



PRACTICAL CHALLENGES IN HANDLING UNSTRUCTURED DATA USING ARTIFICIAL INTELLIGENCE

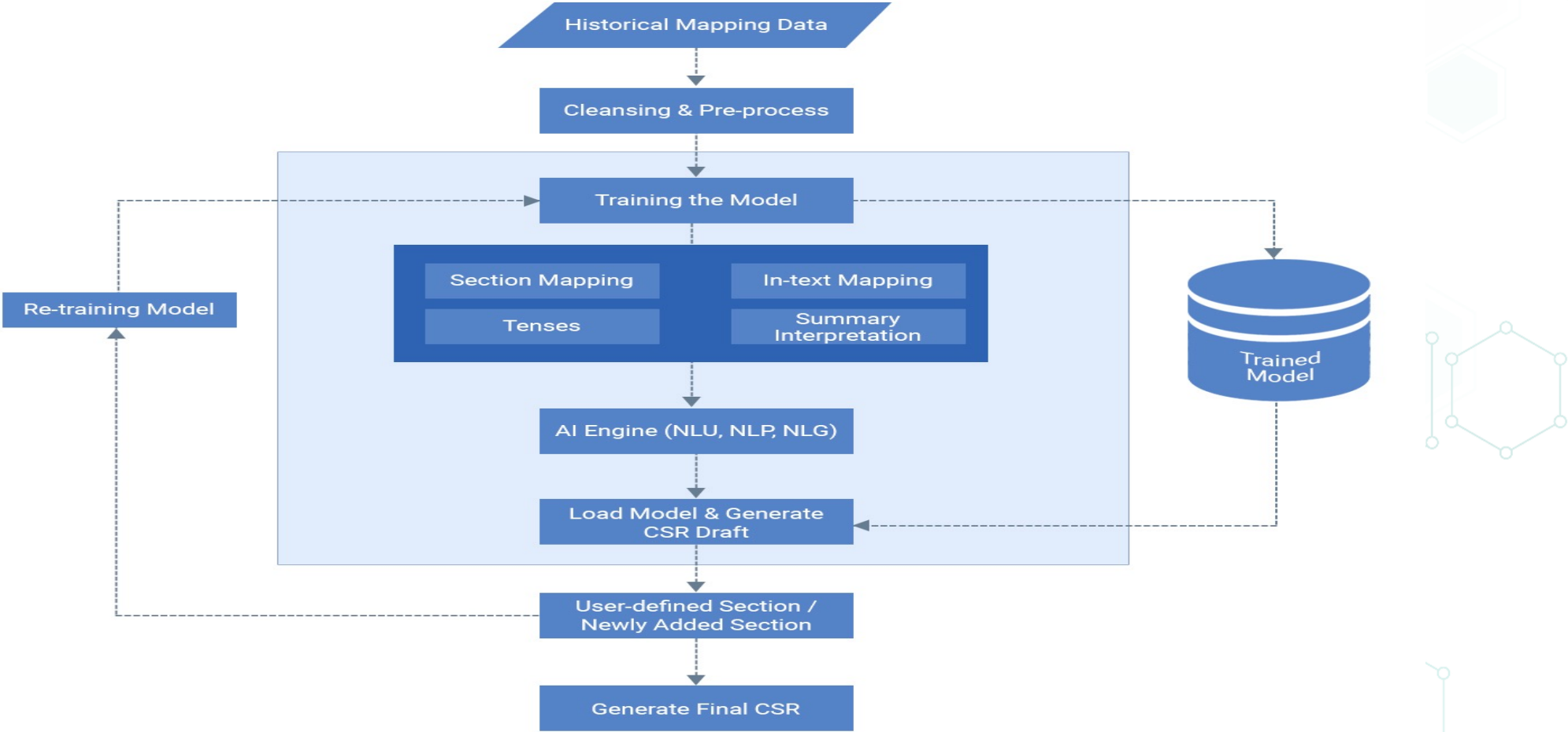
WHERE DID WE MANAGE UNSTRUCTURED DATA ?

- ❖ We created an application to author clinical study report (CSR) using AI
- ❖ Both inputs to the application and the output CSR are unstructured data (pdf/word documents)
- ❖ Creating Clinical Study Report (CSR) describes the end points, data collection and safety & efficacy analysis of the clinical study.
- ❖ Authoring of CSR is highly manual and time consuming
- ❖ The content of CSR is constructed from Protocol, SAP, Safety narratives, In-text tables and medical writer's interpretation of study results.
- ❖ The ML prediction algorithm along with NLP/NLG is used to identify/effectively used to read/write the texts from other documents such as Protocol, SAP, In-text Table & etc.,
- ❖ Copy and paste information from multiple source documents to relevant sections of the CSR.

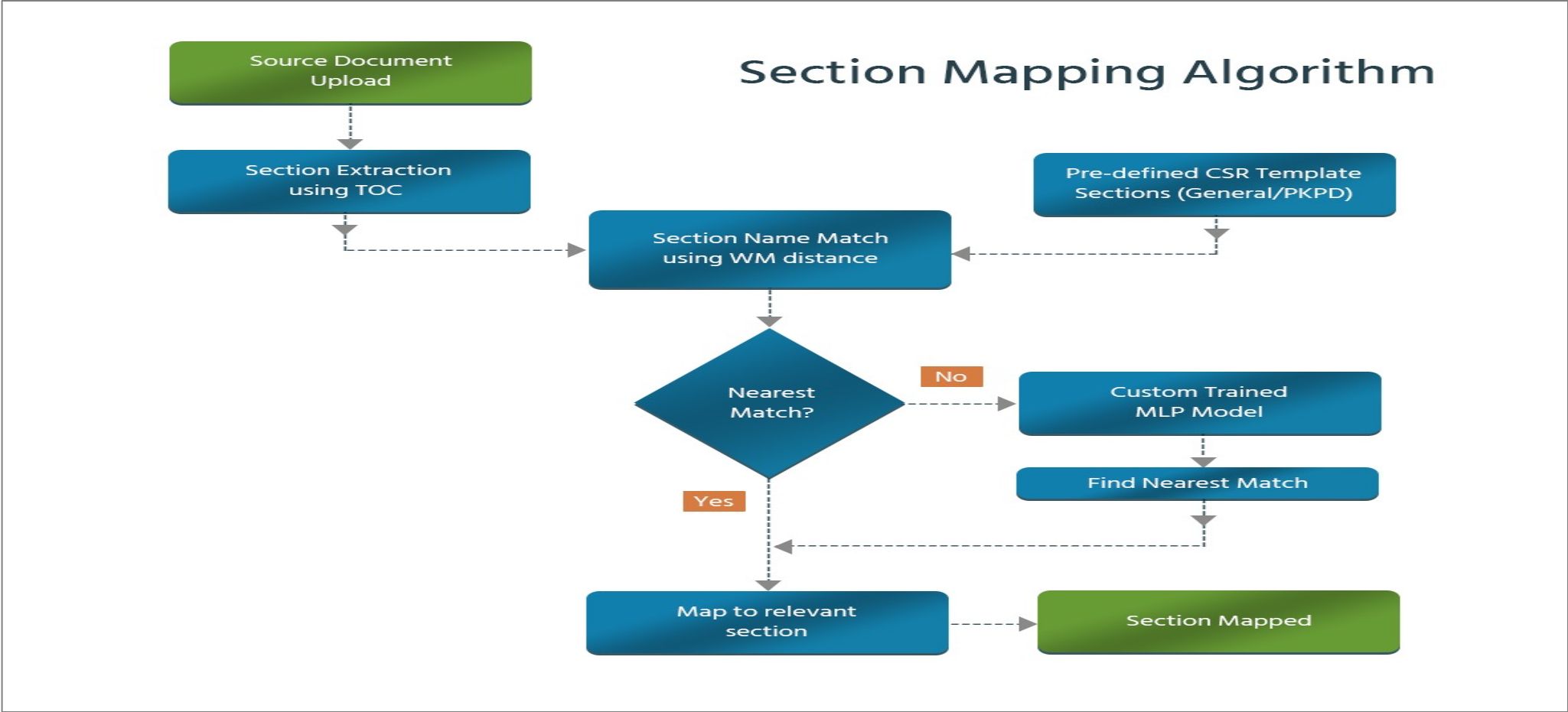
CHALLENGES IN HANDLING UNSTRUCTURED DATA

- ❖ Extracting data from source documents (PDF/Word, formatting and special symbols, subscript & superscripts and diagrams)
- ❖ Summarizing large paragraphs using NLU & NLG for CSR synopsis
- ❖ Reading the data from the unstructured TLFs & table details extraction of data and interpretation in simple English
- ❖ Tense conversion
- ❖ Section mapping challenges
- ❖ Optimization for performance in ML models

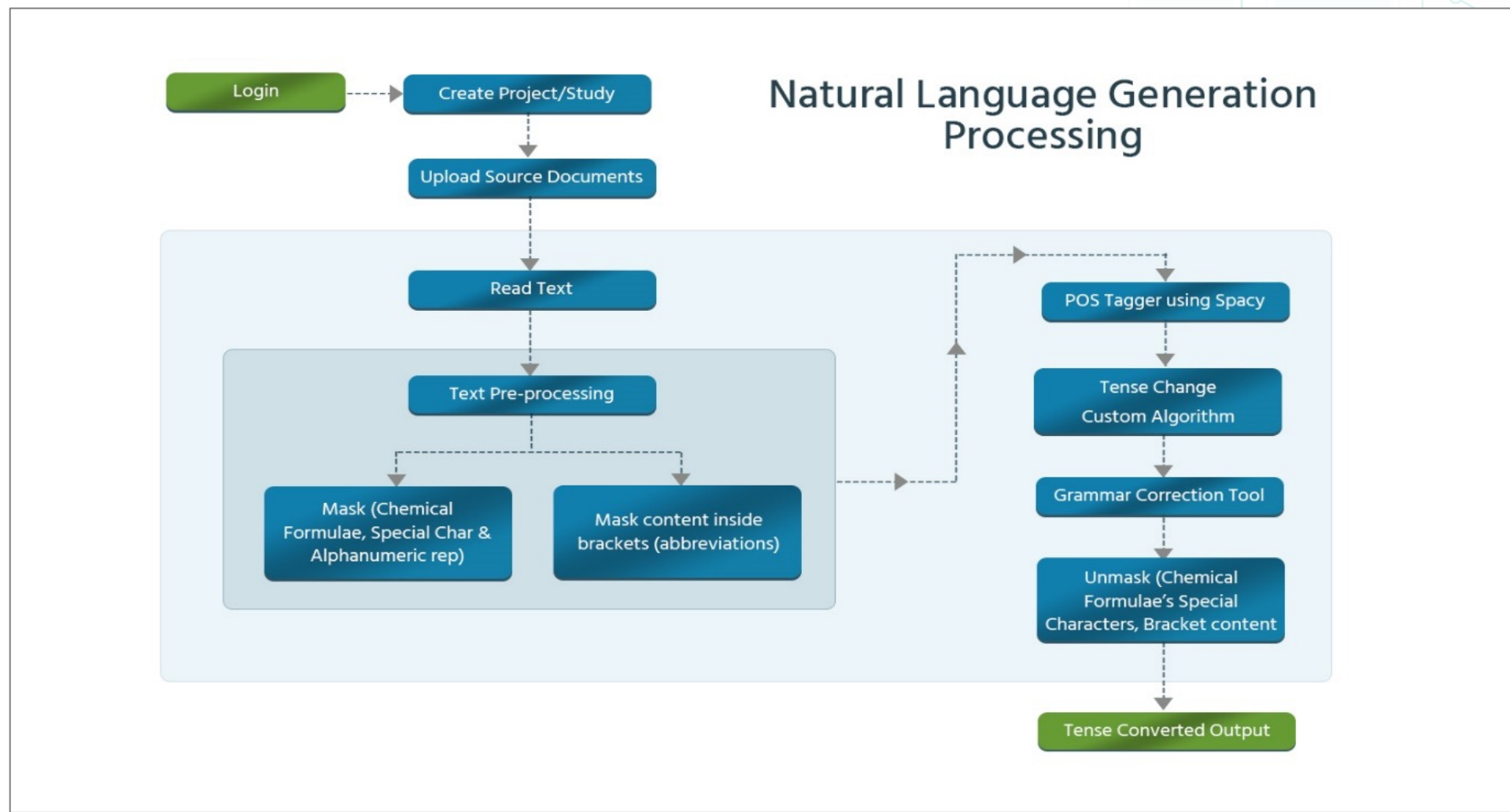
HOW AI IS APPLIED



COPYING TEXT FROM SOURCE TO CSR SECTIONS



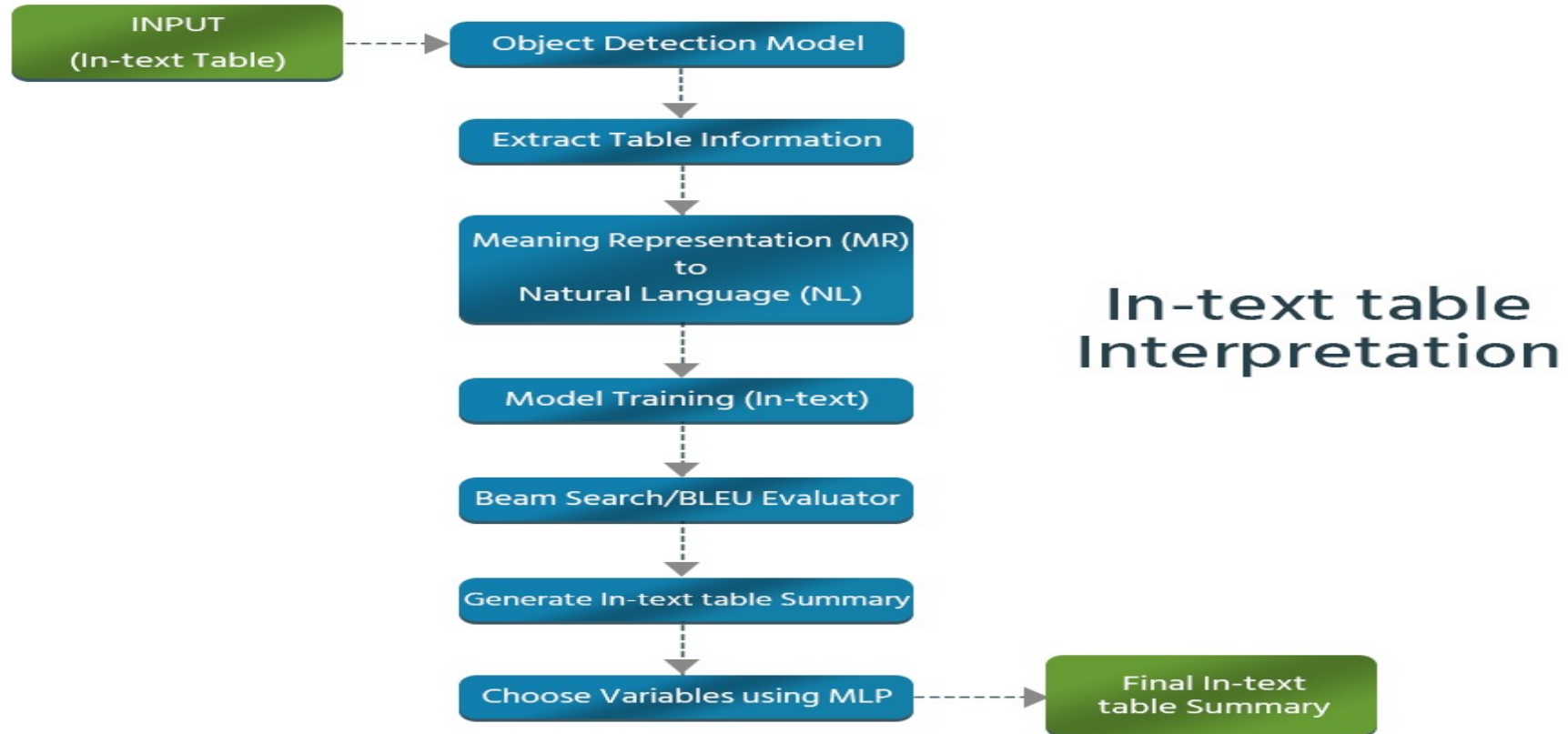
NATURAL LANGUAGE GENERATION (NLG)



NATURAL LANGUAGE GENERATION (NLG) – CONT'D

- ❖ Mask the chemical formulae's, special characters, alphanumeric representation
- ❖ Use the spacy tokenizer to tokenise each word to get the POS tags to detect the verbs
- ❖ Based on the pos tags detection change the tense of the word with respect to the sentence based on English grammatical rules algorithm custom made
- ❖ Apply the grammar correction tool to ensure the sentence that's been tense changed are grammatically correct
- ❖ Unmask the chemical formulae's, special characters, brackets content to reflect the original content

TLF INTERPRETATION IN SIMPLE ENGLISH

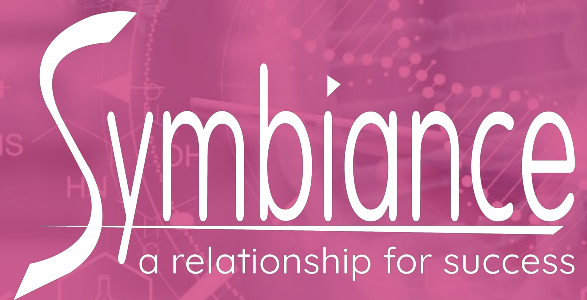


TLF INTERPRETATION IN SIMPLE ENGLISH – CONT'D

- ❖ Detect the in-text table structure using custom trained detectron2 object detection model
- ❖ Extract the table structure information's in data frames using custom algorithms
- ❖ Custom trained sequece2sequence model transferring meaning representations (MR) of a in-text table into a description in natural language (NL). Pytorch is used to build up the seq2seq neural model.
- ❖ Train model on the in-text table interpretation data set on a GPU
- ❖ Implement beam search for evaluation
- ❖ Implement a BLEU evaluator and report BLEU scores
- ❖ Generate the in-text table summary based on the template detected using seq2seq model
- ❖ Use a MLP Model to choose the variables whose values need to be replaced, and final summary text will be generated.

Questions and Comments?





Thank You!!

Farha Fathima Feroze
fferoze@symbiance.com
Product Manager
Mobile: +1 (609) 945 7431