Насонов Евгений Александрович
Харьковский национальный университет радиоэлектроники

МЕТОДЫ ОБРАБОТКИ БОЛЬШОГО КОЛИЧЕСТВА КУЛИНАРНЫХ РЕЦЕПТОВ УЧИТЫВАЮЩИЕ СЕМАНТИЧЕСКУЮ БЛИЗОСТЬ ИНГРЕДИЕНТОВ

Summary. Methods of processing of big food recipe data are considered in this article. Specifics of Russian language text processing are considered when it comes to food recipes. A model food recipe vector representation that considers semantic relationships is developed. A brief survey of available software implementations of Russian language text processing algorithms is given.

Аннотация. В статье рассматриваются методы работы с большим количеством пользовательских кулинарных рецептов и их анализа. Рассматриваются методы работы с тексом на русском языке в применении к конкретной предметной области. Разработана модель векторного представления кулинарных рецептов, учитывающий семантическую близость ингредиентов. Также проводится обзор доступных программных инструментов для решения задач обработки текстов.

Key words: big data, food recipes, semantic relationships, machine learning, natural language processing.

Ключевые слова: большие данные, семантические отношения, кулинарные рецепты, машинное обучение, обработка естественных языков.

Постановка проблемы. Эффективная автоматизированная обработка кулинарных рецептов является достаточно сложной задачей, с которая представляет одновременно как теоретический интерес в области обработки естественных языков, машинного обучения без учителя, а более точно алгоритмов кластеризации, так и практический интерес проектирования и построения веб-систем, которые могли бы производить качественную обработку больших объемов данных и отвечать за пользовательские запросы; также правильный анализ данных позволяет выполнять различные процессы, которые помогают повысить релевантность веб-системы с помощью SEO (Search Engine Optimization) [1]. В случае малоэффективного разрешения задачи обработки кулинарных рецептов веб-система может игнорироваться современными поисковыми системами. Даже попав на такой сайт пользователь столкнется с трудностями вызванными большой вариацией классических кулинарных блюд и ему будет сложно найти искомую информацию среди большого количества плохо сгруппированного проанализированного во многом повторяющегося материала. Непосредственная проблема обработки рецептов кулинарных блюд в первую очередь осложняется классическими задачами обработки естественного языка. Так, необходимо уметь выделять ключевые сущности текста для определения ингредиентов и их значимость в конкретном блюде. Далее, большие данные, которые необходимы для алгоритмов машинного обучения, зачастую приходится собирать в различных источниках пользовательских рецептов, которые, к сожалению, могут содержать разные описки, использовать неологизмы, употреблять синонимы местного вернакуляра. Несомненно сложной задачей является разрешение кореференций, с помощью чего можно было бы более эффективно определять значимость ингредиента. Далее перед нами становится задача удобного представления данных таким образом, чтобы можно было применять различные алгоритмы машинного обучения, а также подбор самих алгоритмов опираясь на специфику данных, их размер, а также характеристики модели представления данных.

Анализ последних исследования и публикаций. Были исследованы методы обработки текстов на естественном языке. В частности, были изучены методы таких сложных задач, как разрешение кореференций [2]; на основании чего можно оценить фактическую сложность задачи. Стоит выделить. Был выполнен сравнительный анализ корпусов русского и английского языков, а также инструментов их обработки, в результате чего можно сказать, что готовые программные решения задач обработки русского языка уступают английскому языку. Среди исследованных инструментов стоит выделить морфологический анализатор pymorphy2 [3], который имеет удобный программный интерфейс на языке Python, а также довольно быстрое время выполнения ключевых функций, таких как лемматизация, которая зачастую является одним из первичных этапов обработки текста и в случае работы с большими объемами данных кулинарных рецептов скорость выполнения таких базовых операций несомненно важна. К сожалению, pymorphy2 не позволяет выполнять контекстную лемматизацию, что потенциально могло бы привести к улучшению точности результатов дальнейших операций с текстом. Для выполнения контекстной лемматизации был испробован инструмент PyMystem3 от компании Yandex, однако скорость его работы оказалась недостаточной для решения задач машинного обучения связанных с большими данными. Также для работы с выделением ключевых сущностей, в нашем случае ингредиентов, помимо разрешения кореференций, нужно уметь работать с синонимами и словосочетаниями. Для работы с синонимами достаточно использовать тезаурус. Работу со словосочетаниями проводить значительно сложнее. Можно использовать словари устойчивых словосочетаний русского языка, однако на

эффективность такого подхода в решении задач обработки кулинарных рецептов надеяться не приходится, т.к. такие словари не учитывают интересующие нас словосочетания-ингредиенты. Более эффективные методы используют статистический анализ биграмм; такой метод получил развитие в статье [4], где были выделены и проанализированны основные метрики статистической сочетаемости лексических единиц, такие как MI, LL, MI3, MS, t-score, а также рассмотрены требования к программному обеспечению, которые актуальны и на данный момент; отмечается нехватка мощных и вариативных инструментов для обработки текста на русском языке. Похожий подход был описан в [5], где задача была рассмотрена непосредственно для имен существительных, что представляет наибольший интерес для работы с кулинарными рецептами, а ещё позднее идеи получили дальнейшее развитие в сторону машинного обучения [6]. Для использования рассмотренных квантитативных подходов несомненно необходимо работать с большими корпусами, во многих случаях желательно иметь размеченные корпусы; удобный программный интерфейс к таковым также необходимы для построения программных систем; в данной статье рассматривался открытый корпус русского языка OpenCorpora.

К вопросу кластеризации кулинарных рецептов обращались в [7], что позволило получить представления о непосредственном опыте работы с кулинарными рецептами, а не с общеязыковыми корпусами. Подход статьи сложно обобщить, т.к. автор производил работу с достаточно специфичным и ограниченным набором данных, в связи с чем были облегчены задачи обработки естественного языка; более того, рассмотренный корпус не является достаточно большим, чтобы его можно было сравнить с теми объемами данных, которые встречаются в промышленных задачах с сотнями тысяч неупорядоченных пользовательских рецептов. Тем не менее, применение методов кластеризации к анализу данных представляет особый интерес.

Выделение нерешенных частей проблемы. Таким образом, можно заключить недостаточность освоения алгоритмов обработки текстов на русском языке в специфичных областях, в частности для обработки кулинарных рецептов. Было отмечено наличие корпусов общего назначения, однако открытым остается вопрос высококачественного источника данных рецептов блюд. Открытым остается вопрос специализированного подхода извлечения полезной информации из текстов кулинарных рецептов. Также была отмечена сложность задачи разрешения кореференций, эффективное и доступное решение которой несомненно поможет лучше справляться с этапом предобработки многих актуальных задач. Существует не так много программных инструментов для выполнения рассмотренных задач, и не все из них имеют современный удобный интерфейс. Были выделены подходы определения коллокаций в тексте на основе биграмм и алгоритмов машинного обучения. Однако остаются проблемы семантической близости слов и словосочетаний, что может представлять большую ценность для построения алгоритмов анализа кулинарных рецептов. Некоторые кулинарные ингредиенты могут принадлежать одному синонимическому либо ассоциативному ряду или же быть разными типами одного и того же ингредиента, например, разными видами сыра. Остается вопрос оптимального учета таких семантических отношений при построении моделей.

Цель статьи. Целью данного исследования является изучение особенностей построения систем, работающих с большими данными пользовательских кулинарных рецептов. Предлагается способ решения задачи семантической близости кулинарных ингредиентов на основе алгоритмов машинного обучения, а также рассмотрены методы построения векторного представления данных рецептов для дальнейшей работы классических алгоритмов. Рассматриваются основные первичные шаги сведения данных к виду, в котором с ними можно будет работать с конкретными задачами. Также рассматриваются современные программные инструменты с задачами компьютерной лингвистики.

Изложение основного материала. Перед рассмотрением способов обработки данных рецептов необходимо оговорить процесс получения самих данных. Так как на данный момент не имеется большого размеченного корпуса данных кулинарных рецептов, то такие данные приходится собирать собственноручно. Существуют сайты, собирающие большое количество пользовательских рецептов. Таким образом, получается, что тексты рецептов могут содержать различную стилистику, лексических ошибки и проч., из-за чего необходимо производить предобработку рецептов, прежде чем двигаться дальше, но уже на этом этапе отмечается возможная потеря точности итоговых моделей. Некоторые сайты предоставляют вместе с текстом рецептов список ингредиентов, которые приводятся в нормальной форме и проверяются администрацией сайта; для исследования был проанализирован ряд веб-ресурсов и был выбран один из таких сайтов [8], который позволяет работать с более, чем 140000 пользовательских рецептов, в каждом из которых выделен перечень используемых ингредиентов в удобном для автоматизированного извлечения виде. В результате анализа данных было установлено, что среднее количество ингредиентов, которое содержится в большинстве рецептов равно 10, что может оказаться полезной информацией для оценки количества параметров различных моделей; также анализ показал, что существуют рецепты, которые состоят всего из одного ингредиента, на

что нужно обращать внимание на этапе предварительной очистки данных от шума в зависимости от цели анализа данных. Также был выполнен анализ существующего программного обеспечения для предварительной обработки текстов на русском языке с точки зрения удобности использования, времени выполнения, а также точности. Отмечается небольшой выбор инструментов для решения таких задач и основным используемым программным модулем стал программный пакет NLTK, который не разрабатывался специально для русского языка, учитывая его лингвистические особенности, но поддерживает его наряду с другими распространенными языками мира. NLTK имеет удобный программный интерфейс и его можно попробовать не устанавливая на локальную рабочую станцию используя облачный сервис Google Colab; пакет библиотек в первую очередь используется для удаления стоп-слов и пунктуации из текста, что выполняется во многих задачах обработки текста и определенно необходимо при рассмотрении специфики предметной области, когда главный интерес представляют имена существительный. Обработка текстов рецепта главным образом необходима для выделения ценности ингредиентов для дальнейшего учета в различных алгоритмах. Если, в отличии от нашего случая, список ингредиентов не дан заранее в удобном для выделения виде, то необходимо произвести нахождение ключевых сущностей в тексте, при чем работа осложняется необходимостью работы со словосочетаниями, которые не обязательно представляют собой устоявшиеся выражения. Для повышения качества обработки, обращаясь к [5], следует производить статистическую анализ биграмм и триграмм, возможно, если анализ словаря возможных ингредиентов покажет, анализ N-грамм; анализ данных показал, что ингредиенты сайта [7] содержат не более четырех слов, однако в полном виде в тексте рецептов не встречаются. Более того, следует обращать внимание на авторский стиль; так, например, было замечено название «лук-рыба» для обозначения смеси рыбы с луком; обращаясь к этому же случаю, нужно не забывать про синонимические ряды, где на самом деле, в рецепте используется не просто рыба, а конкретный ингредиент — камбала. Рассматривая особенности предметной области, можно сказать, что конкретные названия рецептов имеют большую ценность, чем общие, и помогают отделить рецепт от ряда других. Здесь может стать вопрос разработки синонимического словаря для конкретного случая рассмотрения данных кулинарных рецептов, однако в данном исследовании использовалась семантическая близость слов, о которой будет сказано дальше. Для выделения ключевых сущностей, в данном случае ингредиентов, часто используется метрика TF-IDF (term frequency-inverse document frequency). В связи со сложностью одновременной работы с N-граммами, вместо использования TF-IDF, с целью сокращения затрат временных ресурсов на анализ данных, исследование ограничилось использованием векторного представления, основанного на семантических связях.

Представление данных в векторном виде является важным этапом задач обработки естественного языка и в частности нашей задачи, где нужно попытаться дать эффективное численное представление кулинарным рецептам. Простейшей моделью является «Мешок слов» [9], которая не учитывает семантических связей ингредиентов и не может определить, например, что разные сорта вина могут взаимозаменяться и использоваться в похожих блюдах. В ходе исследования было определено наличие 1104 ингредиентов на сайте [8]. Модель можно дополнить метрикой TF-IDF, используя вместо двоичного индикатора наличия ингредиента его вес. Также метрика TF-IDF позволит избавиться от чрезмерного вклада таких ингредиентов, как «кухонная соль», который присутствует в большинстве рецептов и вполне может быть не указан автором ввиду очевидности. Если же не использовать TF-IDF, необходимо заранее проанализировать частоту ингредиентов во всем наборе данных и отсеять вручную, или же установить маленький вес, наиболее частым, вроде соли и муки.

Для учета семантических связей ингредиентов была выбрана модель основанная на подходе word embedding [10]. Модель использует архитектуру нейронных сетей «Автокодировщик» и позволяет получить векторное представление слова в более низком измерении, чем количество всех слов; слова, употребляющиеся в схожем контексте, имеют меньшее косинусное расстояние, чем слова, которые зачастую общего контекста не имеют; именно общий контекст и понимается под семантической близостью слов. Преимуществом является также простота интерфейса и доступность имеющихся реализаций. Данное исследование использует программный пакет word2vec проекта gensim. С помощью данной модели была выполнена попытка получения векторного представления ингредиентов, которая учитывает их семантическую близость. Одним из важнейших параметров модели является измерение конечных векторов. Это число обычно устанавливается эмпирическим образом. В данном исследовании 1104 измерений были сокращены до 75. В данном исследовании размер контекста был установлен равным 10, что соответствует среднему количеству ингредиентов в рецепте и было подобрано исходя из специфики используемых данных. Число было подобрано из соображений скорости обучения модели, а также адекватности представления. Точность представления оценивалась на основе сходимости модели; так, например, при меньших или больших измерениях оставались ингредиенты, «выбросы», которые были значительно отдалены от всех остальных и эту проблему не получалось устранить при разном количестве эпох обучения.

Наличие «выбросов» определялось графическим способом используя метод главных компонент. На рисунке 1 изображены самые семантически-близкие слова ингредиенту «брынза».

```
('фета', 0.7700055241584778),
('сыр адыгейский', 0.7259960174560547),
('сыр сулугуни', 0.6097695231437683),
('моцарелла', 0.5436751842498779),
('сыр мягкий', 0.5195363759994507),
('сыр твердый', 0.5171306729316711),
('сыр голландский', 0.49607977271080017),
('сыр полутвердый', 0.4841534197330475),
('фетаки', 0.4596104323863983),
('пармезан', 0.42259055376052856)],
```

Рисунок 1. Ингредиенты с наименьшим косинусным расстоянием от ингредиента «брынза».

Видно, что слово, которое морфологически никак не показывает принадлежность сырам, было употреблено в одинаковом контексте со многими другими сортами сыра. Таким образом, получается довольно удобный семантический словарь для конкретной предметной области используя лишь на методах машинного обучения и не прибегая к специализированным словарям, которые могут и не содержать всех синонимов с пользовательского сайта. Вспомним вышеприведенный пример с «камбалой» и посмотрим на семантический соседей этого ингредиента на рисунке 2. Таким образом, модель способна понимать, что «камбала» — это рыба. Тем не менее, не все примеры столько хороши; так, например, модель не понимает, что «какао» и «какао-порошок» — семантически схожие элементы. Это можно связать с недостаточностью данных.

```
[('окунь морской', 0.5929579138755798),
 ('рыба', 0.578944981098175),
 ('пангасиус', 0.5736755132675171),
 ('карп', 0.5489171147346497),
 ('форель', 0.5452172756195068),
 ('филе рыбное', 0.537221372127533),
 ('скумбрия', 0.5231660604476929),
 ('треска', 0.5059509873390198),
 ('горбуша', 0.49632734060287476),
 ('тилапия', 0.4793583154678345)])
```

Рисунок 2. Ингредиенты с наименьшим косинусным расстоянием от ингредиента «камбала».

Была получена модель, которая позволяет представлять ингредиенты блюда в векторном пространстве учитывая семантические связи. Эту модель можно подавать на вход другим алгоритмам, например, алгоритмам кластеризации. Для получения векторного представления рецепта можно вычислить среднее арифметическое ингредиентов блюда. Такая модель не будет учитывать ценность отдельного ингредиента в блюде. Модель можно дополнить используя метрику TF-IDF давая вес векторам и вычисляя среднее взвешенное, вместо среднего арифметического.

Выводы и предложения. Были рассмотрены последние публикации обработки текстов русского языка. Были рассмотрены публикации работы с кулинарными рецептами. Также был произведен обзор доступного открытого программного обеспечения, необходимого для решения задач связанных с обработкой больших количеств кулинарных рецептов. В исследовании упоминаются некоторые особенности работы именно с пользовательскими рецептами, а не готовыми размеченными корпусами, а также были предложены методы решения классических задач обработки естественных языков конкретными программными инструментами для конкретной предметной области. Был предложен способ получения векторного представления данных, который основывается на алгоритмах нейронных сетей, и позволяет решать задачу семантической схожести ингредиентов. Полученная модель, при наличии больших качественных данных, позволяет частично избавиться от этапа составления словаря синонимов для конкретной узкой предметной области. Наконец, были предложены способы использования модели в качестве входных данных для других алгоритмов, а также предложена интеграция метода с метрикой TF-IDF. Также рассматриваются возможные улучшения процесса обработки с помощью более продвинутого программного обеспечения, в частности

эффективного разрешения кореференций для улучшения подсчета TF-IDF.

Список литературы:

1. Khorsheed, K & Madbouly, M & Khorsheed, Khattab & Madbouly, Magda & Guirguis, Shawkat. (2015). SEARCH ENGINE OPTIMIZATION USING DATA MINING APPROACH. IX. 184.

2. Азеркович И. Л. Использование мер семантической близости для распознавания кореференции в русском языке / И. Л. Азеркович// Вестник НГУ. Серия: Лингвистика и межкультурная коммуникация. 2019. Т. 17, № 1. С. 65-77. DOI 10.25205/1818-7935-2019-17-1-65-77.

3. Korobov M.: Morphological Analyzer and Generator for Russian and Ukrainian Languages / M. Korobov// Analysis of Images, Social Networks and Texts, pp 320-332 (2015).

4. Захаров, В. П., Хохлова, М. В. (2010). Анализ эффективности статистических методов выявления коллокаций в текстах на русском языке. / В. П. Захаров, М. В. Хохлова// Компьютерная лингвистика и интеллектуальные технологии, 9 (16), 137-143.

5. Хохлова М. В. Сопоставительный анализ статистических мер на примере частеречных преференций сочетаемости существительных. / М. В. Хохлова// Компьютерная лингвистика и вычислительные онтологии: сборник научных статей. Труды XIX Международной объединенной научной конференции «Интернет и современное общество» (IMS-2017), Санкт-Петербург, 21 – 23 июня 2017 г. — СПб: Университет ИТМО, 2017. С. 165-171.

6. Хохлова М. В. Статистический подход применительно к исследованию сочетаемости: от мер ассоциации к машинному обучению. / М. В. Хохлова// Структурная и прикладная лингвистика. Выпуск 13. СПб., 2019. С. 106–122.

7. Лазеева Н.В. Структурные и языковые особенности кулинарных рецептов поваренной книги “Cooking for Friends” г. Рамзи./ Н.В. Лазеева // Инновационная наука. 2016. №3-3 (15).

8. Рецепты и кулинария на Поварёнок.Ру. — URL : https://www.povarenok.ru (дата обращения 10.06.2020).

9. Проскурин А.А., Авсеева О.В. Объектно-ориентированная реализация обработки текста на основе алгоритма continuous bag of words. /А.А. Проскурин, О.В. Авсеева // Объектные системы. 2016. №13.

10. Karyaeva, Maria & Braslavski, Pavel & Sokolov, Valery. (2018). Word Embedding for Semantically Relative Words: an Experimental Study. Modeling and Analysis of Information Systems. 25. 726-733. 10.18255/1818-1015-2018-6-726-733.

УДК 62-974:

Нечитайлов К.П.
магистр кафедры Инженерия процессов, аппаратов, холодильной техники и технологии
Московский государственный университет пищевых производств (Россия, г. Москва)
Феськов О.А.
к.т.н., доцент кафедры Инженерия процессов, аппаратов, холодильной техники и технологии
Московский государственный университет пищевых производств (Россия, г. Москва)
Стефанова В.А.
к.т.н., доцент кафедры Инженерия процессов, аппаратов, холодильной техники и технологии
Московский государственный университет пищевых производств (Россия, г. Москва)

ИССЛЕДОВАНИЕ ПАРАМЕТРОВ РАБОТЫ ХОЛОДИЛЬНОГО АГРЕГАТА С ВОЗДУШНОЙ КАМЕРОЙ

Nechitaylov K. P.
Magister department «Engineering of processes, devices, refrigerating equipment and technology»
Moscow state University of food production (Moscow, Russia)
Feskov O.A.
Candidate of Technical Sciences, department «Engineering of processes , devices, refrigerating equipment and technology»
Moscow state University of food production (Moscow, Russia)
Stefanova V.A.
Candidate of Technical Sciences, department «Engineering of processes, devices, refrigerating equipment and technology»
Moscow state University of food production (Moscow, Russia)