Jiawei Tang (Research Portfolio)

Research Internships

Massachusetts Institute of Technology (MIT), USA (2022/06-08)

Mentor Professor Samuel Madden

Project Use deep learning models to solve the problem of entity resolution. Entity resolution is the task of deciding whether two data records refer to the same real-world object. It has diversified application domains such as banking, insurance, e-commerce, health care, and many others. For example, an e-commerce company wants to know if two products from different suppliers are the same so they can be displayed on the same product page; two banks sharing data need to identify and reconcile common customers.

Contributions I was responsible for system design, implementation, and testing for two tasks: determining the accuracy of foundation models for entity resolution and designing a deep-learning model for generic entity resolution.

Publication **First author** of paper “Generic Entity Resolution Models” accepted by Table Representation Learning Workshop @ NeurIPS 2022, where NeurIPS is one of the most prestigious and competitive international conferences in machine learning and computational neuroscience.

Link `https://openreview.net/pdf?id=tRkVo1jMas`

Qatar Computing Research Institute, Qatar (2021/06-08)

Mentor Dr. Mourad Ouzzani

Project Build an end-to-end data visualization system that acts as a virtual assistant to allow novices to create visualizations through either natural language or speech.

Contributions Designed and implemented two main components: Speech-to-Text which is based on Google Cloud Speech-to-Text Rest API, and Text-to-VIS, which uses an end-to-end neural machine translation model.

Publication **First author** of paper “Sevi: Speech-to-Visualization through Neural Machine Translation” accepted by ACM SIGMOD International Conference on Management of Data, where SIGMOD is a leading international forum for database researchers. I presented and demonstrated this work in SIGMOD 2022 @Philadelphia.

Link `https://dl.acm.org/doi/pdf/10.1145/3514221.3520150`

Tsinghua University, China (2020/06-08)

Mentor Professor Guoliang Li

Project Construct a benchmark of (natural language, data visualization) pairs and use this benchmark to train a deep learning model that translate a natural language query into a data visualization.

Contributions Used Python toolkits to clean and annotate data. Used PyTorch and Transformer models to train a deep learning model to support the translation from natural language queries to data visualizations.

Publication **Co-author** of paper “Natural Language to Visualization by Neural Machine Translation” accepted by IEEE Transactions on Visualization and Computer Graphics 2021, a top journal for data visualization.

Link `https://ieeexplore.ieee.org/document/9617561`

Note: Please see the following three pages for the abstracts of the above three papers.

Generic Entity Resolution Models

Jiawei Tang[*]
American School of Doha
23jtang@asd.edu.qa

Yifei Zuo
University of Science and Technology of China
dune@mail.ustc.edu.cn

Lei Cao
MIT CSAIL/Univeristy of Arizona
lcao@csail.mit.edu

Samuel Madden
MIT CSAIL
madden@csail.mit.edu

Abstract

Entity resolution (ER) – which decides whether two data records refer to the same real-world object – is a long-standing data integration problem. The state-of-the-art results on ER are achieved by deep learning based methods, which typically convert each pair of records into a distributed representation, followed by using a binary classifier to decide whether these two records are a match or a non-match. However, these methods are dataset specific; that is, one deep learning based model needs to be trained or fine-tuned for each new dataset, which is not generalizable and thus we call them *specific ER models*. In this paper, we investigate *generic ER models*, which use a single model to serve multiple ER datasets over different datasets from various domains. In particular, we study two types of generic ER models: Employs foundation models (*e.g.,* GPT-3) or trains a generic ER model. Our results show that although GPT-3 can perform ER with zero-shot or few-shot learning, the performance is worse than specific ER models. Our trained generic ER model can achieve comparable performance with specific ER models, but with much less train data and much smaller storage overhead.

1 Introduction

Entity resolution (ER) (*a.k.a.* record linkage), a fundamental problem of data integration [5] and cleaning [1], has been extensively studied for several decades [7] from different aspects, including: declarative rules [22, 29]; machine learning based methods (or probabilistic methods) [2, 12], deep learning based methods [13, 18, 8]; and crowdsourcing based methods [28]. ER has a wide spectrum of critical applications such as healthcare [7], e-commerce [10], data warehouses [30], etc.

Deep learning for entity resolution. The state-of-the-art results on ER are achieved by deep learning based methods [13, 18, 8]. These methods typically consist of two steps: using a feature extractor (*i.e.,* an encoder) to convert an entity pair into a representation, and then employing a binary classifier to map this representation to a Boolean output as either a match (1) or a non-match (0). These solutions are dubbed as ***specific ER models***, because each model serves only one ER dataset.

Limitations of specific ER models. There are three main limitations.

1. *Need a lot of train data for each new ER dataset.* Existing deep learning based ER solutions, even with pre-trained language models as the encoder (*e.g.,* Ditto [13] uses BERT [6] and RoBERTa [14]), require a lot of labeled train data for each new ER dataset, while labeled train data for ER is expensive to obtain. Although transfer learning [26, 25, 16] and active

[*]Work done while interning at MIT CSAIL.

Table Representation Learning Workshop at NeurIPS 2022.

Sevi: Speech-to-Visualization through Neural Machine Translation

Jiawei Tang† American School of Doha, Qatar
Yuyu Luo* Tsinghua University Beijing, China
Mourad Ouzzani QCRI, HBKU Doha, Qatar
Guoliang Li Tsinghua University Beijing, China
Hongyang Chen Zhejiang Lab Hangzhou, China
23jtang@asd.edu.qa,{luoyy18@mails.,liguoliang@}tsinghua.edu.cn,mouzzani@hbku.edu.qa,hongyang@zhejianglab.com

ABSTRACT

Data visualization is a powerful tool for understating information through visual cues. However, allowing novices to create visualization artifacts for what they want to see is not easy, just as not everyone can write SQL queries. Arguably, the most natural way to specify *what to visualize* is through natural language or speech, similar to our daily search on Google or Apple Siri, leaving to the system the task of reasoning about *what to visualize and how*.

In this demo, we present **Sevi** an end-to-end data visualization system that acts as a virtual assistant to allow novices to create visualizations through either natural language or speech. **Sevi** is powered by two main components: `Speech2Text` which is based on Google Cloud Speech-to-Text Rest API, and `Text2VIS`, which uses an end-to-end neural machine translation model called **ncNet** trained using a cross-domain benchmark called **nvBench**. Both **ncNet** and **nvBench** have been developed by us. We will walk the audience through two general domain datasets, one related to COVID-19 and the other on NBA player statistics, to highlight how **Sevi** enables novices to easily create data visualizations. Because **nvBench** contains `Text2VIS` training samples from 105 domains (*e.g.,* sport, college, hospital, etc.), the audience can play with speech or text input with any of these domains.

CCS CONCEPTS

• **Information systems → Data analytics**; • **Human-centered computing → Visualization**; **Visualization systems and tools**.

KEYWORDS

Speech-to-Visualization; Natural Language-to-Visualization

ACM Reference Format:
Jiawei Tang†, Yuyu Luo*, Mourad Ouzzani, Guoliang Li, and Hongyang Chen. 2022.**Sevi**: Speech-to-Visualization through Neural Machine Translation. In *Proceedings of the 2022 International Conference on Management of Data (SIGMOD '22), June 12–17, 2022, Philadelphia, PA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3514221.3520150

† Work done while interning at QCRI, Qatar.
* Yuyu Luo is the corresponding author.

SIGMOD '22, June 12–17, 2022, Philadelphia, PA, USA.

ACM ISBN 978-1-4503-9249-5/22/06...$15.00
https://doi.org/10.1145/3514221.3520150

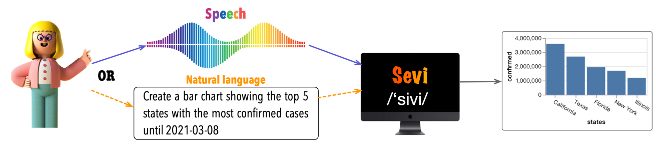

Figure 1: A user provides input in the form of voice (speech) or text (natural language). Sevi translates either input into a visualization.

1 INTRODUCTION

Data is taking the world by storm, transforming virtually every industry, and is playing an important role in our daily lives. It is important to understand the insights that numbers alone cannot tell us. However, it is nontrivial to interpret the massive amounts of information being collected today.

Data visualization plays a key role in communicating information, through the use of visual elements such as bar charts, scatter plots, and histograms [16]. This makes the data more natural for the human mind to comprehend and therefore provides an accessible way for anyone, even those without statistical background, to identify trends, patterns, and outliers within large datasets [8, 9, 21]. In fact, we have been inundated with visual interpretations of the COVID-19 data, from early graphics urging us to flatten the pandemic curve to regularly updated dashboards [5, 7, 12].

Although there are many choices of interactive data visualization tools (*e.g.,* Tableau and Qlik) and easy-to-specify data visualization languages (*e.g.,* Vega-Lite [17] and ggplot2), only experts are able to to create good visualizations. In addition, this assumes that these experts know many details such as the meaning and the distribution of the data, the right combination of attributes, and the right type of charts.

The democratization of data visualization means that anyone can easily create data visualizations without the need to write code and with a very fast learning curve, similar to how Google democratized *search* using a natural language interface. In fact, both commercial vendors (*e.g.,* Tableau's Ask Data [18], Power BI [2], ThoughtSpot [3], and Amazon's QuickSight [1]) and academic researchers [4, 10, 15, 20] have investigated the translation from natural language queries to visualizations (`Text2VIS`). They mainly use statistical phrase-based translation that first employs natural language processing toolkits (*e.g.,* Stanford CoreNLP [14] and NER [6]) to parse a natural language query and produce a variety of linguistic annotations (*e.g.,* parts of speech, named entities, etc.), based on which they then devise algorithms to generate target visualizations.

Natural Language to Visualization by Neural Machine Translation

Yuyu Luo, Nan Tang, Guoliang Li*, Jiawei Tang, Chengliang Chai, Xuedi Qin

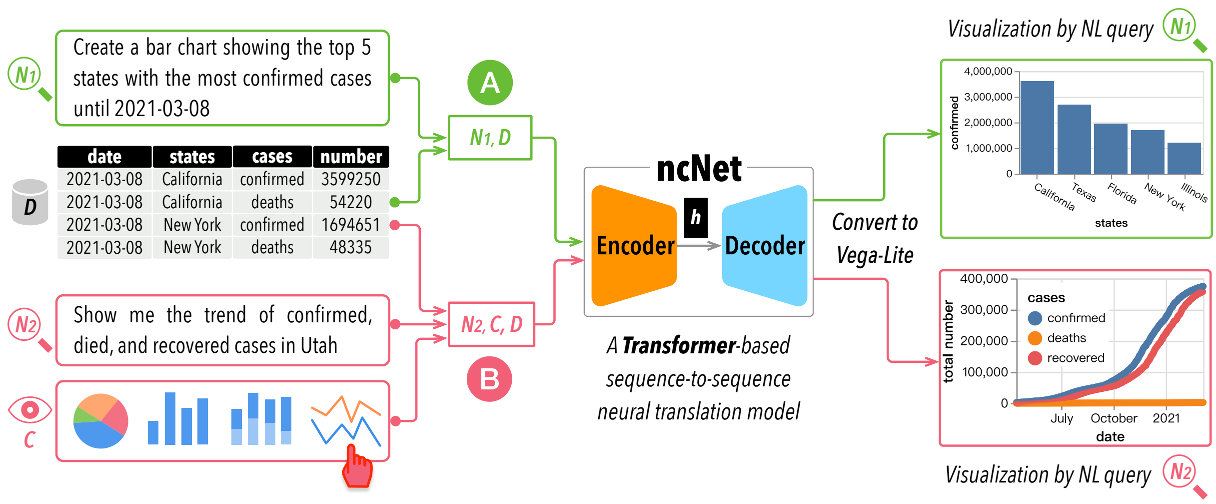

Fig. 1: We present **ncNet**, a Transformer-based sequence-to-sequence model that translates natural language queries to visualizations. It works in two modes. (A) It takes a natural language query N_1 and a dataset D as input, translates them (N_1, D) into a visualization rendered in Vega-Lite. (B) Besides a natural language query N_2 and a dataset D, the user can optionally select a chart template C; **ncNet** will translate the given input (N_2, C, D) into a target visualization.

Abstract— Supporting the translation from natural language (NL) query to visualization (NL2VIS) can simplify the creation of data visualizations because if successful, anyone can generate visualizations by their natural language from the tabular data. The state-of-the-art NL2VIS approaches (*e.g.,* NL4DV and FlowSense) are based on semantic parsers and heuristic algorithms, which are not end-to-end and are not designed for supporting (possibly) complex data transformations. Deep neural network powered neural machine translation models have made great strides in many machine translation tasks, which suggests that they might be viable for NL2VIS as well. In this paper, we present **ncNet**, a Transformer-based sequence-to-sequence model for supporting NL2VIS, with several novel visualization-aware optimizations, including using attention-forcing to optimize the learning process, and visualization-aware rendering to produce better visualization results. To enhance the capability of machine to comprehend natural language queries, **ncNet** is also designed to take an optional chart template (*e.g.,* a pie chart or a scatter plot) as an additional input, where the chart template will be served as a constraint to limit what could be visualized. We conducted both quantitative evaluation and user study, showing that **ncNet** achieves good accuracy in the **nvBench** benchmark and is easy-to-use.

Index Terms—Natural language interface; data visualization; neural machine translation; chart template;

1 Introduction

Natural language interface is a promising interaction paradigm for simplifying the creation of visualizations [32,43,52]. If successful, even novices can generate visualizations simply like a Google search. Not surprisingly, both commercial vendors (*e.g.,* Tableau's Ask Data [46], Power BI [2], ThoughtSpot [3], and Amazon's QuickSight [1]) and academic researchers [7,13,20,33,34,40,42,45,49,50,57] have investigated to support the translation from NL queries to visualizations (NL2VIS).

NL2VIS needs both natural language understanding that uses machines to comprehend natural language queries, and translation algorithms to generate targeted visualization using a visualization language. Natural language understanding is considered an AI-hard problem [56], with many intrinsic difficulties such as ambiguity and underspecification. Many tools from the NLP community, especially based on statistical phrase-based translation [26] and neural machine translation [4,10], have been used to tackle NL2VIS.

The state-of-the-art NL2VIS methods (for example, NL4DV [40] and FlowSense [57]) are statistical phrase-based translation, which treats natural language understanding and machine translation as two steps. They first employ NLP toolkits (for example, NLTK [5], Stanford CoreNLP [37], and NER [12]) to parse an NL query and produce a variety of linguistic annotations (for example, parts of speech, named entities, etc), based on which they then devise algorithms to generate target visualizations. They are good choices when there are not many training datasets to train deep learning models.

We present **ncNet**[1], an end-to-end solution using a Transformer-based sequence-to-sequence (seq2seq) model, which translates an NL query to a visualization. It adopts self-attention to generate a rich repre-

- *Yuyu Luo, Guoliang Li, Chengliang Chai, Xuedi Qin are with the Department of Computer Science, Tsinghua University, China. Email: {luoyy18@mails., liguoliang@, ccl@, qxd17@mails.}tsinghua.edu.cn*
- *Nan Tang is with QCRI, Hamad Bin Khalifa University, Qatar. Email: ntang@hbku.edu.qa*
- *Jiawei Tang is with American School of Doha, Doha, Qatar. Email: 23jtang@asd.edu.qa*
- **Guoliang Li is the corresponding author.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

[1] The code is available at `https://github.com/Thanksyy/ncNet`