DOI: 10.12731/2227-930X-2018-1-106-128
УДК 004.852

КОНСТРУИРОВАНИЕ СИСТЕМЫ ИНТЕЛЛЕКТУАЛЬНОГО ПОИСКА ОТВЕТОВ НА ВОПРОСЫ ОБУЧАЮЩИХСЯ НА ОНЛАЙН-КУРСЕ НА ОСНОВЕ WORD2VEC

Рожкин П.А., Нехаев И.Н., Маркин К.А.

Целью данной работы является разработка системы интеллектуального поиска ответов на вопросы слушателей онлайн-курса среди ранее опубликованных на учебном форуме вопросов-ответов. В настоящее время уже имеются успешные эксперименты по применению систем искусственного интеллекта (IBM WATSON) в онлайн-обучении. В данной работе исследуется возможность построения такой системы с использованием технологии word2vec. Конструируется двухэтапный метод поиска ответа на вопрос с использованием технологии word2vec для векторного представления вопросов и ответов. На первом этапе определяется тематика вопроса и, если она соответствует теме форума, то среди тематических статей форума проводится поиск статей, наиболее релевантных заданному вопросу. Моделировалась реальная ситуация с 16 тематиками и 80 ответами на возможные вопросы в рамках раздела онлайн-курса "Линейная алгебра и геометрия". На основе построенной векторной модели предметной области сконструирована вопросно-ответная система и проведена оценка качества её работы. Подобраны параметры для достижения наилучшего результата классификации вопросов и поиска релевантных ответов. В 83% случаях релевантный ответ на сформулированный вопрос содержался среди топ-3 ответов, которые система предлагала. Рассматриваются вопросы дальнейшего развития применяемых подходов и повышения полезности конструируемой вопросно-ответной системы.

Цель: *разработка системы интеллектуального поиска ответов на вопросы слушателей онлайн-курса среди ранее опубликованных на учебном форуме.*

Методология: *векторизация вопросов и ответов, нейросетевая классификация тематики вопроса, построение рейтинга ответов.*

Результаты: *достижение приемлемой точности в поиске релевантного ответа на вопрос среди имеющихся ответов.*

Практическое применение: *полученные результаты исследования могут быть положены в основу конструирования интеллектуальных помощников учителя на онлайн-курсах.*

Ключевые слова: *сопрооождение обучения на онлайн-курсе, технология word2vec; векторизация вопросов; векторное пространство текстов; классификация тематики вопроса; поиск релевантных ответов*

DESIGNING AI TEACHER ASSISTANT ON ONLINE-COURSE BASED ON WORD2VEC TECHNOLOGY

Rozhkin P.A., Nekhaev I.N., Markin K.A.

The purpose of this work is to develop an AI teacher assistant, who can find answers to online course participants questions among answers previously published at the training forum. Currently, there are already successful experiments on the use of artificial intelligence systems (IBM WATSON) in online training. In this paper, we investigate the possibility of constructing such a system using word2vec technology. A two-stage method for finding an answer to a question is constructed. Method use word2vec technology for vector representation of questions and answers. At the first stage, the subject matter of the issue is determined and, if it corresponds to the theme of the forum, then the articles most relevant to the question are searched. A real situation was simulated with 16 themes and 80 answers to possible questions within the section of the online course "Linear Algebra and Geometry". The question-answer system was designed and its performance was eval-

uated. The parameters have been chosen to achieve the best result. In 83% of the cases, the relevant answer to the formulated question was contained among the top 3 responses that the system offered. The issues of further development of applied approaches and increasing utility of the constructed question-answer system are considered.

***Purpose:** developing an AI teacher assistant, who can find answers to online course participants questions among answers previously published at the training forum.*

***Methodology:** vectorization of questions and answers, neural network classification of the subject matter, construction of the answers rating.*

***Results:** acceptable accuracy in finding a relevant answer to a question are received.*

***Practical implications:** The results of the research can be used as a basis for designing an AI teacher assistant in online courses.*

***Keywords:** online course, e-learning support; word2vec technology; vectorization of questions; vector space of texts; classification of subject matter; search for relevant answers.*

Введение

С развитием интернет-технологий, технологий виртуальной и дополненной реальности онлайн-обучение, онлайн-курсы становятся привлекательными для миллионов обучающихся. Одной из важных составляющих обеспечения качества процесса обучения на массовых онлайн-курсах является технология сопровождения обучения. Чаще всего для организации сопровождения на массовых онлайн-курсах используются форумы. Удобство использования форума для массовых онлайн-курсов заключается в возможности опубликовать свой вопрос и получить ответ на него от других слушателей курса. Минус заключается в том, что каждый публикуемый вопрос требует ответа, хотя на самом деле вопрос может быть и не нов и ответ на него, как правило, уже имеется на форуме. Стандартный поиск форума работает слишком просто, чтобы найти похожий вопрос, если он не сформулирован такими же словами. Необходим более удобное средство поиска дискуссий

на форуме, в соответствии с темой, затронутой в вопросе. В 2016 году подобный помощник учителя на онлайн-курсе был создан с использованием технологии IBM WATSON в George Tech. [7]. Однако это единичный случай и пока нет информации для обобщения и тиражирования данного опыта. Требуется исследование, которое могло бы показать действенность существующих методов машинного обучения и искусственного интеллекта для решения задачи создания помощника учителя на онлайн-курсе.

При построении вопросно-ответных систем обычно используются технологии анализа потока слов Seq2Seq, преобразующие входную последовательность слов вопроса в выходную последовательность слов ответа [6]. Показано, что Seq2Seq улучшают качество своей работы на длинных последовательностях [5, 8]. Поэтому для решения поставленной задачи построения вопросно-ответной системы предлагается использовать рекуррентные нейронные сети RNN LSTM. В условиях онлайн-курса вопросы на форуме часто содержат короткие последовательности слов. Поэтому для поиска ответов на вопросы обучающихся более подходят технологии word embedding в сочетании с методами классификации.

Решить задачу выбора ответа на запросы слушателей курса по определенной тематике значительно сложнее, чем отфильтровать спам или сортировать электронную почту по папкам. Чтобы обучить классификатор необходимо определить возможные ответы на возможные вопросы. Но надо понимать, что даже при ограниченной тематике на онлайн-курсе всевозможных формулировок может быть очень много.

Также необходимо учесть возможность того, что вопрос может быть сформулирован не по теме, быть некорректным или недоопределенным. Т.е. нужны: а) мягкие способы оценки релевантности возможных вариантов ответа, б) необходим предварительный анализ потенциальной возможности подбора ответа и привлечение экспертов-тьюторов, помощников, сопровождающих процесс обучения на курсе в случае невозможности подобрать требуемый ответ.

В настоящее время существует несколько основных подходов к классификации текстов [1].

1. Методы на основе “мешка слов”, например TF-IDF [3] или BM-25 [15]. Предполагается, что значимость n-граммы для определения тематики текста прямо пропорциональна частоте ее появления в тексте и обратно пропорциональна доле текстов, в которых эта n-грамма встречается. Она может нормализоваться, ограничиваться сверху, чтобы избежать присваивания слову слишком большого веса [11].
2. В методах тематического моделирования, например LSA [13], NMF [16] строятся модели текстов в виде вектора принадлежности его различным тематикам и используются методы для понижения размерности векторов тем и векторов текстов с целью максимизировать выбранный показатель оптимальности их представления. В LDA [4], для построения моделей текстов и тем используется предположение о случайном распределении векторов тем и векторов документов.
3. Многочисленные методы на основе Word embeddings[14]. Например, continuous Bag-of-Word (CBOW) использует окружающие слова, но предсказания не зависят от порядка этих слов или skip-gram, который предсказывает окружающие слова, основанные на текущем слове.

В данной работе исследуется применимость одной из известных технологий векторного представления слов Word2Vec, как одна из наиболее распространенных эффективных технологий анализа тематики текста. Технология основана на векторном представлении слов, словосочетаний и самого текста, использует для представления слова его окружение. Таким образом, возможная тематика текста, вопроса определяется используемыми словосочетаниями и предложениями, а не отдельными словами, которые могут встречаться в разных тематиках. Применение данного подхода дает гибкость при анализе неопределенности задаваемого вопроса непосредственно при его формулировании. Это означает возможность построения диалога с мгновенной обратной связью

для формулирующего вопрос слушателя даже без подбора конкретного ответа на него.

План работы

1. Выбор тематик и подготовка соответствующего корпуса текста для моделирования ситуации обучения и проведения исследования
2. Конструирование алгоритма классификации вопросов и ответов по тематикам
3. Исследование различных моделей классификации тематики вопроса и поиска ответа на вопрос.
4. Анализ результатов и выводы.

Моделирование ситуации обучения

Для исследования применимости разрабатываемой модели необходимо было смоделировать реальную ситуацию обучения на онлайн-курсе. В качестве такого курса был выбран курс “Линейная алгебра и геометрия. Часть 2: векторная алгебра”. Т.е. содержательная область была ограничена одним разделом дисциплины. При этом, в соответствии с лекционным материалом было выделено две большие темы (см. рис.1) и 16 подтем:

1: 'Определение поля чисел',

2: 'Определение векторного пространства над полем чисел',

3: 'Линейная зависимость векторов',

4: 'Базис и размерность векторного пространства',

5: 'Разложение векторов по базису',

6: 'Переход от одного базиса к другому',

7: 'Изоморфизм векторных пространств',

8: 'Использованием векторного представления объектов для оценки их близости',

9: 'Метрика векторного пространства',

10: 'Норма векторного пространства',

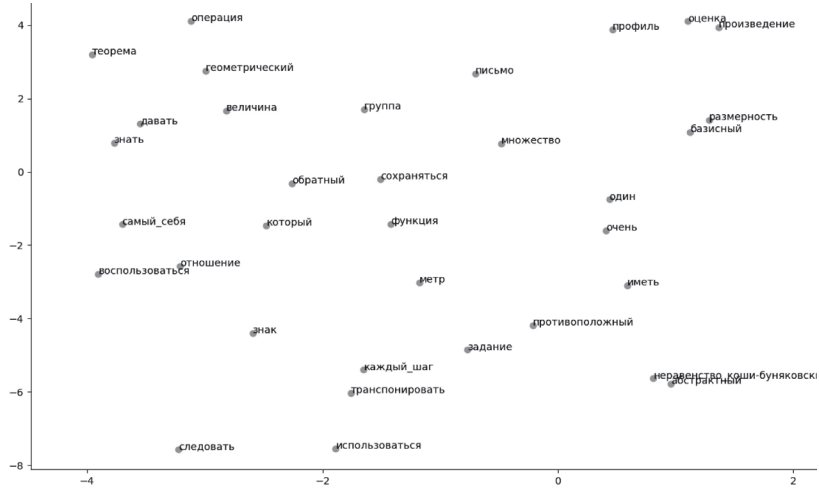
11: 'Скалярное произведение',

12: 'Вычисление скалярного произведения',

13: 'Угол между векторами',

14: 'Ортонормированный базис',
15: 'Процедура ортогонализации',
16: 'Решение задач классификации'.

Для каждой подтемы был выделен текст из курса лекций – одна – три страницы (от 500 до 2500 слов). Текст каждой темы был поделен на смысловые части – возможные ответы на вопросы курса. Всего было выделено 80 частей, в среднем 5 частей на одну подтему.



Исследования были направлены на то, чтобы сконструировать систему классификации, которая бы относила вопросы, которые могли быть сформированы (и формировались на реальном курсе), к той или иной тематике из представленных 16 или к неизвестной тематике, выходящей за рамки рассмотренных.

Кроме этого, исследовалась и возможность применения данной системы векторизации текста для того, чтобы находить наиболее релевантные к сформулированному вопросу текстовые абзацы из соответствующих наиболее релевантных тематик.

Построение алгоритма определения тематики вопроса

Так как цель исследования – разработать интеллектуальную систему сопровождения обучения на онлайн-курсе, а именно, поиска

релевантного ответа на возникающие вопросы студентов на форумах, то необходимо научиться обрабатывать входные вопросы, заданные пользователем на естественном языке. Задача обработки текстов на естественном языке входит в группу задач изучаемых направлением NLP (*Natural Language Processing*).

Основными этапами обработки естественного языка с использованием машинного обучения являются:

1) Лексический анализ текста
2) Векторизация текста
3) Семантический анализ текста

Именно различная реализация данных этапов и определяет качество всей системы обработки естественного языка в целом. Более подробно алгоритм вопросно-ответной системы обработки можно представить так:

1) Ввод пользователем текстового вопроса по теме онлайн-курса на естественном языке
2) Проведение лексической обработки (разбиение на конструкции и слова, а также их обработка – токенизация и лемматизация)
3) Перевод текстовых данных в математическую модель (векторизация с использованием модели word2vec)
4) Семантический анализ обработанного вопроса (классификация темы вопроса и поиск релевантных ответов)
5) Выдача пользователю ответов в виде топ-3 ответов наиболее релевантных заданному вопросу.

Рассмотрим более детально реализацию каждого из выделенных этапов алгоритма и результаты исследования сконструированной системы.

Исследование влияния различных способов лексического анализа на точность классификации.

Лексический анализ делится на два подпункта:

- токенизацию – разбиение конструкций на токены (слова, разделители и т.д.);
- анализ и обработку токенов.

Рассмотрены два варианта системы:

- система без обработки полученных токенов;
- система с обработкой токенов.

В первом случае вопросно-ответная система разбивала конструкцию на токены и пыталась их векторизовать без применения какой-либо обработки, во втором же случае система проводила анализ и обработку токенов, а именно:

- устанавливала теги на токены, определяющие грамматическую и частеречную принадлежность
- на основании частеречной принадлежности проводила исключение токенов, не несущих смысловую для системы нагрузку (союзов, разделителей и цифр)
- нормализовала каждый токен, который не был исключен. В качестве нормальной формы был использован именительный падеж, единственное число.

Система, которая не включала в себя обработку, показала себя значительно хуже. Размер ее векторного пространства составлял 337 векторов-слов, против 260 у системы, которая имела обработку, а точность распознавания на контрольной выборке из 15 вопросов стремилась к 0 (против 83%), при этом большинство вопросов даже не могли быть переведены в вектора. Это означает, что в векторном пространстве системы имелось около 100 векторов-слов, которые: а) просто повторяли друг друга (имели разные формы), б) не несли никакой смысловой нагрузки, входя во множества тем и искажая их классификацию. Если бы пользователь работал с такой системой, то ему приходилось бы использовать жесткие формы вопросов и при этом в большинстве случаев они были бы неверно распознаны.

В целом, анализ и обработка на основе тегирования позволяют уменьшить размерность векторного пространства слов системы и увеличить точность семантического анализа. Их наличие обязательно для вопросно-ответных систем.

Также стоит отметить, что существуют и более простые способы обработки токенов, например вычленение основы слова, но такой метод не учитывает частеречную принадлежность и две разные части речи могут быть восприняты как одна.

Исследование влияния параметров технологии Word2Vec при векторизации слов и словосочетаний на точность классификации.

Векторизация слов является переходным процессом из текстовой информации, которую понимает человек, в числовую информацию (в данном случае в вектора) с которой работают математические модели. Именно по этой причине методы используемые для векторизации должны максимально эффективно переводить текстовую информацию в математическую модель.

В работе применяется метод CBOW. Его преимущество заключается в том, что при создании векторного пространства (ВП) происходит учет контекстной близости слов, т.е. слова, встречающиеся в тексте рядом с одинаковыми словами (а следовательно, имеющие общий смысл), будет иметь близкие координаты векторов-слов. Это преимущество будет являться значимым плюсом для семантического анализа.

В качестве исходных данных для создания векторного пространства системы использован прошедший этап лексического анализа корпус текста состоящий из 80 абзацев тем. Инструмент «word2vec» позволяет задавать набор базовых параметров при создании ВП. Основные параметры, которые следует настроить, это:

- «size» – размерность векторного пространства,
- «min_count» – частота (минимальное количество повторений слова) в корпусе текста для включения его в векторное пространство
- «window» – максимальное расстояние между словами для создания контекстной близости слов

При исследовании рассматривались варианты с 50, 200 и 400 размерными векторными пространствами. В первом случае размерность является слишком малой, чтобы включать в себя обширный объем имеющихся тем и система не могла определить верную тему для большинства вопросов контрольной выборки (см. ниже табл. 1) даже с 30% вероятностью. В последнем случае система практически не изменила точность классификации по сравнению с размерностью равной 200, но несла дополнительную нагрузку

на вычислительную систему, т.к. надо было вычислять вектор размерностью в два раза больше.

Также использовалось разные значения порога «min_count» для принятия решения о добавлении слова в ВП. В связи с тем, что изначальная выборка по некоторым темам была не велика, оптимальным значением порога для данного корпуса стало 3. При пороговом значении 5 система векторизировала не все значимые для поиска релевантных подтем слова и ошибка классификации возрастала.

Оптимальное значение параметров для реализуемой системы оказались такими:

- «size» = 200
- «min_count» = 3
- «window» = 5.

Для анализируемой вопросно-ответной системы (см. пример на рис. 1) векторное пространство (ВП) слов содержит 260 слов и биграмм слов, на основе которых будет рассчитываться вектор языковой конструкции (в данном случае пользовательский вопрос).

Для расчета вектора вопроса – входной текстовой конструкции – использовалось средневзвешенное значение векторов-слов, т.е. математическое выражение вида:

$$vec = \frac{\sum_{i=1}^{count\_word} vec\_word_i}{count\_word}, \quad (1)$$

где vec – вектор конструкции; vecword(i) – вектора слов в конструкции, которые прошли векторизацию, count(word) – кол-во слов конструкции, которые были преобразованы в вектора.

После вычисления вектора конструкции этап векторизации заканчивается и начинается семантический анализ полученного вектора. Семантический анализ - классификация и ранжирование

Для самой классификации тем принято решение использовать отдельную нейросетевую модель для классификации каждой темы, т.к. это позволяет не переобучать всю модель при добавление новых тем, а лишь добавлять новые, тем самым экономя

время системы. Кроме того, исследование подобных систем классификации электронных писем показало, что точность при этом классификации значительно повышается. Конечно, в случае если система начала неверно классифицировать темы или же слова новой темы не находятся в векторном пространстве, то необходимо или пересоздать векторное пространство слов и/или переобучить все модели, а это затратно в вычислительном плане. Тем не менее, выгоды в повышении точности работы системы налицо.

Сами же модели каждой из тем строятся по единым параметрам. Количество нейронов на входе каждой нейросети равняется размерности векторного пространства и составляет 200 нейронов (для данного случая). Выходной слой содержит один нейрон, который определяет вероятность принадлежности заданного вопроса данной теме. Помимо этого, имеется два скрытых слоя, количество нейронов в первом слое 130, во втором 70. В качестве функции активации для входного и скрытых слоев выбран «relu» (выпрямитель), т.к. он менее ресурсоемкий, в отличие от той же сигмоиды, и обеспечивает повышенную скорость сходимости стохастического градиента функции. Для выходного слоя функцией активации является «sigmoid». Использовался алгоритм оптимизации модели – «adam».

Исследование результатов классификации тематики и рейтинга возможных ответов в зависимости от точности и возможных вариаций формулировки вопроса.

Само исследование сконструированной системы поиска ответов состояло в следующем. Система обучалась на выбранном корпусе текста лекций по двум темам, который был сегментирован на 80 текстовых частей, содержащих ответы на возможные вопросы студентов в рамках данных тем и классифицированных по 16 подтемам. От имени студента формулировались возможные вопросы и задавались системе. Система должна была установить вероятность того, что данный вопрос: а) принадлежал данной подтеме б) выбрать топ возможных подтем в) среди ответов из выбранных подтем определить топ-3 наиболее релевантных.

Для оценки качества работы системы классификации вопросов по тематикам (подтемам) рассматривались следующие показатели:

1. Процент верной классификация тематики вопроса; под верной классификацией считаем такой ответ системы, при котором номер указанной экспертом темы имеет самую большую вероятность при ответе системы.
2. Процент верной классификация тематики вопроса; под верной классификацией считаем такой ответ системы, при котором номер указанной экспертом темы содержится в списке наиболее вероятных тем при ответе системы (не более трех тем).

Для оценки данных показателей взята выборка из 15 вопросов (табл. 1), которые не входили в обучающую выборку (для каждой подтемы были сформулированы вопросы, на которые можно было найти ответы в подтеме), и каждому вопросу эксперт сопоставил подтему или список из двух подтем, которые содержали ответы на заданный вопрос. По результатам работы системы (табл.1) оценивалась точность классификации системы.

Таблица 1.

Оценка классификации вопросов

№	Вопрос	Темы (Эксперт)	Темы (Система)
1	Что такое поле чисел	1	1 – 100% 2 – 92%
2	Как вычислить угол между векторами	13	13 –100%
3	Как найти коэффициенты разложения вектора по базису	5,6	6 – 100% 4 – 95%
4	Как выполнить фильтрацию спама	16	16 – 45% 8 – 100% 14 – 50%
5	Как построить матрицу перехода от одного базиса к другому	6	6 – 100%
6	процедура ортогонализации Грама-Шмидта	14	14 – 48% 16 – 44%
7	как найти координаты вектора в ортонормированном базисе	14,15	4 – 68% 6 – 100%
8	как найти скалярное произведения векторов	11,12	11 – 47% 12 – 49% 14 – 50%

Окончание табл. 1.

9	Что такое изоморфизм векторов	7	7 – 100% 4 – 98%
10	линейная зависимость векторов и ее применение	3	3 – 100%
11	методы классификации объектов	8	8 – 100% 7 – 98%
12	Определение векторного пространства	2	3 – 100% 6 – 66%
13	Что означает линейная зависимость векторов	3	3 – 100%
14	Как определить линейную зависимость векторов	3	3 – 100%
15	Как вычислить след матрицы	0	3 – 49% 6 – 100%

Если смотреть по 1-му показателю, то точность классификации составила: 100*9/15 % = 60%

Если смотреть по 2-му показателю, то точность классификации составила: 100*12/15 % = 80%

Из 15 вопросов было верно классифицировано 12, что является обнадеживающим результатом. Также стоит отметить, что на некоторые вопросы система может дать более развернутый ответ, чем эксперт. Например, вопрос №4 относится не только к теме 16, в которой конкретно рассматривается ответ на вопрос, но и к темам 8 и 14, содержащие необходимые для понимания ответа концепты. Конечно, с этой задачей может справиться и эксперт, но для этого надо анализировать всю структуру знаний, используемую в данной дисциплине.

Вопрос №7 не был распознан верно, скорее всего, из-за того, что система больше обратила внимание на окружение таких понятий как “координаты”, “вектора” и “базис”, чем на “ортонормированный”. Справиться с данной проблемой может помочь учет в модели текстовой конструкции (1) векторов биграмм (словосочетаний), а также взвешивание важных для каждой тематики слов. В этом случае система может обратить внимание именно на словосочетание «ортонормированный базис», чем просто на базис.

Независимо от имеющихся недостатков работы системы, можно сказать, что применяемая технология справилась с поставленной задачей классификации тематики сообщений. Это было показано и ранее, когда демонстрировалась способность технологии верно классифицировать электронные сообщения. При количестве имеющихся сообщений, равном 3000 и выделении 4-х тематик сообщений личной почты, система верно классифицировала письма с точностью 97%. При классификации 4450 писем на 5 тематик из почты учреждения, была достигнута точность в 90% [2].

Но для решения задачи поддержки онлайн-обучения на курсе мало обеспечить верную классификацию тем, необходимо еще и дать ответ на поставленный вопрос, если он имеется в материалах курса. Например, в случае верной классификации темы вопроса система должна была бы еще найти релевантные ответы на форуме, относящиеся к данной теме. Для исследования данной возможности выполнялось ранжирование имеющихся текстов по данной тематике и отбор топ-3. Здесь использовались следующие показатели качества работы системы:

1. Процент удачного поиска релевантного ответа на вопрос, когда релевантный ответ находится на 1-м месте в рейтинге выдаваемых ответов;
2. Процент удачного поиска релевантного ответа на вопрос, когда релевантный ответ находится в тройке лучших в рейтинге выдаваемых ответов.

Для оценки данных показателей используем те же самые вопросы, но теперь система должна найти наиболее близкие ответы для него из тех, которые соответствуют выявленной тематике. Также искались ответы только на те вопросы, тематику которых система классифицировала верно. Это позволит понять, на сколько качественно система могла бы искать ответы пользователей среди имеющихся в базе ответов. Всего в экспериментальной базе содержалось 80 ответов и ранжирование проводилось только тех ответов, которые содержались в данной теме.

Для ранжирования использовалось обыкновенное эвклидово расстояние по формуле (2) между вектором вопроса и векторами ответов, которые формировались по одному и тому же алгоритму усреднения (1).

$$d(x.y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}, \quad (2)$$

Здесь x-вектор вопроса, y-вектор ответа, n-размерность векторного пространства.

Таблица 2.

Результаты ранжирования имеющихся ответов

№ вопроса	Номер ответа от эксперта	Номер ответа от системы
1	1	2,1
2	60	62,59,60
3	17	31,25,23
4	78	78,80,77
5	24	33,24,26
6	63	63,69,66
8	58	58,55,56
9	36	36,35,38
10	9	9,8,12
11	41	49,43,47
13	9,8	9,12,8
14	9	9,12,8

Если смотреть по 1-му показателю, то точность работы системы составила: 100*7/12 % = 58%

Если смотреть по 2-му показателю, то точность работы системы составила: 100*10/12 % = 83%

Из полученных данных можно предположить, что если система нашла нужную тему, то она с высокой точностью найдет наиболее релевантный ответ среди ближайших 3-х ответов.

Прежде чем сформулировать выводы по работе системы рассмотрим пример ответа системы на вопрос №3 – “Как найти коэффициенты разложения вектора по базису”. Рассмотрим ответы

системы (31,25,23). Наименьшее расстояние (0.37) имеет следующий ответ (№31):

"Теперь рассмотрим произвольный вектор b. Пусть нам известны его координаты в старом базисе. Будем обозначать координаты вектора b в старом базисе через b_e, а его координаты в новом базисе через b_f. Применим рассмотренную ранее общую схему и составим систему уравнений, связывающую эти координаты. Видим, что, задав матрицу T, матрицу перехода от нового базиса к старому мы можем найти новые координаты, решив систему уравнений с этой матрицей как с матрицей коэффициентов и с правой частью, равной координатам вектора в старом базисе.'"

Данный ответ можно назвать релевантным к вопросу, если внести дополнение в поставленный вопрос «при переходе от одного базиса к другому». Также релевантными можно считать ответы 25 и 23, в которых приводятся примеры расчета координат векторов-функций в базисе степенных функций. Ответа же о том, что такое координаты вектора не оказалось в нужной теме. Он по ошибке был оставлен в теме 4 «Базис и размерность векторного пространства» вместо более уместной темы 5 «Разложение векторов по базису». Прямого же ответа о том, как можно найти координаты вектора, не оказалось совсем в тексте лекции, т.к. подразумевалось, что слушатель знает как находить коэффициенты разложения вектора по системе линейно-независимых векторов. Таким образом, система может помочь найти неточности и пробелы в работе преподавателей, экспертов, если таковые возникнут.

Заключение

Разрабатываемая система помогает найти ответы на большинство рассмотренных вопросов, которые дадут релевантную информацию пользователю. Технология word2vec может быть использована при построении интеллектуальных систем сопровождения обучения на онлайн-курсах. Тем не менее, до внедрения системы в реальные онлайн-курсы остается еще много нерешенных вопросов. Перечислим основные из них.

Как система может определить, что задаваемый вопрос не относится к тематикам данного форума/раздела? Наши попытки задать нерелевантный вопрос привели к парадоксальным результатам – система уверенно относила его к какой-то тематике, если были использованы какие-то понятия или даже просто часто употребляемые слова из данной тематики. Здесь мы приходим к необходимости введения новой темы №0, к которой надо отнести все нерелевантные темы, по которым необходимо обучать систему. Сразу возникнет вопрос об объеме образцов темы 0 и возникнет вопрос о необходимости создания предварительного классификатора и введения двух этапов процедуры классификации тематики вопроса. На первом этапе происходит фильтрация нерелевантных сообщений, а на втором этапе уточнение тематики вопроса.

Как должны меняться параметры процедуры векторизации, векторной модели при увеличении числа тематик и количества ответов? С ростом числа вопросов неминуемо будет рост числа ответов и, возможно, расширяться-уточняться перечень тематик. Надо ли будет адаптировать векторную модель предметной области или достаточно просто переобучать классификаторы с учетом новых образцов?

Как сделать систему корректирующей ошибки экспертов, в том числе и выявляющей нерелевантные ответы на формулируемые вопросы, или отсутствие ответов на возникающие у студентов вопросы?

Как учесть сокращения, синонимы, а также новые слова в вопросах, которые не появлялись ранее в опросах и ответах и не вошли в состав векторной модели?

Как сделать интеллектуального помощника постоянно обучающимся с использованием обратной связи от обучающихся?

Эти вопросы не новы для всех, кто внедряет системы тематического моделирования документов, поиска релевантных документов. Есть много примеров решения подобных вопросов [1, 9, 10, 12]. Надеемся, что внедрение данной системы в онлайн-курсы портала онлайн-образования Поволжского регионального центра компетенций в области онлайн-обучения и новые исследования позволят найти удовлетворительные ответы на поставленные вопросы.

Список литературы

1. Пархоменко П.А., Григорьев А.А., Астраханцев Н.А. Обзор и экспериментальное сравнение методов кластеризации текстов. Труды ИСП РАН, 2017 г., том 29, вып. 2. С. 161–200.
2. Ширяев А.И., Нехаев И.Н. Исследование применимости технологии word2vec для решения задачи классификации электронных почтовых сообщений клиентов // Научному прогрессу – творчество молодых. Материалы XII международной молодежной научной конференции по естественнонаучным и техническим дисциплинам. Йошкар-Ола, 21–22 апреля 2017 года. Часть 3. Йошкар-Ола, ПГТУ. С. 114–116.
3. Aggarwal Charu C, Zhai Cheng Xiang. Mining text data. Springer Science & Business Media, 2012.
4. Blei David M., Ng Andrew Y., Jordan Michael I. Latent dirichlet allocation. Journal of machine Learning research. 2003. Т. 3, № Jan., pp. 993–1022.
5. Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473, 2014.
6. Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In Advances in neural information processing systems, pages 3104–3112, 2014.
7. Jason Maderer, Artificial Intelligence Course Creates AI Teaching Assistant, |MAY 9, 2016, ATLANTA,GA.URL: http://www.news.gatech.edu/2016/05/09/artificial-intelligence-course-creates-ai-teaching-assistant (дата обращения: 27.03.2018).
8. Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. arXiv preprint arXiv:1508.04025,2015.
9. Chawla N.V., Bowyer K.W., Hall L.O., Kegelmeyer W.P. SMOTE: synthetic minority over-sampling technique // Journal of artificial intelligence research. 2002, pp. 321–357.
10. Rehurek R., Sojka P. Software Framework for Topic Modelling with Large Corpora // Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. 2010, pp. 45–50.

11. Salton Gerard, Buckley Christopher. Termweighting approaches in automatic text retrieval. Information processing & management. 1988. T. 24, № 5, pp. 513–523.
12. Sanjeev Arora, Yingyu Liang, Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings // ICLR. 2017.
13. Scott Deerwester, Susan T Dumais, George W Furnas et al. Indexing by latent semantic analysis. Journal of the American society for information science. 1990. T. 41, № 6, pp. 391.
14. Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean, Efficient estimation of word representations in vector space // ICLR Workshop. 2013.
15. Whissell John S., Clarke Charles L.A. Improving document clustering using Okapi BM25 feature weighting. Information retrieval. 2011. T. 14, № 5, pp. 466–487.
16. Xu Wei, Liu Xin, Gong Yihong. Document clustering based on non-negative matrix factorization. Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval. ACM. 2003, pp. 267–273.

References

1. Parhomenko P.A., Grigor'ev A.A., Astrahancev N.A. Obzor i jeksperimental'noe sravnenie metodov klasterizacii tekstov [Review and experimental comparison of text clustering methods]. *Trudy ISP RAN*, 2017. Vol 29, no. 2, pp. 161–200.
2. Shirjaev A. I., Nehaev I. N. *Issledovanie primenimosti tehnologii word-2vec dlja reshenija zadachi klassifikacii jelektronnyh pochtovyh soobshhenij klientov* [The study of the applicability of word2vec technology to solve the problem of classification of e-mail]. Yoshkar-Ola. 2017. No 3, pp. 114–116.
3. Aggarwal Charu C, Zhai Cheng Xiang. Mining text data. Springer Science & Business Media, 2012.
4. Blei David M., Ng Andrew Y., Jordan Michael I. Latent dirichlet allocation. *Journal of machine Learning research*. 2003. Vol. 3, no, pp. 993–1022.

5. Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473,2014.
6. Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *In Advances in neural information processing systems*. 20140, pp. 3104–3112.
7. Jason Maderer, Artificial Intelligence Course Creates AI Teaching Assistant. ATLANTA. 2016. GA.URL: http://www.news.gatech.edu/2016/05/09/artificial-intelligence-course-creates-ai-teaching-assistant (date of access: 27.03.2018).
8. Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. arXiv preprint arXiv:1508.04025,2015.
9. Chawla N.V., Bowyer K.W., Hall L.O., Kegelmeyer W.P. SMOTE: synthetic minority over-sampling technique // *Journal of artificial intelligence research.* 2002, pp. 321–357.
10. Rehurek R., Sojka P. Software Framework for Topic Modelling with Large Corpora. Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. 2010, pp. 45–50.
11. Salton Gerard, Buckley Christopher. Termweighting approaches in automatic text retrieval. *Information processing & management*. 1988. Vol. 24, no 5, pp. 513–523.
12. Sanjeev Arora, Yingyu Liang, Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. ICLR Workshop. 2017.
13. Scott Deerwester, Susan T Dumais, George W Furnas et al. Indexing by latent semantic analysis. *Journal of the American society for information science*. 1990. Vol. 41, no 6, pp. 391.
14. Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean. Efficient estimation of word representations in vector space. ICLR Workshop. 2013.
15. Whissell John S, Clarke Charles LA. Improving document clustering using Okapi BM25 feature weighting. *Information retrieval*. 2011. Vol. 14, no 5, pp. 466–487.
16. Xu Wei, Liu Xin, Gong Yihong. Document clustering based on non-negative matrix factorization. Proceedings of the 26th annual international

ACM SIGIR conference on Research and development in information retrieval. ACM. 2003, pp. 267–273.

ДАННЫЕ ОБ АВТОРАХ

Рожкин Павел Александрович, студент кафедры информационно-вычислительных систем
Поволжский государственный технологический университет
ул. Панфилова, 17, г. Йошкар-Ола, Республика Марий Эл, 424006, Российская Федерация
blackiiifox@gmail.com

Нехаев Игорь Николаевич, доцент кафедры прикладной математики и информационных технологий, кандидат технических наук
Поволжский государственный технологический университет
ул. Панфилова, 17, г. Йошкар-Ола, Республика Марий Эл, 424006, Российская Федерация
nehaevin@volgatech.net

Маркин Кирилл Анатольевич, студент кафедры информационно-вычислительных систем
Поволжский государственный технологический университет
ул. Панфилова, 17, г. Йошкар-Ола, Республика Марий Эл, 424006, Российская Федерация
kirill1997_markin@mail.ru

DATA ABOUT THE AUTHORS

Rozhkin Pavel Aleksandrovich, Student of the Department of Information and Computing Systems
Volga State University of Technology
17, Panfilov Str., Yoshkar-Ola, Mari El Republic, 424006, Russian Federation
blackiiifox@gmail.com

Nekhaev Igor Nikolaevich, Associate Professor of Applied Math and Informational Technologies Cathedra, PhD in Technical Sciences
Volga State University of Technology
17, Panfilov Str., Yoshkar-Ola, Mari El Republic, 424006, Russian Federation
nehaevin@volgatech.net
SPIN-code: 4137-5250

Markin Kirill Anatol'evich, Student of the Department of Information and Computing Systems
Volga State University of Technology
17, Panfilov Str., Yoshkar-Ola, Mari El Republic, 424006, Russian Federation
kirill1997_markin@mail.ru