На правах рукописи

Коновалов Василий Павлович

Методы переноса знаний для нейросетевых моделей обработки естественного языка

Специальность: 05.13.17 —
«Теоретические основы информатики»

АВТОРЕФЕРАТ
диссертации на соискание учёной степени
кандидата технических наук

Долгопрудный — 2022

Работа выполнена в федеральном государственном автономном образовательном учреждении высшего образования «Московский физико-технический институт (национальный исследовательский университет)»

Научный руководитель: **Бурцев Михаил Сергеевич — кандидат физико-математических наук**

Ведущая организация: **Федеральное государственное учреждение «Федеральный исследовательский центр «Информатика и управление» Российской академии наук»**

Защита состоится **15 сентября 2022 г. в 14 часов 00 минут** на заседании диссертационного совета **ФПМИ.05.13.17.008**, созданного на базе федерального государственного автономного образовательного учреждения высшего образования «Московский физико-технический институт (национальный исследовательский университет)» (МФТИ, Физтех)

по адресу: 141701, Московская область, г. Долгопрудный, Институтский переулок, д. 9.

С диссертацией можно ознакомиться в библиотеке МФТИ, Физтех и на сайте организации https://mipt.ru.

Автореферат разослан «_____» _____ 2022 г.

Ученый секретарь диссертационного совета Войтиков Константин Юрьевич

Общая характеристика работы

Актуальность темы. Разработка эффективной модели обработки естественного языка методами машинного обучения с учителем требует соблюдения ряда условий: наличия тренировочной выборки достаточного размера; единства языка тренировочной и тестовой выборок; нахождения тренировочной и тестовой выборок в одном и том же тематическом домене; сходства распределения классов в тренировочной и тестовой выборках. Однако, учитывая изобилие языков, задач и доменов в реальном мире, решать задачи, строго следуя этой парадигме, представляется трудновыполнимым. В связи с этим в рамках стандартного обучения с учителем возникает необходимость в использовании вспомогательного инструмента, позволяющего справляться с современными вызовами.

Методы переноса знаний стали одним из таких инструментов. Перенос знаний позволяет осуществлять передачу знаний из связанных областей, задач и языков в целевую задачу. Перенос знаний допускает, что предметные области (домены), задачи и распределения данных могут быть разными в обучающей и тестовой выборках. В реальном мире мы регулярно сталкиваемся с переносом знаний. Например, умение играть на фортепьяно, несомненно, помогает при освоении электронного органа.

Перенос знаний давно зарекомендовал себя в обработке естественного языка (англ.: Natural Language Processing, NLP). Например, применяя предобученные на неразмеченных корпусах контекстно-независимые векторные представления слов, осуществляется перенос знаний с предобученных векторов на задачи NLP, что в свое время позволило добиться впечатляющих результатов на задачах определения именованных сущностей, исправления грамматических ошибок и многих других.

Метод word2vec, предложенный специалистами компании Google в 2013 году, при обучении контекстно-независимых векторных представлений слов для английского языка, использовал корпус Google News, содержащий 6 миллиардов токенов. Векторные представления GloVe обучались на датасете Common Crawl, содержащем 42 миллиарда токенов. Как известно, чем больше данных для обучения, тем эффективнее будет перенос на целевую задачу. Однако большое количество данных доступно только для популярных языков. В связи с тем, что для непопулярных языков зачастую отсутствуют размеченные корпуса и весьма ограничено количество текстовых данных в открытом доступе, такие языки называют малоресурсными (англ.: low-resource languages, LRL). При этом существует масса причин для того, чтобы исследовать LRL. Разработка NLP инструментов для LRL может иметь серьезные экономические перспективы. Кроме того, исследование LRL препятствует их исчезновению и способствует их популяризации. Но возникает закономерный вопрос: как

обучать векторное представление слов для малоресурсного языка, для которого нет большого количества неразмеченных данных? Чтобы ответить на него, необходимо сравнить классические частотные методы построения векторных представлений слов (PMI, сингулярное разложение матрицы PMI) с нейросетевыми методами (Skip-Gram Negative Sampling (SGNS), Continuous Bag-of-Words (CBoW)).

Применение контекстуально-независимых векторных представлений слов значительно продвинуло NLP. В большинстве задач перенос знаний с предобученных векторов приводил к улучшению качества по сравнению со случайной инициализацией. Однако, серьезным недостатком применения контекстуально-независимых векторных представления является то, что они используются только при инициализации первого слоя нейронной сети, в то время как остальные слои обучаются с нуля на данных целевой задачи.

Современные методы решения задач NLP используют нейросетевые языковые модели, такие как ELMo, BERT и другие, способные генерировать контекстуальные векторные представления слов. Обучение таких моделей проходит в два этапа: предобучение и дообучение. Этап предобучения гарантирует изучение связей между словами, а этап дообучения обеспечивает эффективный перенос знаний для решения целевой задачи, что было показано на датасетах GLUE, SWAG, SQuAD.

Чтобы дальше развивать модели понимания языка, сообществу требуются новые сложные задачи и наборы данных для них. Такими задачами могут стать вопросно-ответные задачи, задачи отслеживания состояния диалога и другие. Под вопросно-ответной задачей подразумевается поиск ответа на вопрос по контексту. Модель отслеживания состояния диалога поддерживает текущее состояние диалога в семантическом представлении.

Стэнфордский вопросно-ответный датасет SQuAD (англ.: The Stanford Question Answering Dataset) на английском языке содержит около ста тысяч примеров в обучающей выборке. Чтобы собирать датасеты такого размера, используются специальные платформы, например, Yandex.toloka[1] или Amazon Mechanical Turk[2]. Это краудсорсинговые платформы, которые позволяют распределять задачи разметки между большим количеством неквалифицированных разметчиков. При этом большие объемы данных и сложность задач часто приводят к ошибкам в разметке. Анализ ошибок вопросно-ответной системы, обученной на датасете русского языка SberQuAD, выявил, что 74% ошибок модели связаны с неправильной разметкой датасета (29% – неполный ответ, 19% – размытый вопрос, 14% – неправильный ответ, 12% – слишком общий вопрос), а не с отсутствием возможности модели правильно отвечать на вопросы по

[1] `https://toloka.yandex.ru/`
[2] `https://www.mturk.com/`

контексту. Одна из задач этой работы – выяснить возможно ли использовать методы переноса знаний для того, чтобы сократить объем требуемой тренировочной выборки без существенной потери качества модели.

Однако, не для всех задач можно найти подходящий набор данных. Например, качество решения задачи отслеживания состояния диалога (англ.: Dialogue State Tracking, DST) до сих пор страдает от нехватки подходящих обучающих датасетов. Популярный DST-датасет MultiWOZ 2.0, содержащий диалоги для семи доменов (отель, такси, ресторан и другие), изобилует ошибками разметки: запоздалая разметка – слот размечается после его первоначального использования в диалоге; мультиразметка – токен размечен как относящийся к нескольким слотам; ошибочная разметка – токен назначен неправильному слоту и другие. Исправление этих ошибок привело к нескольким последовательным версиям: MultiWOZ 2.1, MultiWOZ 2.2, MultiWOZ 2.3, MultiWOZ 2.4, каждая из которых исправляет старые ошибки и вносит новые. Факт наличия нескольких ревизий одного датасета значительно осложняет работу разработчикам диалоговых систем. Для сбора диалоговых данных обычно разрабатывается отдельная полуавтоматическая система симуляции диалога, которая работает в режиме Wizard-of-Oz, когда пользователь считает, что общается с автоматической диалоговой системой, а на самом деле – с другим разметчиком (или человеком из команды сбора данных). Стоимость разработки такой системы ложится тяжелым бременем на сборщиков датасета. Нивелировать часть этих проблем и заметно снизить объем обучающих данных, необходимых для тренировки диалоговой системы, может применение методов переноса знаний.

Целью данной работы является исследование методов переноса знаний при решении проблем обработки естественного языка с помощью нейросетевых моделей.

Для достижения поставленной цели требуется решить следующие **задачи**:

1. Предложить метод обучения контекстно-независимых векторных представлений слов при наличии ограниченной обучающей выборки.
2. Предложить способ сравнения качества контекстно-независимых векторных представлений слов.
3. Сравнить качество обученных векторных представлений слов предложенным методом.
4. Предложить способ экономии ресурсов при сборе обучающей выборки и предобучении языковой модели для решения вопросно-ответной задачи целевого языка.
5. Разработать модель отслеживания состояния диалога при помощи последовательного переноса вопросно-ответной модели.

6. Опубликовать в открытом доступе программный код обученных моделей.

Научная новизна:

1. Выполнено сравнение методов построения контекстно-независимых векторных представлений слов для трех малоресурсных языков.
2. Предложен оригинальный способ внутренней оценки качества обученных векторных представлений слов.
3. Выполнено сравнение различных многоязычных и языко-специфичных нейросетевых языковых моделей для решения вопросно-ответной задачи.
4. Показано, что использование многоязычной обучающей выборки позволяет сократить требуемый размер обучающей выборки целевого языка для решения вопросно-ответной задачи.
5. Разработан оригинальный метод переноса вопросно-ответной модели для отслеживания состояния диалога.

Теоретическая и практическая значимость. Следующие положения относятся к теоретической значимости:

- Установлено, что при наличии обучающей выборки ограниченного размера, контекстно-независимые векторные представления, обученные частотными методами, превосходят векторные представления, обученные нейросетевыми методами.
- Показано, что межъязыковой перенос позволяет использовать общедоступные тренировочные данные английского языка при дообучении M-BERT для вопросно-ответной задачи.
- Экспериментально установлено, что, применяя метод межъязыкового переноса, M-BERT, дообученный с применением многоязычной обучающей выборки, имеет сопоставимое качество с языко-специфичными BERT для вопросно-ответной задачи.
- Разработана оригинальная модель GOLOMB, которая, применяя метод последовательного переноса знаний, использует вопросно-ответную модель SQuAD для отслеживания состояния диалога.

Практическая значимость заключается в следующем:

- Обучены и опубликованы в открытом доступе контекстно-независимые векторные представления для бурятского, эрзянского и коми языков.
- Установлено, что M-BERT, дообученный с применением многоязычной обучающей выборки, имеет сопоставимое качество с языко-специфичными BERT для вопросно-ответной задачи, тем самым отпадает необходимость в вычислительных ресурсах для предобучения языко-специфичных BERT.

– Показано, что межъязыковой перенос позволяет использовать общедоступные тренировочные данные английского языка при дообучении M-BERT для вопросно-ответной задачи, таким образом, отпадает необходимость в сборе полноценного тренировочного датасета целевого языка.
– Применение последовательного переноса вопросно-ответной модели позволяет улучшить качество модели отслеживания состояния диалога.
– В открытый доступ выложены модели, обученные в рамках диссертационной работы. Обученные модели готовы для использования в приложениях.

Методология и методы исследования. В исследовании использовались методы численного эксперимента для анализа задач обработки естественного языка, методы машинного обучения, основы теории вероятностей. При создании моделей для библиотеки с открытым кодом DeepPavlov использовались методы разработки на языках Python, Bash.

Основные положения, выносимые на защиту:

1. Векторные представления, обученные частотными методами, превосходят векторные представления, обученные нейросетевыми методами, при использовании обучающей выборки ограниченного размера.
2. Применяя метод межъязыкового переноса, многоязычный BERT, дообученный с применением многоязычной обучающей выборки, имеет сопоставимое качество, а иногда превосходит языко-специфичные BERT для вопросно-ответной задачи.
3. При наличии большой обучающей выборки модели обучаются быстрее в режиме ранней остановки, чем в режиме фиксированного количества эпох для вопросно-ответной задачи.
4. Метод последовательного переноса позволяет применить вопросно-ответную модель к решению задачи отслеживания состояния диалога, улучшая качество последней.

Достоверность результатов обеспечивается экспериментами при использовании алгоритмов машинного обучения. Модели, обученные в рамках работы, выложены в открытый доступ либо в составе библиотеки DeepPavlov[3], либо отдельно. Таким образом обеспечивается воспроизводимость экспериментов. Кроме того, результаты работы согласуются с результатами, полученными другими авторами.

Апробация работы. Результаты исследования были представлены на следующих семинарах и научных конференциях:

– «XXIV Международная конференция по компьютерной лингвистике и интеллектуальным технологиям Диалог», доклад «Learning Word Embeddings For Low Resource Languages: The Case Of Buryat»,

[3] `https://github.com/deepmipt/DeepPavlov/`

Vasily Konovalov, Zhargal Tumunbayarova, 30 мая – 2 июня 2018, Москва;
- Конференция «Google NLP Summit 2019», постер «DeepPavlov: An Open-Source Library for Conversational AI», Vasily Konovalov, 22 июня 2019, Цюрих, Швейцария;
- Конференция «IA Week», постер «DeepPavlov: An Open-Source Library for Conversational AI», Vasily Konovalov, 17 – 21 ноября 2019, Тель-Авив, Израиль;
- Конференция «AI Journey», постер «Multi-task Dialogue State Tracking», Pavel Gulyaev, Evgenia Elistratova, Vasily Konovalov, Mikhail Burtsev, 8 – 9 ноября 2019, Москва;
- Конференция «AAAI Conference on Artificial Intelligence (AAAI-20)», постер «Goal-Oriented Multi-Task BERT-Based Dialogue State Tracker», 7 – 8 февраля, 2020, Нью-Йорк, США;
- «XXVI Международная конференция по компьютерной лингвистике и интеллектуальным технологиям Диалог», доклад «Exploring the BERT Cross-Lingual Transfer for Reading Comprehension», Vasily Konovalov, Pavel Gulyaev, Alexey Sorokin, Yury Kuratov, Mikhail Burtsev, 17 июня – 20 июня 2020, Москва.

Личный вклад. В работе [1] (индексируется Scopus) автором собран и предобработан корпус, проанализированы методы обучения векторных представлений слов, обучены векторные представления слов, разработана методика анализа качества векторных представлений слов, обученных разными способами. В работе [2] (индексируется Scopus) автором выполнены эксперименты по дообучению моделей, построены кривые обучения и проанализированы результаты. В работах [3] и [4] (индексируется RSCI) автором реализована часть модели отслеживания состояния диалога, проведены эксперименты, выполнен анализ результатов. В работе [5] автором адаптированы модели классификации текста для случая ограниченной тренировочной выборки.

Публикации. Основные результаты по теме диссертации изложены в 5 печатных изданиях, 1 из которых издано в журналах, индексируемых RSCI, 2 — в периодических научных журналах, индексируемых Web of Science и Scopus.

Содержание работы

Во **введении** раскрывается актуальность исследований, проводимых в рамках данной диссертационной работы, дается обзор научной литературы по изучаемой тематике, формулируется цель исследования, ставятся задачи работы, описывается научная новизна и практическая значимость работы.

Первая глава посвящена обзору методов переноса знаний при решении задач обработки естественного языка. В главе подробно описывается архитектура Трансформер, которая раскрыла возможности механизма внимания и обновила лучшие метрики качества на задачах генерации последовательностей: машинный перевод, суммаризация текста и многих других. Рассматривается архитектура BERT, которая представляет собой кодировщик Трансформера, предобученный на задачах восстановления пропущенного слова (англ.: masked language modeling, MLM) и предсказания следующего предложения (англ.: next sentence prediction, NSP). Языковая модель BERT совершила поистине переворот в NLP, позволив использовать предобученную модель для решения прикладных задач, значительно улучшив качество и сократив количество данных, необходимых для обучения.

В сценарии машинного обучения с учителем при обучении модели для задачи и предметной области (домена) предполагается, что предоставлены размеченные данные для той же задачи и домена. Традиционная парадигма машинного обучения с учителем сталкивается с трудностями, когда нет достаточного количества размеченных данных для обучения целевой задачи на целевом домене. Перенос знаний (англ.: transfer learning) позволяет использовать результаты, полученные при решении некоторой исходной задачи на исходном домене, для решения целевой задачи на целевом домене.

Домен D состоит из пространства признаков χ и распределения $P(X)$, где $X = \{x|x_i \in \chi, i = 1,...,n\}$. Задача τ состоит из пространства классов υ и функции принятия решений $f : \tau = \{\upsilon, f\}$. Функция f обучается алгоритмами машинного обучения $f(x_j) = \{P(y_k|x_j)|y_k \in \upsilon, k = 1,...,|\upsilon|\}$.

Исходный домен D_S задачи τ_S содержит размеченную обучающую выборку $D_S = \{(x,y)|x_i \in \chi^S, y_i \in \upsilon^S, i = 1,...,n^S\}$, при этом целевой домен D^T содержит либо ограниченное число размеченных примеров, либо только неразмеченные примеры.

Таким образом, при наличии обучающей выборки для исходных задач и доменов $\{(D_{S_i}, \tau_{S_i})|i = 1,...,m^S\}$ и выборки для целевых задач и доменов $\{(D_{T_i}, \tau_{T_i})|i = 1,...,m^T\}$, перенос обучения использует знания, полученные при решении исходных задач τ_S на исходных доменах D_S, для решения целевых задач τ_T на целевых доменах D_T, улучшая функции принятия решений f^T. Определение дано для общего случая, когда учитывается произвольное количество исходных и целевых задач. Большинство случаев применения переноса знаний имеют дело с $m^S = m^T = 1$.

Таксономия методов переноса знаний, адаптированная к обработке естественного языка, приведена на рисунке 1. Перенос знаний делится на две основные группы: индуктивное обучение и трансдуктивное обучение.

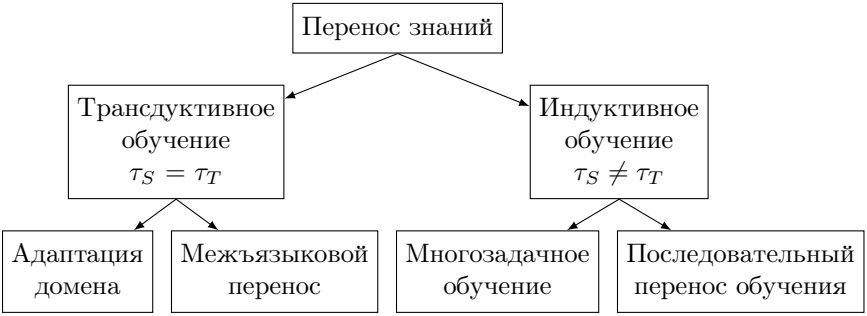


Рис. 1 — Таксономия методов переноса знаний, адаптированная к обработке естественного языка

Цель индуктивного обучения – улучшить функцию f_T на целевом домене D_T, используя знания, полученные из D_S и τ_S, когда $\tau_S \neq \tau_T$. Индуктивное обучение предполагает, что доступна размеченная обучающая выборка для целевой задачи. При этом для исходной задачи не обязательно наличие размеченной обучающей выборки.

При **последовательном переносе** модели исходной и целевой задач обучаются последовательно. То есть сначала обучается модель исходной задачи на размеченной или неразмеченной выборке, затем, используя знания, полученные при обучении на исходную задачу, обучается модель целевой задачи. Например, контекстно-независимые векторные представления, обученные на неразмеченной выборке, в последующем используются для решения многих задач NLP.

Для **многозадачного обучения** доступна обучающая выборка как для исходной задачи, так и для целевой задачи. При этом для обоих задач обучается единая модель с учетом единой функции потерь. Предполагается, что единая модель, обучаясь одновременно на несколько задач, улучшает качество решения целевой задачи.

Трансдуктивное обучение предполагает, что исходная и целевая задачи одинаковы $\tau_S = \tau_D$, однако домены задач могут быть разными.

Межъязыковой перенос знаний подразумевает, что исходная и целевая задачи совпадают, однако используют разные языки.

При **адаптации домена** предполагается, что домены исходной и целевой задач отличаются.

Вторая глава посвящена сравнению методов обучения векторных представлений слов для малоресурсных языков (на примере бурятского, эрзянского, коми).

Разработки в области NLP проделали длинный путь от моделей на регулярных выражениях до моделей, обученных алгоритмами машинного обучения, которые активно используют перенос знаний. Современные

модели NLP требуют внушительных текстовых ресурсов как неразмеченных (для предобучения), так и размеченных (для дообучения). При этом исследования фокусируются на 20 популярных языках из около 7,000 существующих. Для непопулярных языков зачастую отсутствуют базовые NLP инструменты (например, лемматизатор) и размеченные корпуса. Такие языки называют малоресурсными языками (англ.: low-resource languages, LRL). Существует масса причин для того, чтобы исследовать LRL. Например, в Африке и Индии представлены 2,000 LRL, носителями которых являются 2.5 миллиарда человек. Разработка NLP инструментов для такой многочисленной группы людей может иметь серьезные экономические перспективы.

Языки народов России относятся к 14 языковым семьям. Более чем 250 языков являются малоресурсными. Среди них 32 обладают собственными корпусами, снабженными поисковыми системами, но только пять из них доступны по открытой лицензии Creative Commons, для 22 языков лицензия неизвестна.

В главе представлено сравнение методов обучения векторных представления слов для малоресурсных языков, когда лишь небольшой текстовый корпус доступен для обучения. В качестве малоресурсных языков использовались: бурятский, эрзянский, коми. В качестве контекста применялись окна размером 2, 5, 10 токенов с каждой стороны от целевого слова. В каждом случае обучались векторные представления методом положительной точечной взаимной информации (PPMI), а также 50, 100, 500-мерные векторные представления методами сингулярного разложения матрицы PPMI, Skip-Gram с негативным семплированием (SGNS), непрерывного мешка слов (CBoW), глобальных векторов (GloVe).

Многие малоресурсные языки России имеют собственные лингвистические корпуса, например, корпус бурятского языка[4], но не оговаривается, под какой лицензией корпус распространяется. Зачастую такие корпуса отсутствуют в открытом доступе и не доступны по запросу составителям. Поэтому векторные представления слов обучались на статьях Википедии соответствующего языка[5].

Первая часть главы посвящена обзору методов построения векторных представлений слов.

Вторая часть описывает метод, предложенный для внутренней оценки качества (англ.: intrinsic evaluation) векторных представлений. Качество векторных представлений английского языка оценивается на основе следующих наборов данных: RG, WordSim-353, WS-Sim и MEN. Каждый из этих наборов данных состоит из пар слов с соответствующими оценками сходства, присвоенными аннотаторами. Векторное представление оценивается путем вычисления корреляции (коэффициент корреляции Спирмена)

[4] http://web-corpora.net/BuryatCorpus

[5] распространяется под лицензией CC BY-SA и GNU

между сходством векторов и оценкой сходства аннотаторов. Однако эти наборы данных страдают общими недостатками: сравнение несопоставимых понятий, низкая внутренняя согласованность разметчиков. Кроме того, использование рейтинговых шкал может привести к различным ошибкам в аннотациях. В таких методах разные отношения и разные целевые слова оцениваются по одной шкале, например, (*кошка*, *домашнее животное*) vs. (*зима*, *время года*).

Для оценки сформированных векторных представлений использовался набор базовых слов (гиперонимов) с соответствующими гипонимами. Гипероним выражает обобщение значения, тогда как гипоним – ограничение. Например, *фрукт* является гиперонимом по отношению к *яблоко*, *яблоко* – гипонимом к *фрукт*. Для формирования положительных пар использовались разные гипонимы одного гиперонима, для формирования отрицательных – гипонимы разных гиперонимов. Очевидно, что семантическое сходство между положительными парами должно быть больше, чем семантическое сходство между отрицательными парами. Поэтому семантическое сходство между положительными парами берется за 1, в то время как семантическое сходство между отрицательными парами берется за -1. В итоге вычисляется функция потерь на основе косинусной близости между векторами[6]:

$$loss(x,y) = \begin{cases} 1 - cos(x_1,x_2) & \text{if } y = 1, \\ max(0, cos(x_1,x_2)) & \text{if } y = -1. \end{cases}$$

В таблице 1 приведены значения функции потерь при определении качества векторных представлений. Таким образом чем значение меньше, тем векторное представление лучше.

Векторные представления, сформированные методом SVD, существенно превосходят все остальные методы при наличии лишь ограниченной обучающей выборки. Этот факт подтверждается прошлыми работами. Кроме того, для метода SVD качество векторов меньшей размерности превосходит качество векторов большей размерности.

Учитывая недостатки методов внутренней оценки качества, указанные в прошлых работах, данный метод оценки подходит лишь для того, чтобы сравнить между собой векторные представления, обученные разными методами.

Внутренняя оценка качества векторных представлений слов дает общую информацию о методах, что не означает, однако, что методы, показавшие лучший результат на внутренней оценке качества, будут наиболее эффективными при решении конкретных задач. Внешняя оценка качества (англ.: extrinsic evaluation) применяется непосредственно к задачам NLP. В качестве такой задачи рассматривается определение частей речи (POS),

[6] `https://pytorch.org/docs/stable/generated/torch.nn.CosineEmbeddingLoss.html`

Таблица 1 — Внутреннее сравнение качества векторных представлений слов, сформированных разными методами.

Метод	win	бурятский			эрзянский			коми		
	dim	2	5	10	2	5	10	2	5	10
PPMI	-	0.441	0.428	0.417	0.435	0.417	0.393	0.421	0.397	0.385
SVD	50	**0.287**	0.291	0.294	0.257	0.235	**0.223**	0.209	**0.189**	0.194
	100	0.296	0.301	0.309	0.271	0.244	0.227	0.213	0.208	0.218
	500	0.385	0.391	0.393	0.311	0.291	0.277	0.282	0.282	0.296
CBoW	50	0.417	0.363	0.325	0.369	0.362	0.35	0.298	0.259	0.234
	100	0.419	0.364	0.329	0.373	0.358	0.357	0.301	0.258	0.238
	500	0.426	0.372	0.334	0.376	0.364	0.358	0.3	0.258	0.234
SGNS	50	0.355	0.325	0.324	0.343	0.317	0.307	0.294	0.279	0.284
	100	0.359	0.332	0.338	0.346	0.319	0.307	0.298	0.277	0.286
	500	0.363	0.33	0.336	0.346	0.326	0.305	0.296	0.276	0.288
GloVe	50	0.404	0.358	0.333	0.395	0.36	0.344	0.378	0.337	0.33
	100	0.406	0.373	0.354	0.388	0.381	0.357	0.383	0.35	0.348
	500	0.413	0.378	0.367	0.388	0.379	0.358	0.382	0.357	0.351

которая ранее уже применялась для внешней оценки качества векторных представлений.

Модели POS обучаются и тестируются на датасетах универсальных зависимостей (англ.: Universal Dependencies, UD), которые подходят для морфосинтаксического, семантического исследования типологически разных языков.

Для классификации частей речи используется модель на основе долгой краткосрочной памяти (англ.: Long short-term memory, LSTM). Слой векторного представления инициализируется обученными векторами (векторы зафиксированы и не обновляются во время обучения). Качество модели измеряется с помощью F-меры. Для оценки качества моделей используется перекрестная проверка (англ.: cross-validation) по 10 блокам, как в предыдущих работах.

В таблице 2 приведены результаты сравнения векторных представлений слов на задаче определения частей речи. Для всех видов векторных представлений увеличение размерности вектора приводит к улучшению качества целевой модели. В некоторых случаях увеличение окна контекста приводит к улучшению результата.

Частотные методы значительно превосходят нейросетевые методы, что можно связать с нехваткой обучающей выборки для нейросетевых методов. Это показывает, что не надо слепо следовать рекомендации и строить векторные представления слов, используя только нейросетевые методы. Качество факторизации SVD-500 сопоставимо с векторами PPMI. Таким образом, целесообразно использовать именно факторизованные векторы SVD вместо разряженных многомерных векторов PPMI. Полученные результаты подтверждают, что SGNS превосходит CBoW на задаче определения частей речи.

Таблица 2 — Внешнее сравнение качества векторных представлений слов, обученных разными методами. Для сравнения используется задача классификации частей речи с метрикой потокенная F-мера.

Модель	win	бурятский			эрзянский			коми		
	dim	2	5	10	2	5	10	2	5	10
PPMI	-	57.4	57.2	**57.6**	**54.7**	54.4	54.2	45.2	45.6	**46.3**
SVD	50	39.1	41.1	41.9	40.9	44.5	44.6	37.4	37.9	38.0
	100	44.8	49.2	50.4	46.4	49.4	48.8	38.9	39.7	41.9
	500	56.1	56.3	57.0	52.8	53.1	53.6	45.2	44.4	44.6
CBoW	50	29.4	28.6	30.1	32.1	34.0	36.2	30.0	33.0	33.7
	100	30.4	29.9	32.5	35.0	36.1	38.6	32.7	34.1	35.0
	500	32.6	34.2	35.1	37.0	37.9	39.9	35.6	35.9	36.6
SGNS	50	31.2	31.8	26.6	37.0	37.2	35.6	28.4	29.4	26.6
	100	34.9	36.6	37.6	39.7	39.5	40.9	33.8	34.3	36.3
	500	38.3	40.7	41.7	42.0	43.1	44.1	37.6	37.6	38.4
GloVe	50	34.9	34.6	35.4	39.5	41.0	40.8	34.0	34.4	34.4
	100	38.1	38.9	40.9	43.1	44.4	45.5	34.9	36.1	36.7
	500	40.4	45.0	48.9	46.1	48.4	49.7	36.8	38.3	39.1

Результаты главы опубликованы в работе «Learning Word Embeddings For Low Resourse Languages: The Case Of Buryat» [1]. Все модели и векторные представления выложены в открытый доступ[7].

В **третьей главе** применяется метод межъязыкового переноса знаний для вопросно-ответной модели.

Межъязыковой перенос знаний подразумевает, что для обучения модели на целевом языке используются данные другого языка. Такое возможно, если модель применяет универсальное, не зависящее от языка, векторное представление.

Появление в 2016 году открытого Стэнфордского вопросно-ответного датасета (англ.: Stanford Question Answering Dataset , SQuAD) сделало возможным создание QA систем на основе неструктурированного текста. В отличие от других вопросно-ответных датасетов, SQuAD не предлагает выбрать ответ из списка предопределенных ответов, а предоставляет возможность самостоятельно найти ответ на вопрос в статье. Датасет SQuAD содержит 107,785 пар вопрос-ответ, которые были сформулированы разметчиками Амазон на основе 536 статей Википедии. Датасет случайным образом поделен между обучающей (80%), валидационной (10%) и тестовой (10%) выборками. Тестовая выборка не выложена в открытый доступ и используется для оценки обученных моделей[8].

14 сентября 2017 года департамент обработки данных Сбербанка анонсировал вопросно-ответное соревнование с денежным призом[9]. Для

[7] https://github.com/vaskonov/burvec
[8] https://rajpurkar.github.io/SQuAD-explorer/
[9] https://github.com/sberbank-ai/data-science-journey-2017

этого соревнования Сбербанк предоставил вопросно-ответный датасет на русском языке SberQuAD (англ.: Sberbank Question Answering Dataset), содержащий около 50 тысяч обучающих примеров, 15 тысяч валидационных примеров, 25 тысяч тестовых примеров. Тестовый набор данных не был выложен в открытый доступ и использовался для оценки моделей в соревновании. Датасет SberQuAD собран и размечен следуя методологии создания датасета SQuAD. Ниже приведен пример из датасета.

Контекст: *«В 1996 году на свет появилась овца Долли — первое клонированное млекопитающее. Долли была получена методом терапевтического клонирования: в соматическую клетку взрослой овцы вместо её собственного ядра поместили ядро из ооцита другой овцы и получившуюся гибридную клетку стимулировали электрошоком, после чего клетка начала делиться...»*

Вопрос: *«Первое клонированное млекопитающее?»*

Ответ: *«овца Долли»*

Анализ ошибок с помощью системы проверки правописания от «Яндекса»[10] обнаружил 2,646 грамматических ошибок в тренировочной выборке и 287 – в тестовой. Кроме того, в 385(51) вопросах в тренировочной(тестовой) выборке была обнаружена частица *ли*, которая предполагает *да/нет* ответ, что не соответствует правилам датасета SQuAD (ответом должна быть подстрока из текста). После обучения нескольких разнотипных моделей авторами был произведен анализ ошибок. Отобрав 100 вопросов, на которых ошиблись все модели, авторы вручную провели анализ. Результаты анализа приведены в таблице 3.

Большая часть ошибок связана с ошибками разметки, такими как: *неполный ответ* – ответ аннотатора, который содержит только часть правильного ответа; *некорректный вопрос* – вопрос, который является следствием неправильной интерпретации контекста аннотатором; *неправильный ответ* – аннотатор ошибся при выделении ответа на вопрос; *общий вопрос* – вопрос слишком общий, чтобы ответом являлась подстрока контекста; *нет ответа* – ответ на вопрос отсутствует в контексте; *да/нет* вопрос – недопустимый правилами да/нет вопрос. Остальные ошибки возможно отнести к ошибкам QA модели, которая не смогла справится с теми или иными феноменами естественного языка. После исправления найденных ошибок датасет был выложен в открытый доступ[11]. В откорректированном варианте датасет использовался в прошлых работах.

[10] `https://yandex.ru/dev/speller/`

[11] `http://files.deeppavlov.ai/datasets/sber_squad_clean-v1.1.tar.gz`

Таблица 3 — Анализ 100 случайно отобранных ошибок вопросно-ответной модели, обученной на датасете SberQuAD. В верхней части таблицы представлены ошибки, обусловленные некорректной разметкой, в нижней части – ошибки дообученной модели

Категория ошибки	%
неполный ответ	29
неконкретный вопрос	19
неправильный ответ	14
общий вопрос	12
нет ответа	3
да/нет	3
разрешение кореференции	12
грамматические ошибки	6
аргументация (reasoning)	6
парафразы	3

Вопросно-ответный датасет китайского языка DRCD собран и размечен специалистами компании Delta в 2018 году. Датасет содержит более 30,000 вопросов на основе 10,014 параграфов из 2,108 статей Википедии. Тренировочная выборка содержит 26,936 примеров, валидационная выборка – 3,524 и тестовая выборка – 3,493. В отличие от предыдущих датасетов, тестовая выборка датасета DRCD была выложена в открытый доступ[12].

Для обучения вопросно-ответной модели применяется решение на основе BERT и реализованное в библиотеке DeepPavlov [5]. Обучение эффективной NLP модели начинается с подбора подходящей модели BERT, предобученного на целевой язык. Установлено, что модель на основе языко-специфичного BERT превосходит модель на основе многоязычного BERT (M-BERT). Адаптированный M-BERT под русский язык – RuBERT превзошел оригинальный M-BERT на датасетах русского языка: RuSentiment – задача определения тональности, SberQuAD - вопросно-ответная задача, ParaPhraser – задача определения парафраз. При этом предобучение на русский язык задействовало восемь графических ускорителей Tesla P100 в течение двух дней.

Аналогичные результаты были получены для французского языка. Предобученный BERT для французского языка – CamemBERT превзошел M-BERT на датасетах французского языка French Treebank (FTB), а также на французской части XNLI датасета. Предобучение CamemBERT задействовало 256 графических ускорителей NVidia V100 (32 Гб) в течение одного дня. Таким образом, языко-специфичный BERT хоть и улучшает качество NLP моделей на целевом языке, однако его предобучение требует внушительных вычислительных ресурсов, которыми обладают не все академические вычислительные центры или коммерческие организации.

[12] `https://github.com/DRCKnowledgeTeam/DRCD`

Языко-специфичные BERT доказали свое превосходство над M-BERT при использовании обучающей выборки целевого языка. При этом в прошлых экспериментах не использовалось преимущество M-BERT над языко-специфичными BERT, а именно M-BERT позволяет применять многоязычную обучающую выборку. В связи с тем, что современная компьютерная лингвистика фокусируется на ограниченном числе языков, на момент решения задачи для целевого языка уже вероятно существует аналогичный датасет для другого языка (чаще всего английского языка). Следовательно, первая цель эксперимента ответить на вопрос: возможно ли применение M-BERT для экономии на сборе и разметке обучающей выборки целевого языка? Для этого необходимо сравнить качество моделей на основе языко-специфичных BERT и многоязычных BERT в условиях, когда последний дообучен с применением обучающей выборки английского языка.

Предобучение BERT для целевого языка требует внушительных вычислительных ресурсов. Соответственно, следующая цель – выяснить, устранит ли дообучение M-BERT на многоязычных данных необходимость в предобучении языко-специфичного BERT для вопросно-ответной задачи.

Для дообучения QA моделей используются два языка: язык, для которого существует обучающая выборка достаточного объема (во многих случаях это английский язык), и целевой язык – язык тестовой выборки (английский, русский, китайский). В таком случае M-BERT, предобученный на 104 языках, видится чрезмерно универсальным инструментом. Техника языковой фильтрации, предложенная в работе «Load What You Need: Smaller Versions of Multilingual BERT» в 2020 году специалистами компании Geotrend, позволяет убирать векторные представления субтокенов невостребованных языков из M-BERT. Качество дообученых отфильтрованных моделей сопоставимо с оригинальным многоязычным M-BERT на задаче XNLI, при этом отфильтрованные модели более компактные. Дополнительная цель экспериментов – сравнить качество дообученных отфильтрованных моделей с дообученным M-BERT на QA задачах английского, русского и китайского языков.

В данных экспериментах не рассматривались модели на основе дистиллированного BERT, поскольку по результатам прошлых работ дистиллированный BERT существенно уступал оригинальному в решении задач обработки естественного языка, включая разработку вопросно-ответных систем.

В следующей части главы описаны эксперименты по дообучению M-BERT и языко-специфичного BERT в двух режимах. В режиме фиксированного количества эпох (3 эпохи) и в режиме ранней остановки (early stopping, patience=10), то есть модель дообучается до тех пор, пока она

улучшает свое качество на валидационной выборке. Результаты экспериментов для трех языков (английский на SQuAD, русский на SberSQuAD, китайский на DRCD) приведены в таблице 4.

Таблица 4 — Сравнение моделей на основе M-BERT с языко-специфичными BERT для трех датасетов: английский (SQuAD), русский (SberQuAD), китайский (DRCD) в двух режимах: фиксированный режим (3 эпохи), ранней остановки (patience=10)

Модель(обуч выборка)	Тест	Фикс режим		Реж ран ост	
		F1	EM	F1	EM
BERT(SQuAD)	SQuAD	**87.29**	**79.86**	86.67	78.56
M-BERT(SQuAD)		87.2	79.47	87.2	79.47
M-BERT$_{\text{en}}$(SQuAD)		87.02	78.76	87.08	79.07
RuBERT(SberQuAD)	SberQuAD	83.24	64.44	83.41	64.63
M-BERT$_{\text{ru}}$(SberQuAD)		82.41	63.58	82.41	63.58
M-BERT(SQuAD)		73.42	41.89	73.42	41.89
M-BERT$_{\text{en-ru}}$(SQuAD)		73.09	42.54	71.76	40.35
M-BERT(SberQuAD)		82.54	63.32	82.54	63.32
M-BERT$_{\text{en-ru}}$(SberQuAD)		82.51	63.66	82.51	63.66
M-BERT(SberQuAD+SQuAD)		**83.55**	**64.69**	83.44	63.84
M-BERT$_{\text{en-ru}}$(SberQuAD+SQuAD)		83.45	64.46	83.37	64.22
ChBERT(DRCD)	DRCD	88.37	82.51	88.17	82.45
M-BERT$_{\text{ch}}$(DRCD)		87.88	82.19	87.88	82.19
M-BERT(SQuAD)		75.26	62.34	76.41	61.66
M-BERT$_{\text{en-ch}}$(SQuAD)		75.45	61.57	75.42	60.67
M-BERT(DRCD)		88.06	82.79	88.06	82.79
M-BERT$_{\text{en-ch}}$(DRCD)		87.56	81.77	87.94	82.55
M-BERT(DRCD+SQuAD)		**89.99**	**84.49**	89.14	83.65
M-BERT$_{\text{en-ch}}$(DRCD+SQuAD)		89.27	83.47	89.17	83.74

Для экспериментов использовались следующие гиперпараметры: шаг обучения – $2 \cdot 10^{-5}$, размер батчей – 8, оптимизатор – Adam with weight decay, максимальная длина последовательности – 512.

Модели английского языка Модель, дообученная на основе M-BERT, уступает модели, дообученной на BERT в фиксированном режиме. Однако в режиме ранней остановки, модель на основе M-BERT превзошла модель на основе BERT, при этом она дообучалась дольше, чем модель на основе BERT. Модель, дообученная на M-BERT$_{\text{en}}$, незначительно отстает от оригинальной M-BERT, однако M-BERT$_{\text{en}}$ менее требовательна к ресурсам по сравнению с M-BERT. В режиме ранней остановки M-BERT$_{\text{en}}$ дообучается в два раза быстрее, чем M-BERT, при этом достигая сопоставимого с M-BERT качества.

Модели русского языка Модели, дообученные только на английском SQuAD, значительно отстают от остальных моделей, что

говорит о важности дообучения на данных целевого языка. Модели M-BERT(SberQuAD+SQuAD) и RuBERT(SberQuAD) имеют сопоставимое качество, но с небольшим преимуществом M-BERT. Все обозначенные модели имеют сопоставимое качество как в фиксированном режиме, так и в режиме ранней остановки, однако в режиме ранней остановки модели дообучались в среднем быстрее (меньшее количество батчей). Например, модель M-BERT(SberQuAD+SQuAD) в фиксированном режиме дообучалась 46,000 батчей, а в режиме ранней остановки в два раза быстрее – 21,000 батчей. Аналогично остальные модели, кроме M-BERT(SQuAD), для которой потребовалось 17,500 батчей вместо 15,000, что существенно не отразилось на качестве. Фильтрованные модели M-BERT незначительно отстают от оригинальной M-BERT, но при дообучении на многоязычной обучающей выборке отставание сокращается до незначительного. Однако $\text{M-BERT}_{\text{en-ru}}$ содержит на 36% меньше обучаемых параметров и на 243 Мбайт легче, то есть $\text{M-BERT}_{\text{en-ru}}$ менее требователен к ресурсам, чем M-BERT, при незначительной потере в качестве.

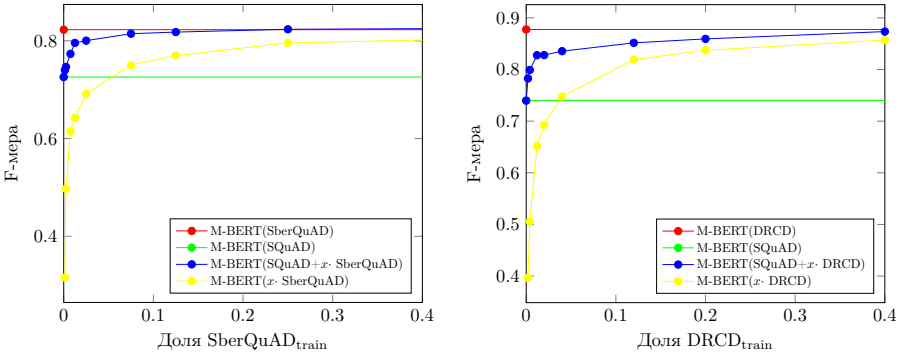
Модели китайского языка Как и в случае с русским языком, необходимо отметить важность дообучения на тренировочных данных целевого языка (все модели дообученные на тренировочной выборке DRCD значительно превосходят M-BERT(SQuAD)). Модель M-BERT(DRCD+SQuAD) превосходит все остальные, что ставит под сомнение целесообразность предобучения языко-специфичных BERT при наличии многоязычной обучающей выборки для вопросно-ответной задачи. Модели, дообученные в режиме ранней остановки, не показывают существенной разницы в качестве по сравнению с моделями, дообученным в фиксированном режиме. Однако, как и в случае с русским языком, при наличии большой многоязычной обучающей выборки, модели, дообученные в режиме ранней остановки, дообучаются быстрее. Фильтрованные модели M-BERT незначительно отстают от оригинальной M-BERT, но это отставание несущественное. При этом $\text{M-BERT}_{\text{en-ch}}$ содержит на 37% меньше обучаемых параметров и на 249 Мбайт легче.

Таким образом, дообучая M-BERT на многоязычной обучающей выборке, представляется возможным обойтись без предобучения языко-специфичного BERT. Однако, как показали эксперименты, наличие обучающей выборки целевого языка необходимо для обучения качественной вопросно-ответной модели. При этом предыдущие эксперименты не ответили на вопрос, какой объем обучающей выборки на целевом языке необходим, поскольку при разработке вопросно-ответной системы задача разработчиков – как можно эффективнее использовать уже доступные данные и минимизировать затраты на сбор и разметку языко-специфичных данных.

Цель экспериментов в этой части – показать, какой объем обучающей выборки на целевом языке необходим для дообучения многоязычного BERT. Для этого были построены кривые обучения (рисунок 2). Модели

обучались независимо для каждого размера обучающей выборки. Обучение происходило в режиме ранней остановки.

Кривые обучения моделей, дообученных на основе M-BERT при тестировании на $\text{SberQuAD}_{\text{test}}$ (русский), представлены на рисунке 2а и на $\text{DRCD}_{\text{test}}$ (китайский) на рисунке 2б.



а) б)

На оси X – доля тренировочной выборки целевого языка, на оси Y – F-мера модели. Зеленая кривая – общая модель для обоих языков M-BERT(SQuAD), которая дообучалась только на $\text{SQuAD}_{\text{train}}$. Красная кривая – модель M-BERT, дообученная целиком на тренировочной выборке целевого языка. Желтая кривая – модель M-BERT, дообученная на доле тренировочной выборке x целевого языка. Синяя кривая – модель M-BERT, дообученная на доле тренировочной выборке x целевого языка совместно с полноразмерным $\text{SQuAD}_{\text{train}}$ английского языка.

Рис. 2 — Кривые обучения моделей, дообученных на основе M-BERT для русского и китайского языков

Добавление полноразмерного английского $\text{SQuAD}_{\text{train}}$ в обучающую выборку значительно улучшает качество модели по сравнению с обучением только на ограниченной выборке набора ($x\cdot$ $\text{SberQuAD}_{\text{train}}$, $x\cdot$ $\text{DRCD}_{\text{train}}$). Модель русского языка, обученная на объединенном наборе данных с 10-15% триплетов на целевом языке, имеет сопоставимое качество с моделью, обученной на полноразмерной обучающей выборке русского языка. Для китайского языка требуется 25-30% триплетов для получения сопоставимого качества. Таким образом, за счет применения переноса знаний и использования общедоступной обучающей выборки другого языка представляется возможным сэкономить на сборе и разметке полноразмерного вопросно-ответного набора данных для целевого языка.

Результаты главы опубликованы в работе «Exploring the BERT Cross-Lingual Transfer for Reading Comprehension» [2]. Модели, обученные в рамках этой главы, выложены в открытый доступ в виде конфигурационных файлов библиотеки DeepPavlov[13].

Четвертая глава посвящена переносу вопросно-ответной модели для решении задачи отслеживания состояния диалога. Диалоговые системы (ДС) – это системы, взаимодействующие с пользователем на естественном языке. Отслеживание состояния диалога (англ.: Dialogue State Tracking, DST) является ключевой функцией диалоговых систем. DST отвечает за определение текущего состояния диалога. Состояние диалога включает в себя намерения пользователя (интенты), пары (слот, значение), соответствующие целям диалога и извлеченные из реплик пользователя. Например, если целью пользователя является заказ такси (интент: *GetRide*), то слотами могут быть *destination*, *number_of_riders*. Кроме того, DST хранит историю состояний ДС (историю взаимодействия системы и пользователя). На рисунке 3 приведен пример отслеживания состояния многодоменного диалога.

В главе подробно описаны виды диалоговых систем, компоненты диалоговых систем, а также приводится обзор диалоговых наборов данных: NegoChat, ATIS, Multi-WOZ. Подробно рассматриваются модели отслеживания состояния диалога, такие как StateNet, HyST, PtrNet, TRADE, BERT-DST.

Для обучения модели DST используется наиболее крупный на данный момент схемоориентированный корпус многодоменных диалогов SGD, собранный и размеченный в 2019 году специалистами компании Google. Корпус содержит диалоги, составляющие 26 сервисов (сервис объединяет в себе несколько однотипных целей), принадлежащих 16 доменам. Чтобы измерить способность ДС к адаптации под новые домены, тестовые наборы содержат сервисы и домены, которых не было в тренировочной выборке. Для каждого сервиса, в SGD предоставляется схема, дающая краткие описания сервиса, всех его слотов и интентов на естественном языке. Схемоориентированный подход предполагает, что при определении состояния диалога используются описания намерений и слотов из схемы. Таким образом, обеспечивается обобщение на новые домены, не участвовавшие в процессе обучения. Схема для сервиса цифрового кошелька приведена на рисунке 4.

Разработанная модель GOLOMB (GOaL-Oriented Multi-task BERT-based dialogue state tracker) представляет собой многозадачную модель на основе BERT для отслеживания состояния целеориентированного диалога. Модель GOLOMB реализована на основе вопросно-ответной модели. Чтобы применить подход, реализованный в рамках обучения вопросно-ответной модели, необходимо переформулировать задачу DST в терминах

[13] `https://github.com/vaskonov/cross_lang_squad`

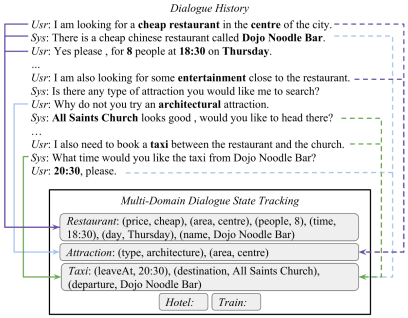


Рис. 3 — Пример отслеживания состояния многодоменного диалога

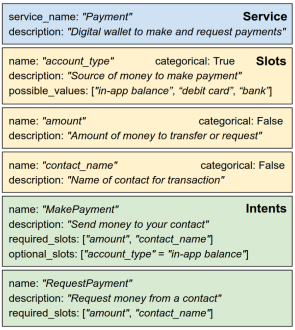


Рис. 4 — Схема для сервиса цифрового кошелька

вопросно-ответной задачи, когда ответ на вопрос извлекается из текста. В качестве контекста, по которому задается вопрос, используется история диалога. «Вопросом» является текстовое описание домена, слота и интента, для которых необходимо найти «ответ» – значение слота, представленное в истории диалога. Таким образом, для определения состояния диалога модель решает задачу поиска подстроки в строке и ряд классификационных задач. Для каждой классификационной задачи существует свой классификатор, реализованный как полносвязный слой нейронной сети без функции активации. Такой подход позволяет одновременно обучаться на информации из истории диалогов и из схемы, описывающих выделяемые значения. Кроме того, он устойчив к изменениям в схеме и не требует дообучения модели для новых намерений и слотов. Модель не требует предварительно рассчитанного векторного представления схемы. Реализованный подход превосходит базовую модель по основной метрике *общая целевая точность* (joint goal accuracy), достигая 53.97% на валидационных данных. Архитектура модели GOLOMB приведена на рисунке 5.

Для дообучения использовалась предварительно обученная модель BERT (bert-large-cased-whole-word-masking-finetuned-squad[14]) с 24 слоями размерностью 1024, 16 головами самовнимания и 340 миллионами параметров. Алгоритм оптимизации AdamW обновляет параметры модели. Начальный коэффициент обучения (learning rate) оптимизатора модели – $3.5 \cdot 10^{-5}$. Модель обучается 5 эпох с размером батча 8 и шагом обновления градиента (gradient accumulation step) – 12 на графическом процессоре Tesla V100 32GB. Реализация модели основана на библиотеке Transformers от компании HuggingFace.

Результаты оценки модели GOLOMB показаны в таблице 5. Предложенная модель превосходит базовую модель по *общей* и *усредненной*

[14] `https://huggingface.co/bert-large-uncased-whole-word-masking-finetuned-squad`

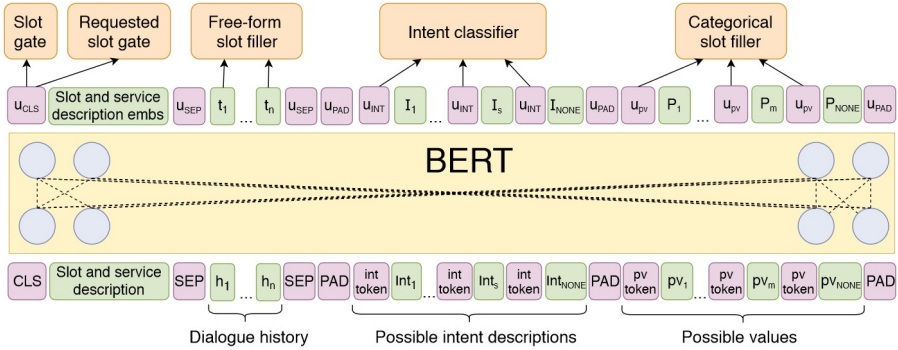


Рис. 5 — Архитектура модели GOLOMB. **Классификатор наличия слота** (slot gate) решает должен ли слот попасть в обновленное состояние диалога. **Классификатор требуемых слотов** (requested slot gate) предсказывает был ли слот запрошен пользователем. **Классификатор намерений** (intent classifier) выбирает активное намерение (intent) пользователя. В зависимости от того, является ли слот категориальным или некатегориальным, используются разные классификаторы. Для некатегориального слота используется **заполнитель слотов подстроками** из текста (free-form slot filler), который выбирает позиции начала и конца значения слота в истории диалога. Для категориального слота **классификатор категориальных слотов** (categorical slot filler) выбирает значение слота среди представленных возможных значений

целевой точности, но базовая модель показывает лучшие результаты на *F-мере запрашиваемых слотов* и *точности намерений*. Возможная причина значительного превосходства базовой модели в определении активных намерений заключается в том, что базовая модель использует выход [CLS] токена для предсказания намерения. Попытка использовать выход токена [CLS] для классификации намерений в предложенной модели приводила к улучшению точности определения намерений, при этом основная метрика (*общая целевая точность*) ухудшалась.

Таблица 5 — Сравнение качества между базовой моделью и предложенной моделью на валидации и на тесте

	Нам. точность active intent acc	Запр. слоты F1 req slot F1	Усред. цел. точность average goal accuracy	Общая цел. точность joint goal accuracy
Базовая модель(вал сет)	**90.8**	**97.3**	74	41.1
GOLOMB(вал сет)	66	96.9	**81.7**	**53.9**
GOLOMB(тест сет)	74.7	97.1	75	46.5

Результаты главы опубликованы в работе «Goal-Oriented Multi-Task BERT-Based Dialogue State Tracker» [3].

В **заключении** приведены основные результаты работы, которые заключаются в следующем:

1. Произведен обзор методов переноса знаний при решении проблем обработки естественного языка.
2. Произведено внутреннее и внешнее сравнение контекстно-независимых векторных представлений слов, обученных разными способами на ограниченной обучающей выборке для трех малоресурсных языков. Обученные векторные представления выложены в открытый доступ.
3. Разработана вопросно-ответная модель на основе BERT с применением библиотеки transformers от HuggingFace.
4. Показано, что межъязыковой перенос позволяет использовать общедоступные обучающие данные английского языка при дообучении M-BERT для вопросно-ответной задачи, таким образом, отсутствует необходимость в сборе полноценного обучающего датасета целевого языка.
5. Экспериментально установлено, что применяя метод межъязыкового переноса, M-BERT, дообученный с применением многоязычной обучающей выборки, имеет сопоставимое качество с языко-специфичными BERT для вопросно-ответной задачи, тем самым отпадает необходимость в вычислительных ресурсах для предобучения языко-специфичных BERT.
6. Разработана оригинальная модель GOLOMB, которая использует вопросно-ответную модель SQuAD для отслеживания состояния диалога.
7. Модели, обученные в рамках работы, доступны в открытом доступе.

Публикации автора по теме диссертации

1. *Konovalov, V.* Learning Word Embeddings For Low Resourse Languages: The Case Of Buryat / V. Konovalov, Z. Tumunbayarova // Komp'juternaja Lingvistika i Intellektual'nye Tehnologii. — 2018. — С. 331—341.
2. Exploring the BERT Cross-Lingual Transfer for Reading Comprehension / V. Konovalov [и др.] // Komp'juternaja Lingvistika i Intellektual'nye Tehnologii. — 2020. — С. 445—453.
3. Goal-oriented multi-task bert-based dialogue state tracker / P. Gulyaev [и др.] // arXiv preprint arXiv:2002.02450. — 2020.
4. Отслеживание состояния целеориентированного диалога на основе БЕРТ / П. А. Гуляев [и др.] // Труды МФТИ. — 2021. — Т. 13, № 3. — С. 48—61.
5. Deeppavlov: An open source library for conversational ai / M. Burtsev [и др.] //. — 2018.

Коновалов Василий Павлович

Методы переноса знаний для нейросетевых моделей обработки естественного языка

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

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