MedTable: Extracting Disease Types from Web Tables

Maria Koutraki and Besnik Fetahu

L3S Research Center, Leibniz University of Hannover, Germany
koutraki@l3s.de fetahu@l3s.de

Abstract. Diseases and their symptoms are a frequent information need for Web users. Diseases often are categorized into sub-types, manifested through different symptoms. Extracting such information from textual corpora is inherently difficult. Yet, this can be easily extracted from semi-structured resources like tables. We propose an approach for *identifying* tables that contain information about *sub-type classifications* and their *attributes*. Often tables have *diverse* and *redundant* schemas, hence, we align *equivalent* columns in disparate schemas s.t. information about diseases are accessible through a *unified* and a *common* schema. Experimental evaluation shows that we can accurately identify tables containing disease sub-type classifications and additionally align equivalent columns.

1 Introduction

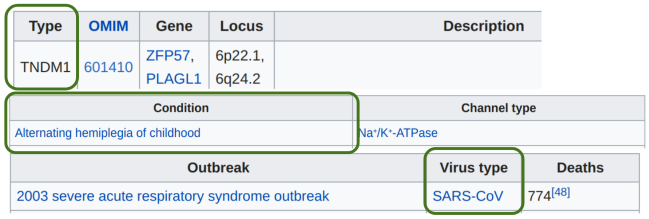
Publicly available medical resources like PubMed[1], or MedQuad [1] (a Q&A dataset about *disease information*), serve as a training ground for Q&A systems with use cases such as symptoms and disease identification [2]. Yet, these repositories are mostly unstructured and require extensive efforts for reasoning over concepts like disease, or different types that diseases or genetic syndromes may have.

On the other hand, for structured resources, like the Disease Ontology (DO)[2] or the classification schema International Classification of Diseases (ICD)[3], accessing information is trivial, however, coverage is limited. DO uses up to seven features to describe a disease (e.g. ID, name, description) while ICD provides only a textual description and a link to the parental disease in the taxonomy.

Recent efforts, focused on harnessing information from Web tables, show that tables are rich in information coverage (e.g. more than 4k medical articles in Wikipedia). Often, tables can be interlinked with each other according to their topic similarity [6], thus, producing even a richer landscape of information that can be extracted from tables.

In this work, our main aim is to harness information from tables containing medical information about *disease sub-type* classifications (e.g. `Arthritis` has two common types `Osteoarthritis` and `Rheumatoid Arthritis`) and their

[1] https://www.ncbi.nlm.nih.gov/pubmed/
[2] https://disease-ontology.org/
[3] https://www.who.int/classifications/icd/en/



Type	OMIM	Gene	Locus	Description
TNDM1	601410	ZFP57, PLAGL1	6p22.1, 6q24.2	

Condition	Channel type
Alternating hemiplegia of childhood	Na^+/K^+-ATPase

Outbreak	Virus type	Deaths
2003 severe acute respiratory syndrome outbreak	SARS-CoV	774[48]

Fig. 1. Examples of three different disease tables with equivalent columns marked in green.

characteristics and *symptoms*, to enrich existing medical corpora like MedQuad, such that Q&A application can provide more faceted answers from the rich table structures, contrary to the short and ambiguous summaries in MedQuad dataset.

To do so, we address two problems. First, from the tables' corpus, TableNet [6], we *identify* tables that contain information about disease *sub-type classifications*, and second, due to diversity of table schemas, we *align* related or equivalent columns. These steps ensure our goal to provide a common schema, which allows for a unified access to all tables containing disease classification related information. In this paper, we make the following contributions:

- an approach for identifying tables about disease types or genetic syndromes;
- an approach for aligning columns that refer to equivalent or related concepts;
- a corpus under a common schema for tables related to disease type classifications.

2 MedTable: Table Identification and Column Alignment

In this section, we present our approach *MedTable* and describe the two main steps for generating the corpus of tables containing disease classification related information.

2.1 Table Identification

Our testbed for tables is the TableNet[6], with more than 3M tables. However, only a small portion of tables is of interest, namely, containing information related to possible (-sub)types of a disease. In the following, we describe the features that we construct for building a supervised machine learning model for classifying tables into either containing *(-sub)types* of diseases or *not*. Since our tables' corpus consists of tables and the corresponding Wikipedia articles from where they are extracted, we consider the following two feature categories.

Article level features. The choice of article features is to consider the context in which a table occurs. This is necessary as some tables are underspecified and the actual information can be interpreted only in conjunction with the article information [4,5]. We consider the Wikipedia *article name*, *section label*, and the *average word representation* of a section's text [11]. Contextual information is necessary as the models learn to distinguish between tables that have similar structures, but topically are highly divergent.

Table related features. Even though context is important, another set of crucial features are extracted from the tables themselves. We additionally consider the *column names* as one of our features. The intuition here is that column names provide crucial hints on the information that the column stores.

2.2 Column Alignment

After having classified tables whether they contain (sub-)type diseases information, the objective here is to *align* columns that are semantically related or equivalent. This is a necessary step, as table schemas across tables are not standardized and often columns with the same information are named differently (cf. Fig.1). Furthermore, column names are ambiguous, and as such a simple lexical match is insufficient.

For that reason we follow a similar approach to those used on schema matching for knowledge graphs [8,9,3]. For a column pair $\langle c_i, c_j \rangle$ from two disparate table schemas, we extract features from the columns, namely the cell values from the respective tables they are extracted, and train a supervised model that classifies them into either *equivalent* or *not*.

We consider the following column features. First, from the *column heading* we construct an average word representation based on GloVe pretrained embeddings [11], correspondingly, we measure the cosine similarity of such representations for the columns c_i and c_j. Second, since column names can be ambiguous, hence, we consider features that are computed based on the column cell values. For columns whose cell values are already interlinked to Wikipedia entities, we consider the average node embedding representation from all instances, by training the graph embeddings based on node2vec [7] on the Wikipedia's anchor graph. That is, for the pair $\langle c_i, c_j \rangle$, we compute the cosine similarity of such representations. Third, for cell values that are simple literals (i.e. numbers, strings etc.), we consider the jaccard similarity of the corresponding values, and in the case of numerical values, we compute the Kullback-Leibler divergence from the corresponding probability distributions of the cell values.

3 Evaluation

The evaluation setup and approach of this work is available for download[4].

3.1 Dataset & Ground Truth

Diseases Dataset. We collect all the Wikidata (WD)[5] instances of class *Disease* (wd:Q12136), resulting in 11k instances. From the resulting subset, we consider only those that have a corresponding article in the English Wikipedia (WP) resulting in 4386 pages, out of those 327 contain tables. We additionally investigated the diseases' ontology from the BioSNAP Datasets [10], resulting in an additional 17 diseases that did not exist in the WD corpus. Finally, our dataset consists of 344 WP disease articles.

[4] https://github.com/koutraki/medtable [5] Accessed 17.04.2019

TableNet – Data. From the 344 WP pages, we extract 764 tables from the TableNet [6], consisting of 5,738 rows in total, with 990 distinct columns.

Ground Truth. We manually constructed the ground-truth for both classification steps in our approach. For the first step, **table classification.**, we annotated all the 764 tables of our dataset, resulting in 190 relevant tables, and the remainder are not related to (sub-)type disease classification. Whereas, for the second step, **column alignment.**, we randomly sampled a set of 350 column pairs from the tables to assess which columns can be aligned, which resulted in 66 aligned column pairs, whereas the remainder of column pairs did not represent equivalent columns.

3.2 Results and Discussion

In this section we discuss the obtained results for both steps in our approach. In both cases, we train a *logistic regression* model based on the described feature sets in Section 2.1 and 2.2. We evaluate the performance of our models based on evaluation metrics such as: *precision* – P, *recall* – R and F1. Evaluation results in Table 1 and 2 correspond to 5-fold cross validation.

Table 1. Table classification results.

	P	R	F1
article features	0.82	0.68	0.74
table features	0.86	0.53	0.66
all	**0.87**	**0.73**	**0.80**

Table 2. Column alignment results.

	P	R	F1
`equiv.`	0.867	0.703	0.78
`non-equiv.`	0.922	0.970	0.95

Table Classification. Table 1 shows the results of the table classification step, and the feature ablation. Note how the two feature sets are complementary, in that, article features provide better coverage, which was our initial intuition as well by capturing contextual information from the articles which describe the diseases listed in a table. On the other hand, table features are more accurate predictors of tables that contain (sub-)type disease information. This is mostly attributed to specific columns that are often to describe disease classifications and describing their symptoms (e.g. *"Type"* the table contains relevant information to (sub-)types of the disease). Jointly, the model is able to achieve high classification performance with an overall score of F1=0.80.

Column Alignment. Table 2 shows the classification results for the column alignment step. Note here that the two classes are highly imbalanced, with the `equiv` class representing only 18% of the dataset. The achieved results are highly satisfactory, reaching a high F1 score of 0.78. This allows us to align columns that are equivalent across disparate table schemas, and thus, offer a unified way to access the disease (sub-)type classifications and their descriptions through a common schema.

4 Conclusions

We present MedTable, an approach for identifying tables about (sub-)types of diseases and correspondingly aligning columns that represent equivalent concepts. We performed an evaluation on the TableNet corpus, where we evaluated on manually constructed ground-truth. We identified nearly 200 relevant tables and were able to align 18% of columns as equivalent. The generated corpus will be made publicly available and can serve for Q&A approaches in the medical domain.

References

1. Abacha, A.B., Demner-Fushman, D.: A question-entailment approach to question answering. arXiv e-prints (January 2019), https://arxiv.org/abs/1901.08079
2. Abacha, A.B., Shivade, C., Demner-Fushman, D.: Overview of the mediqa 2019 shared task on textual inference, question entailment and question answering. In: BioNLP (2019)
3. Biswas, R., Koutraki, M., Sack, H.: Exploiting equivalence to infer type subsumption in linked graphs. In: The Semantic Web: ESWC 2018 Satellite Events - ESWC 2018 Satellite Events, Heraklion, Crete, Greece, June 3-7, 2018, Revised Selected Papers. Lecture Notes in Computer Science (2018)
4. Biswas, R., Koutraki, M., Sack, H.: Predicting wikipedia infobox type information using word embeddings on categories. In: Proceedings of the EKAW 2018 Posters and Demonstrations Session (2018)
5. Biswas, R., Türker, R., Moghaddam, F.B., Koutraki, M., Sack, H.: Wikipedia infobox type prediction using embeddings. In: Proceedings of the First Workshop on Deep Learning for Knowledge Graphs and Semantic Technologies (DL4KGS) (2018)
6. Fetahu, B., Anand, A., Koutraki, M.: Tablenet: An approach for determining fine-grained relations for wikipedia tables. In: The World Wide Web Conference, WWW 2019 (2019)
7. Grover, A., Leskovec, J.: node2vec: Scalable feature learning for networks. In: KDD (2016)
8. Koutraki, M., Preda, N., Vodislav, D.: SOFYA: semantic on-the-fly relation alignment. In: Proceedings of the 19th International Conference on Extending Database Technology, EDBT 2016, Bordeaux, France, March 15-16, 2016, Bordeaux, France, March 15-16, 2016 (2016)
9. Koutraki, M., Preda, N., Vodislav, D.: Online relation alignment for linked datasets. In: The Semantic Web - 14th International Conference, ESWC 2017, Portorož, Slovenia, May 28 - June 1, 2017, Proceedings, Part I. Lecture Notes in Computer Science (2017)
10. Marinka Zitnik, Rok Sosič, S.M., Leskovec, J.: BioSNAP Datasets: Stanford biomedical network dataset collection. http://snap.stanford.edu/biodata (Aug 2018)
11. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: EMNLP (2014)