

УДК 004.8

ГРАФОВЫЕ МОДЕЛИ ДЛЯ КОНТЕКСТНОГО ПРОГНОЗИРОВАНИЯ НАМЕРЕНИЙ В ДИАЛОГОВЫХ СИСТЕМАХ

© 2023 г. Д. П. Кузнецов^{1,*}, Д. Р. Леднева^{1,**}

Представлено академиком РАН А.И. Аветисяном

Поступило 31.08.2023 г.

После доработки 15.09.2023 г.

Принято к публикации 15.10.2023 г.

В статье представлена инновационная методология прогнозирования намерений в диалоговых системах на основе графового подхода. Методология заключается в создании графовых структур, представляющих диалоги, с целью отображения контекстной информации. На основе анализа результатов, полученных на различных наборах данных с открытым и закрытым доменом, авторы демонстрируют, что использование графовых моделей в сочетании с текстовыми энкодерами существенно повышает точность прогнозирования намерений. Основное внимание уделяется исследованию влияния различных графовых архитектур и энкодеров на производительность предложенного подхода. Экспериментальные результаты подтверждают превосходство графовых подходов по метрикам точности и вычислительных ресурсов над другими методами по метрике $Recall@k$ (MAR). Данная работа раскрывает новое направление в предсказании намерений в диалоговых системах с использованием графов, внося важный вклад в область обработки естественного языка и машинного обучения.

Ключевые слова: предсказание интенгов, диалоговые системы, графовые нейронные сети

DOI: 10.31857/S2686954323601896, **EDN:** GEBNHK

1. ВВЕДЕНИЕ

Диалоговые системы стали неотъемлемой составляющей повседневной жизни, находя применение в различных сферах. Они представляют собой программное обеспечение, способное коммуницировать с пользователем в процессе диалога, при этом эмулируя поведение, аналогичное реакции человека на информацию, предоставленную пользователем. Одной из фундаментальных задач в моделировании диалоговых систем является определение намерения, выражаемого участником диалога в его реплике, и предсказание намерения, которое будет лежать в основе последующей реплики диалога. В данной исследовательской работе предлагается инновационный подход к решению задачи прогнозирования намерения в диалоговых системах, базирующийся на использовании графовых структур.

Диалог может быть интерпретирован как последовательность намерений, где каждое намерение отражает конкретный запрос или утверждение, высказываемое участниками диалога. Такая

парадигма позволяет представить некоторый набор диалогов в виде полного ориентированного многодольного диалогового графа, вершины которого описывают намерения участников диалога. Каждая доля построенного графа представляет поведение участника через вероятностную модель переходов между его намерениями и намерениями других участников диалога. Рассматривая каждый диалог как путь или подграф многодольного графа, можно преобразовать задачу предсказания намерения следующего высказывания в задачу классификации, где намерение следующего высказывания является целевой переменной. Подобный подход позволяет улучшить точность задачи предсказания намерений, что приводит к более результативному взаимодействию между диалоговой системой и пользователем. Помимо этого, представленный метод может быть особенно важным в контексте решения задачи выбора ответа, где использование графовой структуры диалогов приводит к значительному повышению точности [1], что подчеркивает актуальность и перспективность исследования.

Наш вклад заключается в следующем: (i) Разработка инновационных графовых методов для русского языка, которые обеспечивают точное предсказание намерений следующего высказывания в диалоге. (ii) Оценка производительности нескольких современных текстовых кодировщиков и графовых моделей на различных русско-

¹Московский физико-технический институт (национальный исследовательский университет), Долгопрудный, Московская обл., Россия

*E-mail: kuznetsov.den.p@phystech.edu

**E-mail: da.led@mail.ru

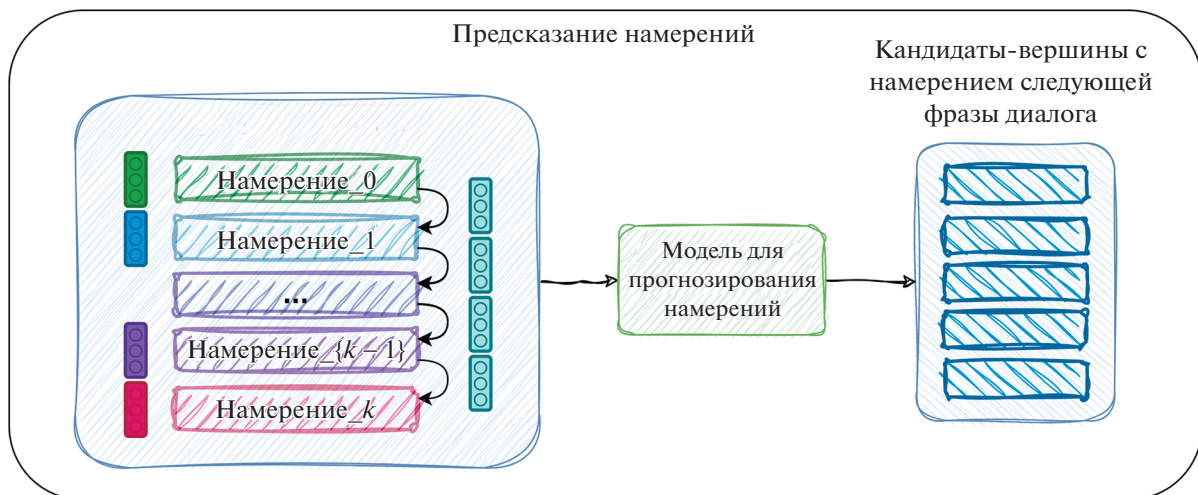


Рис. 1. Прогнозирование намерения следующего высказывания в диалоге с помощью модели, предсказывающей намерение на основании подграфа диалога, а также признаков ребер и вершин.

язычных наборах данных с целью решения задачи предсказания намерений. (iii) Анализ полученных результатов, позволяющий оценить эффективность предложенных подходов на русско-язычных диалоговых системах.

Код доступен здесь (*анонимизированная ссылка*).

2. СОПУТСТВУЮЩИЕ РАБОТЫ

Текстовые энкодеры. Выбор подходящего текстового энкодера играет важную роль в формировании качественных векторных представлений для диалоговых фраз. В последние годы исследования в области обработки естественного языка привели к появлению разнообразных архитектур для кодирования текстовой информации [2]. Одним из наиболее динамично развивающихся направлений стало использование трансформерных моделей, таких как BERT и GPT. Такие языковые модели продемонстрировали впечатляющие результаты в анализе контекста, выявлении семантических связей и генерации текста, что раскрывает новые перспективы для повышения качества диалоговых систем.

Автоматическое построение диалогового графа. Как уже было упомянуто, для решения задачи прогнозирования намерений требуется построение многодольного диалогового графа.

Построение многодольного диалогового графа начинается с формирования вершин, которые в дальнейшем могут быть интерпретированы как намерения участников диалога, исходя из методологии построения графа. Для этого реплики диалогов разбиваются на кластеры с применением двухэтапного алгоритма кластеризации, базирующегося на алгоритме Ллойда (метода К-средних) [3] и семантических и контекстуальных век-

торных представлениях. Подробное описание алгоритма кластеризации и способов получения векторных представлений представлено в разделе “Методология”.

После построения вершин графа следующим шагом является установление связей между ними. Итоговый результат представляется в виде графа, где кластеры выступают в качестве вершин, а направленные ребра обозначают вероятности переходов между кластерами. Важно подчеркнуть, что каждый кластер представляет собой вершину многодольного диалогового графа и задает свое уникальное намерение.

Прогнозирование намерений. Каждый диалог можно интерпретировать как подграф в многодольном диалоговом графе или последовательность переходов между вершинами многодольного графа, соответствующими различным намерениям, содержащимся в репликах диалога. А именно, каждая реплика диалога несет в себе определенное намерение участника диалога, что позволяет определить, к какой вершине многодольного графа реплика диалога относится. Таким образом, при помощи представления диалога в форме ориентированного подграфа диалогового графа, используя методы классификации графов становится возможным предсказать намерение следующего высказывания в диалоге (см. рис. 1).

Для решения задач графовой классификации наиболее распространенным подходом является применение Графовых Нейронных Сетей (GNN) [4]. Обычно графовые методы разделяют на две основные категории: гомогенные и гетерогенные. Гомогенные методы, включая Графовые Сети Внимания (GAT) [5], пользуются высокой популярностью за счет их способности улавливать зависимости между вершинами графа. В рамках ге-

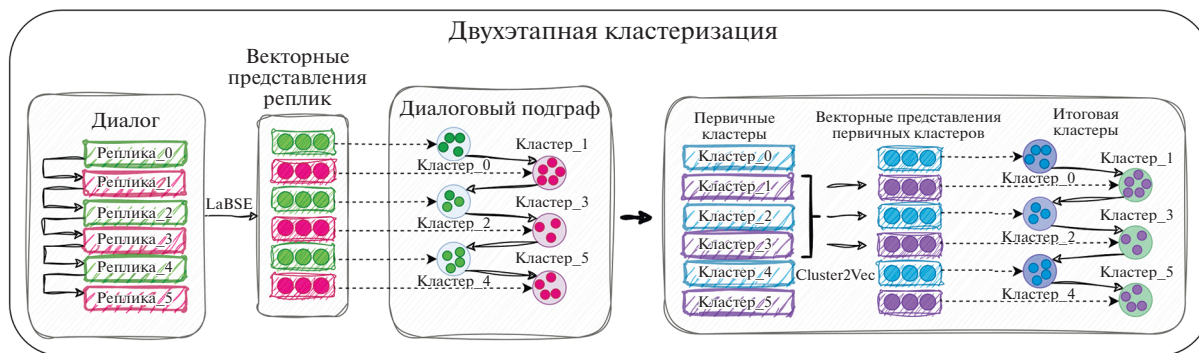


Рис. 2. Двухэтапный алгоритм кластеризации диалоговых высказываний. На первом этапе используется метод кластеризации К-средних для объединения семантически близких высказываний. На втором этапе для кластеров формируются контекстуальные векторные представления с использованием Cluster2Vec. Этот процесс приводит к созданию вершин многодольного диалогового графа.

терогенных методов используются Графовые Трансформерные Сети (GTN) [6], которые выявляют информативные связи между вершинами, создавая на основе исходной графовой структуры новые структуры графа, называемые метапутями.

3. МЕТОДОЛОГИЯ

Автоматическое построение диалогового графа.

Для эффективного решения задач, связанных с классификацией графов и предсказанием намерений в диалоговых системах с разным числом участников и разнообразными ролями, необходимо сформировать многодольный полный ориентированный диалоговый граф [7]. Такой граф будет отражать различные роли в диалоговой системе, поведение участников диалога и их взаимодействие.

К примеру, в диалоговых системах с закрытым доменом принято выделять две роли — пользователя и менеджера. Соответственно, в диалоговом графе целесообразно сформировать две доли, каждая из которых соответствует одной из ролей. В случае диалоговых систем с открытым доменом, как правило, присутствует только одна роль — роль участника диалога. Однако существуют исключения, например, набор диалоговых данных Matreshka, который, несмотря на открытый домен, включает две разные роли: пользователя и бота. Это подчеркивает важность гибкости методологии в обработке разнообразных диалоговых сценариев.

Построение вершин многодольного диалогового графа осуществляется при помощи двухэтапного алгоритма кластеризации (см. рис. 2). На первом этапе происходит группирование реплик с использованием аппроксимации метода k-means из библиотеки FAISS. В результате образуются первичные кластеры, содержащие диалоговые высказывания, обладающие схожими век-

торными представлениями и семантической структурой. На втором этапе для каждого из первичных кластеров создаются контекстуальные векторные представления с применением метода Cluster2Vec. Суть данного метода заключается в интерпретации диалога как последовательности номеров кластеров, к которым принадлежат соответствующие реплики диалога. На основе сформированных последовательностей происходит обучение Word2Vec, где числа, представляющие номера кластеров, выполняют роль “слов”. Подход Cluster2Vec позволяет получить векторные представления первичных кластеров, которые включают в себя контекстную информацию, описывающую окружение, в котором встречаются реплики каждого кластера в рамках диалогов. Далее, с использованием аппроксимации метода k-means из библиотеки FAISS на основе контекстуальных векторных представлений осуществляется объединение первичных кластеров в итоговые кластеры — будущие вершины многодольного диалогового графа. Таким образом, вершины диалогового графа содержат в себе высказывания с идентичной семантикой, которые встречаются в диалогах в схожих контекстах.

Предварительная обработка данных. Этап предварительной обработки данных включает в себя подготовку набора диалоговых данных для обучения моделей прогнозирования намерений в диалогах. Для предсказания следующего намерения используется информация из последних m высказываний диалога с использованием метода скользящего окна длиной m . Если история диалога на момент предсказания короче, чем m , добавляется нулевая вершина, указывающая на отсутствие высказывания в истории. Каждый фрагмент диалога из получившейся выборки представляется в виде ориентированного подграфа многодольного диалогового графа (см. рис. 3). Вершины подграфа соответствуют вершинам многодольного диа-

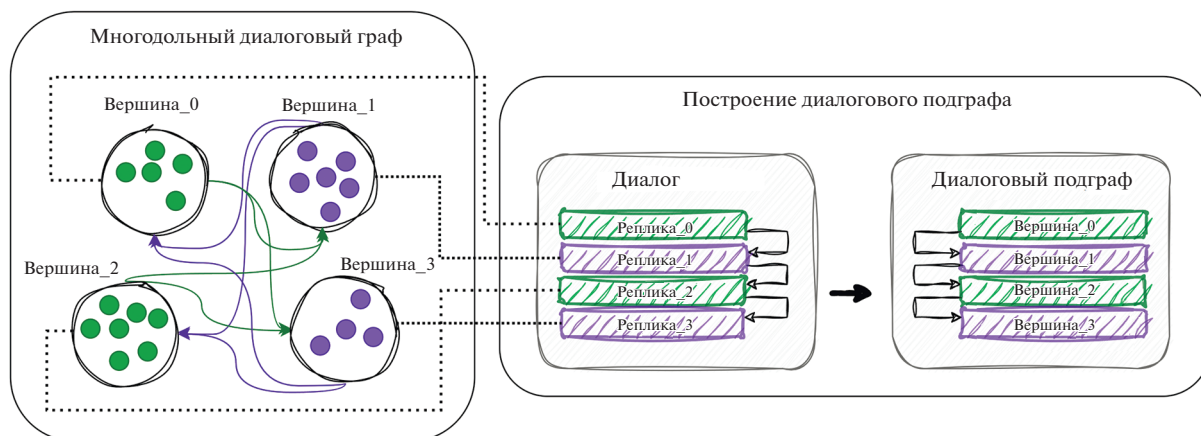


Рис. 3. Представление фрагмента диалога в виде подграфа многодольного диалогового графа. Вершины в подграфе соответствуют тем вершинам многодольного графа, которые содержат высказывания фрагмента диалога.

логового графа, в которых расположены высказывания фрагмента диалога, на основе которого построен подграф. Далее создаются необходимые признаки для вершин и ребер каждого подграфа.

Метрики. В данном исследовании метрики *Recall@k* и *MAR (Mean Average Recall)* использовались для точности прогнозирования намерения следующего высказывания предложенными подходами. Для каждого построенного по диалогу из тестовой выборки графа присваивается значение 1, если вершина с намерением следующей реплики входит в топ- k вершин, сформированный на основе предсказанных вероятностей, и значение 0 – в противном случае. Затем для каждого k полученные значения усредняются по всем репликам и диалогам. Для оценки распределения метрики *Recall@k* и определения, насколько точно предсказываются вершины-кандидаты, использовались различные значения k : {1, 3, 5, 10}. Метрика *MAR* является средним арифметическим по значениям $k \in \{1, 3, 5, 10\}$ метрики *Recall@k*.

Кроме того, чтобы выделить различия в кластерах, сформированных для разных ролей в наборах данных с закрытым доменом, мы представляем отдельные результаты для прогнозирования намерений пользователя и диалоговой системы.

4. ПРЕДЛАГАЕМЫЕ ПОДХОДЫ

В этом разделе представлен обзор использованных подходов для прогнозирования намерения следующего высказывания в диалоге. Главной задачей каждого подхода является предсказание по диалоговому подграфу намерения следующего высказывания диалога, что аналогично предсказанию вершины в многодольном диалоговом графе, так как каждая вершина в многодольном диалоговом графе представляет уникальное намерение. В математической форме постановку задачи можно представить следующим образом: дан диалог $U =$

$= \{u_1, u_2, \dots, u_t\}$, t – количество высказываний в диалоге, u_i – i -е высказывание в диалоге. Для каждой реплики диалога u_i известна вершина v_i с намерением реплики, соответственно, диалог представляется в форме ориентированного подграфа многодольного диалогового графа $G = (V, E)$, где $V = \text{unique}(\{v_1, v_2, \dots, v_t\})$ – это набор вершин подграфа, а $E = \{(v_1, v_2), (v_2, v_3), \dots, (v_{t-1}, v_t)\}$ – это набор ребер подграфа.

Markov Chain. Базовый подход, в котором на основе набора диалоговых данных вычисляются вероятности перехода из одной вершины в вершины других долей многодольного диалогового графа. Самые вероятные вершины для перехода из текущей вершины графа рассматриваются как потенциальные кандидаты для вершины с намерением следующей реплики в диалоге.

Encoder. Другим базовым подходом является использование предварительно обученных языковых моделей. В данном исследовании для этой цели был выбран энкодер LaBSE, который продемонстрировал лучшие результаты при формировании вершин многодольного диалогового графа. В рамках данного метода с применением текстового энкодера получаются векторные представления как предыдущих высказываний из истории диалога, так и потенциальных последующих реплик. На основе векторных представлений фраз из истории диалога осуществляется предсказание потенциальных последующих высказываний в диалоге и их намерений. Формально подход может быть представлен следующим образом: k кандидатов-вершин многодольного графа с намерением следующей реплики диалога u_{t+1} определяются как

$$P(s) = \text{LaBSE}(u_t) \cdot \text{LaBSE}(s)^T, \quad (1)$$

$$\text{candidates} = \arg \max_{|\text{unique}(c(s_i))| = k, s \in S} P(s), \quad (2)$$

где S — реплики из всех диалогов обучающей выборки, $c(s_i)$ — функция, по реплике возвращающая номер вершины многодольного графа, в котором содержится реплика. Полученные на первом шаге значения $P(s)$, $s \in S$ сортируются по убыванию, также для каждой реплики из обучающей выборки определяется номер кластера. Выбор кандидатов-вершин происходит с начала отсортированного списка значений $P(s)$ до тех пор, пока количество уникальных кандидатов-вершин не станет равным k .

Encoder-MAP. С целью проведения всестороннего сравнительного анализа был разработан метод, который представляет собой русскоязычную контекстуальную модель, включающий в себя дообучение языковой модели для задачи предсказания последующей реплики на основе истории диалога. В данной концепции LaBSE выступает в качестве базовой модели, которая дополняется слоем Feed Forward Neural Network, содержащим отдельные матрицы весов для векторных представлений истории диалога и потенциальных следующих реплик. Выбор кандидатов-вершин происходит по аналогии с предсказанием кандидатов-вершин в рамках подхода Encoder.

CatBoost. Результаты экспериментов показали, что среди библиотек градиентного бустинга наилучшей эффективностью для прогнозирования намерений обладает CatBoost. В рамках реализации подхода для каждого подграфа формируется вектор, который представляет собой объединение признаков всех вершин, входящих в этот подграф. Затем полученный вектор передается в качестве входных данных алгоритму для решения задачи предсказания намерений.

Message Passing. Графовые Нейронные Сети представляют собой один из наиболее широко распространенных методов в решении задач, связанных с анализом графовых структур. Среди многочисленных моделей графовых нейронных сетей особый интерес представляют Графовые Сети Внимания (GAT). Графовые Сети Внимания оперируют механизмом внимания, который позволяет определить степень важности сообщений, поступающих из соседних вершин, для агрегации информации между вершинами.

FastGTN. В графах, сформированных на основе диалогов, вершины имеют разные типы, зависящие от того, какая вершина многодольного диалогового графа соответствует текущей вершине. Для решения задач, связанных с такими графами, могут быть применены Графовые Трансформерные Сети (GTN). В представленной работе используется FastGTN — усовершенствованная реализация GTN.

5. ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ

В данном разделе приведены результаты применения различных подходов на наборах диалоговых данных с открытым доменом (см. табл. 1) и наборах диалоговых данных с закрытым доменом (см. табл. 2), где подходы оцениваются с использованием метрики *MAR* (*Mean Average Recall*).

В разделе обсуждения результатов можно выделить следующие ключевые аспекты исследования.

Сравнительная таблица результатов. В разделе обсуждения результатов представлено сравнение (см. табл. 3) различных подходов на основе подробных результатов оценивания с использованием метрики *Recall@k*, $k \in \{1, 3, 5, 10\}$. Каждому методу присваивается единица, если его результаты по конкретной метрике являются лучшими, в противном случае — 0. Затем суммируются значения по всем метрикам для каждого подхода и датасета.

Преимущество моделей на основе графов. Из таблицы сравнения видно, что наилучший результат продемонстрировал подход, использующий метод Message Passing (MP). Однако Graph Transformer Network (GTN) также продемонстрировал высокий уровень производительности, превосходя градиентный бустинг и текстовые подходы. Следует отметить, что Message Passing и Graph Transformer Networks также выделяются более высокой скоростью выполнения и низкими требованиями к вычислительным ресурсам в сравнении с другими подходами. Это подчеркивает их эффективность и практическую применимость при использовании диалоговых графов для предсказания намерений.

Превосходство графовых моделей над текстовыми подходами. Графовые модели продемонстрировали превосходство над текстовыми архитектурами в решении задачи прогнозирования намерений в диалогах. В частности, графовые модели показали результат лучше как в сравнении с простым использованием текстового кодировщика, так и в сравнении с дополнительно обученным текстовым кодировщиком Encoder-MAP. Это подтверждает важность учета структурных взаимосвязей между элементами диалога.

Асимметричность ролей в диалогах. В процессе анализа метрик на наборах данных с закрытым доменом было выявлено значительное различие в результатах между метриками пользователя и метриками диалоговой системы. Это различие обусловлено асимметричностью ролей, которые субъекты играют в диалоге, и подчеркивает важность учета асимметричности ролей в будущих исследованиях.

Таким образом, результаты данного исследования подтверждают важность применения гра-

Таблица 1. Результаты экспериментов по метрике Mean Average Recall на наборах данных открытого домена. Средняя метрика — среднее значение метрики пользователя и метрики диалоговой системы. Для стабильности результатов все методы были обучены на 3 разных наборах кластеров, полученные результаты были усреднены

Подход	# Параметров	Отн. время обучения	Датасет		Toloka Persona Chat Rus	Russian Dialogues SUBSET	Matreshka		
			# Кластеров				Пользователь	Диалоговая система	Среднее
			Первый этап	Второй этап					
Markov Chain	10K	0.13	200	30	43.13 ± 0.88	35.62 ± 3.36	45.19 ± 1.01	60.78 ± 1.37	52.19 ± 0.60
			400	60	31.59 ± 0.64	22.03 ± 0.37	35.80 ± 0.78	48.41 ± 0.47	41.46 ± 0.45
			800	120	22.37 ± 0.52	15.21 ± 0.42	26.97 ± 0.73	35.05 ± 1.07	30.67 ± 0.35
Encoder	471M	0.50	200	30	35.23 ± 1.02	32.89 ± 1.86	31.54 ± 1.60	54.55 ± 1.03	43.05 ± 1.15
			400	60	23.09 ± 0.58	21.73 ± 0.75	20.94 ± 0.77	40.74 ± 0.80	30.84 ± 0.57
			800	120	15.55 ± 0.45	15.70 ± 0.36	14.36 ± 0.61	27.76 ± 0.73	21.06 ± 0.64
Message Passing	471M + 3.7M	0.47	200	30	49.18 ± 0.76	37.67 ± 1.67	51.45 ± 1.36	72.08 ± 0.96	60.81 ± 0.87
			400	60	37.95 ± 1.14	24.59 ± 0.53	41.37 ± 0.84	59.96 ± 0.53	49.79 ± 0.60
			800	120	27.58 ± 0.31	18.20 ± 0.80	30.61 ± 0.86	46.94 ± 0.81	37.94 ± 0.68
CatBoost	471M+2.2M	1.00	200	30	48.10 ± 0.54	36.67 ± 1.29	52.26 ± 1.55	72.40 ± 0.53	61.41 ± 0.81
			400	60	36.68 ± 0.58	24.20 ± 0.67	42.01 ± 0.83	59.08 ± 1.16	49.70 ± 0.68
			800	120	26.40 ± 0.53	17.34 ± 0.37	32.99 ± 1.17	45.26 ± 0.43	38.52 ± 0.69
FastGTN	471M+1.9M	0.49	200	30	48.02 ± 1.09	37.99 ± 2.26	52.36 ± 2.08	71.75 ± 0.83	61.20 ± 0.89
			400	60	36.95 ± 0.69	23.11 ± 1.06	40.09 ± 0.97	59.34 ± 1.03	48.79 ± 0.92
			800	120	25.77 ± 0.29	17.64 ± 0.69	32.36 ± 0.71	47.08 ± 0.57	39.05 ± 0.51
Encoder-MA	471M+2M	0.78	200	30	38.92 ± 1.12	35.75 ± 2.40	40.60 ± 1.55	64.88 ± 1.69	52.74 ± 1.44
			400	60	27.32 ± 1.43	20.94 ± 1.47	24.45 ± 1.47	50.82 ± 1.22	37.64 ± 0.97
			800	120	16.54 ± 1.36	12.99 ± 0.56	16.49 ± 0.77	35.47 ± 1.86	25.98 ± 1.08

Таблица 2. Результаты экспериментов по метрике Mean Average Recall на наборах данных закрытого домена. Средняя метрика — среднее значение метрики пользователя и метрики диалоговой системы. Для стабильности результатов, все методы были обучены на 3 разных наборах кластеров, полученные результаты были усреднены

Подход	# Параметров	Отн. время обучения	Датасет		Russian MultiWOZ			Telecom Domain Dataset			Russian FoCus			Russian Taskmaster		
			# Кластеров		Пользователь	Диалоговая система	Среднее	Пользователь	Диалоговая система	Среднее	Пользователь	Диалоговая система	Среднее	Пользователь	Диалоговая система	Среднее
			Первый этап	Второй этап												
Markov Chain	10K	0.13	200	30	57.27 ± 0.70	71.45 ± 0.72	64.37 ± 0.48	39.11 ± 2.72	51.55 ± 1.34	45.33 ± 1.71	44.82 ± 1.81	48.75 ± 0.74	46.79 ± 1.09	52.17 ± 0.60	51.70 ± 0.53	52.01 ± 0.53
			400	60	44.24 ± 1.31	59.90 ± 0.53	52.07 ± 0.63	24.43 ± 1.96	36.58 ± 2.13	30.51 ± 1.20	34.31 ± 2.64	38.50 ± 0.67	36.40 ± 1.50	44.62 ± 0.30	43.48 ± 0.20	44.13 ± 0.24
			800	120	29.91 ± 1.24	41.31 ± 0.53	35.61 ± 0.64	16.36 ± 3.98	20.24 ± 0.78	18.30 ± 1.80	28.28 ± 0.27	28.10 ± 0.36	28.19 ± 0.24	36.65 ± 0.64	35.58 ± 0.37	36.18 ± 0.42
Encoder	471M	0.50	200	30	28.08 ± 2.10	55.09 ± 1.45	41.59 ± 1.55	18.03 ± 1.92	37.61 ± 3.09	27.82 ± 2.48	24.27 ± 0.81	46.87 ± 0.58	35.57 ± 0.62	31.27 ± 0.10	35.24 ± 0.96	33.25 ± 0.45
			400	60	16.43 ± 1.06	39.30 ± 1.69	27.87 ± 0.97	8.64 ± 1.14	20.43 ± 0.50	14.54 ± 0.69	17.82 ± 0.44	34.67 ± 0.33	26.24 ± 0.22	20.11 ± 0.25	25.48 ± 0.29	22.80 ± 0.22
			800	120	9.41 ± 0.40	24.02 ± 0.66	16.71 ± 0.51	5.69 ± 1.97	9.91 ± 0.46	7.80 ± 1.10	11.46 ± 0.27	25.34 ± 0.32	18.40 ± 0.25	13.48 ± 0.16	16.37 ± 0.31	14.93 ± 0.23
Message Passing	471M+3.7M	0.47	200	30	63.34 ± 1.49	81.48 ± 0.12	72.41 ± 0.75	39.89 ± 1.03	63.74 ± 1.24	51.82 ± 0.87	56.56 ± 0.36	64.24 ± 0.74	60.40 ± 0.37	63.74 ± 0.50	68.72 ± 0.30	66.28 ± 0.14
			400	60	49.74 ± 0.73	71.27 ± 1.01	60.50 ± 0.56	25.15 ± 3.44	44.03 ± 0.71	34.59 ± 1.47	44.90 ± 0.24	54.95 ± 0.45	49.93 ± 0.31	58.06 ± 0.70	64.34 ± 0.20	61.26 ± 0.40
			800	120	36.68 ± 0.61	52.94 ± 0.51	44.81 ± 0.48	15.41 ± 2.40	25.22 ± 0.88	20.31 ± 0.97	42.58 ± 1.65	44.52 ± 0.36	43.55 ± 0.75	49.59 ± 0.54	54.71 ± 0.57	52.20 ± 0.40
CatBoost	471M+2.2M	1.00	200	30	65.90 ± 1.52	80.21 ± 0.85	73.05 ± 0.62	42.10 ± 1.70	62.67 ± 0.70	52.38 ± 0.60	46.47 ± 0.60	62.52 ± 0.40	54.49 ± 0.44	64.51 ± 0.13	69.66 ± 0.35	67.14 ± 0.19
			400	60	51.55 ± 0.75	68.66 ± 0.57	60.11 ± 0.49	25.57 ± 1.21	44.52 ± 2.28	35.05 ± 1.15	44.22 ± 0.76	51.99 ± 0.28	48.11 ± 0.37	57.13 ± 0.26	63.42 ± 0.48	60.34 ± 0.18
			800	120	36.36 ± 0.57	49.98 ± 0.95	43.17 ± 0.64	17.34 ± 3.35	27.20 ± 1.20	22.27 ± 1.22	38.62 ± 3.90	40.80 ± 0.41	39.71 ± 2.11	49.20 ± 0.37	54.96 ± 0.69	52.15 ± 0.37
FastGTN	471M+1.9M	0.49	200	30	63.85 ± 1.43	80.97 ± 0.90	72.41 ± 0.72	41.09 ± 1.58	61.75 ± 1.25	51.41 ± 1.28	52.65 ± 1.42	63.77 ± 0.57	58.20 ± 0.73	64.48 ± 0.56	69.44 ± 0.42	67.01 ± 0.37
			400	60	51.96 ± 0.76	71.78 ± 1.13	61.87 ± 0.83	26.01 ± 1.95	43.36 ± 1.32	34.69 ± 0.76	42.00 ± 0.76	54.04 ± 0.89	48.02 ± 0.64	57.09 ± 0.55	63.33 ± 0.35	60.27 ± 0.28
			800	120	36.63 ± 0.88	51.88 ± 1.32	44.25 ± 0.91	18.97 ± 3.25	26.95 ± 0.96	22.96 ± 1.20	39.77 ± 2.41	42.85 ± 0.49	41.30 ± 0.97	49.42 ± 0.35	53.43 ± 0.42	51.48 ± 0.36
Encoder-MAP	471M+2M	0.78	200	30	48.14 ± 1.37	68.39 ± 2.39	58.27 ± 1.00	27.17 ± 4.00	42.91 ± 2.13	35.04 ± 2.70	43.78 ± 1.16	53.78 ± 0.88	48.78 ± 0.74	54.10 ± 0.63	54.52 ± 0.51	54.30 ± 0.47
			400	60	36.38 ± 0.98	57.83 ± 1.19	47.11 ± 0.87	11.40 ± 0.78	25.05 ± 1.15	18.23 ± 0.93	30.27 ± 0.88	42.55 ± 0.44	36.41 ± 0.45	45.03 ± 0.57	44.47 ± 0.57	44.75 ± 0.40
			800	120	22.38 ± 0.53	38.05 ± 0.59	30.22 ± 0.50	7.52 ± 1.77	14.21 ± 2.81	10.86 ± 1.31	24.21 ± 3.36	27.99 ± 0.89	26.10 ± 1.69	35.44 ± 0.68	32.70 ± 1.02	34.07 ± 0.81

Таблица 3. Сравнение подходов прогнозирования намерений на основе подробных результатов исследования

Dataset	Markov Chain	Encoder	Message Passing	CatBoost	FastGTN	Encoder-MAP
Russian MultiWOZ	0	0	10	5	12	0
Telecom Domain Dataset	0	0	7	12	11	0
Russian FoCus	0	0	12	2	6	0
Russian Taskmaster	0	0	9	8	7	0
Matreshka	0	0	12	11	10	0
Toloka Persona Chat Rus	0	0	12	6	8	0
Russian Dialogues SUBSET	6	0	10	10	9	3
Summary	6	0	72	54	63	3

фовых моделей для анализа диалоговых данных и их потенциал в улучшении точности и эффективности в задачах прогнозирования намерений в диалогах.

6. ЗАКЛЮЧЕНИЕ

Исследование демонстрирует превосходство графовых моделей над текстовыми архитектурами в контексте задачи прогнозирования намерений в диалогах. Графовые модели и метод градиентного бустинга демонстрируют сопоставимое качество, однако Message Passing и Graph Transformer Networks выделяются более высокой скоростью работы и требуют меньших затрат вычислительных ресурсов. Мы обнаружили различия в метриках, оценивая модели на данных с несколькими ролями в диалогах, что подтверждает гипотезу о важности формирования отдельных кластеров, отображающих намерения, для каждой роли.

Текстовые подходы, такие как Encoder-MAP, демонстрируют низкие результаты в сравнении с графовыми моделями, что подчеркивает важность извлечения структурных зависимостей из данных. Также дополнительное обучение повышает эффективность текстовых подходов, но по сравнению с моделью на основе цепей Маркова качество текстовых архитектур оказывается ниже.

Приложение

А. СРАВНЕНИЕ ТЕКСТОВЫХ КОДИРОВЩИКОВ

Для создания вершин многодольного диалогового графа требуется сформировать векторные представления высказываний из набора диалоговых данных, на основе которых будет построен граф. В этом исследовании мы воспользуемся векторными представлениями диалоговых реплик, полученными с применением текстового энкодера Language-agnostic BERT Sentence Em-

bedding (LaBSE) [8]. Выбор текстового энкодера LaBSE обусловлен результатами сравнительного анализа современных архитектур энкодеров предложений (см. табл. 4). В рамках данного анализа была проведена оценка способности текстовых энкодеров учитывать контекст при преппроессинге векторных представлений высказываний. Исследование продемонстрировало, что точность предсказания намерений в значительной степени зависит от качества векторных представлений предложений.

В. ПОДРОБНЫЕ РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ

В данном разделе приведены подробные результаты применения различных подходов на наборах диалоговых данных с закрытым доменом (см. табл. 5) и наборах диалоговых данных с открытым доменом (см. табл. 6), где подходы оцениваются с использованием метрики $Recall@k$, $k \in \{1, 3, 5, 10\}$.

С. ЭКСПЕРИМЕНТЫ

Наборы диалоговых данных. В данном исследовании подходы были оценены на различных наборах данных, как с открытым, так и с закрытым доменом. Среди наборов данных с закрытым доменом присутствовали MultiWOZ, FoCus и Taskmaster, переведенные с английского языка на русский язык при помощи модели NLLB для машинного перевода. Кроме того, использовался набор данных Telecom Domain Dataset, содержащий диалоги, связанные с телекоммуникациями.

Наборы диалоговых данных с открытым доменом включали набор диалоговых данных Toloka Persona Chat Rus, содержащий диалоги на тему знакомств, а также набор диалоговых данных Russian Dialogues, содержащий диалоги на общие темы. Учитывая, что общий объем датасета Russian Dialogues составляет более миллиона диалогов, что требует значительных вычислительных

Таблица 4. Сравнение качества векторных представлений реплик диалогов с использованием различных текстовых кодировщиков на датасете Russian MultiWOZ. Исследование влияния выбора кодировщиков предложений на производительность трех основных подходов: Message Passing, Encoder, Markov Chain

Модель	LASER	Rubert-Tiny	MMPNet	LaBSE	DistilRuBERT	SBERT-large	RuBERT
# Параметров	52M	11.9M	278M	471M	135M	427M	178M
Encoder							
Recall@1	10.95 ± 0.353	7.39 ± 0.260	20.35 ± 1.791	15.75 ± 0.643	9.53 ± 0.238	14.26 ± 0.359	8.71 ± 0.924
Recall@3	27.25 ± 0.932	24.24 ± 0.547	43.31 ± 0.518	35.25 ± 1.662	27.56 ± 0.538	35.98 ± 0.049	20.35 ± 0.912
Recall@5	37.94 ± 0.508	34.73 ± 0.938	53.88 ± 0.758	47.84 ± 2.355	39.60 ± 1.140	47.33 ± 0.256	29.47 ± 0.725
Recall@10	56.44 ± 0.902	55.00 ± 1.067	69.41 ± 0.095	67.51 ± 1.521	59.75 ± 1.343	64.69 ± 0.121	48.82 ± 0.526
Message Passing							
Recall@1	40.87 ± 0.546	37.61 ± 2.028	44.31 ± 1.909	45.67 ± 0.468	39.25 ± 1.113	40.20 ± 1.597	28.60 ± 0.361
Recall@3	67.30 ± 0.536	64.18 ± 0.378	70.01 ± 0.716	71.57 ± 0.819	66.03 ± 0.675	69.31 ± 0.902	54.47 ± 0.825
Recall@5	78.39 ± 0.316	75.27 ± 0.660	80.18 ± 0.550	80.98 ± 0.776	77.05 ± 0.661	80.37 ± 0.663	67.03 ± 0.785
Recall@10	91.16 ± 0.281	89.07 ± 0.438	91.50 ± 0.097	91.43 ± 0.920	89.69 ± 0.554	91.61 ± 0.385	83.63 ± 0.395
Markov Chain							
Recall@1	30.62 ± 0.186	27.46 ± 1.086	34.00 ± 0.516	35.68 ± 0.665	27.97 ± 0.573	32.17 ± 1.419	19.80 ± 0.822
Recall@3	55.48 ± 0.440	53.29 ± 0.034	60.00 ± 0.779	61.45 ± 0.588	55.17 ± 0.497	60.82 ± 0.938	42.71 ± 1.472
Recall@5	68.72 ± 0.628	65.85 ± 0.271	71.55 ± 0.495	73.37 ± 0.499	67.59 ± 0.970	73.19 ± 0.694	55.39 ± 1.189
Recall@10	83.89 ± 0.364	82.84 ± 0.490	86.55 ± 0.188	86.97 ± 0.168	84.01 ± 0.756	86.82 ± 0.339	73.77 ± 1.527

ресурсов для оценки методов, в исследовании было использовано подмножество датасета Russian Dialogues. Также использовался набор диалоговых данных Matreshka, представляющий собой датасет открытого домена, отличительной особенностью которого является явное выделение не одной, а двух ролей — бота и пользователя.

Количество кластеров. На первом и втором этапах кластеризации в исследовании использовались разные конфигурации количеств кластеров: 200 и 30, 400 и 60, 800 и 120 кластеров соответственно. Эти значения были выбраны согласно методологии, описанной в статье [7]. Тем не менее оптимальное количество кластеров для каждого набора данных следует подбирать на основе его индивидуальных характеристик.

Детали реализации. Все модели обучались с применением оптимизатора Adam, способствующему более быстрой сходимости моделей. Для агрегации информации из вершин графа во всех графовых подходах использовался модуль объединения. Кроме того, для выполнения задачи классификации графов во все графовые подходы был добавлен линейный слой.

Для предотвращения переобучения были применены два метода: ранняя остановка, который

отслеживает производительность модели на валидационной выборке и останавливает обучение при отсутствии улучшений, и планировщик Reduce Learning Rate on Plateau, который автоматически снижает скорость обучения. Гиперпараметры были настроены с использованием значений по умолчанию, которые адаптировались под конкретную структуру диалоговых графов.

Чтобы учесть возможную зависимость качества подходов от конкретных конфигураций разбиения диалоговых реплик на кластеры, все подходы обучались и оценивались на трех различных наборах кластеров. Затем полученные значения метрик для различных наборов кластеров были усреднены, чтобы получить более надежную оценку качества подходов.

D. ОГРАНИЧЕНИЯ

Хотя исследование представляет многообещающие результаты, следует учесть следующие ограничения:

Мультиязычность и разнообразие доменов. Одним из главных ограничений нашего исследования является его ориентированность на русскоязычные диалоговые наборы данных. Хотя такое

Таблица 5. Результаты экспериментов по метрике $Recall@k$, $k \in \{1, 3, 5, 10\}$ на наборах данных закрытого домена. Средняя метрика – среднее значение метрики пользователя и метрики диалоговой системы. Для стабильности результатов все методы были обучены на 3 разных наборах кластеров, полученные результаты были усреднены

Подход	# Параметров	Относительное время обучения	Dataset			Russian MultiWOZ			Telecom Domain Dataset			Russian FoCus			Russian Taskmaster		
			# Кластеров	Метрика		Пользователь	Диалоговая система	Среднее	Пользователь	Диалоговая система	Среднее	Пользователь	Диалоговая система	Среднее	Пользователь	Диалоговая система	Среднее
Markov Chain	10K	0.13	200	30	Recall@1	28.26 ± 0.670	43.09 ± 0.662	35.68 ± 0.665	17.05 ± 3.689	22.68 ± 0.590	19.86 ± 2.114	21.58 ± 2.149	24.18 ± 0.989	22.88 ± 1.172	24.13 ± 0.343	25.32 ± 0.557	24.82 ± 0.450
						51.91 ± 0.464	70.98 ± 1.552	61.45 ± 0.588	33.09 ± 3.830	46.30 ± 2.433	39.69 ± 2.438	39.55 ± 1.648	42.35 ± 0.485	40.95 ± 0.959	46.16 ± 0.679	45.09 ± 0.215	45.71 ± 0.444
			400	60	Recall@5	65.91 ± 1.486	80.83 ± 0.502	73.37 ± 0.499	42.92 ± 2.647	60.43 ± 1.284	51.67 ± 1.889	49.63 ± 1.489	53.72 ± 0.729	51.68 ± 1.016	59.52 ± 0.864	58.09 ± 0.709	58.88 ± 0.773
						83.02 ± 0.174	90.92 ± 0.164	86.97 ± 0.168	63.36 ± 0.723	76.81 ± 1.064	70.09 ± 0.400	68.53 ± 1.935	74.77 ± 0.745	71.65 ± 1.215	78.87 ± 0.499	78.30 ± 0.633	78.63 ± 0.472
			800	120	Recall@10	20.14 ± 1.245	30.49 ± 1.151	25.32 ± 0.652	9.36 ± 2.166	14.47 ± 1.789	11.91 ± 1.908	16.69 ± 0.985	19.76 ± 0.427	18.22 ± 0.691	21.58 ± 0.507	22.21 ± 0.247	21.98 ± 0.349
						39.46 ± 0.703	56.54 ± 0.111	48.00 ± 0.399	19.24 ± 0.798	31.27 ± 2.309	25.25 ± 0.816	30.43 ± 1.650	34.14 ± 1.048	32.29 ± 1.337	40.34 ± 0.276	39.06 ± 0.062	39.80 ± 0.151
			800	120	Recall@5	51.08 ± 1.267	69.64 ± 0.582	60.36 ± 0.583	27.66 ± 2.875	41.82 ± 2.499	34.74 ± 1.342	39.22 ± 3.831	42.81 ± 0.624	41.01 ± 2.175	50.95 ± 0.166	48.39 ± 0.282	49.75 ± 0.216
						66.28 ± 2.034	82.91 ± 0.273	74.59 ± 0.881	41.47 ± 2.019	58.77 ± 1.933	50.12 ± 0.743	50.90 ± 4.108	57.28 ± 0.566	54.09 ± 1.801	65.62 ± 0.261	64.25 ± 0.229	64.98 ± 0.237
			800	120	Recall@1	11.71 ± 0.667	17.29 ± 0.544	14.50 ± 0.542	7.49 ± 2.655	9.24 ± 0.611	8.37 ± 1.571	19.68 ± 0.393	12.44 ± 0.249	16.06 ± 0.257	16.63 ± 0.502	18.43 ± 0.740	17.62 ± 0.617
						25.28 ± 1.320	34.44 ± 0.459	29.86 ± 0.645	12.38 ± 2.467	16.60 ± 0.620	14.49 ± 0.984	25.80 ± 0.058	24.79 ± 0.260	25.30 ± 0.107	32.68 ± 0.932	31.86 ± 0.245	32.35 ± 0.455
Encoder	471M	0.50	200	30	Recall@5	34.05 ± 1.295	46.15 ± 0.933	40.10 ± 0.427	18.98 ± 5.512	22.17 ± 0.680	20.58 ± 2.506	30.21 ± 0.388	31.62 ± 0.647	30.91 ± 0.393	41.68 ± 0.563	39.85 ± 0.138	40.83 ± 0.262
						48.62 ± 1.688	67.36 ± 0.197	57.99 ± 0.928	26.59 ± 5.275	32.94 ± 1.212	29.77 ± 2.125	37.42 ± 0.253	43.54 ± 0.297	40.48 ± 0.207	55.59 ± 0.577	52.16 ± 0.368	53.91 ± 0.328
			400	60	Recall@10	7.78 ± 1.646	23.72 ± 1.211	15.75 ± 0.643	4.09 ± 1.187	12.42 ± 3.138	8.26 ± 2.162	7.02 ± 0.117	21.02 ± 0.934	14.02 ± 0.420	10.26 ± 0.119	14.78 ± 1.527	12.52 ± 0.727
Encoder	471M		200	30	Recall@3	20.32 ± 2.075	50.18 ± 1.470	35.25 ± 1.662	12.62 ± 1.451	30.65 ± 2.545	21.64 ± 1.976	17.58 ± 0.212	40.63 ± 0.666	29.10 ± 0.378	24.32 ± 0.181	29.46 ± 1.047	26.89 ± 0.477

Таблица 5. Продолжение

Подход	# Параметров	Относительное время обучения	Датасет			Russian MultiWOZ			Telecom Domain Dataset			Russian FoCus			Russian Taskmaster		
			# Кластеров	Метрика		Пользователь	Диалоговая система	Среднее	Пользователь	Диалоговая система	Среднее	Пользователь	Диалоговая система	Среднее	Пользователь	Диалоговая система	Среднее
				Первый этап	Второй этап												
			400	60	Recall@5	31.45 ± 2.734	64.23 ± 1.989	47.84 ± 2.355	19.95 ± 2.011	42.89 ± 3.061	31.42 ± 2.487	25.96 ± 0.326	53.90 ± 0.460	39.93 ± 0.327	35.82 ± 0.026	40.23 ± 0.662	38.03 ± 0.320
						52.77 ± 1.944	82.25 ± 1.140	67.51 ± 1.521	35.45 ± 3.042	64.48 ± 3.600	49.97 ± 3.313	46.51 ± 2.598	71.92 ± 0.246	59.22 ± 1.359	54.66 ± 0.079	56.50 ± 0.615	55.58 ± 0.276
						4.78 ± 0.351	12.49 ± 1.277	8.64 ± 0.642	1.79 ± 0.203	5.40 ± 0.093	3.60 ± 0.147	5.50 ± 0.304	15.39 ± 0.271	10.44 ± 0.265	6.89 ± 0.374	11.39 ± 0.298	9.14 ± 0.269
						12.00 ± 1.073	33.20 ± 2.642	22.60 ± 1.341	6.10 ± 1.055	14.69 ± 0.462	10.40 ± 0.702	12.95 ± 0.413	29.43 ± 0.194	21.19 ± 0.171	16.14 ± 0.246	21.77 ± 0.380	18.95 ± 0.311
						18.36 ± 1.183	46.91 ± 1.688	32.64 ± 0.925	9.48 ± 1.269	22.87 ± 0.926	16.18 ± 0.778	18.45 ± 0.481	38.86 ± 0.417	28.65 ± 0.035	22.19 ± 0.253	28.52 ± 0.313	25.36 ± 0.267
						30.59 ± 1.627	64.61 ± 1.145	47.60 ± 0.957	17.19 ± 2.014	38.76 ± 0.537	27.97 ± 1.152	34.38 ± 0.563	55.00 ± 0.422	44.69 ± 0.424	35.23 ± 0.131	40.25 ± 0.178	37.74 ± 0.052
						2.37 ± 0.222	6.48 ± 0.611	4.42 ± 0.369	0.96 ± 0.044	2.26 ± 0.438	1.61 ± 0.206	3.93 ± 0.221	10.48 ± 0.402	7.20 ± 0.311	4.40 ± 0.151	7.26 ± 0.174	5.83 ± 0.145
						6.51 ± 0.344	18.25 ± 0.655	12.38 ± 0.498	4.58 ± 2.179	6.30 ± 0.390	5.44 ± 1.079	9.12 ± 0.227	21.78 ± 0.385	15.45 ± 0.285	10.63 ± 0.205	13.76 ± 0.268	12.20 ± 0.215
						10.38 ± 0.443	27.17 ± 0.517	18.78 ± 0.443	6.88 ± 3.234	10.37 ± 0.412	8.63 ± 1.799	12.77 ± 0.244	29.11 ± 0.346	20.94 ± 0.219	14.89 ± 0.116	18.45 ± 0.256	16.67 ± 0.185
						18.39 ± 0.574	44.16 ± 0.861	31.27 ± 0.717	10.33 ± 2.425	20.70 ± 0.602	15.52 ± 1.313	20.03 ± 0.377	39.98 ± 0.155	30.01 ± 0.175	23.99 ± 0.171	26.01 ± 0.552	25.00 ± 0.357
						35.65 ± 0.979	55.68 ± 0.176	45.67 ± 0.468	17.02 ± 0.573	33.27 ± 0.754	25.15 ± 0.658	30.26 ± 0.418	35.21 ± 0.881	32.74 ± 0.608	36.95 ± 0.472	43.18 ± 0.210	40.16 ± 0.187
						60.03 ± 1.724	83.11 ± 0.104	71.57 ± 0.819	34.16 ± 1.030	60.35 ± 1.692	47.25 ± 1.306	52.68 ± 0.322	60.47 ± 0.737	56.57 ± 0.527	59.69 ± 0.837	65.29 ± 0.337	62.54 ± 0.275
Message Passing	471M + 3.7M	0.47	200	30	Recall@1	71.49 ± 1.471	90.47 ± 0.125	80.98 ± 0.776	43.63 ± 1.498	74.26 ± 1.540	58.95 ± 1.216	63.43 ± 0.369	73.05 ± 0.935	68.24 ± 0.290	71.38 ± 0.501	76.25 ± 0.509	73.85 ± 0.012
						86.21 ± 1.799	96.65 ± 0.076	91.43 ± 0.920	64.77 ± 1.005	87.10 ± 0.983	75.94 ± 0.294	79.86 ± 0.313	88.23 ± 0.425	84.04 ± 0.061	86.94 ± 0.195	90.15 ± 0.127	88.56 ± 0.067
						24.46 ± 1.175	40.85 ± 2.178	32.65 ± 0.934	9.03 ± 4.425	20.17 ± 0.438	14.60 ± 2.049	24.12 ± 0.220	31.15 ± 0.220	27.64 ± 0.165	34.35 ± 0.587	42.30 ± 0.286	38.41 ± 0.430

Таблица 5. Продолжение

Подход	# параметров	Относительное время обучения	Датасет		Russian MultiWOZ			Telecom Domain Dataset			Russian FoCus			Russian Taskmaster			
			# Кластеров	Метрика	Пользователь	Диалоговая система	Среднее	Пользователь	Диалоговая система	Среднее	Пользователь	Диалоговая система	Среднее	Пользователь	Диалоговая система	Среднее	
																	Первый этап
Cat-Boost	471M +2.2M	1.00	200	30	Recall@1	37.95 ± 2.477	54.34 ± 1.970	46.14 ± 0.784	17.79 ± 1.731	32.71 ± 1.193	25.25 ± 0.610	22.23 ± 0.388	34.73 ± 0.334	28.48 ± 0.277	37.50 ± 0.210	43.23 ± 0.270	40.45 ± 0.134
					Recall@3	62.60 ± 1.257	81.42 ± 0.765	72.01 ± 0.523	35.52 ± 0.601	59.44 ± 0.438	47.48 ± 0.102	40.95 ± 0.598	58.67 ± 0.533	49.81 ± 0.563	60.46 ± 0.080	66.50 ± 0.359	63.55 ± 0.212
			400	60	Recall@5	74.62 ± 1.363	89.18 ± 0.371	81.90 ± 0.543	46.23 ± 1.858	72.10 ± 0.201	59.16 ± 0.828	50.24 ± 0.791	70.56 ± 0.279	60.40 ± 0.516	72.40 ± 0.123	77.81 ± 0.404	75.15 ± 0.266
					Recall@10	88.42 ± 0.975	95.90 ± 0.307	92.16 ± 0.628	68.88 ± 2.629	86.42 ± 0.981	77.65 ± 0.855	72.44 ± 0.624	86.11 ± 0.444	79.27 ± 0.392	87.69 ± 0.100	91.10 ± 0.354	89.42 ± 0.162
			400	60	Recall@1	26.14 ± 0.999	37.99 ± 0.246	32.07 ± 0.468	11.84 ± 1.828	21.29 ± 1.124	16.56 ± 1.093	23.78 ± 0.158	28.64 ± 0.267	26.21 ± 0.062	33.83 ± 0.257	40.89 ± 0.562	37.45 ± 0.264
					Recall@3	47.63 ± 1.152	66.91 ± 0.776	57.27 ± 0.835	21.12 ± 1.038	38.71 ± 2.798	29.92 ± 1.571	40.15 ± 0.450	48.11 ± 0.399	44.13 ± 0.059	54.03 ± 0.279	60.65 ± 0.349	57.42 ± 0.042
			400	60	Recall@5	58.73 ± 0.363	79.32 ± 0.778	69.03 ± 0.223	27.74 ± 1.154	49.98 ± 3.025	38.86 ± 1.243	50.33 ± 1.363	58.00 ± 0.259	54.16 ± 0.769	63.99 ± 0.111	69.90 ± 0.312	67.02 ± 0.109
					Recall@10	73.72 ± 0.475	90.43 ± 0.468	82.07 ± 0.416	41.59 ± 0.840	68.09 ± 2.155	54.84 ± 0.673	62.62 ± 1.050	73.22 ± 0.182	67.92 ± 0.580	76.68 ± 0.386	82.23 ± 0.683	79.49 ± 0.320

Таблица 5. Продолжение

Подход	# Параметров	Относительное время обучения	Датасет			Russian MultiWOZ			Telecom Domain Dataset			Russian FoCus			Russian Taskmaster		
			# Кластеров	Метрика		Пользователь	Диалоговая система	Среднее	Пользователь	Диалоговая система	Среднее	Пользователь	Диалоговая система	Среднее	Пользователь	Диалоговая система	Среднее
				Первый этап	Второй этап												
Fast-GTN	47M +19M	0.49	200	800	120	Recall@1	20.41 ± 1.727	17.88 ± 1.047	6.99 ± 2.444	12.73 ± 0.889	9.86 ± 1.011	23.53 ± 2.643	20.60 ± 0.544	22.06 ± 1.557	28.09 ± 0.180	36.03 ± 1.224	32.15 ± 0.674
						Recall@3	43.57 ± 0.805	37.42 ± 0.272	15.35 ± 3.264	22.60 ± 1.112	18.98 ± 1.078	36.89 ± 4.462	37.03 ± 0.178	36.96 ± 2.310	46.04 ± 0.321	52.05 ± 0.689	49.12 ± 0.284
						Recall@5	58.40 ± 0.576	50.12 ± 0.463	18.91 ± 4.342	29.75 ± 1.692	24.33 ± 1.338	42.61 ± 4.515	46.13 ± 0.234	44.37 ± 2.340	54.89 ± 0.354	60.27 ± 0.541	57.63 ± 0.237
						Recall@10	77.55 ± 0.682	67.27 ± 0.779	28.10 ± 3.361	43.70 ± 1.121	35.90 ± 1.469	51.43 ± 3.966	59.46 ± 0.703	55.45 ± 2.244	67.79 ± 0.625	71.50 ± 0.288	69.68 ± 0.277
			400	800	60	Recall@1	54.12 ± 2.012	44.70 ± 1.041	18.51 ± 1.491	31.93 ± 2.106	25.22 ± 1.725	29.39 ± 1.626	34.98 ± 0.795	32.18 ± 0.775	37.34 ± 0.298	43.68 ± 0.917	40.60 ± 0.382
						Recall@3	82.97 ± 0.730	71.69 ± 0.287	34.79 ± 1.472	58.37 ± 1.474	46.58 ± 1.250	46.29 ± 1.495	60.01 ± 0.799	53.15 ± 0.772	60.69 ± 0.739	66.39 ± 0.201	63.60 ± 0.379
						Recall@5	90.28 ± 0.589	81.31 ± 0.909	46.35 ± 2.826	71.26 ± 1.109	58.80 ± 1.881	58.24 ± 1.025	72.36 ± 0.438	65.30 ± 0.730	72.25 ± 0.630	77.00 ± 0.286	74.65 ± 0.410
						Recall@10	96.51 ± 0.287	91.93 ± 0.662	64.70 ± 0.535	85.42 ± 0.321	75.06 ± 0.269	76.66 ± 1.518	87.71 ± 0.262	82.18 ± 0.630	87.66 ± 0.590	90.68 ± 0.293	89.18 ± 0.313
			800	120	120	Recall@1	41.15 ± 2.529	34.18 ± 1.638	10.69 ± 0.666	20.14 ± 1.481	15.42 ± 0.466	23.77 ± 0.112	30.53 ± 0.691	27.15 ± 0.310	33.67 ± 0.595	40.37 ± 0.523	37.11 ± 0.146
						Recall@3	71.73 ± 1.006	59.76 ± 0.866	22.36 ± 1.576	36.37 ± 0.697	29.36 ± 0.509	36.91 ± 0.287	50.20 ± 1.159	43.56 ± 0.440	53.68 ± 0.729	60.48 ± 0.263	57.15 ± 0.230
						Recall@5	82.25 ± 0.719	70.79 ± 0.565	29.50 ± 2.981	49.15 ± 0.965	39.33 ± 1.295	44.54 ± 0.342	60.60 ± 0.767	52.57 ± 0.314	63.91 ± 0.511	69.88 ± 0.149	66.95 ± 0.308
						Recall@10	92.00 ± 0.285	82.76 ± 0.266	41.49 ± 2.588	67.78 ± 2.145	54.63 ± 0.789	62.77 ± 2.316	74.81 ± 0.947	68.79 ± 1.486	77.12 ± 0.384	82.60 ± 0.462	79.88 ± 0.418
			800	120	120	Recall@1	23.66 ± 1.150	19.91 ± 1.110	9.51 ± 3.272	12.64 ± 1.088	11.07 ± 1.094	26.16 ± 0.565	22.23 ± 0.489	24.19 ± 0.072	27.85 ± 0.214	32.94 ± 0.749	30.48 ± 0.482
						Recall@3	45.75 ± 1.754	38.57 ± 1.202	16.25 ± 2.474	22.25 ± 1.258	19.25 ± 0.658	37.31 ± 2.867	39.13 ± 0.295	38.22 ± 1.286	46.12 ± 0.231	50.32 ± 0.285	48.29 ± 0.244
						Recall@5	59.16 ± 1.464	50.63 ± 0.918	21.00 ± 4.002	29.75 ± 0.858	25.38 ± 1.638	42.95 ± 2.948	48.46 ± 0.424	45.71 ± 1.281	55.27 ± 0.433	59.14 ± 0.380	57.26 ± 0.397

Таблица 5. Окончание

Подход	Относительное время обучения	Датасет			Russian MultiWOZ			Telecom Domain Dataset			Russian FoCus			Russian Taskmaster		
		# Кластеров	Метрика		Пользователь	Диалоговая система	Среднее	Пользователь	Диалоговая система	Среднее	Пользователь	Диалоговая система	Среднее	Пользователь	Диалоговая система	Среднее
		Первый этап	Второй этап													
					56.86 ± 0.744	78.93 ± 0.894	67.90 ± 0.393	29.11 ± 3.267	43.15 ± 0.650	36.13 ± 1.391	52.64 ± 3.258	61.57 ± 0.745	57.10 ± 1.258	68.43 ± 0.523	71.31 ± 0.254	69.90 ± 0.322
Encoder-MAP	0.78	200	30	Recall@10	18.12 ± 0.343	32.38 ± 5.128	25.25 ± 2.556	6.33 ± 0.942	14.77 ± 1.082	10.55 ± 0.754	18.87 ± 4.014	25.55 ± 0.904	22.21 ± 1.665	27.01 ± 0.919	24.94 ± 0.486	25.98 ± 0.680
					41.45 ± 1.650	66.58 ± 2.381	54.02 ± 0.477	20.35 ± 5.046	35.15 ± 2.808	27.75 ± 3.220	39.13 ± 0.188	47.54 ± 0.804	43.33 ± 0.425	48.86 ± 1.040	48.76 ± 0.676	48.81 ± 0.627
					55.16 ± 2.084	80.99 ± 1.586	68.08 ± 0.414	30.40 ± 3.212	49.64 ± 2.211	40.02 ± 2.210	48.91 ± 0.167	61.38 ± 0.896	55.14 ± 0.364	62.17 ± 0.147	63.66 ± 0.807	62.91 ± 0.345
					77.84 ± 1.394	93.60 ± 0.480	85.72 ± 0.546	51.61 ± 6.799	72.07 ± 2.416	61.84 ± 4.604	68.23 ± 0.278	80.64 ± 0.934	74.43 ± 0.498	78.34 ± 0.420	80.70 ± 0.088	79.52 ± 0.247
		400	60	Recall@10	13.39 ± 0.087	24.56 ± 2.046	18.98 ± 1.066	2.88 ± 0.297	7.12 ± 0.126	5.00 ± 0.100	14.18 ± 0.382	19.76 ± 0.574	16.97 ± 0.099	21.79 ± 0.210	20.18 ± 0.662	20.99 ± 0.231
					31.35 ± 0.772	54.61 ± 1.678	42.98 ± 0.703	7.28 ± 0.777	18.28 ± 0.753	12.78 ± 0.763	26.08 ± 0.528	37.56 ± 0.548	31.82 ± 0.158	41.08 ± 0.774	40.70 ± 0.959	40.89 ± 0.860
					41.98 ± 1.458	68.20 ± 0.421	55.09 ± 0.682	11.99 ± 0.403	27.82 ± 1.632	19.90 ± 0.997	32.63 ± 0.349	48.39 ± 0.494	40.51 ± 0.326	51.57 ± 0.332	51.23 ± 0.130	51.40 ± 0.230
					58.81 ± 1.609	83.93 ± 0.611	71.37 ± 1.034	23.45 ± 1.655	47.00 ± 2.090	35.23 ± 1.872	48.18 ± 2.247	64.49 ± 0.158	56.33 ± 1.202	65.69 ± 0.950	65.75 ± 0.510	65.72 ± 0.275
		800	120	Recall@10	6.47 ± 0.226	12.31 ± 0.488	9.39 ± 0.270	1.42 ± 0.305	3.71 ± 1.000	2.56 ± 0.492	12.48 ± 3.255	10.57 ± 0.568	11.53 ± 1.847	15.85 ± 0.339	14.93 ± 0.727	15.39 ± 0.483
					16.95 ± 0.187	30.25 ± 0.611	23.60 ± 0.290	4.86 ± 2.251	9.91 ± 2.606	7.38 ± 1.462	22.12 ± 3.878	23.33 ± 1.072	22.72 ± 2.006	31.45 ± 0.712	29.01 ± 1.365	30.23 ± 1.030
					25.19 ± 0.395	43.86 ± 0.701	34.53 ± 0.520	7.97 ± 1.952	16.02 ± 3.687	11.99 ± 1.549	26.26 ± 3.758	31.85 ± 1.143	29.06 ± 1.878	40.55 ± 0.982	36.55 ± 1.285	38.55 ± 1.131
					40.91 ± 1.302	65.80 ± 0.578	53.35 ± 0.933	15.84 ± 2.559	27.18 ± 3.961	21.51 ± 1.752	35.98 ± 2.535	46.20 ± 0.758	41.09 ± 1.022	53.92 ± 0.697	50.30 ± 0.685	52.11 ± 0.614

Таблица 6. Результаты экспериментов по метрике $Recall@k$, $k \in \{1, 3, 5, 10\}$ на наборах данных открытого домена. Средняя метрика – среднее значение метрики пользователя и метрики диалоговой системы. Для стабильности результатов все методы были обучены на 3 разных наборах кластеров, полученные результаты были усреднены

Подход	# Параметров	Относительное время обучения	Датасет			Toloka Persona Chat Rus	Russian Dialogues SUBSET	Matreshka		
			# Кластеров		Метрика			Пользователь	Диалоговая система	Среднее
			Первый этап	Второй этап						
Markov Chain	10K	0.13	200	30	Recall@1	19.44 ± 0.429	12.27 ± 1.098	17.57 ± 0.709	30.56 ± 1.918	23.27 ± 1.275
					Recall@3	37.07 ± 0.588	27.60 ± 2.993	37.47 ± 1.970	57.01 ± 2.179	46.25 ± 0.132
					Recall@5	47.91 ± 1.201	38.57 ± 4.256	51.84 ± 0.903	70.46 ± 0.737	60.29 ± 0.568
					Recall@10	68.09 ± 1.300	64.03 ± 5.110	73.87 ± 0.444	85.08 ± 0.664	78.94 ± 0.439
					Recall@1	15.57 ± 0.304	10.06 ± 0.486	15.88 ± 0.408	22.35 ± 0.354	18.75 ± 0.283
			400	60	Recall@3	26.89 ± 0.601	17.41 ± 0.166	30.27 ± 1.164	44.82 ± 0.039	36.77 ± 0.648
					Recall@5	34.74 ± 0.757	23.00 ± 0.241	40.21 ± 1.033	56.22 ± 0.572	47.39 ± 0.435
					Recall@10	49.17 ± 0.895	37.65 ± 0.599	56.85 ± 0.503	70.27 ± 0.902	62.93 ± 0.425
					Recall@1	10.32 ± 0.382	8.56 ± 0.077	12.70 ± 0.325	16.25 ± 0.447	14.33 ± 0.125
					Recall@3	19.08 ± 0.411	13.02 ± 0.235	23.32 ± 0.603	31.50 ± 1.070	27.10 ± 0.570
Encoder	471M	0.50	200	30	Recall@5	24.80 ± 0.736	16.08 ± 0.466	30.23 ± 0.293	39.23 ± 1.128	34.38 ± 0.400
					Recall@10	35.28 ± 0.549	23.18 ± 0.907	41.63 ± 1.703	53.24 ± 1.654	46.86 ± 0.287
					Recall@1	13.00 ± 0.163	10.24 ± 0.783	10.11 ± 1.262	27.33 ± 1.288	18.72 ± 1.217
					Recall@3	28.67 ± 0.870	25.45 ± 1.911	27.47 ± 2.456	48.56 ± 0.997	38.01 ± 1.622
					Recall@5	39.94 ± 1.531	37.00 ± 2.524	34.94 ± 1.977	61.96 ± 1.140	48.45 ± 1.436
			400	60	Recall@10	59.31 ± 1.498	58.89 ± 2.225	53.66 ± 0.703	80.35 ± 0.703	67.01 ± 0.333
					Recall@1	8.80 ± 0.452	6.65 ± 0.297	7.55 ± 0.774	14.35 ± 0.821	10.95 ± 0.357
					Recall@3	19.11 ± 0.426	16.76 ± 0.358	17.17 ± 0.850	33.79 ± 1.173	25.48 ± 0.980
					Recall@5	25.73 ± 0.511	24.33 ± 1.001	23.87 ± 0.660	49.07 ± 0.413	36.47 ± 0.469
					Recall@10	38.72 ± 0.913	39.20 ± 1.338	35.16 ± 0.811	65.74 ± 0.812	50.45 ± 0.484
			800	120	Recall@1	5.86 ± 0.338	5.10 ± 0.072	4.00 ± 0.694	8.09 ± 0.527	6.05 ± 0.611
					Recall@3	12.31 ± 0.362	12.66 ± 0.654	10.37 ± 0.759	22.11 ± 0.489	16.24 ± 0.573
					Recall@5	17.32 ± 0.484	17.65 ± 0.412	17.39 ± 0.413	31.11 ± 0.828	24.25 ± 0.608

Таблица 6. Продолжение

Подход	# Пара- метров	Относи- тельное время обучения	Датасет			Toloka Persona Chat Rus	Russian Dialogues SUBSET	Matreshka		
			# Кластеров	Метрика	Пользо- ватель			Диалоговая система	Среднее	
										Первый этап
						26.73 ± 0.618	27.41 ± 0.307	25.70 ± 0.564	49.73 ± 1.095	37.71 ± 0.756
Message Passing	471M+3.7M	0.47	200	30	Recall@10	21.70 ± 0.212	12.89 ± 0.654	23.90 ± 1.362	44.40 ± 1.745	33.05 ± 0.438
					Recall@3	43.29 ± 0.284	29.22 ± 1.414	45.26 ± 0.801	70.09 ± 1.391	56.50 ± 1.054
					Recall@5	56.07 ± 1.107	41.94 ± 1.993	57.40 ± 2.108	81.19 ± 0.535	68.30 ± 1.312
					Recall@10	75.66 ± 1.429	66.62 ± 2.637	79.26 ± 1.165	92.64 ± 0.164	85.39 ± 0.684
					Recall@1	17.19 ± 1.132	9.34 ± 0.504	18.91 ± 0.484	32.08 ± 0.082	24.80 ± 0.224
			400	60	Recall@3	33.11 ± 0.909	19.31 ± 0.522	36.24 ± 1.609	56.90 ± 1.061	45.52 ± 1.329
					Recall@5	43.39 ± 0.900	26.65 ± 0.412	46.41 ± 0.693	68.22 ± 0.646	56.31 ± 0.549
					Recall@10	58.13 ± 1.618	43.06 ± 0.677	63.92 ± 0.589	82.63 ± 0.326	72.51 ± 0.289
			800	120	Recall@1	12.12 ± 0.728	7.76 ± 0.293	14.47 ± 0.208	23.99 ± 0.759	18.76 ± 0.431
					Recall@3	23.65 ± 0.226	15.06 ± 0.579	25.59 ± 0.886	42.60 ± 1.091	33.26 ± 0.878
					Recall@5	31.18 ± 0.015	20.11 ± 1.187	33.83 ± 0.904	53.47 ± 0.977	42.61 ± 0.544
					Recall@10	43.36 ± 0.285	29.86 ± 1.156	48.53 ± 1.456	67.69 ± 0.414	57.11 ± 0.873
CatBoost	471M+2.2M	1.00	200	30	Recall@1	21.62 ± 0.632	12.59 ± 1.332	23.31 ± 0.312	44.82 ± 0.624	33.04 ± 0.182
					Recall@3	42.44 ± 0.745	28.37 ± 1.838	45.88 ± 2.122	70.73 ± 0.246	57.20 ± 1.152
					Recall@5	54.32 ± 0.579	41.16 ± 1.175	60.10 ± 3.374	81.22 ± 0.689	69.75 ± 1.474
					Recall@10	74.01 ± 0.212	64.55 ± 0.812	79.76 ± 0.396	92.82 ± 0.566	85.66 ± 0.416
					Recall@1	16.83 ± 0.800	10.35 ± 0.132	19.07 ± 0.660	31.25 ± 1.728	24.48 ± 0.969
			400	60	Recall@3	31.96 ± 0.609	19.32 ± 0.274	36.45 ± 1.143	56.25 ± 1.479	45.41 ± 1.277
					Recall@5	41.61 ± 0.442	26.14 ± 0.497	48.60 ± 0.723	67.72 ± 0.664	57.22 ± 0.056
					Recall@10	56.32 ± 0.462	41.00 ± 1.777	63.93 ± 0.805	81.09 ± 0.766	71.70 ± 0.431
			800	120	Recall@1	11.39 ± 0.586	8.42 ± 0.237	15.20 ± 0.524	23.23 ± 0.179	18.89 ± 0.359
					Recall@3	22.63 ± 0.669	14.47 ± 0.215	28.59 ± 1.024	40.87 ± 0.346	34.12 ± 0.732
					Recall@5	29.81 ± 0.369	18.81 ± 0.100	38.04 ± 1.712	50.96 ± 0.622	43.79 ± 1.086

Таблица 6. Окончание

Подход	# Пара- метров	Относи- тельное время обучения	Датасет			Toloka Persona Chat Rus	Russian Dialogues SUBSET	Matreshka		
			# Кластеров		Метрика			Пользо- ватель	Диалоговая система	Среднее
			Первый этап	Второй этап						
						41.77 ± 0.479	27.67 ± 0.929	50.12 ± 1.435	65.97 ± 0.592	57.27 ± 0.573
FastGTN	471M+1.9M	0.49	200	30	Recall@10	21.38 ± 0.527	12.42 ± 0.511	22.59 ± 3.288	44.29 ± 2.176	32.42 ± 0.771
					Recall@3	41.86 ± 0.766	28.09 ± 1.775	47.97 ± 0.896	70.16 ± 0.651	58.05 ± 0.763
					Recall@5	54.37 ± 1.315	43.12 ± 3.102	60.08 ± 2.376	80.70 ± 0.235	69.51 ± 1.168
					Recall@10	74.45 ± 1.767	68.34 ± 3.662	78.82 ± 1.754	91.84 ± 0.247	84.80 ± 0.857
					Recall@1	16.61 ± 0.615	8.74 ± 0.644	18.78 ± 0.207	31.79 ± 2.618	24.51 ± 1.269
			400	60	Recall@3	32.07 ± 0.516	18.38 ± 1.134	35.20 ± 1.196	55.56 ± 0.921	44.39 ± 1.008
					Recall@5	42.05 ± 0.631	25.54 ± 1.369	43.82 ± 1.254	67.83 ± 0.294	54.70 ± 0.801
					Recall@10	57.09 ± 1.007	39.79 ± 1.097	62.56 ± 1.228	82.17 ± 0.273	71.54 ± 0.587
			800	120	Recall@1	11.30 ± 0.361	6.91 ± 0.425	15.69 ± 0.531	23.85 ± 0.434	19.43 ± 0.245
					Recall@3	21.70 ± 0.206	14.34 ± 0.496	27.92 ± 0.618	42.36 ± 0.879	34.48 ± 0.638
					Recall@5	28.92 ± 0.138	19.32 ± 0.659	36.05 ± 0.188	54.02 ± 0.663	44.21 ± 0.260
					Recall@10	41.16 ± 0.436	30.00 ± 1.172	49.77 ± 1.513	68.09 ± 0.321	58.10 ± 0.894
Encoder- MAP	471M+2M	0.78	200	30	Recall@1	13.03 ± 1.743	10.21 ± 0.352	13.05 ± 2.103	32.63 ± 4.105	22.84 ± 2.883
					Recall@3	31.23 ± 1.295	26.41 ± 1.993	34.43 ± 2.887	61.25 ± 1.390	47.84 ± 1.725
					Recall@5	44.73 ± 1.014	40.08 ± 2.861	48.29 ± 0.798	75.42 ± 1.200	61.86 ± 0.965
					Recall@10	66.69 ± 0.445	66.31 ± 4.390	66.63 ± 0.415	90.21 ± 0.045	78.42 ± 0.201
					Recall@1	9.61 ± 1.603	5.81 ± 0.665	7.98 ± 0.927	20.50 ± 1.646	14.24 ± 0.602
			400	60	Recall@3	21.15 ± 1.861	15.20 ± 1.469	19.82 ± 1.387	45.61 ± 0.960	32.71 ± 0.553
					Recall@5	31.17 ± 1.000	23.25 ± 1.521	27.53 ± 1.535	59.67 ± 1.243	43.60 ± 1.310
					Recall@10	47.34 ± 1.244	39.49 ± 2.233	42.49 ± 2.034	77.50 ± 1.014	60.00 ± 1.403
			800	120	Recall@1	5.36 ± 1.737	3.62 ± 0.115	5.12 ± 0.050	14.04 ± 1.576	9.58 ± 0.804
					Recall@3	13.17 ± 1.339	9.53 ± 0.283	12.79 ± 0.654	29.81 ± 1.297	21.30 ± 0.914
					Recall@5	18.40 ± 1.479	14.36 ± 0.549	18.15 ± 0.631	40.35 ± 2.168	29.25 ± 0.901
					Recall@10	29.22 ± 0.894	24.46 ± 1.296	29.89 ± 1.732	57.68 ± 2.416	43.78 ± 1.728

допущение оправдано для текущих целей исследования, оно ограничивает возможность обобщения подходов на мультязычные диалоговые системы. Для того, чтобы более полно оценить эффективность наших графовых подходов в различных языковых средах, необходимо провести исследования на наборах диалоговых данных на других языках. Другим важным ограничением является ограниченный набор доменов, представленный в наборах диалоговых данных, использовавшихся в наших экспериментах. В будущих исследованиях стоит рассмотреть эффективность наших подходов на более широком спектре доменов, чтобы лучше понять их универсальность и применимость.

Ограниченное количество ролей в диалогах. Используемые в исследовании наборы диалоговых данных охватывают небольшое число ролей участников диалога. Подобное ограничение может не полностью отражать многообразие реальных диалогов, к примеру, диалогов в социальных сетях и мессенджерах. Проведение экспериментов на более крупных и разнообразных наборах данных поможет достичь более всесторонней оценки эффективности представленных подходов.

Выбор текстового кодировщика. Выбор кодировщика предложений имеет ключевое значение, особенно при анализе специализированных наборов данных. Результаты экспериментов подчеркнули важность выбора конкретного текстового кодировщика. Последующие исследования должны сосредоточиться на адаптации текстовых кодировщиков к конкретным областям для дальнейшего совершенствования подходов.

Число кластеров. В данной работе использованы фиксированные количества вершин многодольного диалогового графа. Однако определение оптимального числа кластеров является сложной задачей, зависящей от конкретного набора диалоговых данных. Уменьшение числа кластеров может улучшить метрики за счет меньшего количества кандидатов для вершины с намерением следующей реплики, но также может повлиять на качество вершин с точки зрения их соответствия намерениям участников диалога.

КОНФЛИКТ ИНТЕРЕСОВ

Авторы заявляют об отсутствии конфликта интересов.

БЛАГОДАРНОСТИ

Мы хотели бы выразить благодарность Лаборатории Нейронных Систем и Глубокого обучения Деер-Павлов за их неоценимую помощь и поддержку в ходе исследований и разработок.

ИСТОЧНИКИ ФИНАНСИРОВАНИЯ

Публикация подготовлена за счет средств гранта на поддержку исследовательских центров в сфере искусственного интеллекта, в том числе в области “сильного” искусственного интеллекта, систем доверенного искусственного интеллекта и этических аспектов применения искусственного интеллекта, предоставленного АНО “Аналитический центр при Правительстве Российской Федерации” в соответствии с соглашением о предоставлении субсидии (идентификатор соглашения о предоставлении субсидии 000000D730321P5Q0002) и договором с МФТИ от 1 ноября 2021 г. № 70-2021-00138.

СПИСОК ЛИТЕРАТУРЫ

1. Grigory Minakov, Mumtozbek Akhmadjonov, Denis Kuznetsov. 2023. NEUROINFORMATICS 2023. Dialogue Graphs: Enhancing Response Selection Through Target Node Separation, pages 39–53.
2. Niklas Muennighoff, Nouamane Tazi, Loic Magne, Nils Reimers. MTEB: Massive text embedding benchmark. // Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. Dubrovnik, Croatia, May. Association for Computational Linguistics. 2023. P. 2014–2037.
3. Steinley D. K-means clustering: a half-century synthesis. Br J Math Stat Psychol. 2006. V. 59. Pt 1. P. 1–34. <https://doi.org/10.1348/000711005X48266>
4. Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, Maosong Sun. Graph neural networks: A review of methods and applications, AI Open. 2020. V. 1. P. 57–81, ISSN 2666-6510. <https://doi.org/10.1016/j.aiopen.2021.01.001>
5. Veličković P., et al. Graph Attention Networks. 6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings. OpenReview.net, 2018. <https://doi.org/10.17863/CAM.48429>.
6. Yun Seongjun, Minbyul Jeong, Sungdong Yoo, Seung-hun Lee, Sean S. Yi, Raehyun Kim, Jaewoo Kang, Hyunwoo J. Kim. “Graph Transformer Networks: Learning Meta-path Graphs to Improve GNNs.” Neural networks : the official journal of the International Neural Network Society. 2021. V. 153. P. 104–119.
7. Nagovitsin M., Kuznetsov D. DGAC: Dialogue Graph Auto Construction Based on Data with a Regular Structure. In: Kryzhanovsky B., Dunin-Barkowski W., Redko V., Tiumentsev Y. (eds) Advances in Neural Computation, Machine Learning, and Cognitive Research VI. NEUROINFORMATICS. Studies in Computational Intelligence, vol 1064. Springer, Cham, 2022. https://doi.org/10.1007/978-3-031-19032-2_52
8. Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Ariavazhagan, Wei Wang. Language-agnostic BERT sentence embedding. // Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022. P. 878–891, Dublin, Ireland, May. Association for Computational Linguistics.

GRAPH MODELS FOR CONTEXTUAL INTENTION PREDICTION IN DIALOG SYSTEMS

D. Kuznetsov^a and D. Ledneva^a

^a*Moscow Institute of Physics and Technology (National Research University), Dolgoprudny, Moscow Region, Russian Federation*

Presented by Academician of the RAS A.I. Avetisyan

The paper introduces a novel methodology for predicting intentions in dialogue systems through a graph-based approach. This methodology involves constructing graph structures that represent dialogues, thus capturing contextual information effectively. By analyzing results from various open and closed domain datasets, the authors demonstrate the substantial enhancement in intention prediction accuracy achieved by combining graph models with text encoders. The primary focus of the study revolves around assessing the impact of diverse graph architectures and encoders on the performance of the proposed technique. Through empirical evaluation, the experimental outcomes affirm the superiority of graph neural networks in terms of both *Recall@k* (*MAR*) metric and computational resources when compared to alternative methods. This research uncovers a novel avenue for intention prediction in dialogue systems by leveraging graph-based representations.

Keywords: intent prediction, dialogue systems, graph neural networks