Московский государственный университет имени М. В. Ломоносова
Факультет вычислительной математике и кибернетики
Кафедра математических методов прогнозирования

Штыков Павел Дмитриевич

Построение обобщенного графа диалога

КУРСОВАЯ РАБОТА

Научный руководитель:
д.ф.-м.н., профессор
А. Г. Дьяконов

Москва, 2022

Содержание

1 Введение

Обработка естественного языка (NLP) является ключевой задачей в машинном обучении, а обработка диалогов является важной ее частью. Одной из слабо изученных областей в обработке диалогов является проблема построения и представления общей структуры диалога.

Естественно предположить, что у диалогов из одной области может быть некоторая общая структура. Так же естественно представлять эту структуру в виде графа. Такой граф позволяет представить информацию о корпусе однородных диалогов в сжатой форме, подходящей как для визуализации, так и для встраивания в более сложные диалоговые системы. В данной работе мы предложим формализацию понятия обобщенного графа диалога и несколько базовых способов его построения и визуализации.

2 Существующие подходы

К сожалению, нам не удалось найти большое число работ связанных с данной темой. Существует серия статей [9], [6], в которой предлагается два способа построения графа диалога. Первый основан на использовании вариационных автокодировщиков в паре с рекуррентной нейронной сетью [2]. Во второй статье авторы добавили в свою архитектуру механизм внимания [11], что позволило им улучшить результат. Результат их работы можно увидеть на Рис. 1.

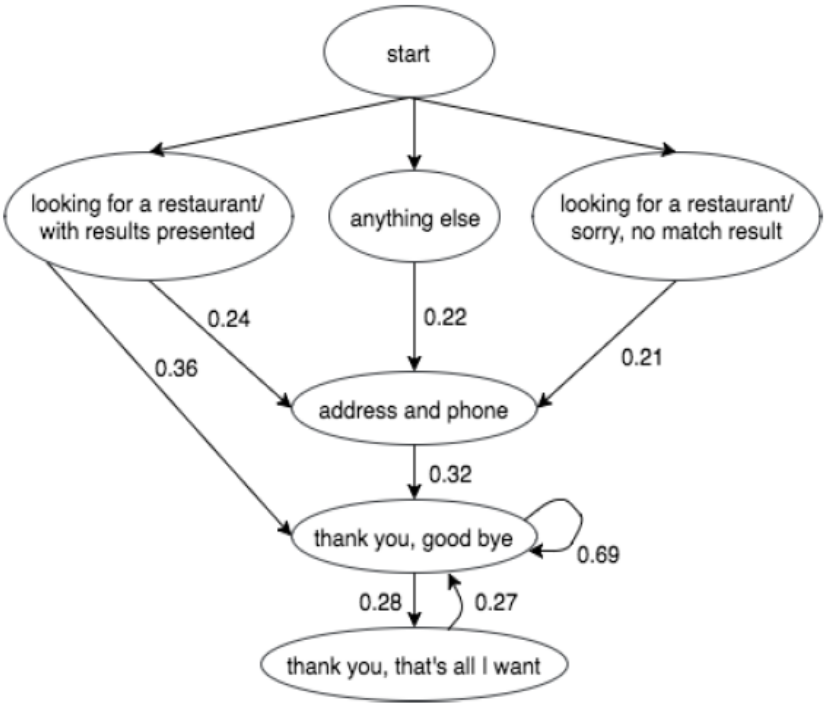

Рис. 1: Пример обобщенного графа диалога взятого из статьи [9]

В другой работе [5] авторы для построения графа применяют алгоритм классификации SCAN [10], работающий без учителя. Данный алгоритм может самостоятельно разметить кластеры, эту разметку авторы и использовали для описания вершин графа. К сожалению, в данной статье использовался приватный датасет и многие тонкости были опущены.

Во всех упомянутых выше статьях авторы сравнивают свои подходы с методом k-ближайших соседей, однако не приводят самого алгоритма основанного на простой кластеризации. В данной работе мы предложим пример такого алгоритма. Также мы приведем уточнения определения обобщенного графа диалога. На наш взгляд, граф, соответствующий такому определению, проще анализировать и визуализировать.

3 Постановка задачи

Введем определение обобщенного графа диалога.

Определение 1. *Назовем обобщенным графом диалога пару* $T = (G, p(u|v))$*, где:*

- $G = (V, E)$ *— направленный взвешенный граф. С каждым ребром в графе* G *ассоциирована вероятность перехода по этому ребру:* $e_{i,j} \sim p(v_i|v_j)$*. При этом сумма вероятностей ребер выходящих из каждой вершины равна* 1*:* $\sum_j p(v_i|v_j) = 1$
- $u \in U$ *— единичное высказывание, а* U *— пространство 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*.*

Такое определение не ограничивает нас в выборе модели для его построения. 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 диалога *«пользователь»-«система»*) — введением раскраски вершин, т.е. дополнительной функции $\phi(v)$, ставящей в соответствие каждой вершине некоторый персональный идентификатор пользователя (ID). Однако необходимо ввести дополнительные ограничения. Так как высказывания пользователей чередуются, то логично потребовать, чтобы вершины разных цветов не были инцидентны. Также основное определение не запрещает петли. В персонализированном графе их стоит запретить.

Дополнительно, пусть $D = \{d_1, d_2, \ldots, d_{|D|}\}$ — выборка диалогов, где d_i — один диалог. Каждый диалог является 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 диалогами. Однако, в общем случае нет ограничения на использование разметки.

Добавим к каждому диалогу x_i технические высказывания *начала* и *конца* диалога ($BEGIN$ и END). Аналогичные вершины добавим и в граф. Вероятность $p(u|v)$ для этих вершин будет вырождена в соответствующих точках в пространстве высказываний U. Это необходи-

мо для более ясной конструкции графа и соблюдения ограничения на сумму вероятностей ребер, исходящих из вершины: $\sum_j p(v_i|v_j) = 1$.

4 Предложенный метод

Для более простой и понятной работы с высказывания перейдем в промежуточное пространство — воспользуемся эмбеддингом (*embedding*):

$$Embedding : U \to M$$

В качестве эмбеддинга мы использовали предобученную сиамскую нейронную сеть [7], где в качестве основной сети использовалась дистиллированная версия RoBERTa [4], [8]. В дальнейшем, если не оговорено другого, под высказыванием u мы будем подразумевать его эмбеддинг $Embedding(u)$.

Пространство M метрическое (в нашем случае это $\mathbb{R}^{768}$) с косинусной метрикой, отражающую семантическую близость высказываний, следовательно в нем можно воспользоваться метрическими метода кластеризации.

Теперь приведем алгоритм построения обобщенного графа диалога T, для высказываний в пространстве эмбеддингов M.

► Пусть существует некоторый алгоритм кластеризации a:

$$a : M \to V$$

В данном случае множество вершин V есть множество кластеров. Соответственно может быть вычислена *дискретная* вероятность принадлежности каждого высказывания к каждой вершине:

$$a(u) = p(v|u)$$

При этом кластеризация может быть как жесткой (например методом k-ближайших соседей (k-NN)), так и мягкой (например смесью гауссиан (GMM)).

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 $p(u)$ частота встречаемости высказывания u во всем корпусе диалогов. Заметим, что вероятность $p(u)$ не одинакова для всех высказываний, так как в корпусе могут встречаться диалоги с одинаковыми высказываниями.

Нам осталось построить в графе ребра и найти вероятности, ассоциированные с ними. Введем в пространстве высказываний U граф $\hat{G}$, подобный графу G, т.е. ориентированный взвешенный граф с вероятностями, ассоциированными с ребрами:

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

Данный граф строиться напрямую по выборке диалогов, и ребра в нем имеют смысл апостериорной вероятности встретить ответ u_j на высказывание u_i. Соответственно матрица смежности $\hat{A}$ графа $\hat{G}$ определяется как:

$$\hat{A} = (\hat{a}_{ij}), \quad \hat{a}_{ij} = p(u_j|u_i), \quad i, j = \overline{1, |U|}$$

Аналогично матрица смежности вводиться и для основного графа G:

$$A = (a_{ij}), \quad a_{ij} = p(v_j|v_i), \quad i, j = \overline{1, |V|}$$

Так как в нашем случае совместные распределения $p(u|v)$ и $p(v|u)$ дискретны, то они могут быть представлены в виде матриц. Следовательно матрица смежности A получается простым перемножением трех матриц:

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

Мы закончили построение обобщенного графа диалога T. ■

Заметим, что данный алгоритм построения обобщенного графа применим не только в случае использования кластеризации в пространстве эмбеддингов, но и в случае использования любого другого алгоритма способного оценить апостериорные вероятности $p(v|u)$ (например с помощью латентного размещения Дирихле (LDA) или нейронной сети, решающей задачу от начала до конца без промежуточного использования эмбеддингов). Кластеризация была выбрана, как наиболее простой метод.

4.1 Метрики качества

Предложим базовые метрики качества построенного графа.

Нам хотелось бы, чтобы апостериорные вероятности $p(u|v)$ как можно меньше пересекались между собой. Следующая метрика принимает значение 1, когда дискретные распределения не пересекаются между собой для каждой пары вершин в графе G, и 0, когда распределения совпадают полностью:

$$M_1(T) = 1 - \frac{1}{C^2_{|V|}} \cdot \sum_{i<j} \frac{\sum_{k=1}^{|U|} \min\{p(u_k|v_i),\, p(u_k|v_j)\}}{\sum_{k=1}^{|U|} \max\{p(u_k|v_i),\, p(u_k|v_j)\}}$$

TODO: *Функция правдоподобия...*

5 Эксперименты

Нам не известен датасет полностью соответствующий поставленной задаче, т. е. датасет, состоящий из диалогов, имеющих некоторую общую *известную* структуру. Поэтому для проведения экспериментов нами было выбрано два стандартных датасета, в которых, на наш взгляд, можно предположить наличие общей структуры. Первый датасет — *Customer Support on Twitter* [1], в котором собраны ответы официальных аккаунтов технической поддержки крупных американских компаний. Мы выбрали из него подмножество сообщений аккаунтов шести разных авиакомпаний, чтобы сделать датасет более однородным. Второй датасет — *DailyDialog* [3], в котором собраны обычные диалоги из повседневной жизни на разные темы. Для экспериментов мы взяли диалоги на тему отношений.

Для начала рассмотрим t-SNE визуализацию пространства эмбеддингов для диалогов (Рис. **??** и **??**). Для диалогов из обоих датасетов заметна некоторая кластерная структура. Это дает нам некоторое подтверждение наличия общей структуры у корпуса диалогов.

[caption]tsne_twitter.png[caption]

TODO: *Сравнение разных кластеризаторов и эмбеддингов по метрике*

Построим и визуализируем графы для обоих датасетов. На рис. 2 представлен обобщенный граф для диалогов технической поддержки авиакомпаний. В качестве базового алгоритма кластеризации использовался простой метод k-ближайших соседей. Количество вершин равно 5. В качестве меток вершин были выбраны 3 слова с самым большим значением Tf-Idf.

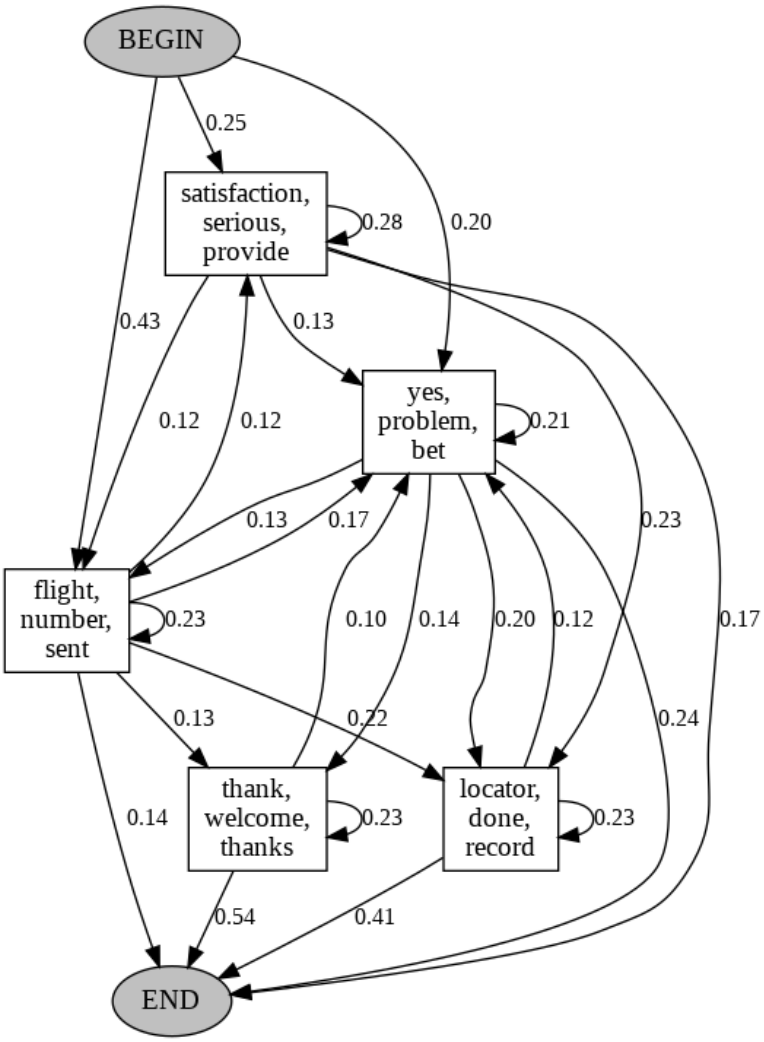

Рис. 2: Обобщенный граф диалогов технической поддержки авиакомпаний в Twitter

TODO: *Тут будут еще графы... Для большего количества вершин тоже... Я не успел написать :(((*

6 Заключение

TODO: ...

Список литературы

[1] Customer support on twitter. `https://www.kaggle.com/datasets/thoughtvector/customer-support-on-twitter`. Доступ: 05.05.2022.

[2] Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel, Aaron C. Courville, and Yoshua Bengio. A recurrent latent variable model for sequential data. *CoRR*, abs/1506.02216, 2015.

[3] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. DailyDialog: A manually labelled multi-turn dialogue dataset. pages 986–995, November 2017.

[4] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.

[5] Apurba Nath and Aayush Kubba. TSCAN : Dialog structure discovery using SCAN. *CoRR*, abs/2107.06426, 2021.

[6] Liang Qiu, Yizhou Zhao, Weiyan Shi, Yuan Liang, Feng Shi, Tao Yuan, Zhou Yu, and Song-Chun Zhu. Structured attention for unsupervised dialogue structure induction. *CoRR*, abs/2009.08552, 2020.

[7] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. 11 2019.

[8] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108, 2019.

[9] Weiyan Shi, Tiancheng Zhao, and Zhou Yu. Unsupervised dialog structure learning. *CoRR*, abs/1904.03736, 2019.

[10] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. Scan: Learning to classify images without labels. 2020.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. 2017.