氏　　名　宇治橋 善史

学位(専攻分野)　博士(情報学)

学 位 記 番 号　総研大甲第 2363 号

学位授与の日付　2022 年 9 月 28 日

学位授与の要件　複合科学研究科 情報学専攻
学位規則第6条第1項該当

学 位 論 文 題 目　Synthesizing Tabular Transformations from Examples using Transformer-based Neural Networks

論 文 審 査 委 員　主　査　高須 淳宏
情報学専攻 教授
相澤 彰子
情報学専攻 教授
石川 冬樹
情報学専攻 准教授
片山 紀生
情報学専攻 准教授
加藤 弘之
情報学専攻 助教
阿久津 達也
京都大学 化学研究所 教授

（様式３）

博士論文の要旨

氏　　名 宇治橋　善史

論文題目 Synthesizing Tabular Transformations from Examples using Transformer-based Neural Networks

Data analysis with a large size of data from various public or enterprise data sources allows us to discover new knowledge or rules that no one have found so far. Data analysis thus has recognized as a promising process for decision makings by business persons or policy makings by governments in these days.

Data preparation task is an essential process as the first step of data analysis. Data from various sources is often in a non-relational or unstructured form and is not able to be directly imputed into a downstream information system, like a database or visualization systems. Data analysts transform such raw data into a desired format that can be easily consumed by other systems. Such a data preparation task regularly involves reformatting data values, reconstructing layout of tables, looking up values with keys from other tables, and integrating multiple data sources, often requiring programming skills. Therefore, it is laborious and time-consuming task for data analysts who have limited coding skills. It is generally said that data analysts spend more time for preparing data than analyzing it: up to 80% of data analyst's time.

Programming by example (PBE) is a technique that makes this troublesome task easier for data analysts by automatically generating programs for data transformation. PBE is one research field of program synthesis. In a program synthesis problem, a program synthesizer takes a program specification from users and then automatically generates a program according to the specification. Meanwhile, in a PBE problem, a program synthesizer takes an input-output example as a specification from users and then automatically generates a program consistent with the input-output example. The system with a PBE program synthesizer should solve the data preparation problem. It allows a user to synthesize a program through specifying a desired transformation by providing input-output example. The user only needs to know how to describe the transformed data without knowing any particular transformation operation or programming code.

In past years, researchers have been studied PBE using techniques based on non-neural algorithms such like graph search algorithm or version space algebra. Many researchers have begun to study PBE using techniques based on ML approach in recent years, inspired from many successes of machine learning (ML) or neural network models. One example of such study, RobustFill, has been proposed as a neural network model that generates a string transformation program. It employs an encoder-decoder model that consists of two long-short-term-memory (LSTM) architectures. Once the model is trained via supervised training, it translates an input-output strings into the corresponding transformation program.

Although RobustFill shows a feasibility of ML-based PBE system and possibility of its application to string transformation, it does not support both of syntactic (namely string) and layout

transformations. Data preparation consists of syntactic transformations where each cell content of a table re-formatted and layout transforms where the layout of a table is re-constructed. We call these transforms as tabular transformation collectively.

Our goal of this dissertation is to realize an ML-based PBE for tabular transformations. This is the first ML-based PBE for tabular transformations to the best of our knowledge. Furthermore, our experiments show that our neural model outperforms the existing non- neural PBE system for tabular transformations, thus indicating that our neural approach to synthesizing tabular transformation programs is promising. Our contributions are as follows.

First, we propose a new ML-based PBE system for tabular transformations. The ML- based PBE system is an encoder-decoder model based on the Transformer neural network which is said as the state-of-the-art translation neural network. Since tabular transformations have more intricate data structures and complicated transformations than string transformations, it requires larger expressive power, namely the larger number of parameters, of the neural network models. The LSTM network which is used in conventional ML-based PBE systems is difficult to have such large parameters, because the LSTM network spends much longer time to train its parameters due to its sequential processing feature of recurrent neural networks. Thus, the LSTM cannot have such an expressive power that is capable of learning tabular transformations in a practical training time. To address this shortage, we propose a Transformer-based model as an ML-based PBE system instead of the LSTM.

Next, we propose an embedding method that embeds two-dimensional tabular data into the Transformer neural network model. We introduce tabular positional encodings which encodes the positions of each location of the tabular data to deal with the tabular data in the Transformer model properly. This method allows us to embed each row index, column index and the local position in a cell of an input-output example tables to represent two (or more)-dimensional positions in the Transformer network. Our experiments show that the tabular positional encodings improve the performance of Transformer-based model through learning and capturing the structure of two-dimensional tabular data.

Finally, we propose two decoding methods, multistep beam search and Program Vali- dation (PV)-Beam Search. Our Transformer-based model generates a sequence which corresponds to a program for an input-output example tables from the Transformer de- coder in the encoder-decoder model. The Transformer decoder firstly outputs how likely each program component occurs according to the input-output examples provided, and next it generates the most likely programs from the likelihoods. Generally, the generations of the most likely programs from the likelihoods are performed by beam search. However, the beam search is not a technique designed for program generation, thereby causing an inefficient exploration of the program search space. Our proposed two variants of beam search are optimized for the program generation task, hence generating correct programs at higher probability than the original beam search. Our experimental results ensure that the Transformer-based model with proposed decoding method outperforms the conventional state-of-the-art PBE system for tabular transformations.

Form8 : Separate Sheet
（様式８・別紙１）

Results of the doctoral thesis screening

博士論文審査結果

Name in Full
氏 名　　宇治橋　善史

Title
論文題目　　Synthesizing Tabular Transformations from Examples using Transformer-based Neural Networks

出願者は，Programing by Example(PBE)の手法を利用して表形式データの入出力例から，データ変換をするプログラムを生成する深層学習モデルを提案した．従来の文字列を対象とした PBE の深層学習モデルで実現されていなかった，Transformer モデルをベースにした 2 次元構造データのエンコード法や効率的なプログラム空間の探索法を提案し，例から表形式データ変換プログラムを高精度に生成する点に新規性がある．

本学位論文は全 7 章より構成され英語で書かれている．

第１章では，データ分析や利活用に必要となるデータプレパレーションが分析者の高いハードルになっている背景を述べ，この課題を解決する技術として本研究が対象とする表形式データ変換 PBE を実現する深層学習モデルの重要性を示し，モデル構築における本博士論文の貢献を述べている．

第２章では，本博士論文が対象とする表形式データ変換 PBE を定義している．まず，入力および出力の表形式データ対が与えられた時に，入力表形式データを出力表形式データに変換するオペレーション列を生成する問題として形式的に定義している．また，変換に用いるオペレーションの集合を定義し，その表現力について述べている．

第３章では，近年の関連研究の動向を述べるとともに，特に本研究と関係の深い，PBE，エンコード・デコードモデル，2 次元構造データ埋め込み，ビームサーチなどの関連技術の研究を概観している．

続く２つの章で本博士論文の主たる貢献を述べている．まず第４章で表形式データの 2 次元構造を深層学習モデルで扱うための埋め込み方式を提案している．本博士論文では自然言語処理の分野で高い性能を実現している Transformer をベースにし，表形式データの２次元構造を扱うために，位置埋め込み層として Tabular Positional Encoding を提案している．表形式データの各セルに表中の位置の情報を埋め込むもので，その有効性を実験により示している．

第５章では提案モデルを用いてプログラムを効率的に生成するプログラム空間の探索法を提案している．探索法としてビームサーチを用いることが多いが，本博士論文が扱うプログラム生成では精度に課題があることを実験的に示し，この課題を解決する multi-step beam search と PV-beam search を提案し，その有効性を実験的に示している．

６章では，４章と５章で提案した手法を組み合わせたプログラム生成法を既存の手法と比較し，従来法よりも高い性能が得られることを実験的に示している．続いて，提案する PBE システムに与える表形式データ対と精度の関係について議論している．まず，表形式データ対のサイズと精度の関係を評価する generalization 指標で提案手法を評価し，既存

手法よりも高い性能を有していることを実験的に示している．さらにプログラム生成に失敗するケースの分析やデータプレパレーションシステムへ組み込む際の課題について考察している．

第 7 章で以上の結果と今後の研究への指針をまとめている．

公開発表会では博士論文の主要な貢献を中心に発表が行われた．その後に行われた論文審査会および口述試験では，審査員からの質疑に対して適切に回答がなされた．

質疑応答後に審査委員会を開催し審査委員で議論を行った．博士論文審査の結果，出願者は情報学分野の十分な知識と研究能力を持つと認められた．研究内容は、データプレパーレーションのための PBE に効果的な表形式データの埋め込み表現と表形式データ変換プログラム空間の探索アルゴリズムを提案するものであり新規性が認められる．既存手法に比べプログラム生成精度が向上することを実験的に示しており実用面での貢献も期待できる．また、本学位論文の成果は，国際学術雑誌に 1 編，査読付き国際会議に 1 編の主著論文が採択されており，学術的な貢献も認められる．以上の理由により，審査委員会は本学位論文が学位の授与に値すると判断した．