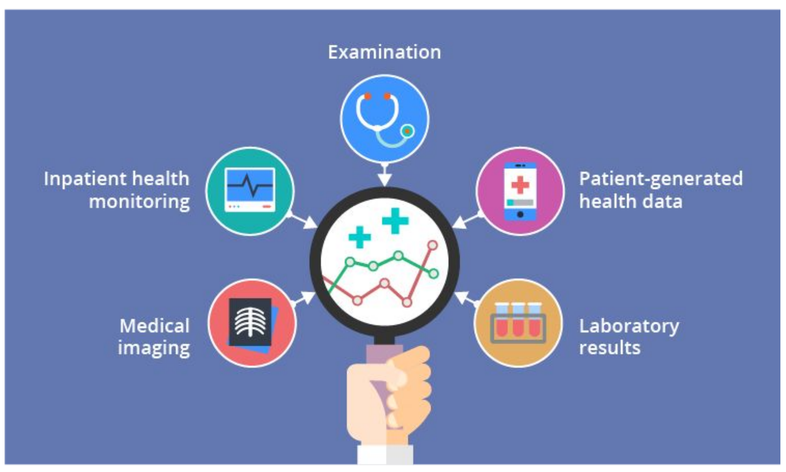
MedTable: Extracting Disease Types from Web Tables

Maria Koutraki and Besnik Fetahu

Motivation

Medical Data



Web Tables

	MERS-CoV	SARS-CoV	SARS-CoV-2
Disease	MERS	SARS	
Outbreaks	2012, 2015, 2018	2002–2004	
Epidemiology			

Feature	Type 1 diabetes	Type 2 diabetes
Onset	Sudden	Gradual
Age at onset	Mostly in children	Mostly in adults
Body size	Thin or normal[39]	Often obese
Ketoacidosis	Common	Rare
Antibodies	Usually present	Absent
Genous insulin	Low or absent	Normal, decreased or increased
Concordance identical twins	50%	90%
Prevalence	~10%	~90%

Nitrogenous base	Ribonucleoside	Deoxyribonucleoside
Adenine	Adenosine A	Deoxyadenosine dA
Guanine	Guanosine G	Deoxyguanosine dG

s (cellulose...)	Polycondensation	Glycosidic bond
cis-1,4-polyisoprene and trans-1,4-utta-percha	Polyaddition	
s, nucleic acids (DNA,		Phosphodiester bond

- Web documents and other medical domain related corpora have an ever increasing amount of medical related topics that form valuable resources for several tasks, e.g. Q&A
- Diseases and their symptoms are a frequent information need for Web users -- categorized into sub-types, manifested through different symptoms
- **But,** most of this information is in **textual form**

- Tables contain rich factual information on diverse topics.
- Tables represent a valuable resource for complex question answering scenarios, or knowledge base population.

Objective: harness information from Wikipedia tables about disease sub-type classifications, their characteristics and symptoms

Task#1: Table Identification

Classifying Wikipedia tables into either **containing (-sub)types** of diseases or **not**

Feature	*Description*
entity	Wikipedia entity page (disease)
sec_level	level of the section containing a table
sec_label	section title
sec_emb	avg. word embeddings of section text
columns	table's column names

	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**

Task#2: Column Alignment

Align columns that are semantically related or equivalent

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]

For a column pair <ci , cj> train a supervised model that classifies them into either **equivalent** or **not**

Feature	*Description*
columns_emb	avg. word embeddings of column heading (GloVe)
value_embl	avg. node embeddings of entity cell values (node2vec)
string_val	Jaccard similarity of string values
numeric_val	Kullback-Leibler divergence

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

MedTable Data & Ground-Truth

- 344 Wikipedia pages disease articles
- 764 tables from TableNet, consisting of 5,738 rows in total, with 990 distinct columns

- Manually constructed ground-truth for both tasks
  Task#1: annotated the 764 tables of our dataset
  → 190 relevant tables
  Task#2: randomly sampled a set of 350 column pairs
  → 66 equivalent

Contact:
Besnik Fetahu
fetahu@L3S.de
@FetahuBesnik

Maria Koutraki
koutraki@L3S.de
@mairyl0u

Forschungszentrum L3S Research Center