АННОТАЦИЯ
диссертационной работы Акановой Акерке Сапаровны
«Нейрокомпьютерная система семантического анализа текста»,
представленной на соискание степени доктора философии (PhD)
по специальности: 6D075100 «Информатика, вычислительная техника и управление»

Актуальность работы

Казахстан является одним интенсивно развивающихся стран, который стремиться соответствовать уровню мирового образовательного пространства. Так же свои коррективы в мировое образовательное пространство внесла пандемия короновируса, которая заставила всех представителей образовательного общества обратиться к информационно-коммуникационным технологиям. Пандемия вынудила коммуницировать образовательное общество через компьютерные технологии, и тем самым показала ценность быстрого обмена информации, основным видом которой является информация, представленная в виде текстов на языке данной страны. У нас казахский язык. Слова в нем состоят из основы и добавляемых к ней аффиксов, которых может быть несколько, что позволяет из одного слова получить до 60 новых слов.

Отсюда, автоматическая обработка текста на казахском языке с целью семантического анализа позволит выделить из совокупности текстов наиболее существенные, концептуальные моменты, важные для данной конкретной задачи и сократит трудозатраты преподавателя на проверку текстовых работ. Да, сейчас казахский язык в компьютерной лингвистике достиг больших результатов. Исследования ученых в области автоматизации обработки текстов на казахском языке с применением нейронной сети заслуживают внимания, но видятся недостаточными. До сих пор отсутствуют модели обработки текстов на казахском языке без морфологического анализа, без словарей онотолии, что было исследовано и изложенно в данной работе.

Таким образом актуальность темы исследования обусловлена задачей создания новых моделей классификации по темам содержания текстов на казахском языке, решение которой позволяет достичь качественного определения семантической близости содержания текста заданной теме.

Цель диссертационной работы

Целью работы является создание модели и алгоритмов, применяемых для решения задачи определения семантической близости содержания текста заданной теме путем глубокого обучения выходных данных тематической модели, которая способствует созданию нейрокомпьютерной системы для семантического анализа текстов на казахском языке.

Для достижения заданной цели в рамках исследования выполняются следующие *задачи:*

1) Исследовать и провести анализ методов и моделей семантического анализа текста, применяемых при автоматической обработке текста.

2) Провести анализ технологии, способствующей глубокому обучению нейронных сетей.

3) Разработать словарь словоформ для автоматичекой обработки текстов на казахсом языке

4) Разработать алгоритм усечения аффиксов для слов казахского языка.

5) Разработать тематическую модель документов на казахском языке.

6) Разработать многослойную нейронную сеть для проведения обучения результата тематической модели.

7) Разработать нейрокомпьютерную систему для семантического анализа текста

Методы исследования

В ходе исследования применялись различные методы для решения поставленных задач. Широко применялись методы автоматической обработки текстов для проведения семантического анализа, методы аналитических исследований и математической статистики, в том числе методы статистической проверки гипотез (статистических критериев), основанных на распределении Стьюдента. Для построения математической модели обучения глубоких нейронных сетей использовался метод распространения обратной ошибки, а также инструменты компьютерного моделирования. В процессе выполнения экспериментальной части работы были использованы средства автоматизации математических расчетов и инструменты для визуализации результатов на базе языка программирования Python.

Объектом исследованияв данной диссертации является автоматическая обработка текстов на казахском языке.

Предметом исследованияв данной диссертации является Классификация содержания текстов на казахском языке по темам.

Научная новизна данного исследования заключается в разработке модели нейронной сети в исследованиях компьютерной лингвистки, требующей определения точности семантики используемого текста, основанной на алгоритме усечения аффиксов слов для создания словаря словоформ при решении задач классификации текстов на казахском языке.

Достоверность и обоснование. Обоснованность и достоверность результатов подтверждаются проведенным всесторонним анализом литературных источников по теме исследования и проведенным экспериментом по обучению нейронной сети. На основании трудов известных зарубежных и отечественных ученых в области компьютерной лингвистики в исследовании выбраны и применены модели и методы для автоматизации обработки текста на основе нейронных сетей. Произведен учет всех факторов при вычислении вероятности распределений тем в документе и слов в темах, функции ошибок в обучении нейронной сети, оптимизатора обучения и метрики. Достоверность выполненных работ обосновывается публикацией выводов исследования в рецензируемом журнале и внедрением нейрокомпьютерной системы в учебный процесс.

Теоретическая значимостьзаключается в применении алгоритма стемминга и нейронной сетевой модели в исследованиях, связанных с обработкой текста на казахском языке.

Практическая значимостьзаключается в применении:

– алгоритма усечения аффиксов для слов на казахском языке, который применяется для создания словаря словоформ при тематическом моделировании (LDA);

– модели нейронной сети с 4-мя слоями для глубокого обучения выходных данных LDA-модели:

а) выбраны оптимальные алгоритмы (слои) нейронной сети;

б) определены оптимальные параметры для компиляции нейронной сети, которые влияют на результат обучения нейронной сети.

– нейрокомпьютерной системы для семантического анализа текстового материала на казахском языке на основе тематической модели и построенной модели нейронной сети.

Алгоритм усечения аффиксов слов на казахском языке и обученная нейронная сеть получили свою практическую реализацию в нейрокомпьютерной системе, которая применяется для семантического анализа текстовых работ на казахском языке. Данный продукт внедрен в учебный процесс ПГУ имени С. Торайгырова.

Положения, выносимые на защиту:

– алгоритм усечения аффиксов слов на казахском языке;

– тематическая модель документов на казахском языке.

– модель нейронной сети для обучения векторов семантического анализа текстов на казахском языке;

– глубокое обучение нейронной сети

– нейрокомпьютерная система для анализа текстов на казахском языке и вероятностное определение семантической близости заданной теме.

Апробация работы. Апробация работы проходила на базе Павлодарского государственного университета имени С. Торайгырова, и был получен акт внедрения. Основные публикации докладывались на международных научных конференциях, публиковались в научных журналах, разработанные электронные продукты были зарегистрированы в государственном реестреправ на объекты, охраняемые авторским правом:

- Материалы Республиканской научно-практической конференции «Инновации в профессиональном образовании: проблемы и перспективы» приуроченной к 80-летию профессора БолатаӘбдікәрімұлы, февраль, 2019 г, Астана.

- Technology Audit and Production Reserves. DOI: 10.15587/2706-5448.2020.217613

- Свидетельство о внесений сведений государственный реестр прав на объекты, охраняемые авторским правом № 8300от 20 февраля 2020 года и №18050 от 27 мая 2021

Результаты диссертации опубликованы в 10 работах. Из них 1 монография, 3 статьи в журналах, рекомендованных Комитетом по контролю

в сфере образования и науки МОН РК, 1 статья в отечественном научном издании, 1 статья в международном научном издании, вошедший в базу данных Scopus, 1 статья в зарубежном научном издании, 1 работы в материалах международных и республиканских конференций, 2 свидетельства об авторском праве.

Личный вклад автора.

Основные экспериментальные и теоретические результаты, полученные в ходе проведения диссертационного исследования, получены автором самостоятельно. В публикациях в составе коллектива соавторов, соискатель является основным автором, которому принадлежат основные идеи при получении, обобщении и анализе достигнутых результатов. Структура диссертации включает в себя следующие разделы: вводная часть, основная часть (три главы), заключение, список использованных источников и приложения. Работа изложена на 113 страницах компьютерного текста, включает 30 рисунков, 15 таблиц и 199 наименований библиографических источников.

Основные результаты исследования.

В результате исследования научных источников приведена **классификация методов**, применяемых при автоматической обработке текста (рисунок 4) в целом.

Описан **метод распространения ошибки** в глубоком нейронном обучении, который был применен в разработке нейрокомпьютерной системы семантического анализа текста на казахском языке.

Проведен **сравнительный анализ технологий распознавания объектов на основе нейронных сетей,** где были описаны сравнительные характеристики 13 технологий – ApacheSinga, Caffe, Deeplearning4j, Keras, MicrosoftCognitiveToolkit, MXNet, NeuralDesigner, OpenNN, Theano, Torch, TensorFlow, NLTK, Gensim. Из приведенного анализапри классификации текстовых данныхболее убедительным является применение технологииGensim. Остальные технологии лучше подходят для распознавания изображений, видео и звука.

Разработан алгоритм усечения аффиксов на основе стеммера Портера для создания эталонного словаря словоформ. Создание словоформ можно делать разными методами, но самый распространенный – это стемминг. Стемминг – это процесс извлечения основы слова с помощью отсечения окончаний и суффиксов. На языке программирования Python реализована работа стеммера по усечению аффиксов для текстов на казахском языке.

Построена тематическая модель при помощи LDA для обработки текста на казахском языке. Для создания тематической модели используется корпус текстов на казахском языке, который состоит из коллекции документов. Затем корпус текстов подвергается токенизации, что дает нам разделение текста на слова. При помощи стеммера проводится отсечение отоснов суффиксов и окончаний и создается словарь. Слова из полученного словаряизвлекаются при помощи биграмм, которые определяют признак приближенности слов друг к другу, после полученных результатов биграммы

определяются темы с помощью технологии bag-of-words (BOW), в данной работе использовался готовый алгоритм doc2bow, который находит приближенные по значению ключевые слова. Затем при помощи алгоритма LDA ключевые слова распределяются по темам и темы распределяются по документам. В результате LDA получаем ключевые слова с весами относительно тем, распределение тем с весами по документам.

Результат полученной LDA-модели оценивается путем проведения глубокого обучения при помощи построенной нейросетевой модели. Таким образом, для обучения была построена нейронная сеть с 4 слоями. Входной слой Embedding() – преобразует ключевые слова с весами и темы с весами в векторные данные. Второй слой Spatial Dropout1D() – производит регуляризацию нейронов в сети. Третий слой LSTM – содержит внутри себя еще два слоя регуляризатора: один обычный dropout, а второй – рекуррентный dropout. Четвертый слой –выходной плотный слой Dense. Каждый слой выполняет свою функцию относительно полученных данных в нейронной сети. Каждый слой представляет собой алгоритм обработки нейронов путем вычисления взвешенной суммы. Процесс обучения начинается с обратного распространения ошибки – процесса вычисления полученной ошибки в обратном направлении, то есть от выхода к входу.

Разработана нейрокомпьютерная система обработки текстовых данных на казахском языке. Приведены этапы разработки нейрокомпьютерной системы:

1.Создание корпуса текстовна казахском языке.

2. Создание словаря основ путем применения алгоритма усечения аффиксов для казахского языка.

3. Обновление словаря выбранными биграммами или униграммами относительно тем (ключевые слова).

4.Создание тематической модели(LDA).

5. Создание многослойной нейронной сети.

6. Проведение обучения многослойной нейронной сети.

7.Связь задача и обученной модели нейронной сети.

8.Создание интерфейса для человеко-машинного взаимодействия.

9.Апробация нейрокомпьютерной системы на проведение семантического анализа текстов на казахском языке.

Разработана архитектура нейрокомпьютерной системы семантического анализа текста. Апробация нейрокомпьютерной системы заключается в проверке текстового материала на казахском языке на соответствие его содержания тематике текста. На экспериментальном уровне был рассчитан коэффициент корреляции, который показал линейную зависимость между результатами экспертов и самой программой, что подтверждает достоверность разработанных моделей и обученной нейронной сети. Программа была внедрена в университете Торайгырова и получен акт внедрения.