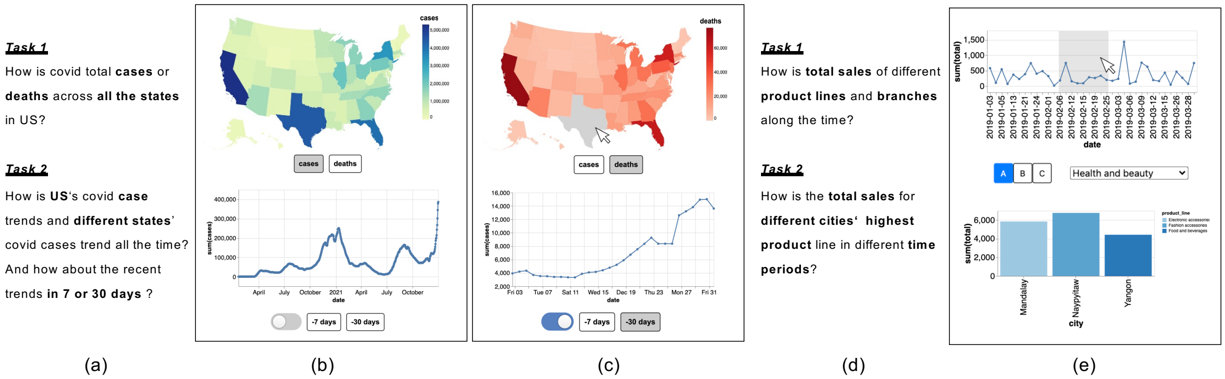
PI2-NL: Interactive Visualization Interface Generation from Natural Language Tasks

Roy G. Biv[*]
Starbucks Research

Ed Grimley[†]
Grimley Widgets, Inc.

Martha Stewart[‡]
Martha Stewart Enterprises
Microsoft Research



Task 1

How is covid total **cases** or **deaths** across **all the states** in US?

Task 2

How is **US**'s covid **case** trends and **different states**' covid cases trend all the time? And how about the recent trends **in 7 or 30 days** ?

Task 1

How is **total sales** of different **product lines and branches** along the time?

Task 2

How is the **total sales** for **different cities‘ highest product** line in different **time periods**?

Figure 1: (a) shows two natural language tasks for COVID-19 dataset. (b) is the interface generated for the tasks listed in (a). (c) shows that users can interact with interface to study the deaths distribution by clicking the "deaths" on the middle buttons and the Texas recent 30 days' trend by clicking the state "Texas" on the Map visuallization and clicking the button "-30 days" at the bottom. (d) shows two natural language tasks for a supermarket sales dataset. (e) is the interactive interface generated for tasks in (d). The top visualization shows the total sales of task 1 where the branch is specified by the button widget and the product line is specified by the dropdown widget. The bottom visualization shows each city's highest product line's sale and the time period is specified by the brushing over the first visualization.

Abstract

We propose PI2-NL, the first system to automatically generate interactive multi-visualization interface from natural language tasks.

1 Introduction

Interactive visualization interfaces (or simply *interfaces*) play a critical role in nearly every stage of data management—including data cleaning [13], wrangling [8], modeling [5], exploration [2, 10], and communication [4, 6].

Such interfaces require considerable expertise and trial-and-error to design and implement because the charts, interactions, and layout should be chosen to support the underlying analysis task [9]. Our recent work PI2 [3, 12] uses SQL as proxy for analysis task, and is the first system to automatically generate a fully interactive multi-visualization interface from SQL analysis. Under the hood, PI2 proposes DIFFTREEs as a compact representation of an anlysis task, models interface generation as a schema mapping problem, and search for an optimal interface mapping given a cost model. PI2 helps designer automatically and effectively translate analysis, i.e. SQL queries into interfaces so that user can focus on the analysis without worrying about complicated interface design and implementation.

Although SQL is ubiquitous in data analysis, it is still hard for normal people, especially non-programmers to write. Specifying the analysis task in natural language would be a much more accessible and promising way for everyone. Recently, advancements in translation based NLP technologies [1, 11] have vastly improved accuracy of producing schema-aware SQL queries from natural language question. Combined with PI2, we propose **PI2-NL, the first system to automatically generate interactive multi-visualization interface from natural language tasks.** PI2-NL finetunes Codex [] model to output DIFFTREE representation from input natural language task and use PI2-NL to generate interactive interface for the input analysis. Yiru: doubel check NLP model description

Below are two end-to-end examples:

Example 1 (Covid Analysis) *Given the COVID-19 dataset, a user is interested in two tasks - "How is covid total cases or deaths across all the states in US?", and "How is US's covid case trends and different states' covid cases trend all the time? And how about the recent trends in 7 or 30 days ? " as shown in Figure 1(a).* PI2-NL *will generate an interactive visualization interface in Figure 1(b). The map visualization corresonding to the first task. Users can interact with the middle button widget to choose to show the death distribution in Figure 1(c). The line chart in Figure 1(b) shows the US overall trend. Users can toggle on to specify the date range e.g. recent 30 days and click on the map visualization to filter specific state, e.g. "Texas". After these interactions, Figure 1(c)'s bottom line chart shows the Texas' recent 30 days covid cases trend.*

Example 2 (Sales Analysis) *For the supermarket sales dataset [7], in Figure 1(d), a user writes two analysis tasks - "How is total sales of different product lines and branches along the time?", and "How is the total sales for different cities‘ highest product line in different time periods?" Figure 1(e) shows the interface where the above visualization shows the total sales of task 1 and the bottom visualization shows each city's highest product line's sale. Notice*

[*] e-mail: roy.g.biv@aol.com
[†] e-mail: ed.grimley@aol.com
[‡] e-mail: martha.stewart@marthastewart.com

that the task 2' analysis is a complilcated query analysis which has to first find each city's highest product line in different period and then calculte the total sales. This results in complicated subqueris rather simple SPJA queries. For the interface, users can interact the button widget and the dropdown widget to specify the branch and product line. Brushing over the first visualizaiton will specify the time period for the second analysis task. With such an interactive interface, users can easily explore different branches, product lines and periods.

As we can see, with PI2-NL, users can purely focus on write analysis task in natural language, and PI2-NL will automatically return a fully interactive multi-visualization interfaces that can perform the data analysis task.

There is also a line of work which answer natural language task or dialoge with visualizaiotn

they focus on visualizaito specification around one certain question,

PI2-NL is beyond this in that it takes consideration of multi view, it considers interface characteristic

yet, PI2-NL clearly differ them they mainly focus on single visualization specify / what's more, since PI2-NL use sql DIFFTREE as intermediate representation, it can express complicated query where these work can not.

Above all, this paper contribute PI2-NL

which is able to we organize paper in the following related work system overview

2 RELATED WORK

2.1 NL to vis

2.2 NLP model to SQL

NLP model to predict sql

3 PI2-NL SYSTEM

4 DISCUSSION

Attribute ambiguity

5 CONCLUSION

REFERENCES

[1] M. Chen, J. Tworek, H. Jun, Q. Yuan, H. P. d. O. Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

[2] Y. Chen and S. Huang. Tsexplain: Surfacing evolving explanations for time series. *Proceedings of the 2021 International Conference on Management of Data*, 2021.

[3] Y. Chen and E. Wu. Pi2: End-to-end interactive visualization interface generation from queries. In *Proceedings of the 2022 International Conference on Management of Data*, pp. 1711–1725, 2022.

[4] FiveThirtyEight. All posts tagged data visualization. https://fivethirtyeight.com/tag/data-visualization/, 2021.

[5] Google. Facets - know your data. https://pair-code.github.io/facets/, 2021.

[6] iCheck. icheck. icheckuclaim.org, 2021.

[7] Kaggle. Dataset: Supermarket sales. `https://www.kaggle.com/aungpyaeap/supermarket-sales`, 2021.

[8] S. Kandel, A. Paepcke, J. Hellerstein, and J. Heer. Wrangler: interactive visual specification of data transformation scripts. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2011.

[9] T. Munzner. *Visualization analysis and design*. CRC press, 2014.

[10] D. Murray. Tableau your data!: Fast and easy visual analysis with tableau software. 2013.

[11] T. Scholak, N. Schucher, and D. Bahdanau. Picard: Parsing incrementally for constrained auto-regressive decoding from language models. *arXiv preprint arXiv:2109.05093*, 2021.

[12] J. Tao, Y. Chen, and E. Wu. Demonstration of pi2: Interactive visualization interface generation for sql analysis in notebook. In *Proceedings of the 2022 International Conference on Management of Data*, SIGMOD '22, p. 2365–2368. Association for Computing Machinery, New York, NY, USA, 2022. doi: 10.1145/3514221.3520153

[13] E. Wu and S. Madden. Scorpion: Explaining away outliers in aggregate queries. 2013.