Docs2KG

Unified Knowledge Graph Construction from Heterogeneous Documents Assisted by Large Language Models

Pascal Sun

Computer Science and Software Engineering

The University of Western Australia

November 22, 2024

Table of Contents

1. The Problem and Opportunity
2. Stage 1: Unstructured Data Pre-process
3. Stage 2: Knowledge Graph Construction
4. Q & A

The Problem and Opportunity

The Problem and Opportunity

- **Fact**: Most of enterprise knowledge reside in unstructured documents of heterogeneous formats, especially PDFs.
- **Problem**: LLM can not easily access them.
- **Opportunity**: Get them structured and accessible for LLM agents.

Core Challenges to Address

We have two stages to construct knowledge graphs from docs:

- Unstructured Data Pre-process
  - PDFs, Excel, Web Pages, and Emails.
- Knowledge Graph Construction from unstructured data

Stage 1: Unstructured Data Pre-process

Unstructured Data Pre-process

- **Web Content (HTML)**
  - Well-structured and standardized format
  - Forms foundation for modern search engines
- **PDF Documents**
  - Digitally generated: Direct text extraction
  - Scanned documents: Requires OCR processing
- **Email Communications**
  - We have plain text or HTML formats
- **Spreadsheet Data**
  - Appears structured but complex scenarios create challenges

Stage 1: Data Preprocessing Pipeline

Challenge: Processing diverse data formats through a unified pipeline

- **Unified Output Format**
  - Markdown as universal container
  - figures and tables will be output separately
- **Dual Processing Paths**
  - Image Pipeline: scanned documents, excel also in this pipeline
  - Markdown Pipeline: For native digital content

Docs2KG Overview

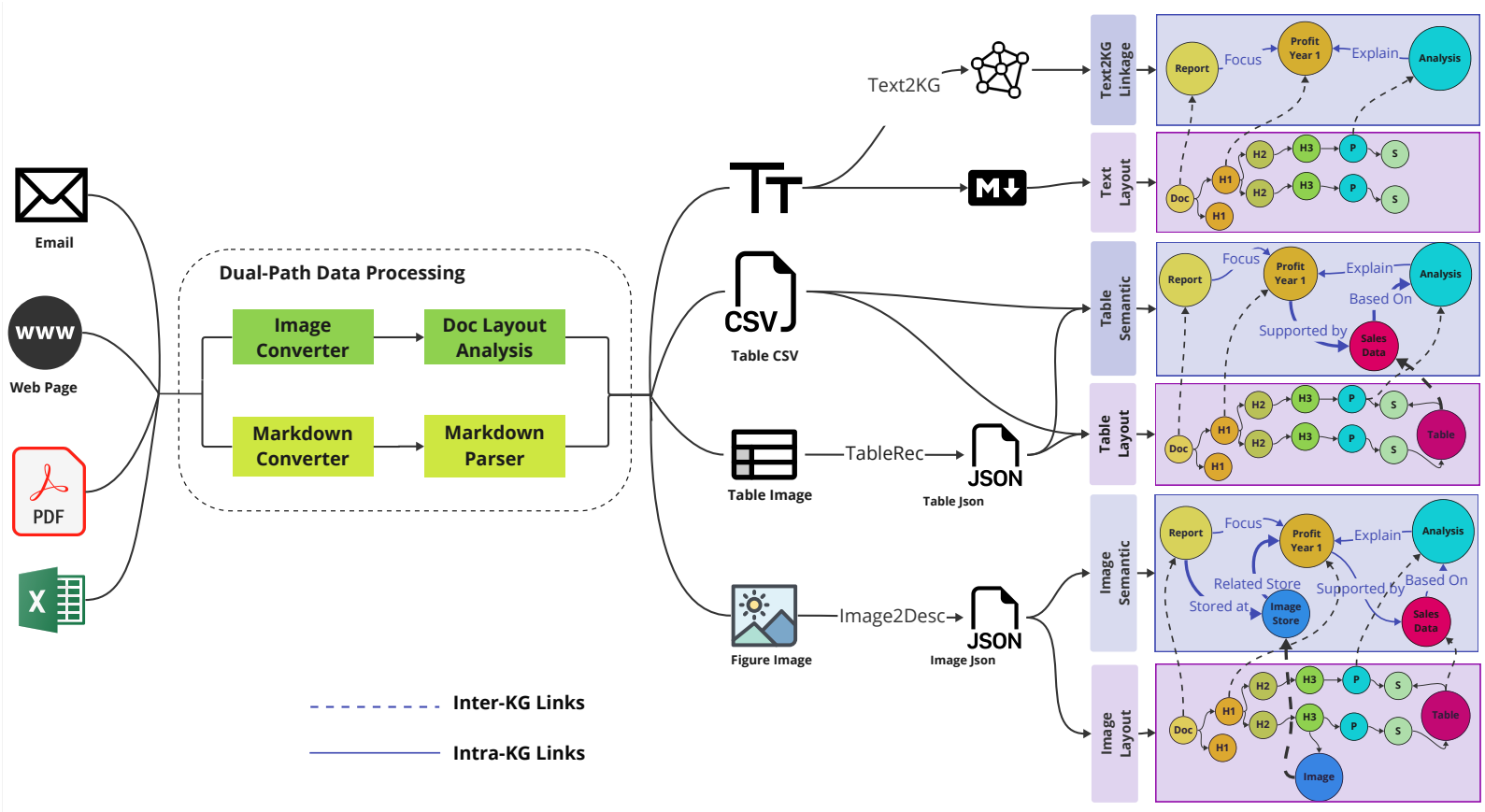

Figure: Docs2KG Overview

Stage 2: Knowledge Graph Construction

Stage 2: Knowledge Graph Construction - Key Questions

1. **Starting Point?**
   - Without established ontology, where should we begin?
2. **Where to stop?**
   - When has our knowledge graph reached **adequate** completeness?

Stage 2: Knowledge Graph Construction - Starting point

- Do you have an ontology?
  - If yes, then use ontology in the prompt to build one.
- If not: Do you have an entities list?
  - You normally do, within your relational databases, or your company daily operations.
  - These are the **entities of interest** for your company.
  - Use them to link documents together
- If not: We had to rely on the bottom up KG construction strategy
  - Ask LLM to generate a list of entity types (ontology) for you.
  - Human in the loop to improve it.

Stage 2: Knowledge Graph Construction - Where to stop?

Challenge: No established metrics for KG completeness

Current Evaluation Strategy

Measured through downstream applications

Geology Domain Validation

1. **Query Capability**
   - Success rate of Cypher queries
2. **RAG Performance**
   - Integration with one-hop graph search
   - Result quality assessment

Q & A

Further Development Plan[1]

Current Features

- **Python Package Delivery**
  - **Input:** File(s)
  - **Output:** JSON importable into Neo4j
- **Prompting with OpenAI**
- **Image Path Processing**
  - Utilizing LLM and PaddleOCR

Work in Progress

- **Command-Line Support**
  - Inputs: Files and optional ontologies
- **Neo4j Output Enhancement**
  - Generating Neo4j-supported JSON format
- **Public LLM Integration**
  - Support for publicly available LLMs

[1] https://docs2kg.ai4wa.com/

Thank You For Your Attention!

Contact me: pascal.sun@research.uwa.edu.au