A DATA DICTIONARY BASED APPROACH TO SEMANTIC TABULAR MAPPING

Matthew Johnson

Submitted in Partial Fullfillment of the Requirements
for the Degree of

DOCTOR OF PHILOSOPHY

Approved by:
Deborah L. McGuinness, Chair
James A. Hendler
Mohammed J. Zaki
Nicholas R. Del Rio

Department of Computer Science
Rensselaer Polytechnic Institute
Troy, New York

[August 2024]

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGMENT

This work has been supported by the Air Force Research Laboratory's Civilian Academic Degree Program. Research reported in this dissertation was supported by the National Institutes of Health (NIH) under award number U2CES026555. Publicly available data used in this work was generated through grants supported by the NIH as part of the Human Health Exposure Analysis Resource (HHEAR). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

This work would not have been possible without the guidance of my adviser, Deborah McGuinness, who taught me how to stand on my own as a researcher. I will never forget what I have learned here. I owe a special thanks to Nicholas Del Rio, who always made time to discuss a problem or help me focus my ideas. I also want to thank the rest of my committee, James Hendler and Mohammed Zaki, for their questions and guidance, which have elevated my research. I'm lucky to have worked with the Tetherless World Constellation faculty, staff, and students. I have learned much from our discussions, debates, and reading clubs. It has been an honor growing alongside you, and I look forward to seeing what you do in the future.

Finally, I am deeply grateful to my family and friends for their unwavering support during the long work days. Your love and understanding have been my anchor, and I wouldn't have reached this point without you.

ABSTRACT

Knowledge graphs are an important technology that enables a wide variety of analytics and data visualizations across an enterprise. However, creating knowledge graphs or adding to an existing knowledge graph can be challenging because data is often stored in a semi-structured form within tables that do not capture the full context of the data. To fill the context gap many data publishers include a data dictionary that aims to capture the meaning of schema elements, typically with text descriptions. These descriptions are helpful for humans to understand the data for integration tasks but are challenging for machines to interpret.

Previous work has focused on integrating tables into an existing knowledge graph using data-level alignments without the additional context provided by data dictionaries. While these data-level alignment algorithms have proven successful on synthetic datasets, they struggle to make accurate alignments on real-world datasets that exhibit complex structures. Humans overcome these issues by leveraging context information from data dictionaries to understand the groups and relationships among the entities within a table. Recently, data publishers have started using this metadata to create semantic data dictionaries (SDDs) that formally capture alignments between tabular data. These alignments allow data publishers to convert tabular data into Resource Description Framework (RDF) triples and create or integrate data into a knowledge graph. However, SDDs require authors to have domain knowledge and experience in ontology modeling, which creates a barrier to entry for users.

In this thesis, our goal is to improve the field of data integration by exploring algorithms that leverage context information from data dictionaries to align complex tabular data to ontology classes and properties. To achieve this, we address three key research questions: *Can algorithms effectively use context from data dictionaries to improve alignment on complex tables? Are alignment algorithms that leverage data dictionaries competitive with data-level alignment algorithms on simple tables? What type of data dictionary descriptions are well suited for alignment algorithms?* For the first research question, we developed the Semantic Data Dictionary Generator (SDD-Gen), a tabular alignment algorithm that generates SDDs by leveraging context information from data dictionaries. We show the effectiveness of SDD-Gen by comparing the performance against the current state of the art on complex tables. For the second research question, we developed a methodology for generating representative artificial data dictionaries using large language models. We use this methodology to generate data dictionaries for a popular tabular alignment dataset and show that SDD-Gen is as effective as the data-level algorithms on simple tables. For the final research question, we developed an evaluation framework to determine the type of data dictionary description best suited for tabular alignment. We show that intensional descriptions that define the conditions needed to be a member of a column are most effective and improve the reusability of data dictionaries.

CHAPTER 1
INTRODUCTION

1.1 Overview

In 2012, Google popularized the term knowledge graph and defined it as a network of named entities where nodes represented different entities and edges captured the relationships between them [1]. While the term is relatively new, the concept of the semantic web has been around much longer [2]. This technology has been critical because it enables the integration of diverse datasets, improves the capabilities of search engines [1], powers question-answering services [3], and can be used to generate explanations for inference engines [4]. As a result, knowledge graphs have become an important tool for both industry and governments. Unfortunately, creating or modifying an existing knowledge graph can be challenging because knowledge graphs encode meaning through structure and typically require statements to be unambiguous and logically consistent [5]. In contrast, most source data is stored in semi-structured forms, such as tables which use locality to encode relationships within the data. This encoding is typically optimized to the data publisher's original task, which often requires additional context for third parties to interpret and reuse a table for subsequent unanticipated tasks. The semantic tabular interpretation task addresses this problem by aligning table components to ontological terms and integrating tabular data into knowledge graphs using these alignments. However, efforts to automate semantic tabular interpretation have focused on data-level alignments within the table and, as a result, struggle with tables that have complex encodings. In this dissertation, we introduce a new algorithm for semantic tabular interpretation that leverages additional context from data dictionaries to make tabular alignments, enabling us to align data from complex tables.

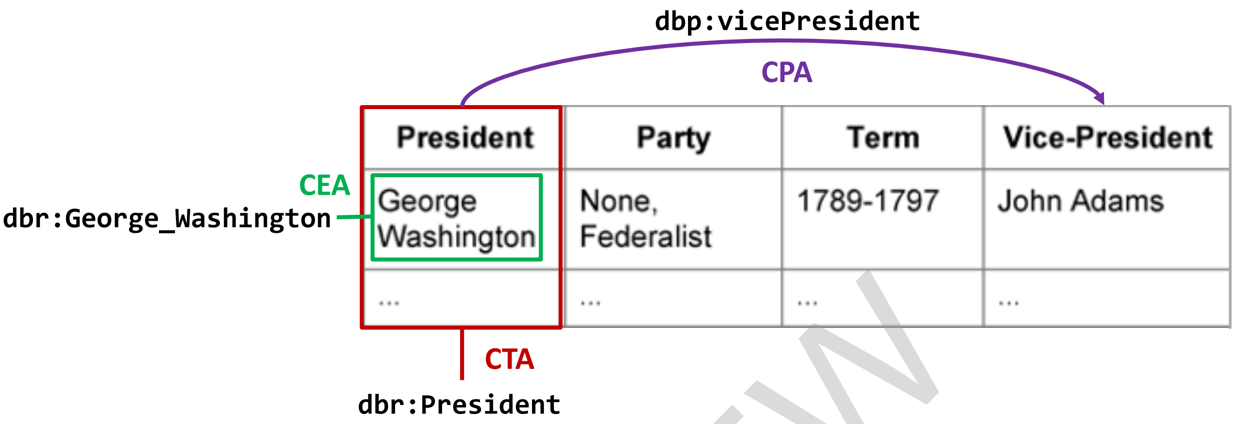
1.2 Semantic Tabular Interpretation

The semantic tabular interpretation problem focuses on integrating tables into an existing knowledge graph and is often broken down into three subtasks:

- Cell-Entity Annotation (CEA): Matching a cell to an entity within a knowledge graph

Portions of this chapter have previously appeared as: M. Johnson, J. A. Stingone, S. Bengoa, J. Masters, and D. L. McGuinness, "Complex semantic tabular interpretation using sdd-gen," in *18th Int. Conf. on Semantic Comput. (ICSC)*. 2024, pp. 317-322.

- Column-Type Annotation (CTA): Assigning an ontology class to a column
- Column-Property Annotation (CPA): Assigning an ontology property between two columns



President	Party	Term	Vice-President
George Washington	None, Federalist	1789-1797	John Adams
...	...	...	...

Figure 1.1: Example of the three semantic tabular interpretation subtasks over a table

Previous work automating semantic tabular interpretation has focused on developing algorithms that integrate tables into an existing knowledge graph using data-level alignments. These algorithms generally align named entities in the table to a knowledge graph entity. These alignments are then used to vote on a column's upper-level class and the relationships between subject and non-subject columns. The yearly Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab) competition at the International Semantic Web Conference aims to evaluate the performance of semantic tabular interpretation algorithms. Their work has created a common framework to systematically evaluate alignment algorithms, which has led to real progress upon which we build. SemTab has demonstrated that data-level alignment algorithms can be effective at semantic tabular interpretation [6], [7], [8], [9]; however, there are several limitations to these solutions.

One key limitation of current solutions is that they require an overlap of named entities within the knowledge graph and table. This means that data-level alignment algorithms can't be used to bootstrap new knowledge graphs, and tables without named entities can't be aligned. There are also open questions on what percentage of entity overlap is needed for these algorithms to be effective, which further adds to the complexity of the problem. Second, tables are expected to be in third normal form, where every other column is a property of the subject column [10]. For instance, in Figure 1.1, the “President” column is the subject column because “Party”, “Term”, and “Vice-President” are all properties of a president. This format requirement is severely limiting; in 2008, researchers extracted 14.1 billion HTML tables using Google’s general-purpose web crawl and estimated that only around 1.1% of tables contain high-quality relational data that met these requirements [11]. For this study, tables were considered to contain high-quality relational data if they were in 3rd normal form where the first column is the subject column, have schema labels, have finite rows, be no less than two rows or two columns, and the content of the table didn’t include visual layout information. Some of the tables in this study were rejected because the content contained non-interesting information, such as web page layout information. Still, others were rejected because they didn’t meet the strict table formatting requirements. We cover the definition of table complexity in great detail in section 1.3 and this study in section 2.2. These restrictions have meant that the SemTab competition has often been forced to evaluate algorithms against synthetic datasets [6], [7], [8], [9].

1.3 Table Complexity

The original definition of what tables should be considered for the semantic tabular interpretation problem required tables to have finite rows, schema labels, a coherent set of domain-appropriate attributes, no less than two rows or two columns, contain relational data, and have the first column be a subject column [11]. The SemTab competition loosened this definition by allowing any column to be the subject column [6]. These restrictions have led data-level alignment algorithms to dominate artificial datasets and struggle with real-world datasets [8]. This is because, in the real world, tables are more complex. In this dissertation, we consider two dimensions of table complexity: 3rd normal form and the number of named entity columns.

In relational databases, 3rd normal form is a schema design approach that simplifies data management. This is done by requiring every non-prime column to be transitively dependent on the prime column. This restriction works the same way in semantic tabular interpretation, requiring every non-subject column to be a property of the subject column. However, it’s common for data publishers to have multiple subject columns or an implied subject in scientific data such as biomedical studies. For example, the data dictionary in table 1.1 describes a table covering various birth statistics. These columns contain properties for the mother and child, where the child is not explicitly identified with a column.

Table 1.1: A data dictionary describing the columns of a table on birth statistics

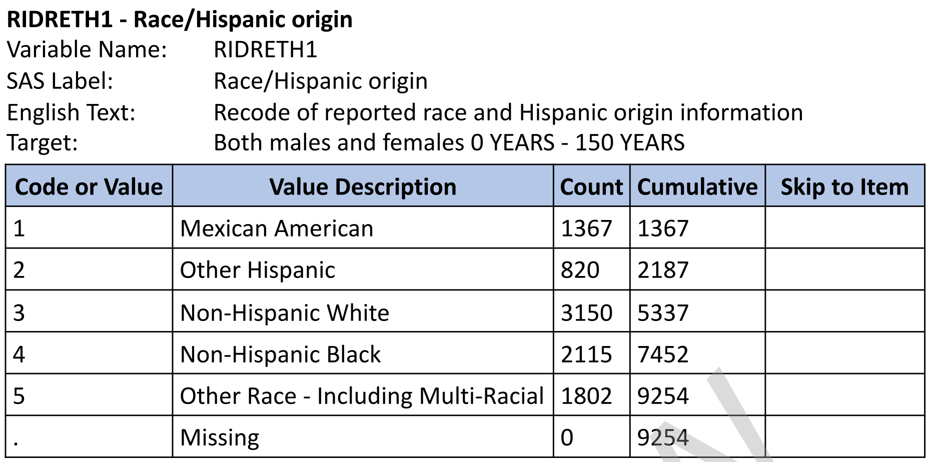
Column Name	Description
pid	The mother's participant ID
birthweight	The child's birth weight
weight	The mother's weight
gender	The child's gender
educat	Mother's highest education level

The other complexity dimension is the number of named entities within a table. Named entities provide an example of class membership within a column. This is the primary data signal that all data-level alignment algorithms rely on to solve the CTA and CPA subtasks. However, it's common for most columns to be numerical in study data. For example, in the table described by the data dictionary in table 1.1, only "educat" and "gender" would contain named entities. The Semtab challenge has begun to explore datasets with fewer named entity columns. However, there are still open questions on how much named entity overlap is sufficient for semantic tabular interpretation [8]. In this dissertation, we consider a table to be complex if it has more than one subject column, real or implied, and most of the columns contain numerical data.

1.4 Data Dictionary

One way to address the issue of complex table alignment is to consider the additional context around the table that asserts groups and relationships among the entities present in the table [12]. To capture this context, many data publishers include data dictionaries alongside their tabular data to provide background context. A data dictionary is a metadata file that can include information such as a description of the data, the relationships between data, the origin of the data, the data usage, and the format of the data [13]. The goal is to allow third-party users to understand and reuse published data easily. Unfortunately, there is no metadata standard that data dictionaries adhere to, and as a result, they come in many shapes and sizes. Common data dictionary fields include text descriptions of schema elements, data ranges that capture the min and max data values, data units, and codebooks

[14], which contain templated codes for organizing text data. In this dissertation, we require data dictionaries to contain a text description of each schema label within the data.



RIDRETH1 - Race/Hispanic origin

Variable Name:	RIDRETH1
SAS Label:	Race/Hispanic origin
English Text:	Recode of reported race and Hispanic origin information
Target:	Both males and females 0 YEARS - 150 YEARS

Code or Value	Value Description	Count	Cumulative	Skip to Item
1	Mexican American	1367	1367	
2	Other Hispanic	820	2187	
3	Non-Hispanic White	3150	5337	
4	Non-Hispanic Black	2115	7452	
5	Other Race - Including Multi-Racial	1802	9254	
.	Missing	0	9254	

Figure 1.2: Data dictionary sample from NHNANES 2017-2018 Demographic dataset (©2024 IEEE)

In figure 1.2, we provide a segment of a data dictionary from the Centers for Disease Control and Prevention's National Center for Health Statistics' (NCHS) National Health and Nutrition Examination Survey (NHANES) 2017-2018 Demographics dataset [15] as an example. The figure shows the schema element "RIDRETH1", provides a text description of the element "Recode of reported race and Hispanic origin information," and a codebook that maps races to different numbers within the data. This example highlights the need for data dictionaries. If a third party wanted to reuse this data, it would be challenging to determine the data element from the schema label alone.

1.5 Semantic Data Dictionary

Data dictionaries aid humans in reusing a published dataset, but they typically fall below the standards laid out in the FAIR guiding principles [16], and the wide variety of descriptions can make it difficult for a machine to understand the semantic alignments within the data [17]. To address these issues, previous works have developed Semantic Data Dictionaries (SDDs) [17], which capture context through semantic annotations by aligning tabular data to well-defined terms. This specification allows data publishers to align their schema to domain ontologies and capture explicit and implicit variables within the data. Using a tabular dataset and a semantic data dictionary, users can generate knowledge graph fragments using standard formats such as Resource Description Framework (RDF) to create or augment a knowledge graph [18], [19], [20].

Semantic data dictionaries align each element within the data to properties of several top-level ontologies. These ontologies form the foundation for the resulting model and are used to describe the alignments to domain ontologies. In theory, any ontology can be used as a top-level ontology however, Rashid et al. [18] recommend using well-known ontologies such as Semanticscience Integrated Ontology (SIO) [21] or Basic Formal Ontology (BFO) [22]. In practice, many use the default top-level ontologies introduced by Rashid et al. [18] which include properties from RDF [23], Resource Description Framework Schema (RDFS) [24], SIO, and PROV Ontology (PROV-O) [25]. To limit the scope of this dissertation, we will only consider semantic data dictionaries that use the default top-level ontologies, but the methodologies discussed can be adapted to other top-level ontologies.

SDD alignments assign a domain ontology class to the columns and virtual columns of a table. The virtual columns are the implied entities within the data dictionary and are helpful because in study data, we often find there are multiple implicit subjects with which data is associated.

Table 1.2: A subset of data dictionary descriptions from NHNANES 2017-2018 Demographic dataset

Column	**Description**
SEQN	Respondent sequence number.
RIDAGEYR	Age in years of the participant at the time of screening. Individuals 80 and over are topcoded at 80 years of age.
RIDAGEMN	Age in months of the participant at the time of screening. Reported for persons aged 24 months or younger at the time of exam (or screening if not examined).
RIAGENDR	Gender of the participant.
RIDRETH1	Recode of reported race and Hispanic origin information

We have included a portion of the data dictionary from the National Center for Health Statistics' NHANES 2017-2018 Demographics dataset [15] in table 1.2 and the corresponding semantic data dictionary in tables 1.3, 1.4 to highlight some of the features introduced. This semantic data dictionary uses SIO [21] as its top-level ontology and the Human Aware Science Ontology (HAScO) [26], SIO, the Human Health Exposure Analysis Resource (HHEAR)

Table 1.3: A subset of semantic data dictionary columns for NHNANES 2017-2018 Demographic dataset (©2024 IEEE)

Column	Attribute	attributeOf	Unit	Time
SEQN	hasco:originalID [original ID]	??participant		
RIDAGEYR	sio:SIO_001013 [age]	??participant	sio:SIO_000428 [year]	??screening
RIDAGEMN	sio:SIO_001013 [age]	??participant	sio:SIO_000429 [month]	??screening
RIAGENDR	sio:SIO_010029 [biological sex]	??participant		
RIDRETH1	hhear:00609 [Race or Ethnicity Combined]	??participant		

Table 1.4: A subset of semantic data dictionary virtual columns for NHNANES 2017-2018 Demographic dataset (©2024 IEEE)

Column	Entity	Role	Relation	inRelationTo
??participant	sio:SIO_000485 [human]	sio:SIO_000883 [subject role]		
??screening	nhanes:00020 [Screening time]			??participant

ontology [27], and the NHANES ontology [28] for the domain ontologies. To improve the readability, we have separated the semantic data dictionary into two separate tables and have added brackets with the RDF label to all ontology alignments.

In table 1.3, the first column shows all the data column labels from the corresponding table. It aligns their relationship to the top-level ontology terms Attribute, attributeOf, Unit, and Time. The Attribute column aligns the class label to a domain ontology attribute type. The attributeOf property assigns that attribute to an entity within the data. The Unit column captures the metric used for the attribute. Finally, the Time column captures when the measurement took place. Throughout table 1.3 several virtual columns are referenced with the ?? notation. Table 1.4 shows the virtual column declarations that allow users to assign an Entity, Role, Relation, and inReleationTo alignment to a virtual column.

While semantic data dictionaries are more declarative, and thus often more useful in computational systems, they require additional work. The optimal SDD authors are domain experts who have a deep understanding of the data as well as experience in ontology modeling. The crucial task of aligning concepts to ontology terminology underscores the precision required in this process, often creating a barrier to entry for users. Traditionally, this is overcome by creating a multidisciplinary team or additional training, but this can be a significant investment to align datasets.

1.6 Contributions

In this thesis, we seek to improve the general area of data integration by furthering the semantic tabular interpretation field. We explore algorithms that leverage context information from data dictionaries to perform semantic tabular interpretation. Previous research has focused on data-level alignment algorithms, which align named entities and infer class and relations. While data-level alignment has been effective on simple tables [6], [7], [8], [9], there are still open questions on its effectiveness on complex tables [8]. We hypothesize that semantic tabular interpretation algorithms that leverage context information from data dictionaries will be more effective on complex tables. In the same way that a data dictionary provides context for humans, it will allow algorithms to account for variables and relations not explicitly represented in the table. We answer three research questions about semantic tabular interpretation: *Can alignment algorithms effectively use context from data dictionaries to improve alignment on complex tables? Are alignment algorithms that leverage data dictionaries competitive with data-level alignment algorithms on simple tables? What type of data dictionary descriptions are well-suited for alignment algorithms?*

For the first research question, we developed the Semantic Data Dictionary Generator (SDD-Gen), a tabular alignment algorithm that generates SDDs by leveraging context information from data dictionaries. To evaluate the performance of the SDD-Gen we created a new alignment benchmark for complex tables called Semantic Data Dictionary Dataset (SDD-Dataset) by combining biomedical study data from the National Institute of Environmental Health Sciences' (NIEHS) HHEAR program [29], the National Center for Health Statistics' (HCHS) NHANES survey [30], and the National Cancer Institute's (NCI) The Cancer Genome Atlas (TCGA) [31]. We compare the performance of SDD-Gen versus KGCODE-Tab [32], the state-of-the-art data-level semantic tabular interpretation algorithm from Semtab 2022 [9]. We found that SDD-Gen is twelve times more accurate than

KGCODE-Tab on the SDD-Dataset.

For the second research question, we developed a methodology for generating artificial data dictionaries using large language models. We use this methodology to generate data dictionaries for the BiodivTab dataset [33], a simple table benchmark developed from real-world biodiversity research studies and one of the first non-artificial datasets used for SemTab. We show that the SDD-Gen is as effective as the data-level algorithms on simple tables. We found that SDD-Gen would have placed fourth out of eight algorithms.

For the final research question, we developed an evaluation framework to determine the type of data dictionary description best suited for tabular alignment. We meticulously reviewed data dictionaries from the SDD-Dataset, a comprehensive collection containing over 1,600 data dictionary descriptions from the HHEAR, NHANES, and TCGA studies. This exhaustive review led us to identify and define three common types of data dictionary descriptions: lexical, intensional, and extensional. Lexical descriptions define a column by expanding the column name or acronym and associating the column with a subject. Intensional descriptions define the conditions needed to be a member of that column and associate the column with a subject. Finally, extensional descriptions use codes or data examples to describe the column. Our evaluation framework uses Large Language Models (LLM) to normalize descriptions for the SDD dataset and generates three new datasets consisting of only a single description type. We train three new models only on a single description type and compare their performance to determine which is better suited for tabular alignment. Using our evaluation framework, we show that intensional descriptions are most effective for tabular alignment and improve the reusability of data dictionaries.

1.7 Evaluation

Our primary goal across all experiments was to reuse the experimental procedure established by the SemTab competition. This was to follow established community standards and directly compare the results with the SemTab competition. Across all experiments, CTA was the subtask used to measure performance because it is the same task in SDD creation and SemTab competition. We dig into this issue in greater detail in section 3.2. Standard F1-score was used as the primary evaluation metric for all experiments, and precision was used as the secondary metric in case of a tie.

$$Precision = \frac{correctAnnotations}{submittedAnnotations} \tag{1.1}$$

$$Recall = \frac{correctAnnotations}{groundtruthAnnotations} \tag{1.2}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{1.3}$$

In the first experiment, we wanted to evaluate the effectiveness of data dictionary context versus data-level algorithms on complex tables. To do this, we compared the alignment accuracy of SDD-Gen (data dictionary) against KGCODE-Tab (data-level) on the SDD-Dataset (complex tables) and measured the F1 Score, precision, and recall. In the second experiment, we wanted to evaluate the effectiveness of data dictionary context versus data-level algorithms on simple tables. To do this, we compared the alignment accuracy of SDD-Gen (data dictionary) against SemTab 2022 competitors (data-level) on the BiodivTab dataset (simple tables) and measured the F1 Score, precision, and recall. In the third experiment, we wanted to determine the most effective data dictionary description type. To do this, we took the SDD-Dataset and created three new datasets by normalizing them to single description type. We then trained three new SDD-Gen models on a single description type and measured the F1 Score, precision, recall, and α.

1.8 Motivation

The original motivation for this research project came from the NIEHS HHEAR project [29]. The goal of the project was to combine environmental exposure data across multiple studies, enabling cross-study analytics. However, the main bottleneck was the semantic tabular alignment process, which often required authors to be domain experts with a good understanding of the data and experience in ontology modeling to align concepts properly. Existing algorithms were ineffective because of the lack of named entities within the study data and the complex relations between columns. To reduce the barrier to entry, we developed a web app called the SDD-Editor [34] that leverages suggestions from SDD-Gen to link tabular data to a set of ontologies using data dictionary descriptions [35]. The SDD-Editor provides users with an excel-like interface, which allows them to review suggestions or make manual alignments. Figure A.1 shows a screenshot of the SDD-Editor in action. Although our efforts in the HHEAR project focused on creating semantic data dictionaries to capture these alignments, others have applied these techniques to generate the NHANES knowledge graph [36], which demonstrates that algorithms developed here can be applied to other semantic alignment tasks.

1.9 Dissertation Outline

This dissertation is organized as follows. In this chapter, we introduced the semantic tabular interpretation problem, explained how humans have attempted to address the context gap, and described our approach to solving the problem. In Chapter 2, we present a literature review that explores the history of semantic tabular interpretation, the various types of solutions explored, and the progress made under the SemTab competition. In Chapter 3, we introduce the SDD-Gen's alignment algorithm, explore the creation of the SDD-Dataset a complex table dataset, and dive into our experiments comparing SDD-Gen performance on complex tables. In Chapter 4, we introduce BiodivTab, one of the first SemTab benchmarks created from non-synthetic data, explore how to make artificial data dictionaries using large language models, and take a look at our experiments comparing SDD-Gen performance on simple tables. In Chapter 5, we explore what types of descriptions are best suited for tabular alignment, introduce our evaluation framework, description normalization using large language models, and compare the SDD-Gen performance on different description types. In Chapter 6, we discuss the limitations of our contributions and future research directions for semantic tabular interpretation. Finally, in Chapter 7, we summarize our findings and make final recommendations.