На правах рукописи

ЕРИМБЕТОВА
Айгерим Сембековна

ЛИНГВИСТИЧЕСКОЕ И АЛГОРИТМИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ПРОЦЕССА ИНФОРМАЦИОННОГО ПОИСКА НА ОСНОВЕ ГРАММАТИКИ СВЯЗЕЙ, В ТОМ ЧИСЛЕ ДЛЯ ТЮРКСКИХ ЯЗЫКОВ

Специальность 05.13.17 – Теоретические основы информатики

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата технических наук

Новосибирск – 2019

Работа выполнена в Федеральном государственном бюджетном учреждении науки Институте систем информатики им. А.П. Ершова Сибирского отделения Российской академии наук (ИСИ СО РАН)

Научный руководитель:
Мурзин Федор Александрович
кандидат физико-математических наук, зам. директора по научной работе, Федеральное государственное бюджетное учреждение науки Институт систем информатики им. А.П. Ершова Сибирского отделения Российской академии наук (ИСИ СО РАН)

Официальные оппоненты:
Жижимов Олег Львович
доктор технических наук, ведущий научный сотрудник лаборатории информационных ресурсов, Федеральное государственное бюджетное учреждение науки Институт вычислительных технологий Сибирского отделения Российской академии наук (ИВТ СО РАН)

Саломатина Наталья Васильевна
кандидат физико-математических наук, старший научный сотрудник, Федеральное государственное бюджетное учреждение науки Институт математики им. С.Л. Соболева Сибирского отделения Российской академии наук
(ИМ СО РАН)

Ведущая организация:
Федеральное государственное бюджетное образовательное учреждение высшего образования «Алтайский государственный технический университет им. И.И. Ползунова» (АлтГТУ)

Защита состоится «27» февраля 2020 года в ___ ч. ___ 00 мин. на заседании диссертационного совета Д 219.005.02 в Федеральном государственном бюджетном образовательном учреждении высшего образования «Сибирский государственный университет телекоммуникаций и информатики» (СибГУТИ) по адресу: 630102, г. Новосибирск, ул. Кирова, 86 ауд. 625.

С диссертацией можно ознакомиться в библиотеке СибГУТИ и на сайте: http://www.sibsutis.ru/science/postgraduate/dis_sovets/.

Автореферат разослан « » 20 г.

Ученый секретарь
диссертационного совета,
к.т.н., доцент  И.В. Нечта

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы. Проблема информационного поиска является актуальной. Прежде всего, это обусловлено колоссальным объемом информационных ресурсов. В настоящее время поиск и анализ текстовой информации является важной задачей в области компьютерной лингвистики. В условиях стремительного роста объемов информационных ресурсов возникает необходимость повышения качества информационного поиска и методов обработки текстов на естественном языке в информационно-поисковых системах (ИПС). Это, в свою очередь, приводит к необходимости совершенствования алгоритмов поиска и ранжирования документов, так, чтобы они были способны учитывать семантику поступающих запросов.

Многие исследователи склоняются к необходимости проведения глубокого семантического анализа текстов для создания их семантических образов, на основе которых можно было бы проводить тонкое ранжирование документов. Этот подход, несомненно, наиболее разумный, однако требует тщательной и долгой работы над созданием подходящих инструментов для автоматической обработки текстов. В частности, может потребоваться детальное описание различных областей знаний. Поэтому имеет смысл также поиск частичных решений, одно из которых представлено в данной работе.

Основная цель – построение алгоритмов, которые, проникая в структуру текста, могли бы получить правильную оценку соответствия (адекватности) текста поисковому запросу, исходя из контекста поискового запроса, не ограничиваясь ключевыми словами, близостью, или частотой.

Следующим актуальным направлением исследований в области информационного поиска являются задачи автоматической тематической фиксации анализируемого текста и резюмирования. Чаще всего для оценки информативности различных элементов текста используют статистический подход, основанный на частотных характеристиках слов или словосочетаний. В результате пользователь получает список наиболее значимых предложений исходного текста, и вес предложения определяется как сумма частот, входящих в него значимых слов или количеством связей между данным предложением и предложениями, находящимися слева и справа от него. Известны также позиционные методы, в которых информативность предложения зависит от его положения в тексте и индикаторные методы, основанные на функциональной идентификации фраз первичного документа с помощью индексации их специальными словами, называемыми маркерами или индикаторами.

Поиск по аннотациям значительно быстрее и упрощает определение релевантности текста поисковому запросу, чем при извлечении необходимых сведений из полных текстов. Однако большинство разработок носят

экспериментальный характер, многие из них недоступны, и многие существующие системы не поддерживают агглютинативные языки при формировании аннотаций (IBM Text Miner, Oracle Text, TextAnalyst).

В рамках данной работы обсуждается задача оценки качества в предположении, что содержание реферата зависит от предпочтений пользователя и регулируется с помощью запроса. Исследуется существующий интересный алгоритм, предложенный Нираджем Кумаром, учитывающий такой фактор, как порядок слов, а не просто их близость, а также применяющий весьма интересную методику определения соответствия двух текстов с использованием понятия центральности по близости. Предлагается обобщение алгоритма с использованием знаний о грамматических структурах. На макроуровне алгоритм может быть представлен следующим образом.

1. Проводится предварительная обработка текста, могут удаляться отдельные элементы, специальные обозначения, не поддерживаемые символы.
2. Специальным образом по текстам формируются графы.
3. Вычисляются веса слов на основе частотных характеристик или с учетом грамматической структуры текста.
4. Предполагается, что отдельные абзацы текста могут отражать информацию по различным темам. Вычисляются веса абзацев (условно можно сказать, тематических фрагментов), исходя из весов входящих в них слов.
5. Вычисляется оценка релевантности абзацев в тексте (потенциальных фрагментов реферата) и текстов, являющихся тематическими эталонами с учетом грамматической структуры и с применением понятия центральности по близости.
6. Вычисляется окончательная оценка на основе полученных ранее оценок релевантности и весов тем.

При решении технической проблемы разработки систем автоматического реферирования необходимо в первую очередь решить научную проблему, связанную с разработкой метода формирования контента реферата, адекватно отражающего смысл текста. К настоящему времени большинство предлагаемых систем автоматического реферирования используют метод составления выдержек, т.е. выделяют и выбирают оригинальные фрагменты из исходного документа и соединяют их в короткий текст.

Такие программы могут использоваться различными организациями и отдельными пользователями, которые регулярно ищут в сети различного типа информацию: научную, технологическую, политическую и социально-экономическую, военную и т.д.

Резюмируя, можно сказать, что актуальность темы обусловлена: необходимостью разработки новых и совершенствования имеющихся алгоритмов поиска и ранжирования документов, способных учитывать семантику

поступающих запросов; наличием научных проблем, связанных с поиском и анализом текстовой информации; вариативностью лексики, омонимия и синтаксическая синонимия (перефразирование); необходимостью разработки быстрых алгоритмов поиска и анализа, которые могут применяться для больших текстовых коллекций; тем что, алгоритмы информационного поиска и анализа часто скрываются разработчиками.

Степень проработанности темы. Теоретической основой послужили научные работы, содержащие исследования по агглютинативным языкам, грамматикам связей, синтаксическим анализаторам текстов на естественном языке, методам сравнения предложений и определения тем тексов, алгоритмам на графах и математической логике.

Наиболее важными из них являются работы следующих авторов: D. Sleator, D. Temperley, L. Vepstas, N. Kumar, Lotfi Zadeh, J. Lafferty, G. Salton, Н.В. Лукашевич, Н.Н. Леонтьева, Г.С. Осипов, И.В. Соченков, В.Ф. Хорошевский, И.В. Ефименко, Özlem İstek, Eşref Adalı и ряда других.

В Казахстане прикладные возможности морфологических и синтаксических анализаторов применительно к тюркским языкам в системах машинного перевода и проблемы семантического анализа при автоматической обработке исследовались в работах А.А. Шарипбаева, Г.Т. Бекмановой, Т.Г. Балова, У.А. Тукеева, Ж. Жуманова, Д. Рахимовой, Е.Н. Амиргалиева, О.Ж. Мамырбаева, Р.Р. Мусабаева и ряда других исследователей.

Цель и задачи исследования. Основная цель диссертации – разработка нового лингвистического и алгоритмического обеспечения технологий информационного поиска и анализа текстовой информации с учетом синтаксиса и элементов семантики, в том числе, для тюркоязычных текстов.

Более конкретно, цель состоит в том, чтобы разрабатываемые методы могли позволять информационно поисковой системе сопоставлять конструкции естественного языка и в ряде случаев отождествлять даже перефразированные варианты предложений, основываясь на анализе их синтаксических структур.

Таким образом, можно сопоставить поисковый запрос и текст, взятый из сети Интернет или других источников, для определения релевантности (соответствия) текста поисковому запросу.

Второй аспект цели состоит в разработке методов, позволяющих определять темы текстов.

В диссертации предполагается, что алгоритмы основываются на использовании диаграмм связей, создаваемых программным приложением Link Grammar Parser (далее, LGP).

В соответствии с поставленной целью в диссертационной работе решаются следующие **задачи.**

1. Исследование методов повышения качества информационного поиска на основе грамматики связей, в том числе с учетом перефразирований предложений.

2. Разработка системы связей (морфологических и синтаксических) для тюркских языков, и реализация прототипов программной системы Link Grammar Parser для казахского и турецкого языков.

3. Анализ моделей определения тем текстов на естественном языке, обобщение алгоритма Нираджа Кумара с использованием диаграмм Link Grammar Parser.

4. Реализация программного инструментария для анализа текстов: алгоритмы построения графов по предложениям, определение степени близости предложений, подсчет различных характеристик.

Соответствие диссертации паспорту специальности. Диссертация соответствует области исследований специальности 05.13.17 – Теоретические основы информатики: пункт 5 «Разработка и исследование моделей и алгоритмов анализа данных, обнаружения закономерностей в данных и их извлечениях, разработка и исследование методов и алгоритмов анализа текста, устной речи и изображений»; пункт 6 «Разработка методов, языков и моделей человеко-машинного общения; разработка методов и моделей распознавания, понимания и синтеза речи, принципов и методов извлечения данных из текстов на естественном языке»; пункт 12 «Разработка математических, логических, семиотических и лингвистических моделей и методов взаимодействия информационных процессов, в том числе на базе специализированных вычислительных систем».

Методы исследования

В основном, применялись методы, относящиеся к информационным технологиям и используемые при обработке текстов на естественном языке, а также методы из теории графов и математической логики. В диссертации привлечен довольно обширный материал из классической и математической лингвистики.

Научная новизна работы заключается в следующем:

- На основе грамматики связей разработаны алгоритмы сопоставления предложений с целью определения их похожести с учетом перефразирований. Рассмотрено обобщение алгоритма Нираджа Кумара определения тем текстов.
- Предложена система связей (морфологических и синтаксических) для тюркских языков. Прототипы программной системы Link Grammar Parser для казахского и турецкого языков обладают богатым набором связей.
- Созданный инструментарий позволяет проводить широкомасштабное тестирование и совершенствование алгоритмов информационного поиска на

естественном языке, в том числе на казахском и турецком языках, дающих высокую степень адекватности результата запросу.

Теоретическая значимость. Предложенные в диссертации способы использования семантико-синтаксических отношений между смысловыми единицами предложения, получаемых на основе диаграмм системы Link Grammar Parser, для вычисления степени близости естественно-языковых конструкций, включая разного рода перефразирования, представляют собой большой теоретический интерес. В частности, это позволяет повысить качество определения релевантности текста поисковому запросу и классификации текстов по темам. Реализован прототип системы Link Grammar для казахского и турецкого языков. Перед реализацией была предложена система связей для тюркских языков, что также является значимым теоретическим результатом.

Практическая значимость. Результаты работы могут быть применены в автоматизированных системах акцепции информации из текстов на естественном языке, интеллектуальных системах поиска информации в сети, при построении систем автоматического резюмирования, электронных переводчиках и словарях и в системах безопасности, например, работающих с банковской информацией.

Основные этапы исследования выполнены в рамках проектов и грантов: Грант Министерства образования и науки Республики Казахстан на 2015-2017 гг. № 46 «Разработка информационно-поискового тезауруса (с учетом морфологии казахского языка) в полнотекстовых базах данных по ИТ-технологиям»; Грант Министерства образования и науки Республики Казахстан на 2018-2020 гг. № AP05133550 «Модели и методы семантического анализа и представления смысла текста в компьютерной лингвистике»; Грант РФФИ на 2018-2019 гг. «Модели и методы создания информационных систем поддержки научных исследований, интегрированных в открытое семантическое пространство» (№ 18-07-01457-А); Интеграционный проект СО РАН на 2018-2020 гг. (№ АААА-А18-118022190008-8).

Положения, выносимые на защиту. На защиту выносятся следующие новые научные результаты:

1. Методы повышения качества информационного поиска на основе грамматики связей.
2. Система связей для тюркских языков и прототип программной системы Link Grammar Parser для казахского и турецкого языков.
3. Обобщенный алгоритм Нираджа Кумара, дополнительно использующий систему Link Grammar Parser.
4. Специализированный программный инструментарий для анализа текстов на естественном языке.

Степень достоверности результатов. Достоверность результатов подтверждена строгой математической формализацией основных положений

диссертационного исследования и результатами экспериментальных исследований на основе разработанных программных средств, реализующих предложенные методы, структуры данных и алгоритмы.

Апробация результатов исследования. Основные результаты диссертации докладывались автором на 20 научных конференциях, среди них: Вторая Российско-Тихоокеанская Научная Конференция по Компьютерным Технологиям и Приложениям (RPC-2017); Международная конференция «Computational and Information Technologies in Science, Engineering and Education» (CITech-2018); Russian summer school in information retrieval (RuSSIR-2018); 12-я Международная Ершовская конференция по информатике. Рабочий семинар «Наукоемкое программное обеспечение» (PSI-19).

Основные результаты диссертации докладывались и обсуждались также на научных семинарах в Институте систем информатики им. А.П. Ершова СО РАН, Новосибирском государственном университете, Сибирском государственном университете телекоммуникаций и информатики, Стамбульском техническом университете (Стамбул), Евразийском национальном университете им. Л.Н. Гумилева (Астана), Институте информационных и вычислительных технологий КН МОН РК (Алматы), Казахском национальном университете им. аль-Фараби (Алматы), Казахстанско-Британском техническом университете (Алматы), Карагандинском государственном техническом университете (Караганда).

Публикации соискателя по теме диссертации. По материалам диссертации опубликованы более чем 35 научных работ, из них: 1 монография, 4 работы в изданиях рекомендуемых ВАК, 6 работ индексируемых в WoS и/или Scopus. Получено 1 свидетельство о регистрации программного обеспечения.

Личный вклад. Содержание диссертации и основные положения, выносимые на защиту, отражают персональный вклад автора в опубликованные работы. В основном работа выполнялась совместно с научным руководителем. Наибольший вклад автором диссертации внесен в разработку алгоритмов для анализа текстов на тюркских языках и в создание программного обеспечения в целом.

Объем и структура работы. Диссертация состоит из введения, четырех глав, заключения и восьми приложений. Полный объем диссертации составляет 122 страницы, включая 18 рисунков и 16 таблиц. Список литературы содержит 122 наименования.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении обосновывается актуальность исследований, проводимых в рамках данной диссертационной работы, приводится обзор научной литературы по изучаемой проблеме, формулируется цель, ставятся задачи работы, излагается научная новизна и практическая значимость представляемой работы.

Первая глава посвящена алгоритмам информационного поиска.

Основная задача – сопоставление предложений и анализ их на похожесть. Предполагаем, что L – множество слов некоторого естественного языка. Для любого слова $x \in L$ обозначим $Norm(x)$ его нормализованную форму. Запись $Syn(x, y)$ обозначает, что x, y – синонимы.

Возникают два вида эквивалентностей:

1) $x_1 \approx x_2 \leftrightarrow x_1 = x_2 \vee Syn(x_1, x_2)$;

2) $x_1 \equiv x_2 \leftrightarrow Norm(x_1) = Norm(x_2)$.

Предложение рассматриваем, как вектор с компонентами из слов $\bar{x} =< x_1, \ldots, x_n >$. Функция $Norm$ может быть естественно распространена на предложения $Norm(\bar{x}) =< Norm(x_1), \ldots, Norm(x_n) >$. Текст $T =< \bar{x}_1, \ldots, \bar{x}_n >$ есть последовательность предложений.

Пусть запись $\bar{x} \models P(x_i, x_j)$ обозначает, что в схеме разбора предложения $\bar{x} =< x_1, \ldots, x_n >$ посредством анализатора Link Grammar Parser имеется коннектор типа P, идущий от слова x_i к слову x_j. Знак $\models$ означает, что фактически мы имеем дело с моделью. Основным множеством модели является множество пар $\{< 1, x_1 >, \ldots, < n, x_n >\}$. Так как одно и то же слово может входить в предложение два и более раз, то это приводит к необходимости рассмотрения именно пар, а не отдельных слов. В силу сказанного выше корректным является даже обозначение $\bar{x} \models \varphi$, где φ – формула, например исчисления предикатов первого порядка. Фактически $\bar{x}$ одновременно является обозначением и для вектора, и для модели.

Предположим, что даны два предложения $\bar{x} =< x_1, \ldots, x_n >$, $\bar{y} =< y_1, \ldots, y_m >$. Интерес представляют функции f такие, что

$$dom(f) \subseteq \{1, \ldots, n\},\ range(f) \subseteq \{1, \ldots, m\}$$

с дополнительными свойствами типа: $f(i) = j \rightarrow x_i \approx y_j$, $f(i) = j \rightarrow x_i \equiv y_j$ и другие подобные им.

При сопоставлении двух предложений, точнее при анализе их на близость, осуществляется проверка ряда логических свойств. Например, пусть $f(i_1) = j_1, f(i_2) = j_2$. Теперь приведены примеры такого рода свойств.

1. Инвариантность коннектора
$\bar{x} \models P(x_{i_1}, x_{i_2}) \rightarrow \bar{y} \models P(y_{j_1}, y_{j_2})$.
2. Замена коннектора на дизъюнкцию других
$\bar{x} \models P(x_{i_1}, x_{i_2}) \rightarrow \bar{y} \models \bigvee_t Q_t(y_{j_1}, y_{j_2})$.
3. Расщепление коннектора на два коннектора
$\bar{x} \models P(x_{i_1}, x_{i_2}) \rightarrow \exists k\ (\bar{y} \models Q(y_{j_1}, y_k) \wedge R(y_k, y_{j_2}))$.
4. Расщепление коннектора на два коннектора с инверсией
$\bar{x} \models P(x_{i_1}, x_{i_2}) \rightarrow \exists k\ (\bar{y} \models Q(y_{j_2}, y_k) \wedge R(y_k, y_{j_1}))$.

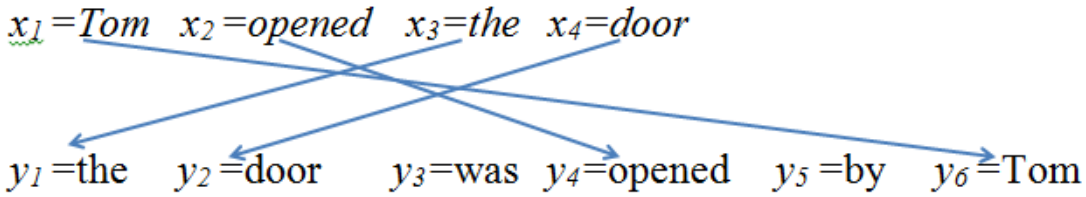
Принимая во внимание, что $\bar{y}$ является обозначением для соответствующей модели, формула из третьего пункта может быть переписана в виде $\bar{x} \models P(x_{i_1}, x_{i_2}) \rightarrow \bar{y} \models \exists y Q(y_{j_1}, y) \wedge R(y, y_{j_2})$. По аналогии может быть записана формула из четвертого пункта.

Рассмотрим пример анализа двух предложений, одно из которых является перефразированным вариантом другого. Ниже показан результат работы синтаксического анализатора Link Grammar Parser:

```
      +------>WV------>+-----Os-----+
      +---Wd--+---Ss---+      +Ds**c+
      |       |        |      |     |
      LEFT-WALL Tom.m opened.v-d the door.n

     +-------------->WV------------>+
     +-----Wd-----+                 |
     |      +Ds**c+--Ss--+----Pv---+--MVp-+-Js+
     |      |     |      |         |      |   |
LEFT-WALL the door.n was.v-d opened.v-d by Tom.m
```

Действие функции f можно видеть ниже:



Таким образом, имеем $f(1)=6, f(2)=4, f(3)=1, f(4)=2$.
При этом отображении получаем:
1) $Norm(opened) = Norm(opened)$ или, что то же самое $opened \equiv opened$;
2) коннектор Ds**c сохраняются, т.е. они инвариантны;

3) $\bar{x} \models \mathrm{Ss}(Tom, opened) \rightarrow \bar{y} \models \mathrm{MVp}(opened, by) \wedge \mathrm{Js}(by, Tom)$, т.е. имеет место расщепление коннектора Ss с инверсией;

4) $\bar{x} \models \mathrm{Os}(opened, door) \rightarrow \bar{y} \models \mathrm{Ss}(door, was) \wedge \mathrm{Pv}(was, opened)$, аналогично имеет место расщепление с инверсией, но другого коннектора Os.

Резюмируя можно сказать, что в нашем распоряжении имеются правила вида

$$R_i : \bar{x} \models \varphi_i(x_1, x_2) \rightarrow \bar{y} \models \psi_i(y_1, y_2).$$

Далее строится функция f, и проводится анализ, встречаются ли индексы $i_1, i_2, j_1 = f(i_1), \; j_2 = f(i_2)$ такие, что на конкретных словах из предложений $\bar{x}, \bar{y}$ выполнено правило R_i, т.е. $\bar{x} \models \varphi_i(x_{i_1}, x_{i_2}) \rightarrow \bar{y} \models \psi_i(y_{j_1}, y_{j_2})$. Для простоты можно говорить, что правило выполняется на паре $< i_1, i_2 >$.

Рассмотрим множество всех таких пар $< i_1, i_2 >$, на которых выполнено одно из правил. Обозначим это множество I, и пусть его мощность $|I| = n$. Отметим, что анализатор Link Grammar Parser допускает между двумя словами наличие только одного коннектора. Поэтому будет выполняться не более, чем одно правило.

Далее пусть n_1, n_2 – количество коннекторов, получающихся в результате анализа предложений $\bar{x}, \bar{y}$ соответственно. В качестве меры похожести двух предложений можно ввести $\mu_0(\bar{x}, \bar{y}) = n / \max(n_1, n_2)$ или $\mu_1(\bar{x}, \bar{y}) = 2n / (n_1 + n_2)$.

В конце сравнения, как обычно, используется порог, в нашем случае порог был взят равным 0,5.

Во второй главе речь идет о системе связей для тюркских языков и реализации прототипов программной системы Link Grammar Parser для казахского и турецкого языков.

Этапы разработки

1. Определение начальных форм слов, построение всех основных форм слов по начальной форме.
2. Построение морфологических связей для казахского и турецкого языков.
2.1. Выделение словоизменительных аффиксов.
2.2. Выделение личных и временных аффиксов.
2.3. Построение набора морфологических связей в словаре.
3. Построение синтаксических связей для казахского и турецкого языков.
3.1. Выделение аффиксов времен глаголов.
3.2. Построение синтаксических связей между словами и группами слов.

Для того чтобы построить набор морфологических связей было рассмотрено большое количество сочетаний аффиксов. В итоге в системе используется более 1100 морфологических связей. Синтаксических связей предложено 18, и они являются универсальными для всех тюркских языков. Предложенные морфологические связи легко адаптировать для любого тюркского языка, т.к.

наборы аффиксов в них практически одинаковые. Ниже приведены некоторые примеры коннекторов и спецификаций.

Аффиксы времен глаголов казахского языка

Вид времен глагола	Аффиксы	Связь
Субъективное прошедшее время	ыпты, іпті	{Vas+}
Результативное прошедшее время	қан, ған, кен, ген	{Var+}
Категорическое прошедшее время	ты, ті, ды, ді	{Vac+}
Конкретное настоящее время	п, ып, іп, а, е	{Vr+}
Переходное будущее время	ады, еді	{Vft+}
Предположительное будущее время	ар, ер	{Vfs+}
Целенаправленное будущее время	мақ, мек, пақ, пек	{Vfg+}

Аффиксы времен глаголов турецкого языка

Вид времен глагола	Аффиксы	Связь
Настоящее продолженное время	yor, iyor, ıyor, uyor, üyor, öyor	{Vpc+}
Субъективное прошедшее время	mış, miş, muş, müş	{Vas+}
Категорическое прошедшее время	dı, di, du, dü, tı, ti, tu, tü	{Vac+}
Категорическое будущее время	acak, ecek, aca, ece	{Vfd+}
Настоящее будущее время	r, ır, ir, ur, ür, ar, er	{Vfp+}

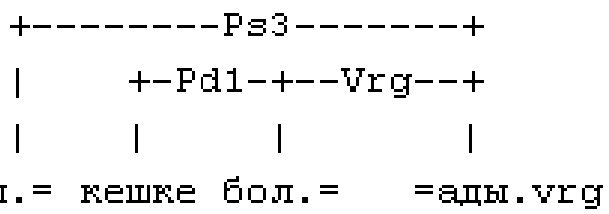
Синтаксические связи для казахского и турецкого языков

AS	Определение при подлежащем
AO	Определение при дополнении
E	Обстоятельство при сказуемом
J	Соединяет послелог с глаголом
OV	Прямое дополнение при сказуемом
OJV	Косвенное дополнение при сказуемом
S	Соединяет подлежащее и сказуемое

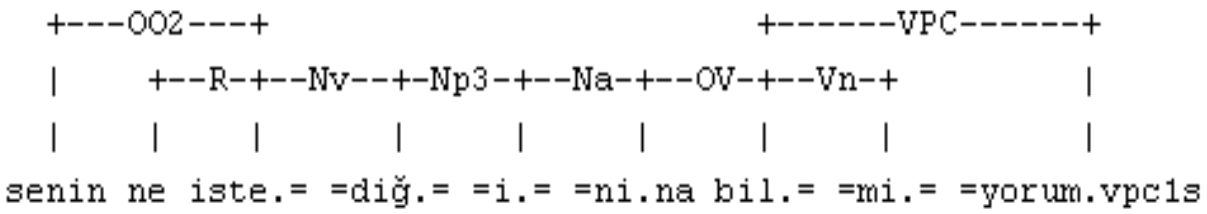
Примеры спецификаций в словаре LGP для турецкого языка

Спецификация в LGP	Значение спецификации
<N_S>: {AS–} & {OV+} & S+	Имя существительное в предложении может выступать в роли подлежащего, к которому относятся определение и/или дополнение, сказуемое всегда будет справа
<N_O>: {AO–} & {OV+} & {OJV+}	Существительное может выполнять функцию дополнения, слева от которого также может быть определение, а справа может находиться послелог и сказуемое
<V_P>: {EI–} & {OV–} & {OJV–} & {S–}	Глагол может выступать в предложении в качестве сказуемого, слева от которого может быть подлежащее, дополнение (прямое или косвенное) или обстоятельство.

Примеры синтаксического разбора предложений на казахском и турецком языках



Предложение, содержащее глагол с аффиксом целенаправленного будущего времени



Предложение, содержащее глагол с аффиксом настоящего продолженного времени

Третья глава посвящена тематическому анализу текстов. Исследования базируются на подходе Нираджа Кумара.

Структура алгоритма.

1. Дан текст и заранее подготовленный образец текста (эталон) по фиксированной тематике, эталон называется темой.
2. Текст разбивается на блоки, каждый из которых состоит из набора предложений взятых из текста не обязательно следующих друг за другом.

3. Методы разбиения текста на блоки могут быть различными, простейший метод – это разбиение по абзацам, другой вариант – с помощью агломеративных процедур, учитывающих лексику.
4. Блоку из текста и эталону сопоставляются определенным образом графы. Вершины графа соответствуют словам. Ребра имеют пометки, в которых «зашито» довольно много информации.
5. Специальным образом выделяются значимые слова из фрагмента текстов и из эталонов с использованием, так называемого метода ссылочного ранжирования.
6. Граф, построенный по фрагменту текста (блоку) сравнивается с графом, построенным по образцу и вычисляются оценки их схожести путем сравнения соответствующих им так называемых центральностей по близости для значимых слов.

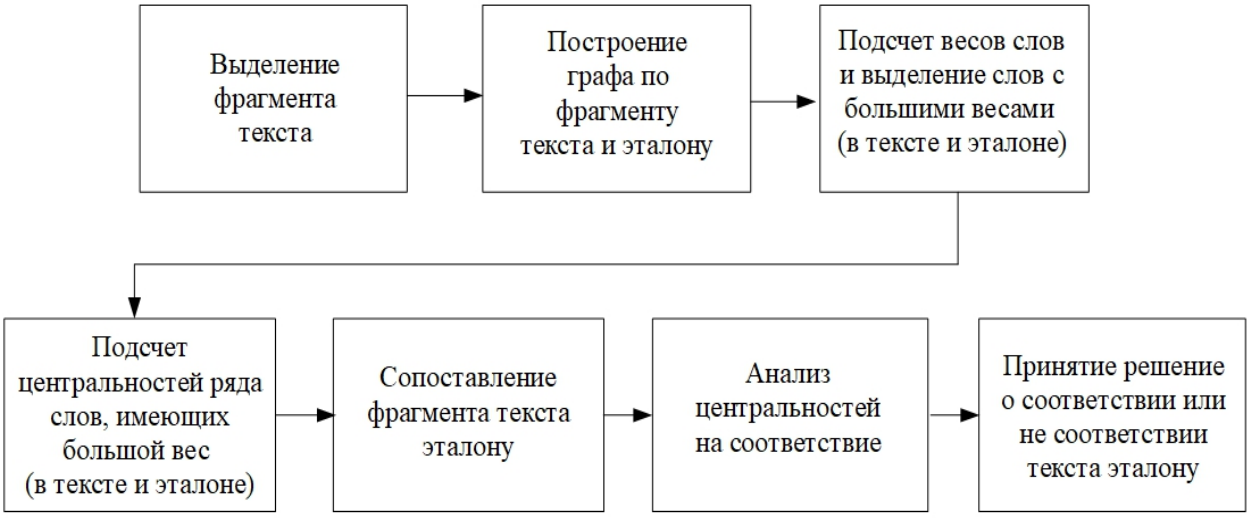


Рис. 1. Структурная схема алгоритма; блок определения тем

Ниже приведены формулы для рекурсивного вычисления весов вершин.

Для каждой вершины V_i обозначим $IN(V_i)$ – множество вершин, которые ссылаются на нее (*предшественники*), а $OUT(V_i)$ – множество вершин, на которые ссылается, сама (*потомки*). Тогда ранг каждой вершины графа (вес каждого слова) можем определить по формуле:

$$S(V_i) = \frac{1-\lambda}{N} + \lambda \sum_{V_i \in IN(V_i)} \frac{S(V_j)}{\left|OUT(V_j)\right|}$$

где, $S(V_i)$ – ранг вершины (вес слова) V_i;

$S(V_j)$ – ранг вершины V_j, из которой связь направлена в вершину V_i;

$OUT(V_j)$ – количество потомков вершины V_j;

N – количество вершин графа;

λ – коэффициент затухания (damping factor), используется фиксированная величина, равная «0,85».

Таким образом, в целом тексте выделяются наиболее важные, можно назвать, ключевые слова.

В четвертой главе описана программная реализация наиболее важных алгоритмов и их тестирование.

Программный продукт состоит из двух частей: набора функциональных исполняемых файлов и отдельного визуального модуля.

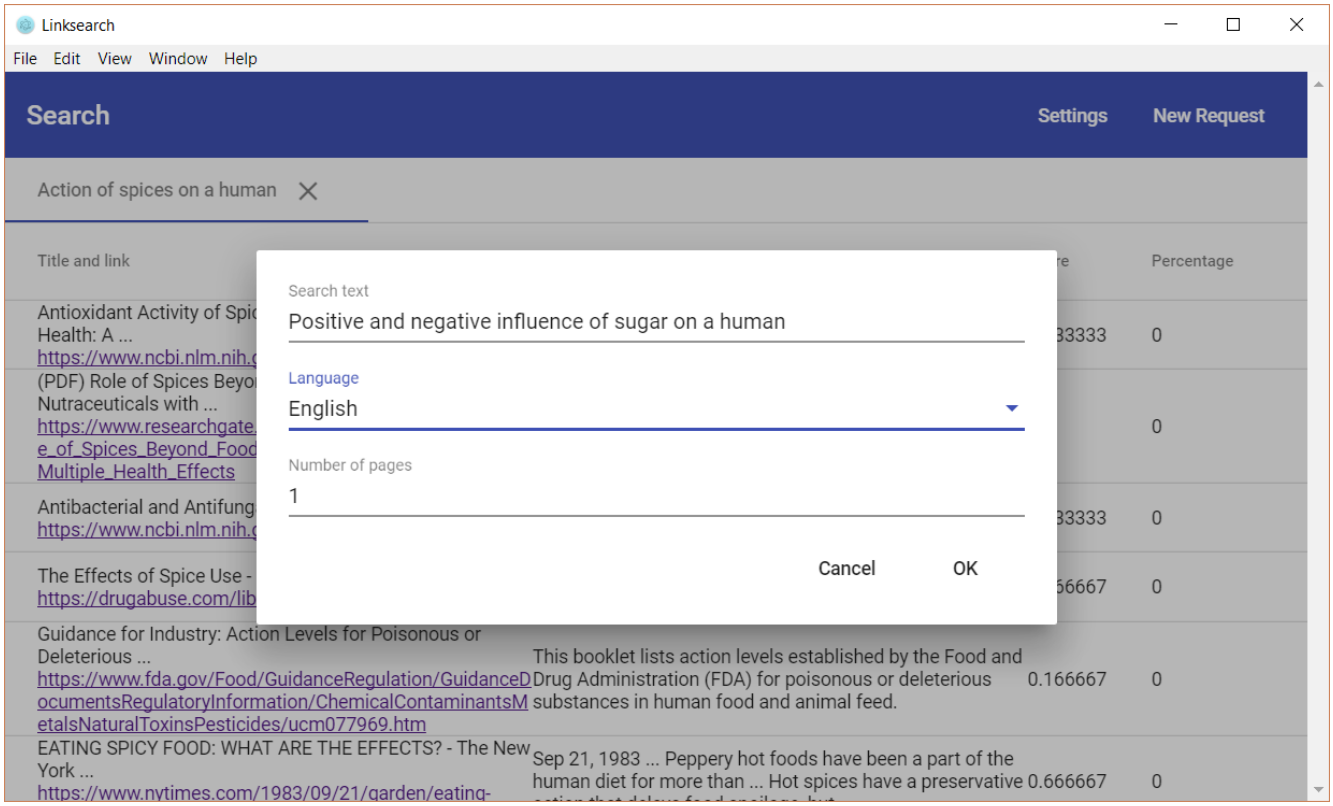


Рис. 2. Главное окно программы

В типовом варианте интерфейс программной системы по данному запросу получает от поисковой системы (например, Google) набор ссылок и коротких сниппетов, проводит их предварительную обработку и передает далее на обработку, цель которой произвести фильтрацию данных, полученных из сети.

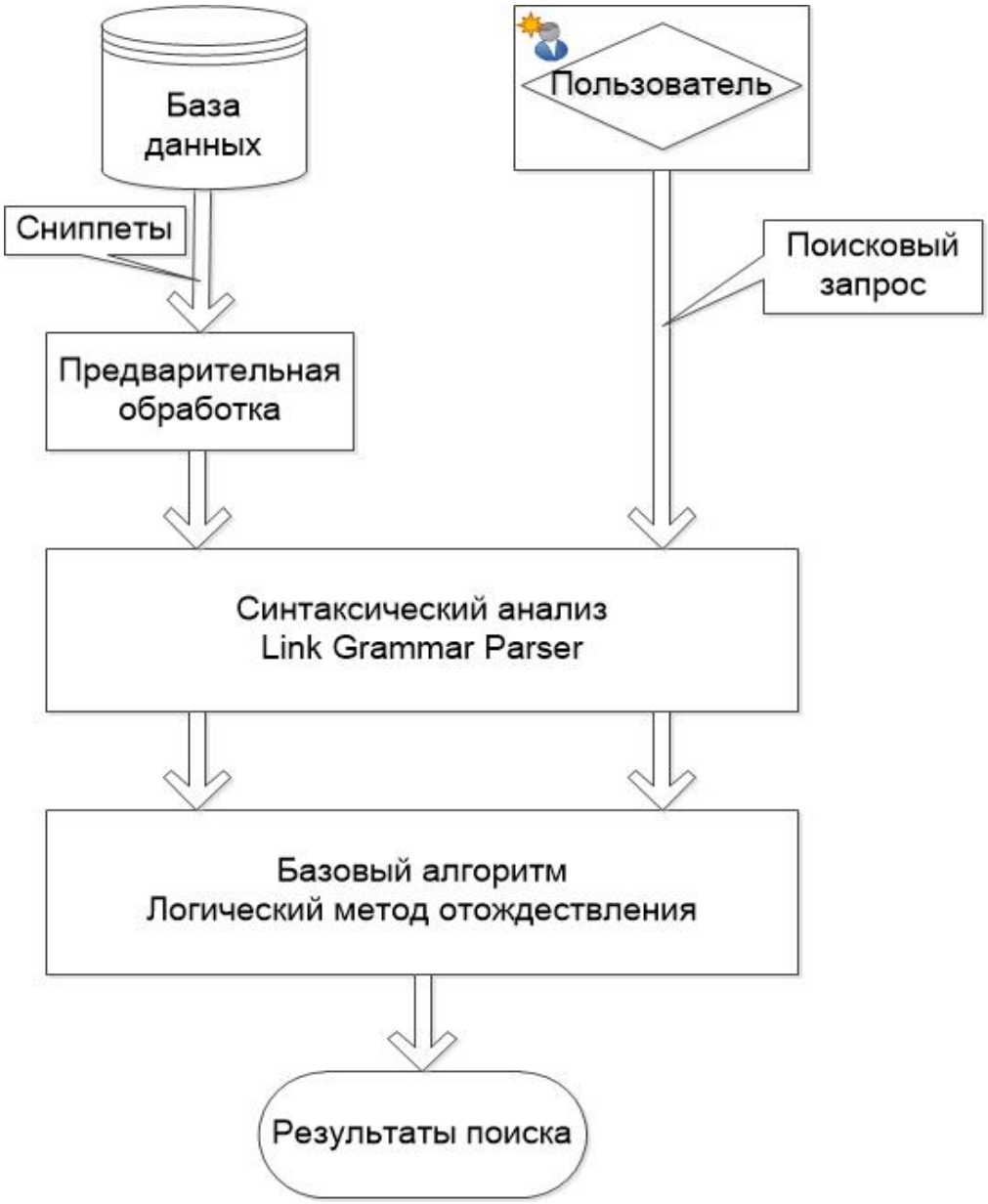


Рис. 3. Схема работы программной системы; блок определения релевантности

Ниже приведен пример тестирования. Запросы относятся к темам, связанным с информационными технологиями.

Сначала опишем основные количественные характеристики. Далее приведены результаты тестирования.

Тестирование для русского языка

n_0 – общее количество ссылок и сниппетов, полученных от Google;

n_1 – количество релевантных ссылок, одобренных системой;

n_2 – количество релевантных ссылок, пропущенных системой;

n_3 – количество не релевантных ссылок, одобренных системой.

№	Текст тестового запроса	n_0	n_1	n_2	n_3
1.	Основные понятия теории информации Шеннона и понятие энтропии	140	20	0	0
2.	Коды, исправляющие ошибки, их связь с избыточностью и энтропией	100	14	3	0
3.	Статистические модели источников сообщений	160	0	0	0
4.	Приложения статистики в лингвистике, исследовании музыки и в генетике	112	15	0	8
5.	Элементы криптографии, коды открытого ключа и метод эллиптических кривых	120	19	0	2
6.	Алгоритмические основы систем символьных преобразований, сличение с образцом	140	16	0	9
7.	Источники и типы изображений, классификация алгоритмов обработки изображений	110	16	0	7
8.	Представление изображений в компьютерных системах	160	20	0	0
9.	Технические средства ввода изображений, фото-приемные матрицы и линейки на основе приборов с зарядовой связью	93	18	1	1
10.	Форматы изображений и цветовые пространства	195	20	0	0
11.	Меры близости изображений, анализ перепадов яркости и гистограмм	90	20	0	0
12.	Поиск объектов на изображениях и области применения алгоритмов поиска в робототехнике и в системах безопасности	90	20	0	0

Выводы по результатам тестирования

1. По каждому запросу загружались списки адресов с их описанием, которые поисковики обычно выдают пользователю. По этим коротким описаниям (сниппетам; англ. snippet) производилась оценка ресурса.
2. Для сравнения с поисковой системой (а именно с системой Google) была составлена статистика. Система оставляла релевантные ссылки по ее мнению, отбрасывая нерелевантные по ее же мнению.
3. В итоге работы экспертов выяснилось, что на проведенных тестах в среднем из 100-150 ссылок, полученных из поискового сервиса Google, система оставляла 10–20 качественных релевантных ссылок.
4. Небольшое количество нерелевантных ссылок (2–4) система ошибочно принимала за релевантные и примерно такое же количество релевантных ссылок отбрасывала.
5. Это показывает, что система смогла произвести фильтрацию данных на хорошем уровне.

Оценка качества информационного поиска

Точность (precision).

Определяется как отношение числа релевантных документов, найденных ИПС, к общему числу найденных документов:

$$Precision = \frac{|D_{rel} \cap D_{retr}|}{|D_{retr}|}.$$

Полнота (recall).

Отношение числа найденных релевантных документов, к общему числу релевантных документов:

$$Recall = \frac{|D_{rel} \cap D_{retr}|}{|D_{rel}|}.$$

Выпадение (fall-out).

Выпадение характеризует вероятность нахождения нерелевантного ресурса и определяется, как отношение числа найденных нерелевантных документов к общему числу нерелевантных документов:

$$Fall\text{-}out = \frac{|D_{nrel} \cap D_{retr}|}{|D_{nretr}|}.$$

В действительности, для оценки качества информационного поиска используются и другие характеристики. Можно указать, по крайней мере, 7 таких показателей. В наших примерах в среднем получались следующие значения: *Prec*=0,75; *Rec*=0,89; *Fall-out*=0,25.

В четвертой главе описана также программная реализация процесса тематической классификации.

В заключении подведены итоги, сформулированы основные полученные результаты исследования.

В приложениях представлены результаты тестирования, приводятся таблицы и рисунки.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ

1. Предложены новые методы повышения качества информационного поиска на основе грамматики связей, в том числе с учетом перефразирований предложений. Методы базируются на использовании диаграмм, генерируемых синтаксическим анализатором Link Grammar Parser.

2. Проведен анализ работ по агглютинативным языкам, и как результат, разработана представительная система связей для тюркских языков, на основе которой реализованы прототипы программной системы Link Grammar Parser для казахского и турецкого языков.

3. Исследованы модели определения тем текстов на естественном языке, связанные с ними графы, соответствующие понятия и оценки качества. Основой

послужили работы Нираджа Кумара и др. Рассматривались тексты на русском, английском, казахском и турецком языках.

4. Реализован программный инструментарий для анализа текстов на естественном языке, включающий различные алгоритмы: определения степени близости предложений, построения графов по предложениям, вычисления весов слов, центральностей и других характеристик. Созданный инструментарий позволяет проводить широкомасштабное тестирование и совершенствование алгоритмов.

Отметим, что настоящее исследование имеет большие перспективы развития. Возможны различные вариации рассмотренных в диссертации алгоритмов, и необходимо их дальнейшее исследование и тестирование, чтобы выбрать наилучшие из них.

СПИСОК ПУБЛИКАЦИЙ ПО ТЕМЕ ДИССЕРТАЦИИ

Монография

1. *Батура Т.В., Бакиева А.М.,* ***Еримбетова А.С.,*** *Мурзин Ф.А., Сагнаева С.К.* Грамматика связей, релевантность и определение тем текстов // Институт систем информатики им. А.П. Ершова СО РАН. – Новосибирск: Изд-во СО РАН, 2018. ISBN 978-5-7692-1632-9. – 91 с.

Статьи в журналах из перечня ВАК

2. *Бакиева А.М., Батура Т.В.,* ***Еримбетова А.С.,*** *Митьковская М.В., Семенова Н.А.* Исследование грамматики связей на примере казахского и турецкого языков // *Вестн. Новосиб. гос. ун-та. Серия: Информационные технологии*. 2016. Т. 14, № 3. С. 5–14.
3. *Федотов А.М., Тусупов Д.А., Самбетбаева М.А.,* ***Еримбетова А.С.,*** *Бакиева А.М., Идрисова А.И.* Модель определения нормальной формы слова для казахского языка // *Вестн. Новосиб. гос. ун-та. Серия: Информационные технологии*. 2015. Т. 13, № 1. С. 107–116.
4. *Мурзин Ф.А., Тусупова М.Д.,* ***Еримбетова А.С.*** Filling up Link Grammar Parser dictionaries by using Word2vec techniques // *Совм. вып. по матер. Междунар. конф. «Вычислительные и информационные технологии в науке, технике и образовании», Вестник ВКГТУ им. Д. Серикбаева, Вычислительные технологии.* Усть-Каменогорск – Новосибирск, 2018. Т. 1, 4.3. С. 169–176.
5. *Батура Т.В., Ефимова Л.В.,* ***Еримбетова А.С.,*** *Касекеева А.Б., Мурзин Ф.А.* Временные и пространственные понятия в текстах на естественном языке и их исследование. *Вестник СибГУТИ.* – Новосибирск, 2019. №3, С. 27-35

Статьи в изданиях, индексируемых в Scopus и/или Web of Science

6. ***Yerimbetova A.S.,*** *Murzin F.A., Batura T.V., Sagnayeva S.K., Semich D.F., Bakiyeva A.M.* Estimation of the degree of similarity of sentences in a natural language based on using the Link Grammar Parser program system // *Journal of Theoretical and Applied Information Technology*, 2016. Vol. 86. N. 1. P. 68–77.
7. ***Yerimbetova A.S.,*** *Murzin F.A., Batura T.V., Sagnayeva S.K., Tazhibayeva S.Zh., Bakiyeva A.M. Link Grammar Parser for Turkic Languages and algorithms for* estimation the relevance of documents // *2016 IEEE 10th International Conference on Application of Information and Communication Technologies (AICT-2016)*. 12-14 October 2016, Baku, Azerbaijan. 2016. pp. 104-107.
8. *Batura T.V., Murzin F.A., Semich D.F., Sagnayeva S.K., Tazhibayeva S.Zh., Bakiyev M.N.,* ***Yerimbetova A.S.,*** *Bakiyeva A.M.* Using the Link grammar parser in the study of Turkic languages // *Eurasian journal of mathematical and computer applications*. ISSN: 23066172. Astana: L.N. Gumilyov Eurasian National University, 2016. V. 4. Iss. 2. pp. 14–22.
9. *Fedotov A.M., Tussupov J., Sambetbayeva M.,* ***Yerimbetova A.S.,*** *Idrisova I.* Development and implementation of a morphological model of kazakh language // *Eurasian journal of mathematical and computer applications*. ISSN: 23066172. Astana: L.N. Gumilyov Eurasian National University, 2015. V. 3. Iss. 3. pp. 69–79.
10. *Fedotov A.M., Tusupov J.A., Sambetbayeva M.A., Sagnayeva S.K., Bapanov A.A., Nurgulzhanova A.N.,* ***Yerimbetova A.S.*** Using the thesaurus to develop it inquiry systems // *Journal of Theoretical and Applied Information Technology*, 2016. Vol.86. Iss. 1. P.44-61.
11. ***Yerimbetova A.S.,*** *Sagnayeva S.K., Murzin F.A., Tussupov J.A.* Creation of tools and algorithms for assessing the relevance of documents // *RPC 2018. Proceedings of the 3rd Russian-Pacific Conference on Computer Technology and Applications [8482202] Institute of Electrical and Electronics Engineers* Inc.. https://doi.org/10.1109/RPC.2018.8482202

Прочие публикации по теме диссертации

12. *Batura T.V., Murzin F.A., Bakiyeva A.M.,* ***Yerimbetova A.S.*** The methods of estimation of the degree of similarity of sentences in a natural language based on the link grammar // *Bulletin of the Novosibirsk Computing Center. Series: Computer Science.* 2014. Is. 37. P. 55–69. URL: http://bulletin.iis.nsk.su/files/article/batura_v8.pdf
13. *Batura T.V., Murzin F.A., Semich D.F., Bakiyeva A.M.,* ***Yerimbetova A.S.*** On some graphs connected with texts in a natural language, link grammar and the summarization process // *Bulletin of the Novosibirsk Computing Center. Series: Computer Science.* 2015. Iss. 38. p. 37–49.

14. *Мурзин Ф.А., Батура Т.В.,* ***Ерімбетова А.С.,*** *Бакиева А.М.* Методы определения степени близости предложений на естественном языке на основе грамматики связей // *Наука и мир. Волгоград: Научное обозрение*, 2015. № 3 (19). Т. 2. С. 61–67.
15. ***Ерімбетова А.С.,*** *Ефимова Л.В.* Анализ текстов на естественном языке с помощью синтаксического анализатора Link Grammar Parser и семантической компоненты системы Dialing // *Труды XVI Всероссийской конференции молодых ученых по математическому моделированию и информационным технологиям (YM-2015)*. 2015. Красноярск, Россия, 28-30 октября 2015. С. 71-72.
16. *Batura T.V., Bakiyeva A.M.,* ***Yerimbetova A.S.,*** *Mit'kovskaya M.V., Semenova N.A.* Methods of constructing natural language analyzers based on Link Grammar and rhetorical structure theory // *Bulletin of the Novosibirsk Computing Center. Series: Computer Science.* 2016. Is. 40. pp. 37–51. URL: http://bulletin.iis.nsk.su/files/article/batura_3.pdf
17. *Murzin F.A., Batura T.V., Semich D.F., Sagnayeva S.K., Bakiyeva A.M.,* ***Yerimbetova A.S.,*** *Mit'kovskaya M.V., Semenova N.A.* Research of link grammar for kazakh and turkish languages // *Вестник КазНИТУ. Алматы*, 2016. № 4 (116). С. 684–691.
18. ***Ерімбетова А.С.,*** *Бакиева А.М.* Модели определения релевантности текста и задача реферирования // *Материалы 54-й Международной научной студенческой конференции, МНСК* – 2016, г. Новосибирск, 16 - 20 апреля 2016 г, С. 167.
19. *Бакиева А.М.,* ***Ерімбетова А.С.*** Исследование грамматики связей на примере турецкого и казахского языка // *Материалы 54-й Международной научной студенческой конференции, МНСК* – 2016, г. Новосибирск, 16 - 20 апреля 2016 г, С. 163.
20. *Batura T.V., Murzin F.A., Semich D.F.,* ***Yerimbetova A.S.,*** *Bakiyeva A.M.* Link Grammar Parser and estimation of the document relevance to the search query // *Марчуковские научные чтения 20 - 2017 (MSR 2017). Тезисы. Новосибирск: Омега Принт*, 2017. Новосибирск, 25 июня–14 июля 2017 г. С. 200.
21. *Мурзин Ф.А.,* ***Ерімбетова А.С.,*** *Сагнаева С.К., Батура Т.В., Бакиева А.М., Семич Д.Ф.* Алгоритмы и программные инструменты для определения релевантности текста поисковому запросу и определения тем текстов // *Труды Международной конференции «Актуальные проблемы чистой и прикладной математики». Алматы: ИМиММ*, 2017. Алматы, 22–25 августа 2017 г. С. 141–142.
22. *Мурзин Ф.А.,* ***Ерімбетова А.С.,*** *Батура Т.В., Бакиева А.М., Семич Д.Ф., Ефимова Л.В.* О новых инструментах поиска информации на основе грамматики связей // *Интеллектуальный анализ сигналов, данных и знаний: методы и средства. Сборник статей Всероссийской научно-практической*

конференции с международным участием. Новосибирск: НГТУ, 2017. С. 161–166.
23. *Мурзин Ф.А., Батура Т.В.,* ***Ерімбетова А.С.,*** *Бакиева А.М., Семич Д.Ф., Ефимова Л.В.* О системе поиска информации на основе грамматики связей // *Труды XVI Российской конференции «Распределенные информационно-вычислительные ресурсы. Наука – цифровой экономике» (DICR-2017)*. Новосибирск: ИВТ СО РАН, 4–7 декабря 2017 г., С. 100–114.
24. *Murzin F.A., Sagnaeva S.K.,* ***Yerimbetova A.S.,*** *Sambetbaeva M.A.* Agglutinative languages with a link grammar // *Вестник КазАТК*, 2016. № 2 (97). С. 62–67.
25. *Мурзин Ф.А., Сагнаева С.К.,* ***Ерімбетова А.С.,*** *Дайырбаева Э.Н.* Разработка системы связей для тюркских языков // *Вестник КазАТК* № 3 (102), 2017, С. 102–107
26. ***Ерімбетова А.С.,*** *Абдалиев Б.Ж.* Проблемы построения функциональной модели тюркских языков // *Сборник материалов XII Международной научной конференции студентов и молодых ученых «Наука и образование* - 2017». г. Астана, Казахстан. С. 652–655.
27. ***Ерімбетова А.С.*** Link Grammar Parser и оценка релевантности документа для поискового запроса // *Материалы XVI Всероссийской конференции молодых ученых по математическому моделированию и информационным технологиям. Иркутск: ИДСТУ СО РАН*, 21–25 августа 2017. С.76.
28. ***Ерімбетова А.С.*** Определение тем текстов. *XIII Международная научная конференция студентов, магистрантов и молодых ученых «ЛОМОНОСОВ – 2017»* 14–15 апреля 2017. С.43–44.
29. ***Ерімбетова А.С.,*** *Дайырбаева Э.Н.* Агглютинативті тілдер үшін LINK GRAMMAR PARSER // *Материалы XLI Международной научно-практической конференции КазАТК им. М. Тынышпаева на тему: «Инновационные технологии на транспорте: образование, наука, практика»* (3-4 апреля 2017 г.), том 2. С. 155–159.
30. *Александров К.В.,* ***Ерімбетова А.С.*** Разработка и анализ новых технологий поиска информации // *Материалы 56-й Международнрй научной студенческой конференции МНСК-2018, Информационные технологии, Новосибирск, Россия*. 22-27 апреля 2018 г., 115 стр.
31. *Sambetbayeva M.A.,* ***Yerimbetova A.S.,*** *Daiyrbayeva E.N.* Models and methods of creating information systems integrated into the open semantic space // *The Bulletin of Kazakh Academy of Transport and Communications named after M. Tynyshpayev*, 2018. Vol. 106, No.3. P. 134–140.
32. ***Ерімбетова А.С.,*** *Абдалиев Б.* Қазақ тіліндегі сөздердің қалыптасу формасын анықтау // *Материалы III Международной научно-практической интернет-конференции «Проблемы и перспективы развития современной науки в*

странах Европы и Азии». Сборник научных работ, Переяслав-Хмельницкий, 2018. С.152–155.

33. *Murzin F.,* ***Yerimbetova A.S.,*** *Tussupova M., Aleksandrov K.* Development and analysis of technologies of searching information relevant to the search query using linguistic support // *Proceedings of the XX International Conference «Data Analytics and Management in Data Intensive Domains»*. M., Russia, 2018. P. 207–214.
34. ***Ерімбетова А.С.,*** *Мурзин Ф.А.* Разработка и анализ технологий определения релевантности текста поисковому запросу // *Информационные технологии и системы. Труды Седьмой Всероссийской научной конференции с международным участием. Ханты-Мансийск, Россия*, 2019 г. С. 152-156.
35. ***Ерімбетова А.С.,*** *Батура Т.В., Мурзин Ф.А., Сагнаева С.К., Смаилова У.М., Тишбек Ж.Ж.* Инструментарий для определения семантической близости текстов на агглютинативных языках // *Материалы V Международной научно-практической конференции «Global Science and Innovations 2019: Central Asia»*, г. Астана, 2019. С. 106-109.
36. *Tussupova M., Murzin F.A.,* ***Yerimbetova A.S.,*** *Tazhibayeva S.Zh.* Grammatical categories determination with the use of machine learning // *Materials of the International scientific conference "Theoretical and applied questions of mathematics, mechanics and computer science" dedicated to the 70th anniversary of the doctor of physical and mathematical sciences, Professor Murat Ibraevich Ramazanov, Karagandy*, 2019. – 119 p.
37. *Батура Т.В., Ефимова Л.В.,* ***Ерімбетова А.С.,*** *Касекеева А.Б., Мурзин Ф.А.* Анализ временных и пространственных понятий, встречающихся в текстах на естественном языке // *12-я Международная Ершовская конференция по информатике. Труды семинара «Наукоемкое программное обеспечение»*, 2019, Новосибирск, Россия. С. 53-58

Свидетельство о регистрации программы для ЭВМ

38. ***Ерімбетова А.С.****, Батура Т.В. Мурзин Ф.А., Сагнаева С.К., Бакиева А.М.* Свидетельство о государственной регистрации прав на объект авторского права Министерства Юстиции Республики Казахстан «Қазақ және түрік тілдеріне арналған LINK GRAMMAR PARSER синтаксистік талдағышы» запись в реестре № 743 от 17.04.2017 г. (Синтаксический анализатор LINK GRAMMAR PARSER для казахского и турецкого языков).

Еримбетова А.С.

ЛИНГВИСТИЧЕСКОЕ И АЛГОРИТМИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ПРОЦЕССА ИНФОРМАЦИОННОГО ПОИСКА НА ОСНОВЕ ГРАММАТИКИ СВЯЗЕЙ, В ТОМ ЧИСЛЕ ДЛЯ ТЮРКСКИХ ЯЗЫКОВ

Автореферат

Подписано в печать
Формат бумаги 60 × 90 1/16
Объем 1,5 уч.-изд. л.
Тираж 120 экз.
Отпечатано в ЗАО РИЦ «Прайс-курьер»
630090, г. Новосибирск, ул. Русская, 41, тел. +7(383)373-18-76
Заказ № 124