# **TAB2KNOW:** BUILDING A KNOWLEDGE BASE FROM TABLES IN SCIENTIFIC PAPERS

### BENNO KRUIT, HONGYU HE, JACOPO URBANI







## TABLES IN SCIENTIFIC PAPERS

- Structured information about scientific process
  - similar structure across documents
- Could support reviews or search
- Examples of tables for **human readers**

How do we automatically process tables that were not designed for automatic processing?

TAB2KNOW: BUILDING A KNOWLEDGE BASE FROM TABLES IN SCIENTIFIC PAPERS

## PROBLEMS

- Tables must be **extracted** from PDFs
  - reconstruct from PDF text-boxes!
- Every author uses different **conventions** 
  - e.g. structure, jargon, layout, formats
- No **Knowledge Base** to link concepts
- Interpretation is goal-specific
  - Both construction and querying must be user-oriented and flexible

TABLE I. RANKING OF SUBMITTED METHODS TO TA

Method Name	Recall (%)	Precision (%)	F-score
USTB_TexStar	82.38	93.83	87.74
TH-TextLoc	75.85	86.82	80.96
I2R_NUS_FAR	71.42	84.17	77.27
Baseline	69.21	84.94	76.27
Text Detection [15], [16]	73.18	78.62	75.81
I2R_NUS	67.52	85.19	75.34
BDTD_CASIA	67.05	78.98	72.53
OTCYMIST [7]	74.85	67.69	71.09
Inkam	52.21	58.12	55.00

## TAB2KNOW

- A system for constructing and querying a Knowledge Graph of information extracted from tables in scientific papers
  - 1. Structural foundation: simple graph of extracted structure
  - 2. Semantic layer: predicted types of tables and columns
  - 3. Entity layer: similar cells resolved to entity clusters
- Based on user-written rules and queries
  - used as weak supervision for machine learning models

TAB2KNOW: BUILDING A KNOWLEDGE BASE FROM TAB2ES IN SCIENTIFIC PAPERS

WEAK SUPERVISION



- Hand-labeling data is expensive and inflexible
  - ... so write **labeling functions** instead!
- Snorkel (Ratner et al, VLDB2020)
  - DryBell @ Google (Bach et al., SIGMOD 2019)
  - Overton @ Apple (Christopher Ré et al., ArXiv 2019)
- Aggregate weak signals for training ML models
  - exploit labeling function correlations and sparsity

TAB2KNOW: BUILDING A KNOWLEDGE BASE FROM TAB

## WEAK SUPERVISION





## **EXAMPLE: TABLE TYPES**

$l_1$ - $l_2$	#S	$\#l_1$ -W	$\#l_2$ -W	$#l_1-V$	$#l_2$ -V
en-de	1.9M	55M	52M	40k	50k
en-fr	2.0M	50M	51M	40k	50k
en-es	1.9M	49M	51M	40k	50k

(a) Input

Туре	Example Words					
Offensive	disgusting, filthy, nasty,					
	rude, horrible, terrible, aw-					
	ful, worst, idiotic, stupid,					
	dumb, ugly, etc.					
Non-offensive	help, love, respect, believe,					
	congrats, hi, like, great,					
	fun, nice, neat, happy,					
	good, best, etc.					

Models	Rerank size	Beam size	GMV	Latency
miDNN	50	-	2.91%	9%
miRNN	50	5	5.03%	58%
miRNN+att.	50	5	5.82%	401%

(b) Observation

$lpha_c$	DP concentration parameter for each $c \in V$
$P_0(e c)$	CFG base distribution
$oldsymbol{x}$	Set of non-terminal nodes in the treebank
S	Set of sampling sites (one for each $x \in \boldsymbol{x}$ )
S	A block of sampling sites, where $S \subseteq \mathcal{S}$
$\boldsymbol{b} = \{b_s\}_{s \in \mathcal{S}}$	Binary variables to be sampled ( $b_s = 1 \rightarrow$
	frontier node)
z	Latent state of the segmented treebank
m	Number of sites $s \in S$ s.t. $b_S = 1$
$oldsymbol{n} = \{n_{c,e}\}$	Sufficient statistics of $\boldsymbol{z}$
$\Delta n^{S:m}$	Change in counts by setting $m$ sites in $S$

(d) Other

#### (c) Example

gs of the Twenty-Seventh International Join Figure Aenthe or anthe letensy ingense with respect to rerank size.

Figure 6: The search latency increase w

s of words2and the corresponding C	A KNOV&LED#SE B#GEVFF#GANTABLES #M-SCIENTI	FICMOdeBEF	Rerank size
ngual contexts. Law c is defined in	cost of automodely we compare the pulline for the later of	miDNN	50
word embeddings can be learned	differentantodel2.0 The lasonary of the haselines 21 son And	miRNN	50
ncy-based joint learning objective	the relatives laten on increase of our models over the source line	miRNN+att.	50
	are shown in Figure 4		

#### el regularization

he DepBiWE model and enhance d embeddings by making full use ation in the parallel corpus with a n in terms of phrase-level semantic

-tree of the parallel sentence pairs, phrase p as a word pair  $(w, w_{dr^{-1}})$ , f w and  $dr^{-1}$  denotes their inverse e representation of a dependencyted as the sum of two word vec--1. By incorporating the phrase-, we encourage the representations ases to be close, as we can derive hrases from the aligned words in For example, (*review*, *word*) and igure 1 are aligned as d

ency-phrase pairs are i select loser the embeddings f e pushed together. By 1 ed dependency-phrases ation term can be writt

$$\sum_{p_j^{l_2})\in D_p} ||\mathbf{p}_i^{l_1} - \mathbf{p}_j^{l_2}||^2,$$

and dependency-phras orpus. The regularizat  $\boldsymbol{\chi}$ objective in Equation ( gs (DepBiWE+R), where  $\gamma_R$  is a rol the contribution of the phrase-

s called RepBIWE. LE: IABLE Table iguine iguic -. epropraces inspectata: ertailes But the barene seofenci RDANs, and #IRWN openation to the new prophokanal p. two menallolaan rates in is 399, ase lateries of mirenne variabilities size reases 400% over the baseline, from 21 ms to 105 ms. Although the RNN mod-Lype 1 CNN the RNN modelanactive sorresponding computational cost of the granter of the mealele arst huga weben we can be it a start of the big for the big of the bi tational cost is the major drawback of our RNN models d, For RNN models, we use beam and first figue, find a good rank-

inghset loanotendike beten orizet wo belefterenetarespectee to the ino f scheckanleargeget: woransulaveleilesegnetrationatic artitespoundation the sBtoWncbettextranklithe 10EPts.0Bufklargeheetalgetleadtepespecctimetutational const LWe startied and software for laterespondcingstowBthWebpasetItantoeatepsizeenAyababedrasilingana shoobedind Figureations light the spectral set of the transformation of the set of the transformation of the set of geowhed Bingshand Mike iBoikabesseetsesminddetingsete beend fizen

#### distinct ?table where { SERVICE bds:search {

?matchedValue bds:search "example"

#### ?table dct:title ?matchedValue

corung to men term nequencies in the training corpus. The words with low frequencies for all languages are mapped to

#### Table 2: The GMV increase

Beam s

5

5

models is 5. Results in Table 2 sh fluence aware ranking framework bi increase over the baselbase distribution good GMV increase with only a littl miRNN+attention Stible amplingheite latency grows too fast. The same ing good GMV shorease with much le miRNN+attention. frantierenode,) if con expensive, the miDNN model is a g where mild latency increase is accept is  $p_{e}^{n}$  Sufficient statistics of is preferred.<sup>*n*</sup>  $\Delta n^{S:m}$ Change in counts by

#### Conclusion

Table 5: DP-TSG model notation. In this paper, we point out the importance between items in e-commerce ind as such  $z \equiv \langle c \rangle$ t time. We incorr intersection p ince generation p ique start symbol on a large e-comp achotsewfragme ofstheceaselineyra oduces a signific

tion treechanism for RNN segurenge

Dut the computational

refore verifies th r.nodes are *inter* items. We also hisedosestinger Al .hout seycheotanju

## **EXAMPLE: TABLE TYPES**



Labeling functions (M functions)

- Two options for aggregating labels:
  - Majority Vote
  - Snorkel Model
- Many options for ML model, but must not overfit!

## **ENTITY RESOLUTION**

- Vlog: Large-scale reasoning on contexts of cell values
  - e.g. column header, column type, author, ...
  - If similar, merge cell values into entity clusters

 $ceNoTypLabel(X,L), ceNoTypLabel(Y,L) \rightarrow X \approx Y$  $eNoTypLabel(X,C,L), eNoTypLabel(Y,C,L) \rightarrow X \approx Y$  $eTableLabel(X,T,L), eTableLabel(Y,T,L) \rightarrow X \approx Y$  $eTypLabel(X,S,L), eTypLabel(Y,S,M), STR\_EQ(L,M) \rightarrow X \approx Y$  $eAuthLabel(X,A,L), eAuthLabel(Y,A,M), STR\_EQ(L,M) \rightarrow X \approx Y$ 

## **SUMMARY**

TABLE I.

Method Name	Recall (%)	Precision (%)	F-score	
USTB_TexStar	82.38	93.83	87.74	
TH-TextLoc	75.85	86.82	80.96	
I2R_NUS_FAR	71.42	84.17	77.27	N
Baseline	69.21	84.94	76.27	
Text Detection [15], [16]	73.18	78.62	75.81	
I2R_NUS	67.52	85.19	75.34	
BDTD_CASIA	67.05	78.98	72.53	(1) Table Extrac
OTCYMIST [7]	74.85	67.69	71.09	
Inkam	52.21	58.12	55.00	

**RANKING OF SUBMITTED METHODS TO TASK 1.1** 

Method Name	Recall (%)	Precision (%)	F-score
USTB_TexStar	82.38	93.83	87.74
TH-TextLoc	75.85	86.82	80.96
I2R_NUS_FAR	71.42	84.17	77.27
Baseline	69.21	84.94	76.27
Text Detection [15], [16]	73.18	78.62	75.81
I2R_NUS	67.52	85.19	75.34
BDTD_CASIA	67.05	78.98	72.53
OTCYMIST [7]	74.85	67.69	71.09
Inkam	52.21	58.12	55.00

#### **Input: PDF Figure**



#### **SPARQL** Queries



Rule 1

Rule 2 Rule 3 . . .





Naïve KB



(3) Entity Linking

#### Table type classification

#### Header detection

Method Name	Recall (%)	Precision (%)	F-score
USTB_TexStar	82.38	93.83	87.74
TH-TextLoc	75.85	86.82	80.96
I2R_NUS_FAR	71.42	84.17	77.27
Baseline	69.21	84.94	76.27
Text Detection [15],[16]	73.18	78.62	75.81
I2R_NUS	67.52	85.19	75.34
BDTD_CASIA	67.05	78.98	72.53
OTCYMIST [7]	74.85	67.69	71.09
Inkam	52.21	58.12	55.00

Column type classification



#### **Output: KB (with linked entities)**

#### **Assets**

**Rules** 

## **RESULTS: TABLE TYPES**

- Gold standard: 400 sampled tables, manual annotation
- 4 table types, 39 label queries
- Features from table caption, header cells and body cells

Model	Prec.	Recall	F1	AUC
SVM	0.71	0.79	0.74	0.86
$\mathrm{LR}$	0.72	0.79	0.74	0.84
NB	0.80	<b>0.82</b>	0.79	0.91

ML model performance on Tab2Know data

## **RESULTS: COLUMN TYPES**

on Tab2Know data

- 22 column types
- **55** label queries

Task	MV	Snorkel
Table Types	0.50	0.71
Column Types (Our corpus)	0.56	0.49
Column Types (Tablepedia)	0.39	0.65

### Accuracy of label aggregation

Model	Prec.	Recall	F1		Model	Prec.	Recall	F1
NB	0.52	0.48	0.47	-	Yu et al. [31]	0.82	0.81	0.81
SVM	0.58	0.56	0.53		NB	0.84	0.82	0.81
$\mathbf{LR}$	0.58	0.56	0.53		$\operatorname{SVM}$	0.90	0.89	0.89
					$\mathrm{LR}$	0.92	0.91	0.91
ML model performance					ML mod	del pe	rforma	nce

on Tablepedia data

## **RESULTS: ENTITY RESOLUTION**

- 3 entity creation rules + 4 entity merging rules
- 65% entities are sensible, 97% mergers are good



Number of entities per rule

Examples

6

## **RESULTS: KNOWLEDGE GRAPH**

- 143k PDFs from Semantic Scholar
  - 73k tables extracted
  - **23**M links in graph
- Demo