На правах рукописи

Грибков Егор Игоревич

НЕЙРОСЕТЕВЫЕ МОДЕЛИ НА ОСНОВЕ СИСТЕМЫ ПЕРЕХОДОВ ДЛЯ ИЗВЛЕЧЕНИЯ СТРУКТУРИРОВАННОЙ ИНФОРМАЦИИ О ПРОДУКТАХ ИЗ ТЕКСТОВ ПОЛЬЗОВАТЕЛЕЙ

Специальность 05.13.17 —
«Теоретические основы информатики»

Автореферат
диссертации на соискание учёной степени
кандидата технических наук

Томск — 2020

Работа выполнена в федеральном государственном бюджетном образовательном учреждении высшего образования «Томский государственный университет систем управления и радиоэлектроники» (ТУСУР).

Научный руководитель – **Ехлаков Юрий Поликарпович**
доктор технических наук, профессор

Официальные оппоненты: **Пимонов Александр Григорьевич,** доктор технических наук, профессор, заведующий кафедрой прикладных информационных технологий Кузбасского государственного технического университета им. Т.Ф. Горбачева (г. Кемерово)

Спицын Владимир Григорьевич, доктор технических наук, профессор отделения информационных технологий Инженерной школы информационных технологий и робототехники Национального исследовательского Томского политехнического университета

Ведущая организация – Федеральное государственное автономное образовательное учреждение высшего образования **«Новосибирский государственный технический университет»**

Защита состоится 24 декабря 2020 г. в 15 часов 15 минут на заседании диссертационного совета Д 212.268.05 ТУСУРа по адресу: 634050, г. Томск, пр. Ленина, 40.

С диссертацией можно ознакомиться на официальном сайте https://postgraduate.tusur.ru/urls/cte13km6 и в библиотеке ТУСУРа по адресу: 634045, г. Томск, ул. Красноармейская, 146.

Автореферат разослан _____ 2020 г.

Ученый секретарь
диссертационного совета  Костюченко Евгений Юрьевич

Общая характеристика работы

Актуальность темы. На протяжении последнего десятилетия в связи с ростом доступности интернета и созданием множества интернет-ресурсов вроде социальных сетей и форумов наблюдается бурное развитие методов для автоматизации обработки текстов на естественном языке, в том числе текстов, в которых потребитель выражает свое мнение о приобретенном товаре или услуге. Это приводит к лавинообразному распространению позитивной и негативной информации о потребительских свойствах продукта, что может послужить как увеличить, так и обрушить продажи, сделать её продукцию менее конкурентоспособной в будущем. Основной формой передачи информации в интернете являются текст. Однако несмотря на письменную форму тексты в интернете чаще всего носят неформальный характер и изобилуют расхождениями с письменной нормой (употребление сленга, жаргонизмов, просторечных слов, игнорирование правил пунктуации, использование редукции). Отмеченные особенности текстовой информации пользователей, а также большие объемы порождаемой пользователями информации делают ручной анализ трудновыполнимой задачей. В этой связи автоматизация обработки и анализа содержимого текстов на естественном языке для извлечения информации о товарах является актуальной задачей. Для качественного решения этой задачи широкое распространение получили методы на основе машинного обучения.

Степень разработанности темы. Общие вопросы применения методов машинного обучения для обработки текстов исследовались в работах таких ученых как К. Дайер, Й. Голдберг, К.Д. Мэннинг, Н. Клахбеннер, З. Янг. Анализ пользовательских текстов является предметом интереса в работах Б. Пана, П.Д. Терни, Н. В. Лукашевич, Е. В. Тутубалиной, В. Вана. В работах Т.Т. Тета, И. Андроцопулиоса, Д. Вагнера и др. отмечается, что методы анализа тональности дают слишком обобщенную оценку объектам интереса пользователя и не предоставляют детальной информации об аспектах – составных частях, атрибутах или характеристиках оцениваемых пользователями объектов. В своих работах они предлагают расширенную постановку задачи, получившей название аспектно-ориентированного анализа тональности (АОАТ), в которой требуется определять в текстах аспекты и их тональности. Решение задачи в такой постановке позволяет получить представление о сложных продуктах с большим количеством эксплуатационных характеристик, в которых потребители могут положительно оценивать одни качества, но высказывать смешенные эмоции относительно других в пределах одного текста. В работах С. Джаббары, П. Симиано, А. Катияра, О. Ирсой рассматривается проблема извлечения из текстов пользователей мнений, представленных в виде

составных объектов, содержащих помимо аспектов связанные с ними оценочные высказывания.

Однако существующие методы извлечения из текстов пользователей структурированной информации производят извлечение отдельных составляющих и их объединение в результирующую структуру с помощью многокомпонентных моделей, которые настраиваются на решение задачи независимо друг от друга. Это приводит к эффекту распространения ошибки между отдельными компонентами и отрицательно сказывается на точности модели. Устранению данных недостатков посвящены методы структурного предсказания, позволяющие предсказывать структуру объекта в рамках единой модели с учетом структурных взаимосвязей между сущностями в тексте. Это устраняет риск распространения ошибки при передаче промежуточных результатов между компонентами и повышает конечную точность предсказаний. Одним из таких методов, описанных в работах Й. Ниврe, Й. Голдберга, Н. А. Смита, К. Дайера, является подход на основе системы переходов, сводящий задачу структурного предсказания к предсказанию последовательности действий, в результате исполнения которых будет получен искомый объект. Он отличается линейной сложностью получения предсказаний, гибкостью при определении структуры объектов и возможностью использования стандартного аппарата нейронных сетей для извлечения признаков из текстов.

Целью диссертационной работы является развитие методов предсказания составных объектов с использованием нейронных сетей в части извлечения структурированной информации из пользовательских текстов на естественном языке.

Для достижения цели необходимо решить следующие **задачи**.

1) Выявить специфику задачи извлечения и анализа структурированной информации из текстов пользователей с точки зрения методов обработки естественного языка.

2) Провести анализ современных методов обработки естественного языка на основе машинного обучения.

3) Разработать нейросетевую модель на основе системы переходов для извлечения составных объектов и их атрибутов из текстов отзывов на естественном языке.

4) Разработать нейросетевую модель и алгоритм для извлечения и анализа пользовательских мнений из текстов отзывов о потребительских свойствах товаров и подготовить обучающий набор данных.

5) Разработать нейросетевую модель и алгоритм обработки запросов пользователей на этапе эксплуатации и сопровождения программного продукта, подготовить обучающий набор данных.

6) Провести экспериментальное исследование предложенных моделей на материале подготовленных наборов данных.

7) Реализовать предложенные модели и алгоритмы в виде модулей программного комплекса и провести их практическую апробацию и внедрение.

Объектом исследования являются неструктурированные тексты на естественном языке, в которых пользователи высказывают свои мысли, мнения, замечания об опыте эксплуатации различных продуктов, а также свои пожелания и запросы производителям.

Предметом исследования являются модели для извлечения и анализа структурированной информации из текстов на естественном языке на основе нейронных сетей с применением подхода на основе системы переходов.

Теоретическую и методологическую базу исследования составили труды ведущих российских и зарубежных специалистов в области обработки естественного языка, лингвистики, машинного обучения и нейросетевых методов. Информационной базой являются материалы, опубликованные в периодической печати, учебной и научной литературе, сети Интернет.

Методы исследования. Диссертационная работа опирается на методы обработки естественного языка, построения и обучения нейронных сетей, методы предсказания структурированных данных.

Область исследования диссертационной работы соответствует указанным в паспорте специальности 05.13.17 «Теоретические основы информатики» пунктам:

– п. 5 «Разработка и исследование моделей и алгоритмов анализа данных, обнаружения закономерностей в данных и их извлечениях разработка и исследование методов и алгоритмов анализа текста, устной речи и изображений»;

– п. 6 «Разработка методов, языков и моделей человекомашинного общения; разработка методов и моделей распознавания, понимания и синтеза речи, принципов и методов извлечения данных из текстов на естественном языке».

Научная новизна полученных в диссертационной работе результатов.

1) Предложена оригинальная нейросетевая модель на основе системы переходов для извлечения составных объектов и их атрибутов из текстов на естественном языке, позволяющая одновременно предсказывать структуру объекта и значения его атрибутов, с возможностью адаптации под конкретные задачи через задание множеств, описывающих семантику фрагментов и атрибутов.

2) На основе предложенной модели разработана оригинальная модель для извлечения и анализа мнений из текстов пользовательских отзывов о продуктах, отличающаяся от известных моделей использованием подхода на основе системы переходов и позволяющая получить лучшие

показатели качества извлечения частей составных объектов: 0,795 F_1 – при определении фрагментов, 0,723 F_1 – при определении отношений, 0,631 F_1 – при определении атрибутов.

3) На основе предложенного метода разработана оригинальная модель для анализа запросов пользователей на этапе эксплуатации и сопровождения программного продукта, отличающаяся от известных моделей использованием подхода на основе системы переходов и позволяющая получить лучшие показатели качества извлечения частей составных объектов: 0,633 F_1 – при определении фрагментов, 0,693 F_1 – при определении отношений.

Теоретическая ценность работы заключается в развитии методов обработки естественного языка, в частности методов предсказания объектов со сложной структурой с использованием моделей на основе системы переходов и нейросетевого подхода в задачах, связанных с обработкой текстов мнений пользователей о продуктах.

Практическая значимость работы обуславливается возможностью использования разработанных моделей и программных средств в следующих случаях:

1) при анализе текстов отзывов пользователей о продуктах маркетологами компаний как для определения сильных и слабых сторон собственных продуктов и продуктов фирм-конкурентов, а также последующей модификации комплекса маркетинговых мероприятий для улучшения положения продукта на рынке;

2) на этапе эксплуатации и сопровождения программного продукта специалистами службы технической поддержки пользователей для обеспечения эксплуатации программного продукта в соответствии с его техническими характеристиками и развития продукта в соответствии с предложениями пользователей и требованиями рыночной ситуации;

3) при сравнении альтернативных предложений потенциальными покупателями товаров интернет-магазина «AliExpress» посредством сервиса «Quiddi.ru» с целью выбора товара, оценка качества которого в наибольшей степени подкреплена информацией их отзывов других покупателей.

Результаты диссертационного исследования использованы:

– в ФГБОУ ВО «ТУСУР» при выполнении государственного задания Министерства науки и высшего образования РФ, проект FEWM-2020-0036 «Методологическое и инструментальное обеспечение принятия решений в задачах управления социально-экономическими системами и процессами в гетерогенной информационной среде».

– в учебном процессе кафедры автоматизации обработки информации (АОИ) ТУСУРа при чтении курса лекций и проведении практических занятий по дисциплинам «Интеллектуальные вычислительные системы», «Анализ больших данных» при подготовке магистров по направлению 09.04.04 — «Программная инженерия»;

– при реализации коммерческих продуктов компании ООО «ТомскСофт»: программной системы для извлечения и анализа мнений о потребительских свойствах товаров «Quiddi Semantics» (свидетельство о регистрации программы для ЭВМ №2019612276 от 14.02.2019), программной системы для обработки запросов пользователей на этапе эксплуатации и сопровождения программного продукта «Quiddi Support Analyst» (свидетельство о регистрации программы для ЭВМ №2020614799 от 24.04.2020).

Достоверность полученных результатов обусловлена корректным применением аппарата нейронных сетей при разработке модели, а также количественным сравнением предложенных моделей с аналогами. Адекватность предложенных в работе моделей и алгоритмов подтверждается результатами их практического использования в коммерческих программных продуктах компании ООО «ТомскСофт».

Публикации. Основные результаты по теме диссертации изложены в 4 журналах, рекомендованных ВАК, 4 – в тезисах докладов. Получены свидетельства о регистрации программ для ЭВМ №2019612276 от 14.02.2019 и №2020614799 от 24.04.2020.

Апробация результатов работы. Основные результаты диссертационной работы докладывались на конференциях различного уровня. Среди них:

1) всероссийская научная конференция молодых ученых «Наука. Технологии. Инновации» (03–07 декабря 2018 г., г. Новосибирск, НГТУ);

2) международная научно-практическая конференция «Электронные средства и системы управления» (20–22 ноября 2019 г., г. Томск, ТУСУР);

3) международная научно-техническая конференция студентов, аспирантов и молодых ученых «Научная сессия ТУСУР» (2019-2020 гг., г. Томск, ТУСУР).

Личный вклад. Автором самостоятельно выполнены анализ современных методов обработки естественного языка на основе машинного обучения, предметной области, теоретическое и экспериментальное исследование разработанных моделей и алгоритмов, проектирование и реализация подсистем обучения моделей и анализа текста в составе программных систем «Quiddi Semantics» и «Quiddi Support Analyst». Совместно с научным руководителем разработаны содержательная и математическая постановки задач, предложены структуры классификаторов типов мнений пользователей и типов информативных фраз о программном обеспечении. Сервис для разметки текстов разработан Трошиным М.В., подсистемы для сбора информации и обращений пользователей разработаны совместно Трошиным М.В. и Пекарских Е.А.

Основные положения, выносимые на защиту.

1) Нейросетевая модель на основе системы переходов для извлечения составных объектов и их атрибутов из текстов на естественном

языке, позволяющая одновременно извлекать фрагменты объектов и определять взаимосвязи между ними с возможностью адаптации к конкретной предметной области через задание множеств, определяющих смысловое наполнение фрагментов составных объектов и из атрибутов.

2) Нейросетевая модель для извлечения и анализа пользовательских мнений из текстов отзывов о потребительских свойствах товаров, разработанная на основе предложенной модели общего вида и обеспечивающая точность определения фрагментов 0,795 F_1, отношений — 0,723 F_1, атрибутов – 0,631 F_1.

3) Нейросетевая модель для обработки запросов пользователей на этапе эксплуатации программного продукта, разработанная на основе предложенной модели общего вида и обеспечивающая точность определения фрагментов 0,633 F_1, отношений — 0,693 F_1.

Объем и структура работы. Диссертация состоит из введения, трёх глав, заключения и двух приложений. Полный объём диссертации составляет 128 страниц, включая 37 рисунков и 20 таблиц. Список литературы содержит 127 наименований.

Основное содержание работы

Во **введении** проводится обоснование актуальности темы исследования, приведен обзор описанных в научной литературе результатов по теме диссертации, сформулирована цель и поставлены необходимые для её решения задачи, описаны объект, предмет и использованные методы исследования, изложены научная новизна, теоретическая и практическая значимость работы. Описаны сведения об апробации и внедрении полученных результатов исследования.

В **главе 1** отмечается, что сопровождение продукта после выхода на рынок является самым длинным в рамках жизненного цикла этапом, в ходе которого потребители могут сталкиваться с различными трудностями при эксплуатации продукта, и производителю необходимо оперативно решать задачи:

– технической поддержки потребителей, для обеспечения эксплуатации продукта в соответствии с его техническими характеристиками;

– развития продукта в соответствии с предложениями потребителей и требованиями рыночной ситуации;

– модификации комплекса маркетинговых мероприятий для улучшения положения продукта на рынке.

Исходной информацией для решения этих задач являются запросы и пожелания пользователей при использовании продукта, а также общедоступные текстовые каналы: почтовые сервисы, службы поддержки, интернет-форумы, социальные сети, страницы интернет-магазинов, торговые агрегаторы. Приводятся особенности порождаемой пользователями

информации. Делается вывод о необходимости автоматизации обработки и анализа содержимого текстов на предмет наличия и извлечения информации о продуктах и использовании для этих целей методов на основе машинного обучения.

Приводится обзор методов анализа пользовательских текстов на естественном языке, отмечается, что современные методы, использующиеся для решения расссмотренных задач, не позволяют учитывать структурные взаимосвязи между извлекаемыми сущностями и подвержены проблемам распространения ошибки, что приводит к низкой точности извлечения объектов при работе от «начала до конца». Отмечается, что наиболее перспективным направлением решения обозначенной проблемы является применение методов, позволяющих предсказывать структуру объекта в рамках единой модели с учетом структурных взаимосвязей между сущностями в тексте, что обеспечивают более высокую точность по сравнению с моделями, представленными последовательно соединенными компонентами.

Учитывая лингвистические особенности текстов о продуктах и услугах, предлагается объединить преимущества моделей на основе системы переходов с признаками, получаемыми с помощью рекуррентных и свёрточных нейронных сетей, для решения задачи извлечения структурированной информации о мнениях, запросах и пожеланиях потребителей из этих текстов.

В **главе 2** предлагается комплекс оригинальных нейросетевых моделей на основе системы переходов для извлечения структурированной информации о продуктах из текстов пользователей. Рассматривается общая задача извлечения составных объектов из текста и два варианта её адаптации под конкретные предметные области.

Модель для извлечения составных объектов и их атрибутов из текстов на естественном языке

Пусть заданы:
$\mathbf{w} = (w_1,...,w_N)$ – последовательность слов исходного текста.
$Lbl = \{lbl_1, ..., lbl_{n_{Lbl}}\}$ – множество типов фрагментов.
$A = a_1, ..., a_{n_A}$ – множество атрибутов.
$V(a) = \{v^a_1, ..., v^a_{n_{V(a)}}\}$ – множество допустимых значений атрибута a.
Необходимо найти:
$O = \{o_1, ..., o_{n_O}\}$, – множество составных объектов.
Составной объект определим как пару $o = (SP, R)$, где:
$SP = \{sp_{n_{SP}}\}$ – множество фрагментов объекта.
$R = \{r_{n_R}\}$ – множество связей между фрагментами.
$sp = (\{w_j, w_{j+1}, ..., w_{j+m}\}, lbl)$ – фрагмент, где – j индекс начала фрагмента в тексте $\mathbf{w}$, m – длина фрагмента, lbl – метка фрагмента.

$r = (sp_1, sp_2, av = \{(a_k, v_m^{a_k}), (a_p, v_g^{a_p}), ...\})$ – отношение между фрагментами, где sp_1, sp_2 – участвующие в отношении фрагменты, av – множество назначеных значений атрибутов ($\forall k,p : a_k \neq a_p$, если $k \neq p$).

Пример формальной структуры составного объекта в общем виде с двумя фрагментами и одной связью приведен на рисунке 1.

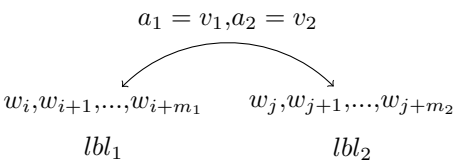

Рис. 1 — Пример формальной структуры составного объекта

Модель для извлечения составных объектов и их атрибутов на основе переходов определяется как $(C,Y,A(C))$, где C – конфигурация системы, Y – множество переходов, изменяющих конфигурацию, $A(C)$ – функция, задающая множество переходов, доступных для исполнения в текущей конфигурации. Конфигурация модели задается кортежем $C = (B, S, L, H)$, где B – список необработанных слов исходного текста, S – список извлеченных фрагментов, L – список отношений между фрагментами, H – история совершенных переходов. Начальную и целевую конфигурации предлагается определить следующим образом:

$$C_0 = ((w_1,...,w_N), \emptyset, \emptyset, \emptyset),$$
$$C_T = (\emptyset, (s_1,...,s_{|S|}), (l_1,...,l_{|L|}), (y_1,...,y_T))$$

Множество доступных переходов задается в следующем виде:

$$Y = \{Shift, Start(lbl), Add(lbl), Link(n_1,n_2), Attr(a,v), End\}.$$

Изменения, вносимые в конфигурацию C_t при совершении определенного перехода на шаге t, представлены в таблице 1.

Таблица 1 — Изменение конфигурации при совершении переходов

t	y_t	$t+1$
$w\|B$	$Shift$	B
$w\|B;S$	$Start(lbl)$	$B;S\|(\{w\};lbl)$
$w\|B;S\|(\{...,x\};lbl)$	$Add(lbl)$	$B;S\|(\{...,x,w\};lbl)$
L	$Link(n_1,n_2)$	$L\|(S_{n_1}, S_{n_2}, \emptyset)$
$L\|(S_{n_1}, S_{n_2}, \{...,(a',v')\})$	$Attr(a,v)$	$L\|(S_{n_1}, S_{n_2}, \{...,(a',v'),(a,v)\})$

Функция $A(C)$, определяющая допустимость совершения переходов в конкретной конфигурации, задана в виде набора условий и представлена в виде таблицы 2.

Таблица 2 — Условия допустимости совершения переходов

Переход	Условие
$Shift$	$B \neq \emptyset$
$Start(lbl)$	$B \neq \emptyset$
$Add(lbl)$	$B \neq \emptyset \wedge S \neq \emptyset \wedge type(S_{-1}) = lbl$
$Link(n_1,n_2)$	$\exists S_{n_1} \wedge \exists S_{n_2} \wedge (n_1,n_2) \notin L$
$Attr(a,v)$	$L \neq \emptyset \wedge \forall v \nexists (a,v) \in L_{-1}$
End	$B = \emptyset$

Тогда процесс извлечения составных объектов и их атрибутов может быть представлен следующей последовательностью шагов:

Шаг 1 На основе исходного текста $\mathbf{w}$ определить начальную конфигурацию C_0, задать $t = 1$.

Шаг 2 Рассчитать набор признаков конфигурации $\phi(C_{t-1})$.

Шаг 3 Подать набор признаков в классификатор f и предсказать наиболее вероятный переход с учетом условий допустимости из 2: $\hat{y}_t = f(\phi(C_{t-1}))$. Если $\hat{y}_t = End$, то $t = t + 1$ и переход на шаг 5.

Шаг 4 Получить конфигурацию C_t из C_{t-1} в соответствии с предсказанным переходом по правилам из таблицы 1 и перейти на шаг 2.

Шаг 5 Конец предсказания, преобразовать полученную последовательность переходов $\mathbf{y}$ в составной объект.

В качестве f будет использовать вероятностный классификатор следующего вида:

$$p_\theta(\hat{y}_t|C_{t-1}) = \text{softmax}_{\hat{y} \in A(C_t)}(\mathbf{W}\phi(C_{t-1}) + \mathbf{b}), \tag{1}$$

где $\phi(C_t)$ – вектор признаков конфигурации, $\mathbf{W}$ и $\mathbf{b}$ – параметры классификатора.

Вектор признаков конфигурации формируется конкатенацией признаков отдельных её элементов:

$$\phi(C_t) = [\phi(B_t); \phi(S_t); \phi(H_t)]\,. \tag{2}$$

Основной для получения признаков B и S служат контекстно-зависимые векторные представления входной последовательности слов:

$$\mathbf{h}_i^B = F(E(w_1, w_2, ..., w_N), i), \tag{3}$$

где E – отображение последовательности слов в последовательность векторных представлений, F – преобразование последовательности векторных представлений в контекстно-зависимую форму. В качестве преобразования F в работе предлагается использовать свёрточную или рекуррентную нейронную сеть. Вектор $\phi(B_t)$ образован конкатенацией первых n_B контекстно-зависимых признаков элементов списка B на шаге t:

$$\phi(B_t) = \left[\mathbf{h}^B_{t(1)}; \mathbf{h}^B_{t(2)}; ...; \mathbf{h}^B_{t(n_B)}\right], \tag{4}$$

где $t(i)$ – позиция i-го элемента B в последовательности представлений h^B на шаге t.

Для формирования вектора $\phi(S_t)$ сначала рассчитываются признаки индивидуальных элементов h^S по формуле:

$$\mathbf{h}_i^S = \left[\left(\frac{1}{e(i) - b(i)} \sum_{j=b(i)}^{e(i)} \mathbf{h}_j^B \right); E^{Label}(i) \right], \tag{5}$$

где $b(i)$ и $e(i)$ – индексы, соответствующие началу и концу i-го фрагмента текста в последовательности $\mathbf{h}^B$, $E^{Label}(i)$ - векторное представление типа i-го фрагмента. Затем n_S последних элементов последовательности конкатенируются, образуя вектор $\phi(S_t)$:

$$\phi(S_t) = \left[\mathbf{h}_{-1}^S; \mathbf{h}_{-2}^S; ...; \mathbf{h}_{-n_s}^S \right]. \tag{6}$$

Для формирования вектора признаков истории совершенных переходов $\phi(H_t)$ используется скрытое состояние последнего шага сети LSTM, примененной к последовательности векторных представлений действий:

$$\phi(H_t) = \text{LSTM}(E^{Act}(H_1),..., E^{Act}(H_t))_t \tag{7}$$

С учетом вышеизложенного архитектура нейронной сети может быть представлена в виде диаграммы, представленной на рисунке 2.

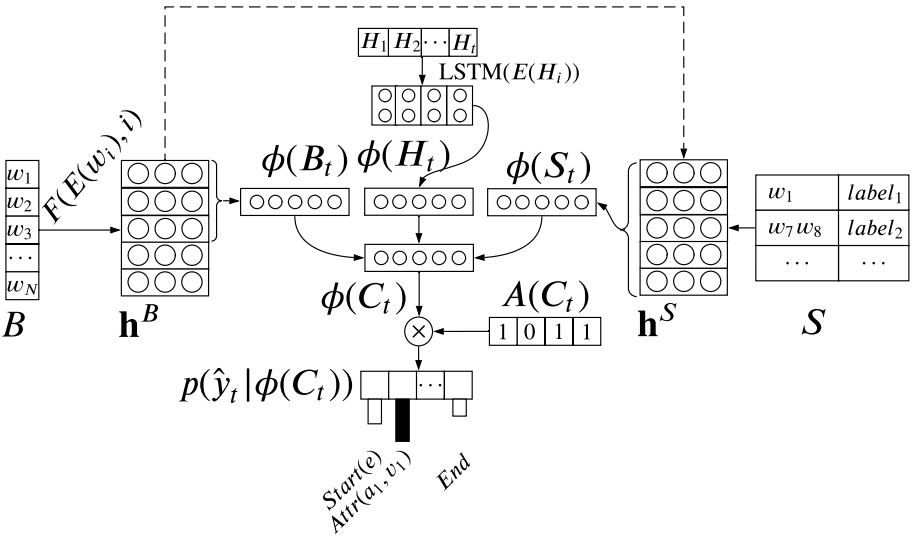

Рис. 2 — Архитектура нейронной сети

Обучение рассмотренной модели производится при помощи метода максимального правдоподобия. В качестве ошибки используется сумма

значений перекрестной энтропии, рассчитанной для каждого шага предсказания отдельно:

$$\mathcal{L}(\mathbf{y}, \hat{\mathbf{p}}) = \sum_{t-1}^{T} \mathcal{L}_t(y_t, \hat{\mathbf{p}}_t) = \sum_{t=1}^{T} -\log p_\theta(\hat{y}_t | C_{t-1})_{y_t}. \quad (8)$$

Градиент ошибки распространяется по сети с помощью процедуры обратного распространения ошибки через структуру. Для оптимизации параметров используется адаптивный метод Adam, развивающий идеи стохастического градиентного спуска.

Для адаптации описанных модели и алгоритмов для решения конкретных прикладных задач необходимо в соответствии со смысловым наполнением фрагментов, связей и атрибутов извлекаемых объектов задать состав множеств Lbl, A, $V(a)$ и функции $A(C_t)$.

При оценке качества работы модели были использованы показатели метрики F_1 для фрагментов, атрибутов, и связей. Так как границы фрагментов однозначно определить тяжело даже человеку, для их оценки используется вариант метрики F_1, учитывающий частичные совпадения.

Предложенная модель для извлечения составных объектов была положена в основу решений двух прикладных задач извлечения структурированной информации из текстов пользователей: извлечения и анализа пользовательских мнений из текстов отзывов о потребительских свойствах товаров; обработки запросов пользователей на этапе эксплуатации программного продукта.

Нейросетевая модель для извлечения и анализа пользовательских мнений из текстов отзывов о потребительских свойствах товаров

Пусть задано множество текстов отзывов покупателей T, содержащих мнения из множества O о наборе продуктов P. Мнением будем называть четверку (Аспект, Описание, Тональность, Тип). Аспектом будем называть важную характеристику или упоминание продукта. Описанием – последовательность слов, содержащую высказанное пользователем мнение о некотором аспекте. Тональность задана категориальной шкалой, состоящей из трех уровней: положительной(+), нейтральной(0), негативной(−). Множество типов мнений зададим в виде иерархического классификатора, представленного на рисунке 3. В целях практической апробации будем использовать укрупненные классы: {Товар, Продавец, Доставка}.

Имея исходные множества T, P, O и заданную структуру классификатора, необходимо извлечь из новых текстов отзывов $\hat{T}$ множество мнений пользователей $\hat{O}$ о новом наборе товаров $\hat{P}$. Для решения задачи необходимо адаптировать описанную ранее нейросетевую модель путем

Рис. 3 — Классификатор типов мнений пользователей

определения состава множеств Lbl, A, $V(a)$ согласно смысловому наполнению введённого выше понятия мнения:

$$Llb = \{\text{Аспект}, \text{Описание}\},$$
$$A = \{\text{Тональность}, \text{Цель}\},$$
$$V(\text{Тональность}) = \{+,0,-\},$$
$$V(\text{Цель}) = \{\text{Товар}, \text{Продавец}, \text{Доставка}\}.$$

Экспериментальное исследование модели для извлечения мнений проводилось на материале отзывов покупателей из интернет-магазина «AliExpress» на русском языке для трех самых многочисленных категорий товаров: бытовая техника, дом и авто, одежда. При сравнении использовалось несколько моделей:

– базовая модель, запоминающая фрагменты в обучающем множестве и извлекающая подстроки максимальной длинны, совпадающие с запомненными фрагментами; Отношения строились между ближайшими парами фрагментов совместимых типов.

– многокомпонентная гибридная модель на основе свёрточных и рекуррентных нейронных сетей (Hybrid-NN);

– модель на основе двунаправленной LSTM и условного случайного поля (LSTM-CRF) для предсказания фрагментов, совмещенная с компонентами для предсказания отношений и атрибутов из многокомпонентной модели;

– два варианта предложенной модели на основе системы переходов с многослойной свёрточной (Trans-CNN) и многослойной двунаправленной LSTM (Trans-LSTM) сетями в качестве преобразования F.

При расчете оценок качества использовалась процедура кросс-валидации с тремя блоками, каждый блок содержал тексты из одной категории. Модель обучалась на текстах из двух категорий, оценка качества рассчитывалась на третьей категории. Целью такой процедуры является уменьшение влияния фактора запоминания моделью конкретных форм фрагментов и связей между ними на показатели качества.

Усредненные по трем категориям значения F_1 приведены в таблице 3. Жирным выделена лучшая точность, курсивом – следующая за ней. По всем рассмотренным показателям лучшим вариантом оказался

Таблица 3 — Результаты сравнения качества моделей на данных «AliExpress» (F_1)

Модель	Аспект	Описание	Отнош.	Тонал.	Цель
Базовая	0,569	0,611	0,318	–	–
Hybrid-NN	0,753	0,763	0,661	0,447	0,549
LSTM-CRF	*0,771*	*0,782*	*0,713*	*0,484*	*0,591*
Trans-CNN	0,770	0,787	0,699	0,532	0,659
Trans-LSTM	**0,788**	**0,802**	**0,723**	**0,578**	**0,684**
Сред. Trans-LSTM	0,795		0,723	0,631	

Trans-LSTM. Улучшение относительно следующей наилучшей альтернативой составило: при извлечении аспектов и описаний – 1,84% и 2,57%, отношений – 1,41%, атрибутов тональности и цели – 19,43% и 15,75%. Итоговая точность при определении фрагментов составила $0{,}795F_1$, отношений – $0{,}723F_1$, атрибутов – $0{,}631F_1$. Высокие показатели роста качества определения атрибутов составных объектов свидетельствуют о том, что предлагаемая модель на основе системы переходов позволяет модели лучше интегрировать информацию о составных частях объекта. Примеры полученных моделью предсказаний приведены ниже.

Пример 1

И $[\text{Носки}]_{A1}$ $[\text{как на картинке}]^{+:\text{тов}}_{O1}$, $[\text{хорошего качества}]^{+:\text{тов}}_{O1}$, $[\text{приятные на ощупь}]^{+:\text{тов}}_{O1}$

П $[\text{Носки}]_{A1}$ $[\text{как на картинке}]^{+:\text{тов}}_{O1}$, $[\text{хорошего}]^{+:\text{тов}}_{O2}$ $[\text{качества}]_{A2}$, $[\text{приятные}]^{+:\text{тов}}_{O3}$ $[\text{на ощупь}]_{A3}$

Пример 2

И $[\text{Чехлы}]_{A1}$ $[\text{ооооочень крутые}]^{+:\text{тов}}_{O1}$, $[\text{подходят на любые стулья}]^{+:\text{тов}}_{O1}$, на наши $[\text{сели вообще идеально}]^{+:\text{тов}}_{O1}$, $[\text{очень рекомендуем}]^{(+:\text{тов}),(+:\text{пр})}_{O(2,3)}$ $[\text{товар}]_{A2}$ и $[\text{продавца}]_{A3}$!

П $[\text{Чехлы}]_{A1}$ $[\text{ооооочень крутые}]^{+:\text{тов}}_{O1}$, $[\text{подходят на любые стулья}]^{+:\text{тов}}_{O1}$, на наши сели вообще идеально,$[\text{очень рекомендуем}]^{(+:\text{тов}),(+:\text{пр})}_{O2}$ $[\text{товар}]_{A2}$ и $[\text{продавца}]_{A2}$!

Нейросетевая модель для обработки запросов пользователей на этапе эксплуатации программного продукта

Пусть задано множество текстов запросов пользователей T, содержащих информативные фразы (ИФ) O о программных продуктах P. Информативной фразой (ИФ) мы будем называть тройку (Функция, Описание, Тип). Функцией будем называть как функциональные возможности ПО, элементы графического интерфейса, так и упоминания самого ПП. Описанием – последовательность слов, содержащую высказанное пользователем мнение о функции, её состоянии или свои пожелания, описание сложившейся ситуации. Множество типов зададим в соответствии с иерархическим классификатором, представленным на рисунке 4. В целях практической апробации будем использовать укрупненные классы: положительная оценка функции, отрицательная оценка функции, ошибка, запрос.

Рис. 4 — Классификатор типов информативных фраз

Имея исходные множества T, P, O и заданную структуру классификатора ИФ, необходимо извлечь из новых текстов запросов $\hat{T}$ информативные фразы $\hat{O}$ о новом наборе программных продуктов $\hat{P}$. Для решения задачи необходимо адаптировать описанную ранее нейросетевую модель путем определения состава множеств Lbl, A, $V(a)$ согласно смысловому наполнению ИФ:

$$Lbl = \{\text{Функция}, +, -, \text{Ошибка}, \text{Запрос}\}, A = \emptyset, V(a) = \emptyset.$$

Экспериментальное исследование предложенной модели осуществлялось на основе обращений пользователей магазина приложений «Google Paly Market» на русском языке. Сбор обращений производился из девяти категорий мобильных приложений: автомобили и транспорт, карты и навигация, медицина, музыка и аудио, персонализация, финансы, шоппинг, образование, видеоплееры и редакторы.

В таблице 4 приведены показатели F_1 для извлечения всех типов ИФ (усредненный), связей между ними, которые усреднены по всем 9 категориям приложениям. Жирным выделен лучший результат, курсивом – следующий за ним. По всем рассмотренным показателям лучшим

Таблица 4 — Результаты сравнения качества моделей

Модель	Функция	Описание	Отношение
Базовая	0,192	0,157	0,275
Hybrid-NN	0,464	0,468	0,396
LSTM-CRF	*0,650*	*0,552*	*0,455*
Trans-CNN	0,601	0,551	0,654
Trans-LSTM	**0,675**	**0,592**	**0,693**
Сред. Trans-LSTM	0,633		0,693

вариантом оказался Trans-LSTM. Улучшение относительно следующей наилучшей альтернативой составило: при извлечении функций и описаний – 3,92% и 7,21%, отношений – 52,34%. Итоговая точность при извлечении фрагментов составила $0{,}633F_1$, отношений – $0{,}693F_1$. Это позволяет подтвердить её эффективность для извлечения информативных фраз о программных продуктах.

Примеры сравнения разметки с предсказанием приведены ниже.

Пример 1

И [Сделайте]$_{\text{З1}}$ [поиск по номеру автомобиля]$_{\text{О1}}$, как у Яндекса, или [qr код на лобовом стекле или двери]$_{\text{О1}}$ - [иногда очень сложно]$_{-2}$ [найти машину]$_{\text{О2}}$ пальцем на карте.

П [Сделайте]$_{\text{З1}}$ [поиск по номеру автомобиля]$_{\text{О1}}$, как у Яндекса, или [qr код]$_{\text{О1}}$ на лобовом стекле или [двери]$_{\text{О2}}$ - [иногда очень сложно найти машину пальцем на карте]$_{-2}$.

Пример 2

И [Напрягает необходимость периодически]$_{-1}$ [начинать работу «с начала»]$_{\text{О1}}$ после того, как [пропадают прежде настроенные]$_{\text{Ош2}}$ [экраны рабочего стола]$_{\text{О2}}$ – ведь, это ещё один-два часа рутинных по сути операций.

П [Напрягает]$_{-1}$ необходимость периодически [начинать работу «с начала» после]$_{\text{О1}}$ того, как [пропадают]$_{\text{Ош2}}$ прежде настроенные [экраны]$_{\text{О2}}$ рабочего стола – ведь, это ещё один-два часа рутинных по сути операций.

Качественный анализ полученных результатов говорит о возможности практического применения предложенной модели для обработки запросов пользователей при эксплуатации и сопровождении программных продуктов.

Глава 3 рассматривает вопросы практической апробации и внедрение предложенных моделей и алгоритмов. Проведен обзор основных представленных на рынке программных продуктов для обработки текстов не естественном языке. Проведенный анализ показал, что из всех рассмотренных продуктов Только два из шести рассмотренных продуктов содержат реализации методов аспектно-ориентированного анализа. Исходя из этих соображений, было принято решение о реализации предложенных моделей извлечения структурированной информации из текстов пользователей в виде оригинальных программных продуктов.

Нейросетевая модель для извлечения и анализа пользовательских мнений из текстов отзывов о потребительских свойствах товаров была положена в основу программной системы (ПС) «Quiddi Semantics», предназначенной для сбора и обработки текстов отзывов магазина «AliExpress». ПС состоит из трех подсистем: сбора информации, обучения модели; анализа отзывов. Получено свидетельство о регистрации программы №2019612276 от 14.02.2019. Разработанное программное обеспечение было использовано компанией ООО «ТомскСофт» при разработке агрегатора товаров «Quiddi.ru», предоставляющего извлеченную из отзывов покупателей магазина «AliExpress» информацию о товарах. На момент написания диссертации система была использована для анализа отзывов о более чем 7 миллионах товаров.

Нейросетевая модель для обработки запросов пользователей на этапе эксплуатации программного продукта положена в основу прототипа программной системы «Quiddi Support Analyst» для сбора и анализа обращений пользователей программных продуктов. ПС реализована в виде микросервисной архитектуры и состоит из нескольких веб-сервисов: сервиса сбора обращений, сервиса разметки и обучения, сервиса анализа сообщений, сервиса предоставления отчетов. На данный момент ПС «Quiddi Support Analyst» находится в стадии рабочего прототипа, на который получено свидетельство о регистрации программы для ЭВМ №2020614799 от 24.04.2020.

Результаты практической апробации предложенных нейросетевых моделей, реализованных в виде компонентов программных систем «Quiddi Semantics» и «Quiddi Support Analyst» позволяют сделать вывод об их работоспособности и практической пригодности при решении задач по извлечению мнений и запросов пользователей из текстов, что подтверждается актом на внедрение и использовании программных систем в компании ООО «ТомскСофт».

Основные результаты работы

В результате выполнения диссертационной работы были получены следующие теоретические и практические результаты.

1) Проведен анализ современного состояния исследования в области анализа пользовательских текстов на естественном языке, в ходе которого выявлено преимущество методов, направленных на извлечение информации о продуктах в структурированной форме; обозначена проблема низкой точности описанных в научной литературе методов, обусловленная использованием многокомпонентных моделей. Сделан вывод о необходимости использования методов, позволяющих предсказывать структуру объекта в рамках единой модели в задачах, связанных с извлечением структурированной информации о продуктах из текстов.

2) Обозначены особенности употребления языка в текстах пользователей о продуктах, которые оправдывают применение современных нейросетевых методов. Рассмотрены существующие методы структурного предсказания и их преимущества перед альтернативами. В частности, выделен класс методов, использующий системы переходов, который позволяет использовать стандартные методы обучения и вывода нейронных сетей, обеспечивая при этом возможность выражать сложные структурные признаки путем использования конфигураций – структур данных, хранящих промежуточное представление предсказываемого объекта.

3) Предложена оригинальная нейросетевая модель на основе системы переходов для извлечения составных объектов и их атрибутов из текстов на естественном языке, позволяющая одновременно извлекать фрагменты объектов и определять взаимосвязи между ними с возможностью адаптации к конкретной предметной области через задание множеств, определяющих смысловое наполнение фрагментов составных объектов и их атрибутов.

4) Предложенная модель легла в основу оригинальной нейросетевой модели для извлечения и анализа мнений из текстов пользовательских отзывов о продуктах, отличающаяся от известных моделей использованием подхода на основе системы переходов. Для экспериментального исследования модели подготовлен набор данных, состоящий из текстов отзывов интернет-магазина «Ali Express» на русском языке. Исследование модели показало более высокое качество извлечения мнений предложенной моделью по сравнению с рассмотренными альтернативами. Качественный анализ полученных результатов говорит о возможности практического применения данной модели для извлечения и анализа мнений пользователей о потребительских свойствах товаров.

5) Предложенная модель легла в основу оригинальной нейросетевой модели для анализа запросов пользователей на этапе эксплуатации и сопровождения программного продукта, отличающаяся от известных моделей

использованием подхода на основе системы переходов. Для проведения экспериментального исследования модели предложен набор запросов из магазина приложений «Google Play Market» на русском языке. Результаты экспериментального исследования позволяют говорить о более высоком качестве анализа запросов предложенной моделью по сравнению с альтернативами.

6) Предложенные нейросетевые модели были положены в основу разработанных программных систем «Quiddi Semantics»(свидетельство о регистрации программы для ЭВМ №2019612276 от 14.02.2019) и «Quiddi Support Analyst»(свидетельство о регистрации программы для ЭВМ №2020614799 от 24.04.2020). Программные системы внедрены и используются в компании ООО «ТомскСофт».

7) Результаты диссертационного использованы в ФГБОУ ВО «ТУСУР» при выполнении государственного задания Министерства науки и высшего образования РФ, проект FEWM-2020-0036 «Методологическое и инструментальное обеспечение принятия решений в задачах управления социально-экономическими системами и процессами в гетерогенной информационной среде»; в учебном процессе кафедры автоматизации обработки информации (АОИ) при чтении курса лекций и проведении практических занятий по дисциплинам «Интеллектуальные вычислительные системы», «Анализ больших данных» при подготовке магистров по направлению 09.04.04 — «Программная инженерия»

Публикации автора по теме диссертации

В изданиях из списка ВАК РФ

1. *Грибков, Е. И.* Нейросетевая модель на основе системы переходов для извлечения составных объектов и их атрибутов из текстов на естественном языке / Е. И. Грибков, Ю. П. Ехлаков // Доклады ТУСУР. — 2020. — Т. 23, № 1. — С. 47—52.
2. *Грибков, Е. И.* Нейросетевая модель на основе системы переходов для извлечения и анализа тональности пользовательских мнений / Е. И. Грибков, Ю. П. Ехлаков // Искусственный интеллект и принятие решений. — 2020. — № 1. — С. 99—110.
3. *Ехлаков, Ю. П.* Модель извлечения пользовательских мнений о потребительских свойствах товара на основе рекуррентной нейронной сети / Ю. П. Ехлаков, Е. И. Грибков // Бизнес-информатика. — 2018. — Т. 46, № 4. — С. 7—16.
4. *Грибков, Е. И.* Нейросетевая модель для обработки запросов пользователей на этапе эксплуатации и сопровождения программного продукта / Е. И. Грибков, Ю. П. Ехлаков // Бизнес-информатика. — 2020. — Т. 14, № 1. — С. 7—18.

Свидетельства о государственной регистрации программы для ЭВМ

5. «Quiddi Semantics» / Е. И. Грибков [и др.] // Свидетельство о государственной регистрации программы для ЭВМ №2019612276 от 14.02.2019.
6. «Quiddi Support Analyst» / Е. И. Грибков [и др.] // Свидетельство о государственной регистрации программы для ЭВМ №2020614799 от 24.04.2020.

В сборниках трудов конференций

7. *Грибков, Е. И.* Набор данных и модель глубокого обучения для анализа текстов отзывов пользователей / Е. И. Грибков, Ю. П. Ехлаков // Наука. Технологии. Инновации. Сборник научных трудов. В 9-ти частях. — Новосибирск : НГТУ, 2018. — С. 180—184.
8. *Грибков, Е. И.* Модель обработки запросов пользователей на этапе эксплуатации и сопровождения программного продукта / Е. И. Грибков, Ю. П. Ехлаков // Электронные средства и системы управления. Материалы докладов Международной научно-практической конференции. — Томск : В-Спектр, 2019. — С. 141—143.
9. *Грибков, Е. И.* Модель извлечения структурированных объектов и их атрибутов из текстов на естественном языке / Е. И. Грибков // Сборник избранных статей научной сессии ТУСУРа (Томск, 22–24 мая 2019 г.): в 2 ч. — Томск : В-Спектр, 2019. — С. 54—56.
10. *Грибков, Е. И.* Модель на основе системы переходов для извлечения составных объектов из текстов / Е. И. Грибков, Ю. П. Ехлаков // Сборник избранных статей научной сессии ТУСУР, Томск, 13–30 мая 2020. — Томск : В-Спектр, 2020. — С. 52—55.

Грибков Егор Игоревич

НЕЙРОСЕТЕВЫЕ МОДЕЛИ НА ОСНОВЕ СИСТЕМЫ ПЕРЕХОДОВ ДЛЯ ИЗВЛЕЧЕНИЯ СТРУКТУРИРОВАННОЙ ИНФОРМАЦИИ О ПРОДУКТАХ ИЗ ТЕКСТОВ ПОЛЬЗОВАТЕЛЕЙ

Автореф. дис. на соискание ученой степени канд. тех. наук

Подписано в печать ___.___._____. Заказ № _____
Формат 60×90/16. Усл. печ. л. 1. Тираж 100 экз.
Типография _____