болғым келеді* *бейімдегісі келді*	Выражение желательной формы

ПРИЛОЖЕНИЕ Г

Основные словообразовательные залоговые суффиксы глаголов казахского языка

Суффикс	Залог	Пример
–н, -ын, -ін	возвратный	*жу-ын-у* (умыться, умываться), *ора-н-у* (укутаться) *көр-ін* (показался), *көр-ін-ді* (виднелся)
-лан, -лен, -дан, -ден, -тан, -тен	возвратный	*намыс-тан-у* (устыдиться), *шат-тан-у* (радоваться)
-сын, -сін, -қан, -кен	возвратный	
-л, -ыл, -іл	возвратный	Образуют как страдательный так и возвратный залог *бу-ыл-у* (задыхаться), *түй-іл-у* (нахмуриться)
-лык, -лік, -дық, -дік, -тік, -ық, -ік, -лығ, -іг, -ліг	возвратный	*Бу-лығ-у* (задыхаться), *іл-іг-у* (попасть)
-с, -лас, -лес	возвратный	*орна-лас-у (*устроиться*), қате-лес-у* (ошибаться)
-л, -н	страдательный	*жина-**л**-ды* (быть убранным)
-ыл, -іл, -ын, -ін	страдательный	*Оқ-**ыл**-ды* (быть прочитанным)
-лын, -лін, -ныл, -ніл	страдательный	*же-лін-у*(быть созданным), *қолда-ныл-у*(быть использованных) *пайдала-ныл-у*
-с, ыс, -іс	совместно-взаимный	
-лас, -лес, -дас, -дес, -тас, -тес	совместно-взаимный	*мұн-дас-у* (делиться своими радостями) *сыр-лас-у* (делиться своими горестями), *қас-тас-у* (враждовать друг с другом)
-ылыс, -ныс, -ыныс, -ініс -тығыс и др.	совместно-взаимный и возвратный	*сапыр-ыл-ыс-у* (вмешиваться), *байла-н-ыс-у* (связаться), *ұғ-ын-ыс-у* (объяснять друг другу) *соқ-тығ-ыс-у* (столкнуться)
-стыр, -стір, -ластыр, -лестір	совместно-взаимный и принудительный	*жара-с-тыр-у* (заставить соревноваться друг с другом), *таны-с-тыр-у* (познакомить их друг с другом)
-ландыр, -тендір, -лендір	возвратный и понудительный	*Индустрия-лан-дыр, коллектив-тен-дір, ірі-лен-дір*
-дастыр	принудительный	*Колхоз-дас, колхоз-дас-тыр, колхоз-*

		дас-тыр-ыл
-т, -ыт, -іт, -дыр, -дір, -тыр, -ғыз, -гіз, -қыз, -кіз, -ар, -ер, -ыр, -ір, -қар, -кер, -ғыр, -дар, -сет	обудительный	
-ғыздыр, -гіздір, -дырғыз, -діргіз	сложные аффиксы	
-ындыр, -індір, -ландыр, -лендір, -тандыр, -дендір и	К возвратному залогу + понудительный	
-ылыс, -лыс, -ініс, -ліс , -ығыс, -тығыс, --лікіс, -ықыс, -ікіс, -ныс, -ніс, -ыныс, -ініс	К возвратному залогу + совместно-взаимный залог	
-лін, -лын, -нілл	К возвратному + страдательный	
-тырыл, дырыл, -ғызыл, -сетіл	Понудительный + срадательный	
-ттыр	Понудительный + совместно взаимный	
-арыс, -іріс	понудительный	
-стыр, -стір, -ластыр, -лестір	Совместно-взаимный+ понудительный	

ПРИЛОЖЕНИЕ Д

Формализация грамматического оформления наклонения казахского языка.

• **повелительное наклонение**				
лицо	числ	Особенности основы	**Аффикс**	Пример
1	ед.	к основе деепричастия или глагола на **–а, -е**	**-йын** **-йін**	шақыр-**а-йын** (позову-ка) көтер-**е- йін** (подниму-ка) ен-**е-йін** (войду-ка) ки-**е-йін** (надену-ка)
1	мн.	к основе деепричастия или глагола на –а, -е	**-йық** **-йік**	Тарт-**а-йық** (потянем же) қос-**а-йық (**присоединим же) жат-**а-йық** (ляжем [те])
2	ед	Совпадает с основной формой глагола		Жет (достигай) Өт (проходи) Қорға (Защити)
2	Ед вежл	Присоединение к основе глагола аффикса	**-ңыз** **-ңіз**	Тында-**ңыз** (послушайте) кейіме-**ңіз** (не огорчайтесь)
2	мн	Прибавление к форме единственного числа аффикса множественного числа, а после аффикса ң	**-дар** **-ң-дар** **-ң-дер**	таста-**ң-дар** (бросайте) ойла-**ң-дар (**подумайте)
2	мн вежл	К вежливой форме ед. числа прибавляется аффикс множественного	**-ңыз-дар** **-ңіз-дер**	тоқта-**ңыз-дар** (остановитесь) жәрдемдесі-**ңіз-дер** (помогите)
3	ед	Прибавление к основе глагола аффикса	**-сын** **-сін**	бер-**сін** (пусть сообщит) тапсыр-**сын** (пусть вручат)
3	мн	3лицо не имеет формы множественного числа		
		Присоединение аффиксов, к повелительному наклонению, определяет	**–шы, -ші**	Ала--**йын–шы** (беру-ка), Айт-**шы** (скажи-ка), айты-**ңыз-шы** (скажи-

		действие как безотлогательное		ка)
• **настоящее конкретное время изъявительного наклонения**				
Лицо	число	Особенности основы	Вспомогательный глагол	Пример
1	Ед., мн.	деепричастие на **–п**	отыр, тұр, жатыр, жүр	Мен жазып отыр-**мын,** Біз жазып отыр-**мыз**
2	Ед., мн.	деепричастие на **–п**	отыр, тұр, жатыр, жүр	Сен келе жатыр-**сың,** Сендер келе жатыр-**сың-дар**
2	Ед., мн., вежл	деепричастие на **–п**	отыр, тұр, жатыр, жүр	Көріп тұр-**сыз,** Көріп тұр-**сыз-дар**
3	Ед., мн.	деепричастие на **–п**	отыр, тұр, жатыр, жүр	Катысып жүр
• **настоящее переходное время изъявительного наклонения**				
лицо	число	Особенности основы	аффиксы	Примеры
1. 2. 3.	Ед., мн.	деепричастия на –а, -е, -й, -и	Личные окончания глагола: -мын, -сыз, -сың, -мыз, -міз, -сыңдар, -сендер, -сіздер, -сыздар, -ды, -ай-ды	(ол бар-а-ды, он пойдет), колхозшылар егін **жинайды** (колхозники убирают уражай)
• **будущее предположительное время изъявительного наклонения**				
лицо	число	Особенности основы	аффиксы	
1. 2. 3.	Ед., мн.	деепричастия на –ар, -ер, -ір, -ир	Личные окончания глагола: -мыз, -міз, сыз, -сың, -сыңдар, -сендер, -сіздер, -сыздар	
• **неопределенное будущее время изъявительного наклонения**				
лицо	число	Особенности основы	аффиксы	
1.	Ед.,	Глаголы на –	Личные окончания	Бар-**мақ-пын**

2. 3.	мн.	**мақ, -мек – пақ, -пек, -бақ, -бек**	глагола: -мыз, -міз, сыз, -сың, -сыңдар, -сендер, -сіздер, -сыздар Может быть: **–шы, -ші** (придает действию обязательность)	(собираюсь пойти), Ол театрға бар-**мақ** (он собирается идти в театр), сат-**пақ-шы-мын,** жолық-**пақ-шы-мын.** Сен аудармақ –**сың,** Сендер аудармақ **-сың-дар,** Сіз аудармақ –**сыз,** Сіздер аудармақ **-сыз-дар,** Олар аудармақ
лицо	число	Особенности формирования	Вспомогательный глагол	Примеры
1. 2. 3.	Ед., мн.	Личные окончания принимает вспомогательный глагол	Еді, екен	бітірмекші еді, әңгімелеспекші едім
• **давно прошедшее время изъявительного наклонения**				
лицо	число	Особенности основы	аффиксы	Пример
		Деепричастие на –п	Личное окончание глагола: **-пін, –піз, -сің –сіндер, -сіз – сіздер, -ті –ті, -ты. (**3лицо**)** Может быть добавлен: **-мыс, -міс (**предположительность)	жүргізіппіз, жүргізіпсің, жүргізіпті, жеңіпті-міс, қайтып келіпті
• **обычное прошедшее время изъявительного наклонения**				
лицо	число	Особенности основы	аффиксы	
		Причастие прошедшего времени на **ған, ген қан, кен, ға, ге, қа, ке**	Личное окончание глагола: **- мын, -мін –быз, -сың – сыңдер**	

• **недавно прошедшее (окончательное) время изъявительного наклонения**				
		Особенности основы	аффиксы	
		Глагольная основа	Притяжательные аффиксы: **–ды, -ді, -ты, -ті** Множественное число добавляет – **к, -қ, -м, -ң**	көрін-**ді-м,** көрін-**ді-к,** Сен көрін-**ді-ң,** сіздер көрін-**ді-ңіз-дер,** ол [олар] көрін-**ді**
лицо	число	Основное слово предиката	Вспомогательные глаголы	
		существительное	[емес/ жоқ] **еді**	етікші еді
1	Ед, мн.	деепричастие на **–п,** причастие на **ған, ген, қан, кен**	[емес/ жоқ] едім, [емес/ жоқ] едік	Мен алған емес едім, біз алған емес едік
2	Ед, мн.	деепричастие на **–п,** причастие на **ған, ген, қан, кен**	[емес/ жоқ] едің, [емес/ жоқ] едіңдер, [емес/ жоқ] едіңіз, [емес/ жоқ] едіңіздер	Сен алған емес едің, сендер алған емес едіңдер, сіз алған емес едіңіз, сендер алған емес едіңіздер
3	Ед, мн.	деепричастие на **–п,** причастие на **ған, ген, қан, кен**	[емес/ жоқ] еді	Ол алған емес еді, олар алған емес еді
• **незаконченное прошедшее время изъявительного наклонения**				
лицо	число	Основное слово предиката	Вспомогательный глагол	Примеры
1	Ед, мн.	глаголы состояния (**отыр, тұр, жатыр, жүр**)	[емес/ жоқ] едім, [емес/ жоқ] едік	Келе жатыр едім, Біз келе жатыр едік
2	Ед, мн.	глаголы состояния	[емес/ жоқ] едің, [емес/ жоқ]	Сен келе жатыр едің, Сендер келе жатыр

		(**отыр, тұр, жатыр, жүр**)	едіңдер, [емес/ жоқ] едіңіз, [емес/ жоқ] едіңіздер	едіңдер, Сіз келе жатыр едіңіз, Сіздер келе жатыр едіңіздер
3	Ед, мн.	глаголы состояния (**отыр, тұр, жатыр, жүр**)	[емес/ жоқ] еді	Ол келе жатыр еді, Олар келе жатыр еді
1. 2. 3.	Ед, мн.	причастие на **атын, етін, йтын, йтін**	[емес/ жоқ] едім, [емес/ жоқ] едік, [емес/ жоқ] едің, [емес/ жоқ] едіңдер, [емес/ жоқ] едіңіз, [емес/ жоқ] едіңіздер, [емес/ жоқ] еді	танымайтын еді, естітін еді
• **форма намерения прошедшего времени изъявительного наклонения**				
		Основное слово предиката	Вспомогательный глагол	Примеры
1. 2. 3.	Ед, мн.	причастие на **мақ, мақшы**	[емес/ жоқ] едім, [емес/ жоқ] едік, [емес/ жоқ] едің, [емес/ жоқ], [емес/ жоқ] едіңіз, [емес/ жоқ] едіңіздер, [емес/ жоқ] еді	Жазбақшы едім, жазбақшы едік, жазбақшы едің, жазбақшы едіңдер, жазбақшы едіңіз, жазбақшы едіңіздер, жазбақшы еді
• **сложная сослагательная форма**				
ли цо	число	Основное слово предиката	Вспомогательный глагол	Примеры
1. 2. 3.	Ед., мн.	причастие на **–ар, -ер, -р**	едім, едік, едің, едіңіз, едіңіздер, еді	Айтар едім, айтар едің, айтар едік, айтар едіңіздер
• **Желательное наклонение пасивного типа**				
Ли цо	число	Особенности основы	аффиксы	Примеры
1	Ед.,мн.	причастие прошедшего времени на	**-мын, мыз, быз**	

		ғай, гей, қай, кей		
2	Ед.,мн.	причастие прошедшего времени на **ғай, гей қай, кей**	**-сың сыз –сыңдар, сыздар**	
3	Ед.,мн.	причастие прошедшего времени на **ғай, гей қай, кей**	-	
лицо	число	Основное слово предиката	Вспомогательный глагол	
1. 2. 3.	Ед.,мн.	основного глагола на -ғай, -гей, -қай, -кей	едім, едік, едің, едіңіз, едіңіздер, еді	Бітіргей едім, жеткей еді
• **Желательное наклонение утвердительного (или стремительного) типа**				
	число	Основное слово предиката	аффиксы	Вспомогательный глагол
1. 2. 3.	Ед., мн.	Основа глагола на **-ғы, -гі, -қы, -кі**	Притяжательные аффиксы: **- мыз, -міз, -ң -ңыз -ңіз -сы -сі**	**келеді** тыңда-**ғы-м келеді,** тыңда-**ғы-ң келеді,** тыңда-**ғы-ларыңыз келеді,** тыңда-**ғы-сы келеді**
	число		Вспомогательное слово	Вспомогательный глагол
1. 2. 3.	Ед., мн.	Основа глагола на , **-са, -се**	**игі**	едім, едік, едің, едіңіз, едіңіздер, еді Ал-**са игі еді,** Ал-**са игі еді**
• **Условное наклонение**				
лицо	число		Возможны личные окончания:	
1	Ед, мн.	Основа глагола	**-м [шы],-қ[шы],**	Ал-**са-м,** Ал-**са-қ** бер-**се-**

		на **-са, -се**	**-к[шы]**	**к,** айтса-**м-шы**, Айтса-**қ-шы**
2	Ед, мн.	Основа глагола на **-са, -се**	**-ң[шы], -ңыз[шы], –ңіз[шы], (+-дар[шы], -дер[шы])**	Ал-**са-ң,** ал-**са-ңыз,** ал-**са-ң-дар,** ал-**са-ң-дар,** айтса-**ңыз-шы**
3	Ед, мн.	Основа глагола на **-са, -се**	-	Ал-**са,** бер-**се**
лицо	число		Возможны личные окончания:	Вспомогательный глагол
1. 2. 3.	Ед., мн.	Основа глагола на **-са, -се**	**-м,-қ, -к, -ң, -ңыз, –ңіз,**	**екен, еді,** частица **ғой** мен көр-се-м екен, мен көр-се-м ғой, сіздер көр-**се-ңіз-дер екен,** сіздер көр-**се-ңіз-дер ғой**

ПРИЛОЖЕНИЕ Е

Классификация основных словообразовательных афиксов существительных казахского языка.

шы, ші тық; лық; лік; дық; дік; тік дас; лас; лес; дес; тас; тес қар; қор; кер; гер; кес; паз; сақ; ман қи ғи қай ғай гей қой ғой ша ше шық шік шақ шек пана хана	Аффиксы отглагольного образования
қыш кіш ғыш ғіш қаш ғаш қы кі ғы гі ыш іш ш ыс іс с ық уық уік ік ақ қ лақ лек дақ дек тақ тек ғақ гек ек к мақ мек пақ пек ма ме ба бе па пе ым ім м	Аффиксы отглагольного образования
шылық шілік шылдық шілдік ластық лестік дастық дестік тастық тестік лылық лілік дылық ділік тылық тілік герлік кестік қорлық паздық сақтық шақтық сыздық сіздік ғыштық гіштік	Сложные аффиксы
ша ше шақ шық шік қан тай ке ш жан сымақ	Аффиксы экспрессивной оценки – уменьшительно-ласкательный и уничижительный оттенки