УДК 519.7 DOI 10.17223/20710410/X/1

ПОСТРОЕНИЕ И ВИЗУАЛИЗАЦИЯ ОБОБЩЁННОГО ГРАФА ДИАЛОГА ПО НАБОРУ ДИАЛОГОВ

П. Д. Штыков, А. Г. Дьяконов

МГУ имени М. В. Ломоносова, г. Москва, Россия

E-mail: shtykov.pa@gmail.com, djakonov@mail.ru

Предлагается определение обобщённого графа диалога, с помощью которого описывается структура диалога по корпусу однородных диалогов. Задача построения такого графа является актуальной в современном разговорном искусственном интеллекте, однако работ с конкретными результатами мало, часто не даётся полного описания алгоритмов, не выкладывается код с их реализацией. В настоящей работе предложен метод построения обобщённого графа диалога, который был реализован на языке программирования Python и выложен в открытый доступ. Были проведены эксперименты на открытых данных и описаны их результаты.

Ключевые слова: *диалоговая система, обработка естественного языка, граф, граф диалога, кластеризация, представления*

A GENERALIZED DIALOGUE GRAPH CONSTRUCTION AND VISUALIZATION BASED ON A CORPUS OF DIALOGUES

P. D. Shtykov, A. G. Dyakonov

MSU "M. V. Lomonosov", Moscow, Russia

A definition of a generalized dialogue graph is proposed, with the help of which the structure of a dialogue is described according to the corpus of homogeneous dialogues. The task of constructing such a graph is relevant in modern conversational artificial intelligence, however, there are few works with meaningful results, often a complete description of the algorithms is not given and the code with the implementation is not published. In this paper, a method for constructing a generalized dialogue graph is proposed, which was implemented in the Python programming language and made publicly available. Experiments were carried out on open data and the results were described.

Keywords: *dialogue system, NLP, graph, dialogue graph, clustering, representations*

Введение

Обработка естественного языка (Natural Language Processing, NLP) является одним из ключевых направлений в области искусственного интеллекта. При этом одной из важнейших задач в NLP является обработка и понимание диалогов. В данной работе исследуется общий подход к представлению и анализу диалогов, а также представлению общей структуры диалога, идея которого не нова, но публикаций по его реализации крайне мало.

Предположим, что у нас есть достаточное число диалогов из некоторой узкой доменной области. Например, это диалоги работников колл-центра банка с клиентами

(здесь диалоги ещё и проблемно-ориентированные — task-oriented dialogs). Естественно предположить, что их можно разбить на группы одноцелевых диалогов, например «предложение новой услуги», «перевыпуск банковской карты» и т.п. Также логично, что каждой группе соответствует граф-диалога. Собственно, у каждого работника колл-центра есть чёткая инструкция по общению с клиентом и ей соответствует, например, такой граф:

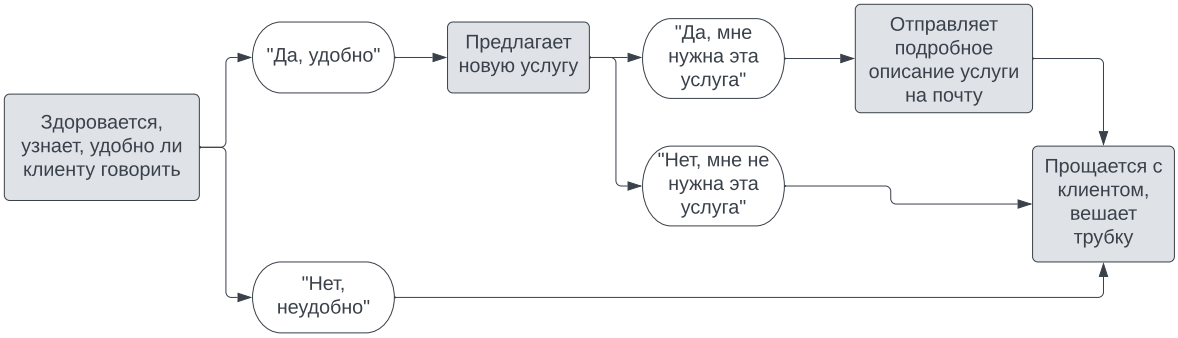


Рис. 1. Пример графа диалога работника колл-центра с клиентом

По одному диалогу в общем случае невозможно восстановить граф диалога, поскольку не реализуются все варианты прохода по этому графу. По нескольким диалогам, которые соответствуют одному графу, это уже возможно. Теоретически (хороших практических реализаций этой идеи нет) можно восстановить набор графов по корпусу диалогов из некоторой узкой доменной области. Отметим специфику этой задачи:

— даже если графов несколько, у них есть схожие вершины (например, вершина «поздороваться оператору»),
— не всегда диалог может идти по задуманному графу, возможны отклонения (например, пользователь просит повторить, отказывается от услуги и просит помощи в чём-то и т.п.),
— не всегда текущий ответ пользователя однозначно определяет переход на новую вершину графа (например, пользователь может попросить отключить услугу и закрыть счёт, оператор сам решает, с чего начать, и это определяет следующую вершину).

Отметим, что граф диалога автоматически построенный по корпусу однородных диалогов позволяет представить информацию в сжатой форме, подходящей как для визуализации, так и для встраивания в сложные диалоговые системы. В данной работе предложена формализация понятия обобщённого графа диалога и базовые способы его построения и визуализации (рассмотрим построение одного графа, а не набора).

1. Существующие подходы

Базовая идея построения графа диалога состоит в поиске некоторой общей структуры диалогов в однородной выборке. Такую структуру чаще всего представляют в виде ориентированного графа, вершины которого отражают темы реплик, а дуги — переходы между ними (см. рис. 2). Приведём работы по изучению таких графов диалогов.

В [1], [2] предпринимались попытки обнаружения «структуры диалога», но без явного построения графа. Самой цитируемой из работ про графы диалогов является серия статей [3], [4], в которой предлагаются два способа построения такого графа.

Первый основан на использовании рекуррентного варианта вариационных автокодировщиков [5] (Variational Recurrent Neural Network, VRNN), результат работы этого алгоритма представлен на рис. 2. Во второй статье авторы добавили в данную архитектуру механизм внимания [6], что позволило им улучшить качество (Structured-Attention Variational Recurrent Neural Network, SVRNN).

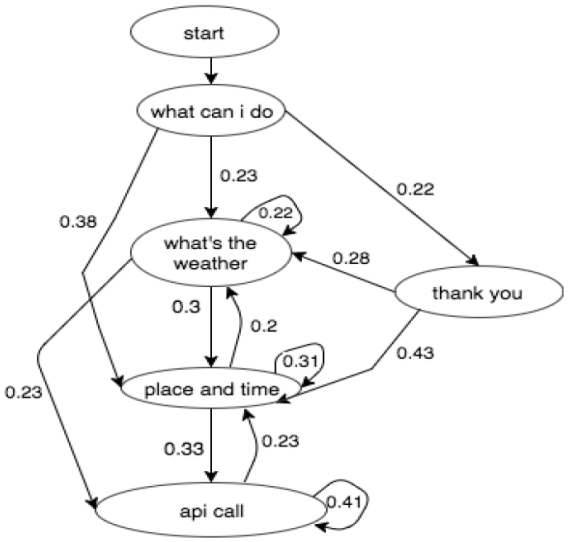
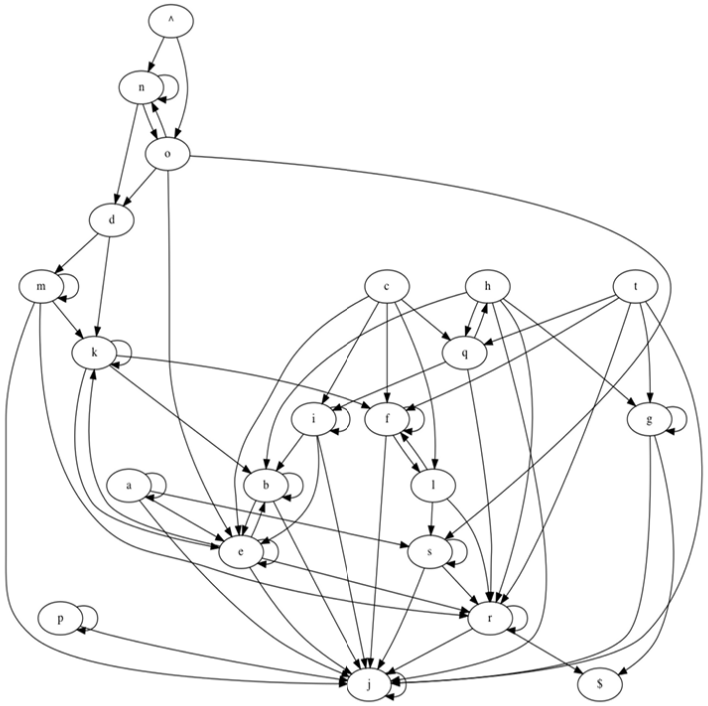


Рис. 2. Пример обобщенного графа диалога взятого из статьи [3]

В последние годы появляются статьи на более амбициозные темы, например в [7] строится двухуровневый граф диалога для «открытого домена» (open domain dialogs). Обычно структура диалога выявляется для проблемно-ориентированных диалогов (task-oriented dialogs), здесь же рассматривались более лексически разнообразные диалоги. В [7] используется довольно сложная техника: DVAE-GNN (Discrete Variational Auto-Encoder with Graph Neural Network).

На русском языке можно отметить работы [8], [9]. Первая выгодно отличается от многих наличием реализации в открытом доступе, в ней признаки, извлеченные из диалогового графа используются при генерации реплик диалоговой системой.

В данной работе мы будем больше ориентироваться на метод TSCAN (text SCAN) из статьи [10], в которой для построения графа диалога применяется алгоритм классификации изображений с самообучением — SCAN (Semantic Clustering using Nearest Neighbors) [11]. Авторы адаптировали данный алгоритм для работы с текстами, использовав в качестве представления (эмбеддинга) для предварительной задачи (pretext task) стандартный трансформер BERT [12], один из примеров графа, полученного алгоритмом, приведён на рис. 3. В данной работе будет исследована применимость представления (эмбеддинга), более подходящего для семантической кластеризации — SBERT (SentenceBERT, [13]). Однако в [10] авторы используют закрытый набор данных для сравнения алгоритма с более простыми методами, в частности с алгоритмом кластеризации k-средних (k-means) [14]. Кроме этого, авторы не предоставляют ни подробного алгоритма, ни его программного кода. В данной работе будет приведен пример такого алгоритма и проведены эксперименты с разными методами кластеризации. Также в работе дополнено определение обобщенного графа диалога для его более простого дальнейшего анализа и визуализации.



cluster-name	manual-label
^	start-node
n	self-introduction
o	rpc-inquiry
d	brand-intro
m	product-info
k	payment-inquiry
f	amount-info
l	date-reminder
r	number-intimation
g	payment-date-inquiry
$	end-node

Рис. 3. Пример обобщенного графа диалога взятого из статьи [10]

2. Постановка задачи

Definition 1. Назовем обобщённым графом диалога пару $T = (G, p(u|v))$, где:

— $G = (V, E)$ — направленный взвешенный граф, каждой дуге которого сопоставлена вероятность перехода по ней: $e_{i,j} \sim p(v_i|v_j)$, при этом сумма вероятностей дуг выходящих из каждой вершины равна 1: $\sum_i p(v_i|v_j) = 1$,

— $u \in U$ — единичное высказывание, а U — множество всех высказываний во всех диалогах,

— $p(u|v)$ — плотность вероятности (либо функция вероятности в случае дискретного пространства U) отнесения высказывания u к текущей вершине v.

Данное определение не ограничивает нас в выборе модели для построения обобщённого графа диалога. Дополнительное требование наличия функции $p(u|v)$ позволит нам вычислять статистики полезные для визуализации и дальнейшего использования графа (например, самое вероятное предложение или самые частотные слова среди предложений ассоциированных с текущей вершиной).

Также такой граф достаточно просто обобщается на случай персонализированных диалогов (например, диалогов вида *«пользователь»* - *«система»*) — введением раскраски вершин, т.е. дополнительной функции $\phi(v)$, ставящей в соответствие каждой вершине некоторый персональный идентификатор пользователя (ID). Однако для корректности необходимо ввести дополнительные ограничения: смежные вершины не должны быть одинаково окрашены (так как высказывания пользователей чередуются), в графе нет петлей (заметим, что основное определение их не запрещает). В данной работе мы будем строить простой граф диалога, без дополнительной раскраски.

Пусть $D = \{d_1, d_2, \ldots, d_{|D|}\}$ — выборка диалогов, каждый диалог d_i является упорядоченным набором из нескольких высказываний: $d_i = \{d_i^1, d_i^2, \ldots, d_i^n\}$, $d_i^j \in U$. В данной работе мы будем работать с неразмеченными корпусами диалогов. В общем

же случае нет ограничения на использование разметки. Добавим к каждому диалогу d_i технические высказывания: BEGIN (начала) и END (конца), аналогичные вершины добавим и в граф диалога. Вероятность $p(u|v)$ для этих вершин будет вырождена в соответствующих точках в пространстве высказываний U. Это необходимо для более ясной конструкции графа и соблюдения ограничения на сумму вероятностей ребер, исходящих из вершины: $\sum_i p(v_i|v_j) = 1$. Рассмотрим задачу построения обобщённого графа диалога по набору D.

3. Предложенный метод

Для решения предварительной задачи (pretext task) воспользуемся представлением (эмбеддингом):

$$\text{Embedding} : U \to M = \mathbb{R}^n, \quad n \in \mathbb{N}.$$

Для реализации представления мы использовали предобученную сиамскую нейронную сеть [13] с разными базовыми сетями (подробнее в разделе 4). В дальнейшем, если не оговорено другого, под высказыванием u мы будем подразумевать его представление $\text{Embedding}(u)$. При этом $n = 768$ или $n = 384$ в зависимости от использованной базовой сети в SBERT. В пространстве M введена косинусная мера сходства, отражающая семантическую близость высказываний, это позволяет использовать в пространстве M простые методы кластеризации для объединения близких по смыслу высказываний.

Теперь опишем предлагаемый алгоритм построения обобщенного графа диалога $T = (G, p(u|v))$, для высказываний в пространстве представлений M. Пусть выбран некоторый алгоритм кластеризации a:

$$a : M \to V.$$

Будем считать, что множество полученных кластеров (или, для удобства, их номеров) и есть множество вершин V нашего графа. Соответственно, может быть вычислена *дискретная* вероятность $p(v|u)$ принадлежности каждого высказывания u к каждой вершине v . При этом кластеризация может быть как жесткой, например методом k-средних (k-means), так и мягкой, например смесью гауссиан (Gaussian mixture model, GMM) [15]. Зная $p(v|u)$, можно вычислить $p(u|v)$, используя теорему Байеса:

$$p(u|v) = \frac{p(v|u)p(u)}{\sum_{i=1}^{|U|} p(v|u_i)p(u_i)},$$

где $p(u)$ — вероятность встречаемости высказывания u во всем корпусе диалогов. На практике мы будем пользоваться оценками вероятностей: частотами. Заметим, что вероятность $p(u)$ не одинакова для всех высказываний, так как в корпусе могут встречаться диалоги с повторяющимися высказываниями.

Нам осталось определить в графе G дуги и найти вероятности, ассоциированные с ними. Введём на множестве высказываний U граф $\hat{G}$, подобный графу основному G, т.е. ориентированный взвешенный граф с вероятностями, ассоциированными с дугами:

$$\hat{G} = (\hat{V}, \hat{E}), \quad \hat{V} \subset M, \quad \hat{E} \subset \hat{V} \times \hat{V}, \quad \hat{e}_{i,j} \sim p(u_j|u_i).$$

Данный граф строится напрямую по выборке диалогов, и дуги в нём имеют смысл апостериорной вероятности встретить ответ u_j на высказывание u_i. Схему совместного размещения обоих графов G и $\hat{G}$ можно увидеть на рис. 4. Соответственно, матрица смежности $\hat{A}$ графа $\hat{G}$ определяется как:

$$\hat{A} = (\hat{a}_{ij}), \quad \hat{a}_{ij} = p(u_j|u_i)$$

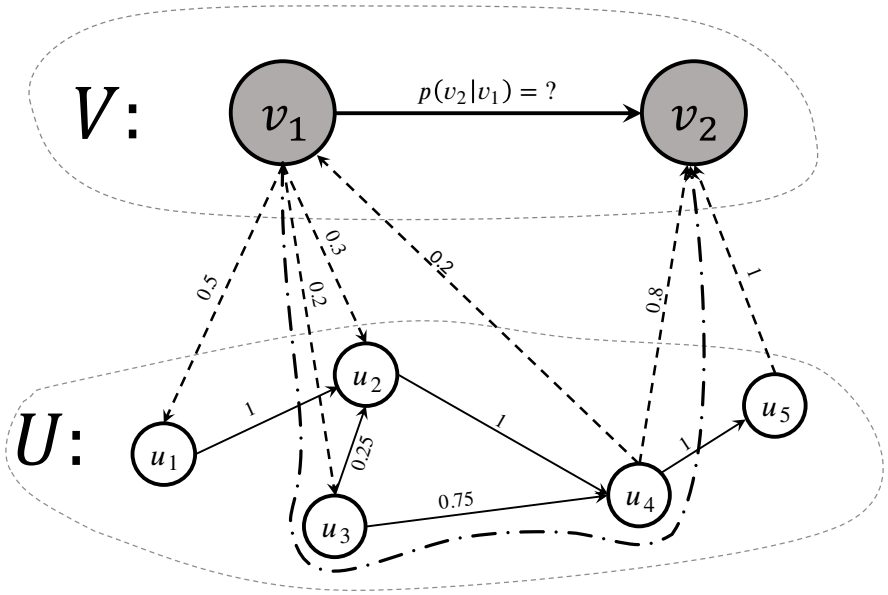


Рис. 4. Пример двух графов: G (сверху) в пространстве вершин V и $\hat{G}$ (снизу) в пространстве высказываний U.

для $1 \leqslant i, j \leqslant |U|$. Теперь, зная вероятности $p(u|v)$, $p(u|u)$ и $p(v|u)$, становится просто вычислить вероятности дуг $p(v|v)$ в графе G:

$$p(v_j|v_i) = \sum_{\alpha,\beta} p(v_j|u_\beta)p(u_\beta|u_\alpha)p(u_\alpha|v_i),$$

то есть вероятность перехода из вершины v_i в вершину v_j графа G равна сумме вероятностей всех простых путей из v_i в v_j, проходящих через одиночные пары высказываний в графе $\hat{G}$ вида (u_α, u_β). Пример такого пути выделен на рис. 4 штрихпунктирной линией.

Вероятностная матрица смежности взвешенного графа G имеет вид

$$A = (a_{ij}), \quad a_{ij} = p(v_j|v_i)$$

для $1 \leqslant i, j \leqslant |U|$. Так как в нашем случае совместные распределения $p(u|v)$ и $p(v|u)$ дискретны, то они могут быть представлены в виде матриц, поэтому способ вычисления матрицы смежности A может быть представлен в явной матричной форме:

$$A = p(u|v) \cdot \hat{A} \cdot p(v|u).$$

Мы закончили описание построения обобщенного графа диалога T. Заметим, что описанный метод построения применим не только в случае использования кластеризации в пространстве представлений (эмбеддингов), но и в случае использования любого другого алгоритма способного оценить апостериорные вероятности $p(v|u)$ (например с помощью латентного размещения Дирихле (Latent Dirichlet allocation, LDA) или нейронной сети, решающей задачу от начала до конца без промежуточного использования представлений). Кластеризация была выбрана, как наиболее простой метод. Исходный код алгоритма и экспериментов доступен по ссылке [16].

4. Эксперименты

4.1. Данные для экспериментов

Для многих задач в анализе данных и машинном обучении есть стандартные наборы данных (датасеты), на которых отслеживается качество предложенных решений и определяется лучшее текущее решение — SotA (state of the art), например на ресурсе paperswithcode.com. Постановка задачи, рассматриваемая в данной работе, достаточно нова: стандартных наборов данных почти нет, т.е. наборов диалогов, имеющих некоторую общую *известную* структуру. Описанным критериям удовлетворяет лишь корпус STAR (Schema-Guided Dialog Dataset for Transfer Learning) [17], на котором не тестировались упомянутые выше методы. Поэтому для проведения экспериментов нами было выбрано два известных датасета, для которых можно предположить наличие общей структуры. Первый датасет — «Customer Support on Twitter» [18], в котором собраны ответы официальных аккаунтов технической поддержки крупных компаний. Для обеспечения однородности из него выбрано подмножество сообщений аккаунтов шести разных авиакомпаний США. Второй датасет — «DailyDialog» [19], в котором собраны обычные диалоги из повседневной жизни на разные темы. Для экспериментов были взяты диалоги на тему работы, как наиболее однородные. В результате для экспериментов подготовлены два набора данных: 8081 диалог в среднем по 3.6 высказывания в диалоге и 1924 диалога в среднем по 7.5 высказывания. Примеры диалогов из двух наборов приведены на рис. 5.

Twitter Customer Support

— @AlaskaAir it says you open at 5:15 @317258 where is everyone? #hellooooo https://t.co/WePfUANLsZ

— @429415 @317258 Ticket counter opens at 615 is what I see on our website.

— @AlaskaAir @317258 all good! They just showed up thanks Andre

— @429415 That is good news

DailyDialog

— Everything's gone wrong.

— I know, it's not as I had planned.

— What are we going to do now?

— I'll speak to Bob, he'll be able to help us

Рис. 5. Примеры диалогов из датасетов Twitter Customer Support и DailyDialog

Видно, что датасет «Twitter Customer Support» очень зашумлен, так как переписка в соцсети Twitter публична и в неё часто вмешиваются третьи лица. Поэтому в корпусе были оставлены только те диалоги между компаниями и пользователями, которые соответствуют схеме:

«система» – «пользователь N» – «система» – «пользователь N» и т.д.

Из высказываний были убраны все идентификаторы пользователей, а идентификаторы компаний были заменены на единый токен *«companyname»*. После чего была применена следующая предобработка текста:

— приведение всего текста к нижнему регистру,
— удаление слов с цифрами,
— удаление ссылок,
— лемматизация с помощью пакета NLTK [20],
— удаление стоп-слов (использовался набор стоп-слов из пакета NLTK).

Датасет «DailyDialog» зашумлён существенно меньше, поэтому к нему была применена только предобработка описанная выше.

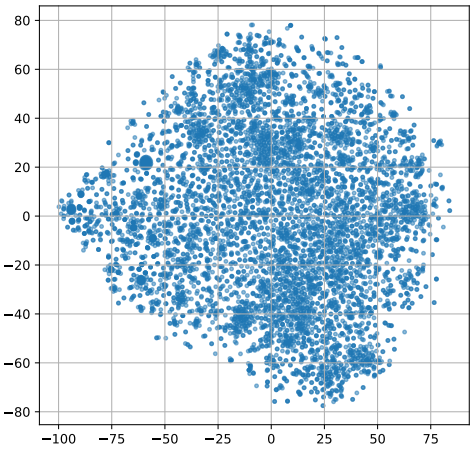


Рис. 6. Пространство эмбеддингов для датасета DailyDialog

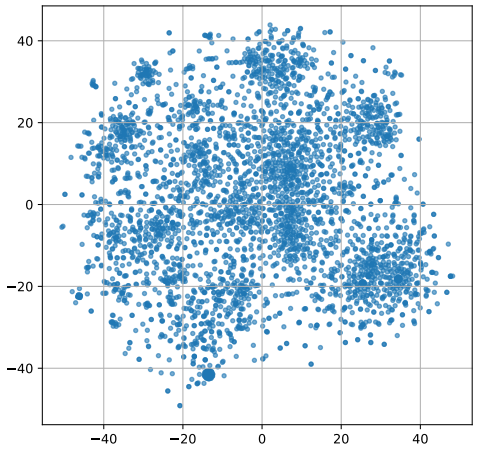


Рис. 7. Пространство эмбеддингов для датасета Twitter Customer Support

4.2. Оценка качества кластеризации

Для начала рассмотрим визуализацию пространства представлений (эмбеддингов), полученных с помощью SBERT с базовой нейронной сетью Distill-RoBERTa [21]. Для визуализации пространство отображено на плоскость помощью t-SNE (t-distributed stochastic neighbor embedding) [22] с перплексией равной 50, см. рис. 6 и 7. Для диалогов из обоих датасетов заметна кластерная структура, однако кластеры имеют небольшие размеры и между ними много шума. Это может привести к зашумлению и самого графа диалога.

Измерим качество кластеризации. В таблице 1 переставлены результаты работы алгоритма с 5-ю вершинами в графе (т.е. при 5 кластерах) в зависимости от следующих параметров:

— Базовая модель в SBERT: MPNet [23], DistillRoBERTa [21], [24] и MiniLM [25],
— Кластеризатор: k-средних (k-means), смесь гауссиан (GMM) и SCAN [10].

Так как авторы статьи [10] не указали оптимальные гиперпараметры для SCAN, он обучался для каждой конфигурации с нуля со следующими стандартными гиперпараметрами:

— количество голов классификатора: 1,
— темп обучения: $1e^{-5}$,
— количество эпох: 15.

Для графов с 10-ю и 15-ю вершинами аналогичная статистика приведена в таблицах 2, 3.

	Model	Twitter Customer Support				DailyDialog			
		Silh.	C.-H.	D.-B.	Entr.	Silh.	C.-H.	D.-B.	Entr.
$\|V\| = 5$	MPNet_KMeans	**0.052**	1043.9	3.912	0.535	0.011	**292.3**	5.166	0.712
	RoBERTa_KMeans	0.043	1003.3	4.43	0.606	**0.024**	281.8	5.176	0.719
	MiniLM_KMeans	0.045	**1054.0**	3.869	0.523	0.023	286.5	5.287	0.724
	MPNet_GMM	0.036	940.4	4.152	0.524	-0.001	253.3	5.517	0.694
	RoBERTa_GMM	0.034	988.7	4.544	0.617	0.022	278.1	5.156	0.71
	MiniLM_GMM	0.044	1046.8	**3.856**	**0.519**	0.01	273.4	**4.74**	**0.663**
	MPNet_SCAN	0.042	1117.515	4.169	0.576	0.017	230.752	5.992	0.717
	RoBERTa_SCAN	0.037	947.797	4.569	0.625	0.024	258.829	5.563	0.718
	MiniLM_SCAN	0.036	901.678	4.493	0.624	0.021	247.22	5.596	0.705

Таблица 1

Результаты сравнения качества кластеризации для графа с 5-ю вершинами для исследованных наборов данных

	Model	Twitter Customer Support				DailyDialog			
		Silh.	C.-H.	D.-B.	Entr.	Silh.	C.-H.	D.-B.	Entr.
$\|V\| = 10$	MPNet_KMeans	0.04	672.1	4.018	0.659	0.016	**210.4**	4.422	0.778
	RoBERTa_KMeans	0.041	634.1	4.147	0.675	**0.021**	195.1	4.394	0.758
	MiniLM_KMeans	**0.054**	**693.4**	3.827	0.668	0.015	198.3	4.482	0.759
	MPNet_GMM	0.037	652.9	**3.669**	**0.597**	−0.002	195.5	4.899	0.776
	RoBERTa_GMM	0.036	617.6	3.992	0.611	0.018	185.1	4.724	0.767
	MiniLM_GMM	0.026	668.6	3.807	0.623	0.009	182.8	**4.373**	**0.73**
	MPNet_SCAN	0.032	626.664	4.569	0.663	0.014	176.676	5.117	0.809
	RoBERTa_SCAN	0.031	562.999	4.465	0.68	0.021	165.634	5.414	0.809
	MiniLM_SCAN	0.03	585.485	4.278	0.679	0.021	175.363	5.003	0.81

Таблица 2

Результаты сравнения качества кластеризации для графа с 10-ю вершинами для двух датасетов: Twitter Customer Support и DailyDialog

В качестве показателей качества кластеризации использовались следующие базовые для неразмеченных данных: коэффициент силуэта (Silh.) [26], индекс Калински-Харабаса (C.-H.) [27] и индекс Дэвиса-Болдина (D.-B.) [28]. Также введём дополнительный показатель для оценки структуры графа. Нам хотелось бы, чтобы граф был «более детерминированным»: если случайно блуждать по графу согласно приписанным дугам вероятностям, то в идеале переходы должны быть детерминированны (т.е. только одна исходящая из вершины дуга имеет вероятность 1, а остальные — 0). Для этого будем измерять среднюю нормализованную энтропию (Entr.):

$$H(G) = \frac{1}{|V|} \sum_{i=1}^{|V|} \frac{\sum_{j=1}^{|V|} p(v_j|v_i) \ln p(v_j|v_i)}{\log |V|}.$$

Соответственно, чем меньше энтропия, тем более детерминирован граф.

Из таблицы 1 видно, что мы не смогли в точности повторить результаты авторов статьи TSCAN [10] их же методом: наша реализация SCAN-кластеризатора уступает стандартным алгоритмам k-средних и смеси гауссиан по всем показателям. Возмож-

	Model	Twitter Customer Support				DailyDialog			
		Silh.	C.-H.	D.-B.	Entr.	Silh.	C.-H.	D.-B.	Entr.
$\|V\| = 15$	MPNet_KMeans	0.038	530.2	3.704	0.659	0.018	**167.7**	4.329	0.796
	RoBERTa_KMeans	0.04	490.5	3.936	0.651	**0.023**	155.2	4.27	0.785
	MiniLM_KMeans	**0.049**	**533.7**	3.781	0.668	0.018	160.8	4.234	0.788
	MPNet_GMM	0.015	508.8	3.711	0.626	0.003	158.4	4.404	0.781
	RoBERTa_GMM	0.014	466.8	3.79	**0.623**	0.016	150.7	**4.14**	**0.772**
	MiniLM_GMM	0.03	509.1	**3.663**	0.637	0.014	155.6	4.341	0.775
	MPNet_SCAN	0.024	475.495	4.457	0.693	0.005	136.29	4.779	0.832
	RoBERTa_SCAN	0.024	447.426	4.418	0.694	0.023	135.245	4.937	0.829
	MiniLM_SCAN	0.024	439.935	4.43	0.696	0.022	140.048	5.132	0.826

Таблица 3

Результаты сравнения качества кластеризации для графа с 15-ю вершинами для двух датасетов: Twitter Customer Support и DailyDialog

но, это связано с неправильно подобранными гиперпараметрами (которые авторы не указали в статье).

4.3. Визуализация графов

Построим и визуализируем графы, полученные предложенным методом. Для всех графов использовалась лучшая (по результатам из табл. 1) модель для данного количества вершин и для данного набора данных. В качестве маркировки вершин будем использовать 4-е слова из высказываний, которые соответствуют вершине, с самым большим значением Tf-Idf (TF — term frequency, IDF — inverse document frequency) [29]. Tf-Idf-представления строились для двухсот наиболее вероятных для данной вершины высказываний. Для удобства визуализации дуги не помечаются вероятностями, вероятность отображается толщиной дуги: чем толще дуга, тем больше вероятность перехода по ней. Также были убраны дуги с вероятностями меньше 0.1: $p(u_j|u_i) < 0.1$. Визуализация графов производилась с помощью пакета GraphViz [30].

На рис. 8 и 9 представлены графы с 5-ю вершинами, составленные по обоим датасетам, на рис. 10 и 11 — графы с 10-ю вершинами, на рис. 12 и 13 — графы с 15-ю вершинами. В целом, графы с 5-ю вершинами выглядят приемлемыми для анализа, дуги с разной толщиной позволяют понять, какие диалоги наиболее вероятны. Хотя Tf-Idf-представление и является простым инструментом маркировки вершин, оно позволяет понять тему высказываний, которые соответствуют вершине. Графы с 10-ю и 15-ю вершинами становятся почти полносвязными, наиболее вероятные диалоги на них менее заметны, интерпретировать такие графы сложнее.

Заключение

Мы формализовали понятие графа диалога, предложили и исследовали на двух наборах данных простой алгоритм построения графа диалога с помощью кластеризации в пространстве представлений SBERT. Были проведены сравнения простых методов кластеризации со SCAN. В результате получены изображения графов диалогов пригодные для визуального анализа. Отметим, что работа оставляет довольно большой задел для будущих исследований:

— автоматическое определение числа вершин в графе (простейший вариант решения — использование алгоритма DBSCAN, см., например, [8], [9]),

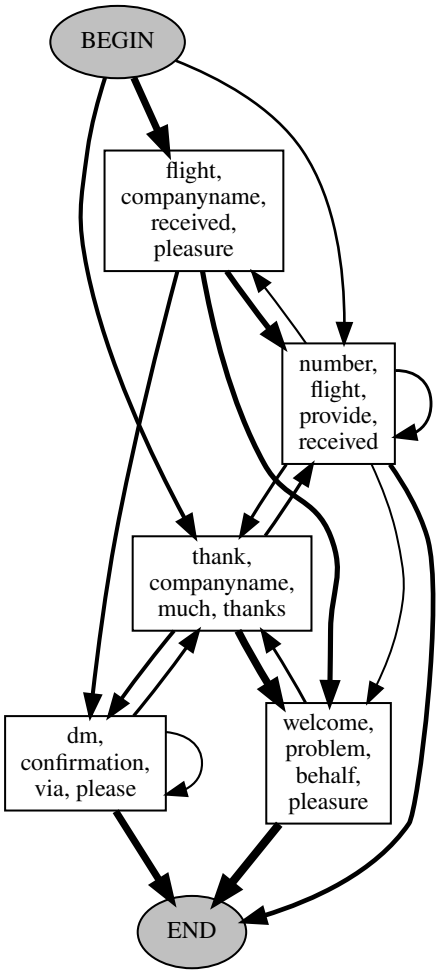


Рис. 8. Граф диалога с 5-ю вершинами для датасета Twitter Customer Support

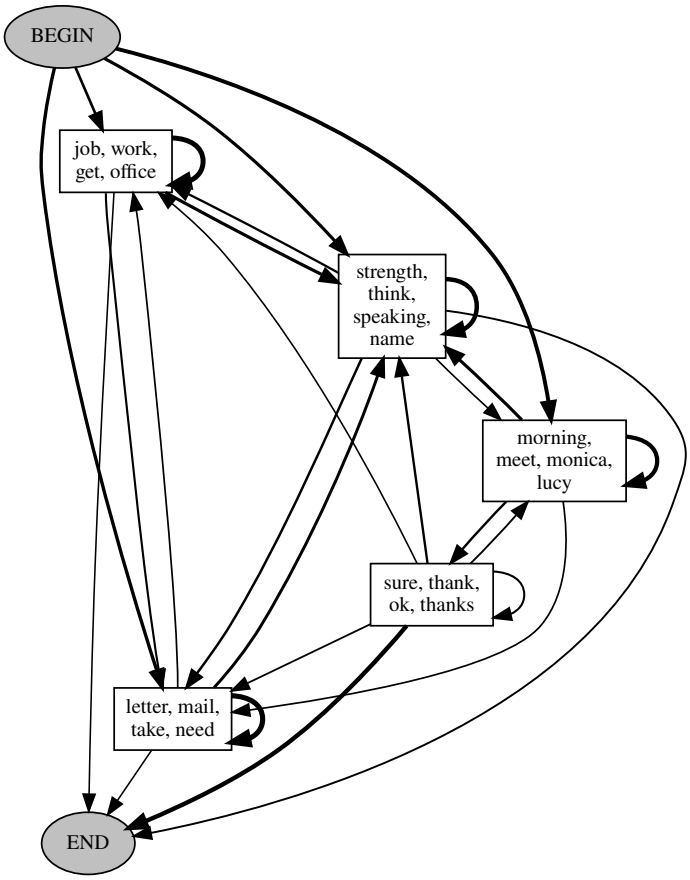


Рис. 9. Граф диалога с 5-ю вершинами для датасета DailyDialog

— автоматическое определение высказываний, которые не соответствуют вершинам (например, отклонения от темы, здесь также можно использовать DBSCAN),
— автоматическая пометка вершин (хотелось бы, чтобы оно производилось полноценным предложением, здесь напрашивается применить технику реферирования, в [8] рассматривался вариант использования высказывания, чьё представление наиболее близко к центру соответствующего кластера),
— исследование оптимального представления высказываний и оптимальной кластеризации (это более эффективно решается с наборами данных, заточенных под решаемую задачу, например, если графы диалогов построены экспертами или заданы изначально как в STAR [17]),
— проблема сравнения графов диалогов (в идеале сравниваются ответы разных алгоритмов, а не промежуточные результаты их работы),
— проблема построения нескольких графов (как описано во введении, по-видимому, такая постановка задачи ранее не рассматривалась).

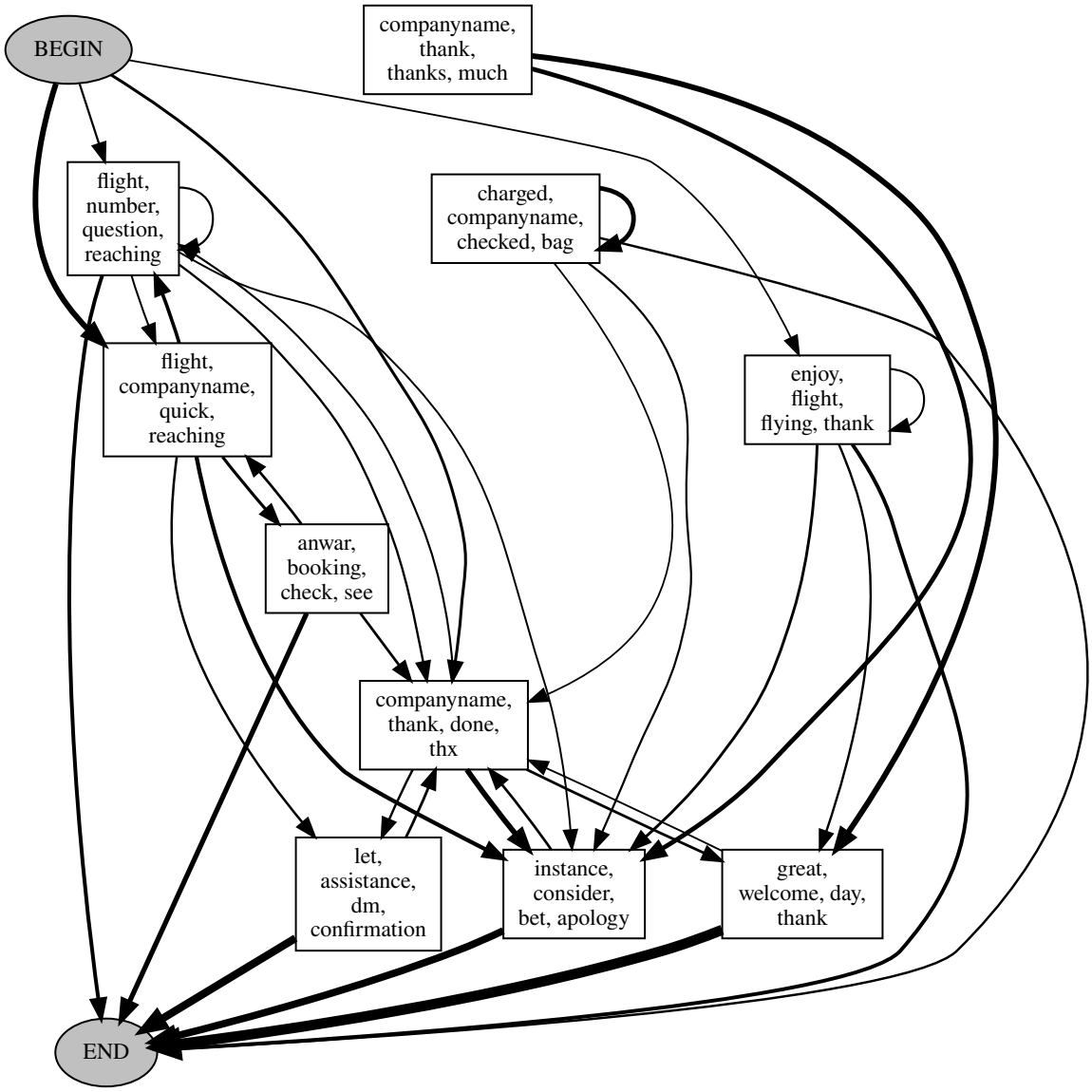


Рис. 10. Граф диалога с 10-ю вершинами для датасета Twitter Customer Support

ЛИТЕРАТУРА

1. *Chotimongkol A.* Learning the structure of task-oriented conversations from the corpus of in-domain dialogs // Ph.D. thesis. Carnegie Mellon University. 2008.
2. *Tang D., Li X., Gao J., Wang C., Li L. and Jebara T.* Subgoal Discovery for Hierarchical Dialogue Policy Learning // EMNLP. 2018.
3. *Shi W., Zhao T. and Yu Z.* Unsupervised Dialog Structure Learning // ArXiv. 2019. V. abs/1904.03736.
4. *Qiu L., Zhao Y., Shi W., Liang Y., Shi F., Yuan T., Yu Z. and Zhu S.* Structured Attention for Unsupervised Dialogue Structure Induction // ArXiv. 2020. V. abs/2009.08552.
5. *Chung J., Kastner K., Dinh L., Goel K., Courville A. and Bengio Y.* A Recurrent Latent Variable Model for Sequential Data // ArXiv. 2015. V. abs/1506.02216.
6. *Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser L. and Polosukhin I.* Attention Is All You Need // ArXiv. 2017. V. abs/1706.03762.
7. *Xu J., Lei Z., Wang H., Niu Z., Wu H., Che W. and Liu T.* Discovering Dialog Structure Graph for Open-Domain Dialog Generation // ArXiv. 2020. V. abs/2012.15543.

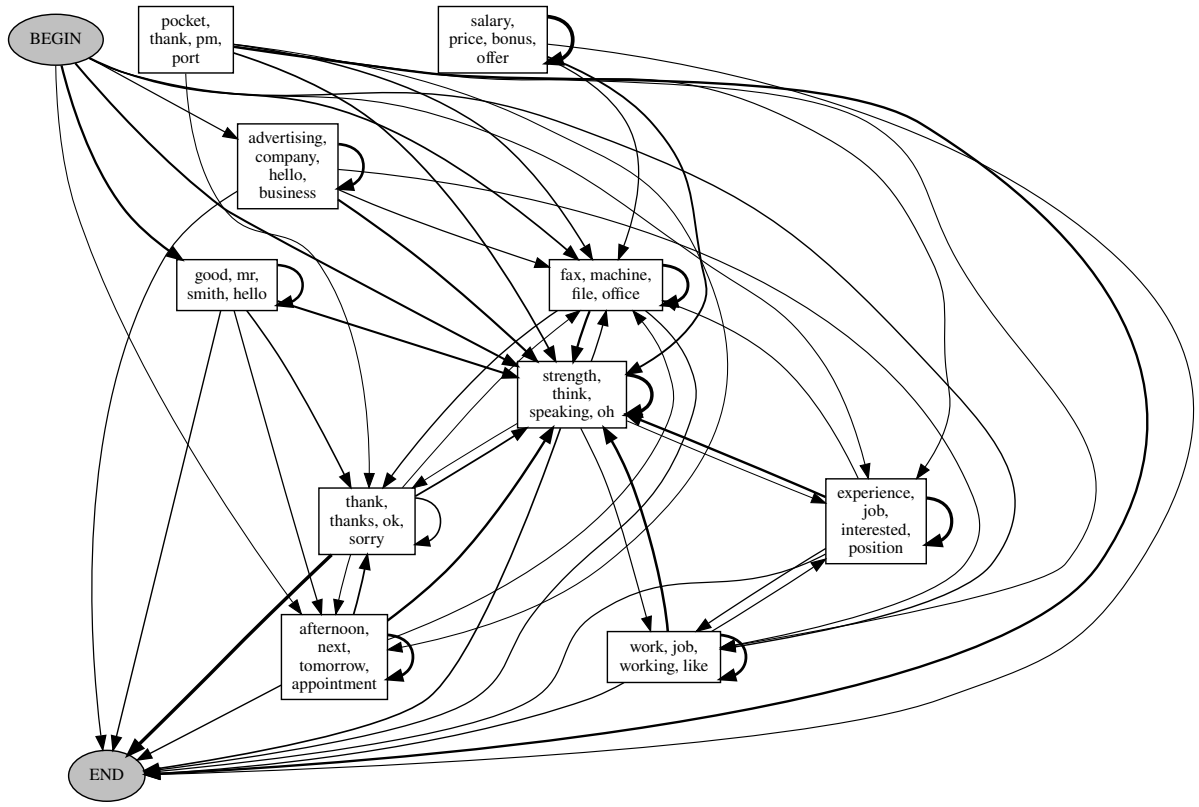


Рис. 11. Граф диалога с 10-ю вершинами для датасета DailyDialog

8. *Юсупов И. Ф., Трофимова М. В. и Бурцев М. С.* Построение и использование диалогового графа для улучшения оценки качества в целенаправленном диалоге. // ТРУДЫ МФТИ. 2020. Т. 21. № 3. С. 75–86.
9. *Фельдина Е. А. и Махныткина О. В.* Автоматическое построение дерева диалога по неразмеченным текстовым корпусам на русском языке // Научно-технический вестник информационных технологий, механики и оптики. 2021. Т. 21. № 5. С. 709–719.
10. *Nath A. and Kubba A.* TSCAN : Dialog Structure discovery using SCAN // ArXiv. 2021. V. abs/2107.06426.
11. *Van Gansbeke W., Vandenhende S., Georgoulis S., Proesmans M. and Van Gool L.* SCAN: Learning to Classify Images without Labels // ArXiv. 2020. V. abs/2005.12320.
12. *Devlin J., Chang M.-W., Lee K. and Toutanova K.* BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // ArXiv. 2018. V. abs/1810.04805.
13. *Reimers N. and Gurevych I.* Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks // ArXiv. 2019. V. abs/1908.10084.
14. *Bishop C.* Pattern recognition and machine learning. New York: Springer, 2006. 424 p.
15. *Bishop C.* Pattern recognition and machine learning. New York: Springer, 2006. 110 p.
16. `https://github.com/PavelShtykov/generalized_dialogue_graph` — Построение и визуализация обобщённого графа диалога по набору диалогов. 2022.
17. *Mosig J., Mehri S. and Kober T.* STAR: A Schema-Guided Dialog Dataset for Transfer Learning // ArXiv. 2020. V. abs/2010.11853.
18. `www.kaggle.com/datasets/thoughtvector/customer-support-on-twitter` — Customer Support on Twitter. 2022.

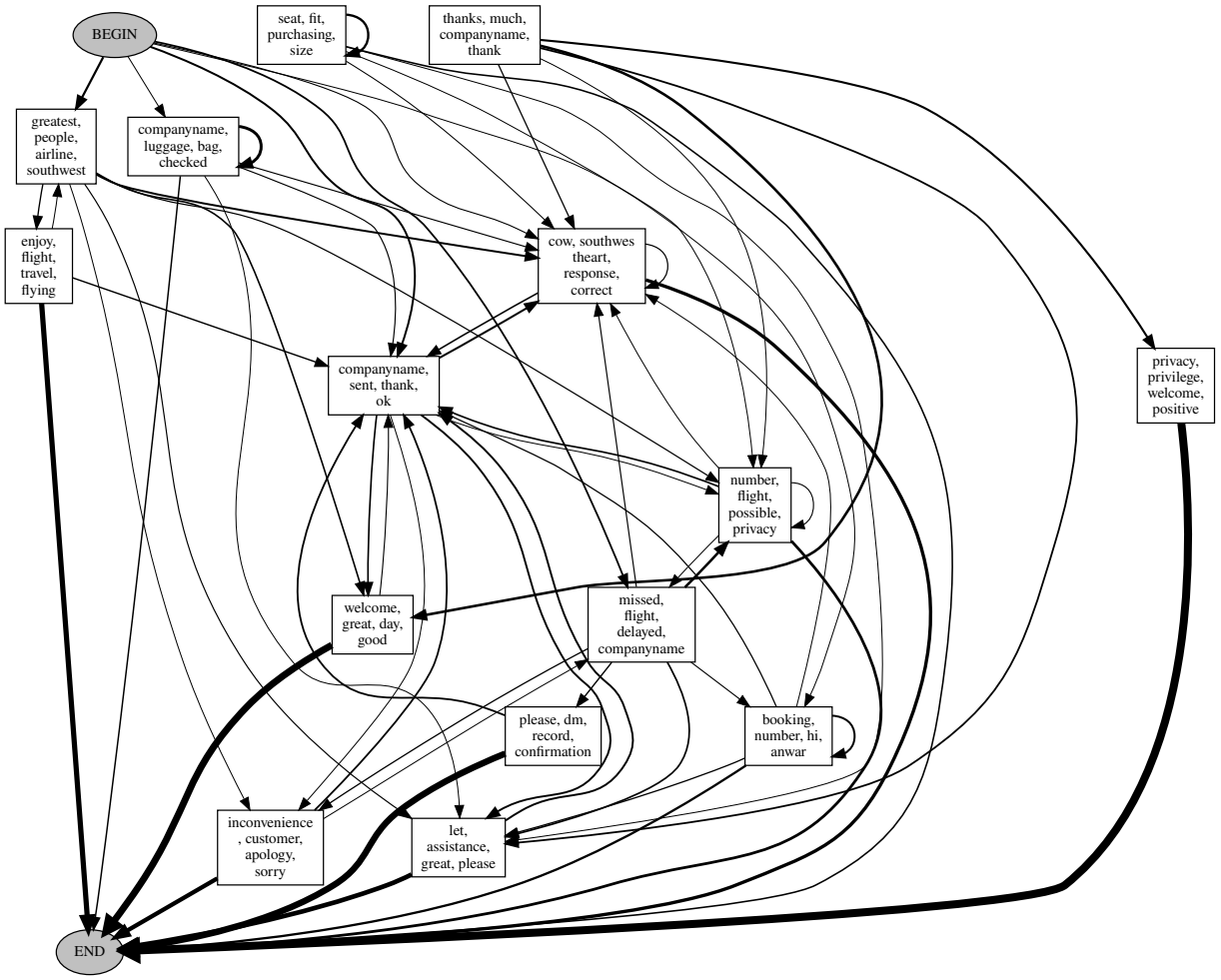


Рис. 12. Граф диалога с 15-ю вершинами для датасета Twitter Customer Support

19. *Li Y., Su H., Shen X., Li W., Cao Z. and Niu S.* DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset // Proceedings of the Eighth International Joint Conference on Natural Language Processing. 2017. V. 1, P. 986–995
20. `https://www.nltk.org` — Natural Language Toolkit. 2022.
21. *Liu Y., Ott M., Goyal N., Du J., Joshi M., Chen D., Levy O., Lewis M., Zettlemoyer L. and Veselin S.* RoBERTa: A Robustly Optimized BERT Pretraining Approach // ArXiv. 2019. V. abs/1907.11692.
22. *Van der Maaten L. and Hinton G.* Viualizing data using t-SNE // Journal of Machine Learning Research. 2008. V. 9. P. 2279–2605.
23. *Song K., Tan X., Qin T., Lu J. and Liu T.-Y.* MPNet: Masked and Permuted Pre-training for Language Understanding // ArXiv. 2020. V. abs/2004.09297.
24. *Sanh V., Debut L., Chaumond J. and Wolf T.* DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter // ArXiv. 2019. V. abs/1910.01108.
25. *Wang W., Wei F., Dong L., Bao H., Yang N. and Zhou M.* MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers // ArXiv. 2020. V. abs/2002.10957.
26. *Rousseeuw P. J.* Silhouettes: A graphical aid to the interpretation and validation of cluster analysis // Journal of Computational and Applied Mathematics. 1987. V. 20. P. 53–65.

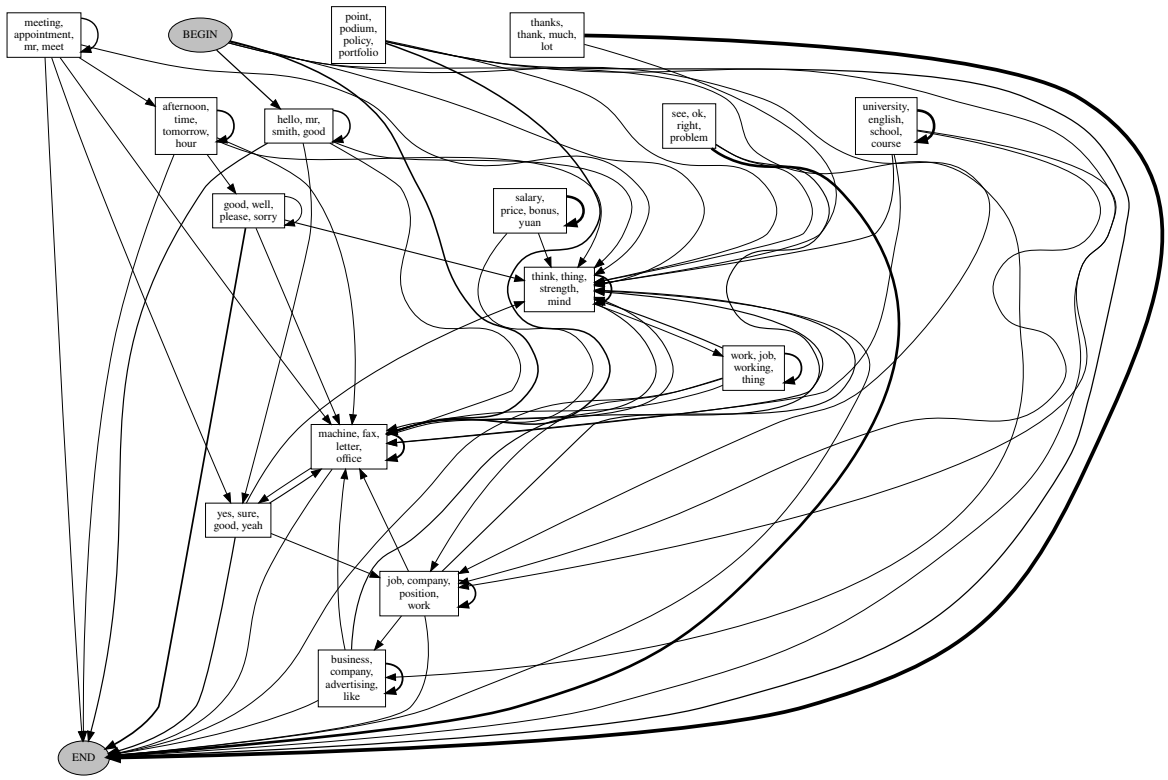


Рис. 13. Граф диалога с 15-ю вершинами для датасета DailyDialog

27. *Caliński T. and Harabasz J.* A Dendrite Method for Cluster Analysis // Communications in Statistics - Theory and Methods. 1974. V. 3. №. 1. P. 1–27.
28. *Davies D. L. and Bouldin D. W.* A Cluster Separation Measure // IEEE Transactions on Pattern Analysis and Machine Intelligence. 1979. V. 1. №. 2. P. 224–227.
29. *Spärck K. J.* A statistical interpretation of term specificity and its application in retrieval // Journal of Documentatio. 2004. V. 60. P. 493–502.
30. `https://graphviz.org` — Graphviz: open source graph visualization software. 2022.

REFERENCES

1. *Chotimongkol A.* Learning the structure of task-oriented conversations from the corpus of in-domain dialogs. Ph.D. thesis, Carnegie Mellon University, 2008.
2. *Tang D., Li X., Gao J., Wang C., Li L. and Jebara T.* Subgoal Discovery for Hierarchical Dialogue Policy Learning. EMNLP, 2018.
3. *Shi W., Zhao T. and Yu Z.* Unsupervised Dialog Structure Learning. ArXiv, 2019, vol. abs/1904.03736.
4. *Qiu L., Zhao Y., Shi W., Liang Y., Shi F., Yuan T., Yu Z. and Zhu S.* Structured Attention for Unsupervised Dialogue Structure Induction. ArXiv, 2020, vol. abs/2009.08552.
5. *Chung J., Kastner K., Dinh L., Goel K., Courville A. and Bengio Y.* A Recurrent Latent Variable Model for Sequential Data. ArXiv, 2015, vol. abs/1506.02216.
6. *Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser L. and Polosukhin I.* Attention Is All You Need. ArXiv, 2017, vol. abs/1706.03762.
7. *Xu J., Lei Z., Wang H., Niu Z., Wu H., Che W. and Liu T.* Discovering Dialog Structure Graph for Open-Domain Dialog Generation. ArXiv, 2020, vol. abs/2012.15543.
8. *Yusupov I. F., Trofimova M. V. and Burtsev M. S.* Postroyeniye i ispol'zovaniye dialogovogo grafa dlya uluchsheniya otsenki kachestva v tselenapravlennom dialoge [Unsupervised graph

extraction for improvement of multi-domain task-oriented dialogue modelling]. TRUDY MFTI, 2020, vol. 21, no. 3, pp. 75–86. (in Russian)

9. *Feldina E. A. and Makhnytkina O. V.* Avtomaticheskoye postroyeniye dereva dialoga po nerazmechennym tekstovym korpusam na russkom yazyke [Automatic construction of the dialog tree based on unmarked text corpora in Russian]. Nauchno-tekhnicheskiy vestnik informatsionnykh tekhnologiy, mekhaniki i optiki, 2021, vol. 21, no 5, pp. 709–719. (in Russian)
10. *Nath A. and Kubba A.* TSCAN : Dialog Structure discovery using SCAN. ArXiv, 2021, vol. abs/2107.06426.
11. *Van Gansbeke W., Vandenhende S., Georgoulis S., Proesmans M. and Van Gool L.* SCAN: Learning to Classify Images without Labels. ArXiv, 2020, vol. abs/2005.12320.
12. *Devlin J., Chang M.-W., Lee K. and Toutanova K.* BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv, 2018, vol. abs/1810.04805.
13. *Reimers N. and Gurevych I.* Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. ArXiv, 2019, vol. abs/1908.10084.
14. *Bishop C.* Pattern recognition and machine learning. New York, Springer, 2006. 424 p.
15. *Bishop C.* Pattern recognition and machine learning. New York, Springer, 2006. 110 p.
16. `https://github.com/PavelShtykov/generalized_dialogue_graph` — A generalized dialogue graph construction and visualization based on a corpus of dialogues, 2022.
17. *Mosig J., Mehri S. and Kober T.* STAR: A Schema-Guided Dialog Dataset for Transfer Learning. ArXiv, 2020, vol. abs/2010.11853.
18. `www.kaggle.com/datasets/thoughtvector/customer-support-on-twitter` — Customer Support on Twitter, 2022.
19. *Li Y., Su H., Shen X., Li W., Cao Z. and Niu S.* DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. Proceedings of the Eighth International Joint Conference on Natural Language Processing, 2017, vol. 1, pp. 986–995
20. `https://www.nltk.org` — Natural Language Toolkit, 2022.
21. *Liu Y., Ott M., Goyal N., Du J., Joshi M., Chen D., Levy O., Lewis M., Zettlemoyer L. and Veselin S.* RoBERTa: A Robustly Optimized BERT Pretraining Approach. ArXiv, 2019, vol. abs/1907.11692.
22. *Van der Maaten L. and Hinton G.* Viualizing data using t-SNE. Journal of Machine Learning Research, 2008, vol. 9, pp. 2279–2605.
23. *Song K., Tan X., Qin T., Lu J. and Liu T.-Y.* MPNet: Masked and Permuted Pre-training for Language Understanding. ArXiv, 2020, vol. abs/2004.09297.
24. *Sanh V., Debut L., Chaumond J. and Wolf T.* DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. ArXiv, 2019, vol. abs/1910.01108.
25. *Wang W., Wei F., Dong L., Bao H., Yang N. and Zhou M.* MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. ArXiv, 2020, vol. abs/2002.10957.
26. *Rousseeuw P. J.* Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 1987, vol. 20, pp. 53–65.
27. *Caliński T. and Harabasz J.* A Dendrite Method for Cluster Analysis. Communications in Statistics - Theory and Methods, 1974, vol. 3, no. 1, pp. 1–27.
28. *Davies D. L. and Bouldin D. W.* A Cluster Separation Measure. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1979, vol. 1, no. 2, pp. 224–227.
29. *Spärck K. J.* A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation, 2004, vol. 60, pp. 493–502.
30. `https://graphviz.org` — Graphviz: open source graph visualization software, 2022.

Штыков Павел Дмитриевич — бакалавр, МГУ имени М. В. Ломоносова, г. Москва. E-mail: **shtykov.pa@gmail.com**

Дьяконов Александр Геннадьевич — доктор физико-математических наук, МГУ имени М. В. Ломоносова, г. Москва. E-mail: **djakonov@mail.ru**