

From Surface to Semantics:

Semantic Structure Parsing for Table-Centric Document Analysis

Xuan Li, Jialiang Dong, Raymond Wong

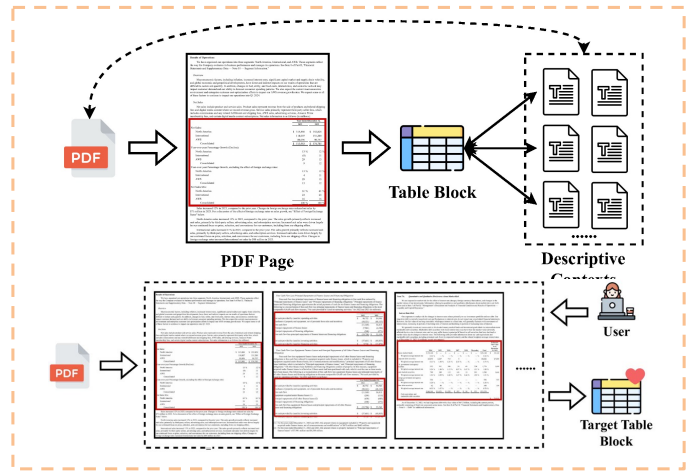
University of New South Wales, Sydney, Australia

Introduction

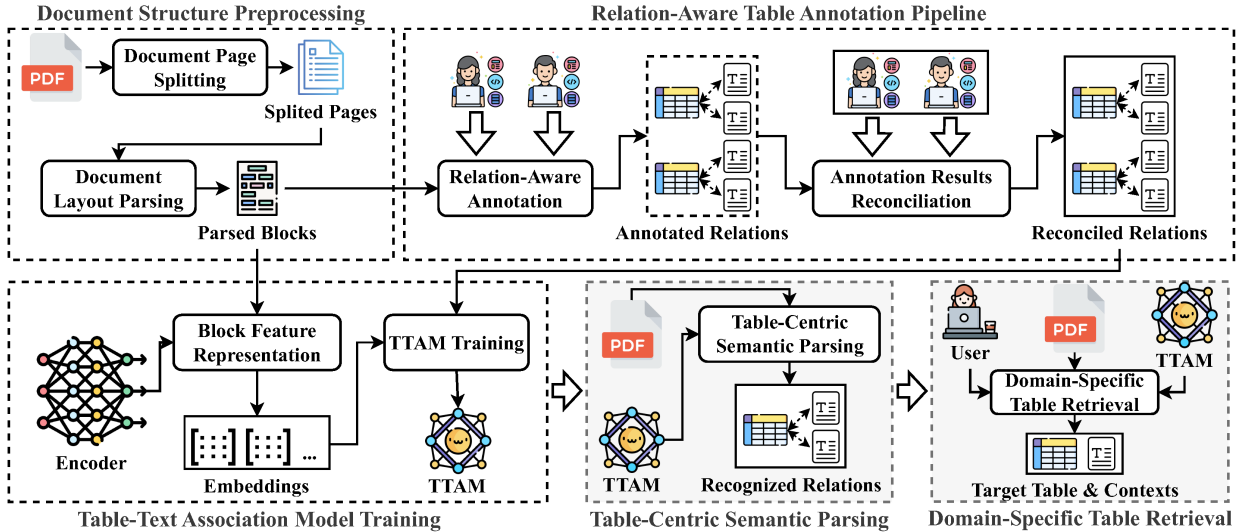
Tables are key carriers of structured information in documents such as financial reports and research papers. However, most existing document analysis systems only perform surface-level layout and text extraction, lacking a deep understanding of semantic relationships between tables and their descriptive text. We propose DOTABLER — a table-centric semantic parsing framework that bridges this gap. DOTABLER integrates document layout analysis and OCR with a new Table-Text Association Model (TTAM) to identify which paragraphs semantically describe each table.

Main Contribution

- **We design DoTabler**, which integrates a complete document preprocessing pipeline and semantic relationship analysis of PDF elements, enabling both semantic structure parsing and domain-specific table retrieval.
- **We propose the first PDF semantic-level dataset**, to effectively analyze relationships between tables and their contextual content.
- **We conduct a comprehensive evaluation** of DoTabler on nearly 4,000 pages of real-world PDF documents, demonstrating its superior performance and practical utility in application.



DOTABLER Pipeline



Experiment Results

Table 1: Details of the Constructed Dataset

Source	#PDF	#Page	#Table Block	#Text Block
arXiv	125	2,408	741	1,101
PubMed Central	102	1,544	320	523
Sum	227	3,952	1,061	1,624

Table 2: Performance Evaluation of Table-Text Block Linking

Scheme	TP	FP	TN	FN	Precision	Recall	F1
GPT-4o	168	19	450	338	89.84	33.20	48.48
Gemini-2.0	373	59	410	133	86.34	73.72	79.53
Claude-3.5	316	31	438	190	91.07	62.45	74.09
BERT	426	35	434	80	92.41	84.19	88.11
BART	444	50	419	62	89.88	87.75	88.80
RoBERTa	455	50	419	51	90.10	89.92	90.01

Table 3: Results of Table-Centric Document Semantic Parsing

Scheme	All Correct	POS Correct	NEG Correct	#Sum
GPT-4o	109	119	183	193
Gemini-2.0	113	147	159	
Claude-3.5	114	137	170	
DoTABLER	128	166	155	

Table 4: Results of Domain-Specific Table Retrieval

Scheme	Retrieval Recall@K
DoTABLER @K=1	71.70
DoTABLER @K=2	84.91
DoTABLER @K=3	88.68

Table 5: Overall Time Cost of Distinct Schemes

Scheme	Mean Time Cost (s)	Median Time Cost (s)
GPT-4o	0.8008	0.6201
Gemini-2.0	0.7434	0.5763
Claude-3.5	1.7054	1.4214
DoTABLER	0.0035	0.0031