ВЫЧИСЛИТЕЛЬНАЯ ОБРАБОТКА КАЗАХСКОГО ЯЗЫКА

СБОРНИК НАУЧНЫХ ТРУДОВ

ҚАЗАҚ УНИВЕРСИТЕТІ
БАСПА ҮЙІ

КАЗАХСКИЙ НАЦИОНАЛЬНЫЙ УНИВЕРСИТЕТ имени АЛЬ-ФАРАБИ

ВЫЧИСЛИТЕЛЬНАЯ ОБРАБОТКА КАЗАХСКОГО ЯЗЫКА

Сборник научных трудов

Алматы
«Қазақ университеті»
2020

УДК 811.512.122
ББК 81.2Қаз-923
В 27

Рекомендовано к изданию Ученым советом факультета информационных технологий (протокол № 13 от 30 июня 2020 г.)

Сборник выполнен в рамках проекта грантового финансирования научных исследований МОН РК AP05132950 «Разработка информационно-аналитической поисковой системы данных на казахском языке»

Рецензенты:
PhD, доцент, кандидат физико-математических наук ***К.С. Дуйсебекова***
кандидат технических наук ***Л.С. Копболсын***

Под редакцией
PhD Рахимовой Д.Р.

В 27 **Вычислительная** обработка казахского языка: сборник научных трудов / под редакцией Рахимовой Д.Р. – Алматы: Қазақ университеті, 2020. – 147 с.

ISBN 978-601-04-4698-4

В данном сборнике представлены научные разработки в области обработки казахского языка.

Предназначен для преподавателей, научных сотрудников, магистров и студентов факультетов информационных технологий и филологий.

УДК 811.512.122
ББК 81.2Қаз-923

ISBN 978-601-04-4698-4

СОДЕРЖАНИЕ

ВВЕДЕНИЕ 4

Глава 1. РАЗРАБОТКА ОНОТОЛОГИЧЕСКОЙ МОДЕЛИ ГРАММАТИКИ КАЗАХСКОГО ЯЗЫКА (Шарипбай А.А., Муканова А.С., Ергеш Б.Ж., Разахова Б.Ш., Елибаева Г.К.) 6

Глава 2. К ВОПРОСУ О РАЗРАБОТКЕ КОРПУСА КАЗАХСКОГО ЯЗЫКА (Мадиева Г.Б., Бектемирова С.Б., Мамбетова М.К.) 34

Глава 3. РАЗРАБОТКА МЕДИА-КОРПУСА КАЗАХСКОГО ЯЗЫКА (Мансурова М.Е., Мадиева Г.Б., Қадырбек Н.Қ., Қыргызбаева М.Е.) 48

Глава 4. РАЗРАБОТКА АЛГОРИТМОВ ИЗВЛЕЧЕНИЯ КЛЮЧЕВЫХ СЛОВ И СЕМАНТИЧЕСКОГО АНАЛИЗА ДАННЫХ НА КАЗАХСКОМ ЯЗЫКЕ (Рахимова Д., Турганбаева А.О., Жуманов Ж.М.) 63

Глава 5. МОДЕЛИ И МЕТОДЫ СЕНТИМЕНТ АНАЛИЗА ТЕКСТОВ НА КАЗАХСКОМ ЯЗЫКЕ (Ергеш Б.Ж., Шарипбай А.А., Бекманова Г.Т.) 86

Глава 6. МЕТОД ИДЕНТИФИКАЦИИ КРИМИНАЛЬНОГО ЗНАЧЕНИЯ ТЕКСТОВ НА КАЗАХСКОМ ЯЗЫКЕ, БАЗИРУЮЩИЙСЯ НА VSM (Мамырбаев О.Ж., Хайрова Н.Ф., Мухсина К.Ж., Колесник А.С.) 104

Глава 7. РАЗРАБОТКА МОДЕЛИ ПОСТ-РЕДАКТИРОВАНИЯ В МАШИННОМ ПЕРЕВОДЕ КАЗАХСКОГО ЯЗЫКА (Рахимова Д.Р., Турарбек А.Т., Пазылхан Н.М.) 121

ВВЕДЕНИЕ

Обработка естественного языка (*Natural Language Processing, NLP*) – общее направление искусственного интеллекта и математической лингвистики. Оно изучает различные прикладные задачи, связанные с информационными технологиями анализа и синтеза естественных языков. Применительно к искусственному интеллекту анализ означает понимание языка, а синтез – генерацию грамотного текста. Решение этих проблем будет означать создание более удобной формы взаимодействия компьютера и человека. Сегодня NLP применяется во многих сферах, в том числе в голосовых помощниках, автоматических переводах текста и фильтрации текста.

В области NLP проводятся исследования и разработки различных задач. Среди этих задач, можно выделить следующие:

– Распознавание текста, речи, синтез речи;

– Морфологический анализ (слова);

– Сбор и хранение лингвистических ресурсов (корпус, словари, тезаурус и др.);

– Синтаксический разбор, токенизацию предложений;

– Извлечение отношений, определение языка, анализ эмоциональной окраски;

– Аннотацию документа, перевод, анализ тематики текста;

– Информационный поиск, машинный перевод и др.

Область обработки естественного языка в мировой науке является достаточно зрелой, разработаны многие формальные модели, методы и алгоритмы. Существуют высокого уровня различные прикладные программные системы по анализу и обработке естественных языков, сбору и хранению лингвистических данных.

Активная интеграция Казахстана в мировое сообщество и увеличивающимся объемом информационных потоков между нашей страной и ее зарубежными партнерами, реальная потреб-

ность для различных слоев населения в информационных технологиях в повседневной жизни возрастает. В данной книге представлены научные исследования и разработки в области обработки казахского языка. Научное издание состоит из семи глав различных работ научных групп из исследовательских институтов, университетов и лабораторий Республики Казахстан. В каждой главе представлены различные научные исследования, разработанные алгоритмы, модели и технологии по различным тематикам обработки казахского языка. Представленные научные направления являются результатами долгих и трудоемких работ. Издание предназначено для преподавателей, научных сотрудников, магистров и студентов факультетов информационных технологий и филологии.

Глава 1

РАЗРАБОТКА ОНТОЛОГИЧЕСКОЙ МОДЕЛИ ГРАММАТИКИ КАЗАХСКОГО ЯЗЫКА

Аннотация. *В настоящее время объем информационных ресурсов на естественном языке резко увеличиваются. Обработка таких ресурсов требует наличия лингвистических баз данных и знаний. Для их создания требуются языки разметки и онтологические модели предметных областей. Эти задачи являются актуальными в области компьютерной лингвистики. Для этих целей предлагается разработать метаязык и онтологическую модель грамматики казахского языка. В данном разделе описаны онтологические модели морфологии и синтаксиса казахского языка, которые являются частью исследований в рамках проекта «AP05132249 Разработка электронных тезаурусов тюркских языков для создания систем многоязычного поиска и извлечения знаний».*

1.1. Введение

В связи с резким увеличением объема информации на естественных языках в интернете и социальных сетях исследования и разработки в области компьютерной лингвистики становятся чрезвычайно актуальными. Обработка информационных ресурсов на естественном языке требует наличия текстовых корпусов и тезаурусов. Для их создания и обработки требуются языки разметки и онтологические модели предметных областей. Существующие языки разметки в основном содержат концепты романо-германских и славянских языковых групп.

Метаязык предназначен для разметки текстов естественных языков. Метаязык обладает следующими свойствами: с помощью его языковых средств можно выразить все, что выразимо средствами объектного языка, и обозначить все знаки, выражения объектного языка, для которых имеются имена; на метаязыке можно говорить о свойствах выражения объектного языка и отношениях между ними; на нем можно сформулировать определе-

ния, обозначения, правила образования и преобразования для выражений объектного языка [1]. Имеются известные системы разметки, такие, как Penn Treebank [2], система разметки CLAWS [3], который применяется для разметки Британского корпуса, система разметки Брауновского корпуса, в американском национальном корпусе [4, 5] используются несколько систем разметки, про системы разметки более подробно описаны в [6, 7]. Все эти системы в основном применяются для разметки английского языка.

Эти метаязыки не приспособлены для описания тюркских языков, которые имеют много концептов, отличных от концептов вышеуказанных языковых групп. Поэтому создание единого метаязыка для разметки текстов тюркских языков (UniTurk) является актуальной задачей для обработки тюркских языков. Такой язык позволит унифицировать разметки, облегчить их понимание и использовать общее программное обеспечение, а также позволит производить сравнительный анализ лингвистических концептов тюркских языков.

А также кроме языка разметки, для компьютерной обработки любых естественных языков, требуются формализация их грамматических (морфологических и синтаксических) правил, разработка алгоритмов анализа и синтеза слов и предложений по этим правилам, программная реализация всех этих алгоритмов, создание тезаурусов по предметным областям аналогично WordNet [8], построение текстовых корпусов (база данных размеченных текстов) и других программ для анализа и обработки текстов.

Для формализации грамматических правил используется онтология. Онтология – это концептуальная схема, состоящая из множества понятий и множества утверждений об этих понятиях, на основе которых можно описывать классы, отношения, свойства, функции и индивиды.

Можно также сказать, что онтология – это база знаний, потому что если добавить интерпретирующие функции к структурно-семантической модели, то она станет базой знаний [9].

Все онтологические модели построены в среде Protégé [10], который позволяет упростить процесс создания, загрузки, изменения и преобразования базы знаний, а также предоставить ее в общее пользование в виде совместного просмотра и редактирования.

1.2. Основная часть

1.2.1. Разработка унифицированной системы разметки для тюркских языков

В настоящее время существует больше десяти электронных корпусов для тюркских языков: корпус казахского языка [11, 12, 13, 14]; корпус татарского языка ‘Туган тел’ [15], корпус турецкого языка [16]; корпус крымско-татарского языка [17], чувашского языка [18] и др. Одним из основных компонентов этих корпусов является система морфологической, синтаксической и семантической разметок, базовой среди которых является система морфологической разметки, которая разнится в некоторых корпусах. Поэтому необходимость создания единой системы разметок не вызывает сомнения. Единая система разметок позволит унифицировать разметки, облегчить их понимание и использовать общее программное обеспечение, а также проводить различные исследования по лингвостатистическому сравнительному анализу среди тюркских языков [19-26]. Создание унифицированной системы разметки (метаязыка) для тюркских языков (UniTurk) является актуальной задачей.

Для создания такой системы разметок необходим общий ресурс с которым могли бы работать все разработчики тюркских электронных корпусов. Такой ресурс мог бы выполнять роль справочной системы как для разработчиков, так и для пользователей тюркских электронных корпусов. Наиболее подходящими компонентами такого ресурса, которые соответствуют требуемым условиям являются онтологические модели грамматики тюркских языков.

В результате научно-исследовательской работы по разработке системы разметки предложена система единой разметки понятий морфологии и синтаксиса тюркских языков (казахский, татарский, кыргызский, узбекский и турецкий) [27, 28].

Метаязык морфологии и синтаксиса тюркских языков представляет собой унифицированный перечень специальных лингвистических разметок – тэгов, необходимых для разметки морфологических и синтаксических понятий, и будут в онтологической модели грамматики тюркских языков. Подробное описание тегов приведены в промежуточных отчетах НИР [29, 30].

Созданные лингвистические ресурсы позволяют, с одной стороны, способствовать взаимопониманию терминологии между тюркскими языками, а с другой стороны, стать многоязычной базой данных, которая будет использоваться в системах многоязычного поиска, машинного перевода между тюркскими языками, автореферирование тюркских текстов, а также в информационно-справочных и обучающих системах.

1.2.2. Разработка онтологической модели морфологии казахского языка

Прикладная онтология «Морфология казахского языка» основана на принципах общей онтологии и построена в среде Protégé, так как именно в Protégé можно описывать не только понятия, но и конкретные объекты, а также имеет богатый набор операторов – например, пересечение, объединение и отрицание. Protégé основан на логической модели, которая позволяет создавать определения, соответствующие неформальному описанию. Логическая модель позволяет использовать рассуждения, которые могут проверить все ли утверждения и определения в онтологии взаимно согласуются и могут также выяснить, какие концепции соответствуют заданным определениям [10, 31].

Онтологическая модель морфологии казахского языка состоит из отдельных индивидов, свойств и классов [35, 36]. В ходе работы первым делом необходимо было создать классы. Все спроектированные классы морфологических правил казахского языка отображены в окне «Class hierarchy» (рис. 1.1).

Классы интерпретируются как множества, элементами которых являются индивиды. Они описываются, используя формальные (математические) конструкции, которые декларируют требования для членства в классе.

Классы могут быть организованы в иерархию отношений вида «подкласс-суперкласс», которая так же известна, как таксономия. Подклассы специализируют (т.е. являются подмножествами) своего суперкласса. Например, рассмотрим классы «Грамматика» и «Морфология».

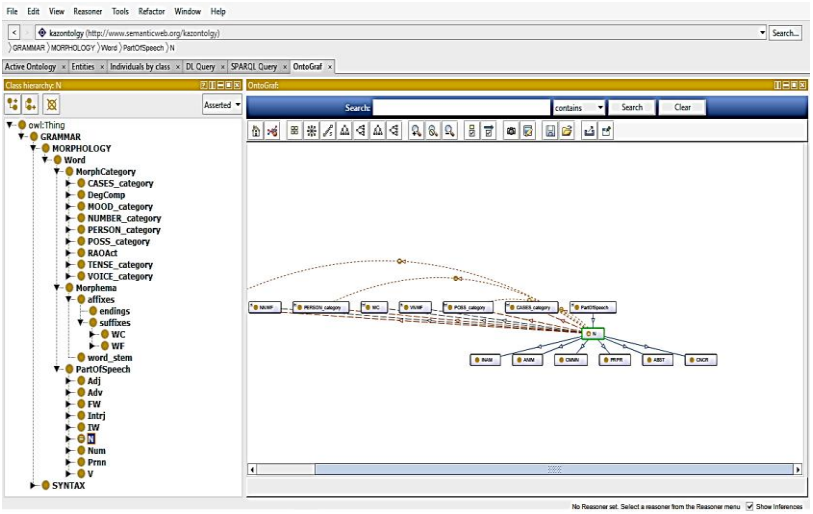

Рис. 1.1. Классы и подклассы в онтологии «Морфология казахского языка»

«Морфология» определяется как раздел Грамматики, и поэтому может быть подклассом Грамматики (таким образом, Грамматика – суперкласс класса «Морфология»). Это означает, что все «морфологические признаки» – это грамматика, все члены класса «Морфология» – члены класса «Грамматика» (рис. 1.2).

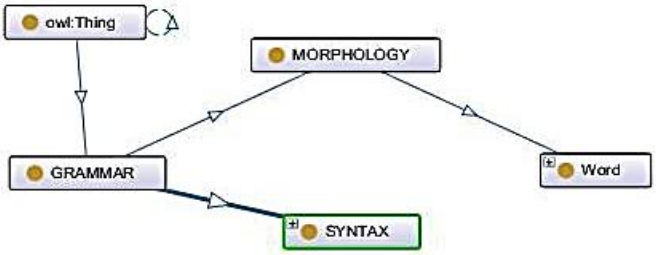

Рис. 1.2. Пример представления классов

После создания классов прописывались в них поля – свойства. Свойства объектов – определяют некоторые отношения между двумя объектами (классы, индивиды). К примеру, для концепта «Adj» (Имя прилагательное) характерны следующие свой-

ства: типом может быть либо «качественные имена прилагательные» (Qual), либо «относительные имена прилагательные» (Rel) в котором изменяются смысловые значения прилагательных. Поэтому вводим функциональное свойство «hasSemanticType» (имеет семантическое значение), и описываем его с помощью следующего ограничения: hasSemanticType some (Qual or Rel) (рис. 1.3).

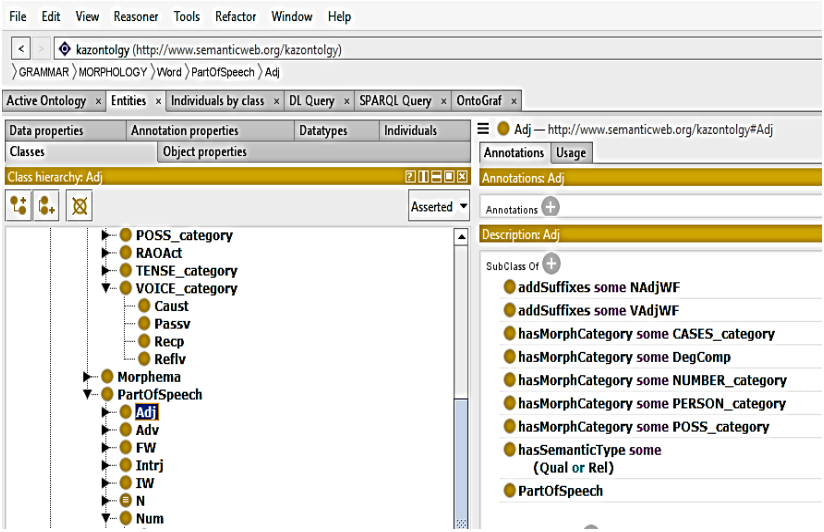

Рис. 1.3. Пример свойств семантических значений имен прилагательных казахского языка

В онтологии «Морфология казахского языка» также будет использоваться другой тип свойства – свойства аннотации. Свойства аннотации используются для добавления информации (метаданные – данные о данных) для классов, отдельных индивидов и свойств объектов. Там содержится информация о том, чем является этот концепт, какие грамматические категории или морфологические единицы он описывает, какие элементы или подклассы он может включать. На рисунке 1.4 представлен пример свойства аннотации, который содержит информацию об имени числительном (Num) казахского языка.

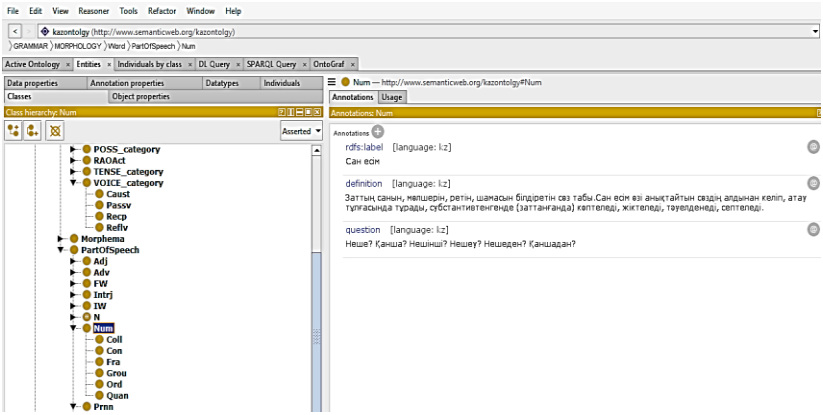

Рис. 1.4. Пример свойства аннотации имени числительного казахского языка

Далее, после обработки классов и свойств были добавлены индивиды соответствующих классов. Индивиды, представляют собой конкретные объекты интересующей предметной области, это основные, нижне-уровневые компоненты онтологии. Рисунок 1.5 показывает пример представления индивидов исходного падежа (ABL) имени существительного (N) казахского языка, мы представляем отдельных индивидов как ромбики в диаграммах.

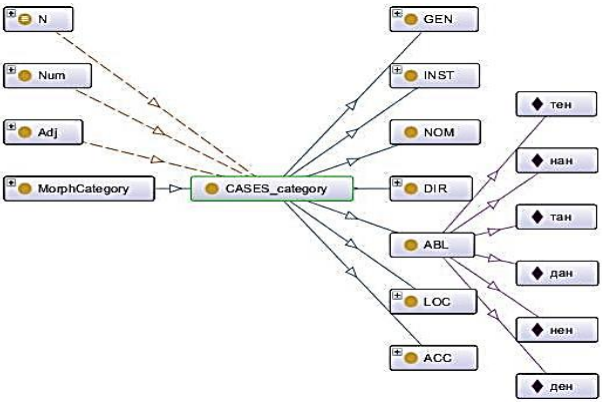

Рис. 1.5. Изображение индивидов исходного падежа (ABL) имен существительных (N) казахского языка

После обработки всех классов, свойств и индивидов была построена онтологическая модель «Морфология казахского языка». В отчете НИР [29] подробно описаны формальные модели частей речи казахского языка по отдельности.

Таким образом, построенная онтологическая модель «Морфология казахского языка» (рис. 1.6) включает в себя все компоненты и совокупности, связанные с морфологическими признаками казахского языка [32-34].

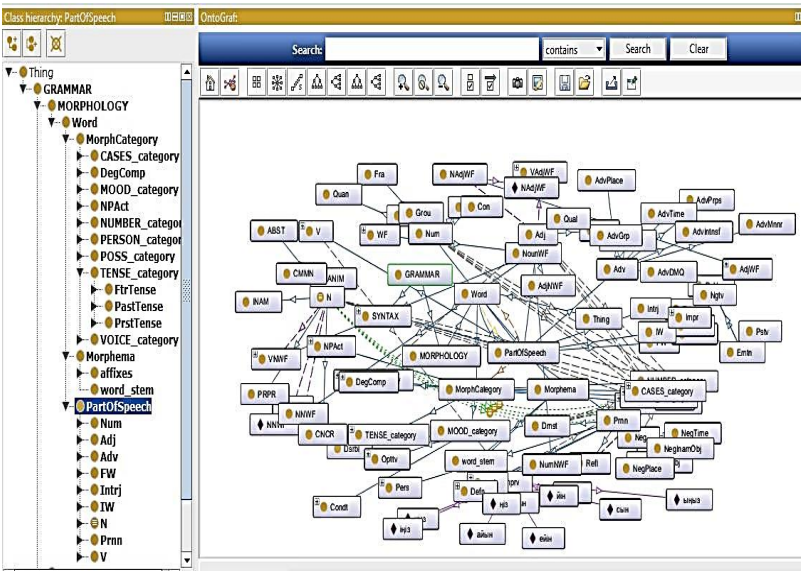

Рис. 1.6. Онтологическая модель «Морфология казахского языка»

Разработанная онтологическая модель «Морфология казахского языка» может давать ответы на введенные разработчиком запросы, основанные на синтаксисе Protégé, извлекающие данные на основе сбора всей информации о конкретном классе, свойстве или экземпляре класса. Например, выполним следующий запрос:

```
SELECT ?subject ?object
WHERE { ?subject rdfs:subClassOf ?object }
```

В результате выполнения запроса переменной ?subject в соответствие будут установлены классы, для которых характерно наличие свойства subClassOf. То есть классы, которые являются потомками для других классов (например, класс, определяющий ANIM (одушевленные существительные), является подклассом по отношению к классу, определяющему N (имя существительное) (рис. 1.7) и классов, которые являются подклассами ограничений (рис. 1.8).

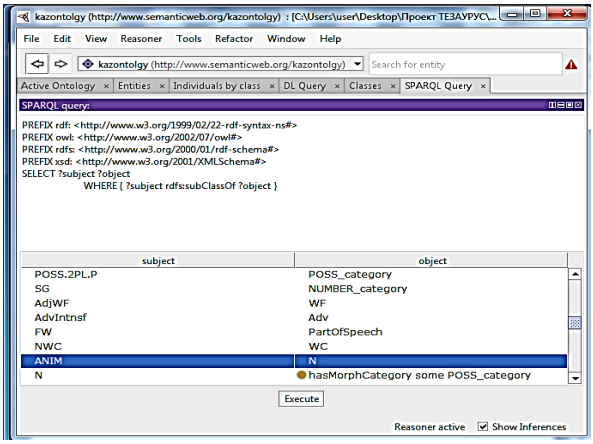

Рис. 1.7. Образцы потомков класса N

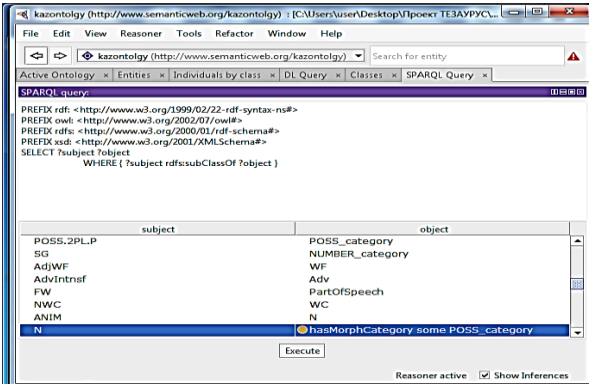

Рис. 1.8. Образцы потомков для ограничения классов существительных содержащих морфологическую категорию с окончанием принадлежности

Выполним запрос, который вернет только те классы, которые являются подклассами «Сан есім». В результате выполнения запроса будут возвращены потомки класса, для которого установлена метка «Сан есім» @kz (рис. 1.9).

```
SELECT ?subject ?class
WHERE { ?subject rdfs:subClassOf ?class.
?class rdfs:label «Сан есім»@kz}
```

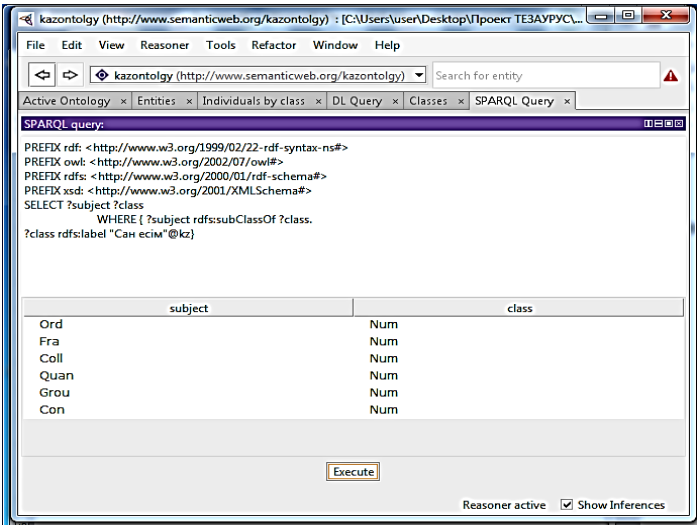

Рис. 1.9. Потомки класса NUM

Выполним запрос, используя информацию о комментарии (рис. 1.10).

```
SELECT ?class
WHERE {
?class rdfs:comment «Сөздің өзіне тән мағынасы бар ең ұсақ бөлшегі» @kz}
```

Построенная онтологическая модель поможет пользователю получить обширную информацию о морфологических признаках

казахского языка, а также имеет большое влияние для разработки тезауруса и программы машинного перевода тюркских языков.

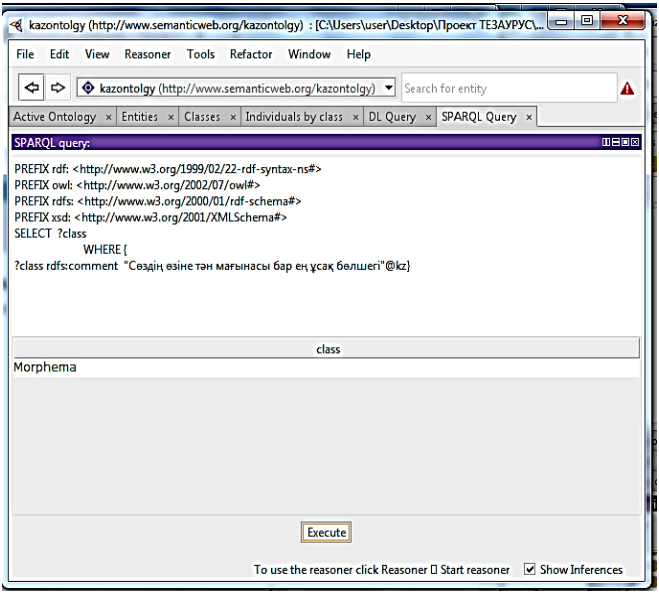

Рис. 1.10. Пример класса, которому соответствует заданный комментарий

123. Разработка онтологической модели синтаксиса казахского языка

В данном разделе описана онтологическая модель синтаксиса словосочетаний казахского языка, формализованы синтаксис предложений и показаны их деревья составляющих, а также построена онтологическая модель синтаксиса предложений с учетом семантики их составляющих.

Прикладная онтология «Синтаксис казахского языка» состоит из отдельных индивидов, свойств и классов, а также функций интерпретации, заданных на концептах или отношениях онтологии [37, 38]. Все спроектированные классы синтаксических понятий казахского языка отображены на рисунке 1.11.

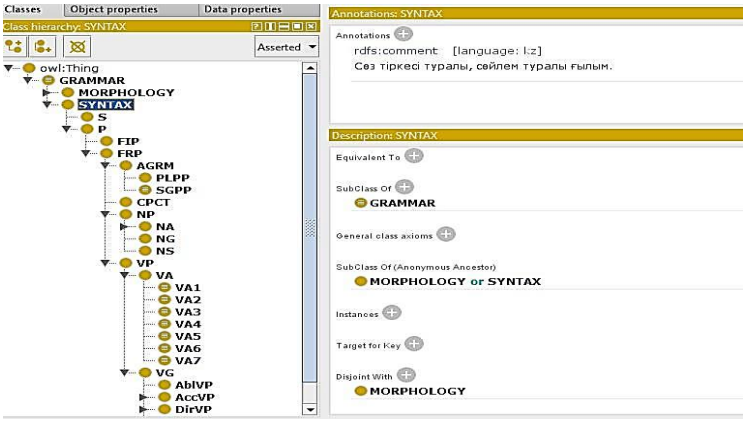

Рис. 1.11. Онтология «Синтаксис казахского языка»

На рисунке 1.12 показаны виды связи слов казахского языка.

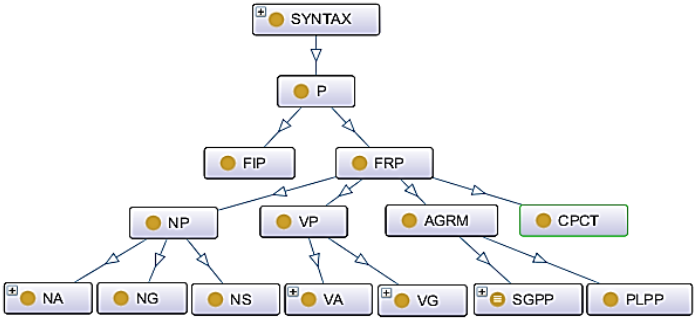

Рис. 1.12. Виды связи слов казахского языка

На рисунках 1.13, 1.14 показаны фрагменты реализации именных словосочетаний казахского языка в среде Protégé.

На рисунке 1.13 представлена реализация именного примыкания (главное слово – имя существительное, зависимое слово – имя прилагательное). Например, этому правилу в казахском языке соответствуют словосочетания *«қызыл алма»*, *«атты адам»*. Для этого должны быть выполнены следующие необходимые и достаточные условия:

$$NA2 \equiv \exists hasDependent\ (Adj) \sqcap \exists hasHead\ (N), \quad (1.1)$$

где: NA2 – именное примыкание (имя прилагательное + имя существительное);

hasDependent – имеется зависимое слово;

Adj – имя прилагательное;

hasHead – имеется главное слово;

N – имя существительное.

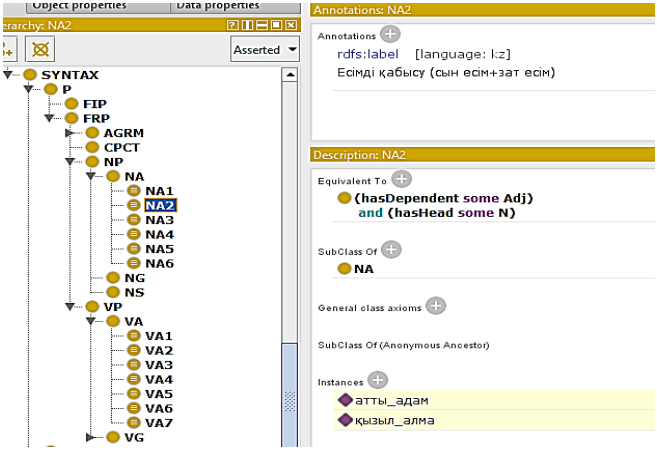

Рис. 1.13. Именное примыкание
(имя прилагательное + имя существительное)

Именное примыкание позволяет определять словосочетания *«қызыл алма»*, *«атты адам»*, которые являются индивидами концепта P (словосочетание) в категории NA2.

Для синтаксической категории DirNG2 – именное управление (имя числительное в направительном падеже + имя существительное) должны быть выполнены следующие необходимые и достаточные условия:

$$DirNG2 \equiv \exists hasDependent\ (AdjDir) \sqcap \exists hasHead\ (N), \quad (1.2)$$

где: DirNG2 – именное управление (имя числительное в направительном падеже + имя существительное);

hasDependent – имеется зависимое слово;
AdjDir – имя числительное в направительном падеже;
hasHead – имеется главное слово;
N – имя существительное.

При запуске резонера (решателя) в среде Protégé, словосочетания *«жақсыға дос»* определяется как именное управление в категорию DirNG2, что показано на рисунке 1.14.

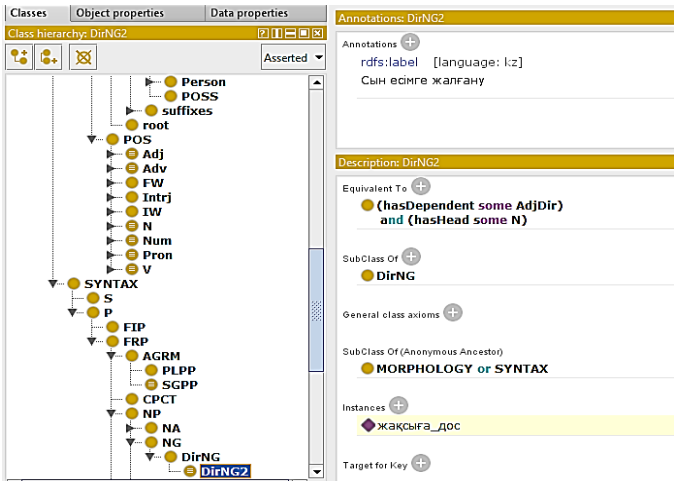

Рис. 1.14. Именное управление
(имя числительное в направительном падеже + имя существительное)

Таким образом, онтологическая модель именных словосочетаний казахского языка имеет вид, который представлен на рисунке 1.15.

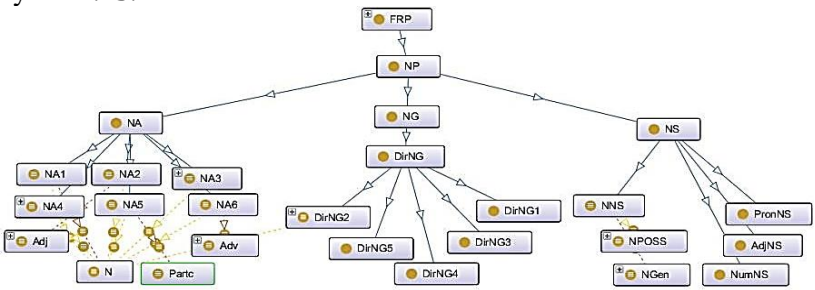

Рис. 1.15. Онтологическая модель именных словосочетаний казахского языка

На рисунках 1.16 и 1.17 показаны реализации синтаксических правил глагольных словосочетаний казахского языка в среде Protégé.

Например, для синтаксической категории VA1 (примыкание глагола с наречием) должны быть выполнены следующие необходимые и достаточные условия:

$$NVA1 \equiv \exists hasDependent\ (Adv) \sqcap \exists hasHead\ (V), \quad (1.4)$$

где: VA1 – глагольное примыкание (наречие + глагол);

hasDependent – имеется зависимое слово;

Adv – наречие;

hasHead – имеется главное слово;

V – глагол.

При запуске резонера в среде Protégé, словосочетания *«балаша күлді»*, *«кеше келді»*, которые являются индивидами концепта P (словосочетание) определяются как глагольные примыкания в категорию VA1, которые представлены на рисунке 1.16.

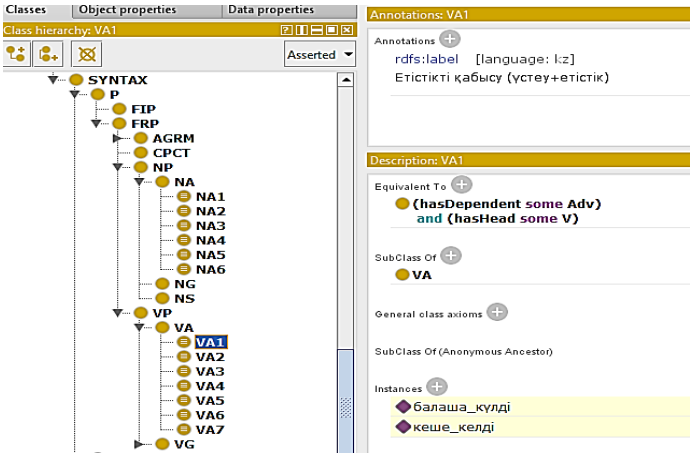

Рис. 1.16. Глагольное примыкание (наречие + глагол)

На рисунке 1.17 представлена реализация синтаксического правила словосочетания на казахском языке *«далаға қарады»*. По синтаксическому правилу это словосочетание относится к гла-

гольному управлению (главное слово – глагол, зависимое слово – имя существительное в направительном падеже). Для этого примера необходимым и достаточным условием являются следующие:

$$DirVG1 \equiv \exists hasDependent\ (NDir) \sqcap \exists hasHead(V), \quad (1.5)$$

где: DIRVG1 – глагольное управление (имя существительное в направительном падеже + глагол);

hasDependent – имеется зависимое слово;

NDir – имя существительное в направительном падеже;

hasHead – имеется главное слово;

V – глагол.

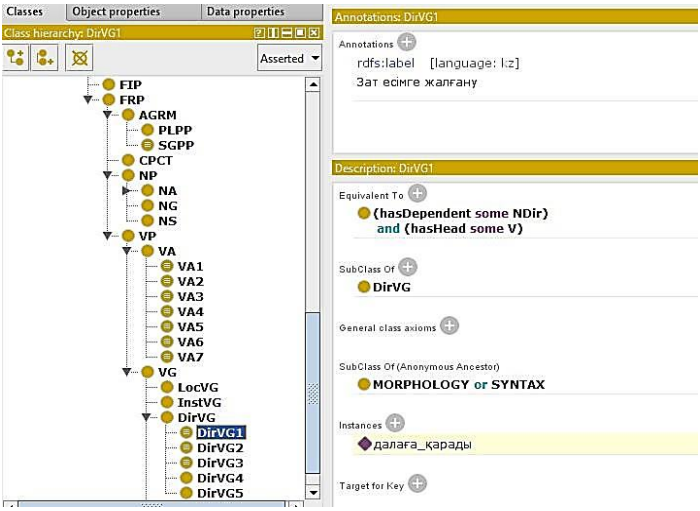

Рис. 1.17. Глагольное управление

Онтологическая модель глагольных словосочетаний казахского языка представлена на рисунке 1.18.

Таким образом, онтологическую модель словосочетаний казахского языка можно представить как на рисунке 1.19, которая включает в себя все понятия и взаимосвязи словосочетаний согласно синтаксическим правилам казахского языка.

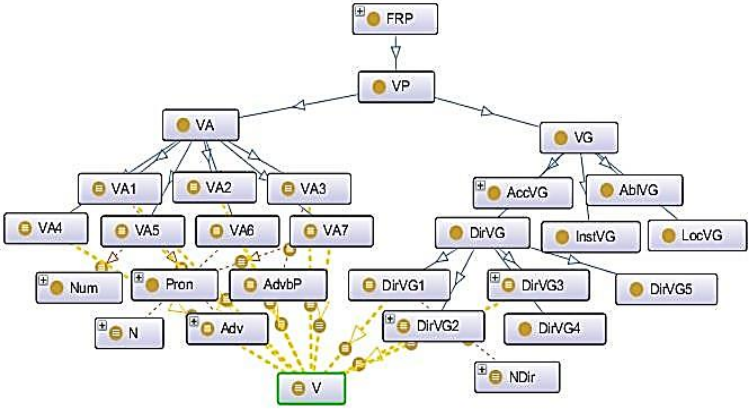

Рис. 1.18. Онтологическая модель глагольных словосочетаний казахского языка

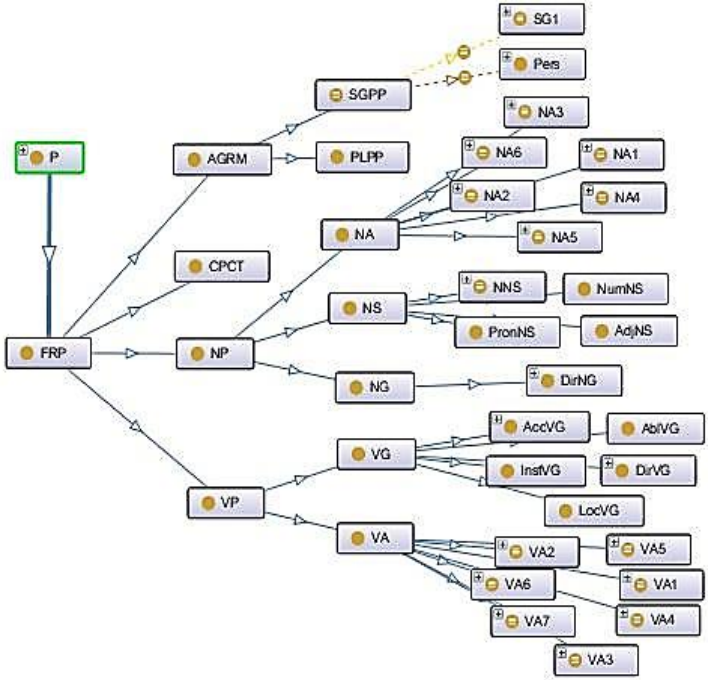

Рис. 1.19. Фрагмент онтологической модели словосочетаний казахского языка

Для построения онтологической модели синтаксических правил предложений казахского языка сначала были построены деревья их составляющих по формализации этих правил с помощью контекстно-свободной (КС) грамматики [6, 40, 41] и они использованы при создании онтологической модели предложений казахского языка.

Структуру предложений можно представить из двух частей: именное, глагольное. Синтаксис повествовательных простых предложений казахского языка можно описывать с помощью КС-грамматики [38].

Например, пусть заданы следующие простые предложения казахского языка:

1. «Самат сабаққа дайындалды» – «Самат подготовился к уроку»;
2. «Астана тез көркейді» – «Астана быстро преобразилась»;
3. «Самат атасынан кеше келді» – «Самат от дедушки вчера вернулся»;
4. «Апамның балалары хабарды жеткізді» – «Дети бабушки донесли известие».

Чтобы описать структуры этих предложений правила вывода КС грамматики записываются в следующем виде:

$$S \to NP \mid VP, NP \to N \mid N \mid Adj \mid Adv, VP \to N \mid V \mid Adv \mid NP \mid VP\}.$$

Используя правила этой грамматики деревья составляющих вышеуказанных предложений казахского языка представлены на рисунках 1.20-1.23:

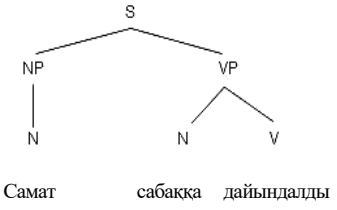

Рис. 1.20. Дерево составляющих S(NP(N),VP(N,V))

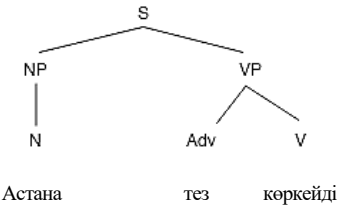

Рис. 1.21. Дерево составляющих S(NP(N), VP(Adv,V))

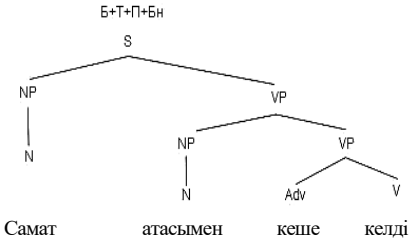

Рис. 1.22. Дерево составляющих S(NP(N), VP(NP(N), VP(Adv,V)))

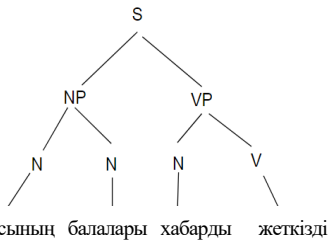

Рис. 1.23. Дерево составляющих S(NP(N,N), VP(N,V))

В казахском языке имеются 19 видов структуры простых повествовательных предложений, которые формализованы в работах с помощью КС-грамматики [40, 41]. На основании этой грамматики можно построить онтологические модели простого повествовательного предложения казахского языка, представленные на рисунках 1.24-1.30.

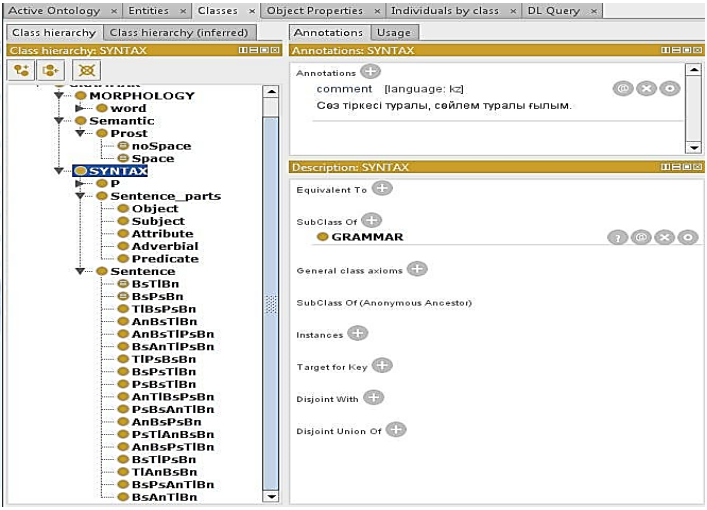

Рис. 1.24. Фрагмент структуры онтологической модели предложения казахского языка

На рисунке 1.26 представлен тип предложения *BsTlBn* (Бастауыш, Толықтауыш, Баяндауыш). Для этого типа предложения

должны быть выполнены следующие необходимые и достаточные условия:

$$BsTlBn \equiv \exists hasNP(PL\ or\ Pers) \sqcap \exists hasVP\ (DirVP1 \sqcap \left(\exists hasDependent\ \left(NDir \sqcap \left(\exists hasRoot(root \sqcap (\forall isSpace\ noSpace))\right)\right)\right)) \quad (1.7)$$

$BsTlBn$ – типы предложения состоящие из подлежащего, дополнения и сказуемого (Бастауыш, Толықтауыш, Баяндауыш);

$hasNP$ – имеет именное словосочетание;

PL – слова во множественном числе;

$Pers$ – личные местоимения;

$hasVP$ – имеет глагольное словосочетание;

$DirVP1$ – глагольное управление (имя существительное в направительном падеже + глагол);

$hasDependent$ – имеется зависимое слово;

$NDir$ – существительное в направительном падеже;

$hasRoot$ – имеет корень;

$root$ – корень;

$isSpace$ – является ли словом с пространственным значением;

$noSpace$ – непространственная семантика.

Если все условия соблюдены, тогда при запуске резонера в среде Protégé, предложение *«Балалар тамаққа тойды»*, который является индивидом концепта Sentence (предложение) определяется как тип предложения $BsTlBn$, так как существительное «тамақ» является непространственным существительным (рис. 1.25), а также будут выполнены необходимые и достаточные условия этого типа предложения.

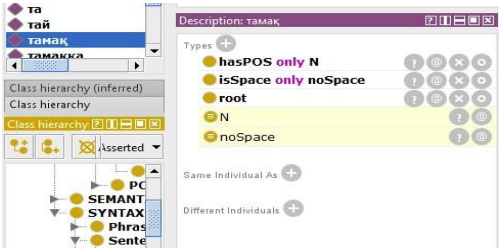

Рис. 1.25. Определение слова «тамақ» как непространственное существительное

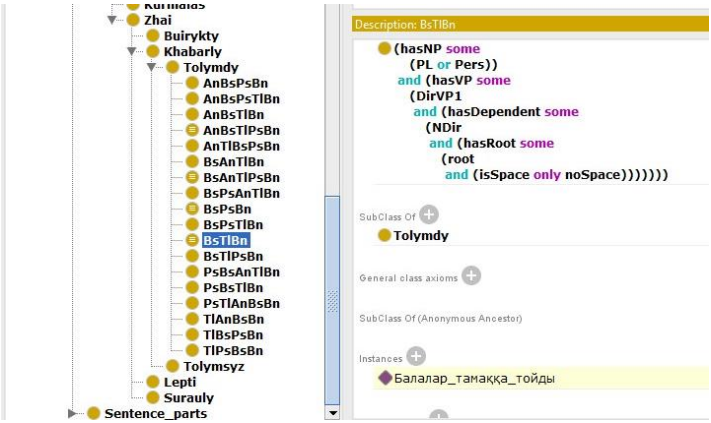

Рис. 1.26. Онтологическая модель предложения казахского языка типа BsTlBn

На рисунке 1.27 представлены правила определения типа предложения *AnBsTlPsBn* (Анықтауыш, Бастауыш, Толықтауыш, Пысықтауыш, Баяндауыш) в онтологии. Для этого типа предложения должны быть выполнены следующие необходимые и достаточные условия:

$$AnBsTlPsBn \equiv \exists hasNP\ (NA2) \sqcap \exists hasNP\ (NAcc) \sqcap \exists hasVP(VA1) \quad (1.9)$$

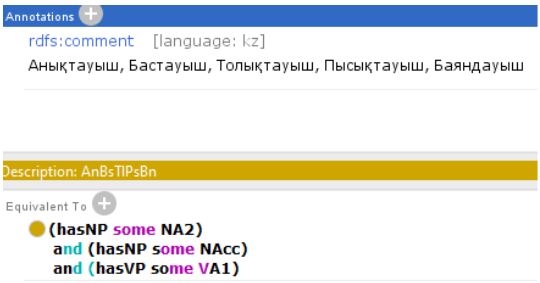

Рис. 1.27. Правила определения типа предложения *AnBsTlPsBn* в онтологии

Если все условия соблюдены, тогда при запуске резонера в среде Protégé, предложение «*Уздік оқушы сабақты тез түсінді*»,

который является индивидом концепта Sentence (предложение) определяется как тип предложения *AnBsTlPsBn*. Онтологическая модель предложения казахского языка типа *AnBsTlPsBn* показано на рисунке 1.28.

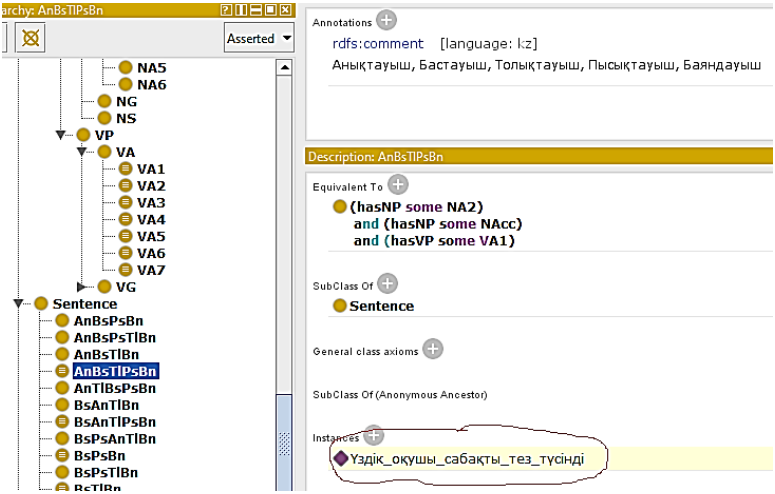

Рис. 1.28. Онтологическая модель предложения казахского языка типа *AnBsTlPsBn*

На рисунке 1.29 представлены правила определения типа предложения *BsAnTlPsBn* (Бастауыш, Анықтауыш, Толықтауыш, Пысықтауыш, Баяндауыш) в онтологии. Для этого типа предложения должны быть выполнены следующие необходимые и достаточные условия:

$$BsAnTlPsBn \equiv \exists hasNP\ (N) \sqcap \exists hasNP\ (NA1Loc) \sqcap \exists hasVP(VA4) \quad (1.10)$$

Если все условия соблюдены, тогда при запуске резонера в среде Protégé, предложение «*Абай қалалық дебатта жақсы сөйледі*», который является индивидом концепта Sentence (предложение) определяется как тип предложения *BsAnTlPsBn*. Онтологическая модель предложения казахского языка типа *BsAnTlPsBn* показано на рисунке 1.30.

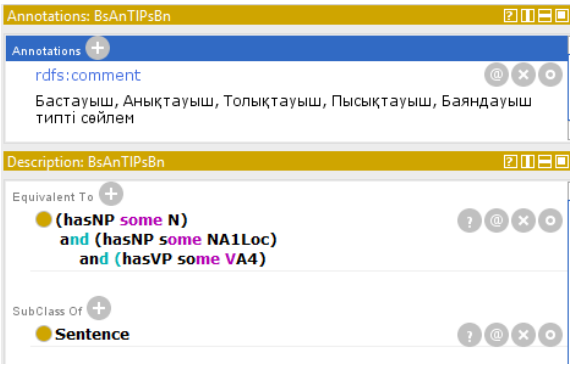

Рис. 1.29. Правила определения типа предложения *BsAnTlPsBn* в онтологии

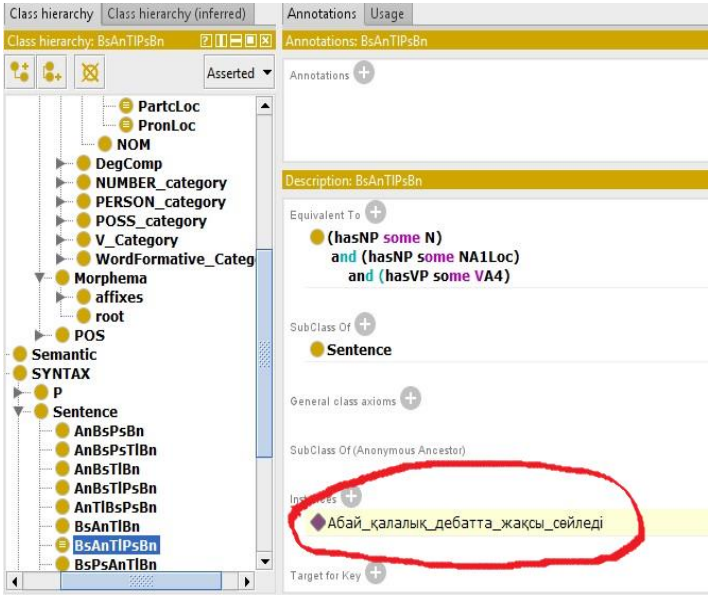

Рис. 1.30. Онтологическая модель предложения казахского языка типа *BsAnTlPsBn*

Известно, что формальная грамматика Хомского позволяет описать синтаксис заданного языка [42], а онтологические модели не только его синтаксис, но и семантику.

Заключение

Полученные результаты можно обобщить путем применения метаязыка для описания грамматических категорий и построения онтологических моделей морфологических и синтаксических правил других тюркских (туркменский, каракалпакский, якутский, башкирский, хакасский и других) языков. Созданный метаязык позволяет унифицировать обозначения концептов всех тюркских языков, а онтологические модели грамматики могут применяться при построении универсальных морфологических анализаторов и синтезаторов для всех тюркских языков.

Разработанный единый метаязык и онтологические модели понятий грамматики казахского языка охватывают все признаки морфологии и синтаксиса казахского языка, однако могут потребовать дополнения и уточнения в случае обнаружения новых правил.

Описанные результаты в дальнейшем могут быть использованы и планируются к использованию при построении онтологических моделей морфологии и синтаксиса, тезаурусов грамматики кыргызского, татарского, турецкого и узбекского языков. Также они могут быть использованы для описания морфологических, синтаксических категорий и онтологических моделей грамматики азербайджанского, балкарского, башкирского, каракалпакского, ногайского, хакасского, туркменского, якутского и других тюркских языков.

Благодарность

Научно-исследовательская работа проведена в рамках проекта «AP05132249 Разработка электронных тезаурусов тюркских языков для создания систем многоязычного поиска и извлечения знаний» по договору №132 от «12 « марта 2018 г.

Литература

1. Залевская А.А. Введение в психолингвистику: учебник. – М.: Российск. гос. гуманит. ун-т, 2000. – 382 с.

2. Marcus, Mitchell P., Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of English: the Penn Treebank // Computational Linguistics. – 1993. – 19(2). – С. 313–330.
3. Garside, R. The CLAWS Word-tagging System// In: R. Garside, G. Leech and G. Sampson (eds), The Computational Analysis of English: A Corpus-based Approach. – London: Longman, 1987.
4. The Open American National Corpus [Электр.ресурс]. – 2018. – URL: http://www.anc.org (дата обращения: 10.10.2018).
5. Ide, N. The American National Corpus: Then, Now, and Tomorrow // Selected Proceedings of the 2008 HCSNet Workshop on Designing the Australian National Corpus: Mustering Languages, Cascadilla Proceedings Project, Sommerville, MA. – 2008.
6. Jurafsky, Daniel and James H. Martin. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. – 2nd Edition. – Prentice-Hall, 2009. – 988 p.
7. Nitin Indurkhya and Fred J. Damerau. Handbook of Natural Language Processing. – 2nd ed. – Chapman & Hall / CRC, 2010. – 704 p.
8. George A. Miller. WordNet: A Lexical Database for English // Communications of the ACM. – 1995. – Vol. 38. – No. 11. – P. 39-41.
9. Цуканова Н. И. Онтологическая модель представления и организации знаний. – М.: Горячая линия – Телеком, 2015. – 272 с.
10. Protégé [Электрон.ресур]. – 2019. – URL: http://protege.stanford.edu (дата обращения: 10.10.2019).
11. Kazakh Language Corpus [Электрон.ресур]. – 2018. – URL: http://kazcorpus.kz/ (дата обращения: 10.09.2018).
12. Makhambetov O., Makazhanov A., Yessenbayev Zh., Matkarimov B., Sabyrgaliyev I., and Sharafudinov A. 2013. Assembling the Kazakh Language Corpus // In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 2013. – P. 1022–1031.
13. Нурхан А.К., Рахимова Д.Р. Исследование и создание размеченного корпуса текстов для казахского языка // Сборник материалов Шестой Международной конференции по компьютерной обработке тюркских языков «TurkLang-2018» (Ташкент, Узбекистан, 18-20 октября 2018 г.), 2018. – С. 127-133
14. Мадиева Г.Б., Уматова Ж.М. Об Алматинском корпусе казахского языка // Вестник КазНУ. Серия «Филология». – Алматы, 2015. – №5 (157). – С. 99-103.
15. Татарский национальный корпус «Туган тел» [Электр.ресурс]. – 2018. – URL: http://tugantel.tatar (дата обращения: 10.10.2018).
16. Turkish National Corpus (TNC) [Электрон.ресур]. – 2018. – URL: http://www.tnc.org.tr/(дата обращения: 10.10.2018).
17. Kubedinova L., Gatiatullin A. Morphological tagging of crimean tatar electronic corpus // Proceedings of the international conference «Turkic languages processing» TurkLang-2015. – Kazan, Tatarstan, 2015. – P. 331-337.
18. Zheltov P. Morphological annotation system for the national corpus of the chuvash language //Proceedings of the international conference «Turkic languages processing» TurkLang-2015. – Kazan, Tatarstan, 2015. – P. 328-331.

19. Sharipbay A., Mukanova A., Yergesh B., Zhetkenbay L., Zulkhazhav A., Yelibayeva G. Ontology modeling of morphological rules of the kazakh and turkish languages // Abstract of the VI international conference «modern problems of applied mathematics and information technology – al-Khorezmiy 2018». – Tashkent, Uzbekistan, 2018. – P. 51-52.
20. Zhetkenbay L., Sharipbay A., Bekmanova G., Kamanur U. Ontological modeling of morphological rules for the adjectives in Kazakh and Turkish languages // Journal of Theoretical and Applied Information Technology, 2016. – Vol. 91. – No.2. – P. 257- 263.
21. Aripov M., Sharipbay A., Abdurakhmonova N., Razakhova B. Ontology of grammar rules as example of noun of Uzbek and Kazakh languages // Abstract of the VI international conference «modern problems of applied mathematics and information technology - al-Khorezmiy 2018». – Tashkent, Uzbekistan, 2018. – P. 37-38.
22. Шарипбай А.А., Ергеш Б.Ж., Елибаева Г.К., Жеткенбай Л., Исраилова Н., Бакасова П. Сравнение онтологических моделей существительных казахского и кыргызского языков // Сборник трудов VI международной конференции по компьютерной обработке тюркских языков TURKLANG-2018. – Ташкент, Узбекистан. – 2018. – С. 182-188.
23. Шарипбай А., Елибаева Г., Муканова А., Жеткенбай Л. Онтологическое моделирование имени прилагательного казахского языка // Сб. трудов VI международной конференции по компьютерной обработке тюркских языков TURKLANG-2018.
24. Шарипбай А., Адалы Е. (Adalı E.), Бекманова Г., Жеткенбай Л. Морфологические правила глаголов казахского и турецкого языков // III международный научный конгресс «Иностранная филология. Социальная и национальная вариативность языка и литературы», 2018 г.
25. Жеткенбай Л., Шарипбай А., Адалы Е. (Adalı E.), Бекманова Г., Қажымұқан Д., Каманур У. Сравнение морфологических правил глагола казахского и турецкого языков // Вестник КазНУ. Серия математика, механика, информатика. – Алматы, 2018. – N.4(100) . – С. 42-51.
26. Жеткенбай Л., Шәріпбай А.Ә. Елибаева Г.К., Муканова А.С., Ергеш Б.Ж. Қазақ және түрік тілдерінің зат есімінің онтологиялық моделі // ҚазҰТЗУ Хабаршысы. – №3. – 2019. – Б. 439-445.
27. Шарипбай А.А., Гатиатуллин А.Р., Ергеш Б.Ж., Қажымұхан Д.А. Разработка единого метаязыка морфологии тюркских языков // Вестник КазНУ. Серия математика, механика, информатика. – Алматы, 2018. – N.4(100) . – С. 78-87.
28. Yelibayeva G., Mukanova A., Sharipbay A., Zulkhazhav A., Yergesh B., Bekmanova G. Metalanguage and Knowledgebase for Kazakh Morphology // Computational Science and Its Applications – ICCSA, 2019. Lecture Notes in Computer Science, vol 11619. Springer, Cham June 2019, DOI: 10.1007/978-3-030-24289-3_51. – Pp. 693-706.
29. «AP05132249 Разработка электронных тезаурусов тюркских языков для создания систем многоязычного поиска и извлечения знаний» [Текст]: отчет о НИР (промежуточ.) / ЕНУ им. Л.Н. Гумилева; рук.

Шарипбай А.А.; исполн. Муканова А.С. и др. – Астана, 2018. – 57 с. – No ГР 0118РК00656. – Инв. №0218РК01298.
30. «AP05132249 Разработка электронных тезаурусов тюркских языков для создания систем многоязычного поиска и извлечения знаний» [Текст]: отчет о НИР (промежуточ.) / ЕНУ им. Л.Н. Гумилева; рук. Шарипбай А.А.; исполн. Муканова А.С. и др. – Нур-Султан, 2019. – 63 с. – No ГР 0118РК00656. – Инв. №0219РК00281.
31. Gruber, T.R. Toward Principles for the Design of Ontologies Used for Knowledge Sharing // International Journal Human-Computer Studies. – 1995. – Vol. 43. – Pp.907-928.
32. Ысқақов А. Қазіргі қазақ тілі (2 басылымы). – Алматы: Ана тілі, 1991. – 384 б.
33. Қазақ грамматикасы. Фонетика, сөзжасам, морфология, синтаксис. – Астана, 2002. – 784 б.
34. Қазақ тілі (Қысқаша грамматикалық анықтағыш). – Алматы: Мемлекеттік тілді дамыту институты, 2010. – 92 бет.
35. Yelibayeva G., Mukanova A., Zhulkhazhav A, Razakhova B., Sharipbay A. Combined morphological analyzer of the Kazakh language based on ontological modeling // 9th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics 2019. – Poznań, Poland. – P. 241-244.
36. Елибаева Г.К., Муканова А.С., Зулхажав А., Жеткенбай Л., Разахова Б.Ш. «Қазақ тілі морфологиясы» онтологиясынан сұратымдар бойынша білімдерді алу // *Педагогика және психология*. – Алматы, 2018. – N.4 . – С. 64-68.
37. Елибаева Г.К., Муканова А.С., Разахова Б.Ш., Ергеш Б.Ж., Жеткенбай Л. Қазақ тіліндегі есімді қабыса байланысқан сөз тіркестерін лингвистикалық белгілеу және формалды моделдеу // Вестник Алматинского университета энергетики и связи. – Алматы, 2019. – N.4(47). – С. 230-236.
38. Sharipbay A., Razakhova B., Mukanova A., Yergesh B, Yelibayeva G. Syntax parsing model of Kazakh simple sentences // In proceedings of the Second International Conference on Data Science, E-Learning and Information Systems (DATA '19). – Dubai, 2019. – Article 54. – 5 p. DOI: https://doi.org/10.1145/3368691.3368745.
39. Оразбаева Ф., Джунусбекова К., Ахметов А. Казахский язык (синтаксис): учебное пособие). – Алматы, 1998. – 191 с.
40. Sharipbaev A.A. Formalization of syntactic rules of the Kazakh language/ A.A. Sharipbaev, B.Sh. Razakhova // Вестник. Специальный выпуск. – Астана: ЕНУ им. Л.Н. Гумилева, 2012. – С. 42-50.
41. Шарипбаев А.А. Контексті бос грамматика арқылы қазақ тілі сөйлемдер жиынының анықталуы // Қазақстан Республикасы Ұлттық ғылым академиясының Баяндамасы. – Алматы, 2005. – № 5. – Б.123-128.
42. Chomsky N. Syntactic Structures. – Berlin - New York: Mouton de Gruyter, 2002. – 118 p.

Шарипбай А.А.
Д.т.н., профессор, Евразийский национальный университет имени Л.Н. Гумилева, ОО «Казахстанская академия искусственного интеллекта», Нур-Султан, Казахстан, e-mail: sharalt@mail.ru;

Муканова А.С.
PhD, доцент, Евразийский национальный университет имени Л.Н. Гумилева, ОО «Казахстанская академия искусственного интеллекта», Нур-Султан, Казахстан, e-mail: asel_ms@bk.ru;

Ергеш Б.Ж.
PhD, Евразийский национальный университет имени Л.Н. Гумилева, ОО «Казахстанская академия искусственного интеллекта», Нур-Султан, Казахстан, e-mail: b.yergesh@gmail.com;

Разахова Б.Ш.
к.т.н., доцент, Евразийский национальный университет имени Л.Н. Гумилева, ОО «Казахстанская академия искусственного интеллекта», Нур-Султан, Казахстан, e-mail: utalina@mail.ru;

Елибаева Г.К.
PhD докторант, Евразийский национальный университет имени Л.Н. Гумилева, ОО «Казахстанская академия искусственного интеллекта», Нур-Султан, Казахстан, e-mail: gaziza_y@mail.ru.

Глава 2

К ВОПРОСУ О РАЗРАБОТКЕ КОРПУСА КАЗАХСКОГО ЯЗЫКА

Аннотация. *Настоящая работа посвящена проблемам построения национального корпуса языков, в том числе корпуса казахского языка. Формирование корпуса на протяжении XX в. – актуальная задача многих современных мировых сообществ, поскольку государственному статусу языка могут соответствовать не только кодифицированные языки, но и все его реализации во всех его стилях и жанрах. В связи с этим актуальной является идея создания национального корпуса казахского языка, которая в настоящее время не нашла своей полной реализации.*

Введение

В рамках Государственной программы функционирования и развития языков (2011-2020) появилась острая необходимость создания Национального корпуса казахского языка.

Формирование Корпуса – актуальная задача многих современных мировых сообществ, поскольку государственному статусу языка могут соответствовать не только кодифицированные языки, но и все его реализации во всех стилях и жанрах. Как отмечает один из создателей Национального корпуса русского языка В. Плунгян, корпус языка – это эффективный и полезный инструмент, особенно в том случае, когда корпус является большим по объему и полным по охвату материала, т.е. представляет собой Национальный корпус языка <…>. Корпус языка – это, в первом приближении, собрание текстов на данном языке, представленное в электронной форме и снабженное научным аппаратом. Аппарат, «встроенный» в корпус, называется «разметкой», или «аннотацией»; корпус тем лучше, чем полнее и совершеннее его аннотация» [1]. Создание корпуса – длительный, трудоемкий процесс, который создается усилиями многих центров и институтов при поддержке государственных программ и информационных ресурсов.

Большинство весомых по своей функциональной значимости, сфере распространения языков мира имеют свои национальные корпуса, различающиеся по полноте и уровню научной обработки текстов. Общепризнанным образцом является Британский национальный корпус (BNC) [2], на который ориентированы многие современные корпуса. Среди корпусов славянских языков выделяется Чешский национальный корпус [3,4]. Как известно, первый большой компьютерный корпус – *Брауновский корпус* (США, 500 фрагментов текстов, 1 млн слов). По его модели создан частотный словарь русского языка под редакцией Л.Н. Засориной (1970), построенный на основе корпуса текстов в 1 млн слов «Словарь представляет собой свод статистических данных о лексическом составе современного русского языка. Словарь составлен на основании обработки средствами вычислительной техники одного миллиона словоупотреблений, что дало около 40 тыс. единиц словаря» [5], а также русский корпус, созданный в Университете Уппсалы (Швеция), который был первым русскоязычным корпусом, созданным в 1980-е годы [6]. Развитие информационных технологий и компьютерных мощностей, способных работать с большими объемами текстов, позволили в 80-е годы XX в. предпринять попытки создать корпуса большего размера: Банк Английского (COBUILD корпус, это сборник английских текстов, в основном, британских, в который включены американские и австралийские тексты. В 2005 году Корпус насчитывал 525 миллионов слов; «Банк английского языка доступен только для небольшой группы исследователей из Университета Бирмингема. Подавляющее большинство людей, которые используют данные Банка Англии за 1990-е годы (период, обсуждаемый ниже), будут делать это через WordBanks Online») [7], Британский национальный корпус, Машинный фонд русского языка (программа комплексной информатизации исследований в русистике, разработанная в начале 1980-х гг. А.П. Ершовым и Ю.Н. Карауловым в Институте русского языка им. В.В. Виноградова Российской академии наук) и др. [8]. Формирование фонда текстов в электронном формате значительно облегчило задачу создания представительных корпусов объемом в десятки и сотни миллионов слов. Однако проблемы по созданию корпусов остаются актуальными, т.к. необходимо решить такие задачи, как инвентари-

зация большого количества текстов, снятие проблем с авторскими правами, приведение текстов в единый формат, классификация корпуса по темам, стилям, жанрам, снятие омонимии. Представительные корпуса существуют (или разрабатываются) для многих языков мира: финского, польского, лезгинского, турецкого, словенского, немецкого, армянского, японского, болгарского и др. [10]. Например, Национальный корпус русского языка содержит более 300 млн словоупотреблений.

2.1. Исследования в области линвистического корпуса

В Казахстане существуют попытки создания Национального корпуса [9], однако, его наполнение до сих пор не достигло ожидаемого, даже минимального, результата. На настоящий момент в рамках научно-исследовательского проекта учеными Казахского национального университета им. аль-Фараби разработана версия корпуса казахского языка в 40 млн словоупотреблений, который был назван как *Алматинский корпус казахского языка* (АККЯ). Эта версия Корпуса создавалась совместными усилиями с учеными Научно-исследовательского университета *Высшая школа экономики*. Для корпуса была адаптирована поисковая система Восточноармянского национального корпуса. Однако размер в 40 млн слов достаточен для лексикографического описания самых частотных слов [10]. В связи с этим необходимо совершенствование Корпуса, увеличение объема текстов различного жанра и стилей, улучшение его поисковой системы, качественной разработки разметок, метаразметок, снятия омонимии, расширение контекста и мн. др. Он должен быть сбалансированным и представительным по объему (сотни миллионов словоупотреблений), оснащенный всеми возможными видами полной и удобной разметки.

Создание полномасштабного корпуса позволит изучать историю казахского языка, обучать и обучаться казахскому языку, осуществлять статистический мониторинг функционирования лексических, грамматических и стилистических языковых средств, работать по лексикографической поддержке современного казахского языка, его стандартизации, создавать словари, учебники, справочные пособия, проводить статистический анализ различ-

ных языковых единиц. АККЯ (Алматинский корпус казахского языка) способен служить современным источником кодификации и стандартизации казахского языка, поскольку в корпусе оказывается зафиксированным письменный язык в его максимально репрезентативном виде. В перспективе намечается фиксировать и звучащую речь.

Казахский язык, его особенность и уникальность, история и современность, будущее, вероятность кардинальных изменений – все это является важным и актуальным не только для лингвистов, но и для специалистов многих отраслей: культуры, экономики, истории, политики и т.д., в том числе языковой. Подобные вопросы обсуждаются и решаются в настоящее время с помощью такого механизма, как корпус языка.

Создание национальных корпусов государственных языков ведущих стран мира возведено в ранг важных историко-культурных и политических мероприятий современности. Понятие *корпус* многими отождествляется с понятием «набора текстов или языковых единиц», что не дает необходимой теоретико-методологической базы для того, чтобы рассматривать корпус не только как феномен, обладающий определенным набором характерологических свойств и признаков, свойственных разным типам, стилям любого языка, но и как феномен идиоэтнического порядка, определяемый особенностями национальной ментальности. Эта проблема с помощью Корпуса была решена для многих хорошо изученных языков мира (английского языка, американского варианта английского языка, немецкого, русского, французского, польского и др.). Формирование Национального корпуса казахского языка – одна из важнейших задач суверенного Казахстана. В рамках государственной программы планируется создание корпуса казахского языка масштабного по объему текстов и тематике подкорпусов, который будет востребован не только отечественными, но и зарубежными потребителями с целью исследования казахского языка, его изучению и обучению. Для реализации этой задачи и был создан АККЯ. Уже сейчас можно говорить о том, что заметен научный и учебный интерес к этому корпусу: его используют при написании докторских, магистерских диссертаций, дипломных проектов. Настоящий корпус используется также в учебных целях в образовательной программе

Компьютерная лингвистика на занятиях по компьютерной, корпусной лингвистике, машинной обработке большого массива языковых данных (например, в учебной дисциплине *Language resources and databases*) и т.п.

Национальный корпус языка – неоценимый инновационный инструмент, сокращающий затраты времени на техническую работу по изучению языковых явлений и за считанные минуты дающий возможность найти справочную информацию. АККЯ – это не просто техническая поддержка лингвистических исследований. Это справочно-информационная база по современному казахскому языку, позволяющая получать ответы на многие вопросы, которые возникают перед любым потребителем, изучающим казахский язык, а также ставить новые проблемы, которые не входили в круг проблем лингвистики прошлых лет, оптимизировать и усовершенствовать работу с языковым материалом. Так, создатели корпуса русского языка отмечают, что с его помощью появилась возможность «поиска по грамматическим критериям автоматически получить примеры, из которых уже вручную можно выбрать наиболее подходящие для наших целей… Корпус дает возможность отсортировать источник примеров». Еще одно преимущество Корпуса – скорость подбора примеров и источников [11]. Кроме того, «действительно, пока основная масса пользователей Корпуса – ученые-исследователи; огромный резерв здесь составляют преподаватели и учащиеся самых разных уровней – от школ до университетов, подготовительных курсов, курсов усовершенствования или второго высшего образования…, где Корпус фактически служит активным инструментом обучения грамматике, стилистике, культуре речи и всему комплексу дисциплин, связанных с русским языком...» [12].

В рамках казахского языкознания и прикладной лингвистики исследование и разработка НККЯ, в том числе и АККЯ, представляет особый интерес, что определяется недостаточной разработанностью проблематики в данной области. Так, считая исследование корпуса казахского языка логическим продолжением традиции его изучения, тем не менее, можно апеллировать только к констатации Корпуса казахского языка, к достаточно ограниченному числу работ, посвященных описанию данного феномена, но не его наличию в полном объеме. Несмотря на достижения

в этой области (попытка составления корпуса с необходимыми разметками, наличие множества научных исследований в виде монографий, диссертаций, учебников казахского языка), границы исследований не выходят за рамки традиционного языкознания, что ограничивает усилия по разработке корпуса или сводит их к механистическому выявлению отличий казахского языка. Нужен современный исследовательский механизм и практический инструмент, которым будет Корпус казахского языка. Помимо этого, обучение языку при помощи компьютерных технологий отходит от традиционных способов подачи материала и фокусируют внимание на тех видах деятельности, которые стимулируют новые подходы, например, аутентичные тексты, к которым можно получить доступ в языковом корпусе. Эти факты свидетельствуют об объективной реальности и актуальности проблемы создания Национального корпуса казахского языка. К настоящему времени казахстанская корпусная лингвистика как научное и прикладное направление не получила своего должного развития.

Обозначенные выше научные лакуны, связанные с разработкой Национального корпуса казахского языка, определяют актуальность предпринятого научного и прикладного исследования.

В результате реализации проекта в казахстанской лингвистике проведено планомерное исследование зарубежного опыта по корпусной лингвистике, формируется отечественное направление по корпусной лингвистике, запланировано создание значительной текстовой базы для наполнения контента Корпуса казахского языка, употребляемого в различных видах дискурса с использованием методов и основных принципов корпусной лингвистики. В рамках проекта обобщается практический и теоретический опыт использования различных Корпусов мира в преподавании языков в Казахстане; изучается казахский язык функционально: в науке, технике, экономике, культуре и т.п. в синхронном срезе на широком фоне социальной, культурной, политической жизни.

Настоящее исследование актуально в рамках внедрения образовательной программы «Вычислительная лингвистика» по проекту Development of the interdisciplinary master program on

Computational Linguistics at Central Asian universities CLASS, выполняемое в рамках программы Эрасмус+. Проект направлен на разработку образовательной программы по специальности «Вычислительная лингвистика» для второй ступени обучения – магистратуры на основе анализа существующих учебных программ в зарубежных вузах с учетом международного опыта. В проекте принимают участие ученые из университетов разных стран: 1. Казахстана – ЕНУ им. Л. Гумилева, Костанайского университета им. А. Байтурсынова, КазНУ им. аль-Фараби; 2. Узбекистана – Ургенчского государственного университета, Самаркандского государственного института иностранных языков и Ташкентского государственного университета узбекского языка и литературы; 3. вузов дальнего зарубежья – University of West Attica, Adam Mickiewicz University in Poznań, University of Porto, University of a Coruna, University of Santiago de Compostela. Создана уникальная возможность для развития единого информационно-образовательного пространства Центральной Азии.

Результаты проекта актуальны в рамках поиска путей совершенствования форм и методов образования студентов, школьников и других слоев населения на основе компьютерной компетентности и проблемно-ориентированного обучения. Проект реализует принципы национальной политики в сфере развития государственного языка и сохранения его богатства; ориентирован на обновление содержания образовательного процесса на основе инновационно-информационных технологий, создание креативных платформ, в которые может быть инсталлирована и на которых могут быть продемонстрированы достоинства АККЯ как открытого информационно-образовательного портала.

2.2. Описание метода

Реализация проекта активизирует новые формы использования компьютерных технологий, формирование у пользователей (специалистов, студентов, магистрантов, докторантов, школьников, учителей и т.п.) новых форм диалога «пользователь-компьютер» для различных видов работ с казахским языком; главное условие эффективности – скорость поиска, устранение

механистической работы с различными текстами, поиском форм слов, подбора необходимого слова в различных реализациях и контекстах, что позволит решить ряд вопросов и достичь существенного эффекта экономии в исследовании, изучении казахского языка, а также коллекционировать его разнообразие и богатство.

Для того чтобы разработать платформу Корпуса и составить репрезентативную текстовую базу для его наполнения в ходе выполнения проекта используется совокупность лингвистических методов:

– выборка и систематизация текстов, инвентаризация текстов по хронологическим, жанровым и стилевым критериям;

– графематический анализ, позволяющий выделить синтаксические и структурные единицы входного текста (абзацы, предложения, словосочетания, отдельные слова, знаки препинания);

– морфологический анализ, предполагающий определить структуру слова, основное слово и его словоформу, отнесение к той или иной части речи для дальнейшей процедуры снятия омонимии;

– синтаксический анализ, позволяющий определить функцию слова в составе предложения, его сочетаемость с другими словами, порядок слов в предложении;

– семантический анализ, необходимый для анализа текста по смыслу, уточнения связи слов, исключающий бессмысленный набор слов.

В современных условиях в данной отрасли значительную помощь оказывает привлечение методов разработки современных корпусов:

– морфологическая разметка: полная морфологическая характеристика каждой словоформы с возможностью определения спорных случаев, имеющих неоднозначное понимание;

– синтаксическая разметка: выделение различных типов синтаксических единиц (предложение, словосочетание);

– семантическая разметка: информация о семантических категориях казахского языка;

– метаразметка (метаинформация о типе текста и его выходных данных).

Кроме того, использованы методы обработки естественного языка (natural language processing (NLP): лексикографическая обработка, токенизация, лемматизация, морфологический анализ) и другие с целью разработки *автоматизированного извлечения информации*; текстовые поиски в крупномасштабных корпусах (конкордансы).

Корпусные методы широко зарекомендовали себя в мировой практике составления корпуса языка, лингвистических исследованиях и преподавании иностранных языков как эффективные инновационные дополнения к традиционным образовательным технологиям.

Содержание, значение, функции этих методов раскрываются на занятиях по таким дисциплинам, как Корпусная лингвистика и компьютерные инструменты, Компьютерная лингвистика, Морфологическая обработка текстов и машинное обучение, Методы и алгоритмы компьютерной лингвистики, которые включены в учебный план образовательной программы Компьютерная лингвистика, предназначенной для магистрантов с лингвистическим образованием. АККЯ используется в этом плане как лингвистический ресурс и инструмент для научных исследований, поскольку АККЯ интересен как для специалистов, преподавателей казахского как родного и иностранного.

Магистранты знакомятся с прикладным значением АККЯ как с информационно-справочной базой, у которой имеются следующие уникальные возможности:

1) обеспечение создания учебников и учебных пособий по казахскому языку текстовым материалом; обеспечение в электронном виде разносторонним языковым материалом процесс обучения казахскому языку;

2) многократное упрощение и ускорение процедур лингвистической обработки массивов текстов на основе современных компьютерных технологий;

3) развитие современных знаний о казахском языке: возможность статистической обработки текстов с целью научного описания строя казахского языка на основе инновационных технологий; формирование базы знаний использования национального корпуса казахского языка;

4) становление теоретико-методологического подхода к организации переводческого процесса;

5) применимость полученных научных результатов: Корпус предоставит широкие возможности для создания различного типа и жанра авторитетных академических и переводных словарей, онлайновых отраслевых, одно-, и двуязычных толковых, терминологических, фразеологических и иных словарей; быстро и эффективно проверять особенности употребления незнакомого слова или грамматической формы у авторитетных авторов и для использования корпусных данных при многих более специальных научных исследованиях.

Алматинский корпус казахского языка уже дает свои первые результаты.

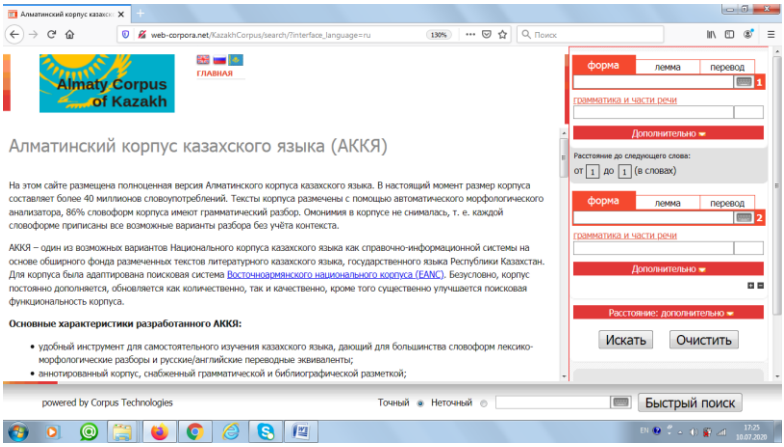

Рис. 2.1. Интерфейс веб-сайта Алматинского корпуса казахского языка

Для корпуса была адаптирована поисковая система Восточноармянского национального корпуса (EANC). http://web-corpora.net/KazakhCorpus/search/ ?interface_language=ru. На этом сайте размещена пилотная версия Алматинского корпуса казахского языка. В настоящий момент размер корпуса составляет более 40 миллионов словоупотреблений. Тексты корпуса были размечены с помощью автоматического морфологического анализатора, 80% словоформ корпуса имеют грамматический разбор.

Во время занятий магистрантам дается информация о том, что тексты АККЯ размечены с помощью грамматического словаря и описания грамматики. Для ускорения разметки программа-анализатор получает на вход не тексты корпуса, а заранее составленный список всех словоформ корпуса, что позволяет каждую словоформу разметить по одному разу. Размеченный список словоформ хранится в виде XML и при необходимости может быть скорректирован вручную (это может быть оправданно для наиболее частотных словоформ в случае, если они получили некорректную разметку). Ниже приведен фрагмент размеченного списка словоформ, который презентуется магистрантам:

<w><ana lex=«да» gr=«CONJ» trans_ru=«и(союз)»></ana>да</w>
<w><ana lex=«бір» gr=«NUM,sg,nom» trans_ru=«один»></ana>бір</w>
<w><ana lex=«де» gr=«CONJ» trans_ru=«вспомогательный глагол»></ana><ana lex=«де» gr=«CONJ» trans_ru=«и(союз)»></ana><ana lex=«деу» gr=«V,imper,2,sg» trans_ru=«говорить; сказать»></ana>де</w>
<w><ana lex=«мен» gr=«PRO,sg,nom» trans_ru=«я»></ana>мен</w>

Размеченный список словоформ и исходные тексты подаются на вход набору специально разработанных скриптов, которые переносят информацию о грамматических разборах в тексты и создают для каждого текста файл формата prs, пригодный для индексации в корпусной платформе. Магистранты могут самостоятельно создавать размеченный список словоформ.

Нужно иметь в виду, что каждой словоформе приписаны все возможные варианты разбора без учёта контекста, т.к. омонимия в корпусе не снималась. Это достаточно трудоемкая задача, которая представляет следующий этап совершенствования корпуса.

Заключение

АККЯ – это одна из версий корпуса казахского языка как справочно-информационной системы на основе обширного фонда размеченных текстов литературного казахского языка, госу-

дарственного языка Республики Казахстан. Безусловно, корпус будет дополняться, обновляться как количественно, так и качественно, кроме того, будет существенно улучшаться поисковая функциональность корпуса. Магистранты, обучающиеся на специальности *Компьютерная лингвистика* уже предпринимают попытки участия в совершенствовании представленного корпуса, выполняют магистерские диссертации по изучению и решению проблем корпуса.

В перспективе основные характеристики АККЯ следующие:

– лингвистически репрезентативный корпус;

– мощный поисковый аппарат для осуществления сложных лексико-морфологических запросов;

– удобный инструмент для самостоятельного изучения казахского языка, дающий для большинства словоформ лексико-морфологические разборы и русские/английские переводные эквиваленты;

– диахронически ориентированный корпус, покрывающий различные периоды истории современного казахского языка;

– диверсифицированный корпус, включающий разножанровые письменные и устные тексты разных типов;

– аннотированный корпус, снабженный грамматической и библиографической разметкой;

– корпус, находящийся в открытом доступе.

– электронная библиотека, включающая более 100 классических произведений казахской литературы.

Благодарность

Работа над проектом Корпуса началась при поддержке ректора КазНУ им. аль-Фараби Г.М. Мутанова. Корпус создается силами кафедры общего языкознания и европейских языков факультета филологии и мировых языков Казахского национального университета им. аль-Фараби под руководством заведующей кафедрой Г.Б. Мадиевой при участии сотрудников факультета филологии НИУ ВШЭ (Москва). Особо хотелось бы отметить вклад в разработку этой версии корпуса к.ф.н. Т.А. Архангель-

ского. Выражаем всем участникам проекта огромную признательность за их нелегкий труд в развитие корпусной лингвистики в Казахстане и международному проекту CLASS, выполняемое в рамках программы Эрасмус+.

Литература:

1. Plungiyan, V.A. Zachem mi delaem Nasionalniy korpus russkogo iyazika? // Otechestvennie zapysky. 2005. – N 2. – P. 296–308 // http://www.strana-oz.ru/2005/2/
2. Британский национальный корпус (BNC) // https://www.english-corpora.org/bnc/
3. Чешский национальный корпус // https://korpus.cz/
4. Nasionalniy korpus russkogo iyazika // http://www.ruscorpora.ru/corpora-intro.html
5. Частотный словарь русского языка (под редакцией Л.Н. Засориной). – М., 1970 // http://project.phil.spbu.ru/lib/data/slovari/zasorina/zasorina.html
6. Захаров В.П. Корпуса русского языка // https://b-ok.asia/book/3061128/2f2520?regionChanged
7. Корпус современного американского английского (COCA) и Word-banks Online // https://www.english-corpora.org/coca/ compare-wbo.asp
8. Русский язык. Энциклопедия русского языка // https:// russkiyyazik.ru/486/
9. http://til.gov.kz/wps/portal
10. https://ru.wikipedia.org/wiki
11. Dobrushina N.R. Kak yspolzobat_Nasionalniy korpus russkogo iyazika v obrazobanyy? // Nasionalniy korpus russkogo iyazika: 2003-2005. – M.: Yndryk, 2005, 308-329.
12. Rahylyna E.B. Korpus kak tvorcheskyy proekt// Nasionalniy korpus russkogo iyazika: 2006-2008. Novie rezultati y perspektyvi. SPb.: Nestor-Ystorya, 2009, 7-26.
13. Алматинский корпус казахского языка // http://web-corpora.net/KazakhCorpus/search/?interface_language=ru

Мадиева Г.Б.
КазНУ им. аль-Фараби, Алматы, Казахстан e-mail: Gulmira.Madiyeva@kaznu.kz

Бектемирова С.Б.
КазНУ им. аль-Фараби, Алматы, Казахстан e-mail: Saule.Bektemirova@kaznu.kz

Мамбетова М.К.
КазНУ им. аль-Фараби, Алматы, Казахстан e-mail: Manshyk.Mambetova@kaznu.kz

Глава 3

РАЗРАБОТКА МЕДИА-КОРПУСА КАЗАХСКОГО ЯЗЫКА

Аннотация. *Целью данной работы является проектирование и разработка медиа-корпуса казахского языка, который представляет собой лингвистический ресурс, доступный на платформе Казахского национального университета им. аль-Фараби. Разрабатываемый медиа-корпус казахского языка состоит из текстов новостного контента и реализован в виде информационно-справочной системы. В разделе описана общая архитектура информационно-справочной системы для автоматического и надежного сбора, хранения и анализа текстов на казахском языке, а также представлены инструменты автоматической предварительной обработки текста для казахского языка. Предлагаемые инструменты могут быть применены в системах автоматического анализа текстов, при создании других языковых ресурсов, таких, как тезаурусы и онтологии.*

3.1. Введение

Национальный корпус языка – неоценимый инструмент, значительно сокращающий затраты на техническую работу по изучению языковых явлений и за считанные минуты дающий возможность найти справочную информацию (статистическую, лексическую, морфологическую, синтаксическую, переводную, иллюстративную, жанровую). Корпус языка – это не просто техническая поддержка лингвистических исследований, это современный инновационный инструмент, справочно-информационная база по искомому языку, позволяющая получать ответы на многие вопросы, которые возникают как перед отечественным, так и зарубежным исследователем, студентом, любым потребителем, изучающим, исследующим, использующим казахский язык. Как отмечают разработчики Национального корпуса русского языка, «национальный корпус обращен ко всем, кто в силу профессии,

по необходимости или из простой любознательности ищет ответ на вопросы об устройстве и функционировании языка, то есть фактически к большинству образованных носителей этого языка и ко всем, изучающим его в качестве иностранного» [1].

Понятие «корпус» многими отождествляется с понятием «набора текстов или языковых единиц», что не дает необходимой теоретико-методологической базы для того, чтобы рассматривать корпус не только как универсальный феномен, т.е. обладающий определенным набором характерных свойств и признаков, присущих разным типам, стилям любого языка, но и как феномен идиоэтнического порядка, определяемый особенностями национальной ментальности, запечатленной в национальной концептуальной картине мира. Эта проблема была решена для многих хорошо изученных языков мира (британского варианта английского языка, американского варанта английского языка, немецкого, русского, армянского, французского, польского, чешского и др.).

Особое место в современной корпусной лингвистике занимают медиа-корпусы. Медиа-корпус языка – информационно-справочная система размеченных медиа-текстов в электронной форме. В базу медиа-текстов включены на основе приема сплошной выборки новостные тексты, опубликованные в средствах массовой информации. Безусловно, медиа-корпус является весьма ценным источником по сбору, анализу какой-либо новостной информации для широкого круга потребителей, которые могут задавать поиск по различным основаниям (ключевым словам, интересующим рубрикам, темам и т.п.). Он может быть и обучающим инструментом для будущих специалистов-журналистов, обозревателей, политиков, специалистов любой медиа-сферы.

В результате работ, выполняемых кафедрой искусственного интеллекта и Big Data совместно с кафедрой общего языкознания и европейских языков факультета филологии и мировых языков Казахского национального университета им. аль-Фараби был разработан медиа-корпус казахского языка на платформе Казахского национального университета имени аль-Фараби (http://corpus.kaznu.kz). На настоящий момент данные для медиа-корпуса собираются с 44 казахоязычных сайтов, из них 10 порталов по чрезвычайным ситуациям, 11 новостных порталов, 13 образовательных порта-

лов, 10 развлекательных ресурсов. Разрабатываемый медиа-корпус казахского языка представляет собой публичный веб-портал, который станет новым инструментом для исследования, анализа, изучения, преподавания казахского языка, предназначенный для широкого круга потребителей на отечественной и мировой арене.

3.2. Развитие тюркской корпусной лингвистики

В рамках казахского языкознания и прикладной лингвистики в Казахстане исследование и разработка Национального корпуса казахского языка представляет особый интерес, что определяется недостаточной разработанностью проблематики в данной области. Несмотря на достижения в этой области (попытка составления корпуса с необходимой разметкой, наличие научных исследований в виде монографий, диссертаций, учебников казахского языка, работы сопоставительного характера, анализирующие отличия разговорного и литературного языков, исследования отдельных его аспектов), поле исследований не выходит за рамки традиционного языкознания, что ограничивает исследовательские усилия по разработке корпуса или сводит их к механистическому выявлению лексических, фонетических и других отличий казахского языка.

Казахский язык относится к классу агглютинативных языков суффиксального типа и входит, как известно, в тюркскую семью языков. Для агглютинативных языков (*agglutinatio* «приклеивание, прилепление») характерно последовательное присоединение различных деривационных или реляционных стандартных, однозначных аффиксов, несущих производное или грамматическое значение, к неизменяемому корню или основе, являющихся носителями лексического значения.

Порядок добавления аффиксов в агглютинативных языках строго определен. Например, в казахском языке для имен существительных к основе слова могут добавляться в строгой последовательности следующие аффиксы: вначале присоединяется словообразующий аффикс, далее окончание множественного числа, затем притяжательное окончание, затем следует падежное окончание, например: *көмек+ші+лер-іміз-ден* (где корневая мор-

фема *көмек* означает «помощь», а образованное при помощи аффикса *-ші* производное слово *көмек+ші* – «помощник», в целом, слово означает «от наших помощников»; последним может присоединяться личное окончание, так как в казахском языке имя существительное может выполнять функцию сказуемого. В таком случае, оно согласуется с подлежащим в лице и числе, поэтому прибавляется личное окончание: *көмекшілердеміз, кабинеттеміз, Гүлнардасың ба*?) [2, 3].

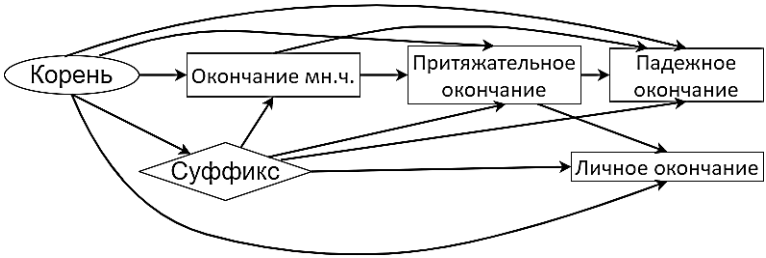

Рис. 3.1. Правило присоединения аффиксов для имен существительных

Тюркская корпусная лингвистика начала интенсивно развиваться лишь с 1990-х годов, поэтому проекты создания общедоступных корпусов тюркских языков особенно актуальны. На сегодняшний день имеется небольшое количество репрезентативных корпусов текстов на тюркских языках. Информация о них доступна на [1], а также на портале Turklang [4]. К наиболее известным корпусам тюркских языков можно отнести:

1) Турецкий национальный корпус объемом 50 миллионов словоупотреблений, который является сбалансированным и репрезентативным корпусом современного турецкого языка. Он состоит из образцов текстовых данных в широком разнообразии жанров, охватывающих период в 20 лет (1990-2009) [5].

2) Башкирский поэтический корпус объемом более 1,8 миллионов словоупотреблений. Он является вторым в мире поэтическим корпусом. Его особенность заключается в том, что корпус состоит из произведений башкирских поэтов XX и начала XXI века [6].

3) Письменный корпус татарского языка объемом более 500 миллионов словоупотреблений при числе различных словоформ – около 5 миллионов [7].

4) Татарский национальный корпус «Туган ел», на 2019 г. объем основного корпуса составляет 194 млн словоформ [8].

5) Национальный корпус Башкирского языка объемом более 20 миллионов словоупотреблений [9].

6) Лингвистический корпус крымскотатарского языка, корпус современного письменного крымскотатарского языка. В состав корпуса входят преимущественно тексты из крымскотатарских газет начала XXI века. В настоящее время корпус содержит 521012 токенов (включая пунктуацию), что составляет около 56752 словоформ [10].

Опыт разработки корпусов тюркских языков положительно повлиял на разработку корпусов казахского языка. К разработанным ранее, но недоступным в настоящее время корпусам казахского языка можно отнести Корпус казахского языка, размещавшийся на портале государственного языка Комитета по языкам Министерства культуры и информации Республики Казахстан [11]; составленный на основе юридических текстов англо-казахский параллельный корпус [12]; Казахский национальный корпус [13].

Из функционирующих в настоящее время разработок корпуса казахского языка можео выделить следующие:

– Национальный корпус казахского языка, созданный усилиями Института языкознания им. А. Байтурсынова. Как отмечается на сайте Института [14], «это первоначальная версия мегапроекта в виде крупномасштабной инновационной открытой информационной системы под названием «Национальный корпус казахского языка», которая содержит 300 миллионов слов электронных текстов казахского языка. В целом корпус текстов, который в несколько сотен раз шире и намного глубже (от средневековья до современности), чем обычная оригинальная версия, представленная в Национальном корпусе казахского языка в идеальной форме, охватывается жанром, стилем, а также обозначениями. В текстовую базу этого проекта были добавлены примеры из 15-томных иллюстраций «Словарь казахского литературного языка». Кроме того, словарный запас объемом 5 млн был взят из текстов художественной литературы, поэзии, драмы, ес-

тественных и гуманитарных наук. Хотелось бы отметить, что в настоящее время функционирует «Қазақ тілі корпусы» по адресу: http://87.255.194.142/, который включает 10 млн словоупотреблений, представляющих тексты пяти стилей языка.

– Корпус казахского языка Kazakh Language Corpus (KLC), созданный силами сотрудников лаборатории National Laboratory Astana Назарбаев Университета совместно с Евразийским университетом им. Л.Н. Гумилева [7]. KLC представляет собой открытый аннотированный казахский корпус, содержащий более 135 миллионов слов и более 400 тысяч документов, классифицированных в пять основных стилистических жанров: литературный, публицистический, официальный, научный и неформальный. Наряду с основным разделом, KLC включает в себя такие разделы, как аннотированный подкорпус, содержащий сегментированные документы с полной морфологической, синтаксической и структурной разметкой текстов, аннотированный речевой подкорпус [15, 16].

– Близкий к полифункциональному корпусу Алматинский корпус казахского языка (АККЯ) [17]. В настоящий момент размер корпуса составляет более 40 миллионов словоупотреблений. Тексты корпуса размечены с помощью автоматического морфологического анализатора, 86% словоформ корпуса имеют грамматический разбор. Омонимия в корпусе не снималась, т.е. каждой словоформе приписаны все возможные варианты разбора без учёта контекста. АККЯ – один из возможных вариантов Национального корпуса казахского языка как справочно-информационной системы на основе обширного фонда размеченных текстов литературного казахского языка, государственного языка Республики Казахстан. Для разработки корпуса была адаптирована поисковая система Восточноармянского национального корпуса (EANC) [18]. Безусловно, корпус постоянно дополняется, обновляется как количественно, так и качественно, кроме того, существенно улучшается поисковая функциональность корпуса.

Проблема создания Национального корпуса казахского языка до сих пор остается актуальной. Одной из задач Государственной программы по реализации языковой политики в Республике Казахстан, принятой на 2020-2025 годы, является реализация

проекта «Национальный корпус казахского языка» [19]. Целью проекта является «разработка и создание Национального корпуса казахского языка как открытой, инновационной, сбалансированной и представительной информационно-справочной системы, оснащенной метаразметкой, аннотированной лингвистической разметкой, поддерживающей функционирование государственного языка и обслуживающей потребности широкого круга пользователей в корректном синхроническом и диахроническом описании казахского языка» [19]. В первый период реализации (2020-2022 гг.) предполагается разработать корпус казахского языка объемом в 30 млн словоупотреблений, во второй период (2023-2025 гг.) – еще 30 млн словоупотреблений. Ответственным за исполнение этого проекта является Министерство культуры и спорта Республики Казахстан.

Проблеме разработки языковых корпусов посвящены ряд конференций, в том числе ICCCI – International Conference on Computational Collective Intelligence, Turklang – International Conference on Computer Processing of Turkic Languages. Авторы данной главы являются регулярными участниками этих конференций [20-23].

Таким образом, разработка корпусов казахского языка является актуальной задачей.

В предлагаемой вниманию читателей работе представлен медиа-корпус казахского языка, реализованный в виде информационно-аналитической системы, доступной по адресу: http://corpus.kaznu.kz/, позволяющей автоматизировать сбор, хранение, поиск и анализ текстов на казахском языке.

3.3. Архитектура медиа-корпуса казахского языка

Для автоматизации сбора, хранения и анализа медиа-текстов на казахском языке была спроектирована и реализована информационная система. Данная система состоит из четырех компонентов:

1) компонент сбора информации;
2) компонент хранения данных;
3) компонент анализа данных;
4) компонент визуализации данных.

Компонентная архитектура информационной системы показана на рисунке 3.2. Использование очередей позволяет системе быть легко масштабируемой и устойчивой к сбоям.

Компонент сбора информации. Задача компонента сбора информации заключается в непрерывном мониторинге сайтов и скачивании новой информации. Компонент реализован с использованием технологий Jsoap, OpenMQ и JavaEE.

Компонент хранения данных. Задача компонента хранения данных заключается в обеспечении стабильного и быстрого доступа к хранилищу данных. В качестве хранилища данных используется NoSQL база данных MongoDB 3.

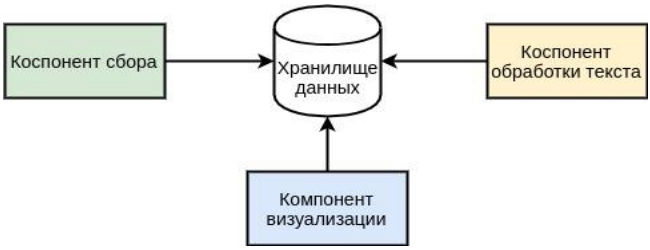

Рис. 3.2. Компонентная архитектура информационной системы

Компонент анализа данных. Задача компонента анализа данных – предобработка накопленных данных. Под предобработкой понимается процесс нормализации текстов, очистки от стопслов, добавление метаданных (табл. 3.1). Также на этом этапе проводится морфологический анализ и разметка текстов. Компонент реализован с использованием технологии Apache Lucene, Apache Spark, Apache Hadoop и MPJ.

Таблица 3.1.

Метаинформация

	Метаданные	**Описание**
1	2	3
1.	date	дата публикации документа
2.	title	заголовок документа
3.	URL	адрес, откуда была взят документ
4.	tags	теги документа

1	2	3
5.	type	тип документа
6.	notificationId	уникальный идентификатор документа
7.	body	текст документа
8.	lang	язык документа
9.	commentInfo	информация про комментарии
10.	commentNumber	количество комментариев
11.	timeStamp	штамп времени
12.	region	регион, указанный в документе
13.	creationDate	дата создания документа
14.	lastChange	время последнего изменения

Компонент визуализации данных. Задача компонента визуализации данных – предоставить пользователю удобный инструмент для поиска по накопленным данным и для отображения поисковой выдачи. Подзадачи компонента – ускорение процесса поиска, улучшение результатов и релевантности поиска, визуализация результатов поиска. Компонент был реализован с использованием технологий HTML5 и Elasticsearch, платформе для обработки данных, построенной на базе Apache Lucene.

3.4. Формат хранения данных

В медиа-корпусе для размеченных текстов применяется язык eXtensible Markup Language (XML). В компоненте сбора информации происходит извлечение релевантного текста из кода HTML страниц с использованием библиотеки Jsoap. Далее данные проходят обработку в компоненте анализа данных. Также на этом этапе проводится морфологический разбор текстов на казахском языке. Морфологический анализатор получает на вход простой текст, а на выходе отдаёт текст в формате XML, с которым в дальнейшем удобно работать, к примеру, легко преобразовать в JSON формат. Формат XML определен при помощи XML Schema Definition (XSD). XSD позволяет эффективно конвертировать данные в любой другой формат, что упрощает обмен данными между системами.

На рисунке 3.3 изображена XSD схема документа, который представляет собой файл описания формата хранения данных.

Эта схема дает возможность удобно обмениваться данными, что является важным при получении размеченных текстов после обработки текстов морфологическим анализатором казахского языка.

На рисунке 3.4 показано выполнение разметки слова в случае, если морфологический анализатор столкнулся с омонимией. Внутри тега «Token», есть атрибут «omonim», в котором указано, имеет ли это слово омоним или нет. В случае если омоним есть, внутри тега «Token» будет два и более тега «Morph», что будет говорить о том, что слово имеет два и более варианта морфологического разбора.

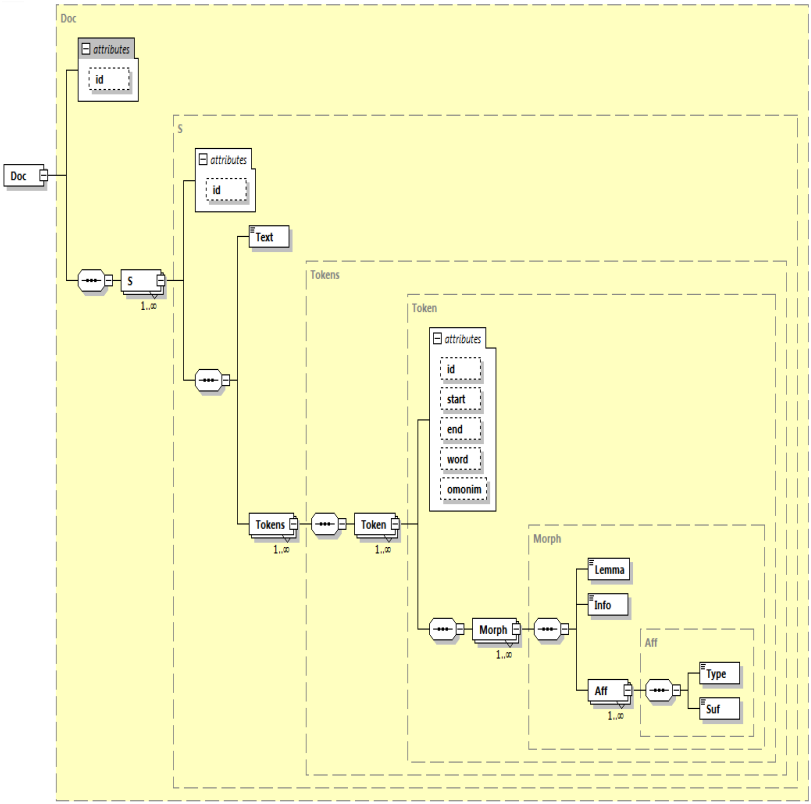

Рис. 3.3. Схема XML-документа

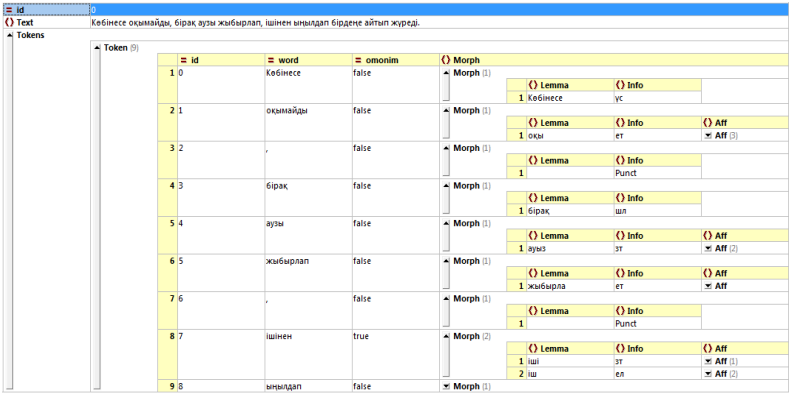

Рис. 3.4. Схема разметки текста

3.5. Морфологическая разметка и постобработка данных

Корпус содержит особую разметку или аннотацию, представляющую собой дополнительную информацию о свойствах входящих в него текстов. Разметка – главная характеристика корпуса; она отличает корпус от простых коллекций (или «библиотек») текстов. Чем богаче и разнообразнее разметка, тем выше научная ценность корпуса [1].

Для постобработки слов в случае неполной морфологической разметки и наличия омонимии разработан специальный интерфейс, с помощью которого эксперт-лингвист может выбрать правильный вариант разбора, или выполнить полную ручную разметку для конкретного слова [24].

На рисунке 3.5 показан интерфейс ручного снятия омонимии, который позволяет выбрать вариант, предложенный морфологическим анализатором, или выполнить новую разметку. При нажатии кнопки «қосу» появляется возможность выбора части речи для анализируемого слова, добавление соответствующего для данной части речи аффикса или нескольких аффиксов. При нажатии кнопки «*тазалау*», можно очистить все аффиксы в случае ошибки и начать добавление сначала. В случае ручного снятия омонимии выбирается переключатель ручной разметки, и добавляются все аффиксы. На рисунке 3.6 показан вариант выполнения ручной разметки.

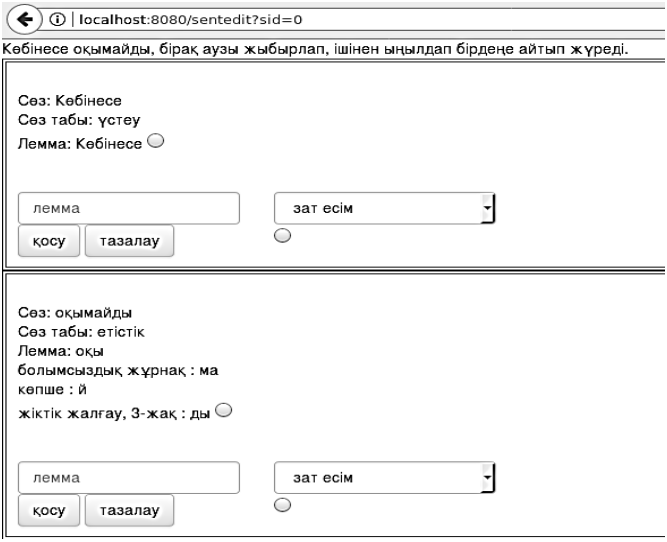

Рис. 3.5. Интерфейс ручного снятия омонимии

Рис. 3.6. Выполнение ручной разметки

3.6 Заключение

В работе описано проектирование и разработка медиа-корпуса казахского языка, состоящего из текстов новостного контента и реализованного в виде информационно-справочной системы. Описана общая архитектура информационно-справочной системы для автоматического и надежного сбора, хранения и анализа текстов на казахском языке. Разработанный медиа-корпус казахского языка позволит:

– предоставлять открытый доступ всем желающим;

– осуществлять поиск по морфологическим параметрам;

– использовать корпус для решения задач обработки естественного языка;

– проводить частотный анализ текстов;

– осуществлять обучение языку, используя переводы слов.

Благодарность

Данная работа выполнена при частичной поддержке МОН РК: проект программно-целевого финансирования О.0856 BR05236340 «Создание высокопроизводительных интеллектуальных технологий анализа и принятия решения для системы «логистика-агломерация» в рамках формирования цифровой экономики РК» (2018-2020 гг.), проект грантового финансирования AP05132933 «Разработка системы извлечения знаний из гетерогенных источников данных для повышения качества принятия решений» (2018-2020 гг).

Литература

1. Национальный корпус русского языка. URL: https://ruscorpora.ru/new/corpora-intro.html
2. Казахский язык. URL: http://www.kaz-tili.kz/lichnie.htm.
3. Грамматический справочник. URL: http://qazaqtili.narod.ru/sprav0.htm
4. Портал по компьютерной обработке тюркских языков TurkLang. URL: http://www.turklang.net/ru/.

5. Национальный корпус турецкого языка, URL: http://www.tnc.org.tr/index.php/en/, дата обращения 2020/07/19; https://ruscorpora.ru/new/corpora-other.html.
6. Башкирский поэтический корпус, URL: http://web-corpora.net/ bash-corpus/search/?interface_language=ru, дата обращения 2020/07/19.
7. Письменный корпус татарского языка, http://corpus.tatar/, дата обращения 2020/07/19;
8. Национальный корпус татарского языка Туган тел. URL:/ http://www.tugantel.tatar/.
9. Национальный корпус башкирского языка. URL: http://bashcorpus.ru/, дата обращения 2020/07/20.
10. Лингвистический корпус крымскотатарского языка // https://korpus.sk/QIRIM/.
11. Портал государственного языка Комитета по языкам Министерства культуры и информации Республики Казахстан. URL: http://til.gov.kz/wps/portal/!ut/p/, дата обращения 2017/03/03.
12. Калдыбеков Т.Е. Англо-Казахский параллельный корпус для статистического машинного перевода // Молодой ученый. – 2014. – №6. – С. 92-95.
13. Қазақстан Республикасы Мемлекеттік Тіл порталы. URL: http://dawhois.com/www/til.gov.kz.html, дата обращения 2020/07/20.
14. Национальный корпус казахского языка. URL: https://tbi.kz/kazcorp
15. Kazakh Language Corpus, URL: http: //kazcorpus.kz/klcweb/en/, last accessed 2020/07/20.
16. Olzhas Makhambetov, Aibek Makazhanov, Zhandos Yessenbayev, Bakhyt Matkarimov, Islam Sabyrgaliyev, and Anuar Sharafudinov. 2013. Assembling the Kazakh Language Corpus. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 1022–1031, Seattle, Washington, USA, October. Association for Computational Linguistics.
17. Алматинский корпус казахского языка. URL: http://web-corpora.net/KazakhCorpus/, дата обращения 2020/07/19.
18. Восточноармянский национальный корпус. URL: http://eanc.net/EANC/search/?interface_language=ru
19. Об утверждении Государственной программы по реализации языковой политики в Республике Казахстан на 2020-2025 годы. URL:
http://prokuror.gov.kz/rus/gosudarstvo/memlekettik-til/ob-utverzhdenii-gosudarstvennoy-programmy-po-realizacii-yazykovoy?language=kk, дата обращения 2020/07/20.
20. Alimzhanov, Y., Mansurova, M. An approach of automatic extraction of domain keywords from the Kazakh text // Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Proc. of the ICCCI-8, Halkidiki, 2016. – P. 555-562.
21. Madina Mansurova, Gulmira Madiyeva, Sanzhar Aubakirov, Zhantemir Yermekov, and Yermek Alimzhanov. Design and Development of Media-

Corpus of the Kazakh Language // Computational Collective Intelligence. N.T. Nguyen et al. (Eds.): ICCCI 2017, Part II, LNAI 10449, pp. 509–518, 2017.
22. Madina Mansurova, Vladimir Barakhnin, Yerzhan Khibatkhanuly, Ilya Pastushkov. Named Entity Extraction from Semi-structured Data Using Machine Learning Algorithms. Computational Collective Intelligence 11th International Conference, ICCCI 2019, Hendaye, France, September 4-6, 2019, Proceedings, Part II (Scopus, Web of Science).
23. Mansurova M., Koibagarov K., Barakhnin V., Soltangeldinova M., Ber-dibekov S. Application of morphological маrkup of Kazakh language to automated filling of the ontology of factographic retrieval system // Известия Кыргызского государственного технического университета им. И. Раззакова Теоретический и прикладной научно-технический журнал. Материалы 4-й Международной конференции по компьютерной обработке тюркских языков «TurkLang 2016». 2016. № 2(38) 61-66 с.
24. Мадиева Г.Б., Мансурова М.Е. Разработка медиа-корпуса казахского языка // Вестник КазНУ. Серия «Филологические науки», 2017 г., № 2. – С. 12-18.

Мансурова М.Е.
КазНУ им. аль-Фараби, Алматы, Казахстан e-mail: Madina.Mansurova@kaznu.kz

Мадиева Г.Б.
КазНУ им. аль-Фараби, Алматы, Казахстан e-mail: Gulmira.Madiyeva@kaznu.kz

Қадырбек Н.Қ.
КазНУ им. аль-Фараби, Алматы, Казахстан e-mail: nurgaliqadyrbek@gmail.com

Қырғызбаева М.Е.
КазНУ им. аль-Фараби, Алматы, Казахстан e-mail: marzhan.kyrgyzbaeva@gmail.com

Глава 4

РАЗРАБОТКА АЛГОРИТМОВ ИЗВЛЕЧЕНИЯ КЛЮЧЕВЫХ СЛОВ И СЕМАНТИЧЕСКОГО АНАЛИЗА ДАННЫХ НА КАЗАХСКОМ ЯЗЫКЕ

Аннотация. *В современном мире цифровых технологий поиск, анализ и обработка текстовых данных являются востребованными процессами в различных сферах деятельности. Извлечение ключевых слов и семантический анализ естественного языка является актуальной задачей классификации, кластеризации, абстрагирования текста и поиска информации. Данная работа посвящена изучению и решению задачи семантического анализа текстов на казахском языке. Для решения этой задачи были изучены и проанализированы исследования в данной области на примерах обработки других естественных языков. Разработаны алгоритмы по извлечению ключевых слов и словосочетаний из текста и семантического анализа казахского языка на основе машинного обучения. Представлены практические результаты экспериментов.*

4.1. Разработка алгоритма извлечения из документов на казахском языке

В настоящее время объемы и динамика информации, которая подлежит обработке в лексикографии и информационном поиске, делают особенно актуальной задачу автоматического извлечения ключевых слов и словосочетаний, которые могут использоваться для создания и развития терминологических ресурсов, а также для эффективной обработки документов: индексирования, реферирования, кластеризации и классификации.

Анализ огромного количества данных может быть упрощен, если у нас будут ключевые слова или словосочетания, которые могут предоставить нам основные характеристики, концепцию и т.д. документа. Соответствующие ключевые слова и словосочетания могут служить кратким изложением документа и помогают нам легко упорядочивать документы и извлекать их на основе их

содержания [1]. Необходимо различать два основных подхода к решению проблемы автоматизации выделения ключевых слов и словосочетаний: назначение ключевых слов и словосочетаний (keyphrase assignment) и их извлечение (keyphrase extraction) [2] [3]. Главное отличие заключается в том, что первый подход позволяет выделять только те ключевые слова и словосочетания, которые содержатся в некотором предусмотренном словаре, а второй подход предполагает выбор ключевой информации непосредственно из текста.

Ключевые слова могут быть назначены вручную или автоматически, но первый подход очень трудоемкий и дорогой. Таким образом, существует необходимость в автоматизированном процессе, который извлекает ключевые слова из документов. Есть готовые программные решения этой задачи для распространенных языков (английский, русский, испанский и т.д.), а для казахского языка только единицы и они не в открытом доступе.

4.1.1. Исследование подходов по извлечению ключевых слов

Методы назначения ключевых слов можно условно разделить на две категории: назначение ключевых слов и извлечение ключевых слов [4, 5, 6, 7]. Оба вращаются вокруг одной и той же проблемы – выбора лучшего ключевого слова. Слова, встречающиеся в документе, анализируются с целью выявления наиболее представительных слов, обычно исследуя свойства источника (то есть частоту, длину) [8]. Существующие методы для автоматического извлечения ключевых слов по предложению Ping-I и Shi-Jen можно разделить на [9]: статистические подходы и подходы машинного обучения.

Необходимо так же акцентировать на четырех категориях, предложенных авторами Zahang и др. в работе [8]: 1) простые статистические подходы; 2) лингвистические подходы; 3) подходы машинного обучения; 4) другие подходы.

Простые статистические подходы включают в себя простые методы, которые не требуют данных обучения. Кроме того, методы не зависят от языка и домена. Статистика слов из документа может быть использована для идентификации ключевых слов:

статистика n-граммы, частота слов, TF-IDF, совпадения слов, дерево PAT (дерево Патрисии; дерево суффиксов или дерево позиций) и т.д. Недостатком является то, что в некоторых профессиональных текстах, таких, как здоровье и медицина, наиболее важные ключевые слова могут появляться только один раз в статье. Использование статистически уполномоченных моделей может непреднамеренно отфильтровать эти слова [9].

В работе [10] предлагается способ оценки терминологичности на базе контрастного подхода. Способ берет за основу известную формулу взвешивания слов TF-IDF, согласно которой вес слова в документе тем выше, чем выше частота его использования в этом документе и чем ниже его разброс по всей коллекции. В новом варианте формулы, который авторы называют «term frequency – inverse domain frequency», оценивается вес слова не в документе, а в целевой коллекции. Согласно новой формуле вес слова тем выше, чем выше относительная частота его использования в целевой коллекции и чем ниже его относительный разброс по всем коллекциям:

$$TF * IDF = TF(t, D) * IDF(t) = \frac{n_{t,D}}{\sum_k n_{k,D}} * \log\left(\frac{|TS|}{|\{d: t \epsilon d\}|}\right), \quad (4.1)$$

где $n_{t,D}$ – это число вхождений слова t в целевую коллекцию D, $\sum_k n_{k,D}$ – это сумма вхождений всех слов в целевую коллекцию D, |TS| – это количество документов во всех используемых коллекциях, $|\{d: t \in d\}|$ – это количество всех документов, в которые слово t входит хотя бы один раз. Таким образом, авторы считают терминами все слова с высокой концентрацией в пределах узкого подмножества документов.

Авторы работы [11] также предлагают оценивать терминологичность слов на базе формулы TF-IDF. Собственный вариант этой формулы они называют контрастным весом (contrastive weight) и определяют его как меру, которая тем выше, чем выше частота употребления слова в целевой коллекции и чем ниже относительная частота его употребления в контрастных коллекциях:

$$Contrastive\ Weight = TF(t,D) * IDF(t) = loglog(f_t^D) * \log\left(\frac{F_{JC}}{\sum_j f_t^i}\right), \quad (4.2)$$

где $f_t{}^D$ – частота употребления слова в целевой коллекции, $\sum_j f_t^j$ – сумма частот всех употреблений слова в контрастных коллекциях, $F_{TC} = \sum_{i,j} f_i^j$ – сумма частот употреблений всех слов во всех коллекциях, включая целевую. Как отмечают сами авторы, контрастный вес значительно лучше оценивает терминологичность слов, чем чистые частоты, однако общая эффективность метода, определенная с помощью F-меры, по их словам, не бросается в глаза.

В работе [12] также развивается идея штрафов и вознаграждений, заложенная в базовой конструкции формулы TF-IDF, и предлагается новый вариант этой формулы, получивший название «term frequency – disjoint corpora frequency». В качестве вознаграждения используется абсолютная частота употребления слова в целевой коллекции, а в качестве штрафа – произведение абсолютных частот употреблений слова в контрастных коллекциях:

$$TF * DCF = \frac{f_t^D}{\prod_{g \in G} 1 + \log\left(1 + f_t^g\right)}, \quad (4.3)$$

где $f_t{}^D$ и $f_t{}^g$ – частоты употреблений данного слова t в целевой и контрастной коллекциях соответственно, G – множество всех контрастных коллекций.

Подходы лингвистики используют лингвистическую особенность слов в основном, предложений и документа. Лексический, синтаксический, семантический и дискурсивный анализ являются одними из наиболее распространенных, но сложных анализов.

Подходы машинного обучения рассматривают контролируемое или неконтролируемое обучение на примерах, но связанная работа по извлечению ключевых слов предпочитает контролируемый подход. Подходы контролируемого машинного обучения создают модель, которая обучается на основе набора ключевых слов. Они требуют ручной аннотации в наборе данных обучения, которая является чрезвычайно утомительной и непоследователь-

ной (иногда запрашивает предопределенную таксономию). К сожалению, авторы обычно присваивают ключевые слова своим документам только тогда, когда они вынуждены это делать. Таким образом, индуцированная модель применяется для извлечения ключевых слов из нового документа. Этот подход включает в себя наивный метод Байеса, *SVM, C4.5, Bagging* и т.д. Таким образом, методы требуют данные обучения и часто зависят от предметной области. Системе необходимо заново изучать и устанавливать модель каждый раз при изменении домена [13, 14]. Индукция модели может быть очень сложной и длительной для массивных наборов данных.

Другие подходы для извлечения ключевых слов в целом объединяют все методы, упомянутые выше. Кроме того, иногда для объединения они включают эвристические знания, такие, как положение, длина, особенности компоновки терминов, HTML и аналогичные теги, форматирование текста и т.д.

Модель векторного пространства (МВП) хорошо известна и является наиболее используемой моделью для представления текста в подходах интеллектуального анализа текста [7, 15, 16]. В частности, документы, представленные в виде векторов признаков, расположены в многомерном евклидовом пространстве. Эта модель подходит для захвата частоты простых слов, однако, структурная и семантическая информация обычно не учитывается. Следовательно, из-за простоты модель векторного про-странства имеет несколько недостатков [17]:1) значение текста и структуры не могут быть выражены, 2) каждое слово не зависит от другого, последовательность появления слов или другие отношения не требуются, 3) если два документа имеют одинаковое значение, но имеют разные слова, сходство не может быть легко вычислено.

Текстовое представление на основе графа известно, как одно из лучших решений, эффективно решающих эти проблемы [17]. График представляет собой математическую модель, которая позволяет очень эффективно исследовать взаимосвязи и структурную информацию. Таксономия основных методов извлечения ключевых слов представлена в иерархической форме на рисунке 4.1 и 4.2.

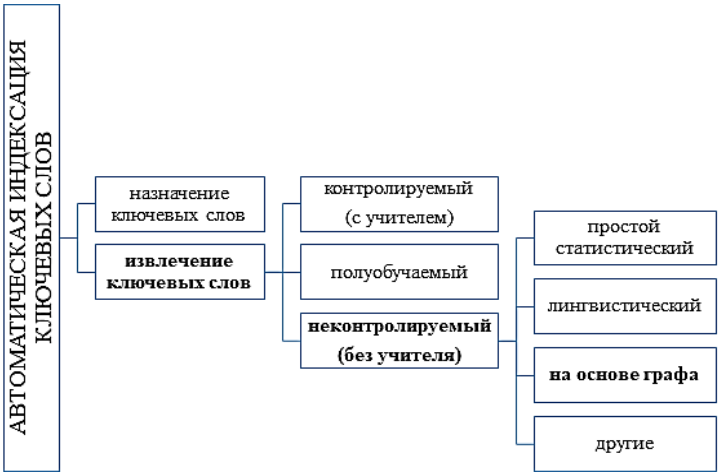

Рис. 4.1. Классификация методов извлечения ключевых слов

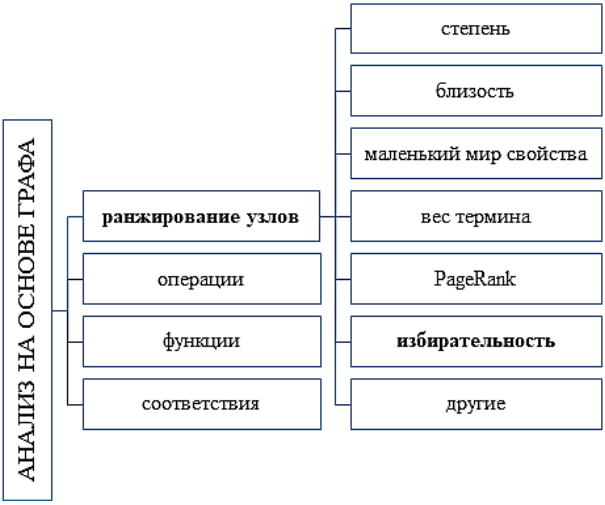

Рис. 4.2. Классификация методов, основанных на графе

Граничное отношение между двумя терминами может быть установлено на многих принципах, использующих различную область текста или отношения для построения графа [17, 18]: 1) слова, встречающиеся вместе в предложении, абзаце, разделе

или документе, добавленные в граф в виде клики; 2) пересекающиеся слова из предложения, абзаца, раздела или документа; 3) слова, встречающиеся в фиксированном окне в тексте; 4) семантические отношения – соединяющие слова, имеющие одинаковое значение, слова, написанные одинаково, но имеющие разное значение, синонимы, антонимы, и т.д.

4.1.2. Разработка алгоритма извлечения ключевых слов, основанного на лингвистических и статистических данных

Алгоритм на рисунке 4.3 извлечения ключевых слов из документов на казахском языке включает в себя 3 этапа:

1) нахождение кандидатов в ключевые слова;

2) выделение признаков;

3) ранжирование.

На первом этапе решаются две задачи: предварительная обработка слов; и разделение текста на отдельные слова и словосочетания.

Первая задача является языко-зависимой, поэтому здесь учитывается морфологическая особенность казахского языка. Для решения этой задачи задействованы система полных окончаний казахского языка (через морфологический анализатор казахского языка разработанный на платформе Apertium выполняем разметку документа), алгоритм стемминга и лемматизации для казахского языка (реализованные на языке программирования Python3). А для второй использован простой подход – процедура токенизации, с помощью которой весь текст разбивается на отдельные слова.

На втором этапе для каждого найденного кандидата в ключевые слова выделяются признаки, по которым оценивается степень его важности. Выделяемые признаки можно разбить на 3 категории: синтаксические признаки, статистические признаки, структурные признаки.

Для выделения синтаксического признака мы использовали морфологический анализ казахского языка. И алгоритм TF-IDF (Term Frequency – Inverse Document Frequency) для определения частотности, то есть для выделения статистических признаков.

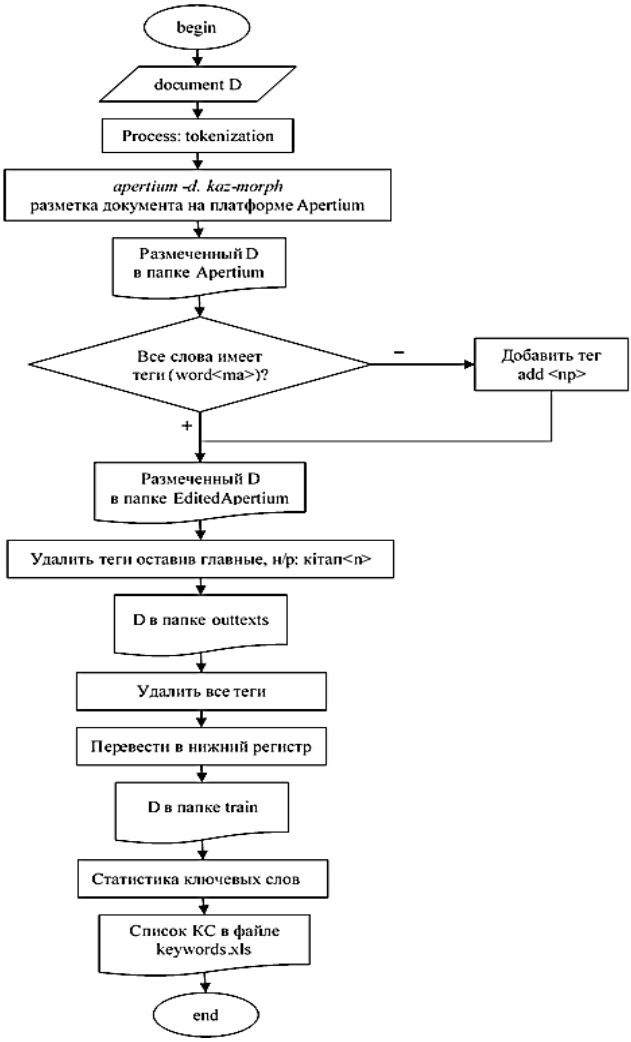

Рис. 4.3. Блок-схема алгоритма извлечения ключевых слов из документов на казахском языке

В программе для биграммного слова указали следующие признаки: существительное+существительное (N+N), прилагательное+существительное (Adj+N), существительное+глагол (N+V),

имя собственное+существительное (Np+N), числительное+существительное (только некоторые слова) (Num+SomeWordsOfTime).

На третьем этапе ранжируем результаты по статистическому признаку и в соответствующей мере к объему текста.

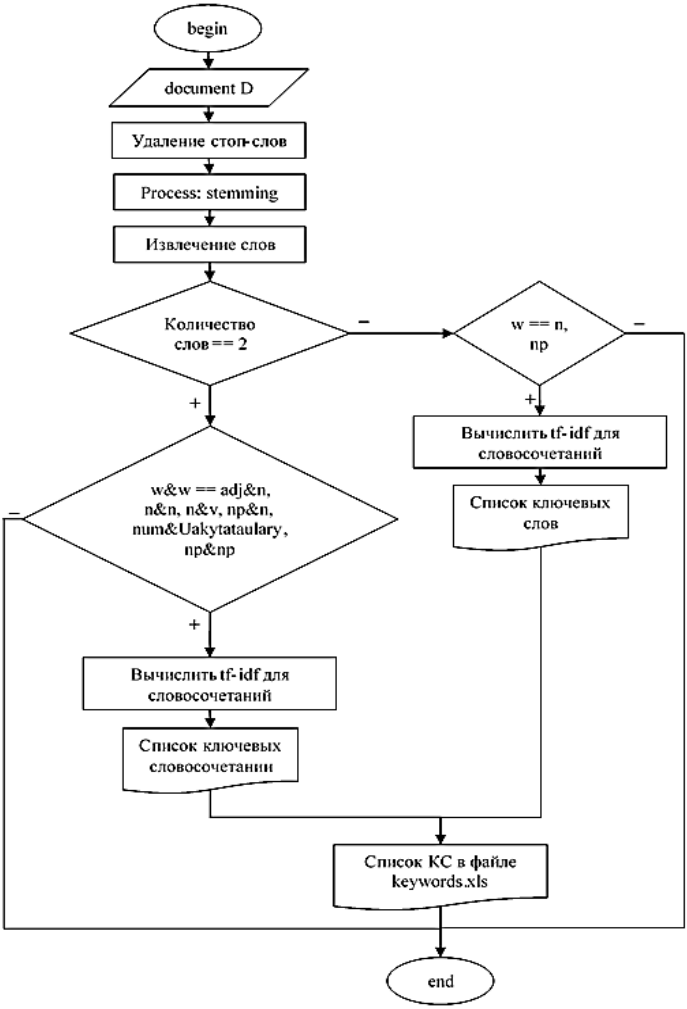

Рис. 4.4. Блок-схема алгоритма вычисления показателя ключевых слов и словосочетаний

4.1.3. Применение подходов по извлечению ключевых слов к текстам на казахском языке

Реализация алгоритма выполнена на языке программирования Python3. Программа тестировалась на подготовленных корпусах различных тематик, приведенных в таблице 4.1. В результате как показано в таблице 4.2 приведены 10 ключевых слов и 5 ключевых словосочетаний по каждому тексту.

Таблица 4.1

Входные данные для тестирования программы

Тексты в корпусе	*Количество предложений*
Қазақстан тарихы (История Казахстана)	484
Манчестер сити	370
Оскар алғандар (Получившие премию Оскар)	309

Таблица 4.2

Полученные результаты тестирования алгоритма извлечения ключевых слов

Тема: Қазақстан тарихы		*Тема:* Манчестер сити	
Ключевые слова и словосочетания	*Показатель tf-idf*	*Ключевые слова и словосочетания*	*Показатель tf-idf*
1	2	3	4
Ғұн	0.013194415852460091	Сити	0.02602130987211911
Тайпа	0.006085581972523395	Манчестер	0.01520522323852743
Қытай	0.004175448054575978	Клуб	0.015048468359779727
Қағанат	0.004175448054575978	Лига	0.005799930513665103
Түрік	0.0038034887328271213	Кубок	0.0053296658774219866
Мемлекет	0.0035182270778650877	Юнайтед	0.005172910998674281
Күлтегін	0.0031733405214777436	Бапкер	0.0048594012411788704
Тонықөк	0.0031733405214777436	Футбол	0.00313509757495411
Ескерткіш	0.002672286754928626	Маусым	0.0029450594945452436
Жорық	0.0033403584436607825	Ойыншы	0.002855815267437812
Түрік қағанат	0.0033403584436607825	Манчестер сити	0.009405292724862329

1	2	3	4
Ғұн мемлекет	0.002839304677111665	Манчестер юнайтед	0.0047026463624311645
Қола дәуір	0.001336143377464313	Есеп жең	0.0029783426962064043
Қазақстан жер	0.0011691254552812739	Есеп жеңіл	0.0021945683024678767
Тас дәуір	0.0010021075330982347	Роберто Манчини	0.0020378134237201712

Полученные результаты экспериментов, выполненных на корпусах, приведенных в таблице 4.3 проанализированы, учитывая граничный коэффициент находки по объему и выявлено, что ключевые слова и словосочетания подобраны правильно. Результаты анализа приведены в таблице 4.4.

Таблица 4.3

Объем корпусов по тематике

Тематика	*Количество предложении*
География	99 659
Космос	12 658
Информатика	26 886
Психология	20 940
Спорт	482
История	37 574
Техника	708

Таблица 4.4

Экспериментальные результаты разработанного алгоритма по определению ключевых слов для казахского языка

Название документа	*Объем документа (предло-жений)*	*Граничный коэффициент ключевых слов*	*Количество ключевых слов*	*Точность нахождения*
1	2	3	4	5
futbol.txt	58	3-8	8	85,71%
boks.txt	59	3-8	8	75%
Samsung.txt	67	3-8	8	87,5%
7tarihitulga.txt	75	3-8	8	100%
LG.txt	97	9-11	10	90%
ginnes.txt	103	9-11	10	80%

1	2	3	4	5
imidj.txt	122	12-14	12	50%
2018biznesmen.txt	152	12-14	12	58,33%
oskaralganjuldyzdar.txt	309	15-17	15	93,33%
GenriFord.txt	343	15-17	15	93,33%
Ttrening	366	15-17	15	86,67%
manchestercity.txt	370	15-17	15	86,67%
BilGeits.txt	452	15-17	15	100%
Kazakhstantarihy.txt	484	15-17	15	100%
AppleInc.txt	531	15-17	15	100%
geoinformatika.txt	821	15-17	15	100%

4.2. Разработка алгоритма семантического анализа текста

Компьютерный семантический анализ тесно связан с задачей понимания текста машиной. Существует много интерпретаций понятия «смысл текста» и задачи его понимания. Например, по Д.А. Поспелову [19] система понимает введенный в нее текст, если с точки зрения человека (или группы экспертов) она правильно отвечает на вопросы, связанные с информацией, заключенной в тексте. Здесь речь может идти не о простом получении фактов, которые явно присутствуют в тексте, а о выявлении скрытых смыслов, которые вносит автор. Д.А. Поспелов выделяет несколько уровней понимания текста, с точки зрения сложности вопросов, на которые должна иметь возможность отвечать интеллектуальная система. Руководствуясь определением из работы [19], можно рассматривать смысл текста как описание знаний, содержащихся в нем, на формальном языке представления знаний, который позволяет решать достаточно широкий круг задач, связанных с анализом текста, а задачу семантического анализа – как трансляцию естественного языка – выражений на язык представления знаний. В качестве языка представления знаний текста на естественном языке могут выступать, например, язык предикатов первого порядка, семантические сети, фреймы, а также онтологий и тезаурусы.

В 60-е и 70-е годы основным подходом к представлению семантики языка был компонентный подход, в рамках которого

значение каждого слова естественного языка должно было быть представлено в виде комбинации семантических универсалий. К середине 80-х годов стало ясно, что общепризнанный набор таких универсалий так и не удалось составить. Альтернативой компонентному подходу в семантике стала реляционная семантика. При этом подходе значения слов языка описываются заданием связей со значениями других слов, а вся понятийная система языка представляется как семантическая сеть [20].

4.2.1. Обзор методов и программных подходов семантического анализа

На данное время программное обеспечение не может заменить высококачественный анализ, который может продумать человек. Однако программы, которые в настоящее время разрабатываются, позволяют сократить время, затрачиваемое на изучение больших баз данных. В связи с этим рассматриваются работы следующих программ для решения задач семантического анализа текста. Программное обеспечение, предлагаемое различными производителями, такие, как *«ООО Семантик», «Томита-парсер (Яндекс),* Семантический аналитик «*JHON», «SummarizeBot API», «TextAnalyst 2.0», «Galaktika-ZOOM», «NLP ISA»* «Наташа» и др. используется в различных предметных областях и для разных языков [21 - 28].

В научных трудах [29-31] описаны базовые идеи информационного поиска. Представлены различные варианты нахождения статистик текста, которые включают в себя подсчет количества вхождений слов в документы и частоту соседства слов, и новые модельные архитектуры для вычисления непрерывных векторных представлений слов из очень больших наборов данных. Было изучено качество векторных представлений слов, полученных различными моделями, по набору синтаксических и семантических языковых задач. В [32] показано применение языковых моделей нейронной сети к задаче расчета семантического подобия для русского языка. Описаны используемые инструменты и корпуса, достигнутые результаты.

Выше представленные программные продукты разработаны для многоресурсных языков как английский, испанский, русский и др. К сожалению, для казахского языка на данный момент в открытом доступе нет программной реализации. Это связано с тем, что казахский язык отличается своими смысловыми и лингвистическими свойствами от других, а также не имеет большие лингвистические ресурсы для проведения прикладных исследований.

4.2.2. Алгоритм семантического анализа текста на казахском языке

Во время цифровых технологий с учетом постоянного роста объема цифровых данных важную роль играет повышение качества информационного поиска за счет использования новых семантических подходов и методов.

Для работы с большими данными разрабатываются разные алгоритмы и методы для машинного решения этой задачи, так как проводить анализ вручную не позволяют объемы данных. Любой естественный язык по-своему сложен, уникален и многогранен, поэтому извлечение данных из документов и текстовых ресурсов представляет собой большую и трудоемкую работу, которая требует предварительной обработки.

На основе проделанных исследований из разработанных моделей, применяемых наиболее для семантического анализа текстовых ресурсов является подход, основанный на машинном обучении. Ниже будет представлен разработанный алгоритм семантического анализа текста на казахском языке и реализация на основе данного подхода. При разработке алгоритма, для сопоставления определенной информации определенному атрибуту, мы остановили свой выбор на нейронной сети (НС) со скрытым слоем (100). Обучение нейронной сети состоит из следующих частей:

– Предобработка текстов. Предобработка текстов состоит из трех этапов: токенизации, удаления стоп-слов, нормализации слов.

– Построение вектора признаков. Вектор признаков – это признак, интересующей нас характеристики. Для одного де-

скриптора признаки брались следующим образом: бралось окно в два слова после, пять до в тексте статьи на месте вхождения элемента. При том, для каждого дескриптора формируется словарь, который отвечает за наличие указанного слова в словаре. Все признаки каждого дескриптора собираются в один и строится вектор признаков.

– Обучение нейронной сети. Сеть обучается путем предъявления каждого входного набора данных и последующего распространения ошибки.

На втором этапе происходило обучение нейронной сети. Для предобработки текста было использованы разработанные модули обработки естественного языка, описанные в разделе 1. После применения данных модулей мы извлекли признаки нашего дескриптора. Затем с помощью извлеченных данных был построен вектор признаков. Построенный вектор признаков сопоставлялся с определенными ключевыми словами, определенный модифицированным методом TF-IDF для казахского языка.

4.2.3. Программное решение и реализация алгоритма

Это одна из самых сложных и востребованных задач, стоящих перед искусственным интеллектом, является NLP (Natural Language Processing). Для решения и реализации задач NLP в данное время существуют несколько программных комплексов и библиотек, которые включает в себя задачи распознавания речи, формирования языка и получения информации и др.

В настоящее время Python является одной из наиболее перспективных программ для решения задач NLP. Библиотеки, написанные на Python, предназначены для решения задач NLP и позволяют моделировать различные языки и функции обработки.

Так же существует много типов библиотек, рассмотрим наиболее известные и применимые для задач обработки текста:

Spacy, NLTK, CoreNLP, StanfordNER и др. Далее в таблице 4.5 представлены сравнение фукциональных возможностей для решения задачи NLP.

Таблица 4.5

Сравнение возможностей библиотек, направленных на решение проблем NLP

Функция	Spacy	NLTK	CoreNLP
Язык программирования	*Python*	*Python*	*Java/Python*
Модели нейронных сетей	+	-	+
Вектор интегрированных слов	+	-	-
Мультиязычная модель	+	+	+
Токенизация	+	+	+
POS-тэгирование	+	+	+
Сегмантация	+	+	+
Парсинг	+	-	+
Выделение именных объектов	+	+	+
Связь между объектами	-	-	-

Изучив технические возможности для реализации алгоритма семантического анализа и обучения нейронной сети авторами будут применены библиотеки Spacy и StanfordNER. Библиотеки StanfordNER и Spacy позволяют моделировать нашу собственную модель. Это также позволяет сделать необходимые конфигурации, в зависимости от специфики рассматриваемого (казахского) языка.

Необходимо определить настройки StanfordCoreNLP [33]: token- токенизировать; ssplit – распределение предложений; pos – определение речи; лемма – найти оригинальную форму каждого слова; ner – выделение именованных объектов; regexner – работа с регулярными выражениями; parse – семантический анализ каждого слова; depparse – определение синтаксиса между словами и предложениями.

Далее на рисунке 4.5 представлен разработанный алгоритм реализации семантического анализа с учетом ключевых слов и описаны работы модулей.

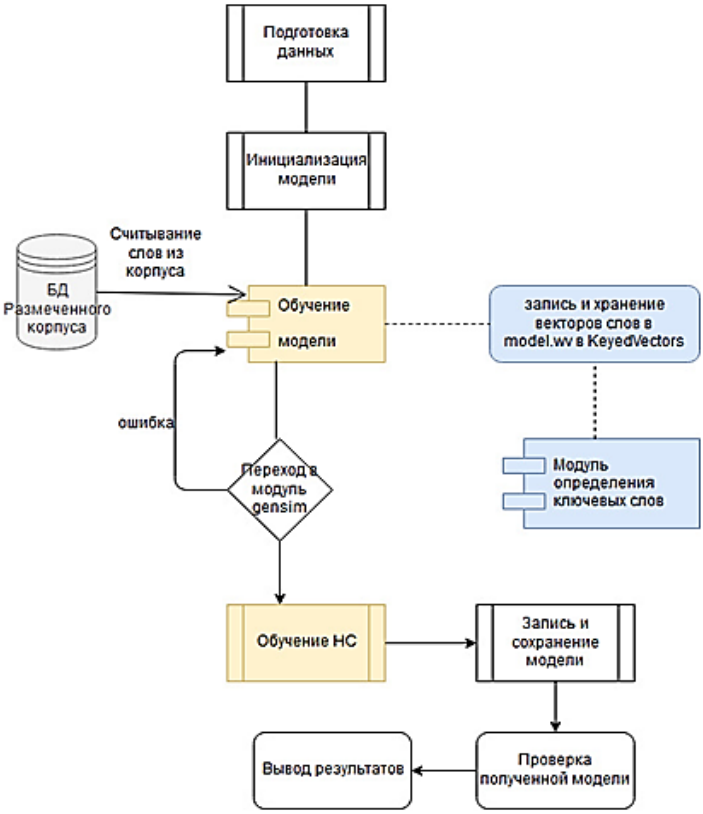

Рис. 4.5. Алгоритм реализации семантического анализа с учетом ключевых слов

На вход подаются текстовые данные. Для обучения модели задаем следующие параметры: Размерность векторов признаков составляет 100; Максимальное расстояние между текущим и предсказанным словом в предложении составляет 5; Минимальный уровень обучения 1; Пороговая частота среза 4 слов.

>>> model = WordVec(sentences, size=100, window=5, min_count=5, workers=4)

Запись инициализированной модели

>>> model.save(fname)

>>> model = WordVec.load(fname) #

Теперь можно проводить обучение с полученной моделью. Для обучения модели был подготовлен одноязычный казахский корпус, который находится в БД SQL. Подробное описание работы по сбору и подготовки одноязычного корпуса показаны в разделах 1-2. При обработке текста моделью выявляются вектора слов, которые хранятся в модуле model.wv в KeyedVectors. Полученные вектора слов так же сопоставляются с ключевыми словами (словосочетаниями) из корпуса текстов с целью дальнейшего использования в качестве возможных значений семантических атрибутов сущностей. Как только модель закончит обучение, вы можете перейти к gensim.models.KeyedVectors в wv:

```
>>> word_vectors = model.wv
>>> del model
```

Модуль gensim.models.phrases автоматически определяет длинную цепочку слов. Этот модуль позволяет нам определять словосочетания путем обучения.

```
>>> bigram_transformer = gensim.models.Phrases(sentences)
>>> model = Word2Vec(bigram_transformer[sentences], size=100, ...)
```

class gensim.models.wordvec.Corpus(*dirname*)

class gensim.models.wordvec.LineSentence(*source,max_sentence_length =10000, limit=None*)

После завершения модуля gensim можно потом приступить к обучению нейронной сети.

```
sentences = LineSentence('myfile.txt')
from gensim.models import Word2Vec # define training data
sentences = [['ұл', 'балалар', 'қыздарға', 'қарағанда', 'мықты', 'болады'],
['Ал', 'қыз', 'балалар', 'ұлдарға', 'қарағанда','нәзік'], ['Қыз', 'әлемнің', 'көркі'],
['Гүл', 'жердің', 'көркі'],
['Қазақстан', 'республикасы', 'тәуелсіз', 'мемлекет']]…
```

В результате полученной обученной модели необходимо проверить полученные данные. Так же можно создать графическую интерпретацию результатов (рис.4.6).

Для обучения модели применили программный пакет NER Stanford. Далее представлен листинг работы с библиотекой NER Stanford и программная реализация

>>> trainFile = train/dummy-kazakh-corpus.tsv serializeTo = dummy-ner-kazakh-french.ser.gz map = word=0,answer=1

useClassFeature=true useWord=true useNGrams=true noMidNGrams=true maxNGramLeng=6 usePrev=true useNext=true

useSequences=true usePrevSequences=true maxLeft=1 useTypeSeqs=true useTypeSeqs2=true useTypeySequences=true wordShape=chris2useLC useDisjunctive= true

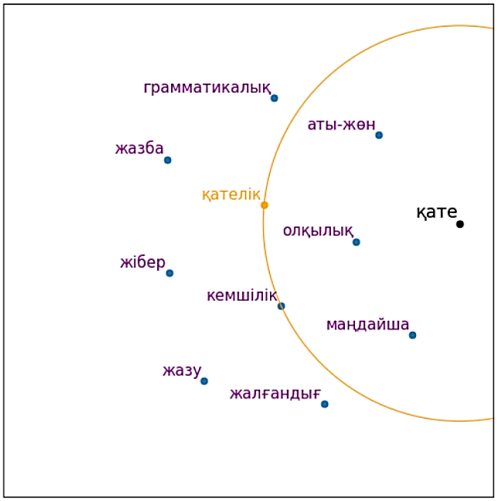

Рис. 4.6. Графическое представление практических результатов векторного пространства

>>> cd stanford-ner-tagger/

java -cp «stanford-ner.jar:lib/*» -mx4g edu.stanford.nlp.ie.crf.CRFClassifier -prop train/prop.txt

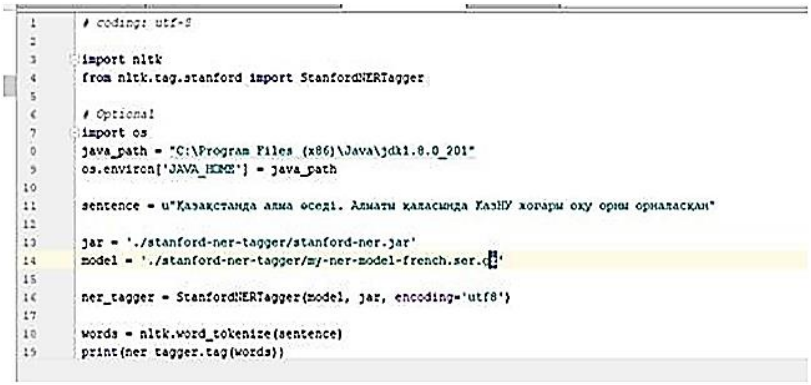

Рис. 4.7. Пример входных данных текста для программы NER

Задача была успешно решена посредством применения морфологического парсера для разметки в текстах частей речи с последующим применением метода машинного обучения семантически связанных ключевых слов (словосочетаний). К набору этих словосочетаний с целью отнесения конкретного словосочетания к определенному атрибуту описываемой в тексте сущности применяется обученная нейронная сеть со скрытым слоем. Таким образом, по набору семантически связанных пар слов, строится онтология для конкретного документа, формирующаяся при работе нейронной сети.

Заключение

По итогам научно-исследовательской работы были получены следующие результаты:

– исследованы методы и современные подходы по извлечению ключевых слов и семантическому анализу текстов;

– разработан модифицированный подход по извлечению ключевых слов и словосочетаний, который будет применен для решения задачи реферирования текстов на казахском языке;

– разработан алгоритм семантического анализа текста на основе машинного обучения.

– разработанные подходы и алгоритмы были применены для обработки текстов на казахском языке.

Благодарность

Исследование выполнено при поддержке Министерства образования и науки Республики Казахстан в рамках научного проекта AP 05132950 «Разработка информационно-аналитической поисковой системы данных на казахском языке».

Литература

1. Шереметьева С.О., Осминин П.Г. Методы и модели автоматического извлечения ключевых слов // Вестник Южно-Уральского государственного университета. – 2015. – № 1. – Т. 12. – С. 76-81.
2. Effective Approaches for Extraction of Keywords // http://www.ijcsi.org/papers/7-6-144-148.pdf: 25.07.2019.
3. Keyword extraction a review of methods and approaches // http://langnet.uniri.hr/papers/beliga/Beliga_KeywordExtraction_a_review_of_methods_and_approaches.pdf: 05.07.2019.
4. Witten I.H., Paynter G.W., Frank E., Gutwin C., NevillManning C.G. Kea: Practical Automatic Keyphrase Extraction // Proceedings of the 4th ACM Conference of the Digital Libraries. – Berkeley, CA. – USA, 1999. – P. 254-255.
5. Turney P.D. Learning to Extract Keyphrases from Text // Technical Report, National Research Council of Canada. – Institute for Information Technology, 1999. – P. 325-328.
6. Medelyan O., Witten I.H. Thesaurus Based Automatic Keyphrase Indexing // Proceedings of the 6th ACM/IEEE-CS JCDL, 2006. – NY, USA, 2006. – P. 296-297.
7. Feldman R., Sanger J. The Text Mining Handbook - Advanced Approaches in Analyzing Unstructured Data. – NY: Cambridge University Press, 2007. – P. 258-305.
8. Zahang C., Wang H., Liu Y., Wu D., Liao Y., Wang B. Automatic Keyword Extraction from Documents Using Conditional Random Fields // Journal of CIS. – 2008. – №4:3. – P. 1169-1180.
9. Chen P., Lin S. Automatic keyword prediction using Google similarity distance // Expert System Application. – 2010. – Vol. 37, issue 3. – P. 1928-1938.
10. Kim S.N., Baldwin T., Kan M-Y. An unsupervised approach to domain-specific term extraction // Proceedings of the Australasian Language Technology Association Workshop. – 2009. – P. 94-98.

11. Basili R. A contrastive approach to term extraction // Proceedings of the 4th Terminological and Artificial Intelligence Conference. – 2001. – P. 154-166.
12. Lopes L., Fernandes P., Vieira R. Estimating term domain relevance through term frequency, disjoint corpora frequency-tf-dcf // Knowledge-Based Systems. – 2016. – Vol. 97. – P. 156-187.
13. Sebastiani F. Machine learning in automated text categorisation // Journal ACM Computing Surveys. – 2002. – №34(1). – P. 1-47.
14. Jones K.S. Informaion retrieval and artificial inteligence // Artificial Intelligence. – 1999. – №114(1-2). – P. 257-281.
15. Berry M.W., Survey of Text Mining II: Clustering, Classification, and Retrieval. – Springer Science & Business Media, 2007. – P. 138-143.
16. Hotho A., Nürnberger A., Paaß G. A Brief Survey of Text Mining // LDV Forum - GLDV Journal for Computational Linguistics and Language Technology. – 2005. – №20(1). – P. 19-62.
17. Sonawane S.S., Kulkarni P.A. Graph based Representation and Analysis of Text Document: A Survey of Techniques // International Journal of Computer Applications. – 2014. – Vol. 96, issue 19. – P. 1-8.
18. Mihalcea R., Radev D. Graph-based Natural Language Processing and Information Retrieval / 1 edition. Cambridge University Press, 2011. – 202 p.
19. Поспелов Д.А. Десять горячих точек в исследованиях по искусственному интеллекту // Интеллектуальные системы (МГУ). – 1996. – Т.1. – No 1-4. – С. 47–56.
20. Алыпанский Г.А., Браславский П.И., Титов П.В. Формирование информационных запросов к машинам поиска интернета на основе тезауруса: семантико-ориентированный подход // Труды VIII Между-нар. конф. по электронным публикациям «EL-Pub2003». – Новосибирск, Академгородок, 2003. – С. 269-270.
21. Семантик // http://semantick.ru/: 14.07.2019.
22. Томита-парсер // http://api.yandex.ru/tomita/: 14.07.2019.
23. В предгорьях семантики // http://dworq.com/: 29.05.2020.
24. AI Data Analysis Technologies for Business // https://www.summarizebot.com/summarization_business.html: 27.05.2019.
25. TextAnalyst ver. 2.0 – Программа для персонального анализа текстов // http://offext.ru/library/data/datakeeping/51.aspx: 19.04.2019.
26. Galaktika-Zoom - аналитическая система для солидных клиентов // https://www.itweek.ru/themes/detail.php?ID=52215: 16.06.2019.
27. Технологии автоматического анализа текстов // http://nlp.isa.ru/: 26.04.2019.
28. GitHub natasha // https://github.com/natasha: 26.04.2019.
29. Manning Ch.D., Raghavan P., Schütze H. Introduction to Information Retrieval. – Cambridge University Press, NY, USA, 2008. – 210 p.
30. Efficient Estimation of Word Representations in Vector Space // https://arxiv.org/pdf/1301.3781.pdf: 10.07.2019.
31. Word2vec Parameter Learning Explained // https://arxiv.org/pdf/1411.2738.pdf: 10.07.2019.

32. Texts in, Meaning out: neural language Models in semantic similarity tasks for russian // https://arxiv.org/ftp/arxiv/papers/1504/1504.08183.pdf: 20.04.2018.
33. The Stanford Natural Language Processing Group // http://nlp.stanford.edu/software/CRF-NER.html: 19.08.2019.
34. Scopus Database // www.scopus.com: 07.08.2019.
35. Thomson Reuters // http://thomsonreuters.com/en.html: 10.08.2019.
36. Etzold J., Brousseau A., Grimm P., Steiner T. Context-aware Querying for Multimodal Search Engines // Proceedings of the 18th international conference on Advances in Multimedia Modeling. – Klagenfurt, Austria, 2012. – P. 728-739.
37. Aixin S., Chii-Hian L. Towards context-aware search with right click // Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval, SIGIR '14. – Gold Coast, Queensland, Australia, 2014. – P. 847-850.

Рахимова Д.Р.
Институт информационных и вычислительных технологий, Алматы, Казахстан e-mail: di.diva@mail.ru

Турганбаева А.О.
Институт информационных и вычислительных технологий, Алматы, Казахстан e-mail: turganbayeva16@gmail.com

Жуманов Ж.М.
Институт информационных и вычислительных технологий, Алматы, Казахстан e-mail: z.zhake@gmail.com

Глава 5

МОДЕЛИ И МЕТОДЫ СЕНТИМЕНТ АНАЛИЗА ТЕКСТОВ НА КАЗАХСКОМ ЯЗЫКЕ

Аннотация. *В современном мире социальные медиа стали частью нашей жизни, данные формируемые в таких системах требует автоматической обработки и анализа. Сентимент анализ является одной из актуальных и интересных задач, которая применяется для анализа данных. Данная работа посвящена изучению и решению задачи сентимент анализа текстов на казахском языке. Для решения этой задачи были изучены и проанализированы исследования по сентимент анализу других естественных языков, создана семантическая база лексических единиц с тональными оценками на казахском языке, построены модели и методы сентимент анализа текстов, реализован алгоритм сентимент анализа текстов на казахском языке.*

5.1. Введение

Сегодня, благодаря широкому распространению интернета, появилась возможность найти необходимую информацию, рекомендации или отзывы людей. Потому что, в настоящее время в Интернете люди открыто высказывают и пишут свои мнения.

С каждым днем в сети появляется новая информация объемом несколько терабайт. Большинство из них являются блогами, твитами, статьями и различной текстовой, аудио и видеоинформацией, отражающей мнения о различных продуктах, товарах, компаниях, фильмах и т.д. Объем информации настолько велик, что их ручная обработка невозможна. Поэтому проблемы создания формальных моделей и программных средств (инструментов) автоматизации и анализа весьма актуальны. Активное развитие современных социальных сетей, блог-платформ и форумов вызывает большой интерес научного сообщества и различных организаций профессиональных IT-специалистов к задачам автоматической обработки и анализа мнений пользователей Интер-

нета. В Казахстане проблемами обработки естественного языка занимаются такие ученые, как А.А. Шарипбай, У.А. Тукеев, Г.Т. Бекманова, Б.Ш. Разахова, Д.Р. Рахимова, О.Ж. Мамырбаев, Ж.А. Есенбаев, М.Х. Карабалаева, А.С. Муканова, А.К. Бурибаева, А.О. Макажанов, Ж.М. Жуманов, Ж.М. Кожирбаев и др.

Сентимент анализ является одним из новых направлений в области обработки естественного языка. При решении этой задачи используются технологии, модели, методы и алгоритмы обработки естественного языка. Исследованиями анализа мнений пользователей и текстовых документов занимаются такие ученые, как Б. Лью, П.Терни, Дж. Вейбе, А. Гельбух, Т. Уилсон, Б. Панг, Т. Насукава, К. Дейв, Е. Камбриа, К. Годдард, Н.В. Лукашевич, И.И. Четверкин, А.Н. Соловьева, П.Ю. Полякова и другие. Кроме того, построены собственные системы крупных корпораций, таких, как Microsoft (Microsoft Azure Text Analytics API, Microsoft Azure Emotion API), Google (Google cloud), Amazon, eBay, SAP, SAS, yandex и др.

Задачи сентимент анализа в основном реализованы для английского, итальянского, русского, китайского и других языков, а исследования, связанные с сентимент анализом текстов на казахском языке, начали проводиться недавно. Таким образом, очень мало готовых инструментов, математических моделей, лингвистических ресурсов для сентимент анализа казахского языка.

В рамках начно-исследовательской работы впервые была создана семантическая база лексических единиц казахского языка с тональными оценками, разработаны модели и методы сентимент анализа, разработан и реализован алгоритм сентимент анализа текстов на казахском языке с использованием гибридного метода на основе формальных правил и словаря [1].

5.2. Основная часть

5.2.1. Проблемы сентимент анализа текстов

Постоянное увеличение объема информации на естественном языке в последнее время в Интернет и социальных сетях

привело к бурному развитию исследований в области компьютерной лингвистики. Компьютерная лингвистика – направление прикладной лингвистики, ориентированное на использование математических методов, компьютерных технологий и программ для организации и обработки данных для моделирования естественного языка [2]. Обработка естественного языка (Natural Language Processing) общее направление компьютерной лингвистики и искусственного интеллекта. К задачам обработки естественного языка относятся: графематический анализ, морфологический синтез и анализ, синтаксический синтез и анализ, семантический анализ, сентимент анализ, машинный перевод, классификация текстов, извлечение информации и др. Несмотря на долгую историю в области лингвистики и обработки естественных языков, исследования, связанные с мнением и настроением людей, стали проводиться только с начала 2000-го года.

Понятие сентимент анализа имеет различные названия, интерпретации, задачи, таких, как например, сентимент анализ, интеллектуальный анализ мнений, поиск мнений, поиск субъективности, анализ настроения и т.д. [3, 4].

Поэтому в этой работе было принято понятие сентимент анализа и в связи с ним были уточнены следующие понятия [1].

Сентимент – эмоциональное отношение автора, выразившего мнение относительно некоторых объектов, (продукт, организация, личность и т.п.), событий (выборы, восстание, война и т.п.), явления (затмение, наводнение и т.п.), процесса (образование, обслуживание и т. п.) или его свойств описанных в тексте.

Мнение – понятие о чём-либо, убеждение, суждение, заключение, вывод, точка зрения или заявление на тему, в которой невозможно достичь полной объективности, основанное на интерпретации фактов и эмоционального отношения к ним.

Сентимент анализ текста – класс методов контент-анализа в компьютерной лингвистике, предназначенный для автоматизированного выявления в текстах эмоционально окрашенной лексики и эмоциональной оценки авторов (мнений) по отношению к объектам, явлениям, событиям, процесссам или их свойствам.

Тональность мнения – признак f, который указывает, что мнение будет положительным, нейтральным или отрицательным. Он также называется полярностью мнения, направлением сентимента или семантическим направлением [5].

Многие исследователи занимаются сентимент анализом, соответственно существует множество различных методов и алгоритмов, которые используются в исследованиях. С одной стороны, в прикладном исследовании сентимент анализа могут применяться методы машинного обучения, методы, основанные на лексиконе или лингвистические методы [4]. С другой стороны, классификация методов сентимента анализа может зависеть от уровня их классификации, например, уровня документа, предложения или аспекта [3]. Сентимент анализ на уровне документа классифицирует полный документ, как положительный или отрицательный. В этом случае считается, что в документе описывается один объект [4, 7]. Сентимент анализ на уровне предложения разделяет каждое предложение в документе, как субъективное или объективное, и классифицирует субъективные мнения на положительные или отрицательные. А на уровне объекта и аспекта определяется отношение объекта к конкретному аспекту, потому что, пользователь может оставить в одном отзыве различные мнения по нескольким аспектам одного объекта. В работах [8-13] отражены результаты исследований по извлечению аспекта.

Методы анализа сентимента можно использовать для различных типов данных, таких, как новости, обзоры, блоги или сообщения в социальных сетях. Каждый вид данных имеет свои особенности, которые необходимо учитывать при сборе, подготовке, предварительной обработке данных и описании объектов.

В настоящее время сентимент анализ находит применение во многих сферах, таких, как отслеживание и анализ отзывов о продукте или компании, определение сторонников или противников политической партии или общественного движения, в различных областях, прогнозирование финансовых доходов. Данные, полученные из социальных сетей и микроблогов (Facebook, Twitter), представляют большой интерес для исследований и приложений, в связи с возможностью публикации в реальном времени отзывов и настроений людей по любым вопросам и доступностью информации в большом количестве.

Решения этих задач так же интересны для некоммерческих и коммерческих организаций, поэтому разрабатываются различные приложения сентимент анализа текстов.

В последние годы наблюдается влияние постов (сообщений) в социальных сетях, сети Интернет на изменение бизнеса, изменения настроения населения относительно политической и социальной системы.

Системы сентимент анализа широко используются в таких областях, как потребительские продукты, услуги, здоровье, финансовые услуги, социальные, политические события. Крупные корпорации, такие, как Microsoft, Google, Hewlett-Packard, SAP, SAS, Yandex разработали собственные системы для анализа мнений. В мире существуют множество решений, приложений для сентимент анализа текстов, но многие решения в основном работают с текстами на английском, итальянском, немецком языках. Например, Google cloud, Microsoft Azure Text Analytics API, Microsoft Azure Emotion API, Social Mention; Sentiment140, «SentiStrength», Semantria, SentiFinder и другие.

Лексические ресурсы, такие, как словари, тезаурусы имеют большую ценность для решения разных задач. Существуют разные лексические ресурсы для определения тональностей текстов, в основном для английского, итальянского, русского. Например, есть такие открытые ресурсы, как WordNet-Affect [14], SentiWordNet [15], SenticNet [16, 17], MPQA Opinion Corpus [18] для английского языка, для русского языка РуСентиЛекс [19]. В работе [1] сделан обзор существующих словарей и тезаурусов. Для казахского языка создан семантический словарь с тональными оценками [20], краткое описание приведено в подразделе 2.8.

Для задач сентимент анализа используются общеизвестные методы, такие, как методы машинного обучения и лингвистические методы [21]. Методы машинного обучения (МО) используют известные алгоритмы МО и лингвистические особенности [4, 22, 23]. Метод, основанный на лексиконе, использует оценочные словари и предопределенный набор терминов сентимента [7, 24, 25]. Этот метод, в свою очередь, делится на методы основанные на словарях и на корпусе. Эти методы используют статистические или семантические методы для определения тональности текста. Также применяются гибридные методы, где применяются несколько методов [26, 27].

Задача сентимент анализа текстов на казахском языке мало изучена. Кроме исследований авторов, имеются работы по сентимент анализу для пары языков (казахскому и русскому) [28, 29, 30].

Решение таких задач, как разработка системы разметки, формализация грамматических правил для компьютерной обработки любого естественного языка, разработка моделей, методов, алгоритмов синтеза и анализа текстов, программная реализация разработанных алгоритмов, создание базы знаний предметной области, создание корпусов являются важными задачами для компьютерной обработки естественных языков.

5.2.2. Признаки, влияющие на определение тональности текстов на казахском языке

Здесь признак обозначает тональный характер или свойство слова или словосочетания. При сентимент анализе текста в качестве признаков используются части речи (существительное, прилагательное, глагол, наречие, междометия), слова отрицания [31]. Например, в казахском языке тональность тексту придают следующие части речи: существительные (жауыздық, соғыс), глагол (тұтқындау, қуанды, ашуланды), прилагательное (әдемі/көріксіз, жақсы/жаман), наречие (нағыз, ең, өте), междометия (алақай!, бәрекелді!, әтеген-ай). В казахском языке слова «емес», «жоқ» являются словами отрицаниями. Как показывают исследования, существительное является аспектом (объектом) обсуждения, а прилагательные в основном определяют семантическую ориентацию (полярность) текста. Тональность оценочного слова может зависеть от контекста и предметной области.

5.2.3. Унифицированная система разметки грамматических правил казахского языка

Система разметок позволит унифицировать разметки, облегчить их понимание и использовать общее программное обеспечение, а также проводить различные исследования.

Эти исследования проводились в рамках проекта «AP05132249 Разработка электронных тезаурусов тюркских языков для создания систем многоязычного поиска и извлечения знаний», некоторые результаты которых были применены в создании правил для сентимент анализа казахских текстов.

Цель создания унифицированной системы разметки заключается в том, чтобы повысить ее практическое применение для других ученых, которые занимаются компьютерной обработкой казахского или других тюркских языков [29].

5.2.4. Моделирование и анализ морфологических правил казахского языка

Морфологический и синтаксический анализ являются предварительными этапами сентимент анализа. Казахский язык относится к тюркской группе языков и характеризуется большим числом словоформ для каждого слова, образованных путем добавления к его концу суффиксов и окончаний. При этом каждый аффикс связан с наборами семантических признаков и порядок добавления аффиксов строго определен. Например, для имен существительных к основе слова в начале добавляется суффикс и далее окончание множественного числа, затем притяжательное окончание, далее следует падежное окончание и последним окончание спряжения [33]. С учетом таких особенностей, были формализованы морфологические правила казахского языка [34-37] и построены онтологические модели в среде Protégé.

На этапе морфологического анализа слова обрабатываются по отдельности, определяются их основы и изменяющие части, такие, как окончания. Морфологический анализ позволяет определить нормальную форму данного слова и набор параметров, присущих этой словоформе.

5.2.5. Формализация структур простого предложения казахского языка

Для решения задачи сентимент анализа необходимо провести синтаксический анализ после морфологического анализа. Для этого формализованы синтаксические правила простых предложений казахского языка и построены их деревья составляющих. Формализация синтаксических правил предложений осуществля-

лась с помощью контекстно-свободной (КС) грамматики Хомского. При синтаксическом анализе определяются составляющие предложения.

Грамматика, в которой всем правилам вывода вида $A \rightarrow \alpha$ накладывается ограничение $A \in N_s$, $\alpha \in (T_s \cup N_s)^*$ называется контекстно-свободной грамматикой (КСГ) – context-free grammar [38].

КСГ G определяется следующими параметрами:

$$G = < N_s, S, T_s, R >,$$

здесь N_s – множество нетерминальных символов; $S \in N_s$ – начальный нетерминальный символ; T_s – множество терминальных символов; R – множество правил вывода вида $A \rightarrow \alpha$, $A \in N_s$ – нетерминальный символ, $\alpha \in (N_s \cup T_s)^*$ – строка символов в неограниченном множестве строк $(N_s \cup T_s)^*$.

Структура простых предложений описана с использованием грамматики составляющих [39]:

S → NP VP|Imit NP VP| Intrj VP| Intrj NP,
NP
→ N |N N |Pron |Num |$AdjP\ N$|N Pron | N Num | N Adv | N Adj | N Conj |
Pron N | Num N | Adv N | Adj N | Part N |
Pron Num | Num Pron | Pron Adv | Adv Pron | Pron Adj | Adj Pron |
Num Num | Num Adj | Adj Num | Adv Num |
Part Pron | Part Num | Part Adj |
Pron N Num Adv | Adj Adj Pron Pron,
VP →
VP | V | V V| NP V | N V | Adv V| Adj V | Pron V | Imit V | Part V | Num V,
AdjP → Adj| Adj Adj |Adv Adj|Adj Conj Adj.

Кроме того, в предложениях с тональными оттенками казахского языка, кроме основных NP (именая фраза) и VP (глагольная фраза) составляющих встречаются, также прилагательная фраза (AdjP), междометия (Intrj). При анализе структуры предложения они также учитываются.

Формализацию с помощью КСГ можно рассмотреть на следующем примере: «Апасынан жақсы хабар келді»:

$N_s = \{S, N, V, NP, VP, Adj\}$;
$T_s = \{А, а, п, с, ы, н, ж, қ, х, б, р, к, е, л, д, і\}$;
$R = \{S \rightarrow NP\ VP, NP \rightarrow N\ |Adj\ N, VP \rightarrow NP\ VP\ |N\ V\}$.

Пример дерева составляющих рассмотренного предложения представлена на рис. 5.1.

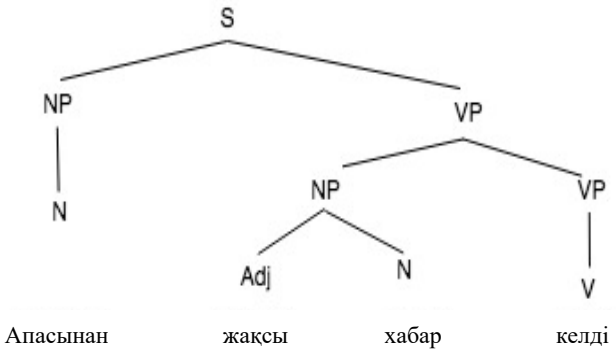

Рис. 5.1. Дерево составляющих S(NP(N),VP(NP(Adj, N),V))

5.2.6. Модель сентимент анализа текстов на казахском языке

Текст – это последовательность лексических единиц, которыми являются слово, устойчивое словосочетание или другая единица языка, способная обозначать предметы, явления, их признаки и т.п. Тональность текста – это эмоциональная оценка автора по отношению к какому-нибудь событию или объекту, представленном в тексте, которая определяется тональностями его лексических единиц.

Модель сентимент анализа текста определяет названия тональных единиц (слов или словосочетаний), тип и их значения (таблица 5.1).

Таблица 5.1

Названия тональных единиц в модели сентимент анализа текста

№	Названия тональных единиц	Тип тональных единиц	Значения тональных единиц
1	очень негативный	целое число	-2
2	Негативный	целое число	-1
3	Нейтральный	целое число	0
4	Позитивный	целое число	1
5	очень позитивный	целое число	2

При построении модели сентимент анализа текстов на казахском языке используются результаты морфологического и синтаксического анализа. Например, с помощью этих словосочетаний можно определить тональность:

1. [N] · [V]
2. [N] · [V] · [Negation]
3. [ADJ] · [N]
4. [ADJ] · [Negation] · [N]
5. [ADJ] · [V]
6. [ADJ] · [V] · [Negation]
7. [ADV] · [ADJ]
8. [ADV] · [N];

здесь, ADJECTIVE – часть речи имя прилагательное, NOUN – имя существительное, Negation – отрицательные слова емес/жоқ, VERB – глагол, ADVERB – наречие.

Кроме того, в тексте могут быть слова междометия, которые могут придавать тональность тексту. Тональность слов определены в базе тональных лексических единиц на казахском языке.

Для проведения сентимент анализа нам необходимо определить лексические единицы сентимент анализа, которыми могут быть слова, фразы и предложения естественного языка. Определяя сентиментальные признаки лексических единиц, мы можем вычислить сентимент всего текста.

Для моделирования сентимент анализа текстов на казахском языке использовалась продукционная модель [40-43]. Для построения продукционной модели введены следующие метаобозначения (таблица 5.2):

Таблица 5.2

Метаобозначения

Обозначение	Назначение
$\alpha, \beta, \gamma, \ldots, \zeta, \xi, \ldots,$	Множество слов языка – Переменные
ω	$\omega = \zeta \cdot \alpha \cdot \beta \cdot \xi$ – лексические единицы (непустое слово или словосочетание)
L	Множество предложений языка
N	Множество имен существительных
Adj	Множество имен прилагательных
$Pron$	Множество местоимений
V_Post	Множество положительных форм глаголов
V_Negt	Множество отрицательных форм глаголов
$AdvIntens$	Множество наречий для усиления
$sent$	Установление сентимента – Предикат
$@$	Слова отрицания емес/жоқ – Константы
$\neg$	Превращения в отрицательную форму – Операция
$\cdot$	Конкатенация – Операция

Ниже приведена модель, основанная на формализованных продукционных правилах для определения тональности лексических единиц (фраз) в тексте на казахском языке:

1) Если лексическая единица сентимент анализа содержит существительное с позитивной тональностью и следующее слово за ним является глаголом (положительной формы) с нейтральной тональностью, то сентимент этой фразы позитивный.

$$\frac{\omega \in L, \omega = \zeta \cdot \alpha \cdot \beta \cdot \xi\ , \alpha \in N, sent(a) = 1, \beta \in V\_Post, sent(\beta) = 0}{sent(\omega) = 1}$$

здесь и далее ζ, ξ – любые цепочки слов, в том числе и пустые. Например, той болды, куанышқа толды.

2) Если лексическая единица сентимент анализа содержит существительное с негативной тональностью и следующее слово за ним является глаголом (положительной формы) с нейтральной тональностью, то сентимент этой фразы негативный.

$$\frac{\omega \in L, \omega = \zeta \cdot \alpha \cdot \beta \cdot \xi\ , \alpha \in N, sent(a) = -1, \beta \in V\_Post, sent(\beta) = 0}{sent(\omega) = -1}$$

например, соғыс болды.

3) Если лексическая единица сентимент анализа содержит глагол (положительной формы) с нейтральной тональностью за существительным с позитивной тональностью, проверятся слово стоящее после глагола, если после глагола встречаются слова отрицания (емес/жок), то сентимент этой фразы негативный.

$$\frac{\omega \in L, \omega = \zeta \cdot \alpha \cdot \beta \cdot \gamma \cdot \xi\ , \alpha \in N, sent(a) = 1, \beta \in V\_Post, sent(\beta) = 0, \gamma = @}{sent(\omega) = -1}$$

например, әділеттілік орнаған жоқ.

4) Если лексическая единица сентимент анализа содержит прилагательное с очень позитивной тональностью и следующее слово за ним является существительным с нейтральной тональностью, то сентимент этой фразы очень позитивный.

$$\frac{\omega \in L, \omega = \zeta \cdot \alpha \cdot \beta \cdot \xi\ , \alpha \in Adj, sent(a) = 2, \beta \in N, sent(\beta) = 0}{sent(\omega) = 2}$$

например, Ардақты ана.

5) Если лексическая единица сентимент анализа содержит прилагательное с очень негативной тональностью и следующее слово за ним является существительным с нейтральной тональностью, то сентимент этой фразы очень негативный.

$$\frac{\omega \in L, \omega = \zeta \cdot \alpha \cdot \beta \cdot \xi\ , \alpha \in A, sent(a) = -2, \beta \in N, sent(\beta) = 0}{sent(\omega) = -2}$$

например, көргенсіз адам.

6) Если лексическая единица сентимент анализа содержит между прилагательным с очень позитивной тональностью и существительным с позитивной тональностью отрицание (емес), то сентимент этой фразы негативный.

$$\frac{\omega \in L, \omega = \zeta \cdot \alpha \cdot \beta \cdot \gamma \cdot \xi\ , \alpha \in Adj, sent(a) = 2, \beta = @, \gamma \in N, sent(\gamma) = 1}{sent(\omega) = -1}$$

например, керемет емес абыройы.

7) Если лексическая единица сентимент анализа содержит прилагательное с позитивной тональностью и следующее слово за ним является глаголом (положительная форма) с нейтральной тональностью, то сентимент этой фразы позитивный.

$$\frac{\omega \in L, \omega = \zeta \cdot \alpha \cdot \beta \cdot \xi\ , \alpha \in Adj, sent(a) = 1, \beta \in V\_Post, sent(\beta) = 0}{sent(\omega) = 1}$$

например, жақсы істейді.

8) Если лексическая единица сентимент анализа перед позитивным прилагательным содержит усилительное наречие и следующее слово за ним является существительным с нейтральной тональностью, то сентимент этой фразы очень позитивный.

$$\frac{\omega \in L, \omega = \zeta \cdot \alpha \cdot \beta \cdot \gamma \cdot \xi\ , \alpha \in AdvIntens, \beta \in Adj, sent(\ \beta) = 1, \gamma = N, sent(\ \gamma) = 0}{sent(\omega) = 2}$$

например, өте әдемі қыз, ең әдемі қала.

Сентимент всего текста определяется как средне-арифметическое величин измерения тональностей лексических единиц (предложений) и правил их сочетаний.

$$\mathrm{sent(L)} = \frac{\sum_{i=1}^{n} \mathrm{sent_i}(\omega)}{n}.$$

5.2.7. Метод сентимент анализа текстов на казахском языке

Для сентимент анализа казахского языка предложен гибридный метод на основе словаря и формальных правил. В качестве словаря используется база лексических единиц с тональными оценками на казахском языке. В качестве правил используются формальные правила, определяющие сентимент текстов на казахском языке с использованием продукционной модели. Каждое правило, которое используется для определения сентимента фрагмента текста, представлено в форме «ЕСЛИ условие, ТО вывод» [1].

5.2.8. Семантическая база оценочных лексических единиц казахского языка

Существующие словари были созданы для других языков, например, английского, русского. А для казахского языка нет доступных лексических ресурсов с тональными оценками. Для решения задачи автоматического определения тональности текста была создана семантическая база лексических единиц казахского языка [20, 31]. База была вручную создана и размечена по тональности по 5-бальной шкале (от -2 до 2). Кроме того, некоторые слова или знаки могут изменить полярность сентимента в зависимости от контекста текста. Для таких случаев применяется знак «+-». Тональность лексических единиц измеряется следующими величинами: -2 – очень негативный; -1 – негативный; 0 – нейтральный; 1 – позитивный; 2 – очень позитивный.

В семантической базе имеется не только список слов и словосочетаний, но и их толкование и синонимы с тональными оценками. База состоит из более 12000 слов и словосочетаний, составленных из слов, относящихся к различным частям речи. Здесь значительная часть прилагательные, которые определяют настроение, тональность текста (рис. 5.2).

Word	meaning/synonym	Pos	SO
арзан	бағасы төмен	Adj	+-
арзанырақ	1. Бағасы төменірек. 2. Ауыс. Оңайырақ, жеңілірек.	Adj	+-
аскаралы	Айдынды, биік.	Adj	1
асыл	Тәрбиелі, текті ақылды адам.	N	2
атаулы	Басты, елеулі, көрнекті.	Adj	1
әсем	Көз жауын алатындай әдемі, көрікті, көркем.	Adj	2
әсемпаз	Кербездікке, сәнқойлыққа әуес	Adj	-1
әуес	Бір нәрсеге үйір, соған құмар, құштар.	Adj	1
әуесқой	1. Әр нәрсеге құмартып қызыққыш. 2. Белгілі бір іс бабында арнайы маман болмаса да соған құмар, бейім.	Adj	1
байланыстырушы	Жалғастырушы, біріктіруші.	Adj	1
байсал	Дел-сал, самарқау, епжар.	Adj	-1
бақандай	Білдей, дырдай, ортан қолдай.	Adj	+-
баяғыдағы	Ерте кездегі, ежелгі.	Adj	+-
бейпарасат	Парасатсыз, ақыл-санасыз.	Adj	-1
бейсауат	Бейсауат.	Adj	-1
бекершілік	1. Қателік, терістік, бос әурешілік. 2. Орынсыздық, оғаштық, жалғандық.	Adj	-1
белсенді	1. Қажырлы, ынталы, жігерлі. 2. Үздіксіз, тынымсыз. 3.Орынсыз белсенділік жасаушы, асыра сілтеуші.	Adj	+-
берекелі	1. Берекесі мол, игілікті. 2. Ауыз бірлігі бар, ынтымақты. 3. Тыңғылықты, тындырымды, ұқыпты. 4. Пайдалы, тиімді.	Adj	1
берекелірек	Берекесі молдау, тиімділеу.	Adj	1

Рис. 5.2. Фрагмент из базы тональных лексических единиц казахского языка

Кроме того, в целях повышения качества анализа текста можно использовать дополнительные обозначения, дающие оттеночную окраску, такие, как эмотиконы, союзные слова, междометия. Пользователи применяют в тексте различные эмотиконы (смайл) для описания эмоций. При сентимент анализе текста возникает необходимость учета значения эмотиконов, так как они тоже влияют на общую оценку. Используются символические и графические формы эмотиконов. В казахском языке междометия также могут сыграть важную роль при сентимент анализе текста, так как выражают различное настроение, чувства человека. Поэтому эти слова также должны быть учтены при анализе. Например, алақай, астапыралла, бәле!, әттеген-ай, мәссаған және т.б. В базе также присутствуют междометия, эмотиконы, размеченные по полярности сентимента.

Заключение

Бурное развитие компьютерных технологий и вычислительных машин позволяет решать сложные задачи в различных сферах жизни. Многие исследователи занимаются исследованиями задач обработки естественных языков. Обработка текстов на естественном языке является одной из актуальных проблем в области информатики, искусственного интеллекта, компьютерной лингвистики.

Результаты, описанные в данной работе, станут полезным ресурсом для ученых, студентов и специалистов, занимающихся компьютерной обработкой казахского языка для решения прикладных задач анализа мнений, классификации текстовых документов, извлечения информации, кластеризации и др. Область применения результатов исследовательской работы очень обширна и может быть использована при оценке, мониторинге, анализе продукции и рекомендательных системах государственных и негосударственных предприятий в области образования, науки, искусства, здравоохранения и других социальных сфер.

Литература

1. Ергеш Б.Ж. Қазақ тіліндегі арнайы мәтіндерді семантикалық талдау моделдері мен әдістері: (PhD) док. ... дис. – Нұр-Сұлтан, 2020. – 103 б.
2. Бекманова Г.Т. Разработка методов звукового распознавания слов на основе их морфологического анализа и синтеза: автореф. ... канд. техн. наук: 05.13.11. – Астана, 2010. – 23 с.
3. Liu B. Sentiment analysis and opinion mining // Synthesis Lectures on Human Language Technologies. – 2012. – Vol. 5(1). – P. 1-167.
4. Pang B., Lee L. Opinion mining and sentiment analysis // Foundations and Trends in Information Retrieval. – 2008. – Vol. 2(1-2). – P. 1-135.
5. Handbook of Natural Language Processing / ed. N. Indurkhya, F.J. Damerau. Ed. – 2nd. – Boca Raton: Chapman and Hall/CRC, 2010. – 676 p.
6. Thelwall M., Buckley K., Paltoglou G. Sentiment in Twitter events // Journal of the American Society for Information Science and Technology, 2011. – Vol. 62(2). – P. 406-418.
7. Turney P. D. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews // In Proceed. of the 40th annual meeting on association for computational linguistics. – Philadelphia, 2002. – P. 417-424.
8. *Popescu A., Etzioni O. Extracting product features and opinions from reviews // In book: Natural language processing and text mining. – London: Springer, 2007. – P. 9-28.*
9. Liu B., Zhang L. A survey of opinion mining and sentiment analysis // *In book: Mining Text Data. – New York:* Springer, 2012. – P. 415-463.
10. Bagheri A., Saraee M., de Jong F. An Unsupervised Aspect Detection Model for Sentiment Analysis of Reviews // *Natural Language Processing and Information Systems.* – Berlin: Springer Berlin Heidelberg, 2013. – P. 140-151.
11. Cilavas G., Korencic D., Snajder J. Aspect-Oriented Opinion Mining from User Reviews in Croatian // *Proceed. BSNLP workshop, ACL 2013.* – Sofia, 2013. – P. 18-23.
12. Zhang L., Liu B. Aspect and Entity Extraction for Opinion Mining // *In book: Data Mining and Knowledge Discovery for Big Data.* – Berlin: Springer Berlin Heidelberg, 2014. – P. 1-40.
13. Poria S., Cambria E., Gelbukh A. Aspect extraction for opinion mining with a deep convolutional neural network // Knowledge-Based Systems. – 2016. – Vol. 108. – P. 42-49.
14. Strapparava C., Valitutti A. Wordnet-affect: an affective extension of wordnet // In Proceed. of the 4th internat. conf. on Language Resources and Evaluation. – Lisbon, 2004. – Vol. 4. – P. 1083-1086.
15. Baccianella S., Esuli A., Sebastiani F. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining // Proceed. of LREC. – Valletta, 2010. – P. 2200-2204.

16. Cambria E. SenticNet 2: A semantic and affective resource for opinion mining and sentiment analysis // Proceed. of AAAI 25th FLAIRS: конф. – Florida, 2012. – P. 202-207.
17. Cambria E. et al. SenticNet 4: A Semantic Resource for Sentiment Analysis based on Conceptual Primitives // Proceed. 26th internat. conf. «Computational Linguistics». – Osaka, 2016. – P. 2666-2677.
18. MPQA Opinion Corpus // http://mpqa.cs.pitt.edu. 16.10.2017.
19. Loukachevitch N., Levchik A. Creating a General Russian Sentiment Lexicon // In Proceed. of Language Resources and Evaluation conf. LREC-2016. – Portorož, 2016. – P. 1171–1176.
20. Ергеш Б.Ж. Сентимент талдауға қажетті лексикалық ресурстар // Қазақстан Республикасы Ұлттық инженерлік академиясының хабаршысы. – 2019. – №3. – С. 55-59.
21. Medhat W., Hassan A., Korashy H. Sentiment analysis algorithms and applications: A survey // Ain Shams Engineering Journal. – 2014. – Vol. 5(4). – P. 1093-1113.
22. Wiebe J., Bruce R.F., O'Hara T.P. Development and use of a gold-standard data set for subjectivity classifications // In Proceed. of the Association for Computational Linguistics (ACL-1999). – Maryland, 1999. – P. 246-253.
23. Wilson T., Wiebe J., Hwa R. Just how mad are you? Finding strong and weak opinion clauses // In Proceed. of nat. conf. on Artificial Intelligence (AAAI-2004). – San Jose, 2004. – P. 761-767.
24. Taboada M., Brooke J., Tofiloski M., Voll K., Stede M. Lexicon-based methods for sentiment analysis // Computational linguistics. – 2011. – Vol. 37(2). – P. 267-307.
25. Пазельская А.Г., Соловьев А.Н. Метод определения эмоций в текстах на русском языке // Компьютерная лингвистика и интеллектуальные технологии: матер. ежегод. междунар. конф. «Диалог». – Бекасово, 2011. – С. 574-586.
26. Wilson T., Wiebe J., Hoffmann P. Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis // Computational linguistics. – 2009. – Vol. 35. – №3. – P. 399-433.
27. Kaity M., Balakrishnan V. An automatic non-English sentiment lexicon builder using unannotated corpus // The Journal of Supercomputing. – 2019. – Vol. 75(4). – P. 2243-2268.
28. Sakenovich N.S., Zharmagambetov A.S. On One Approach of Solving Sentiment Analysis Task for Kazakh and Russian Languages Using Deep Learning // In proceed. of the internat. conf. on Computational Collective Intelligence, ICCCI 2016. – Sithonia: Springer International Publishing, 2016. – Vol. 9876. – P. 537-545.
29. Abdullin Y.B., Ivanov V.V. Deep learning model for bilingual sentiment classification of short texts // https://cyberleninka.ru/article/n/deep-learning-model-for-bilingual-sentiment-classification-of-short-texts. 19.04.2019.
30. Mamykova, Z., Mutanov, G., Sundetova, Z., & Torekul, S. Подходы к разработке информационной системы мониторинга мнений и оценки социального самочувствия // Вестник КазНУ. Серия математика, механи-

ка, информатика. – Алматы, 2019. – No 4 (100). - стр.63-77. doi:10.26577 / JMMCS-2018-4-574.
31. Ергеш Б.Ж. Определение тональности текстов на казахском языке на основе словаря эмоциональной лексики // Матер. 5-я междунар. конф. по компьютерной обработке тюркских языков «TurkLang 2017»: в 2 т. – Казань: Издательство Академии наук Республики Татарстан, 2017. – Т. 1. – С. 62-67.
32. «AP05132249 Разработка электронных тезаурусов тюркских языков для создания систем многоязычного поиска и извлечения знаний» [Текст]: отчет о НИР (промежуточ.) / ЕНУ им. Л.Н. Гумилева; рук. Шарипбай А.А.; исполн. Муканова А.С. и др. – Астана, 2018. – 57 с. – No ГР 0118РК00656. – Инв. №0218РК01298.
33. Қазақ грамматикасы. Фонетика, сөзжасам, морфология, синтаксис / ред. Е. Жанпейісов. – Астана : Астана полиграфия, 2002. – 784 б.
34. Ергеш Б.Ж., Муканова А.С., Шарипбай А.А., Бекманова Г.Т. Формальная модел именных частей речи казахского языка // 14-я национ. конф. по искусственному интеллекту с междунар. участием: тр. конф. – Казань, 2014. – С. 101-107.
35. Шарипбаев А.А., Бекманова Г.Т., Ергеш Б.Ж., Муканова А.С. Онтологическая модел представления морфологических правил казахского языка в виде семантических гиперграфов // Открытые семантические технологии проектирования интеллектуальных систем (OSTIS-2013): матер. 3-й междунар. науч.-техн.й конф. – Минск, 2013. – С. 337-340.
36. Шарипбаев А.А., Бекманова Г.Т., Ергеш Б.Ж., Муканова А.С. Формальная модель морфологических правил казахского языка // Докл. НАН РК. – Алматы: РОО «НАН РК», 2012. – С. 16-22.
37. Муканова, А.С., Ергеш, Б.Ж., Разахова, Б.Ш. Морфологиялық ережелерді онтологиялық моделдеу // Түркі тілдерін компьютерлік өңдеу: 1-ші халық. конф. еңбек. – Астана, 2013. – Б. 188-194.
38. Шәріпбай А.Ә. Тілдер мен автоматтар теориясы: оқулық. – Астана: Л.Н. Гумилев атындағы ЕҰУ, 2013. – 234 б.
39. Sharipbay A., Razakhova B., Mukanova A., Yergesh B, Yelibayeva G. Syntax parsing model of Kazakh simple sentences // Proceed. of the Second internat. conf. on Data Science, E-Learning and Information Systems (DATA '19). – Dubai, 2019. – №54. – P. 1-5.
40. Ергеш Б.Ж., Шарипбай А.А., Бекманова Г.Т. Роль имен прилагательных в определении тональности текста // Тр. междунар. конф. по компьютерной и когнитивной лингвистике TEL-2016. – Казань: Изд-во Казан. ун-та, 2016. – С. 85-89.
41. Yergesh B., Sharipbay A., Bekmanova G., Lipnitskii S. Sentiment analysis of Kazakh phrases based on morphological rules // Journal Of Kyrgyz State Technical University named after I.Razzakov. – Bishkek, 2016. – Vol. 2(38). – P. 39-42.
42. Yergesh B., Bekmanova G., Sharipbay A., Yergesh M. Ontology-based sentiment analysis of kazakh sentences // In book: Lecture Notes in Computer

Science (LNCS). – Cham: Springer International Publishing, 2017. – Vol. 10406. - P. 669-677.
43. Yergesh B., Bekmanova G., Sharipbay A. Sentiment analysis of Kazakh text and their polarity // Web Intelligence– IOS Press. – 2019. – Vol. 17(1). – P. 9-15.

Ергеш Б.Ж.
PhD, Евразийский национальный университет имени Л.Н. Гумилева, Нур-Султан, Казахстан, e-mail: b.yergesh@gmail.com

Шарипбай А.А.
Д.т.н., профессор, Евразийский национальный университет имени Л.Н. Гумилева, Нур-Султан, Казахстан, e-mail:sharalt@mail.ru

Бекманова Г.Т.
к.т.н., PhD, асс. профессор, Евразийский национальный университет имени Л.Н. Гумилева, Нур-Султан, Казахстан, e-mail: gulmira-r@yandex.kz

Глава 6

МЕТОД ИДЕНТИФИКАЦИИ КРИМИНАЛЬНОГО ЗНАЧЕНИЯ ТЕКСТОВ НА КАЗАХСКОМ ЯЗЫКЕ, БАЗИРУЮЩИЙСЯ НА VSM

Аннотация. *Для идентификации криминального содержания текстов казахского языка предлагается подход, основанный на определении семантического подобия входного документа к текстам обученного корпуса, включающего новостные статьи казахских сайтов, которые содержат криминальную информацию. Используемая метрика предлагает вычисление косинусного сходства между текстами корпуса и входным документом на базе Vector Space Model. В работе эмпирически определяется среднее значение коэффициента косинусного сходства, при котором документ идентифицируется как имеющий криминальную окраску. Проведенный эксперимент показал, что F-мера данного подхода идентификации достигает 96 %.*

Введение

Оценка семантического сходства текстов – достаточно емкая и обширная задача, впервые рассмотренная Дж. Солтононом [1] для информационного поиска, сегодня является неотъемлемой частью большинства лингвистических задач, таких, как перефразирование, классификация, создания вопросно-ответных систем, мониторинг социальных и телекоммуникационных сетей и других. В то же время, большинство современных исследований до сих пор сосредоточено на развитии этой области только для английского языка. Несмотря на то, что существует несколько доступных и хорошо работающих приложений для семантического сравнения англоязычных текстов, таких, как WordNet::Similarity или Alchemy API, алгоритмы семантического подобия текстов других языков, в большинстве своем, не завершены.

В то же время, задача определения тематики текста традиционно базируется на подходах классификации или кластериза-

ции. Хорошими и часто используемыми методами классификации текстов являются деревья решений, нейронные сети [2], Random Forest и Support Vector Machine [3], Байесовский метод, определение К-средних [4] и другие подобные методы. Тем не менее все эти методы машинного обучения требуют наличия обученного корпуса, в котором имеется предопределенный набор классов и набор документов, относящихся к этим классам. В случае же, когда необходимо определить узкоспециализированную тематику текста, при отсутствии заранее обученных корпусов с предопределенными классами, задача идентификации тематики существенно усложняется.

В нашем исследовании, для того чтобы определить наличие в текстовых документах казахского языка некоторого криминального смысла мы предлагаем использовать подходы измерения семантического подобия входного текста к текстам обученного корпуса новостных статьей, содержащих криминальную информацию.

6.1. Анализ литературных данных

Поиск семантического подобия текстовой информации становится все более популярным в различных областях научных исследований. Например, в работе [6] методы определения семантического сходства использовались для выявления связи между такими медицинскими объектами, как наркотик и диагноз. Подход основывался на встраивании лекарственного средства в модель рецепта и оценке сходства между ними, через вектора обоих объектов. Данный подход был эмпирически изучен и показывал хорошие результаты в биомедицине [5].

Оценка семантического сходства текстов использовалась также при анализе рынка, в банковском деле и маркетинге. В работе [6] авторы применяли данный подход для определения сходства между различными пресс-релизами банка и оценками их влияния на потенциальных клиентов и финансовый рынок. Это было сделано путем расчета разницы между векторами фиксированной длины пар пресс-релизов. При этом, больший вес присваивался редким словам, а меньший – часто встречающимся.

Методика вычисления семантического сходства между словами текста и лексическим словарем также использовалась в области Sentiment Analysis, в частности, при обработке различных лексических ресурсов. Авторы работы [7] применили данный подход в модели классификации настроений, используя меру семантической близости и встраивания образов.

В работе [8] был использован новый метод определения семантического сходства больших документов – академических статей. Для вычисления степени семантического сходства авторы исследования использовали доменную онтологию, применяемую для вычисления сходства семантических событий. Авторы показали, что использование метода, основанного на онтологиях, имеет преимущества в корреляции, точности и значении F1-score.

В исследовании [9] показан улучшенный способ вычисления коэффициента семантического сходства, основанный на алгоритме random walk.

Особенность данного алгоритма заключается в сравнении распределения каждого текста, полученного при первоначальном случайном проходе по графику из WordNet. Алгоритм позволяет уменьшить значение ошибки по сравнению с обычной векторной моделью.

Проведенный общий анализ позволяет разделить существующие подходы к решению задачи идентификации семантической близости текстовых документов на две большие группы. Первый, традиционный подход базируется на использовании онтологий. Например, онтологический подход использовался методом Резника [10] или расширенным алгоритмом Леска [11]. Однако в подавляющем большинстве случаев такие подходы применялись для идентификации семантического сходства коротких текстовых фрагментов и/или синонимичности слов.

Вторая группа подходов к задаче идентификации семантического сходства текстовых объектов базируется на статистических методах дистрибутивной семантики. Эти методы применяются для измерения семантического подобия слов, подобия документов, а также семантической близости отношений [12]. В то же время, вычисление семантическое подобие крупных текстовых документов по-прежнему основывается только на статистической информации, что явно недостаточно для определения глобальных семантических значений [13].

6.2. Использование VSM в семантическом анализе.

В общем случае Vector Space Model (VSM) представляет семантическую модель представления текстов, в которой каждому документу сопоставлен вектор, отражающий его смысл [14]. Базируясь на идее представления каждого документа в коллекции текстов как точки векторного пространства, модель показывает, что точки, расположенные близко друг к другу в данном пространстве, соответствуют семантически близким документам, а точки, которые расположены далеко друг от друга, соответствуют сильно отличающимся по смыслу документам.

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 применяться для выяснения того, что люди имеют в виду. Говоря другими словами, человеческий интеллект может понимать слова в зависимости от их окружения [15]. Вторая гипотеза, сформулированная Дж. Солтоном для информационного поиска [16], 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.

Если имеется 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 корпуса. Полученная таким образом матрица представляет собой матрицу термин-документ.

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

Например, в результате предварительной лингвистической обработки может быть получен упрощенный текстовый массив специализированных текстов узкой направленности. В упрощен-

ном для понимания виде, массив может включать шесть документов, состоящих из восемнадцати 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*.

Следующим шагом после токенизации и нормализации корпуса текстов является генерация матрицы частот, которая представляет обобщенный вид матрицы *тип токена – документ*, в которой тип токена выступает термином. В данной частотной матрице, показанной в таблице 6.1, элемент соответствует событию появления токена в документе. 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 $f$ij в матрице X – это частота i-того термина wi в j-ом документе dj.

Таблица 6.1

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

Однако простая частота встречаемости – это не лучший способ измерения ассоциации между словами. Используя частотный способ определения ассоциации, не следует учитывать влияние таких слов, как «*емес*», «*әлде*», «*олар*», которые проявляются довольно часто с любыми словами, и при этом не передают 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. Согласно теории информации Клода Шеннона [17] неожиданные события имеют более высокое информационное значение (или содержание информации), чем ожидаемые события. С точки зрения семантической обработки текстов это обозначает, что чем больше частота термина, тем он менее информативен. Гипотеза состоит в том, чтобы неожиданные события, если они разделены двумя векторами, являются более показательными при определении сходства между векторами, чем менее неожиданные события.

Для того чтобы придать больший вес более информативным токенам и уменьшить значение менее информативных токенов, будем использовать взвешивание элементов матрицы с помощью весовой функции PPMI (Positive Pointwise Mutual Information) [18], введенной на базе дистрибутивной гипотезы.

Весовая функция PPMI добавляет больший вес значениям $x$ij, подтверждающим семантическую зависимость между wi и dj и присвоить нулевые значения, если появление wi в документе dj не имеет никакого семантического значения. Таким образом, решается проблема существования стоп-слов в тексте или осуществляется избавление от «информационного шума».

Функция PPMI имеет всегда положительное значение и определяется через меру PMI (Pointwise mutual information).

$$x_{ij} = ppmi_{ij} = \begin{cases} pmi_{\text{ij}}, \text{ если } \text{pmi}_{\text{ij}} > 0 \\ 0, \text{pmi}_{ij} <= 0 \end{cases}. \qquad (6.1)$$

Тогда как мера PMI [18] определяется через вероятностные значения термина и документа в коллекции. Формально PMI можно определить как логарифм отношения оценочной вероят-

ности появления термина в данном документе к произведению оценочных вероятностей появления термина и документа в данной коллекции:

$$pmi_{ij} = \log(\frac{p_{ij}}{p_{i*} p_{*j}}), \quad (6.2)$$

где *p*ij – это оценочная вероятность того, что термин *w*i появится в документе dj, вычисляемая, как частота появления термина в данном документе, нормализованная суммой всех терминов во всех документах коллекции, т.е. размером словаря данной коллекции документов *N*; pi∗ – это оценочная вероятность термина wi, 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 dj, 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 того чтобы решить очевидную проблему смещения веса PMI к более редким событиям будем использовать сглаживания Лапласа при оценке вероятности *p*ij , *p*i∗, и *p*∗j [19].

Сглаживание Лапласа заключается в добавлении к необработанным частотам некоторого положительного числа перед вычислением вероятностей. При этом, каждое *f*ij заменяется на *f*ij +*k*, где $k > 0$. Чем больше константа *k*, тем больше эффект сглаживания. Используя сглаживание Лапласа, заполняя частотную таблицу термин-документ (см. табл. 1), будем считать, что мы видели каждое слово в документе на два раза чаще. Сглаживание Лапласа уменьшает значение pmiij. 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.

В таблице 6.2 показаны значения PPMI вышеприведенного примера, с учетом сглаживания Лапласса *add*-2.

Таблица 6.2

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 [20]. Пусть $x = \langle x1, x2, \ldots, xn \rangle$ и $y = \langle y1, y2, \ldots, yn \rangle$ будут два вектора, из n элементов. Тогда косинус угла Θ между векторами x и y может быть посчитан как скалярное произведение векторов, нормализованное их длинами:

$$\cos(x, y) = \frac{x \bullet y}{\sqrt{x \bullet x}\sqrt{y \bullet y}} = \frac{x}{||x||} \bullet \frac{y}{||y||}. \qquad (6.3)$$

Согласно формуле (3) и исходя из того, что, значения PPMI не могут быть отрицательными, следовательно, значения косинуса между векторами, использующими в качестве координат PPMI, всегда будут лежать в положительном диапазоне [0, 1].

Тогда матрица косинусного сходства PPMI векторов документов рассматриваемого примера (см. табл. 6.1) будет выглядеть так, как показано в таблице 6.3.

Проанализировав полученные результаты можно сделать вывод о том, что из рассмотренного массива документов наиболее

близкими по смыслу являются документы *Doc1* и *Doc2*, так как значение косинуса угла между ними максимально cos (*doc*1, *doc*2) = 0,655787.

Таблица 6.3

Матрица косинусного сходства PPMI векторов документов рассматриваемого примера

	cos (doci, dock), $1 \le k, i \le 6$					
	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

6.3. Информационная технология определения принадлежности документа к узкоспециализированной тематике

Информационная технология определения принадлежности документа к узкоспециализированной области знаний включает в себя три основных этапа, показанных на рисунке 6.1: (1) лингвистическую обработку необработанного корпуса узкоспециализированной тематики (raw corpus) и поступающего на вход документа; (2) этап машинного обучения; (3) этап определения косинусного сходства.

Для определения принадлежности произвольного текста к криминальной тематике был использован созданный корпус, содержащий статьи четырех казахских новостных сайтов *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 информацию о таких криминальных деяниях, как грабежи, угоны машин, убийства, ДТП, чрезвычайные ситуации и происшествия, экстремизм, и другие преступления. Тексты именно данной предметной области и представляют базовый ресурс создаваемого корпуса.

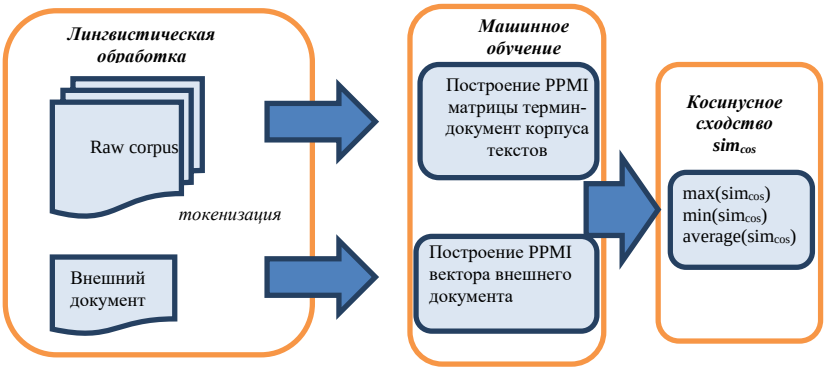

Рис. 6.1. Общая схема используемой технологии

Так как корпус представляет собой необработанные тексты (raw corpus), то на первом этапе создания матрицы термин-документ необходимо осуществить некоторую лингвистическую обработку текста, заключающуюся, в простейшем случае, в токенизации. Токенизация представляет собой разбиение текста на более мелкие части, токены, которыми в общем случае являются слова текста.

На втором шаге лингвистической обработки для большинства NLP приложений, обычно, осуществляется нормализация, заключающаяся в преобразовании внешне разных строк символов к одной и той же форме. Однако в агглютинативных языках, к группе которых относится 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 [21]. Исходя из выше изложенного, предлагаемая технология не использует нормализацию казахских текстов, так как ее использование приведет к резкому уменьшению точности определения близких по смыслу текстов.

Следующий этап предлагаемой технологии представляет собой этап машинного обучения, заключающийся в построении PPMI матрицы термин-документ корпуса текстов и PPMI вектора входного документа.

На третьем этапе вычисляется минимальное, максимальное и среднее значение коэффициента семантической близости, кото-

рый определяется косинусом между вектором документа, поступившим на вход, и векторами документов матрицы термин-документ имеющегося корпуса.

6.4. Эмпирическое определение значение коэффициента косинусного сходства

Для определения значения коэффициента *simcos*, позволяющего отнести входной документ к тематике рассматриваемого корпуса, было проведено два эксперимента. В обоих экспериментах сравнивался смысл входных текстов (test corpus) с документами построенного корпуса криминально значащих текстов (train corpus).

В ходе первого эксперимента вычислялось максимальное (max(simcos)), минимальное (min(simcos)) и среднее (average(simcos)) значение косинусного сходства тестируемого текста doctest и каждого текста обученного корпуса doctrain. Все тестируемые тексты содержали некоторый предопределенный криминальный смысл. В таблице 6.4 приведено максимальное, минимальное и среднее значения косинусного сходства между каждым конкретным криминально окрашенным входным текстом doctest и документами обученного криминального корпуса doctrain.

Таблица 6.4

Фрагмент результатов сравнения произвольных криминально окрашенных текстов и обученного корпуса

Имя файла	Косинусное сходство, simcos		
	max (simcos)	min (simcos)	average (simcos)
1	2	3	4
zakon.kz _2018-08-31_11.33_1.txt	0,92	0,36	0,72
zakon.kz _2018-08-31_12.27_2.txt	0,77	0,35	0,57
zakon.kz _2018-08-31_12.33_3.txt	0,79	0,38	0,67
zakon.kz _2018-08-31_16.24_4.txt	0,78	0,38	0,67
zakon.kz _2018-08-31_16.39_5.txt	0,82	0,36	0,69
zakon.kz _2018-09-01_17.44_6.txt	0,80	0,39	0,68

1	2	3	4
zakon.kz _2018-09-02_10.26_7.txt	0,88	0,27	0,69
zakon.kz _2018-09-02_12.58_8.txt	0,80	0,27	0,60
zakon.kz _2018-09-02_17.00_9.txt	0,87	0,31	0,71
lenta.kz_2018-01-14_13.50_2.txt	0,77	0,50	0,66
lenta.kz_2018-01-14_14.10_3.txt	0,82	0,48	0,68
lenta.kz_2018-01-14_14.40_4.txt	0,77	0,49	0,67
lenta.kz_2018-01-15_15.10_5.txt	0,73	0,34	0,55
lenta.kz_2018-01-16_11.10_6.txt	0,81	0,36	0,68
lenta.kz_2018-01-16_15.00_7.txt	0,88	0,31	0,71
lenta.kz_2018-01-16_18.40_8.txt	0,85	0,33	0,69
nur.kz_2018-10 16_00.09_1.txt	0,79	0,42	0,70
nur.kz_2018-10-17_00.13_2.txt	0,67	0,41	0,59
nur.kz_2018-10-17_00.18_3.txt	0,89	0,37	0,67
nur.kz_2018-10-18_00.18_4.txt	0,74	0,38	0,61
nur.kz_2018-10-19_00.02_5.txt	0,77	0,43	0,62
nur.kz_2018-10-22_00.19_6.txt	0,84	0,43	0,72

Анализ полученных результатов позволяет сделать вывод, что минимальное значение коэффициента косинусного сходства (simcos) между произвольными криминально окрашенными текстами тестового корпуса и документами обученного корпуса не меньше 0,3 (min(simcos)> 0,3), максимальное значение max(simcos)>0,7, а среднее значение average(simcos)>0,55.

В таблице 6.5 показаны результаты второго эксперимента. Минимальное, максимальное и среднее (min(simcos), max(simcos) и average(simcos)) значение коэффициентов косинусного сходства между документами обученного корпуса и входными текстами, относящихся к любой тематике, за исключением криминально специализированной.

Таблица 6.5

Фрагмент результатов сравнения текстов произвольной тематики и обученного корпуса

Имя файла	Косинусное сходство, simcos		
	max(simcos)	min(simcos)	average(simcos)
1	2	3	4
zakon.kz _2018-08-31_1.txt	0,63	0,24	0,41
zakon.kz _2018-08-31_2.txt	0,66	0,25	0,47
zakon.kz _2018-08-31_3.txt	0,69	0,19	0,45

1	2	3	4
zakon.kz _2018-08-31_4.txt	0,67	0,21	0,41
zakon.kz _2018-08-31_5.txt	0,73	0,24	0,48
lenta.kz_2018-09-6.txt	0,60	0,25	0,44
lenta.kz_2018-09-7.txt	0,51	0,15	0,36
lenta.kz_2018-09-8.txt	0,74	0,20	0,53
lenta.kz_2018-09-9.txt	0,67	0,22	0,47
lenta.kz_2018-09-10.txt	0,62	0,23	0,43
lenta.kz_2018-09-11.txt	0,76	0,19	0,50
lenta.kz_2018-09-12.txt	0,75	0,22	0,60
caravan.kz_2018-10-19_13.txt	0,72	0,25	0,49
caravan.kz_2018-10-19_14.txt	0,65	0,19	0,47
caravan.kz_2018-10-19_15.txt	0,66	0,19	0,50
caravan.kz_2018-10-19_16.txt	0,62	0,21	0,38
caravan.kz_2018-10-19_17.txt	0,56	0,28	0,45
caravan.kz_2018-10-19_18.txt	0,62	0,20	0,44
caravan.kz_2018-10-19_19.txt	0,60	0,18	0,39
caravan.kz_2018-10-19_20.txt	0,65	0,24	0,43
caravan.kz_2018-10-19_21.txt	0,73	0,25	0,50
caravan.kz_2018-10-19_22.txt	0,65	0,18	0,44
caravan.kz_2018-10-19_23.txt	0,59	0,23	0,47
caravan.kz_2018-10-19_24.txt	0,56	0,27	0,45
caravan.kz_2018-10-19_25.txt	0,61	0,25	0,48
caravan.kz_2018-10-19_26.txt	0,55	0,20	0,39
caravan.kz_2018-10-19_27.txt	0,70	0,28	0,41

После анализа полученных результатов второго эксперимента был сделан вывод о том, что среднее значение косинусного сходства между текстами обученного корпуса и текстами произвольной тематики находится в пределах $0,35 < average(simcos) < 0,50$. Максимальное и минимальное значения ниже 0,76 и 0,30, соответственно: $max(simcos) < 0,76$ и $min(simcos) < 0,30$.

На базе двух проведенных экспериментов 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.

6.5. Экспертная оценка предложенной технологии определения близости текстов к криминальной тематике

Для проверки правильности эмпирически определенного коэффициента семантического сходства документа к тематике *train corpus* использовалась традиционная метрика полноты, точности F1-мера. В результате проведенного эксперимента были проанализированы 1064 ранее не используемых документов тестового корпуса, из которых 520, заранее определены, как относящиеся к криминально значимой тематике и 544 текстов других тематических направлений.

В таблице 6.6 показан фрагмент результата экспериментальной оценки предложенной технологии определения семантической близости текстов к криминальной тематике.

Таблица 6.6

Фрагмент результата экспериментальной оценки предложенной технологии определения семантической близости текстов к криминальной тематике

Имя файла	априор. инфор.	max (simcos)	min (simcos)	Average (simcos)	Вывод системы
1	2	3	4	5	6
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_18.txt	Крим.	0,71	0,41	0,61	Крим.
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.

1	2	3	4	5	6
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.

При полученной средней гармонической семантической близости документа к узкоспециализированной тематике *F1-мера* ≈96%, полнота отнесения документа к узкоспециализированной криминально тематике *recall* = 98,5%, точность *precision* = 93,4%.

Заключение

В проведенном исследовании определяется узкоспециализированная тематика казахских текстов, при отсутствии предопределенных классов, путем оценки семантического сходства. Мы рассматриваем такую узко тематическую область, как криминальная окрашенность текста и, соответственно, все, что попадает под данную тему. Для исследования был создан специальный корпус криминально окрашенных текстов, которые автоматически собирались с новостных сайтов Казахстана.

В нашем исследовании для того, чтобы избежать проблемы отсутствие подлинной релевантности и получить более точные результаты используется Vector Space Model с PPMI в качестве весовой функции. Данная весовая функция, в отличие от взаимной информации, рассматривающей единичные события, использует MI (взаимная информация), оценивающей среднее значение всех возможных событий.

Полнота разработанной информационной технологии идентификации узкоспециализированной тематики документа составляет более 98%, в то время как точность составляет около 93%, а F-мера = 96%. Полученная полнота технологии несколько выше, чем точность. Что имеет некоторое практическое значение, связанное с тем, что при решении конкретной задачи идентификации

криминально значащих текстов, лучше иметь ошибку первого типа, создающую избыточность криминально значащих документов, чем ошибку второго рода, пропускающую криминально-окрашенные тексты.

Литература

1. Salton, G., Wong, A., & Yang, C.-S.: A vector space model for automatic indexing. Communications of the ACM, 18 (11), 613–620 (1975).
2. Yoon, Kim: Convolutional neural networks for sentence classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1746-1751, Doha, Qatar (2014).
3. Zaidi, N. A. S., Mustapha, A., Mostafa, S. A., Razali, M. N.: A Classification Approach for Crime Prediction. In: Applied Computing to Support Industry: Innovation and Technology, pp. 68-78. Springer, Heidelberg (2019).
4. Sayali, D. Jadhav, Channe, H.: Comparative Study of K-NN, Naive Bayes and Decision Tree Classification Techniques. *In:* Journal of Physics Conference Series, 1142(1):012011 (2016).
5. Bajwa, A. M., Collarana, D., Vidal, M.-E.: Interaction Network Analysis Using Semantic Similarity Based on Translation Embeddings. In: International Conference on Semantic Systems. SEMANTiCS 2019: Semantic Systems. The Power of AI and Knowledge Graphs, pp. 249-255 (2019).
6. Ehrmann, M., Talmi, J.: Starting from a blank page? Semantic similarity in central bank communication and market volatility. ECB Working Paper, No. 2023 (2017).
7. Araque, O., Zhu, G., Iglesias, C. A.: A semantic similarity-based perspective of affect lexicons for sentiment analysis. In: Knowledge-Based Systems, Volume 165, pp. 346-359 (2019).
8. Ming, Liu, Bo, Lang, Zepeng, Gu: Calculating Semantic Similarity between Academic Articles using Topic Event and Ontology. Published in ArXiv (2017).
9. Ramage, D., Rafferty, A.N., Manning, C.D.: Random walks for text semantic similarity. In: Proceedings of the 2009 workshop on graph-based methods for natural language processing, Association for Computational Linguistics, pp. 23–31 (2009).
10. Resnik, P.: Using information content to evaluate semantic similarity in a taxonomy. In: Proceedings of the 14th International Join Conference on Artificial Intelligence, 1995, pp. 448-453, San Mateo, CA (1995).
11. Gad, W.K, Kamel, M.S.: New semantic similarity based model for text clustering using extended gloss overlaps. In International Workshop on Machine Learning and Data Mining in Pattern Recognition, V.7., 23, pp. 663-677 (2009).

12. Turnay, P.D., Pantel, P.: From frequency to meaning: Vector Space Models of Semantics. Journal of Artificial Intelligence Research, 37, pp. 141-188 (2010).
13. Majumder, G., Pakray, P., Gelbukh, A., Pinto, D.: Semantic Textual Similarity Methods, Tools, and Applications: A Survey. Comp. Sist. vol.20, no.4, México (2016).
14. Turnay P.D., Pantel P. From frequency to meaning: Vector Space Models of Se-mantics / Journal of Artificial Intelligence Research 37 (2010). – P. 141-188
15. 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.
16. Солтон Дж. Динамические библиотечно-информационные системы: Пер. с англ. – М.: Мир, 1979. – 557 с.
17. Shannon, C. (1948). A mathematical theory of communication. Bell System Technical Journal, 27, 379–423, 623–656.
18. Pantel, P., & Lin, D. (2002). Document clustering with committees. In Proceedings of the 25th Annual International ACM SIGIR Conference. – Pp. 199–206.
19. 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
20. 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.
21. Johnson, H., & Martin, J. (2003). Unsupervised learning of morphology for English and Inuktitut. In Proceedings of HLT-NAACL 2003. – Pp. 43–45.

Мамырбаев О. Ж.
Институт информационных и вычислительных технологий, Алматы, Республика Казахстан, e-mail:morkenj@mail.ru:

Хайрова Н.Ф.
Национальный технический университет «Харьковский политехнический институт», Харьков, Украина, e-mail: khairova@kpi.kharkov.ua;

Мухсина К. Ж.
Казахский национальный университет имени аль-Фараби, Алматы, Республика Казахстан, e-mail: kuka_ai@mail.ru;

Колесник А.С.
Национальный технический университет «Харьковский политехнический институт», Харьков, Украина, e-mail: kolesniknastya20@gmail.com.

Глава 7

РАЗРАБОТКА МОДЕЛИ ПОСТ-РЕДАКТИРОВАНИЯ В МАШИННОМ ПЕРЕВОДЕ КАЗАХСКОГО ЯЗЫКА

Аннотация. *В работе представлен обзор современных технологий машинного перевода. Работа онлайн-переводчиков, используемых для перевода на казахский язык и обратно. Выявлены ошибки перевода, даны общие преимущества и недостатки онлайн систем машинного перевода на казахском языке. Представлена модель разработки системы пост-редактирования машинного перевода для казахского языка. Применены метод выравнивания и метод максимальной энтропии, проведен анализ процесса пост-редактирования, получены практические результаты.*

Введение

Современный мир и наше ближайшее будущее зависят от прикладных интеллектуальных систем, так как новые технологии развиваются с каждым днем. Одной из задач интеллектуальных систем является машинный (автоматизированный) перевод с одного естественного языка на другой. Машинный перевод (МП) позволяет людям общаться независимо от различия языков, поскольку это устраняет языковой барьер и открывает новые языки общения. Машинный перевод – это новая технология, особый шаг в развитии человека. Этот тип перевода может помочь, когда нужно быстро понять, что ваш собеседник написал или сказал в письме. Конечно, качество такого перевода очень низкое (для некоторых групп языков), но в большинстве случаев основной смысл можно понять. Что нужно сделать, когда необходим полный, смысловой перевод? Тут приходит на помощь ручной перевод, т.е. пост редактирование текста человеком.

В мире существуют специальные языки, используемые в различных сферах деятельности, с помощью которых проводятся деловые переговоры, исследования и другие виды деятельности.

В настоящее время в мире насчитывается более 6000 языков, треть из которых – языки развивающихся регионов, где языковой барьер является проблемой для выхода на зарубежные рынки. Современные разработки в области технологий МП позволили пользователям сделать еще один шаг к решению этой проблемы. Пользователем МП может быть любой сотрудник компании, инженер, юрист, врач, преподаватель, менеджер и т.д., поскольку он может получить качественный перевод многостраничных документов в короткие сроки.

В нашей стране МП казахского языка развивается с 2000-х годов. Один из первых машинный перевод начал исследовать профессор У.А. Тукеев. Ему удалось создать научную школу, которая активно занимается исследованиями в области МП. Среди отечественных учеников можно отметить исследование моделей и методов семантики машинного перевода с русского на казахский язык [1], статистической модели выравнивания англо-казахских слов с использованием алгоритма машинного перевода [2].

Научно-исследовательская группа под руководством профессора У. А. Тукеева работает с 2009 года в рамках проекта программы инициативного исследования «Разработка математической модели и программного продукта компьютерного перевода казахского языка на английский (простые предложения)», в результате которого грамматические формальные модели простых предложений и первая версия программы машинного перевода с казахского на английский язык [3-4]. В 2012–2014 гг. в результате реализации проекта «Разработка эффективных технологий компьютерного перевода казахского языка на английский и русский (и наоборот) на основе методов формальных грамматических и статистических методов» был создан многозначный метод для машинного перевода морфологически сложных естественных языков, таких, как русский и казахский [5-7]. В 2015–2017 годах в результате работы над проектом «Разработка бесплатной / открытой системы машинного перевода с казахского на английский и русский (и наоборот) на основе платформы Apertium» совместно с исследовательской группой профессора М. Форкада из Университета Аликанте (Испания), на основе базы платформы Apertium была разработана открытая система перевода, по грамматическим правилам перевода с казахского на английский,

с казахского на русский (и наоборот) [8-11]. С 2018 года ведутся работы над проектом «Разработка и исследование системы нейронного машинного перевода казахского языка». Однако за последние три или четыре года теория и практика машинного перевода значительно расширились, и было создано новое направление машинного перевода, которое отражает очень высокое качество. Это, в свою очередь, подняло стандарт качества промышленного машинного перевода на новую высоту [12-15].

Пост-редактирование предложения с одного языка на другой тесно связано с машинным переводом. В результате МП всегда есть определенные недостатки, которые можно устранить путем пост-редактирования. Пост-редактирование – обработка текста человеком после МП. Сегодня многие лингвистические провайдеры развивают, методы обучения редакторов и методы пост-редактирования. Основной задачей этой работы является повышение качества перевода с казахского языка и обратно. Модель, построенная на основе постредактирования, должна применяться на практике в форме экспериментальных систем машинного перевода. В этой работе представляется метод постредактирования, основанный на двухэтапных процедурах: первый обнаруживает ошибочные слова в тексте с использованием технологии памяти переводов, второй использует модель максимальной энтропии, чтобы определить, какой альтернативный перевод лучше всего подходит для неправильных слов, определенных на первом этапе. На практическом эксперименте исследуемая модель проверена по разным текстовым корпусам.

7.1. Обзор современных систем машинного перевода для казахского языка и классификация ошибок

Казахский язык является агглютинативным языком со сложным номинативным (морфологическим и синтаксическим) участием полисинтетизма. В связи с развитием нашей страны на глобальном уровне из года в год и ростом внешних связей, различные переводческие программы широко используются при переводе на казахский язык или с казахского языка на другие языки. Различные системы машинного перевода (СМП) все еще не могут

переводить полностью правильно, и есть ошибки в переводе, но область машинного перевода гораздо более развита, чем в предыдущие годы.

Для анализа ошибок МП были выбраны самые популярные системы машинного перевода. При переводе с русского на казахский язык могут возникать различные ошибки, поскольку казахский язык отличается от других языков, он имеет особые характеристики:

– близость лексической структуры;

– закон сингармонизма;

– агглютинация – серия аффиксов;

– отсутствие категории;

– отсутствие вспомогательных слов (предлогов);

– особый порядок слов.

Чтобы определить ошибки, возникающие при машинном переводе, был проведен сравнительный анализ работы перевода простых повествовательных предложений в системах машинного перевода. В таблице 7.1 приведены результаты перевода предложений с казахского на русский. Всего для анализа выбрано 8 программ переводчиков. Таблица 7.1 охватывает первые 4 вида.

Следующим этапом проверки работы СМП был перевод больших объемов научных и публицистических текстов, в Таблице 7.2 представлен пример работы перевода системы Яндекс для различных типов текста на казахскомя языке.

Таблица 7.1

Результат перевода с казахского языка на русский

Исходный текст для перевода на казахском языке	Наименование систем машинного перевода и результаты перевода			
	Google	Яндекс.Переводчик	Tridentsoftware	prof-translate.ru
Мектепке мен ертен барамын Анам мен әпкем жолда тұр Жолда тұрған анам мен әпкем	Я иду в школу завтра Моя мама и сестра уже в пути Мои мама и сестра в пути	И завтра иду в школу Мама и сестра стоят на дороге Мама и сестра, стоявшая на дороге	И завтра иду в школу Мама и сестра стоят на дороге Мама и сестра, стоящие в строке	И завтра иду в школу И завтра с утра иду в школу Мама и сестра стоят на дороге Мама и сестра, стоящие в строке
Мектепке мен ертен барамын Мен мектепке ертен барамын Мен кеше мектепке бармадым Ол кеше мектепке бармады Ол қазір меткепте	Я иду в школу завтра Я иду в школу завтра Я вчера не ходила в школу Он вчера не ходил в школу Она сейчас в школе	Я иду в школу Я иду завтра в школу Я не ходил в школу Этот человек не пошел в школу О қазір мектепе Он сейчас в школе	И завтра иду в школу И завтра иду в школу И вчера я не ходил в школу Он вчера в школу не ходил Он сейчас в школе	И завтра иду в школу И завтра иду в школу И вчера я не ходил в школу Он вчера в школу не ходил Он сейчас в школе
Жолдан қасқырды көрдік Жолда ешкім болған жоқ Бұл айналым жолы Басқа жол болмады	Мы видели волка на дороге На дороге никого не было Это объезд Другого пути не было	С дороги видели волка Никто не остался на дороге Путь этот оборот Не было другого пути	Строки видели волка На дороге никого не было Путь этот оборот Другого пути не было	Строки видели волка На дороге никого не было Путь этот оборот Другого пути не было

Алматы Көктөбе Қарағанды Үштөбе Жадыра Айкүміс	Алматы Коктобе Караганда Не беспокойся Жадыра Серебряный	Г. Алматы Коктобе Г. Караганда Уштобе Жадыра Айкумис	Алматы Коктобе Караганда Уштобе Жадыра Айкумис	Алматы Коктобе Караганда Уштобе Жадыра Айкумис
Желіде кемшіліктер бар болған Артышылықтардың тізімі көрсетілген еді Биіктен аққан ақ сәуле Арналарымыздың түрілері шектеулі Қанаттарымыздың ұзындығы анықталмаған	Были проблемы с сетью Был предоставлен список преимуществ Белый свет течет сверху Типы наших каналов ограничены Длина наших крыльев неизвестна	Есть недостатки в сети Был показан список преимуществ Ао утечки излучения с высоты Ограниченные типы каналов Длина крыльев не определена	Недостатки есть в сети Список преимуществ было указано в Ао утечки излучения с высоты С Арналарымыздың түрілері Длина Қанаттарымыздың не установлена	Недостатки есть в сети Список указанных Артышылықтардың бы Ао утечки излучения с высоты С Арналарымыздың түрілері Длина Қанаттарымыздың не установлена

Таблица 7.3

Примеры работы перевода системы Яндекс для различных типов текста на казахскомя языке

Жанр	Текст	Направление перевода каз. – рус.	Направление перевода каз. – англ.
Научный	Бүгінгі Алматы қаласының геологиялық кезең бойынша ертедегі палеозой дәуірен бергі үштік кезең яғни меозой дәуірінен бергі 60 000 000 жыл бұрын планетадағы Тетис теңізінің түбі болғаны геологилық мәліметтерден белгілі. Оған дәлел Памир тауларының биік шыңдарында балықтардың сүйек қаңқалары табылып жатқандығы бір дәлел болса, екінші дәлел өзіміздің кәдімгі Каратаудың жоталарында әлі зерттелмей жатқан түрлі моллюскалардың миллиондаған қалдықтары дәлел іспеттес.	По геологическим данным, на сегодняшний день на планете было дно Тетисского моря-60 000 000 лет назад, начиная с древней палеозойской эпохи. Об этом свидетельствует тот факт, что на возвышенных вершинах Памирских гор находят костные скелеты рыб, второй доказательством тому являются миллионы останков различных моллюсков, еще не исследуемых на хребтах нашего обычного Каратау.	According to geological data, today the planet had the bottom of the Tethys sea-60,000,000 years ago, starting from the ancient Paleozoic era. This is evidenced by the fact that on the high peaks of the Pamir mountains, fish skeletons are found, and the proof of this is the millions of remains of various mollusks that have not yet been studied on the ridges of our usual Karatau.
Публицистический	Қазақстанда коронавирустан тағы 4 адам қайтыс болды. Өлім нәтижелерін зерттеу жөніндегі комиссия бұл жағдайларды коронавирустық инфекциядан қайтыс болған деп тіркеуге шешім қабылдады.	В Казахстане умерло еще 4 человека от коронавируса. Комиссия по исследованию результатов смерти приняла решение зарегистрировать эти случаи как умершие от коронавирусной инфекции.	In Kazakhstan, 4 more people died from coronavirus. The death investigation Commission decided to register these cases as having died from a coronavirus infection.

Для различных СМП так же были проведены переводы с казахского на русский и английский, а также обратно для выявления и классификации ошибок при переводе на казахскомя языке.

Программы СМП имеют свои преимущества и недостатки, как показано в таблице 7.2. Проанализировав работу каждого из них, мы смогли выявить возможные ошибки при переводе с казахского языка и обратно. На основе ошибок постараемся выяснить, в каких случаях системы машинного перевода допускали ошибки при переводе с казахского на русский.

Таблица 7.2

Сравнительная характеристика работы онлайн СМП для казахского языка

№	СМП	Недостатки	Преимущества
1	Google переводчик	В больших предложениях есть некоторые несоответствия, но в большинстве случаев они очень малы	Хороший и качественный перевод сложных научных текстов, небольшие проблемы соответствия
2	Яндекс переводчик	При переводе больших текстов части предложений прерываются и не переводятся	Качественно переводит очень большие предложения, сложные научные тексты
3	Tridentsoftware	В больших предложениях есть некоторые несоответствия	Хорошие результаты перевода для коротких предложений и слов.
4	Prof-translate	При переводе некоторые предложения не принимаются, возникают трудности при комбинировании длинных предложений	Хорошие результаты перевода для коротких предложений и фраз
5	sozdik.kz	Перевод крупных научных текстов не дает хороших результатов	Хорошие результаты перевода для слов и коротких предложений.
6	translate.zakon.kz	Некоторые предложения не принимаются во время перевода	Хорошие результаты перевода для коротких предложений и слов.

7	elim.kz	В больших предложениях есть некоторые несоответствия	Хорошие результаты перевода для слов и коротких предложений.
8	Promt	Не дает очень хороших результатов при переводе больших научных текстов, многие предложения не согласованы	Показывает хорошие результаты перевода коротких предложений в журналистском стиле

Далее представлены примеры переводов и не корректных случаев перевода для казахского языка , для представленных выше в таблице восьми СМП.

1) Перевод слова «Мен» на казахском языке означает «Я». В некоторых случаях переводчик переводил его как «и». Попробуем перевести в разных предложениях и использовать его во фразах и словосочетаниях. В результате пришли к выводу, что перевод слова был написан с ошибкой, в том случаи, когда используется время. Например, СМП из таблицы 7.2 *Мен бүгін бармадым*, перевели *И сегодня бармадым*. Другие СМП смогли перевести разные версии перевода. Здесь мы обращаем внимание на слово «бармадым». Все СМП с словосочетанием *Мен бүгін* перевели правильно.

2) Перевод многозначных слов. В казахском переводе слово «Жол» является неоднозначным словом и переводится как «Путь», «Дорога», «Линия». Перевод не был выполнен правильно в СМП 1,5,6,7,8 (таблица 7.2). Чтобы проверить перевод, мы постараемся к слову «Жол» добавить новые словосочетания. В разных вариантах перевода это переводится как «Путь» и «Дорога». Лучший перевод (1,7,8). В множественном переводе получается очень много ошибок.

3) Неопределенность рода в переводе. Поскольку в казахском языке нет рода, СМП часто допускали ошибки. Произведем перевод с использованием слова «Ходить». *Ол бармады – Она не ходила* правильный перевод, тут слово «Ол» с казахского языка можно перевести как он, она или оно. Все СМП 2,3,4,5,6,7 перевели в мужском роде, и следующим образом, то есть добавляется предлог «в». Для определения рода добавляем слово «күйеуімен», чтобы определить фамилию. В результате все СМП, кроме

(1), перевели неправильно. Получилось *Он не ходил в куйеуімен*, при правильном переводе должно было быть *Она не пошла с мужем*.

4) Перевод имен собственных. Обнаружены ошибки переводчиков (1,5,8). Имя *Айкумис* переводится как *Серебро*. Подобные имена были переведены правильно. Название города *Уштобе* переводится, как *Не беспокойся*. Все казахстанские СМП имена собственные перевели правильно.

5) Перевод во множественном числе. Слова *Арналарымыздың* и *Қанаттарымыздың* переведены неправильно. СМП 3,4,5,6,7 не смогли перевести. Чтобы определить, относятся ли эти слова к множественному числу, мы пытаемся перевести эти слова в другие вариантах. В результате было обнаружено, что многие СМП не могут переводить, в том случае, когда к словам прибавляется множественное окончание.

6) Ошибки в переводе литературных текстов. При переводе предложения *Биіктен аққан ақ сәуле* только СМП 1 и 8 перевели правильно, остальные переводы были переведены в виде *Ао утечки излучения с высоты*. Слово «Ақ» было неправильно переведено, чтобы выявить ошибку перевода мы пытались перевести его в разные предложениях и словосочетаниях.

Следует использовать морфологический анализ предложений, чтобы избежать различных ошибок при переводе с казахского на русский и обратно. Морфологический анализ является начальным этапом различных задач, связанных с естественным языком, поэтому его фактическая реализация имеет большое значение. Методы морфологического анализа можно разделить на 3 типа:

– анализ словаря аффиксов;

– анализ с использованием словаря аффиксов и оснований;

– анализ с использованием системы словарных слов.

Теперь сделаем морфологический анализ предложения. *Жолдан қасқырды көрдік.*

1. *Жолдан* – откуда? Обстоятельство, исходный падеж

2. *Қасқырды* – кого? Дополнение, винительный падеж

3. *Көрдік* – что сделали? Сказуемое, корень – *көр*, глагол, *ді* – суффикс, *к* – личное окончание, 1-е лицо, множественное число. Здесь *біз* есть скрытый существительное, личное местоимение.

В результате перевода получили *С дороги видели волка* и *Строки видели волка* перевод показывает ошибки переводчиков. Предложения, использованные для русско-казахского перевода и переводчиков, остались в том же порядке. В итоге получили классифицирование ошибок переводчиков.

1) Многозначность слов. Ошибки в переводе родства. Слово «Сестра» переводится на казахский язык как «Әпке» или «Апа», а слово «Сестренка» – «Қарындас». Но в русском языке слово «Младшая сестра» часто используется, на казахский переводится, как «Қарындас» или «Сіңлі». После проверки перевода выяснилось, что все СМП переводили неправильно, и только (1) переводчик правильно перевел слово «Тесть».

2) Перевод слов связанных со временем. В примере слово «Тұру» не переведено, правильная версия – «Тұрған». 1,2,8 СМП перевели правильно.

3) Слова с глаголами. «Не ходил», «Бармадым» или «Бармады».

4) Имена собственные. Перевод имени Айкумис остался неизменным среди СМП 1,2,5 и 7.

5) Слова взятые из других языков. При переводе слова «Проблема» перевел только (1) переводчик, остальные оставили слово без изменений.

6) Несоблюдение порядка написания предложений. В этом случае слова переводятся правильно, но не размещаются правильно, в результате чего общий смысл предложения теряется.

7.2. Разработка системы пост-редактирования в машинном переводе

Постредактирование предложения с одного языка на другой тесно связано с машинным переводом. В результате машинного перевода всегда есть определенные недостатки, которые можно решить путем постредактирования. Постредактирование это – обработка текста человеком после машинного перевода. На сегодняшний день многие лингвистические провайдеры активно развивают эту сферу, развивают методы обучения редакторов и методы постредактирования. Основной целью данной работы

является повышение качества в системе казахско-английского машинного перевода посредством постредактирования. Модель, созданная с помощью постредактирования, должна реализовываться в виде экспериментальных машинных переводческих систем. В данной работе предлагается метод постредактирования, основанный на двухэтапных процедурах: первая выявляет ошибку в тексте с помощью технологии памяти переводов, а вторая выявляет какой альтернативный перевод лучше подходит для неверных слов, определенных на первом этапе с использованием модели максимальной энтропии. Изученная модель в практике проверяется на различных текстовых корпусах.

Когда дело доходит до машинного перевода, компьютер не понимает смысла текста и нюансов языка. Каждая новая структура предложения, фраза, идиомы включены в программу. В зависимости от стиля и цели текста слово может иметь несколько значений. В настоящее время программное обеспечение машинного перевода может распознавать стиль текста и выбирать желаемый перевод соответствующим образом. Сама программа также может предложить несколько переводов для переводчика. Поэтому работа была выполнена с учетом всех этих нюансов.

Основная идея этой работы – найти неправильные слова в английских предложениях, которые переведены с казахских предложений и исправить неправильные английские слова. В результате полученной работы исправляется правильное предложение на английском языке. Рекомендуется использовать технологию памяти переводов, чтобы находить неправильные слова в переведенном тексте, и использовать модель максимальной энтропии для исправления неправильных слов [21]. Система ранее применялась к «обученным», и «рабочим» текстам в параллельных корпусах. Благодаря предварительному обучению системы можно повысить точность перевода, сделать терминологию более удобной и сократить расходы постредактирования.

На рис. 7.1-7.2 можно увидеть модели пост-редактирования машинного перевода.

На рис. 7.1 показано, как выполняется пост-редакционный машинный перевод: сначала текст дается на определенном языке (source), затем этот текст передается на машинный перевод и обновленный словарь, на следующем этапе автоматически перева-

денный текст проверяется постредактором и получается готовый текст.

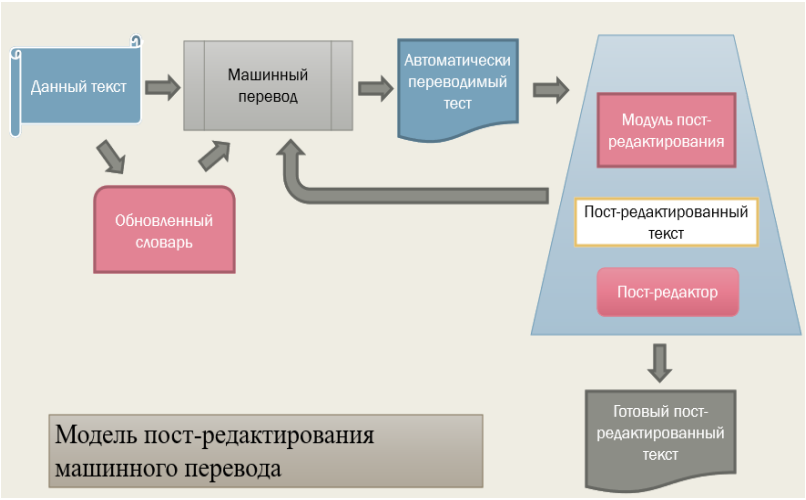

Рис. 7.1. Модель постредактирования машинного перевода

На рисунке 7.2 изображена расширенная модель пост-редактирования машинного перевода.

Рис. 7.2. Расширенная модель постредактирования машинного перевода

На рисунке 7.2. также показана модель пост-редакционного машинного перевода. На этом рисунке изображено, как происходит пост-редактирование машинного перевода в циклическом алгоритме перевода.

Архитектура пост-редактирования машинного перевода показана на рис. 7.3.

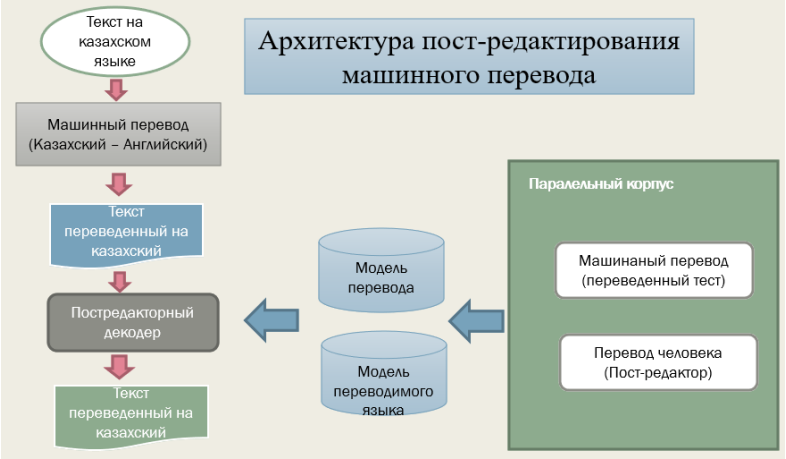

Рис. 7.3. Архитектура пост-редактирования машинного перевода

Архитектура пост-редактирования машинного перевода подробно показывает, как поэтапно реализуется перевод с казахского языка с применением постредакторного декодера.

7.2.1. Методы обработки слов

Предлагаемый нами подход состоит из двух модулей. Функция постредактирования обнаруживает неправильные, неверные слова и чтобы применить анализ определяет текст после перевода и модуль. Метод максимальной энтропии используется для обработки слов и фраз в казахско-английских предложениях. Основная цель модели обучения определяется первым этапом – нахождение неверных, ошибочных слов и исправление этих слов.

Изначально мы берем казахские предложения и переводим их с помощью Google Translate, затем записываем правильный перевод казахских предложений вручную. У нас есть два файла: от переводчика и самостоятельно переведенные казахские предложения. Эти предложения двух файлов (Google Translate и ожидаемый правильный перевод) выровнены с помощью метода выравнивания матричных фраз (Koehn, Statical Machine Translation, стр. 113) [27] и с несовместимыми сегментами и словами. Переведенный казахский текст указывает что определено.

Основная цель второго этапа – исправить предложение заранее определенными неверными словами. Для анализа исходных данных используется небольшое количество (приблизительно 100 000) параллельных предложений на обоих языках. Эти предложения используются при создании таблиц для обучения системы. Мы используем максимальную энтропию для создания кубических таблиц. То есть система обучается по заранее созданным кубическим таблицам. Таблица основана на двуязычном параллельном корпусе и двуязычном словаре.

Индикаторы показывают, что запускается только первый уровень, умножается на два для второго уровня и дает остаток от деления на количество других уровней. Такой подход необязателен для выполнения классификатора, но для понимания теории необходимо понимать разницу между знаком и индикатором, а также разницу в их нумерации.

Классификация проводится по формуле:

$$p(c|d, \lambda) = \frac{exp \sum_i^{n*k} \lambda_i f_i(c,d)}{\sum_{\tilde{c} \in C} exp \sum_i^{n*k} \lambda_i f_i(\tilde{c},d)}. \quad (7.1)$$

В этой формуле [28]:

– f_i – i-ый классификационный индикатор (0 или 1);

– λ_i – вес i-го классификационного индикатора f_i;

– c – гипотеза класса;

– C – сумма всех возможных классов;

– D – классифицированный документ.

У каждого индикатора f_i есть свой вес λ_i, который описывает взаимосвязь между соответствующим классификационным критерием и классом. Чем больше вес, тем сильнее связь. Таким

образом, числитель дроби описывает экспоненту весов для класса-гипотезы, а знаменатель нормирует значение по единице. Самая сложная часть этой формулы – набор весов λ, который приходится определять путем численной оптимизации, о которой мы поговорим позже.

7.2.2. Применение метода в системе пост-редактирования

Метод максимальной энтропии очень тесно связан с другим распространенным алгоритмом машинного обучения – логистической регрессией. То есть при использовании этого метода мы получаем альтернативное слово, которое описано выше и имеет наиболее близкое значение к контексту. Результатом этого подхода является не просто решение о классификации, но вероятность для определенного класса. Одним из преимуществ этой классификации является спецификация возможного распределения классов. Если при использовании текстового машинного перевода с казахского на английский в предложении есть неверные, ошибочные слова должно быть проведено постредактирование текста.

Затем, после нахождения неправильных, ошибочных слов в предложении, проверяем поочередно их перевод, чтобы найти и ввести правильный перевод для использования его в основе метода максимальной энтропии. Общее описание метода следующее:

$$f_i^j = \begin{cases} 1, if\ d = w_i, c = AW_j \\ 0, in\ other\ cases \end{cases} \quad (7.2)$$

$AW_j -$ альтернативные слова(класс), $d -$ классифицированный документ, f_i – i-ый классификационный показатель (0 или 1), j – $\overline{1,n}$;

Для исправления неверных слов используются база данных многозначных словарей и TM (Translation Memory) В качестве примера приведены следующие элементы (таблица 7.4.).

Для практического эксперимента был взят текст с небольшими предложениями. В этом тексте слово *«ана» (мать, мама)* было переведено неправильно, и были определены его эквиваленты:

Таблица 7.4

Примеры синонимов и альтернативных слов казахского языка, встречающиеся в текстах

Альтернативные слова	Пример словосочетания
Ана	Анаңа қызықты кітап ал, ана жақсы көреді
Мама	Мамасының қуанышына айналады
Ене	Енесі пісірген
Шеше	Шешем бәлішпен бір құты маймен
Апа	Апасы бәліш пісіріп

1. Анасы өте жақсы көреді екен (The mother was very fond of it).

2. Бір күні апасы бәліш пісіріп, қызына кешікпеуін айтады (One day, her mother bake a cake, the daughter of late say).

3. Оған енесі пісірген бәлішінен және бір құты май алып бара жатырмын (It'm going to my mother's bälişinen and a bottle of cooking oil).

4. Сен мына жолмен бар, мен ана жолмен жүрейін (You are this way, and the mother in a way).

5. Мен сізге шешем бәлішпен бір құты маймен жіберді (I sent you my cake and a bottle of oil).

6. Ұлы әкесінің мақтанышына, мамасының қуанышына айналады (The pride of his father's, mother's joy becomes).

7. Сен анама қызықты кітап ал (You have an interesting book and a mother).

Мы использовали максимальную энтропию:

$$f^1 = \begin{cases} 1, if\ d = »f_3 \wedge f_4 \wedge f_5», c = AW_1 \\ 0, in\ other\ cases \end{cases} \quad (7.3)$$

$$f^2 = \begin{cases} 1, if\ d = »f_1 \wedge f_2 \wedge f_5», c = AW_2 \\ 0, in\ other\ cases \end{cases} \quad (7.4)$$

$$f^3 = \begin{cases} 1, if\ d = «f_2 \wedge f_5», c = AW_3 \\ 0, in\ other\ cases \end{cases} \quad (7.5)$$

$$f^4 = \begin{cases} 1, if\ d = »f_2 \wedge f_3 \wedge f_4», c = AW_4 \\ 0, in\ other\ cases \end{cases} \quad (7.6)$$

$$f^5 = \begin{cases} 1, if\ d = «f_4 \wedge f_5», c = AW_5 \\ 0, in\ other\ cases \end{cases} \quad (7.7)$$

7.3. Практические результаты

7.3.1. Описание функции обнаружения слов ошибок и их постредактирования

В результате расчета вероятность неправильного разделения слов на части речи были получены следующие значения:

$$\mathrm{P}(AW_1) = 0.427$$
$$\mathrm{P}(AW_2) = 0.998$$
$$\mathrm{P}(AW_3) = 0.713$$
$$\mathrm{P}(AW_4) = 0.855$$
$$\mathrm{P}(AW_5) = 1.142$$

Метод максимальной энтропии, используя характеристики и вес различных альтернативных слов, выбираем наибольшее значение 1,142. Это только определенные части предложения и максимальное значение вероятности, а мы рассмотрим контекст неправильных слов, примененных после использования заполненных кубических таблиц. То есть учитывается не только вероятность частей предложения и требуемых слов, но и значение каждого слова в тексте. Этот метод используется на втором этапе в функциях обнаружения ошибок, которые выполняются путем записи и постредактирования ошибочных слов, при создании таблицы для каждого неправильного слова.

Основная работа состоит из двух этапов (модулей). Первый этап – в переводе с любого казахского предложения найти пра-

вильное слово на английском языке. Эта часть поиска и маркировки неправильных слов или сегментов предложения на английском языке основана на методе памяти переводов. Функция обнаружения некоторых неправильных слов и постредактирования этих неправильных слов на первом этапе связана со списком неправильных слов, чтобы дать подробное описание, он связан с файлом, состоящим из трех типов предложений (Английские предложения после алгоритмов постредактирования, казахские предложения, правильные казахские предложения). Каждая строка выглядит так: *interesting subject, қызықты сабақ, қызықты пән* и т.д. Мы находим неправильные слова, делим их, находим корни и сравниваем их, если взять с примера: сабақ болады. Разработан алгоритм постоператора, в результате которого в файле обнаруживается и записывается только одно неправильное слово со списком многозначных слов. Поскольку нам нужно перечислить как можно больше неправильных слов, мы также рассмотрим многозначные слова. Morphological Analyzer Apertium предназначен для разработки алгоритма разделения слов между корнем и окончанием в казахском языке.

На первом этапе выявление неправильных слов и постредактирование неправильных слов начинается работа с таблицами. Эти таблицы должны быть готовы для любого нового предложения. Чтобы подготовиться, выполняем следующее:

1. Нахождение неправильного казахского слова в казахско-английском словаре, запомните все переводы этого слова.

2. Затем определение основы предложений с этим казахским словом и рассмотрите перевод этих предложений на английский язык.

3. Создание таблицы для каждого неправильного слова и для каждой таблицы определяется перевод слова на казахском языке. То есть, если слово *сабақ* является неправильным и отображается в словаре как *тақырып*, тогда таблица будет заголовком для этого слова. Все эти таблицы сохраняются в виде файла для каждого неверного слова. Вспомнив все синонимы слов, которые встречаются в казахско-английском словаре, записываем эти слова рядом. Для каждого случая подсчитайте, сколько раз они появляются в базе данных из 100 000 предложений. В

корпус можно добавить ещё больше предложений. Таким образом, появляется много таблиц с неверными, ошибочными словами.

7.3.2. Описание тестового этапа

На втором этапе были исправлены ошибочные слова, найденные на первом этапе с помощью метода кубических таблиц на основе максимальной энтропии. Казахско-английский словарь необходим для определения эквивалентов ошибок слов или сегментов перевода на казахский язык. Производился поиск английских и казахских значений этих слов в двуязычном корпусе. Качество постредактрования зависит от контекста данного корпуса, и оно может определять значение предложения на основе контекста и слова. Ниже более подробно описан корпус, с казахскими предложениями, который рассчитан из параллельной кубической таблицы казахско-английских языков.

На этом этапе мы будем использовать предыдущий модуль для перевода любых входных предложений на английский язык.

$$Sentence_{english} \rightarrow Sentence_{kazakh} \rightarrow Sentence_{correct}$$

1. Перевод с помощью СМП (в данном случае применился Переводчик Яндекс [29]).

2. Функциональным этапом обнаружения и постредактирования ошибочных слов является выполнение работы алгоритма Стемминг, который ищет ошибочные слова в файле incorrect.txt, чтобы вычислить вероятность каждого найденного ошибочного слова.

3. Применяются вероятностные алгоритмы.

Когда все слова найдены в таблице, необходимо прочесть переведенное предложение, которое мы исправили. После этого нужно взять только те слова, которые используются в этом предложении, а не использовать весь файл. Например, переведенный текст: *Сабақ кеш болды*.

Нашли в предложении *The lesson as late* неправильно переведенное слово *сабақ*. (В заголовке есть синонимы урок, предмет и

т.д.). Используются только слова, которые есть в переведенном предложении. *Yesterday was a subject*, переведенный как [30]: *Кеше сабақ болды*, определяем слова *кеш* и *бол* с многозначными словами и вычисляем вероятность только для этих слов. Рассчитывается по формуле: P(s) = P(s_1) + P(s_2) + ⋯ + P(s_n) и используется вероятность вышеупомянутых слов в предложениях; *Subject* есть несколько синонимов слова урок: *lesson* (сабақ), *object* (тақырып), *subject* (субьект). Выберите максимальное значение и получите вторую возможность, которая является названием, и поместите это значение в предложение.

4. Чтобы правильно рассчитать слово, найдите предел слова в предложении и используйте морфологический анализатор Apertium. В результате получаем полное предложение с указанным словом после исправления.

Таким образом, пример результата описанной системы направлен на тестовый уровень. В итоге правильно корректируем предложение на английском языке, на первом этапе рассматривается предложение переведенное любым переводчиком с казахского на английский язык, на втором этапе проводятся соответствия для получения качественного перевода.

При развитии алгоритма необходимо дополнить данные, то есть чем точнее предложения, тем точнее информация о редактировании предложений. Также необходимо обновлять и добавлять базу данных TM.

Обратите внимание, что это может быть следующим:

1. Если слово в incorrect.txt не содержит хотя бы одного корня, это предложение не рассматривается, это объясняется, тем что в предложении нет неправильных слов.

2. Слово может быть найдено в incorrect.txt, но если оно не найдено в списке таблиц на этапе идентификации неправильных слов и их редактирования, мы должны найти их, и принять во внимание эти слова, алгоритм создания таблиц должен добавить эти новые слова.

3. В предложении может быть несколько неправильных слов, и тогда мы получим все неправильные слова.

В результате работы, выполненной с использованием этого метода, был сделан небольшой анализ предложений (100 предложений). Различные СМП были использованы для проверки пер-

вого этапа, который выявил неправильные слова или сегменты. По результатам анализа показаны следующие результаты.

В результате у этих трех СМП было найдено количество слов, измененных в соответствии с таблицей, указывающей на использование первого этапа вышеуказанным методом. В соответствии с таблицей было выявлено то, что использование метода памяти переводов позволяет определить слова, которые необходимо изменить.

Таблица 7.5

Процентное соотношение между экспертными показаниями казахско-английского перевода и неверными словами в разработанной системе

Google Translate	Yandex Translate	Promt Translate
11%	13%	16%

Для анализа используем «золотой стандарт» по технологии алгоритма BLEU оценки перевода. В качестве экспертов были задействованы 3 специалиста, которые свободно говорят на исходном языке, а исходный текст – текст на казахском языке. Задача специалистов перевести эти тексты. Предложения, переведенные экспертами, сравниваются с обработанными предложениями в системе. В результате соответствие между переводами экспертов и нашим методом имеет следующее процентное соотношение (табл. 7.5.). Учитывая совпадение, можно более точно сказать о предлагаемой системе и повысить качество перевода.

Заключение

Языковые навыки традиционно были характерной чертой образованных людей. Сегодня, благодаря сочетанию передовых технологий машинного обучения и высококачественных данных, результат автоматического перевода получается, в большинстве случаев качество не уступает искусственному переводу. В целях повышения качества перевода система постоянно обновляется и проверяется, что существенно упрощает и ускоряет работу переводчиков. Проанализированы работы современных технологий машинного перевода. Работа онлайн-переводчиков, исполь-

зуемых для перевода на казахский язык и обратно. Выявлены ошибки перевода, даны общие преимущества и недостатки онлайн систем машинного перевода на казахском языке. Представлена модель разработки системы пост-редактирования машинного перевода для казахского языка. Применены метод выравнивания и метод максимальной энтропии, проведен анализ процесса пост-редактирования, получены практические результаты. Работа сосредоточена на сочетании этих двух методов, применение метода памяти переводов для определения неправильного (некорректного) слова, метод максимальной энтропии для редактирования неправильных слов. Предложенный метод пост-редактирования повышает качество машинного перевода казахского текста. В дальнейшем планируется продолжить исследование в данной области с повышением качества перевода с казахского на английский язык и обратно.

Благодарность

Данная работа была проведена в рамках проекта ИРН AP08052421 «Исследование и разработка системы постредактирования казахского языка в машинном переводе» при поддержке Министерства образования и науки РК.

Литература

1. Рахимова Д.Р. Исследование моделей и методов семантики машинного перевода с русского на казахский язык. Диссертация. – Алматы, 2014 г. https://www.kaznu.kz/content/files/pages/folder14360/%D0%94%D0%B8%D1%81%D1%81%D0%B5%D1%80%D1%82%D0%B0%D1%86%D0%B8%D0%BE%D0%BD%D0%BD%D0%B0%D1%8F%20%D1%80%D0%B0%D0%B1%D0%BE%D1%82%D0%B0%20%D0%A0%D0%B0%D1%85%D0%B8%D0%BC%D0%BE%D0%B2%D0%BE%D0%B9%20%D0%94.%20%D0%A0..pdf (это эл ссылки , так дает интерент ресур)
2. Картбаев А.Ж. Разработка модели и методов сатистического машинного перевода с приложением к казахскому языку. Диссертация. – Алматы, 2018 г. https://www.kaznu.kz/content/files/pages/folder17928/%D0%94%D0%B8%D1%81%D1%81%D0%B5%D1%80%D1%82%D0%B0%D1%86%D0%B8%D1%8F%20%D0%9A%D0%B0%D1%80%D1%82%D0%B1%D0

%B0%D0%B5%D0%B2%20%D0%90%D0%96%20%D0%B7%D0%B0%D1%89.pdf (это эл ссылки , так дает интерент ресур)
3. Zhumanov Zh.M., Tukeyev U.A. Development of machine translation software logical model (translation from Kazakh into English language). Reports of the Third Congress of the World Mathematical Society of Turkic Countries, Volume 1 (June 30 – July 4, 2009) / Edited by Academician Bakhytzhan T. Zhumagulov. – Almaty: Қазақ университеті, 2009. – 356-363 p.
4. Tukeyev U., Zhumanov Zh., Rakhimova D. Features of development for natural language processing. In book «ICT – from theory to practice» edited by M.Milosz. Polish Information Processing Society, Lublin, 2010, 149-174 pp.
5. Tukeyev U., Rakhimova D. Augmented attribute grammar in meaning of natural languages sentences. The 6th International Conference on Soft Computing and Intelligent Systems, and the 13th International Symposium on Advanced Intelligent Systems, SCIS-ISIS2012, Kobe, Japan on November 20-24, 2012, 1080-1085 pp. (Индекс Скопус)
6. Тукеев У.А., Рахимова Д.Р., Байсылбаева К., Умирбеков Н., Оразов Б., Абақан М., Кызырканова С., Көпмағыналық бейнелеу кесте тәсілі негізінде орыс тілінен қазақ тіліне машиналық аудармасының морфологиялық анализбен синтезін құру. түркі тілдерін компьютерлік өңдеу. Бірінші халықаралық конференция: Еңбектері. – Астана: Л.Н. Гумилев атындағы ЕҰУ баспасы, 2013, 182-191.
7. Тукеев У.А. Разработка технологии машинного перевода на основе метода многозначных отображений для морфологически сложных языков. Труды 4-й Международной научно-практической конференции «Информатизация общества». – Астана, 2014, стр. 130-132.
8. Tukeyev, U., Miłosz, M., Zhumanov, Zh. Finite-State Transducers with Multivalued Mappings for Processing of Rich Inflectional Languages. // Lecture Notes in Computer science. New trends in intelligent information and database systems (Vol. 598, pp. 271-280). Springer. 2015. (Индексировано в WoS, Scopus)
9. Tukeyev, U., Automaton models of the morphology analysis and the completeness of the endings of the kazakh language. Proceedings of the international conference «Turkic languages processing» TURKLANG-2015 September 17–19. – Kazan, Tatarstan, Russia, 2015. – Pp. 91-100.
10. 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. – Алматы, Қазақ университеті, 2016. – №2 (89). – P. 110-117.
11. Tukeyev U., Sundetova A., Abduali B., Akhmadiyeva Zh., Zhanbussunov N. Inferring of the morphological chunk transfer rules on the base of complete set of Kazakh endings // Lecture Notes of Artificial Intelligence (LNAI) vol. 9876, Computational Collective Intelligence, Part 2, Springer, 2016, pp.563-574 (Индексировано в WoS, Scopus).

12. Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic Regularities in Continuous Space Word Representations. The Association for Computational Linguistics. In HLTNAACL, p. 746–751 (2013).
13. Nal Kalchbrenner, Phil Blunsom. Recurrent Continuous Translation Models. Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, Washington, USA, p. 1700–1709(2013).
14. Mikel L. Forcado and Ramon P. Neco Recursive Hetero-Associative Memories for Translation. International Work-Conference on Artificial and Natural Neural Networks, IWANN'97 Lanzarote, Canary Islands, Spain, p. 453-462(1997).
15. Ilya Sutskever, Oriol Vinyals, Quoc V. Le Sequence to Sequence Learning with Neural Networks. arXiv:1409.3215v3 [cs.CL](2014).
16. https://ru.wikipedia.org/
17. How does Neural Machine Translation work? | SYSTRAN Blog: [Электронный ресурс]. 2016. URL: http://blog.systransoft.com/how-does-neural-machine-translation-work/
18. eMpTy Pages: A Deep Dive into SYSTRAN's Neural Machine Translation (NMT) Technology: [Электронный ресурс]. 2016. URL: https://kv-empty-pages.blogspot.ru/2016/09/a-deep-dive-into-systrans-neural.html
19. Колганов Д.С., Данилов Е.А. Обзор аналитической, статистической и нейронной технологий машинного перевода // Материалы X Международной студенческой научной конференции «Студенческий научный форум» URL:<ahref=http://scienceforum.ru/2018/article/2018009616»> http://scienceforum.ru/2018/article/2018009616 (дата обращения: 20.05.2020).</p>
20. Эспла, М., Санчес-Мартинес, Ф., Форкада, М.Л .: Көмек беру үшін сөз тіркестерін қолдану компьютерлік аударма пайдаланушылары қай мақсатты сөздерді өзгерту немесе сақтау керектігін белгілеп өңделмеген. Авторы: Еуропалық қауымдастықтың жыл сайынғы 15-ші конференциясының материалдары Машиналық аударма, 81-89 бет, Левен, Бельгия (2011)
21. https://en.wikipedia.org/wiki/Principle_of_maximum_entropy
22. Translator Promt. http://www.promt.ru
23. Google аудармашысы. https://translate.google.kz/#kk/kz
24. http://www.krugosvet.ru/enc/gumanitarnye_nauki/lingvistika/MASHINNI_PEREVOD.html
25. Эспла-Гомис, М., Санчес-Мартинес, Ф., Форкада, М.Л.: Интернеттегі қол жетімді дереккөздерді пайдалану сөздік деңгейдегі машиналық аударма сапасын бағалауға арналған екі тілді ақпарат. Кіріспе: іс Машиналық аударма бойынша Еуропалық ассоциацияның 18-ші жыл сайынғы конференциясының 19-беті 26, Анталия, Түркия (2015).
26. Эспла-Гомис, М., Санчес-Мартинес, Ф., Форкада, М.Л.: мақсатты жақтағы сөздерді өзгертуді ұсыну үшін компьютерлік аударма. Кіріспе: еңбектер Машиналарды аударудың 13-ші саммиті, 19-23 қыркүйек 2011 ж., Сямень, Қытай, 172–179 бет.

27. Koehn: Статистикалық машиналардың аудармасы, б. 113. http://www.statmt.org/book/
28. http://bazhenov.me/blog/2013/04/23/maximum-entropy-classifier.html
29. Аудармашы Яндекс. https://translate.yandex.kz/
30. http://wiki.apertium.org/wiki/Main_Page
31. Кадвелл, Патрик, Шейла Кастильо, Шарон О'Брайен және Линда Митчелл 2016. «Институционалды аудармашылар арасындағы машиналық аударма және пост-редакциядағы адам факторлары». Аударма кеңістігі 5 (2): 222-243.
32. Кадвелл, Патрик, Шарон О'Брайен және Карлос С.К. 2017. «Қарсыласу және орналастыру: кәсіби аудармашылар арасында машиналық аударманы қабылдау факторлары».
33. Кастильо, Шейла, Джосс Моркенс, Федерико Гаспари, Айзер Калисто, Джон Тинсли, Энди Уэй. 2017b. «Нейрондық машиналық аударма қазіргі заманғы жаңа жағдай ма?» Прага Математикалық лингвистиканың хабаршысы 108: 109-120.

Рахимова Д.Р.
КазНУ имени аль-Фараби, Алматы, Казахстан e-mail: diana.rakhimova@kaznu.kz

Турарбек А.Т.
КазНУ имени аль-Фараби, Алматы, Казахстан e-mail: turarbekasem@kaznu.kz

Пазылхан Н.М.
КазНУ имени аль-Фараби, Алматы, Казахстан e-mail: npazylhan@gmail.com

Научное издание

ВЫЧИСЛИТЕЛЬНАЯ ОБРАБОТКА КАЗАХСКОГО ЯЗЫКА

Сборник научных трудов

под редакцией Д.Р. Рахимовой

Редактор *З. Усенова*
Компьютерная верстка *Г. Калиевой*
Дизайн обложки *Б. Малаева*

В оформлении обложки использованы фотографии
с сайта www.google.com

ИБ №13759
Подписано в печать 11.08.2020. Формат 60x84 $^{1}/_{16}$. Бумага офсетная.
Печать цифровая. Объем 9,18 п.л. Тираж 50 экз. Заказ №10393.
Издательский дом «Қазақ университеті»
Казахского национального университета им. аль-Фараби.
050040, г. Алматы, пр. аль-Фараби, 71.

Отпечатано в типографии издательского дома «Қазақ университеті».