НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
январь–февраль 2024 Том 24 № 1 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
January–February 2024 Vol. 24 No 1 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

doi: 10.17586/2226-1494-2024-24-1-41-50
УДК 004.021, 004.827

Алгоритм распознавания омографов на основе евклидовой метрики

Элиса Салаудиновна Израилова[1]✉, Арсланбек Виситович Астемиров[2], Айшат Салауддиновна Бадаева[3], Зелимхан Аюбович Султанов[4], Салаудин Мусаевич Умархаджиев[5], Мохмад-Салех Лейчевич Хехаев[6], Мадина Лечаевна Ясаева[7]

[1,2,3,4,5,6,7] Академия наук Чеченской Республики, Грозный, 364043, Российская Федерация
[1,2,3,4,5,6] Комплексный научно-исследовательский институт им. Х.И. Ибрагимова Российской академии наук, Грозный, 364051, Российская Федерация

[1] uelisa@yandex.ru✉, https://orcid.org/0000-0003-1655-4414
[2] astemirow@mail.ru, https://orcid.org/0009-0003-8416-7587
[3] ayshatbs@gmail.com, https://orcid.org/0009-0009-5351-229X
[4] zelimkhan.sultanov@yandex.ru, https://orcid.org/0009-0001-1391-5475
[5] usmu@mail.ru, https://orcid.org/0000-0002-8283-1515
[6] saleh_1991@mail.ru, https://orcid.org/0009-0009-7672-4148
[7] madina.yasaeva@internet.ru, https://orcid.org/0009-0001-3565-4974

Аннотация

Введение. Проблема разрешения неопределенностей, связанных с омонимией, для чеченского языка стала особенно актуальной после создания систем синтеза речи. Основным недостатком синтезаторов речи на чеченском языке являются ошибки чтения слов-омографов, различающихся долготой/краткостью гласных — долгота таких звуков никак не отображается при письме. Вызывает проблемы воспроизведение дифтонгов, которые обозначаются на письме так же, как близкие им по звучанию монофтонги. Для повышения качества синтезируемой речи на чеченском языке необходима программа автоматического распознавания омографов. Для решения этой проблемы рассмотрена задача устранения неоднозначности смысла слов Word Sense Disambiguation (WSD). **Метод.** Для чеченского языка выбраны алгоритмические (управляемые) методы, работающие на основе заранее размеченной базы данных. Эти методы являются наиболее распространенными при устранении неоднозначности смысла слов. Реализация таких методов возможна при наличии больших размеченных корпусов, которые недоступны для большинства языков мира, в том числе и для чеченского. Чеченский язык относится к малоресурсным языкам, для которых оптимальным подходом, с точки зрения экономии трудовых и временных ресурсов, является полууправляемый гибридный метод распознавания омографов, основанный на использовании алгоритмических и статистических методов. **Основные результаты.** Разработан алгоритм распознавания омографов по шести соседним словам в предложении. Алгоритм реализован в виде программы. Предварительная подготовка исходных данных для работы алгоритма включает разметку предложений по значениям омографов, выполняемую «вручную». Результаты работы программы оценены с использованием общепризнанных метрик точности и составили F1 — 39 %, Accuracy — 45 %. **Обсуждение.** Сравнительный анализ полученных данных с результатами других методов и моделей показал, что точность представленного алгоритма наиболее близка к результатам точности алгоритмов на основе метода Леска. По методу Леска для английского языка получены результаты точности F1 — 41,1% (простой Леск) и 51,1 % (Леск расширенный). Методы, использующие нейросетевые алгоритмы, дают более высокие показатели точности по WSD для большинства языков, однако для их реализации требуется наличие больших корпусов данных, что не всегда доступно для малоресурсных языков, в том числе и для чеченского.

Ключевые слова

графическая омонимия, омографы, WSD, синтез речи, чеченский язык, малоресурсные языки, текстовый корпус

Ссылка для цитирования: Израилова Э.С., Астемиров А.В., Бадаева А.С., Султанов З.А., Умархаджиев С.М., Хехаев М.-С.Л., Ясаева М.Л. Алгоритм распознавания омографов на основе евклидовой метрики // Научно-технический вестник информационных технологий, механики и оптики. 2024. Т. 24, № 1. С. 41–50. doi: 10.17586/2226-1494-2024-24-1-41-50

Homograph recognition algorithm based on Euclidean metric

Elisa S. Izrailova[1]✉, Arslanbek V. Astemirov[2], Ayshat S. Badaeva[3], Zelimkhan A. Sultanov[4], Salaudin M. Umarkhadzhiev[5], Mokhmad-Salekh L. Khekhaev[6], Madina L. Yasaeva[7]

[1,2,3,4,5,6,7] Academy of Sciences of the Chechen Republic, Grozny, 364043, Russian Federation
[1,2,3,4,5,6] Kh. Ibragimov Complex Institute of the Russian Academy of Sciences, Grozny, 364051, Russian Federation

[1] uelisa@yandex.ru✉, https://orcid.org/0000-0003-1655-4414
[2] astemirow@mail.ru, https://orcid.org/0009-0003-8416-7587
[3] ayshatbs@gmail.com, https://orcid.org/0009-0009-5351-229X
[4] zelimkhan.sultanov@yandex.ru, https://orcid.org/0009-0001-1391-5475
[5] usmu@mail.ru, https://orcid.org/0000-0002-8283-1515
[6] saleh_1991@mail.ru, https://orcid.org/0009-0009-7672-4148
[7] madina.yasaeva@internet.ru, https://orcid.org/0009-0001-3565-4974

Abstract

The problem of resolving the uncertainties associated with homonymy for the Chechen language has become especially relevant after the creation of speech synthesis systems. The main disadvantage of speech synthesizers in the Chechen language are errors in reading homograph words that differ in the length / brevity of vowels — the longitude of such sounds is not displayed in any way when writing. The reproduction of diphthongs, which are indicated on the letter in the same way as monophthongs close to them in sound, causes problems. To improve the quality of synthesized speech in the Chechen language, an automatic homograph recognition program is needed. To solve this problem, the article considers the task of eliminating the ambiguity of the meaning of the words WSD (Word Sense Disambiguation). Algorithmic (supervised) methods based on a pre-marked database have been selected for the Chechen language. These methods are the most common solutions for eliminating the ambiguity of the meaning of words. The implementation of such methods is possible in the presence of large marked-up corpora that are inaccessible to most languages of the world including Chechen. The Chechen language belongs to low-resource languages for which the optimal approach from the point of view of saving labor and time resources is a semi-controlled hybrid method of homograph recognition based on the use of algorithmic and statistical methods. The algorithm created by the authors for recognizing homographs by six adjacent words in a sentence is presented. The method is implemented as a program. Preliminary preparation of the initial data for the operation of the algorithm includes marking of proposals by the values of homographs performed "manually". The results of the program were evaluated using generally recognized accuracy metrics and amounted to F1 — 39 %, Accuracy — 45 %. A comparative analysis of the data obtained with the results of other methods and models showed that the accuracy of the algorithm presented in this article is closest to the results of the accuracy of algorithms based on the Lesk method. Using Lesk method for English, the results of F1 accuracy were obtained — 41.1 % (simple Lesk) and 51.1 % (extended Lesk). Methods using neural network algorithms provide higher WSD accuracy rates for most languages; however, their implementation requires large data bodies, which is not always available for low-resource languages, including Chechen.

Keywords

graphic homonymy, homographs, WSD, speech synthesis, Chechen language, low resource languages, text corpus

For citation: Izrailova E.S., Astemirov A.V., Badaeva A.S., Sultanov Z.A., Umarkhadzhiev S.M., Khekhaev M.-S.L., Yasaeva M.L. Homograph recognition algorithm based on Euclidean metric. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2024, vol. 24, no. 1, pp. 41–50 (in Russian). doi: 10.17586/2226-1494-2024-24-1-41-50

Введение

Системы искусственного интеллекта, связанные с обработкой естественных языков, такие как системы синтеза и распознавания речи, автоматические переводчики, морфологические анализаторы, поисковые системы, несомненно, должны включать функцию лексического анализа для устранения многозначности и выявления единственно правильного решения в рамках заданной проблемы.

В настоящей работе рассмотрена проблема устранения неоднозначности омографов в текстах на чеченском языке.

Последние несколько лет в отделе прикладной семиотики Академии наук Чеченской Республики ведутся исследования в области речевых технологий, в результате которых получены практические результаты. В частности, создано онлайн-приложение синтеза чеченской речи на основе базы данных Speeche1 [1] и базы чеченских текстов, являющихся обучающим корпусом для нейросетевых речевых систем. На основе данного корпуса была создана система синтеза чеченской речи, состоящая из различных функциональных модулей: модуль для автоматического преобразования орфографической записи текста в транскрипцию[1], нормализатор для расшифровки числительных и аббревиатур[2], модуль обучения, состоящий из двух нейронных сетей; модуль синтезирования речи на основе вокодера.

После создания синтезатора речи были проведены тестирование, MOS-оценка (Mean Opinion Score) и

[1] Свидетельство № 2019664423, 06.11.2019. Российская Федерация. Elp-Az: свидетельство об официальной регистрации программы для ЭВМ / С.М. Умархаджиев, Э.С. Израилова, А.С. Бадаева, М.Х. Бекаев, З.А. Султанов, Х.И. Асхабов, Я.В. Эльсаев, А.А. Абдулкадыров.

[2] Свидетельство № 2020660527, 04.09.2020. Российская Федерация. TerahYoza: свидетельство об официальной регистрации программы для ЭВМ / С.М. Умархаджиев, Э.С. Израилова, А.С. Бадаева, А.В. Астемиров, З.А. Султанов.

анализ качества синтезируемого программой речевого сигнала [2]. В результате анализа был выявлен недостаток синтезатора — программа не всегда правильно читает слова-омографы, различающиеся долготой/краткостью гласных, а также дифтонгами/монофтонгами. Как показывает опыт создания систем синтеза речи при некорректном определении признаков слова-омографа и последующем синтезе такого слова затрудняется восприятие смысла всего предложения. Для чеченского языка эта проблема особенно актуальна ввиду наличия в нем большого количества омографов. Обусловлено это, как правило, наличием некоторых несоответствий графики, используемой для чеченского языка, особенностям его фонетики и орфоэпии. Главным образом, речь идет об отсутствии специальных средств отображения на письме дифтонгов и долготы гласных. Исходя из этого, было решено провести оптимизацию созданной системы синтеза речи, чтобы решить проблему распознавания омографов в текстах на чеченском языке для их правильного прочтения.

Основная трудность при решении проблемы распознавания омографов — малоресурсность чеченского языка, т. е. отсутствие необходимых текстовых обучающих корпусов и баз данных для реализации различных алгоритмов распознавания омографов.

Для чеченского языка проблема устранения многозначности слов никогда до этого не решалась. В основе настоящей работы лежит исследование существующих методов Word Sense Disambiguation (WSD), при этом учтена специфика чеченского языка, ограниченность ресурсов и особенности чеченской орфографии.

Цель работы — описание особенностей устранения многозначности омографов для чеченского языка, рассмотрение существующих методов WSD и процесса подготовки данных, представление разработанного алгоритма распознавания омографов на основе весового евклидова расстояния между предложениями.

Методы устранения лексической многозначности (WSD)

Все методы WSD делятся на три основные группы. Одну из этих групп составляют методы, основанные на знаниях (knowledge-based methods) и использующие электронные словари. Основным методом первой группы является метод Леска — классический алгоритм разрешения лексической многозначности, основанный на знаниях. Метод предложен Майклом Леском в 1986 году. Идея метода заключается в поиске значения слова в толковых словарях и списках словарных определений с учетом контекста, где это слово использовано [3].

Существует множество модификаций метода Леска, одна из них предлагает, кроме использования самих определений слов, непосредственно входящих в фразу, применить концепцию иерархии в базе WordNet [4]. Еще один метод, основанный на концепции Леска, метод семантической схожести (semantic similarity), включающий разные меры схожести. Общей для всех этих методов является идея того, что схожие слова могут определяться контекстом [5]. В работе [6] рассмотрены современные нейронные расширения метода Леска, основанные на определениях для вычисления смысловых вложений, использующихся в качестве обучающих. Хорошие результаты показала модель SyntagRank, в которой применены пары совпадающих слов с устраненной неоднозначностью, включенные в SyntagNet — ресурс лексико-семантических комбинаций для выполнения современного WSD [7]. Модель интересна тем, что может быть масштабирована для множества языков, которые представлены в многоязычном словаре BabelNet, включающем как лексикографические, так и энциклопедические термины.

Вторая группа методов включает методы машинного обучения с учителем, так называемые контролируемые методы (supervised methods), работающие на основе заранее размеченной базы данных. В эту группу входят методы деревьев решений (decision trees) [8–10], метод опорных векторов (Support Vector Machine) [11], нейронные сети [12, 13].

Методы, использующие нейронные сети с учителем, получили в последнее время наибольшее распространение и наилучшие результаты в устранении неоднозначности. Среди них можно выделить несколько наиболее интересных моделей: модели WSD, полностью управляемые данными, обучающиеся с помощью градиентного анализа [14, 15]; контролируемые модели WSD, использующие глоссы — текстовые определения, закодированные в виде векторов [16, 17]; модели, основанные на знаниях и реляционных графах в WordNet [18, 19].

Методы, применяющие нейросетевые алгоритмы, обеспечивают высокие показатели точности по WSD для большинства языков. В работе [20] получен высокий результат для русского языка с помощью нейросетевой архитектуры ELMo (Embeddings from Language Model), основа которой — вложения из языковых моделей. Используя ELMo, вычислены векторы контекста неконтролируемым образом с помощью двух уровней двунаправленного LSTM (Long Short-Term Memory), принимающих встраивания символов из сверточного слоя в качестве входных данных. Результат точности ELMo для русского языка F1 составил 77 %, а с использованием модели BERT (Bidirectional Encoder Representations from Transformers) — 67 % [21]. В [22] для арабского языка при различных модификациях модели BERT (ArBERT, AraBERTv2) на разных корпусах данных получены результаты точности F1 от 74 до 89 %. В архитектуре модели BERT использован многоуровневый двунаправленный преобразователь-кодер и токены WordPiece, т. е. единицы подслова, что помогает избежать проблемы слов, которых нет в словарном запасе. Для татарского языка модель разрешения морфологической многозначности [23], основанная на рекуррентной нейронной сети с долгой краткосрочной памятью (LSTM), показала результат точности F1 — 79 %, а гибридный метод разрешения многозначности на основе скрытых марковских моделей PurePos — 86 %.

Третью группу методов составляют методы, не требующие ни заранее размеченной базы, ни использования словарей. Они направлены на автоматическое извлечение значений слов и предполагают в качестве основы кластеризацию слов и кластеризацию контек-

ста. Главная идея этих подходов состоит в том, что слова с одинаковыми смыслами будут иметь схожие соседние слова [24, 25].

Ученые Санкт-Петербургского Центра речевых технологий для снятия омонимии в системе синтеза русской речи использовали метод лингвистического анализа, включающий поиск в предложении ключевых слов или выражений [26]. Этот подход включает анализ слов непосредственно рядом с текущим омографом, т. е. анализируется грамматическое окружение и выполняется поиск согласованных слов в предложении. Для формализации этого принципа были введены грамматические правила, увеличивающие условный «вес» словоформы в зависимости от ее окружения. Правила хранятся в формализованном виде, позволяющем быстро оценивать и корректировать работу системы.

Проблема устранения неоднозначности омографов для чеченского языка

В компьютерной лингвистике WSD рассмотрена в виде задачи определения правильного смысла слова в контексте. Проблема устранения многозначности слов на сегодняшний день является одной из актуальных проблем обработки большинства естественных языков. При этом одна из задач, которую требуется решить — неоднозначность омографов. Данную проблему возможно устранить либо посредством полноценного семантического анализа, либо с помощью автономного подхода, который выполняется независимо от лексических и морфологических норм языка.

Для решения проблемы устранения многозначности слов необходимо наличие больших текстовых корпусов, различных словарей и баз данных для реализации методов WSD, как алгоритмических моделей, так и методов машинного обучения. Метки, идентифицирующие слова-омографы и контекстно-зависимые сочетания слов, производятся специалистами вручную, что требует больших временных и кадровых ресурсов. В настоящее время для лексического анализа наиболее часто используется ресурс WordNet [27]. WordNet — большая база данных лексических отношений для английского и других языков. В стадии разработки находятся базы для русского языка.

Основные научные исследования в области WSD и созданные на их основе методы также получены для английского и других языков, достаточно обеспеченных электронными ресурсами. В настоящей работе для проведения сравнительного анализа были изучены материалы по проблеме WSD по некоторым малоресурсным языкам (турецкий, татарский, восточноафриканский, хинди и др.). В чеченском языке много заимствований из русского и арабского языков, поэтому были изучены методы устранения многозначности для этих языков.

Основная проблема, с которой столкнулись авторы данной работы — для чеченского языка не созданы базы данных и текстовые корпусы. Для решения проблемы были собраны максимально разнообразные по тематике и жанрам тексты для формирования всевозможных явлений омонимии и контекстов омонимов. В настоящее время собранный банк текстов насчитывает около 15 млн слов и используется как инструмент для языковых исследований, разработки и настройки различных автоматизированных систем, в том числе и для решения проблемы WSD, в частности для подготовки данных, необходимых при реализации методов распознавания омографов и омонимов [28]. Банк текстов постоянно пополняется новыми текстами.

Для чеченского языка вышеприведенные методы, основанные на знаниях, не являются приемлемыми, поскольку чеченский язык по международной классификации относится к малоресурсным языкам. Основные проблемы, возникающие при разработке прикладных программ для чеченского языка — нехватка электронных ресурсов для обработки языка и речи, в том числе одноязычных корпусов, двуязычных электронных словарей, орфоэпических и грамматических словарей, тезаурусов, толковых словарей, словарей омонимов и других различных лингвистических ресурсов [29].

В результате авторами настоящей работы создана новая база данных для реализации WSD: собран текстовый корпус; извлечены тексты с предложениями, содержащими омографы в различных контекстах; выполнена «вручную» разметка текстов по значениям омографов. Для чеченского языка существуют два основных словаря, которые были оцифрованы в отделе прикладной семиотики Академии наук Чеченской Республики: «Русско-чеченский словарь» А.Г. Карасаева, А.Т. Мациева и «Чеченско-русский словарь» А.Т. Мациева, а также несколько отраслевых русско-чеченских словарей. На основе сведений, содержащихся в этих словарях, разработаны алгоритмы и программы, созданы обучающие корпуса и текстовые базы для проверки орфографии и статистического анализа.

Решение проблемы разрешения графической омонимии в текстах на чеченском языке состоит из следующих функциональных задач:

— подготовка текстового корпуса и словаря омографов;
— сбор лексических признаков омографов и их компоновка в словаре;
— разработка программы для автоматизации поиска и сбора предложений с омографами;
— создание базы предложений, содержащих омографы в различных контекстах;
— разработка программы для автоматизации выборки тегов — слов, однозначно идентифицирующих омографы в контексте;
— создание словаря тегов;
— статистический частотный анализ омографов и тегов;
— создание алгоритма и программной реализации программы распознавания омографов;
— распознавание омографов в тексте при создании автоматических транскрипций;
— применение полученных результатов для автоматического синтеза чеченской речи.

Подготовка данных

В процессе изучения проблемы WSD выполнено исследование алгоритмических (управляемых) методов, работающих на основе заранее размеченной базы

данных (вторая группа методов, рассмотренная в разделе «Методы устранения лексической многозначности (WSD)»), так как они наиболее распространены при устранении неоднозначности слов. Однако для реализации таких методов необходимо наличие больших размеченных корпусов, которые недоступны для большинства языков мира. Так как чеченский язык относится к малоресурсным языкам, оптимальным методом разрешения графической омонимии должен быть тот, который использует для реализации минимальные ресурсы. Реализация такого «экономичного» метода возможна благодаря полууправляемому гибридному методу распознавания омографов.

Полууправляемые методы могут решить проблему распознавания омографов, используя большой размеченный корпус, одна часть которого применена для анализа частотности и индексации тегов, вторая часть — для проверки точности распознавания омографов, т. е. для тестирования созданной программы. В процентном соотношении эти части обычно составляют 80/20, причем большая часть текстового корпуса используется для индексации тегов.

Предлагаемый гибридный метод устранения графической омонимии основан на использовании алгоритмических методов и статистики. Данный метод наиболее подходит для разрабатываемого алгоритма, так как при чтении текста читателю приходится анализировать контекст, чтобы правильно произнести слова-омографы. Статистический метод направлен на анализ контекста омографов по следующим признакам: позиция в предложении; наличие служебных слов и аффиксов; контекст слова справа и слева.

В предлагаемом алгоритме рассмотрен только один признак омографа — контекст слова справа и слева. Алгоритм программы распознавания омонимов рассматривает слова в их контексте, что позволяет сократить список возможных слов до очень ограниченного числа высоковероятных слов-тегов, которые позволяют однозначно идентифицировать омограф в контексте. Далее статистика тегов использована при реализации полученного алгоритма.

Для реализации метода статистического анализа были определены и собраны 100 омографов, наиболее часто встречающихся в литературе и устной речи. Для каждого омографа осуществлен поиск в базе текстов на чеченском языке, в ходе которого выбрана база предложений, содержащих данный омограф. Далее файлы были обработаны созданной по алгоритму программой, проводящей анализ статистики слов-тегов, идентифицирующих в контексте омографы. Программа выполнила индексацию тегов в соответствии с частотой их использования в контексте заданного омографа, показывающая количество повторов соответствующих тегов в базе предложений. Полученная статистика позволила выявить наиболее часто встречающиеся в контексте с омографом теги, из которых был сформирован словарь тегов [30].

База предложений является основой для разработки и реализации предложенного алгоритма. Предложения с омографами собраны автоматически из сформированного банка текстов, результатом программы стала электронная таблица Excel, в которой все предложения с омографами находятся в первом столбце. Дальнейшая работа проведена вручную. База предложений с конкретным омографом была разбита на два столбца в зависимости от значения омографа (в основном по долготе гласной буквы). При письме долгота гласных букв никак не обозначается, поэтому и возникла проблема синтезирования слов-омографов в чеченской речи [30].

В табл. 1 приведен фрагмент таблицы из базы предложений для омографа «бала» со значениями «бала» (с кратким звуком «а») и «бАла» (с долгим звуком «а»).

Для исследования было выбрано 13 наиболее частотных омографов, собраны предложения из созданного авторами текстового корпуса, содержащие эти омографы в различных контекстах. Так, для омографа «дан» база содержит 3150 предложений, «де» — 973, «бала» — 482, «белира» — 800, «яха» — 196, «йолу» — 2130, «бен» — 1511, «дала» — 1301, «дакъа» — 490, «шун» — 543, «цIе» — 2000, «ча» — 288, «лар» — 1171. Общее количество предложений составило 15 035.

Алгоритм AWEN распознавания омографов на основе евклидовой метрики

Рассмотрим разработанный алгоритм AWEN (algorithm, weight, Euclid, number of tags) распознавания омографов на основе евклидовой метрики.

Исходные данные и обозначения. Пусть *om* — омограф (слово, имеющее два значения om_i, где $i = 1, 2$).

Таблица 1. Фрагмент таблицы из базы предложений для омографа «бала»
Table 1. Fragment of a table from the database of sentences for the homograph «бала»

Вариант произношения	
бала (краткий звук)	бАла (долгий звук)
Адаман ойланех цуьнан амалех кхетам бала тарлун цхьацца методикаш хуьлу малхбузен литературехь юьйцуш амма чlагlдойла яц цара дерриг а нийса дуьйцу аьлла	Иштта цхьа дагна бАла lаьткъича бен кху тlаьхьарчу шина баттахь цигаьрка ца узуш волчу lалхас шена цигаьрка а хьарчийна боккха баькхина кlур пехашна кlоргге чууьйзира
Уггар хьалха хьомсара государашх хьомсара государынаш хастам бала беза сийлахь-доккхачу оьрсийн муьтlахьаллина	Болх луш верг а болх беш верг а цуьнан бАла кхочуш ца хилча хала хир ду оцу гlуллакха тlехь кхиаме кхача
Дена шен доьзал мичахь бу хиъчахьана тоуьйту аса-м аьлла шашена ела а къежна хьешана кхача бала дагахь тохаелира хlусамнана	Шен дагара бАла гучу ца болуьйтуш хала садеттара lаббаса

Обозначим буквой p предложение, содержащее омограф om; B_i, где $i = 1, 2$ — две базы предложений $p_{i,j}, j = 1, 2, \ldots, kp_i$, содержащих kp_i предложений. В результате получим два множества предложений:

$$B_1 = \{p_{1,1}, p_{1,2}, \ldots, p_{1,kp1}\} \text{ и } B_2 = \{p_{2,1}, p_{2,2}, \ldots, p_{2,kp2}\}.$$

Буквой N обозначим выбранное количество ближайших к омографу om слов — тегов, по которым алгоритм будет идентифицировать значение омографа. Для удобства выберем число N четным. Обозначим $sp_{i,j,m}$, где $m = 1, 2, \ldots, N$ — слова в предложении $p_{i,j}$. Из них $\frac{N}{2}$ слов $sp_{i,j,m}$, $m = 1, 2, \ldots, \frac{N}{2}$, слева от омографа om и $\frac{N}{2}$ слов $sp_{i,j,m}$, $m = \frac{N}{2} + 1, \frac{N}{2} + 2, \ldots, N$, справа от омографа om.

Такое симметричное разбиение количества тегов необязательно. Отметим, что оптимальность разбиения зависит от порядка слов в предложении в естественном языке. Этот выбор можно сделать экспериментально.

Базы слов. Сформируем две базы S_i, $i = 1, 2$ различных слов $s_{i,k}$, $k = 1, 2, \ldots, ks_i$, содержащихся хотя бы в одном предложении из баз B_i, $i = 1, 2$, где ks_i — количество слов в базе S_i. Будем считать, что каждому слову $s_{i,k}$ из базы S_i соответствуют неотрицательные целые числа $ns_{i,k,m}$, $m = 1, 2, \ldots, N$, равные количеству предложений из базы B_i, в которых слово $s_{i,k}$ совпадает со словом $sp_{i,j,m}$ на m-ом месте в предложении $p_{i,k}$:

$$ns_{i,k,m} = \sum_{\substack{k=1, \\ s_{i,k}=sp_{i,k,m}}}^{kp_i} 1, \; m = 1, 2, \ldots, N, \; i = 1, 2.$$

Таким образом, каждому слову $s_{i,k}$ из баз S_i, $i = 1, 2$, будет соответствовать N-мерный числовой вектор

$$\mathbf{vs_{i,k}} = (ns_{i,k,1}, ns_{i,k,2}, \ldots, ns_{i,k,N}).$$

Заметим, что в некоторых случаях, когда омограф в предложении расположен достаточно близко к началу или к концу предложения, или предложение состоит из малого числа слов, количество тегов может оказаться меньше числа N. В таких случаях соответствующая недостающему слову координата будет равна нулю.

***N*-мерное векторное пространство.** Если каждому предложению $p_{i,j}, j = 1, 2, \ldots kp_i$, из баз B_i, $i = 1, 2$, соответствует числовой вектор

$$\mathbf{vp_{i,j}} = (np_{i,j,1}, np_{i,j,2}, \ldots, np_{i,j,N}),$$

то координаты вектора равны сумме соответствующих координат вектора $\mathbf{vs_{i,j}}$:

$$np_{i,j,m} = \sum_{s_{i,j} \in S_i} \sum_{m=1}^{N} ns_{i,j,m}, \; i = 1, 2, j = 1, 2, \ldots, kp_i.$$

Таким образом, каждое предложение $p_{i,j}$ есть точка N-мерного векторного пространства. В этом пространстве имеем два множества B_1 и B_2 точек $p_{1,j}$ и $p_{2,j}$, $j = 1, 2, \ldots, N$. Идея рассматриваемого алгоритма заключается в том, что необходимо выяснить к какому из двух этих множеств «ближе» содержащее данный омограф тестовое предложение. Исходя из этого, в полученном векторном пространстве необходимо ввести понятие расстояния между предложениями.

Весовое евклидово пространство. Введем в этом пространстве метрику Евклида: расстояние между числовыми векторами $\mathbf{a} = (a_1, a_2, \ldots, a_N)$ и $\mathbf{b} = (b_1, b_2, \ldots, b_N)$ определим формулой

$$\rho(\mathbf{a}, \mathbf{b}) = \sqrt{\sum_{k=1}^{N} (a_k - b_k)^2}. \quad (1)$$

Естественно предположить, что чем дальше тег в предложении от омографа, тем меньше его влияние на значение омографа. Чтобы отразить в алгоритме данное условие, введем в формулу (1) так называемый весовой вектор

$$\mathbf{w} = (w_1, w_2, \ldots, w_N),$$

координаты которого удовлетворяют условиям

$$0 < w_1 \le w_2 \le \ldots \le w_{\frac{N}{2}} = 1 = w_{\frac{N}{2}+1} \ge w_{\frac{N}{2}+2} \ge \ldots \ge w_N > 0,$$

где N — четное число. Тогда весовая метрика (1) примет вид

$$\rho(\mathbf{a}, \mathbf{b}) = \sqrt{\sum_{k=1}^{N} w_k (a_k - b_k)^2}.$$

Тестовое предложение. Пусть tp — тестовое предложение, содержащее омограф om и не входящее в базы B_i, $i = 1, 2$. Необходимо распознать значение омографа om.

Обозначим теги вокруг омографа om в предложении tp через ts_i, $i = 1, 2, \ldots, N$. Координаты $ts_{i,k}$, $k = 1, 2, \ldots, N$, слов ts_i вычислим по обоим базам B_1 и B_2, значения которых равны количеству предложений, в которых слово $ts_{i,k}$ содержится на k-ом месте от омографа:

$$ts_{i,k} = \begin{cases} sp_{1,i,k}, \text{ если } ts_i = sp_{1,i} \\ 0, \text{ если } ts_i \ne sp_{1,i} \end{cases} + \\ + \begin{cases} sp_{2,i,k}, \text{ если } ts_i = sp_{2,i} \\ 0, \text{ если } ts_i \ne sp_{2,i} \end{cases},$$

$$i = 1, 2, \ldots, N, \; k = 1, 2, \ldots, N.$$

Таким образом, каждый тег ts_i в предложении tp имеет N координат:

$$\mathbf{ts_i} = (ts_{i,1}, ts_{i,2}, \ldots, ts_{i,N}), \; i = 1, 2, \ldots, N.$$

Просуммировав соответствующие координаты этих слов, получим координаты тестового предложения:

$$tv_i = \sum_{k=1}^{N} ts_{i,k}, \; i = 1, 2, \ldots, N.$$

Следовательно, тестовому предложению tp соответствует числовой вектор

$$\mathbf{tv} = (tv_1, tv_2, \ldots, tv_N).$$

Финальный этап. Теперь вычислим арифметические средние расстояний от предложения *tp* до предложений из баз B_1 и B_2, полагая, что

$$\rho(tp, p_{i,j}) = \rho(tv, vp_{i,j}), j = 1, 2, \ldots, kp_i, i = 1, 2,$$

$$r_i = \frac{1}{kp_i} \sum_{j=1}^{kp_i} \sqrt{\sum_{k=1}^{N} w_k (tv_k - np_{i,j,k})^2}, i = 1, 2.$$

Сравнивая полученные числа, выбираем значение омографа: если $r_1 < r_2$, то $om = om_1$, если $r_1 > r_2$, то $om = om_2$ и если $r_1 = r_2$, то $om = om_1$ при $kp_1 > kp_2$, иначе $om = om_2$.

Результаты работы программы на основе алгоритма AWEN

Для исходных параметров алгоритма экспериментальным путем установим оптимальные значения: количество тегов $N = 6$, т. е. значение омографа определяется по шести словам в предложении: по трем — слева и по трем — справа от омографа; весовой вектор $\mathbf{w} = (0{,}9; 0{,}95; 1; 1; 0{,}95; 0{,}9)$.

Для тестирования результатов работы алгоритма создан модуль самотестирования программы распознавания омографов. Для модуля выделены случайным образом выбранные части баз B_1 и B_2. На оставшихся частях выполнено обучение. Экспериментально определено оптимальное процентное соотношение между обучающей и тестовой частями баз — 80/20. Цикл обучения на подготовленных базах составил 100 итераций, в итоге программа вывела средние значения для разных метрик, которые отображены в табл. 2.

Таким образом, результаты работы программы, созданной по представленному в работе алгоритму распознавания омографов на основе евклидовой метрики, составили F1 — 39 %, Accuracy — 45 % (табл. 2).

Проведенный сравнительный анализ показал, что результат точности разработанного алгоритма наиболее близок к результатам алгоритмов на основе метода Леска, которые для английского языка варьируются в пределах F1 41–51 %.

Проанализированы результаты точности различных нейросетевых моделей для нескольких языков, наиболее близких к чеченскому. Например, показатель точности модели BERT для русского языка составил F1 — 67 %. Для арабского языка различные модификации этой модели (ArBERT, AraBERTv2) на разных корпусах данных показали результаты метрики F1 от 74 до 89 %. Для татарского языка модель разрешения морфологической многозначности, основанная на рекуррентной нейронной сети (LSTM) дала результат точности F1 — 79 %.

Таблица 2. Точность распознавания омографов, %

Table 2. Accuracy of homograph recognition, %

Омограф	F1	Accuracy
яха	53	49
ча	41	42
дакъа	40	41
белира	66	70
бала	34	36
дала	69	87
де	31	31
цIе	28	32
шун	28	32
йолу	18	20
бен	29	35
лар	31	41
дан	31	41

Методы, использующие нейросетевые алгоритмы, дают более высокие показатели точности по WSD для большинства языков, однако для их реализации требуется наличие больших корпусов данных, что не всегда доступно для малоресурсных языков, в том числе и для чеченского.

Заключение

Заметим, что алгоритм AWEN нетрудно дополнить для идентификации значений не только омографов, но омонимов, имеющих более двух значений. Однако подготовленные текстовые корпуса с предложениями содержат данные только для омографов. Такое ограничение было принято с целью экономии временных ресурсов. Поскольку необходимости в распознавании омонимов в целом для улучшения качества синтеза речи не было, имеющиеся ресурсы были направлены на автоматическое распознавание значений омографов, так как правильность именно их прочтения значительно влияет на качество синтеза речи.

Результаты работы алгоритма AWEN проанализированы и сопоставлены с результатами различных методов для разных языков. Для сравнения полученных оценок приведены значения мер F1 и Accuracy.

Для малоресурсного чеченского языка результаты работы программы, созданной по алгоритму распознавания омографов на основе весовой евклидовой метрики, составили F1 — 39 %, Accuracy — 45 %.

Литература

1. Израилова Э.С. Процесс создания системы синтеза чеченской речи // Известия Российского государственного педагогического университета им. А.И. Герцена. 2020. № 198. С. 171–177. https://doi.org/10.33910/1992-6464-2020-198-171-177
2. Izrailova E.S., Badaeva A.S. Analysis of the speech signal quality of the chechen speech synthesis system // Automatic Documentation and Mathematical Linguistics. 2021. V. 55. N 2. P. 74–78. https://doi.org/10.3103/S0005105521020059
3. Lesk M. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone // Proc. of the 5th Annual International Conference on Systems Documentation. 1986. P. 24–26. https://doi.org/10.1145/318723.318728
4. Banerjee S., Pedersen T. An adapted lesk algorithm for word sense disambiguation using WordNet // Lecture Notes in Computer Science. 2002. V. 2276. P. 136–145. https://doi.org/10.1007/3-540-45715-1_11
5. Lastra-Diaz J.J., Goikoetxea J., Taieb M.A.H., Garcia-Serrano A., Aouicha M.B., Agirre E. A reproducible survey on word embeddings and ontology-based methods for word similarity: linear combinations outperform the state of the art // Engineering Applications of Artificial Intelligence. 2019. V. 85. P. 645–665. https://doi.org/10.1016/j.engappai.2019.07.010
6. Kumar S., Jat S., Saxena K., Talukdar P. Zero-shot word sense disambiguation using sense definition embeddings // Proc. of the 57th Annual Meeting of the Association for Computational Linguistics. 2019. P. 5670–5681. https://doi.org/10.18653/v1/p19-1568
7. Scozzafava F., Maru M., Brignone F., Torrisi G., Navigli R. Personalized PageRank with syntagmatic information for multilingual Word Sense Disambiguation // Proc. of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. 2020. P. 37–46. https://doi.org/10.18653/v1/2020.acl-demos.6
8. Escudero G., Marquez L., Rigau G., Salgado J.G. On the portability and tuning of supervised word sense disambiguation systems: Research report. 2000.
9. Manning C.D., Clark K., Hewitt J., Khandelwal U., Levy O. Emergent linguistic structure in artificial neural networks trained by self-supervision // Proceedings of the National Academy of Sciences. 2020. V. 117. N 48. P. 30046–30054. https://doi.org/10.1073/pnas.1907367117
10. Lin D. Automatic retrieval and clustering of similar words // Proc. of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics. V. 2. 1998. P. 768–774. https://doi.org/10.3115/980691.980696
11. Hadiwinoto C., Ng H.T., Gan W.C. Improved Word Sense Disambiguation using pre-trained contextualized word representations // Proc. of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019. P. 5297–5306. https://doi.org/10.18653/v1/D19-1533
12. Vial L., Lecouteux B., Schwab D. Sense vocabulary compression through the semantic knowledge of WordNet for neural Word Sense Disambiguation // Proc. of the 10th Global Wordnet Conference. 2019. P. 108–117.
13. Scarlini B., Pasini T., Navigli R. SensEmBERT: Context-enhanced sense embeddings for multilingual Word Sense Disambiguation // Proceedings of the AAAI Conference on Artificial Intelligence. 2020. V. 34. N 5. P. 8758–8765. https://doi.org/10.1609/aaai.v34i05.6402
14. Scarlini B., Pasini T., Navigli R. With more contexts comes better performance: Contextualized sense embeddings for all-round Word Sense Disambiguation // Proc. of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020. P. 3528–3539. https://doi.org/10.18653/v1/2020.emnlp-main.285
15. Zhang C.X., Liu R., Gao X.Y., Yu B. Graph convolutional network for word sense disambiguation // Discrete Dynamics in Nature and Society. 2021. V. 2021. P. 2822126. https://doi.org/10.1155/2021/2822126
16. Conia S., Navigli R. Framing Word Sense Disambiguation as a multi-label problem for model-agnostic knowledge integration // Proc. of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. 2021. P. 3269–3275. https://doi.org/10.18653/v1/2021.eacl-main.286

References

1. Izrailova E. Creating a system for synthesizing the Chechen speech. *Izvestia: Herzen University Journal of Humanities & Sciences*, 2020, no. 198, pp. 171–177. (in Russian). https://doi.org/10.33910/1992-6464-2020-198-171-177
2. Izrailova E.S., Badaeva A.S. Analysis of the speech signal quality of the chechen speech synthesis system. *Automatic Documentation and Mathematical Linguistics*, 2021, vol. 55, no. 2, pp. 74–78. https://doi.org/10.3103/S0005105521020059
3. Lesk M. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. *Proc. of the 5th Annual International Conference on Systems Documentation*, 1986, pp. 24–26. https://doi.org/10.1145/318723.318728
4. Banerjee S., Pedersen T. An adapted lesk algorithm for word sense disambiguation using WordNet. *Lecture Notes in Computer Science*, 2002, vol. 2276, pp. 136–145. https://doi.org/10.1007/3-540-45715-1_11
5. Lastra-Diaz J.J., Goikoetxea J., Taieb M.A.H., Garcia-Serrano A., Aouicha M.B., Agirre E. A reproducible survey on word embeddings and ontology-based methods for word similarity: linear combinations outperform the state of the art. *Engineering Applications of Artificial Intelligence*, 2019, vol. 85, pp. 645–665. https://doi.org/10.1016/j.engappai.2019.07.010
6. Kumar S., Jat S., Saxena K., Talukdar P. Zero-shot word sense disambiguation using sense definition embeddings. *Proc. of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 5670–5681. https://doi.org/10.18653/v1/p19-1568
7. Scozzafava F., Maru M., Brignone F., Torrisi G., Navigli R. Personalized PageRank with syntagmatic information for multilingual Word Sense Disambiguation. *Proc. of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2020, pp. 37–46. https://doi.org/10.18653/v1/2020.acl-demos.6
8. Escudero G., Marquez L., Rigau G., Salgado J.G. *On the portability and tuning of supervised word sense disambiguation systems*. Research report, 2000.
9. Manning C.D., Clark K., Hewitt J., Khandelwal U., Levy O. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 2020, vol. 117, no. 48, pp. 30046–30054. https://doi.org/10.1073/pnas.1907367117
10. Lin D. Automatic retrieval and clustering of similar words. *Proc. of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics. V. 2*, 1998, pp. 768–774. https://doi.org/10.3115/980691.980696
11. Hadiwinoto C., Ng H.T., Gan W.C. Improved Word Sense Disambiguation using pre-trained contextualized word representations. *Proc. of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 5297–5306. https://doi.org/10.18653/v1/D19-1533
12. Vial L., Lecouteux B., Schwab D. Sense vocabulary compression through the semantic knowledge of WordNet for neural Word Sense Disambiguation. *Proc. of the 10th Global Wordnet Conference*, 2019, pp. 108–117.
13. Scarlini B., Pasini T., Navigli R. SensEmBERT: Context-enhanced sense embeddings for multilingual Word Sense Disambiguation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, vol. 34, no. 5, pp. 8758–8765. https://doi.org/10.1609/aaai.v34i05.6402
14. Scarlini B., Pasini T., Navigli R. With more contexts comes better performance: Contextualized sense embeddings for all-round Word Sense Disambiguation. *Proc. of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 3528–3539. https://doi.org/10.18653/v1/2020.emnlp-main.285
15. Zhang C.X., Liu R., Gao X.Y., Yu B. Graph convolutional network for word sense disambiguation. *Discrete Dynamics in Nature and Society*, 2021, vol. 2021, pp. 2822126. https://doi.org/10.1155/2021/2822126
16. Conia S., Navigli R. Framing Word Sense Disambiguation as a multi-label problem for model-agnostic knowledge integration. *Proc. of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 2021, pp. 3269–3275. https://doi.org/10.18653/v1/2021.eacl-main.286

17. Amrami A., Goldberg Y. Towards better substitution-based word sense induction // arXiv. 2019. arXiv:1905.12598. https://doi.org/10.48550/arXiv.1905.12598
18. Arefyev N., Sheludko B., Panchenko A. Combining lexical substitutes in neural word sense induction // Proc. of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019). 2019. P. 62–70. https://doi.org/10.26615/978-954-452-056-4_008
19. Vasilescu F., Langlais P., Lapalme G. Evaluating variants of the lesk approach for disambiguating words // Proc. of the Fourth International Conference on Language Resources and Evaluation (LREC'04). 2004.
20. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding // Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019. P. 4171–4186.
21. El-Razzaz M., Fakhr M.W., Maghraby F.A. Arabic Gloss WSD Using BERT // Applied Sciences. 2021. V. 11. N 6. P. 2567. https://doi.org/10.3390/app11062567
22. Kilgarriff A., Rosenzweig J. Framework and results for English SENSEVAL // Computers the Humanities. 2000. V. 34. N 1. P. 15–48. https://doi.org/10.1023/A:1002693207386
23. Гатауллин Р.Р., Гильмуллин Р.А., Хакимов Б.Э. Разрешение морфологической многозначности в корпусе татарского языка на основе статистико-вероятностной модели Purepos и нейросетевой модели LSTM // VI Международная конференция по компьютерной обработке тюркских языков «TurkLang 2018» (труды конференции). Ташкент: Издательско-полиграфический дом «Navoiy Universiteti», 2018. С. 133–138.
24. Haveliwala T.H. Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search // IEEE Transactions on Knowledge and Data Engineering. 2003. V. 15. N 4. P. 784–796. https://doi.org/10.1109/tkde.2003.1208999
25. Peters M.E., Neumann M., Iyyer M., Gardner M., Clark C., Lee K., Zettlemoyer L. Deep contextualized word representations // Proc. of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). 2018. P. 2227–2237. https://doi.org/10.18653/v1/N18-1202
26. Хомицевич О.Г., Рыбин С.В., Аничкин И.М. Использование лингвистического анализа для нормализации текста и снятия омонимии в системе синтеза русской речи // Известия высших учебных заведений. Приборостроение. 2013. Т. 56. № 2. С. 42–46.
27. WordNet: An Electronic Lexical Database // ed. by Ch. Fellbaum. Cambridge, MA: MIT Press, 1998. 423 p.
28. Ясаева М.Л. Создание баз данных чеченских текстов для обработки алгоритмов распознавания омографов компьютерными системами // Всероссийская научно-практическая конференция «Актуальные проблемы исследования родного языка и литературы». Грозный, 2022. С. 65–69.
29. Карпов А.А., Верходанова В.О. Речевые технологии для малоресурсных языков мира // Вопросы языкознания. 2015. № 2. С. 117–135.
30. Израилова Э.С., Астемиров А.В. Статистический контекстный анализ для снятия графической омонимии в текстах на чеченском языке // Материалы Международной научной конференции «Актуальные проблемы развития современной науки» посвященная 30-летию Академии наук Чеченской Республики. Махачкала: Академия наук Чеченской Республики, 2023. С. 478–485.

17. Amrami A., Goldberg Y. Towards better substitution-based word sense induction. *arXiv*, 2019, arXiv:1905.12598. https://doi.org/10.48550/arXiv.1905.12598
18. Arefyev N., Sheludko B., Panchenko A. Combining lexical substitutes in neural word sense induction. *Proc. of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, 2019, pp. 62–70. https://doi.org/10.26615/978-954-452-056-4_008
19. Vasilescu F., Langlais P., Lapalme G. Evaluating variants of the lesk approach for disambiguating words. *Proc. of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, 2004.
20. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. *Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019, pp. 4171–4186.
21. El-Razzaz M., Fakhr M.W., Maghraby F.A. Arabic Gloss WSD Using BERT. *Applied Sciences*, 2021, vol. 11, no. 6, pp. 2567. https://doi.org/10.3390/app11062567
22. Kilgarriff A., Rosenzweig J. Framework and results for English SENSEVAL. *Computers the Humanities*, 2000, vol. 34, no. 1, pp. 15–48. https://doi.org/10.1023/A:1002693207386
23. Gataullin R.R., Gilmullin R.A., Khakimov B.E. Morphological disambiguation in the national corpus of tatar language using Purepos and LSTM models. *VI International Conference on Computer Processing of Turkic Languages "TurkLang 2018" (Proceedings of the Conference)*, Tashkent, Navoiy Universiteti Publ., 2018, pp. 133–138. (in Russian)
24. Haveliwala T.H. Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. *IEEE Transactions on Knowledge and Data Engineering*, 2003, vol. 15, no. 4, pp. 784–796. https://doi.org/10.1109/tkde.2003.1208999
25. Peters M.E., Neumann M., Iyyer M., Gardner M., Clark C., Lee K., Zettlemoyer L. Deep contextualized word representations. *Proc. of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018, pp. 2227–2237. https://doi.org/10.18653/v1/N18-1202
26. Khomitsevich O.G., Rybin S.V., Anichkin I.M. Application of linguistic analysis for text normalization and homonymy resolution in Russian text-to-speech system. *Journal of Instrument Engineering*, 2013, vol. 56, no. 2, pp. 42–46. (in Russian)
27. *WordNet: An Electronic Lexical Database*. Ed. by Ch. Fellbaum. Cambridge, MA, MIT Press, 1998, 423 p.
28. Yasaeva M.L. Creation of databases of Chechen texts for processing homograph recognition algorithms by computer systems. *All-Russian Scientific and Practical Conference "Current Problems of Native Language and Literature Research"*, Grozny, 2022, pp. 65–69. (in Russian)
29. Karpov A.A., Verkhodanova V.O. Speech technologies for under-resourced languages of the world. *Voprosy Jazykoznanija*, 2015, no. 2, pp. 117–135. (in Russian)
30. Israilov E.S., Astemirov A.V. Statistical context analysis program for removing graphic homonymy in texts in the chechen language. *Proceedings of the International Scientific Conference "Current Issues in the Development of Modern Science" theme-based to the 30th anniversary of the Academy of Sciences of the Chechen Republic*, Makhachkala, Chechen Academy of Sciences, 2023, pp. 478–485. (in Russian)

Авторы

Израилова Элиса Салаудиновна — старший научный сотрудник, Академия наук Чеченской Республики, Грозный, 364043, Российская Федерация; младший научный сотрудник, Комплексный научно-исследовательский институт им. Х.И. Ибрагимова Российской академии наук, Грозный, 364051, Российская Федерация, https://orcid.org/0000-0003-1655-4414, uelisa@yandex.ru

Астемиров Арсланбек Виситович — научный сотрудник, Академия наук Чеченской Республики, Грозный, 364043, Российская Федерация; младший научный сотрудник, Комплексный научно-исследовательский институт им. Х.И. Ибрагимова Российской академии наук, Грозный, 364051, Российская Федерация, https://orcid.org/0009-0003-8416-7587, astemirow@mail.ru

Authors

Elisa S. Izrailova — Senior Researcher, Academy of Sciences of the Chechen Republic, Grozny, 364043, Russian Federation; Junior Researcher, Kh. Ibragimov Complex Institute of the Russian Academy of Sciences, Grozny, 364051, Russian Federation, https://orcid.org/0000-0003-1655-4414, uelisa@yandex.ru

Arslanbek V. Astemirov — Scientific Researcher, Academy of Sciences of the Chechen Republic, Grozny, 364043, Russian Federation; Junior Researcher, Kh. Ibragimov Complex Institute of the Russian Academy of Sciences, Grozny, 364051, Russian Federation, https://orcid.org/0009-0003-8416-7587, astemirow@mail.ru

Бадаева Айшат Салауддиновна — научный сотрудник, Академия наук Чеченской Республики, Грозный, 364043, Российская Федерация; младший научный сотрудник, Комплексный научно-исследовательский институт им. Х.И. Ибрагимова Российской академии наук, Грозный, 364051, Российская Федерация, https://orcid.org/0009-0009-5351-229X, ayshatbs@gmail.com

Султанов Зелимхан Аюбович — научный сотрудник, Академия наук Чеченской Республики, Грозный, 364043, Российская Федерация; младший научный сотрудник, Комплексный научно-исследовательский институт им. Х.И. Ибрагимова Российской академии наук, Грозный, 364051, Российская Федерация, https://orcid.org/0009-0001-1391-5475, zelimkhan.sultanov@yandex.ru

Умархаджиев Салаудин Мусаевич — доктор физико-математических наук, доцент, заведующий отделом, Академия наук Чеченской Республики, Грозный, 364043, Российская Федерация; заведующий лабораторией, Комплексный научно-исследовательский институт им. Х.И. Ибрагимова Российской академии наук, Грозный, 364051, Российская Федерация, SC 37089765500, https://orcid.org/0000-0002-8283-1515, usmu@mail.ru

Хехаев Мохмад-Салех Лейчевич — научный сотрудник, Академия наук Чеченской Республики, Грозный, 364043, Российская Федерация; младший научный сотрудник, Комплексный научно-исследовательский институт им. Х.И. Ибрагимова Российской академии наук, Грозный, 364051, Российская Федерация, https://orcid.org/0009-0009-7672-4148, saleh_1991@mail.ru

Ясаева Мадина Лечаевна — научный сотрудник, Академия наук Чеченской Республики, Грозный, 364043, Российская Федерация, https://orcid.org/0009-0001-3565-4974, madina.yasaeva@internet.ru

Ayshat S. Badaeva — Scientific Researcher, Academy of Sciences of the Chechen Republic, Grozny, 364043, Russian Federation; Junior Researcher, Kh. Ibragimov Complex Institute of the Russian Academy of Sciences, Grozny, 364051, Russian Federation, https://orcid.org/0009-0009-5351-229X, ayshatbs@gmail.com

Zelimhan A. Sultanov — Scientific Researcher, Academy of Sciences of the Chechen Republic, Grozny, 364043, Russian Federation; Junior Researcher, Kh. Ibragimov Complex Institute of the Russian Academy of Sciences, Grozny, 364051, Russian Federation, https://orcid.org/0009-0001-1391-5475, zelimkhan.sultanov@yandex.ru

Salaudin M. Umarkhadzhiev — D.Sc (Physics & Mathematics), Associate Professor, Head of Department, Academy of Sciences of the Chechen Republic, Grozny, 364043, Russian Federation; Head of Laboratory, Kh. Ibragimov Complex Institute of the Russian Academy of Sciences, Grozny, 364051, Russian Federation, SC 37089765500, https://orcid.org/0000-0002-8283-1515, usmu@mail.ru

Mokhmad-Salekh L. Khekhaev — Scientific Researcher, Academy of Sciences of the Chechen Republic, Grozny, 364043, Russian Federation; Junior Researcher, Kh. Ibragimov Complex Institute of the Russian Academy of Sciences, Grozny, 364051, Russian Federation, https://orcid.org/0009-0009-7672-4148, saleh_1991@mail.ru

Madina L. Yasaeva — Scientific Researcher, Academy of Sciences of the Chechen Republic, Grozny, 364043, Russian Federation, https://orcid.org/0009-0001-3565-4974, madina.yasaeva@internet.ru

Статья поступила в редакцию 19.04.2023
Одобрена после рецензирования 29.11.2023
Принята к печати 13.01.2024

Received 19.04.2023
Approved after reviewing 29.11.2023
Accepted 13.01.2024