УДК 004.421
DOI: 10.15587/1729-4061.2019.179036

Разработка алгоритма поиска ключевых слов в корпусе текста на казахском языке

А. С. Аканова, Н.Н. Оспанова, Е. В. Кухаренко, Г. М. Абильдинова

Питання семантичного аналізу тексту займає особливе місце в комп'ютерній лінгвістиці. Дослідники даної області мають підвищений інтерес до розробки алгоритму, використання якого дозволить підвищити якість обробки корпусу тексту та ймовірнісне визначення змісту тексту. Результати дослідження застосувань методик, підходів, алгоритмів для семантичного аналізу тексту у комп'ютерній лінгвістиці в міжнародній і казахстанській науці призвели до розробки алгоритму пошуку ключових слів в тексті казахською мовою. Першим етапом алгоритму було складання еталонного словника ключових слів для корпусу тексту українською мовою. Вирішенням цієї проблеми стало застосування алгоритму Портера (стеммера) для корпусу текстів казахською мовою. Реалізація стеммера дозволила виділити унікальні основи слів і отримати еталонний словник, який згодом проіндексували. Наступний крок – це збір навчальних даних із корпусу текстів. Для обчислення ступеня семантичної близькості між словами кожному слову присвоюється вектор відповідних йому словоформ еталонного словника, в результаті якого виходить пара – ключове слово і вектор. І останнім кроком алгоритму є навчання нейронних мереж. При навчанні застосовується метод зворотного поширення помилок, що дозволяє провести семантичний аналіз корпусу тексту і отримати ймовірнісну кількість слів, близьку до очікуваної кількості ключових. Цей процес дозволяє автоматизувати обробку текстового матеріалу шляхом створення цифрових навчальних моделей ключових слів. Алгоритм використовується для розробки нейрокомп'ютерної системи, що буде проводити автоматичну перевірку текстових робіт учнів онлайн курсів. Унікальністю алгоритму пошуку ключових слів є застосування навчання нейронної мережі для текстів казахською мовою. У Казахстані вченими в області комп'ютерної лінгвістики було проведено ряд досліджень на основі застосування морфологічного аналізу, лемматизації та інших підходів і реалізовані лінгвістичні інструменти (в основному словники-перекладачі). Область застосування навчання нейронних мереж для синтаксичного аналізу казахської мови залишається відкритим питанням в казахстанській науці.

Розроблений алгоритм передбачає вирішення однієї з проблем в отриманні ефективного семантичного аналізу тексту казахською мовою

Ключові слова: ключове слово, алгоритм Портера, семантичний аналіз, нейронна мережа

1. Введение

В современных исследованиях [1–3] в области компьютерной лингвистики с применением искусственного интеллекта особое место занимает разработка

методов и инструментов для автоматической обработки текста. Первые системы состояли в основном из больших двуязычных словарей, где слова исходного языка давали один или несколько слов другого языка с учетом синтаксических правил. Эти системы впоследствии «сочли сложными, и подчеркнули необходимость в развитии систематических методов, что привело к созданию правил для синтаксического упорядочения».

В настоящее время исследования в области компьютерной лингвистики достигли уровня высоких интеллектуальных технологий. Чаще исследования направлены на решение задач машинного перевода, задач индексирования, реферирования, классификации и рубрицирования документов при полнотекстовом поиске. Компьютерная лингвистика сочетает знания компьютерных наук и лингвистики. Сложность моделирования естественного языка охватывает морфологический, синтаксический, фонологические уровни языка. Основной проблемой в этой области является создание систем искусственного интеллекта по обработке естественного языка. Компьютерная лингвистика изучает вопросы создания и использования электронных корпусов текстов, создание электронных словарей, тезаурусов, онтологий, машинные переводчики, извлечение информации из текста, автореферирование и построение систем управления знаниями.

Вычислительные задачи и проблемы по компьютерной лингвистике обсуждаются учеными на конференциях, организаторами которых выступает Ассоциация по компьютерной лингвистике (ASL – aclweb.org). Кроме этого, актуальной платформой по обсуждению новых исследований и их результатов являются международная конференция по компьютерной лингвистике «Диалог» (dialog-21.ru) и международная конференции по компьютерной лингвистике и интеллектуальной обработке текста (cicling.org).

Одной из целей ученых, исследующих интеллектуальный анализ текста, автоматизацию обработки текста, семантические отношения слов и предложений в тексте, есть создание интеллектуального инструмента для оценивания эссе, письменных работ и других письменных творческих работ обучающихся. Одним из решений задач таких задач является эффективно функционирующий когнитивный инструмент по автоматической обработке текста – Automated Essay Scoring [4]. В настоящее время существуют четыре вида систем AES, которые широко используются тестирующими компаниями, университетами и государственными школами это: Project Essay Grader (PEG), Intelligent Essay Assessor (IEA) (Measurement Incorporated (MI), США), E-rater (ETS, США) и IntelliMetric(Vantage Learning, США) [5].

Из анализа исследований видим, что широко изученными естественными языками в компьютерной лингвистики остаются английский, русский, немецкий, китайский, французский, испанский, турецкий. Но одним из малоизученных остается казахский язык. Малоизученность задач компьютерной лингвистики по обработке текстов на казахском языке является одной из причин начала исследования по разработке алгоритма поиска ключевых слов. Казахский язык относится к аглютинативным языкам, для которых уже существуют немало алгоритмов поиска ключевых слов и составления словарей, в особенности для кипчакских языков и турецких языков, но глубокое обучение нейронных

сетей при обработке корпусов текста на казахском языке не применялось. Таким образом, применение глубокого обучения нейронной сети для поиска ключевых слов, полнотекстового поиска для корпусов текстов на казахском языке подчеркивает актуальность темы в области компьютерной лингвистики.

2. Анализ литературных данных и постановка проблемы поиска ключевых слов

В области компьютерной лингвистики на данный момент проведены очень много исследований, результатом которого являются переводчики-словари, тезаурусы, поисковые системы интернета. Основой этих результатов являются разработка и применение методов и алгоритмов смыслового (семантического) анализа текстов, извлечения и использования знаний для интеллектуального компьютерного анализа. Исследования в области автоматической обработки текста начинаются с изучения структуры естественного языка, которое включает некоторые виды анализа текста: досемантический, графематический, синтаксический, фрагментационный, морфологический лемматизации текста. Применение выше названных анализов для русского языка можно увидеть на сайте aot.com.

В работе [6] была представлена эволюционная нейродинамическая основа для разработки процесса обучения, основанного на визуальной записи и извлечении нейронных весов посредством нейродинамических опытов при прохождении ссылочного содержимого. Также в работе [7] рассматривается модель извлечения неконтролируемых отношений, которая называется представлением распределения отношений. Представление о распределении отношений направлено на автоматическое изучение векторов сущностей и дальнейшую оценку семантического сходства между этими сущностями. В исследовании [8] предлагается новая методика распознавания отдельных вопросительных слов из речевого запроса одного из южно-индийских языков. В данном исследовании применены преобразование Фурье (FFT) и дискретное косинусное преобразование (DCT) для извлечения признаков, а для классификации и распознавания применена искусственная нейронная сеть (ANN).

Создание интеллектуальной поисковой системы на основе тезауруса был рассмотрен в трудах [9] и предложен подход к созданию семантических метрик и установление семантической связи между определенными терминами. В статье [10] представлен подход классификации соединения, использующий контекстно-семантические функции и алгоритм инкрементного обучения на основе LFNN для классификации текста. Предложенный метод позволяет классификатору динамически изучать модель в динамической базе данных. Этот процесс обучения использует нейронную сеть Back Propagation Lion (BPLion), где он включает нечеткое ограничение и алгоритм Lion (LA) для возможного выбора весов. Исследование в области обработки естественного языка путем ее автоматической корректировки было предложено в работе [11] для текстов на китайском языке. Был представлен алгоритм автоматической проверки корректуры текста.

Нередко используется метод онтологии для реализации классификации данных и связи между данными при семантическом анализе в автоматической обработке текста [12]. В работе [13] рассмотрели мультиагентный подход с взаимодействием двух агентов: первый соответствует значимым единицам извлекаемой информации и второго агента-правила, реализующие пополнения заданной онтологии на основе семантико-синтаксической модели языка. В исследовании [14] применялись семантические сети в извлечении и визуализации знаний, графики глаголов с реляционными графами для реализации логики первого порядка.

В начале 2000-х ученые начали глубже исследовать и применять латентно-семантический анализ в мелкомасштабных корпусах для автоматизированной оценки академических очерков [15]**.** Так же латентно-семантический анализ был применен [16] для реализации метода автоматического суммирования текста, которые используются для оценки релевантности предложения.

Многие задачи семантического анализа текста, такие как поиск текста, суммирование текста и сравнение текста, зависят от извлечения весовых ключевых слов из корпуса текста. В работе [17] предлагают графовую модель текста, позволяющую вычислять частотные характеристики слов текста с учетом расположения пар слов. С учетом такой модели данных в статье предложен алгоритм определения ключевых слов текста. Алгоритм учитывает слова русского языка, удовлетворяющие двум условиям: слово состоит не менее чем из 4 букв; слово распознается морфологическим анализатором как существительное. Основной целью извлечения ключевых слов в компьютерной лингвистике является определение семантической связи слов в разных корпусах текст. Например, исследователями [18] был предложен алгоритм извлечения ключевых слов из патентной документации (PKEA-Patent Keyword Extraction Algorithm), основанный на распределенной модели скип-граммы для классификации патентов. Для достижения цели были использованы стандартные наборы эталонных данных и самодельный набор патентных данных для оценки производительности PKEA.

Некоторыми исследователями [19], для поиска и определения семантически связанных слов, применялись алгоритм оптимизации поиска кукушки в сочетании с алгоритмом генератора ответов, для повышения семантической точности найденных предложений.

На данный момент существует множество исследований о поиске ключевых слов, разработаны разные методы, подходы, а также алгоритм обучения, который должен научить классифицировать как положительные или отрицательные примеры ключевых фраз. С этой целью был специально разработан алгоритм GenEx, который отражен в работе [20] и включает специализированные знания процедурной области, имея наибольший успех в извлечении ключевых слов, чем обычный алгоритм. В результате анализа литературы можно сделать вывод, что каждый из проведенных исследований отражает работу в области автоматической обработки текста, организации семантического анализа текста, применяя при этом разные алгоритмы и модели извлечения ключевых слов. Однако в данных работах остается не затронутым вопрос составления словаря ключевых фраз и базы ключевых слов-слоформ требуемого естественного язы-

ка. И главным моментом для проведения семантического анализа текста на казахском языке стало составление словаря ключевых фраз. Ключевые фразы – это структурные единицы текста, которые в той или иной степени являются важными составляющими при передаче текста. И чаще всего наборы ключевых слов и словосочетаний обычно содержат наиболее важную информацию в понимании смысла текста и формируют общее представление о его содержании.

В результате проведенного исследования можно сказать, что на данный момент остались нерешенными вопросы, связанные с разработкой интеллектуальных инструментов для проведения семантического анализа текстов на казахском языке. Причиной является объективные трудности, связанные с отсутствием алгоритмов поиска ключевых слов из корпуса текста на казахском языке. Вариантом преодоления этих трудностей может быть разработка алгоритма с применением глубокого обучения нейронной сети для казахского языка. Именно глубокое обучение нейронных сетей и методы его реализации предложены в работе [21], где используют этот подход для понимания семантики видео. В работах [22, 23] глубокое обучение нейронных сетей применяют для сквозного обнаружения текста для восстановления и улучшения изображения. Однако использование глубокого обучения нейронной сети для поиска ключевых слов для семантического анализа текста не были отражены в исследованных работах ученых.

Глубокое обучение нейронных сетей широко используются для обработки графических изображений [24]. Обучение искусственных нейронных сетей с помощью метода глубокого обучения (Deep learning) заняла немаловажное место в компьютерной лингвистке, который впервые был использован в 2006 году [25]. На сегодняшний день разработаны методы обучения нейронных сетей, позволяющие быстро и качественно обучать сети, состоящие из ста и более слоев [23].

Все это позволяет утверждать, что целесообразным является проведение исследования, посвященного разработке алгоритма поиска ключевых слов в корпусе текста на казахском языке с применением глубокого обучения нейронной сети. Разработка алгоритма требовалась для дальнейшей разработки нейрокомпьютерной системы с проверкой текстовых работ обучающихся на казахском языке.

3. Цель и задачи и задачи исследования

Целью исследования является разработка алгоритма поиска ключевых слов по тексту на казахском языке.

Для достижения цели были поставлены следующие задачи:

– привести в машиночитаемый вид корпус текстов, который включает определение слов/словоформ казахского языка с помощью стеммера Портера;

– собрать данные и провести обучение нейронной сети.

4. Приведение в машиночитаемый вид корпуса текстов

В качестве корпуса текстов был использован дамп базы Wikipedia на казахском языке по состоянию на апрель 2019 г. Для дальнейшей работы с ним было необходимо составить словарь уникальных словоформ.

Исходный файл дампа был предварительно очищен от служебных слов XML разметки и повторений. В результате был получен словарь из 1062058 слов. Для его сокращения был разработан и применен стеммер, основанный на алгоритме Портера.

Алгоритм стемминга (стеммер Портера используем) часто применяется в подходах по комплексной идентификации слов [26], по манипулированию файлами, поиск и сценарии для конкретных приложений [27], не использует базы основ слов, а лишь, применяя последовательно ряд правил, отсекает окончания и суффиксы, основываясь на особенности языка, в связи с чем работает быстро.

Идея стеммера Портера заключается в том, что существует ограниченное количество формобразующих и словообразующих суффиксов. Стемминг Портера использует множество существующих суффиксов (при этом сложные составные суффиксы разбиваются на простые) и вручную заданные правила. Реализация стеммера для турецкого, румынского, армянского, католанского, греческого, литовского языков размещен в [28]. Отсюда видно, что стеммер для казахского языка не был реализован.

Алгоритм состоит из пяти шагов. На каждом шаге отсекается формо- или словообразующий суффикс и оставшаяся часть проверяется на соответствие правилам (например, для русских слов основа должна содержать не менее одной гласной). Если полученное слово удовлетворяет правилам, происходит переход на следующий шаг. Если нет – алгоритм выбирает другой суффикс для отсечения. Согласно официальному сайту проекта, на первом шаге отсекается максимальный формообразующий суффикс, на втором – буква «и», на третьем – словообразующий суффикс, на четвертом – суффиксы превосходных форм, «ь» и одна из двух «н». То, что стеммер Портера не использует никаких словарей и баз основ, является плюсом для быстродействия и спектра применения (он неплохо справляется с несуществующими словами), и одновременно минусом с точки зрения точности выделения стеммы. Алгоритм часто обрезает слово больше необходимого, а это приводит к затруднению синтеза нормальной формы по получающейся стемме: аманшылық–>ама (при этом действительно неизменяемой частью является аман. При реализации алгоритма Портера для казахского языка в коде программы приводим все суффиксы и окончания казахского языка:

```
_re_all=re.compile(
r"(шалық|шелік|даған|деген|таған|теген|лаған|леген|"
r"дайын|дейін|тайын|тейін|кент|хана"r"ндар|ндер|дікі|тікі|нікі|атын|етін|йты
н|йтін|"
r"гелі|қалы|келі|ғалы|шама|шеме|"
r"мын|мін|бын|бін|пын|пін|мыз|міз|быз|біз|пыз|піз|сың|сің|"
r"сыз|сіз|ңыз|ңіз|дан|ден|тан|тен|нан|нен|нда|нде|дың|дің|тың|"
r"тің|ның|нің|дар|дер|тар|тер|лар|лер|бен|пен|мен|стан|"
r"дай|дей|тай|тей|дық|дік|тық|тік|лық|лік|паз|"
r"ғыш|гіш|қыш|кіш|шек|шақ|шыл|шіл|нші|ншы|дап|деп|"
r"тап|теп|лап|леп|дас|дес|тас|тес|лас|лес|ғар|гер|қар|кер|дыр|"
```

```
r"дір|тыр|тір|ғыз|гіз|қыз|кіз|ған|ген|қан|кен|"
r"ушы|уші|лай|лей|сын|сін|бақ|бек|пақ|пек|мақ|мек|йын|йін|йық|йік|"
r"сы|сі|да|де|та|те|ға|ге|қа|ке|на|не|"
r"ді|ты|ті|ны|ні|ды|ба|бе|па|пе|ма|ме|"
r"лы|лі|ғы|гі|қы|кі|ау|еу|ла|ле|ар|ер|"
r"ып|іп|ша|ше|ші|шы|са|се|"
r"и|й|ы|і)$")
```

Путем проведения через стеммер 1062058слова из дампа, слова присваивается переменной: word="орналасқан" и запускается стеммер. Следовательно, происходит процесс удаления аффиксов стеммером имеющих в базе слов, некоторые примеры процесса приведены в табл. 1.

```
"""
unittest.main()
"""
stemmer=Stemmer()
word="орналасқан"
word=stemmer.stem(word)
print(word)
```

Таблица 1
Процесс удаления аффиксов стеммером имеющих в базе слов

Исходный текст	Ожидаемый вариант	Результат стеммера
Орналасқан	Орналас–қан	Орналас
Деректердің	Дерек+тер+дің	Дерек
Соншалықты	Сонша+лық+ты	Соншалық
Сапалықты	Сапа+лық+ты	Сапалық
Сүңг+и+т+ін	Сүңг+и+т+ін	Сүңгит
Таң+ғы	Таң+ғы	Таң

Полученный словарь был проиндексирован для дальнейшего использования алгоритмом.

Для извлечения ключевых слов из корпуса текста использован подход, основанный на использовании словарей предметных областей (domain dictionary) [29]. Из файла дампа был составлен словарь ключевых терминов, которыми будет оперировать алгоритм. Для этого были использованы темы статей, количество которых составляло около 224 000. В данном подходе ключевое внимание уделяется формированию словаря, в основу которого должны быть включены тщательно отобранные экспертом термины по предметной области. Некоторая часть тем, посвященных персоналиям, населенным пунктам и другие, не подходящие в качестве ключевых слов, в словарь не включалась. Таким образом, длина словаря ключевых терминов составила 50 000 элементов.

5. Обучение нейронной сети

Нам известно, что словарь из 1 062 058 слов является избыточным и при попытке к обучению приведет к значительному увеличению машинного времени на обучение нейронной сети.

По проведенным расчетам определили более близкий размер словаря для качественной работы. После выделения уникальных основ слов размер словаря сократился до 135120 элементов.

Для эксперимента была создана нейронная сеть типа перцептрон. Количество рецепторов перцептрона равно длине словаря основ слов, т. е. 135120. В выходном слое количество нейронов равно длине словаря ключевых терминов значит 50 000.

Выбирая количество элементов скрытого слоя, принималось во внимание, что задача нейронной сети заключается в обобщении входного массива данных. Kevin Swingler [30] рекомендует в таких условиях использовать сужающуюся нейронную сеть, т. е. сеть с количеством нейронов в скрытом слое меньшем, чем во входном.

Например, для предельной ошибки e=0.1 необходимо использовать обучающую последовательность в 10 раз большую количества весов. Эта зависимость описывается формулой:

$$n \geq \frac{\omega}{\varepsilon}. \qquad (1)$$

Согласно формуле (1), где количество тренировочных прмеров (n) равно произведению количество связей (ω) на обратную величину ошибки (1/ε), в результате сокращения получаем отношение количество сзязей (ω) к количеству ошибок (ε).

Отсюда, использование большего количества связей, чем может заполнить обучающий набор данных, вредит обобщающей способности, что было выявлено путем сравнения экспериментально построенных «кривых обучения» (рис. 1, x – ошибки в процентном соотношении, y – количество проводимых экспериментов), соответствующий максимуму обобщающей способности. При этом в скрытом слое сети оказалось 65000 нейронов.

Непосредственно для обучения была подготовлена выборка, в которой каждому ключевому термину соответствовала векторная модель текста вида (2, 0, 4, ..), где первый элемент соответствует числу вхождения в текст первого основания слова из словаря и т. д. Для удобства использования в качестве данных для обучения нейронной сети, частоты вхождения были нормализованы в интервале от 0 до 1.

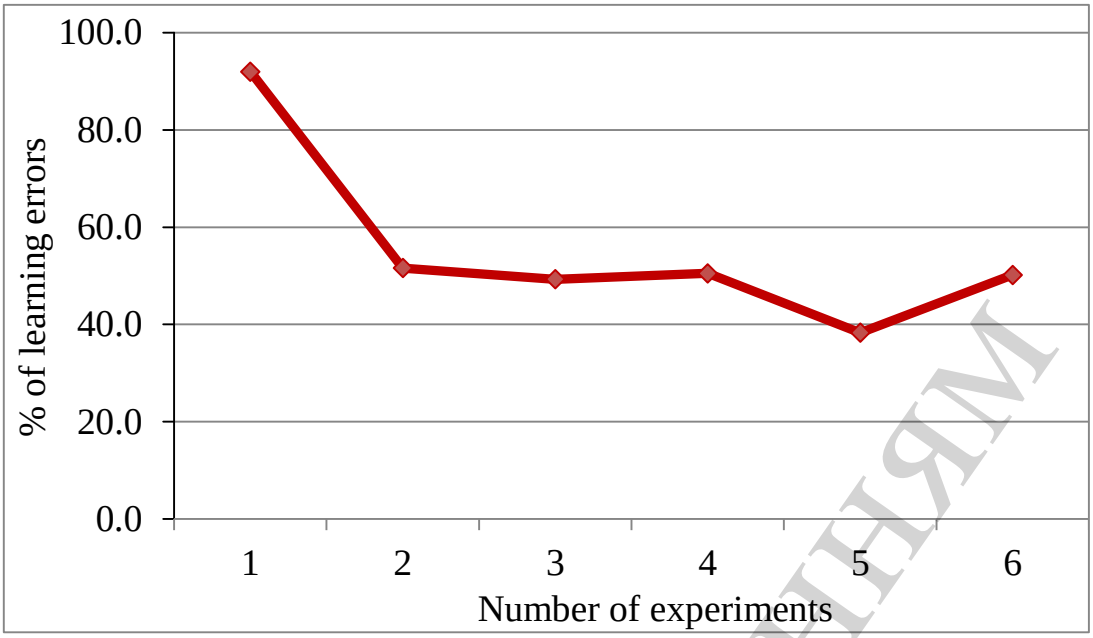

Рис. 1. Размер ошибок (%) обучения при каждом проводимом эксперименте

Для обучения нейронной сети нужно произвести большое количество итерации с корпусом текста, которое сможем реализовать с помощью алгоритма обратного распространения ошибки (error backpropagation algorithm). В этом случае используется обучающее множество корпусов текста с заранее известными ключевыми словами, в результате минимизации ошибок выявляем отличие выходных значений нейронной сети от входных ключевых слов. Данный итеративный градиентный алгоритм используется с целью минимизации ошибки работы многослойного перцептрона и получения желаемого выхода. Обучаем отбору лучших вариантов, который происходит путем сравнения векторной модели исследуемого текста с пороговыми значениями поисковой модели. Векторная модель предполагает сопоставление каждому документу частотного спектра слов и соответственно вектора в лексическом пространстве. В процессе поиска частотный портрет запроса рассматривается как вектор в том же пространстве и по степени близости (расстоянию или углу между векторами) определяются наиболее релевантные документы. В более продвинутых векторных моделях размерность пространства сокращается отбрасыванием наиболее распространенных или редко встречающихся слов, увеличивая тем самым процент значимости ключевых слов. Далее высчитывается релевантность каждого ключевого слова корпуса текстов путем сопоставления ему вектора значений. Определяются их вероятности отнесения к ключевым в соответствии с построенной моделью для приближения показателя к ожидаемому результату.

Фиксируются отличие значений векторов этих параметров для ключевых слов и не ключевых. Далее вычисляется вероятность отнесения каждого слова к группе ключевых и задается ее порог, т. е. модель обучается.

6. Результат алгоритма поиска ключевых слов из корпуса текста на казахском языке

В результате выполнения стеммера Портера создан словарь для поиска ключевых слов на казахском языке, который включает базу основ слов на казахском языке и словарь терминов для обучения нейронной сети. Данная база будет использоваться для разработки системы семантического анализа текста, для дистанционной проверки электронных текстовых работ обучающихся.

Например, выбран корпус текста

Текст 1. *Нейрожелі тәжірибелік мәліметтерді сақтауға және қолдануға табиғи бейімділігі бар параллельді таратылған процессорлар жиыны. Ол екі жағдайда миға ұқсас:*

1) Білім қоры желіні оқыту үрдісінде қалыптасады.

2) Синаптикалық салмақ ретінде анықталған нейрон аралық бірігу күштері есте сақтау үшін қолданылады.

Салмақ – Жер бетіне жақын тұрған денеге әсер ететін ауырлық күшінің сандық шамасы: P=mg, мұндағы m - дене массасы, g - еркін түсу үдеуі (немесе ауырлық күшінің үдеуі). Дененің массасы тұрақты шама, ал g мәні Жер бетіндегі ендікке және теңіз деңгейінен есептелетін биіктікке байланысты (мысалы, Алматы үшін g=9,804 м/с^2) өзгеретіндіктен, оған сәйкес дененің салмағы да өзгереді.

Таблица 2

Словоформы из словаря ключевых слов и ключевых терминов

№	словоформы	№	словоформы	№	термины	№	термины
1	*Нейрожелі*	8	*тұр*	1	*Нейрожелі*	8	*сандық шама*
2	*тәжірибе*	9	*жақын*	2	*заңдылық*	9	*бірігу күштері*
3	*мәлімет*	10	*күші*	3	*ғылым*	10	*әдіс-тәсілдері*
4	*сақта*	11	*сан*	4	*Ауырлық күш*	11	*дәлелдеу*
5	*қолдан*	12	*байланыс*	5	*Синаптикалық салмақ*	12	*Тұрақты шама*
6	*бейім*	13	*тұрақты*	6	*дене салмағы*	13	*Дене массасы*
7	*дене*	14	*заңдылық*	7	*Білім қоры*	14	*Жер беті*

В табл. 2 приведы словоформы (1, 2 столбик) из словаря ключевых слов и ключевые термины (3 ,4 столбик).

При обучении нейронной сети были использованы словари с основами с количеством в 135 120 слов и словарь терминов в 50 000 слов. Нейронная сеть состояла из одного скрытого слоя.

Структура нейронной сети выглядит следующим образом

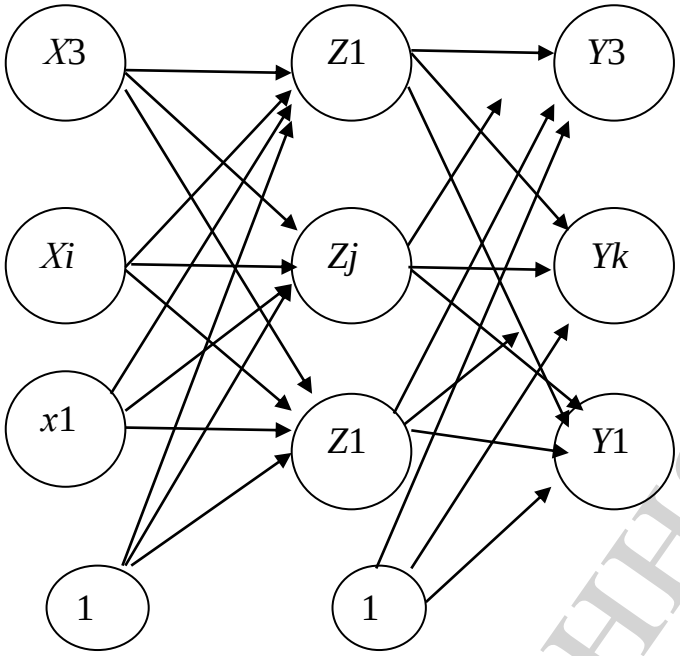

Рис. 2. Нейронная сеть с обратным распространением ошибок с одним скрытым слоем

Здесь нейроны, представляющие собой выходы сети (*Y*), и скрытые нейроны могут иметь смещение (1). Эти смещения служат в качестве весов на связях, исходящих от нейронов, на выходе которых всегда появляется 1. Кроме того, на рисунке стрелками показано перемещение информации в ходе фазы распространения данных от входов к выходам. В процессе обучения сигналы распространяются в обратном направлении. Как уже было сказано ранее, обучение сети включает в себя три стадии: подача на входы сети обучающих данных, обратное распространение ошибки и корректировка весов. В ходе первого этапа каждый входной нейрон (*X*) получает сигнал и широковещательно транслирует его каждому из скрытых нейронов (*Z*). Каждый скрытый нейрон затем вычисляет результат его активационной функции (сетевой функции) и рассылает свой сигнал всем выходным нейронам. Каждый выходной нейрон (*Y*), в свою очередь, вычисляет результат своей активационной функции, который представляет собой ни что иное, как выходной сигнал данного нейрона для соответствующих входных данных. В процессе обучения, каждый нейрон на выходе сети сравнивает вычисленное значение *Y* с предоставленным учителем *T* (целевым значением), определяя соответствующее значение ошибки для данного входного шаблона На основании этой ошибки вычисляется *Qk* (*k*=1, 2…). *Qk* используется при распространении ошибки от *Y* до всех элементов сети предыдущего слоя (скрытых нейронов, связанных с *Yk*), а также позже при изменении весов связей между выходными нейронами и скрытыми. Аналогичным образом вычисляется *Qj* (*j*=1, 2…) для каждого скрытого нейрона *Zj*. Несмотря на то, что распространять ошибку до входного слоя необходимости нет, *Qj* используется для изменения весов связей между нейронами скрытого слоя и входными нейронами. После того как все *Q* были определены, происходит одновременная корректировка весов всех связей.

Таким образом, после получения числовых данные весов был расчитан коэффициент корреляции, который составил 0,99 %. Этот результат показал линейную зависимость входных и выходных данных нейронной сети, что говорит о вероятности соответствия количества слов полученного словаря и количества слов словарей казахского языка. Отсюда, существенные отклонения при обучении нейронной сети не наблюдаются, соответственно достигается желаемый результат. Если сравнить с количеством слов «Толкового словаря казахского языка» – 106 000 слов, то словарь является более подходящим к базе слов казахского языка.

7. Обсуждение результатов исследования по разработке алгоритма поиска ключевых слов

После применения стеммера Портера для поиска ключевых слов в корпусе текстов на казахском языке был получен словарь основ слов в 135 120 словоформ и эталонный словарь ключевых слов (или словарь терминологий) с результатом в 50 000 слов.

Для проведения синтаксического анализа текста на казахском языке был выбран алгоритм Портера. Благодаря его применению были созданы словари для поиска ключевых слов. В работах по разработке лингвистических процессоров для казахского языка были рассмотрены лексико-морфологический и морфологический анализы текста, где для реализации применяли предметную онтологию и словарь суффиксов и аффиксов [31, 32]. Отсюда, исследование семантического анализа текстов на казахском языке при помощи нейронной сети и составление словаря ключевых слов при помощи алгоритма Портера является актуальным.

Данное исследование применимо только для обработки текстовой информации и данные алгоритмы не применимы для другого формата информации (изображения, видео, аудио), что является одним из его недостатков. На основе данного исследования разрабатывается нейрокомпьютерная система, которая позволит проводить семантический анализ текстов и определять, соответствует ли текст заданной теме. Нейрокомпьютерная система будет включать семантический анализатор текста на казахском языке, который можно будет использоваться на онлайн курсах для проверки текстовых работ обучающихся в организациях образования [33].

Если систему наполнить словарями других естественных языков, ее можно применить и к другим языкам.

В связи с переходом казахского языка на латиницу, в будущем следует изучить адаптированность данного исследования к латинице.

7. Выводы

1. Отсюда, в результате работы стеммера Портера создан словарь ключевых слов с итоговым количеством 135 120 словоформ и эталонный словарь ключевых терминов, который включает 50 000 слов. Приведенное количество основ слов и эталонный словарь ключевых слов являются вероятностным приближением к количеству слов толкового словаря казахского языка, в результате

использования их приближенная ошибка составляет 50–90 %. Это является неплохим результатом и обеспечивает возможность определения приближенного количества ключевых слов.

2. В результате подготовки данных для обучения получили пару: ключевое слово и вектор соответствующих ему словоформ. А также фиксированное отличие значений векторов словоформ для ключевых слов и не ключевых. После вычисляется вероятность отнесения каждого слова к группе ключевых и задается ее порог, то есть модель обучается. Определены весы с учетом смещения нейронов во внутреннем слое нейронной сети, при этом корреляционный анализ показал линейную зависимость входных и выходных данных нейронной сети с показателем в 0,99.

Литература

1. Bassiou, N. K., Kotropoulos, C. L. (2014). Online PLSA: Batch Updating Techniques Including Out-of-Vocabulary Words. IEEE Transactions on Neural Networks and Learning Systems, 25 (11), 1953–1966. doi: https://doi.org/10.1109/tnnls.2014.2299806
2. Borschev, V. B., Partee, B. H. (2014). Ontology and Integration of Formal and Lexical Semantics. Proceedings of the international scientific conference on computational linguistics "Dialogue". URL: http://www.dialog-21.ru/digests/dialog2014/materials/pdf/BorschevVBParteeBH.pdf
3. Turdakov, D. Y., Astrakhantsev, N. A., Nedumov, Y. R., Sysoev, A. A., Andrianov, I. A., Mayorov, V. D. et. al. (2014). Texterra: A framework for text analysis. Programming and Computer Software, 40 (5), 288–295. doi: https://doi.org/10.1134/s0361768814050090
4. Attali, Y., Burstein, J. (2006). Automated Essay Scoring With E-rater® V.2. Journal of Technology, Learning, and Assessment, 4 (3). URL: https://ejournals.bc.edu/index.php/jtla/article/view/1650/1492
5. Dikli, S. (2006). Automated Essay Scoring. Turkish Online Journal of Distance Education, 7 (1), 49–62. URL: https://www.researchgate.net/publication/26415982_Automated_Essay_Scoring
6. Rai, A., Kannan, R. J. (2018). Differed Restructuring of Neural Connectome Using Evolutionary Neurodynamic Algorithm for Improved M2M Online Learning. Procedia Computer Science, 133, 298–305. doi: https://doi.org/10.1016/j.procs.2018.07.037
7. Chen, Z., Huang, Y., Liang, Y., Wang, Y., Fu, X., Fu, K. (2017). RGloVe: An Improved Approach of Global Vectors for Distributional Entity Relation Representation. Algorithms, 10 (2), 42. doi: https://doi.org/10.3390/a10020042
8. Sukumar A., R., Sukumar A., S., Shah A., F., Anto P., B. (2010). Key-Word Based Query Recognition in a Speech Corpus by Using Artificial Neural Networks. 2010 2nd International Conference on Computational Intelligence, Communication Systems and Networks. doi: https://doi.org/10.1109/cicsyn.2010.56
9. Литвин, В. В., Мороз, О. В. (2013). Метод контекстного пошуку на основі тезаурусу предметної області. Східно-Європейський журнал передових

технологій, 6 (2 (66)), 22–27. URL: http://journals.uran.ua/eejet/article/view/18700/17065

10. Ranjan, N. M., Prasad, R. S. (2018). LFNN: Lion fuzzy neural network-based evolutionary model for text classification using context and sense based features. Applied Soft Computing, 71, 994–1008. doi: https://doi.org/10.1016/j.asoc.2018.07.016

11. Zhang, H., Jun, Y. (2009). An Algorithm of Text Automatic Proofreading Based on Chinese Word Segmentation. 2009 International Conference on Computational Intelligence and Software Engineering. doi: https://doi.org/10.1109/cise.2009.5364024

12. Kalinichenko, L. A. (2012). Effective support of databases with ontological dependencies: Relational languages instead of description logics. Programming and Computer Software, 38 (6), 315–326. doi: https://doi.org/10.1134/s0361768812060059

13. Garanina, N. O., Sidorova, E. A. (2015). Ontology population as algebraic information system processing based on multi-agent natural language text analysis algorithms. Programming and Computer Software, 41 (3), 140–148. doi: https://doi.org/10.1134/s0361768815030044

14. Bessmertny, I. A. (2010). Knowledge visualization based on semantic networks. Programming and Computer Software, 36 (4), 197–204. doi: https://doi.org/10.1134/s036176881004002x

15. Jorge-Botana, G., León, J. A., Olmos, R., Escudero, I. (2010). Latent Semantic Analysis Parameters for Essay Evaluation using Small-Scale Corpora*. Journal of Quantitative Linguistics, 17 (1), 1–29. doi: https://doi.org/10.1080/09296170903395890

16. Mashechkin, I. V., Petrovskiy, M. I., Popov, D. S., Tsarev, D. V. (2011). Automatic text summarization using latent semantic analysis. Programming and Computer Software, 37 (6), 299–305. doi: https://doi.org/10.1134/s0361768811060041

17. Grigoryeva, E., Klyachin, V., Pomelnikov, Y., Popov, V. (2017). Algorithm of Key Words Search Based on Graph Model of Linguistic Corpus. Vestnik Volgogradskogo Gosudarstvennogo Universiteta. Serija 2. Jazykoznanije, 16 (2), 58–67. doi: https://doi.org/10.15688/jvolsu2.2017.2.6

18. Hu, J., Li, S., Yao, Y., Yu, L., Yang, G., Hu, J. (2018). Patent Keyword Extraction Algorithm Based on Distributed Representation for Patent Classification. Entropy, 20 (2), 104. doi: https://doi.org/10.3390/e20020104

19. Kanagarajan, K., Arumugam, S. (2018). Intelligent sentence retrieval using semantic word based answer generation algorithm with cuckoo search optimization. Cluster Computing. doi: https://doi.org/10.1007/s10586-018-2054-x

20. Turney, P. D. (2000). Learning Algorithms for Keyphrase Extraction. Information Retrieval, 2 (4), 303–304. doi: https://doi.org/10.1023/A:1009976227802

21. Kulhare, S. (2017). Deep Learning for Semantic Video Understanding. A Thesis for the Degree of Master of Science in Computer Engineering. Rochester. URL: https://pdfs.semanticscholar.org/d195/9ba4637739dcc6cc6995e10fd41fd6604713.pdf

22. Ibrahim, A. S. (2017). End-To-End Text Detection Using Deep Learning. Blacksburg. URL: https://vtechworks.lib.vt.edu/handle/10919/81277

23. Lin, X. V., Wang, C., Zettlemoyer, L., Ernst, M. D. (2018). NL2Bash : A Corpus and Semantic Parser for Natural Language Interface to the Linux Operating System. International Conference on Language Resources and Evaluation. URL: https://homes.cs.washington.edu/~mernst/pubs/nl2bash-corpus-lrec2018.pdf

24. Dictionary Based Annotation at Scale with Spark, SolrTextTagger and OpenNLP. URL: https://databricks.com/session/dictionary-based-annotation-at-scale-with-spark-solrtexttagger-and-opennlp

25. Bingel, J., Bjerva, J. (2018). Cross-lingual complex word identification with multitask learning. Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications. doi: https://doi.org/10.18653/v1/w18-0518

26. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., et. al. (2014). Caffe. Proceedings of the ACM International Conference on Multimedia - MM ’14. doi: https://doi.org/10.1145/2647868.2654889

27. Hinton, G. E., Osindero, S., Teh, Y.-W. (2006). A Fast Learning Algorithm for Deep Belief Nets. Neural Computation, 18 (7), 1527–1554. doi: https://doi.org/10.1162/neco.2006.18.7.1527

28. Snowball. URL: https://snowballstem.org/

29. He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). doi: https://doi.org/10.1109/cvpr.2016.90

30. Swingler, K. Applying Neural Networks. A practical Guide. URL: http://matlab.exponenta.ru/neuralnetwork/book4/3_2.php

31. Sharipbaev, A. A., Bekmanova, G. T., Ergesh, B. J., Buribaeva, A. K., Karabalaeva, M. H. (2012). The intellectual morphological analyzer based on semantic networks. Open Semantic Technologies for Intelligent Systems.

32. Койбагаров, К. Ч., Мусабаев, Р. Р., Калимолдаев, М. Н. (2014). Разработка лингвистического процессора текстов на казахском языке. Проблемы информатики, 3, 64–72.

33. Akanova, A., Ospanova, N., Abildinova, G., Ulman, M. (2016). Assessment tools for evaluating knowledge of online students. Proceedings of the 13th International Conference Efficiency and Responsibility in Education 2016, 9–18. URL: https://erie.v2.czu.cz/en/r-13629-proceedings-2016