Towards Question Answering on Irregular Tabular Data Using a Parallel Document Corpus

Raman Chandrasekar
Kelvin Christian

r.chandrasekar@northeastern.edu
Institute for Experiential AI, Northeastern University

Task Focused IR in the Era of Generative AI *Sep 28, 2023*

link

Overview

This is very much **Work In Progress**

- About problem, data collection, approach

Impetus for this work I find exciting,
a real, grounded task

- Table Irregularities
- Related Work
- Broad description/Plan of work
- Share More?

Impetus: QA over Text + Tabular Pharma Info

Years ago, consulted for a pharma company

Task: Paragraph retrieval from corpus of New Drug Application docs submitted to FDA.

[+ Nice-to-have: retrieval from tables]

- Each doc: PDF, more than 300 pages each, with text, figures and tables
- Scanned PDF, not "born-digital" PDF
- Tables:
  - Mostly complex
  - Captions & footnotes often attached
  - Several tables multi-page, or split across pages
  - Some in landscape format

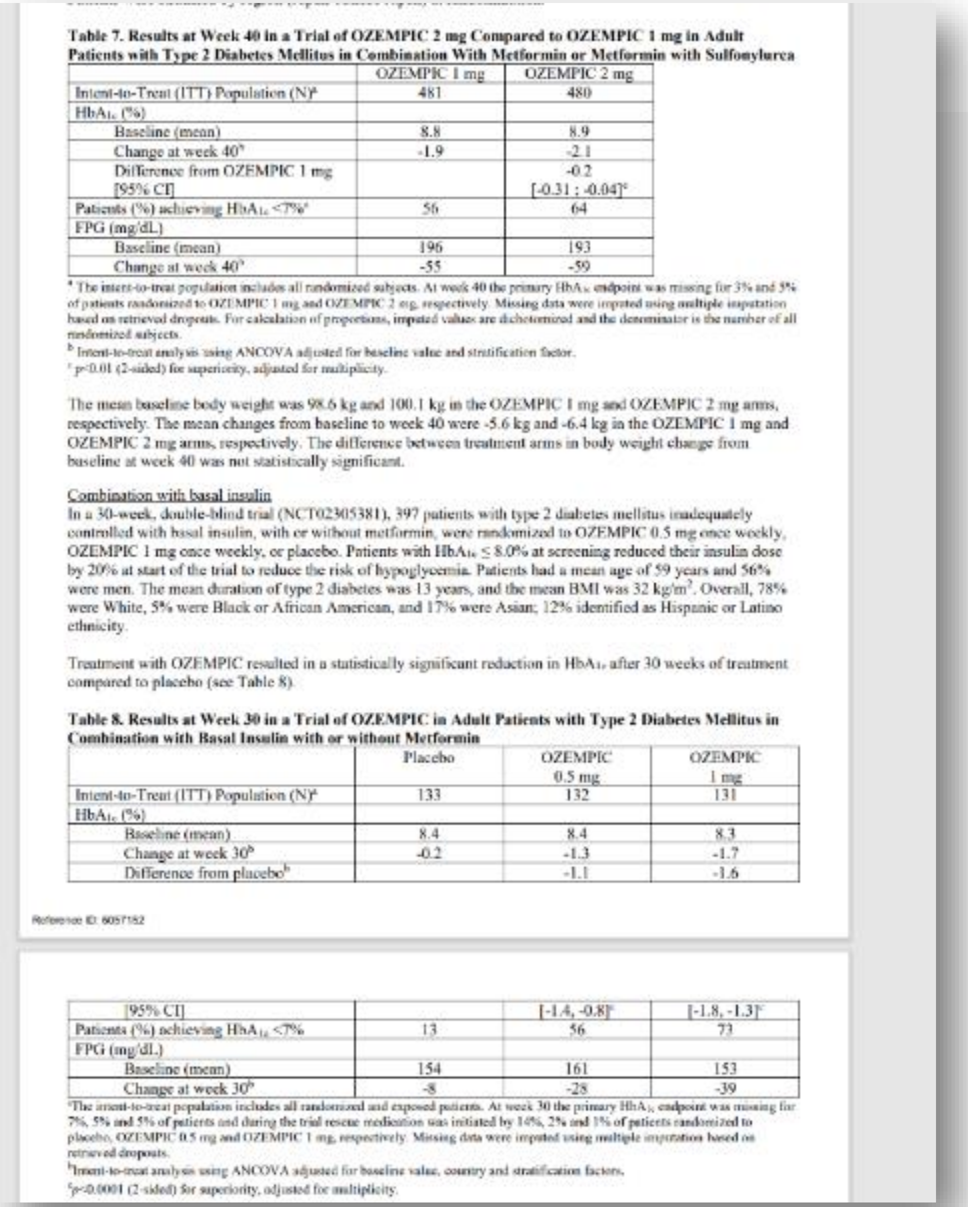


Table 7. Results at Week 40 in a Trial of OZEMPIC 2 mg Compared to OZEMPIC 1 mg in Adult Patients with Type 2 Diabetes Mellitus in Combination With Metformin or Metformin with Sulfonylurea

	OZEMPIC 1 mg	OZEMPIC 2 mg
Intent-to-Treat (ITT) Population (N)[a]	481	480
HbA_{1c} (%)		
Baseline (mean)	8.8	8.9
Change at week 40[b]	-1.9	-2.1
Difference from OZEMPIC 1 mg [95% CI]		-0.2 [-0.31 ; -0.04][c]
Patients (%) achieving HbA_{1c} <7%[a]	56	64
FPG (mg/dL)		
Baseline (mean)	196	193
Change at week 40[b]	-55	-59

[a] The intent-to-treat population includes all randomized subjects. At week 40 the primary HbA_{1c} endpoint was missing for 3% and 5% of patients randomized to OZEMPIC 1 mg and OZEMPIC 2 mg, respectively. Missing data were imputed using multiple imputation based on retrieved dropouts. For calculation of proportions, imputed values are dichotomized and the denominator is the number of all randomized subjects.
[b] Intent-to-treat analysis using ANCOVA adjusted for baseline value and stratification factor.
[c] $p<0.01$ (2-sided) for superiority, adjusted for multiplicity.

The mean baseline body weight was 98.6 kg and 100.1 kg in the OZEMPIC 1 mg and OZEMPIC 2 mg arms, respectively. The mean changes from baseline to week 40 were -5.6 kg and -6.4 kg in the OZEMPIC 1 mg and OZEMPIC 2 mg arms, respectively. The difference between treatment arms in body weight change from baseline at week 40 was not statistically significant.

Combination with basal insulin

In a 30-week, double-blind trial (NCT02305381), 397 patients with type 2 diabetes mellitus inadequately controlled with basal insulin, with or without metformin, were randomized to OZEMPIC 0.5 mg once weekly, OZEMPIC 1 mg once weekly, or placebo. Patients with $HbA_{1c} \leq 8.0\%$ at screening reduced their insulin dose by 20% at start of the trial to reduce the risk of hypoglycemia. Patients had a mean age of 59 years and 56% were men. The mean duration of type 2 diabetes was 13 years, and the mean BMI was 32 kg/m^2. Overall, 78% were White, 5% were Black or African American, and 17% were Asian; 12% identified as Hispanic or Latino ethnicity.

Treatment with OZEMPIC resulted in a statistically significant reduction in HbA_{1c} after 30 weeks of treatment compared to placebo (see Table 8).

Table 8. Results at Week 30 in a Trial of OZEMPIC in Adult Patients with Type 2 Diabetes Mellitus in Combination with Basal Insulin with or without Metformin

	Placebo	OZEMPIC 0.5 mg	OZEMPIC 1 mg
Intent-to-Treat (ITT) Population (N)[a]	133	132	131
HbA_{1c} (%)			
Baseline (mean)	8.4	8.4	8.3
Change at week 30[b]	-0.2	-1.3	-1.7
Difference from placebo[b]		-1.1	-1.6
[95% CI]		[-1.4, -0.8][c]	[-1.8, -1.3][c]
Patients (%) achieving HbA_{1c} <7%	13	56	73
FPG (mg/dL)			
Baseline (mean)	154	161	153
Change at week 30[b]	-8	-28	-39

Reference ID: 6057152

[a] The intent-to-treat population includes all randomized and exposed patients. At week 30 the primary HbA_{1c} endpoint was missing for 7%, 5% and 5% of patients and during the trial rescue medication was initiated by 14%, 2% and 1% of patients randomized to placebo, OZEMPIC 0.5 mg and OZEMPIC 1 mg, respectively. Missing data were imputed using multiple imputation based on retrieved dropouts.
[b] Intent-to-treat analysis using ANCOVA adjusted for baseline value, country and stratification factors.
[c] $p<0.0001$ (2-sided) for superiority, adjusted for multiplicity.

Ozempic (R) Novo Nordisk package insert

What was achieved

[Focusing on tabular info]

- Used multiple hand-crafted tools and heuristics to extract info from tables
- Reasonable results but not perfect, for a variety of reasons

- **Can we do better?**
- With Generative AI, specifically?

Goal now

Goal: Question answering (QA)
on irregular tabular data

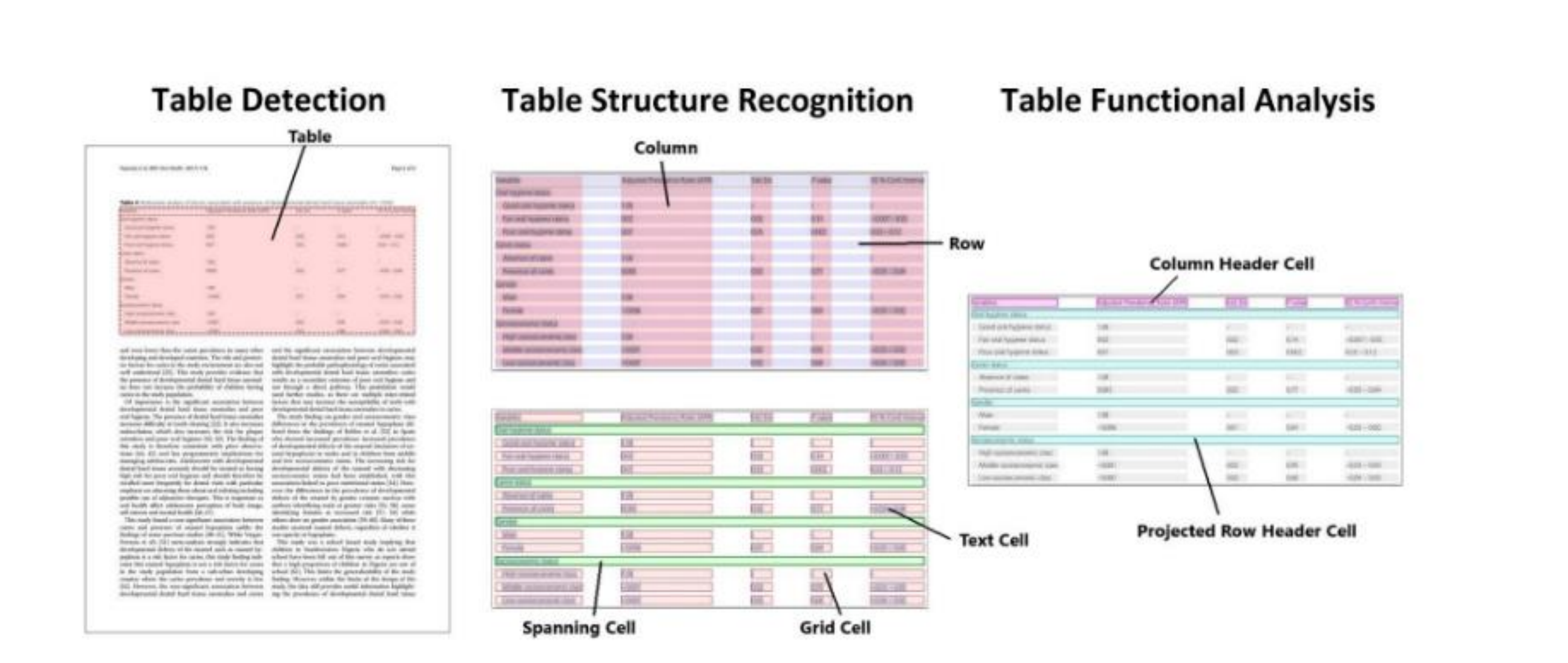
- Lot of work on QA on data in text + regular tables, HTML tables
- Irregular tables common
  - Especially in scientific, medical, and pharma documents
  - Have not received as much attention

Proposal: create a large corpus of tables,
extract & represent canonical table structure,
learn mapping from table to structure,
develop QA on info in these (irregular) tables

Table Extraction Tasks → QA

[https://arxiv.org/pdf/2110.00061.pdf /](https://arxiv.org/pdf/2110.00061.pdf) PubTables-1M

PDF document



QA etc. tasks

Stretch Goal: Document Representation

In addition,
an Interlingua for document representation

allows us to deal with many other issues.

Consider cut-and-paste:

- Google Sheets □ PowerPoint
- PowerPoint equations
  □ on Mac, becomes □ on PC …

- Can we cut and paste across applications?
- If not perfect, close enough with some edits?

Table Irregularities

Headers spanning multiple columns (or rows)

Table 2. Hypoglycemia Adverse Reactions in Placebo-Controlled Trials in Patients with Type 2 Diabetes Mellitus

	Placebo	OZEMPIC 0.5 mg	OZEMPIC 1 mg
Monotherapy			
(30 weeks)	**N=129**	**N=127**	**N=130**
Severe†	0%	0%	0%
Documented symptomatic (≤70 mg/dL glucose threshold)	0%	1.6%	3.8%
Severe† or Blood Glucose Confirmed Symptomatic (≤56 mg/dL glucose threshold)	1.6%	0%	0%
Add-on to Basal Insulin with or without Metformin			
(30 weeks)	**N=132**	**N=132**	**N=131**
Severe†	0%	0%	1.5%
Documented symptomatic (≤70 mg/dL glucose threshold)	15.2%	16.7%	29.8%
Severe† or Blood Glucose Confirmed Symptomatic (≤56 mg/dL glucose threshold)	5.3%	8.3%	10.7%

† "Severe" hypoglycemia adverse reactions are episodes requiring the assistance of another person.

Ozempic (R) Novo Nordisk package insert

Irregular Cells

- Irregular columns or rows, leading to irregular cells
  - multicolumn/multirow
- Color used here to set apart cells. Not our focus.

Formed element	Major subtypes	Numbers present per microliter (μL) and mean (range)	Appearance in a standard blood smear	Summary of functions	Comments
Erythrocytes (red blood cells)		5.2 million (4.4–6.0 million)	Flattened biconcave disk; no nucleus; pale red color	Transport oxygen and some carbon dioxide between tissues and lungs	Lifespan of approximately 120 days
Leukocytes (white blood cells)		7000 (5000–10,000)	Obvious dark-staining nucleus	All function in body defenses	Exit capillaries and move into tissues; lifespan of usually a few hours or days
	Granulocytes including neutrophils, eosinophils, and basophils	4360 (1800–9950)	Abundant granules in cytoplasm; nucleus normally lobed	Nonspecific (innate) resistance to disease	Classified according to membrane-bound granules in cytoplasm
	Neutrophils	4150 (1800–7300)	Nuclear lobes increase with age; pale lilac granules	Phagocytic; particularly effective against bacteria. Release cytotoxic chemicals from granules	Most common leukocyte; lifespan of minutes to days
	Eosinophils	165 (0–700)	Nucleus generally two-lobed; bright red-orange granules	Phagocytic cells; particularly effective with antigen- antibody complexes. Release antihistamines. Increase in allergies and parasitic infections	Lifespan of minutes to days
	Basophils	44 (0–150)	Nucleus generally two-lobed but difficult to see due to presence of heavy, dense, dark purple granules	Promotes inflammation	Least common leukocyte; lifespan unknown
	Agranulocytes including lymphocytes and monocytes	2640 (1700–4950)	Lack abundant granules in cytoplasm; have a simple-shaped nucleus that may be indented	Body defenses	Group consists of two major cell types from different lineages
	Lymphocytes	2185 (1500–4000)	Spherical cells with a single often large nucleus occupying much of the cell's volume; stains purple; seen in large (natural killer cells) and small (B and T cells) variants	Primarily specific (adaptive) immunity: T cells directly attack other cells (cellular immunity); B cells release antibodies (humoral immunity); natural killer cells are similar to T cells but nonspecific	Initial cells originate in bone marrow, but secondary production occurs in lymphatic tissue; several distinct subtypes; memory cells form after exposure to a pathogen and rapidly increase responses to subsequent exposure; lifespan of many years
	Monocytes	455 (200–950)	Largest leukocyte with an indented or horseshoe-shaped nucleus	Very effective phagocytic cells engulfing pathogens or worn out cells; also serve as antigen-presenting cells (APCs) for other components of the immune system	Produced in red bone marrow; referred to as macrophages after leaving circulation
Platelets		350,000 (150,000–500,000)	Cellular fragments surrounded by a plasma membrane and containing granules; purple stain	Hemostasis plus release growth factors for repair and healing of tissue	Formed from megakaryocytes that remain in the red bone marrow and shed platelets into circulation

Nested Tables

Common, especially in business, scientific, medical, pharma documents

Nested Tables

<table>
<tr><th>Header column 1</th><th>Header column 2</th><th>Header column 3</th><th>Header column 4</th></tr>
<tr><td>Row 2 - Item 1</td><td>Row 2 - Item 2</td><td rowspan="2">Row 2: Nested Table 1<table><tr><th>Row 1 Header</th><td>item</td></tr><tr><th>Row 2 Header</th><td>item</td></tr><tr><th>Row 3 Header</th><td>item</td></tr></table></td><td>Row 2 - Item 4
A second line</td></tr>
<tr><td>Row 3: Nested Table 2<table><tr><th>Row 1 Header</th><td>item</td></tr><tr><th>Row 2 Header</th><td>item</td></tr></table></td><td>Row 3 - Item 2</td><td rowspan="2">Row 3 - Item 3</td></tr>
<tr><td>Row 4 - Item 1</td><td>Row 4 - Item 2</td><td>Row 4 - Item 3</td></tr>
<tr><td colspan="4">Row 5 - Last row of outer table</td></tr>
</table>

Table 10. Recommended Storage Conditions for the OZEMPIC Pen

Prior to first use	After first use	
Refrigerated 36°F to 46°F (2°C to 8°C)	Room Temperature 59°F to 86°F (15°C to 30°C)	Refrigerated 36°F to 46°F (2°C to 8°C)
Until expiration date	56 days	

Ozempic (R) Novo Nordisk package insert

Captions & footnotes attached to tables

Table 3. Results at Week 30 in a Trial of OZEMPIC as Monotherapy in Adult Patients with Type 2 Diabetes Mellitus Inadequately Controlled with Diet and Exercise

	Placebo	OZEMPIC 0.5 mg	OZEMPIC 1 mg
Intent-to-Treat (ITT) Population (N)[a]	129	128	130
HbA_{1c} (%)			
Baseline (mean)	8.0	8.1	8.1
Change at week 30[b]	-0.1	-1.4	-1.6
Difference from placebo[b] [95% CI]		-1.2 [-1.5, -0.9][c]	-1.4 [-1.7, -1.1][c]
Patients (%) achieving HbA_{1c} <7%	28	73	70
FPG (mg/dL)			
Baseline (mean)	174	174	179
Change at week 30[b]	-15	-41	-44

[a]The intent-to-treat population includes all randomized and exposed patients. At week 30 the primary HbA_{1c} endpoint was missing for 10%, 7% and 7% of patients and during the trial rescue medication was initiated by 20%, 5% and 4% of patients randomized to placebo, OZEMPIC 0.5 mg and OZEMPIC 1 mg, respectively. Missing data were imputed using multiple imputation based on retrieved dropouts.

[b]Intent-to-treat analysis using ANCOVA adjusted for baseline value and country.

[c]p<0.0001 (2-sided) for superiority, adjusted for multiplicity.

Ozempic (R) Novo Nordisk package insert

Other Variants in Tables

- Tables with and without lines dividing rows and columns
- Tables that have images in some cells (check marks, product images)
- Landscape orientation vs. Portrait orientation
- Text in non-horizontal directions
- Tables split across pages

Related Work

TableBank

- Minghao L et al
  (Microsoft Research Asia ++), 2020
- 417K tables from ArXiv,
  Word and LaTeX markup.
  145K tables with annotations.
- Focus: table detection (TD) &
  table structure recognition (TSR)
  - TD: Used image methods Faster R-CNN with ResNeXt. Tested on 8K images. Best results 98+%
  - TSR: Used image-to-text method in OpenNMT. Metrics: 4-gram BLEU. Results almost 70-78%

Highlighted portions suggest scope to improve perf on tables

Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, Ming Zhou, and Zhoujun Li. Tablebank: Table benchmark for image-based table detection and recognition.
In Proc 12th LREC, pp. 1918–1925, 2020.
https://arxiv.org/pdf/1903.01949.pdf

PubTables-1M

- Smock et al, Microsoft Redmond, Nov 2021
- PDF and XML files (with semantic description, hierarchical doc organization) PubMed Central Open Access (PMCOA)
- Table content and structure as HTML tags
- Detection Transformer (DETR) applied to table detection (TD), structure recognition (TSR), and functional analysis(FA)
- 948K tables for TSR,
  53% with at least one spanning cell,
  multi-page tables not considered
- Perf overall on TD 96.6% AP,
  TSR+FA 91.2%.
  DETR perf on TSR accuracy was ~69.4% on complex tables, overall 81.4%

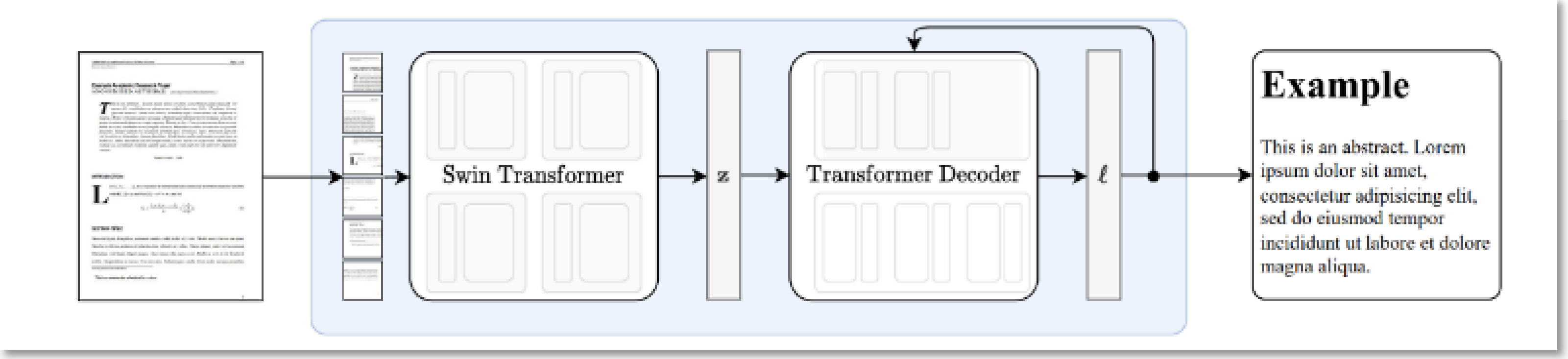
		ΔSDM			
		better	equal	Worse	Sum
ΔSCA	better	19457 (28.9)	12 (0.02)	14654 (21.8)	**34,123 (50.8)**
	equal	1158 (1.7)	21989 (32.7)	1024 (1.5)	**24,171 (36.0)**
	worse	3755 (5.6)	2 (0.003)	5183 (7.7)	**8,940 (13.2)**
	Sum	**24370 (36.2)**	**22003 (32.7)**	**20861 (31.0)**	**67,234 (100.0)**

Smock, Brandon, Pesala, Rohith and Abraham, Robin. "PubTables-1M: Towards comprehensive table extraction from unstructured documents." *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022): 4624-4632. https://arxiv.org/abs/2110.00061

Nougat

- Blecher et al, Meta AI,
  25 Aug 2023
- Neural Optical Understanding for Academic Documents
- arXiv++: PDF and
  LaTeX→ HTML → MD files
- Convert text + tables + equations
  to a version of markdown
  Focus on pages + equations
- Nougat-small (250M)
  Tables: F1   77.3
  Text+Tables+Equations: F1   92.9

  Nougat-base:
  Tables: F1   78.0
  Text+Tables+ Equations : F1   93.1

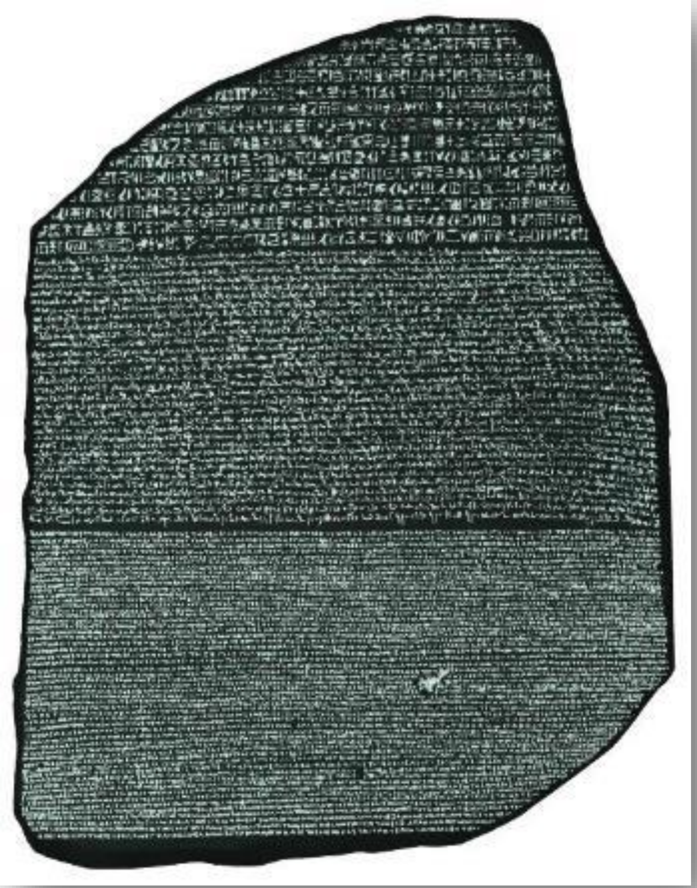


Blecher, Lukas, Guillem Cucurull, Thomas Scialom and Robert Stojnic. "Nougat: Neural Optical Understanding for Academic Documents." *ArXiv* abs/2308.13418 (2023)
https://arxiv.org/abs/2308.13418

Work in progress

Parallel Corpus Methods?

- First plan: Create parallel corpus of arXiv papers, LaTeX ⬄ PDF, over 2 million papers
- Content critical, exact layout not as important
- Treat as a translation task,
  - Extract tables
  - Align LaTeX tables to PDF tables
  - Learn mapping PDF ⬄ LaTeX



Rosetta Stone, 196BC
3 versions of decree in Egyptian-hieroglyphic, Egyptian-Demotic, Ancient Greek

Creating Parallel Corpus: LaTeX files

Extract all tables from LaTeX files

- Multirow, multicolumn, nested tables
- Dealing with LaTeX macros

arXiv Source files

- Size: 2.93 TB
- 2,080,363 files + related (images, style files)
- LaTeX files with at least 1 table: 1,004,368 (~48%)

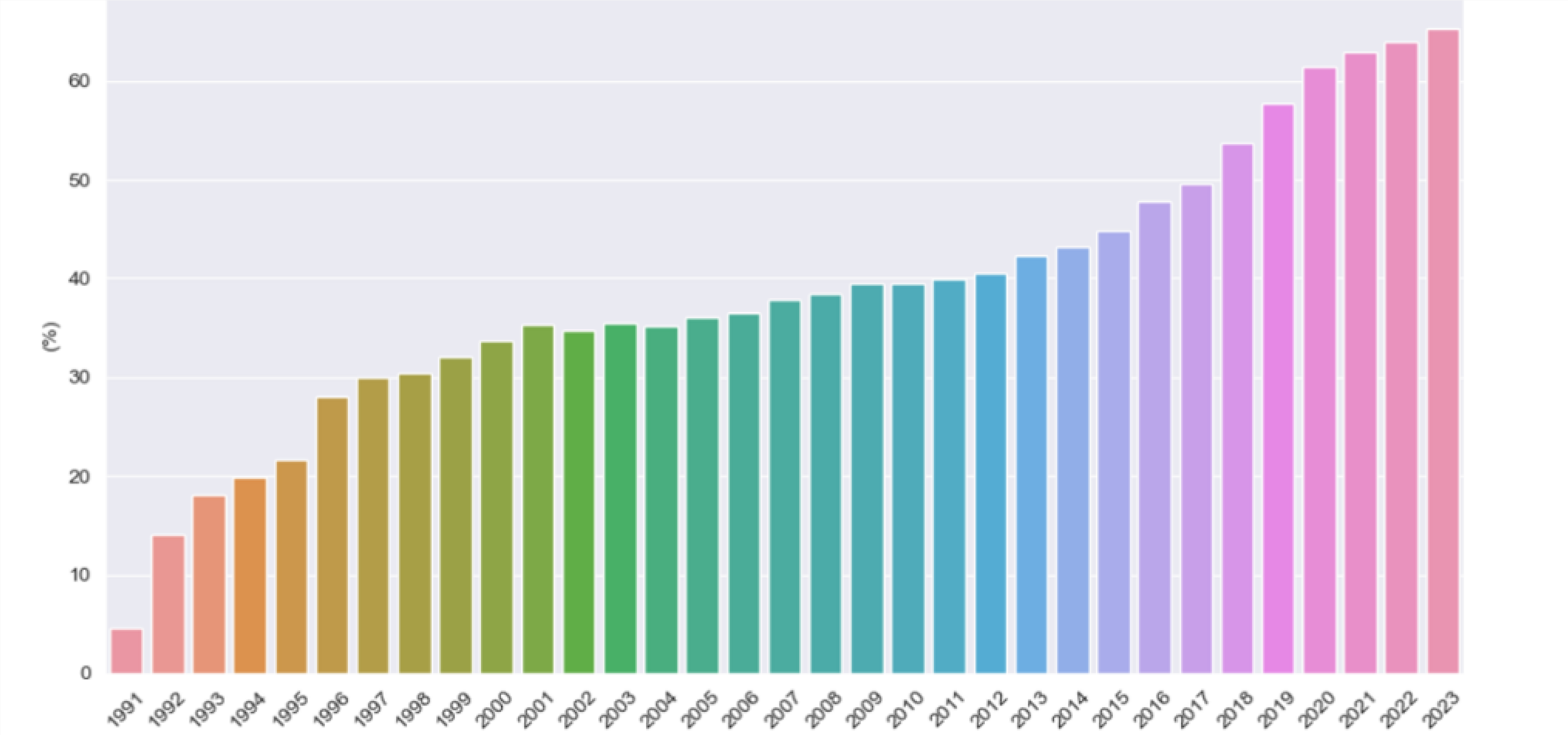
Creating Parallel Corpus: PDF files

- Detect, extract tables from PDFs using Table-transformer
- Confidence threshold: 0.99 to minimize False Positives
  - Heuristics to deal with margin errors
- Extract table content from each image using Tesseract OCR
- Extract text content from the PDF (for later use) using PyMuPDF

arXiv PDF files

- Size: 2.84 TB
- 2,253,795 files + related (images, style files)
- Processed so far: 442,970 (~20%)

Percentage of LaTeX files with tables



Alignment Example

Image from PDF

volume	$8^3 \times 4$		$12^3 \times 4$	$16^3 \times 4$		
β	5.75	5.85	5.75	5.71	5.75	5.85
$\langle n \rangle / V$	0.32	0.06	0.28	0.63	0.30	0.05
$\langle Q^2 \rangle / V$	0.31	0.07	0.28	0.64	0.33	0.05
$\langle n \rangle / \sigma_n$	1.09	0.90	0.92	1.15	1.03	0.83

Table from LaTeX file

```
START
2343 <> b'\\begin{table}[htb]'
2344 <> b'\\begin{center}'
2345 <> b'\\caption{'
2346 <> b'SU(3) data: $n=n_++n_-$ with $n_\\pm$ the number of zero and small non-zero'
2347 <> b'eigenvalues with chirality $\\pm$.'
2348 <> b'$Q=n_+-n_-$ is the topological charge. $\\sigma_n$ is the variance of $n$.'
2349 <> b'The volume normalizations for $n$ and $Q^2$ are per spatial $8^3$ volume.'
2350 <> b'}'
2351 <> b'\\label{tab:su3}'
2352 <> b'\\begin{tabular}{c|cc|c|ccc} \\hline'
2353 <> b'volume            &\\multicolumn{2}{|c|}{$8^3\\times 4$} & $12^3\\times 4$'
2354 <> b'&\\multicolumn{3}{|c}{$16^3\\times 4$} \\\\'
2355 <> b'$\\beta$            & 5.75 & 5.85 & 5.75 & 5.71 & 5.75 & 5.85 \\\\ \\hline'
2356 <> b'$\\<n\\>/V$         & 0.32 & 0.06 & 0.28 & 0.63 & 0.30 & 0.05 \\\\'
2357 <> b'$\\<Q^2\\>/V$       & 0.31 & 0.07 & 0.28 & 0.64 & 0.33 & 0.05 \\\\'
2358 <> b'$\\<n\\>/\\sigma_n$ & 1.09 & 0.90 & 0.92 & 1.15 & 1.03 & 0.83 \\\\ \\hline'
2359 <> b'\\end{tabular}'
2360 <> b'\\end{center}'
2361 <> b'\\end{table}'
END
```

OCR from PDF image

```
123 x4
5.75

8° x 4

5.75 5.85

16° x 4
5.71 5.75

volume

B

(n)/V_  | 0.32 0.06 | 0.28 | 0.63 0.30
(Q?)/V  | 0.31 0.07 | 0.28 | 0.64 0.33
(n)/on  | 1.09 0.90] 0.92 | 1.15 1.03

on

0

ooo
ooo
Ww ot ot ol
```

Alignment Code

Table Alignment, Results

- Match LaTeX tables to table images from PDF
- Leveraging OCR: order of tables in LaTeX file may not be the same as in PDF tables
- Tried several algorithms, settled on Jaccard similarity
  - with conservative threshold

- Total papers processed: 391,623
- Papers with detected tables: 100,322 (~26%)
- Total number of tables: 300,018
- Total aligned tables: 131,485 (~ 43%)

```
START
<> b'\\begin{table}[htb]'
<> b'\\begin{center}'
<> b'\\caption{'
<> b'SU(3) data: $n=n_++n_-$ with $n_\\pm$ the number of zero and small non-zero'
<> b'eigenvalues with chirality $\\pm$.'
<> b'$Q=n_+-n_-$ is the topological charge. $\\sigma_n$ is the variance of $n$.'
<> b'The volume normalizations for $n$ and $Q^2$ are per spatial $8^3$ volume.'
<> b'}'
<> b'\\label{tab:su3}'
<> b'\\begin{tabular}{c|cc|c|ccc} \\hline'
<> b'volume              &\\multicolumn{2}{|c|}{$8^3\\times 4$} & $12^3\\times 4$'
<> b'&\\multicolumn{3}{|c}{$16^3\\times 4$} \\\\'
<> b'$\\beta$             & 5.75 & 5.85 & 5.75 & 5.71 & 5.75 & 5.85 \\\\ \\hline'
<> b'$\\<n\\>/V$           & 0.32 & 0.06 & 0.28 & 0.63 & 0.30 & 0.05 \\\\'
<> b'$\\<Q^2\\>/V$         & 0.31 & 0.07 & 0.28 & 0.64 & 0.33 & 0.05 \\\\'
<> b'$\\<n\\>/\\sigma_n$ & 1.09 & 0.90 & 0.92 & 1.15 & 1.03 & 0.83 \\\\ \\hline'
<> b'\\end{tabular}'
<> b'\\end{center}'
<> b'\\end{table}'
END
```

volume	$8^3 \times 4$		$12^3 \times 4$	$16^3 \times 4$		
β	5.75	5.85	5.75	5.71	5.75	5.85
$\langle n \rangle / V$	0.32	0.06	0.28	0.63	0.30	0.05
$\langle Q^2 \rangle / V$	0.31	0.07	0.28	0.64	0.33	0.05
$\langle n \rangle / \sigma_n$	1.09	0.90	0.92	1.15	1.03	0.83

Next steps

- Finish extraction, alignment
- Wrap up parallel table corpus as shareable resource
  - Deliverable: all tables we can extract with LaTeX, JPG, OCR versions; mapping between LaTeX and PDF versions
- Model tables, using generative AI
- QA on table content
- Where do “humans in the loop” fit in?
- Evaluate

Additional Goal: A Community to Share Even More?

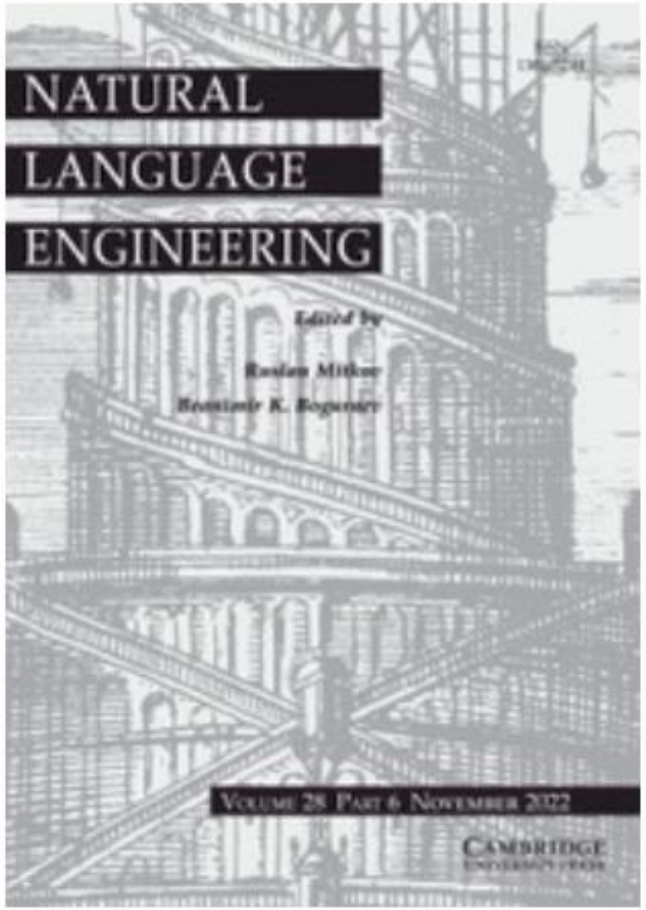
- Similar ideas and approaches, some differences in data, structuring, focus:
  - arXiv, PubMedCOA, …
  - Word/LaTeX, PDF/XML-HTML, PDF/LaTeX-HTML-MD
  - Focus on pages/tables/math/irregular tables, …
  - Lots of underlying tools

  People working on portions will solve the problem!

Collaborate !

- More data is better (Banko & Brill/ACL2001, etc.)
  - Emphasize data collection and sharing, advancing field
  - Can we do something to help in sharing partial solutions, datasets, tools for this and other tasks?
- Huge topic, how we do get others to join in the fun, and address portions of problems?
- Ideas? Please reach out: to me, to like-minded folks
  Discuss tomorrow?

Risks 2.0 and 3.0



Church, Kenneth Ward, Annika Marie Schoene, John E. Ortega, Raman Chandrasekar and Valia Kordoni. “Emerging trends: Unfair, biased, addictive, dangerous, deadly, and insanely profitable.”
Nat. Lang. Eng. 29 (2022): 483-508.

Church, Kenneth Ward and Raman Chandrasekar. “Emerging trends: Risks 3.0 and proliferation of spyware to 50,000 cell phones.”
Nat. Lang. Eng. 29 (2023): 824 - 841.