Navy Requirement Markup Language

STUDENTS: Jenna Flores, Ritik Shrivastava, Bao Van

NAVSEA WARFARE CENTERS Philadelphia

Introduction

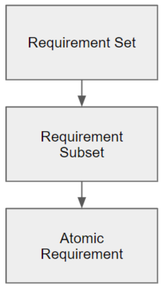
- The NSIN is a network in the U.S. Department of Defense (DOD) aimed at connecting DOD entities with academic and venture partners to innovate new solutions for DOD-member challenges. Navy specifications and requirements are typically presented in narrative form (e.g., paragraphs).
- The narrative requirements are hard to parse visually and nearly impossible to parse using machines in any effective time frame.
- To solve this issue, we created an autonomous system that intakes PDF files and outputs machine-readable JSON files utilizing a machine learning model which can be converted to REQIF or XML for further usage

3.6 Enclosure physical characteristics. The physical characteristics of the equipment shall be as specified in 3.6.1 *Dimensions* through 3.6.3.2 *Work surfaces, console cabinet*.

3.6.1 Dimensions. WCS maximum dimensions shall not exceed 84 inches in height, 96 inches in width, and 50 inches in depth. The dimensions of either the low or high power transfer standard (Paragraph 3.9.2.4) dimensions shall not exceed 14 inches in depth, 19 inches in width, and 5.5 inches in height.

3.6.2 Weight. The weight of the WCS including all covers and accessories shall not exceed 4,200 lbs [based on Navy calibration laboratory maximum floor occupancy load constraint of 125 lbs/ft^2, and maximum occupancy space of 33.3 ft^2 = 4,800 in^2 (96 inches x 50 inches)]. The weight of either the low or high power transfer standard (Paragraph 3.9.2.4) shall not exceed 15 lbs.

Input Document



Nested Structure

Objectives

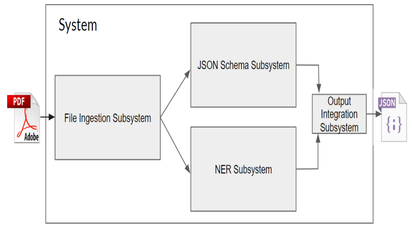
- Develop and define a markup language to capture key characteristics of Navy requirements
- Create a machine learning model to extract entities out of the document paragraph to include in the markup language
- Develop a automated pipeline for processing requirements documents and creating machine readable markup files at scale
- Demonstrate a prototype application of document processing

Requirements

- Use Schema.org standard for markup language
- Identify and process hierarchical structure and their respective paragraph numbers
- Identify and process PDF files containing requirements content
- Shall not store or share data to third party tools
- Adhere to the intellectual property rules and regulations
- ML model shall be able to process PDF files of up to 20 pages
- System shall process a PDF file within 10 minutes

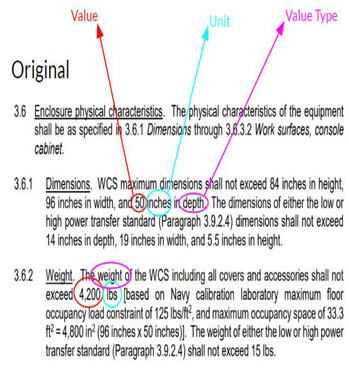
System Design

- File Ingestion Subsystem: Extracting text from PDF and creating text flies
- Reading the text and creating a nested JSON
- Using the trained model to get the entities and their types identified for the content paragraph
- Adding the entity and its class to the sentence and adding them to the JSON as an element which is then given as output and can be converted to REQIF or XML



High Level System Design

Data labeling



Azure Data Labeling

Data Provided:

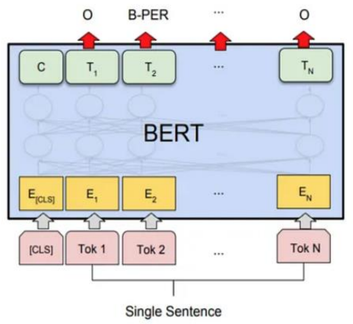
- Multiple PDF files were provided were either original or scans of requirement PDF

Data Preparation:

- Partition text for just the requirement section. This is application specifically
- After extracting content from the input PDF into a text file, we labeled the data on Azure
- BOI tagging was done for the classes 'Value', 'Type' and 'Unit'
- Data was exported in CoNLL format and used in machine learning pipeline

Named Entity Recognition (NER)

- Named Entity Recognition is a subtask of information extraction which identifies and classifies entities in unstructured text into categories
- Makes it easy to identify and classify values and their unit/type pairs
- Optimized BERT base NER, a fine tuned BERT model of Hugging Face to perform NER per our use case
- Fine tuned the existing BERT Base NER using our prepared data (Transfer Learning) to make it domain specific



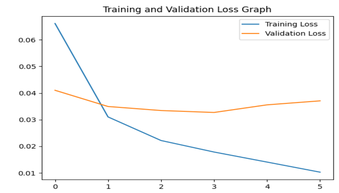
BERT CoNLL NER

ML Model

- Considering unbalanced data for different classes, the F1 score is chosen as matric
- 0 being minimum and 1 being maximum, we achieve an excellent F1 score, i.e., more than 0.8 for class 'unit' and 'value'
- Class 'Type' has a low F1 score because it has the lowest distribution in our data and will improve with more input data to train on
- Validation loss first decreases, then it increases while training loss keeps on decreasing, so the model stops to prevent overfitting. With more training data or data augmentation, validation loss may decrease more

	precision	recall	f1-score	support
type	0.51	0.49	0.50	119
unit	0.81	0.93	0.87	180
value	0.78	0.85	0.81	275
micro avg	0.74	0.80	0.77	574
macro avg	0.70	0.75	0.73	574
weighted avg	0.74	0.80	0.77	574

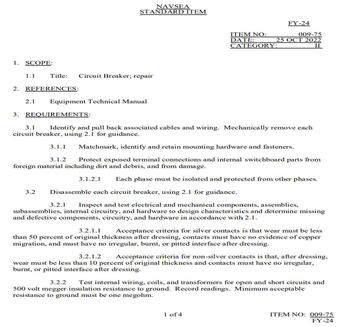
Accuracy Matrix



Training and Validation Loss Curve

Results

- Left is sample pdf, and the right is the resultant JSON file
- We get the output JSON corresponding to input PDF with entities and classes of content identified



NAVSEA STANDARD ITEM

FY-24

ITEM NO: 009-75
DATE: 25 OCT 2002
CATEGORY: II

1. SCOPE:
1.1 Title: Circuit Breaker; repair
2. REFERENCES:
2.1 Equipment Technical Manual
3. REQUIREMENTS:
3.1 Identify and pull back associated cables and wiring. Mechanically remove each circuit breaker, using 2.1 for guidance.
3.1.1 Matchmark, identify and retain mounting hardware and fasteners.
3.1.2 Protect exposed terminal connections and internal switchboard parts from foreign material including dirt and debris, and from damage.
3.1.2.1 Each phase must be isolated and protected from other phases.
3.2 Disassemble each circuit breaker, using 2.1 for guidance.
3.2.1 Inspect and test electrical and mechanical components, assemblies, subassemblies, internal circuitry, and hardware to design characteristics and determine missing and defective components, circuitry, and hardware in accordance with 2.1.
3.2.1.1 Acceptance criteria for silver contacts is that wear must be less than 50 percent of original thickness after dressing, contacts must have no evidence of copper migration, and must have no irregular, burnt, or pitted interface after dressing.
3.2.1.2 Acceptance criteria for non-silver contacts is that, after dressing, wear must be less than 10 percent of original thickness and contacts must have no irregular, burnt, or pitted interface after dressing.
3.2.2 Test internal wiring, coils, and transformers for open and short circuits and 500 volt megger insulation resistance to ground. Record readings. Minimum acceptable resistance to ground must be one megohm.

1 of 4 ITEM NO: 009-75 FY-24

Input Test Sample

Output Test Sample

Results

- Further improving the ML Model to make it compatible with different format of documents
- Scaling up intake file count
- Implement REQIF documentation
- Include option for XML output

ELECTRICAL & COMPUTER ENGINEERING
UNIVERSITY of WASHINGTON

ADVISORS: Haixun Wang, Kevin Murphy, TJ McKelvey

SPONSORS: NATIONAL SECURITY INNOVATION NETWORK (NSIN) / U.S. NAVY