Long Legal Document Entity Extraction with Hierarchical Section-based Representation

Data Science Institute
COLUMBIA UNIVERSITY

Authors: Ning Kang, Xindi Deng, Yihan Chen,
Yuhui Wang, Yuxin Lin
Mentor: Vladimir A. Kobzar

Data Science Capstone Project
with Fu Foundation School of
Engineering and Applied Science

Background

Owing to the intricate and lengthy nature, the reading comprehension of legal documents raises great challenges for both human and AI models. To help readers grasp the main idea in the Merger and Acquisition (M&A) Agreements, the MAUD paper proposes 22 types of questions and aims to find key clauses (i.e., answers) for each question. Each question relates to a deal point type and the answer is a span in the contract. However, the original solution in MAUD did not achieve a satisfying result.

Main idea: Use section hierarchy information to help model understand long legal document.

Introduction

Dataset: This project uses hierarchical representation of text from MarkupMnA and deal point information from MAUD. These two datasets share the same 151 M&A agreements.

- MarkupMnA: CSV formatted contracts containing annotated sections titles and section numbers, derived from HTML.
- MAUD: JSON formatted contracts with 22 questions and relevant answers per agreement. Each question corresponds to a deal point type.

Task: Extractive question-answering task on M&A agreements as example shown in Figure 1.

<u>Contract:</u> ......a material breach of this Agreement; (e) by Parent, if, prior to obtaining the Company Stockholder Approval, (i) the Company Board of Directors shall have effected a Change of Recommendation (whether or not in compliance with this Agreement) or (ii) the Company has materially breachedSection5.3; (f) by either the Company .... Meeting duly convened therefor or at any adjournment or postponement thereof; or (h) by the Company in order to effect a Change of Recommendation and substantially concurrently enter into a definitive agreement providing for a Superior Proposal

<u>Question:</u> Highlight the parts of the text related to deal point "Breach of No Shop" that should be reviewed by a lawyer
<u>Answer:</u> (e) by Parent, if, prior to obtaining the Company Stockholder Approval, (i) the Company Board of Directors shall have effected a Change of Recommendation (whether or not in compliance with this Agreement) or (ii) the Company has materially breachedSection5.3;

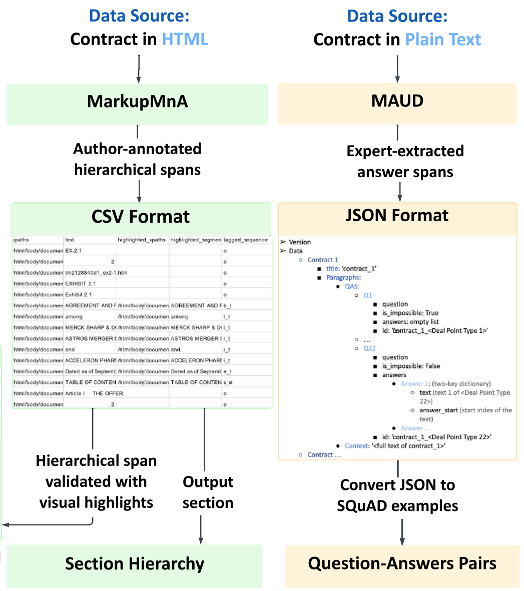
<u>Question:</u> Highlight the parts of the text related to deal point "Lim. on FTR exercise" that should be reviewed by a lawyer
<u>Answer:</u> (h) by the Company in order to effect a Change of Recommendation and substantially concurrently enter into a definitive agreement providing for a Superior Proposal

Figure 1. Example of Task

Data Overview

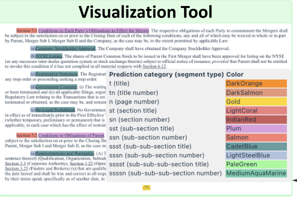
MarkupMnA dataset is in CSV format, containing hierarchy representations that relates text spans to different levels of hierarchy. With the visualization tool, we found that multiple CSV files had inaccurate hierarchy information.

MAUD dataset is a SQuADv2 formatted JSON file, in which a list of answers to 22 questions are recorded for each agreement.



xpaths	text	highlighted_xpaths	highlighted_segmen	tagged_sequence
html/body/documen	EX-2.1			o
html/body/documen	2			o
html/body/documen	tm2128840d1_ex2-1.htm			o
html/body/documen	EXHIBIT 2.1			o
html/body/documen	Exhibit 2.1			o
html/body/documen	AGREEMENT AND F	/html/body/documen	AGREEMENT AND F	b_t
html/body/documen	among	/html/body/documen	among	i_t
html/body/documen	MERCK SHARP & D(	/html/body/documen	MERCK SHARP & D(	i_t
html/body/documen	ASTROS MERGER S	/html/body/documen	ASTROS MERGER S	i_t
html/body/documen	and	/html/body/documen	and	i_t
html/body/documen	ACCELERON PHARM	/html/body/documen	ACCELERON PHARM	i_t
html/body/documen	Dated as of Septemb	/html/body/documen	Dated as of Septemb	e_t
html/body/documen	TABLE OF CONTEN	/html/body/documen	TABLE OF CONTEN	s_st
html/body/documen	Article I THE OFFER			o
html/body/documen	2			o

➢ Version
➢ Data
  ○ Contract 1
    ■ title: 'contract_1'
    ■ Paragraphs:
      ● QAS:
        ○ Q1
          ■ question
          ■ is_impossible: True
          ■ answers: empty list
          ■ id: 'contract_1_<Deal Point Type 1>'
        ○ ...
        ○ Q22
          ■ question
          ■ is_impossible: False
          ■ answers
            ● Answer 1: (two-key dictionary)
              ○ text (text 1 of <Deal Point Type 22>)
              ◇ answer_start (start index of the text)
            ● Answer...
          ■ id: 'contract_1_<Deal Point Type 22>'
      ● Context: '<full text of contract_1>'
  ○ Contract ...


Visualization Tool

Section 7.1 Conditions to Each Party's Obligations to Effect the Merger. The respective obligations of each Party to consummate the Mergers shall be subject to the satisfaction on or prior to the Closing Date of each of the following conditions, any and all of which may be waived in whole or in part by Parent, Merger Sub I, Merger Sub II and the Company, as the case may be, to the extent permitted by applicable Law:

(a) Company Stockholder Approval. The Company shall have obtained the Company Stockholder Approval.

(b) NYSE Listing. The shares of Parent Common Stock to be issued in the First Merger shall have been approved for listing on the NYSE (or any successor inter-dealer quotation system or stock exchange thereto) subject to official notice of issuance; provided that Parent shall not be entitled to invoke this condition if it has not complied in all material respects with Section 6.12.

Prediction category (segment type)	Color
t (title)	DarkOrange
tn (title number)	DarkSalmon
n (page number)	Gold
st (section title)	LightCoral
sn (section number)	IndianRed
sst (sub-section title)	Plum
ssn (sub-section number)	Salmon
ssst (sub-sub-section title)	CadetBlue
sssn (sub-sub-section number)	LightSteelBlue
sssst (sub-sub-sub-section title)	PaleGreen
ssssn (sub-sub-sub-section number)	MediumAquaMarine

-77-

Figure 2. Highlighted Hierarchical Text based on MarkupMnA Annotation for Contract 126

Methods & Models

Main idea: To improve the performance of extractive question-answering task on MAUD, we combined section hierarchy information given by MarkupMnA with original solutions proposed by MAUD.

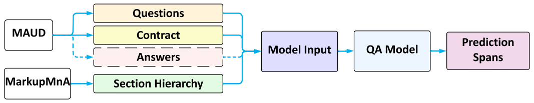


Figure 3. Workflow

Baseline Model: pretrained RoBERTa given by MAUD paper

Method 1: Section-based Sliding Windows

Motivation: For original sliding windows shown in Figure 4, MAUD used a fixed stride length to segment, which causes some subcontracts accross different sections, leading to a loss of the contextually coherence and a challenge for model's understanding.

Contribution: Incorporating section hierarchy information, we propose a new method to create sliding windows. The paragraph length and stride length are customized with the section split points as Figure 5. If the current sliding window covers a section split point, the split point will be considered as the end of current sliding window and the start of the next sliding window.

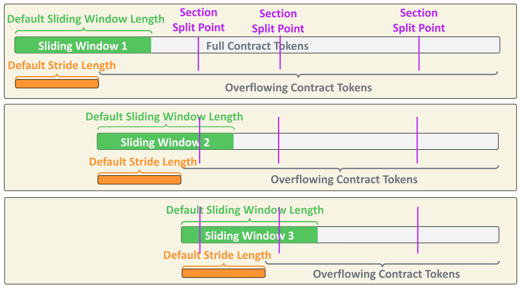


Figure 4. Original Sliding Window

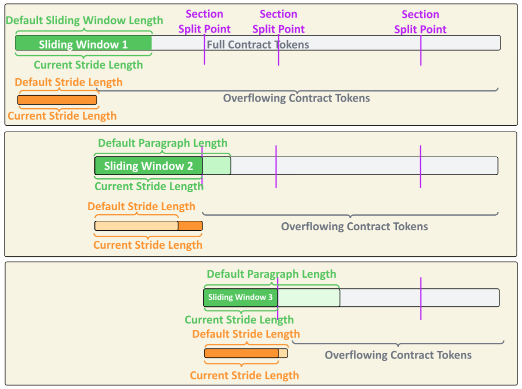


Figure 5. Section-based Sliding Window

Method 2: Section-based Training Set

Motivation: The original model used the data processing functions from Hugging Face. Corresponding to each question-contract pair, there could be many answer spans. However, the functions only extracted the first answer for each question-contract pair. Consequently, the remaining answers are left unused and unexplored.

Contribution: We segmented each contract into multiple sections using the section hierarchy mentioned above. In this case, the functions are able to extract one answer from each question-section pair, and more answer spans can be used in total. Therefore, the model can better harness the potential of the dataset.

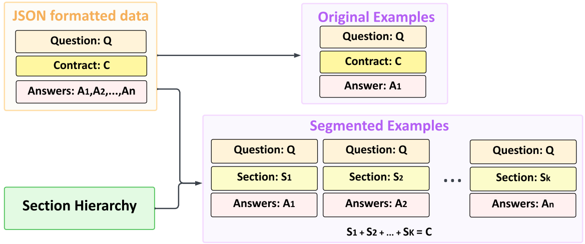


Figure 6. Comparison of Original Examples and Segmented Examples

Results & Conclusion

Result: Our Section-based Sliding Window in method 1 outperforms the original method on F1-score for all cases. The fine-tuned model trained on method 2 achieves a higher F1 score for question-contract pairs without cor answers.

Conclusion: In spite of the concerns existed in the pretrained model, bringing in the hierarchical information could still make a difference, improving the performance of the extraction task.

Future Work: A new embedding of section title could be added to model to take use of global information; The data processing functions could be modified to ensure the use of all the answers.

Method	F1 score			AUPR
	Total	Has Ans	No Ans	
Baseline	52.59	62.76	7.69	0.2186
Section-based Sliding Window	54.08	**63.89**	10.77	**0.2205**
Section-based Training Set	**55.67**	51.55	**73.85**	0.2136

Table 1. F1-score and AUPR

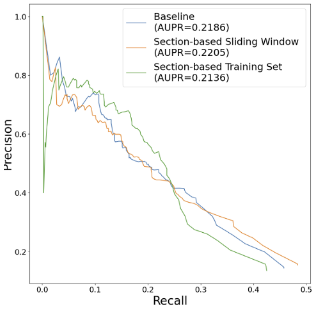


Figure 7. Precision-Recall Curve

Acknowledgement

We would like to express our gratitude towards Prof. Kobzar for his support, NYU team for visualization tool template, and Steven H. Wang for the pretrained model.

References

1. Rao, Sukrit, et al. "MARKUPMNA: Markup-Based Segmentation of M&A Agreements."
2. Wang, Steven H., et al. "MAUD: An Expert-Annotated Legal NLP Dataset for Merger Agreement Understanding." arXiv preprint arXiv:2301.00876 (2023).