MarkupMnA: Markup-Based Segmentation of M&A Agreements

Sukrit Rao[1]*, Pranab Islam[1]*, Rohith Bollineni[1], Shaan Khosla[1], Tingyi Fei[2], Qian Wu[2], Kyunghyun Cho[1], Vladimir A. Kobzar[2]†
[1]New York University [2]Columbia University

Abstract

Developing hierarchical representation of long documents is an important challenge in natural language processing (NLP). This challenge is especially relevant in the legal domain where documents tend to be particularly large. A natural segmentation of a document is given by its section hierarchy, which is typically present in long documents and often marked by special formatting/numbering of section titles. However, current processing techniques generally do not leverage this hierarchy, which leads to a loss of contextually coherent segments of the document. While segmenting long documents into smaller chunks is used to overcome maximum token limitations of language models, this is done either manually or in a manner that does not identify section titles or boundaries. Accordingly, developing a model to identify section titles in a document would lead to a contextually coherent segmentation and make progress towards developing hierarchical representation of long text. Taking stock of these goals, we develop MarkupMnA, a corpus of merger and acquisition (M&A) agreements in the HyperText Markup Language (HTML) with annotated sections titles based on filings by US public companies in the Securities and Exchange Commission's EDGAR database. Using MarkupMnA, we benchmark existing language models jointly pretrained on markup and text information, on the section identification task. Our experiments show that the models that use the markup information achieve promising results on this task and outperform models that use text information only. The agreements contained in MarkupMnA were previously included in the Merger Agreement Understanding Dataset (MAUD), a recent annotated dataset for reading comprehension and information extraction tasks. We hope that this commonality between the datasets will facilitate future work on understanding the effects of segmentation on information extraction and reading comprehension. More broadly, MarkupMnA makes progress towards using NLP models to harness a broad range of other legal documents, and leads to various new research directions in multimodal and legal NLP. The dataset, related code and trained models are available at `https://github.com/MarkupMnA`.

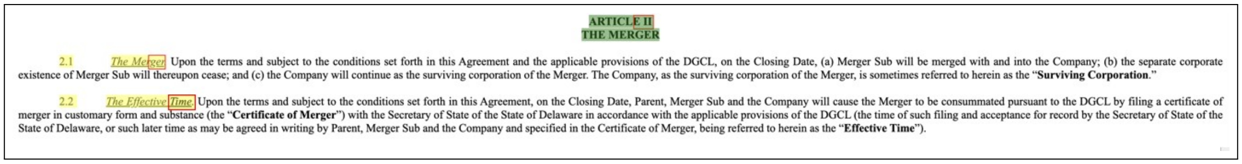
1 Introduction

The broad objective of our work is to make progress towards addressing a fundamental challenge of developing hierarchical representation of long text. Long document understanding is a challenging problem due to the lack of computationally efficient hierarchical representation of long documents and limited ability to process long multimodal input [28]. We focus on legal documents, specifically US public merger and acquisition (M&A) agreements, which tend to be especially large.

*These authors contributed equally to this work

†Corresponding author (`vak2116@columbia.edu`)

Preprint. Under review.



ARTICLE II
THE MERGER

2.1 *The Merger* Upon the terms and subject to the conditions set forth in this Agreement and the applicable provisions of the DGCL, on the Closing Date, (a) Merger Sub will be merged with and into the Company; (b) the separate corporate existence of Merger Sub will thereupon cease; and (c) the Company will continue as the surviving corporation of the Merger. The Company, as the surviving corporation of the Merger, is sometimes referred to herein as the "**Surviving Corporation**."

2.2 *The Effective Time* Upon the terms and subject to the conditions set forth in this Agreement, on the Closing Date, Parent, Merger Sub and the Company will cause the Merger to be consummated pursuant to the DGCL by filing a certificate of merger in customary form and substance (the "**Certificate of Merger**") with the Secretary of State of the State of Delaware in accordance with the applicable provisions of the DGCL (the time of such filing and acceptance for record by the Secretary of State of the State of Delaware, or such later time as may be agreed in writing by Parent, Merger Sub and the Company and specified in the Certificate of Merger, being referred to herein as the "**Effective Time**").

Figure 1: A sample agreement from MARKUPMNA with predictions from MarkupLM. The green and yellow highlights show ground truth annotations of respectively section and subsection titles. The red border marks the tokens (word pieces) that were predicted incorrectly, while the absence of any border indicates correct predictions.

In this paper, we describe MARKUPMNA, a corpus of M&A agreements with annotated sections titles and certain other related information, curated from the HyperText Markup Language (HTML) versions of filings by US public companies in the Electronic Data Gathering, Analysis, and Retrieval (EDGAR) system of the Securities and Exchange Commission (SEC).

A natural segmentation of a document is given by its section hierarchy, which is typically present in long documents and often marked by special formatting/numbering of section titles. However, current NLP models typically process long documents without regard to section boundaries leading to a loss of topical coherence associated with such segmentation. For example, legal documents often exceed the maximum token limitations of many language models (though this limit has increased dramatically in recent models). While segmenting long documents into smaller chunks is commonly used to overcome these limitations, this is typically done either manually or in a manner that does not identify section titles or boundaries.[1] Accordingly, using MARKUPMNA to train models to identify section titles in a long document leads to a contextually coherent segmentation of long documents and makes progress towards developing a hierarchical representation of long text. As an illustration, Figure 1 displays ground truth labels and predictions by one of the models we benchmarked (MarkupLM) in an excerpt from an agreement in MARKUPMNA. Specifically, our main results are the following:

- We release MARKUPMNA, a corpus of 151 US public M&A agreements in HTML with annotated sections titles and certain other related information.
- We benchmark two open source language models jointly pretrained on markup and text information (MarkupLM and XDoc) on the section identification task. We observe the following transition in the learning dynamics as a function of the size of the training sets. When the training sets are small, these multimodal models learn to identify section titles but do not perform well on identifying their depth in the hierarchy of sections and subsections. When the training sets get larger, the model also learns the section hierarchy better.
- As an another ablation study, we have benchmarked RoBERTa, a transformer-based model trained using the plain text of the agreements. Our experiments show that HTML-based models outperform RoBERTa. Specifically, RoBERTa does not appear to learn the section hierarchy as well as the HTML-based models.
- When the text information is masked and the models are trained and tested on the HTML information only – *xpath* identifiers of the document object model (DOM) tree – the results are comparable to the models jointly trained and tested on the text and markup information. This is because the DOM tree captures the section title hierarchy when the section titles and other text are individually formatted with rich visual cues, which is generally the case with US public M&A agreements. We conjecture that for agreements that are not as well formatted as the US public M&A agreements, such as agreements involving smaller and/or different transactions, these trees will be "noisier" and the model will need to rely on the text along with the *xpath* identifiers.

EDGAR has millions of contracts and other long legal documents in HTML, and there is no existing pretrained open-source model that can perform extraction on legal documents of this kind. The specific agreements

[1] Optical character recognition (OCR) metadata and other visual features are sometimes used to improve performance of NLP systems and/or extract section hierarchy. However, the relevant models and datasets appear to be proprietary.

included in MarkupMnA in HTML have been previously included in the MAUD dataset in plain text [31]. The deal point extraction benchmark for the plain text MAUD document is quite challenging ($\sim 20\%$ area under the precision recall curve (AUPRC)). Through future work we hope to leverage segmentation and section titles to improve performance on various downstream extraction and language understanding tasks in the legal domain, such as the MAUD tasks. This would make progress towards unlocking EDGAR and other long legal data to NLP models. Accordingly, our principal conceptual advances are the following:

1. *We make progress towards addressing the lack of title detection datasets and benchmarks for EDGAR documents.* Section title detection in financial disclosures, including those from EDGAR, is part of the annual FinTOC Shared Task: "Financial Document Structure Extraction“. This task is quite similar to ours even though there are some technical differences. The state of the art F1 scores of 90% on the FinTOC title detection task for English prospectuses were achieved by an XGBoost-based classifier; it was trained using manually engineered visual, structural and text features [4, 1]. This score is comparable to the scores of our benchmarks, but the FinTOC datasets and models are not publicly available. Moreover the neural models we use do not require manual feature engineering, as we further discuss in the next paragraph.

2. *We develop a new section title extraction benchmark without manual feature engineering.* We pursue a data-driven approach as an alternative to the heuristic-based approaches in the literature. For example the rule-based construction of the LEDGAR dataset indicates that heuristic-based approaches require manual feature engineering and rejection of documents that do not conform to the heurstics (data cleaning) [29]. The need for a rejection mechanism suggests that when not all documents follow similar formatting rules, heuristic approaches may not generalize very well.

3. *We identify M&A agreements as a compelling use case for developing hierarchical representation of long text in legal domain and beyond.* US public M&A agreements, which tend to be long and well-formatted documents, represent a compelling use case to develop a hierarchical representation of long text, which can be generalized to other legal and non-legal settings. The existing M&A dataset (MAUD) annotates specific extraction regions for specific items of deal information in plain text, but does not provide annotations of section and subsection titles in these agreements. The existing general contract dataset LEDGAR provides individual sections as standalone documents, along with their titles, from various contracts, but does not comprehensively identity each section title for a given contract contained in this dataset.

4. *We make progress towards addressing the lack of HTML-based document layout datasets and any legal document layout datasets.* While there exist a number of document layout datasets and benchmarks using images of documents, we are not aware of any publicly available annotated HTML-based document layout datasets in legal or other domain (there are only HTML-based datasets focusing on specific non-legal information retrieval task from shorter webpages). Also there are no visually rich document layout datasets that focus on legal documents, especially contracts (in markup or image format). Legal documents tend to be much longer and visually different from the documents included in the existing layout datasets.

5. *We provide a starting point for developing a multimodal long document understanding pipeline.* MARKUPMNA (which provides section title annotations in HTML), together with MAUD (which provides deal point annotations in the plain text versions of the same agreements), provides a starting point for researchers interested in exploring multimodal models for long document understanding. Other existing document understanding/layout datasets tend to be shorter, are not multimodal, and/or do not include a benchmark for segmenting them into contextually coherent chunks. Multimodal long document understanding methods developed in the legal domain can be also adapted in other long document tasks, such as book summarization.

2 Related Work

Traditionally, NLP literature on contract analysis has focused on various information extraction tasks. More recent literature combines these tasks with reading comprehension / inference tasks. We briefly review the relevant literature, highlighting connections to section extraction and contract segmentation.

Reference [11] introduced one of the earliest annotated contract datasets. It contains 30 publicly available Australian contracts with annotated titles and clause headings, as well as several other common information extraction items, such as the parties and execution date. The contracts were converted from word into text files, which the authors noted had "the downside of discarding valuable style and layout information found in the word document format."

Reference [7] introduced a contract dataset with annotations of clause headings (993 contracts), as well as another set (2491 contracts) with annotated substantive provisions used for information extraction, such as title, contracting parties and effective date. The *test* set was *manually* segmented into various extraction zones, and specific provisions were extracted during testing from corresponding zones rather than from the full contracts. For example, title, contracting parties and effective date were extracted from the zone manually labelled as "cover page and preamble", while the governing law and jurisdiction information were extracted form the zone manually identified as "governing law" [6]. This dataset was used in subsequent work benchmarking neural models on the information extraction task [8]. Since the contracts in this dataset are nonpublic, to preserve confidentiality, they were provided in an encoded form where each vocabulary word is replaced by its integer identifier.[2]

LEDGAR appears to be the first annotated dataset containing individual clauses from various contracts available on EDGAR (a subset of them was included in the LexGLUE dataset, which also includes other types of legal documents) [29, 9]. The labels of the clauses were extracted from their titles using rules based on certain formatting patterns of HTML tags. Since some clauses and possibly other text were removed during data cleaning, it does not appear possible to automatically identify *all* clause/section titles or to otherwise *fully* segment each contract in this dataset.

Two recent datasets – Contract Understanding Atticus Dataset (CUAD) and Merger Agreement Understanding Dataset (MAUD) – are also based on contracts from EDGAR. CUAD contains 510 commercial contracts of 16 different types [17]. The annotations identify the spans of 41 categories of provisions, like governing law, document name and contracting parties; the extraction task is to identify the span of the provision corresponding to a given category. Since the CUAD dataset is benchmarked using the .txt versions of the contracts, the formatting and other visual information is disregarded. To address the length of these documents, a sliding window is used for various benchmarks. The leading model RoBERTa achieves area under the precision-recall curve (AUPRC) in excess of 40%.

Reference [16] argued that splitting long documents into segments in order to satisfy the maximum token limit of language models, such as the sliding window method mentioned above, leads to the loss of contextual information and negatively impacts the performance on the downstream tasks. Accordingly, they trained a model to split the .pdf versions of the CUAD contracts into individual sections using the text as well as the OCR metadata that captures the visual cues. This approach lead to improved performance on several information extraction and classification tasks. However, the splitting model and the dataset used to train it were not released publicly.

The MAUD dataset is based on 151 M&A agreements for recent acquisitions of US public companies exceeding $200 million in value [31].[3] Each agreement is associated with 7 *deal point categories* that describe how the parties are obligated to complete the transaction, such as Conditions to Closing. Each of these categories is further associated with one of 22 *deal point types*. For example, Absence of Litigation and Accuracy of Target Representations and Warranties deal point types are associated with the Conditions to Closing category. Annotators manually extracted clauses (*deal point texts*) corresponding to each deal point type. Based on this manual extraction, the reading comprehension task entails answering 92 multiple choice questions where each question corresponds to a single deal point type and unique deal point text per

[2] In light of this, it is unclear whether this dataset can be used to develop models for unencoded contracts.

[3] The dataset contains 152 contracts but `contract_148.txt` and `contract_149.txt` appear to be identical.

contract. While none of the deal point texts exceeds the 4096 token limit of the Big Bird model, it appears that further segmentation is necessary to benchmark the other (BERT-based) models, which have the maximum token limit of 512. All these models achieve AUPRC in excess of 50% on the reading comprehension task. On the other hand, the task of extracting the deal point texts from the full MAUD contracts is more challenging with the benchmark model achieving only 19.7% AUPRC. Although all of the MAUD contracts are available in HTML on EDGAR, this benchmark uses their plain text versions and therefore discards formatting information and other visual features.

Reference [20] introduced a dataset of 607 annotated nondisclosure agreements from EDGAR together with a document-level natural language inference (NLI) task. Given a set of hypotheses and a contract, this task entails determining whether each hypothesis is entailed by, contradicted by or not mentioned by the contract. Moreover, if the model determines entailment and contraction, the task further requires the model to extract the span of the evidence supporting such determination from the full contract. NLI is similar to the combination of extraction and reading comprehension tasks described in the context of the MAUD dataset. Instead of identifying the start and end tokens of the evidence, this reference poses evidence identification as a multilabel classification problem over *spans* - in this setting a span refer to a whole sentence. Moreover, the documents are segmented using a segmentation algorithm ensuring that there is at least one segment for each span where the span is not split and receives enough context. Since the segments may be overlapping, the final prediction is given by averaging over overlapping segments.

A similar approach is used in the context of book summarization: to overcome the maximum token limit of language models, the books are segmented randomly and the predictions are averaged over segmentation samples. In this setting, [33] reported that changing the seed in the randomization of segmentation boundaries leads to significant variation in the summaries.

Reference [12] introduced an unsupervised contract segmentation model using a clustering of lines into section titles and other types of text based on manually engineered formatting features of software procurement contracts. The datasets and models used in this work appear to be proprietary.

Section title and table of contents (TOC) extraction from financial disclosure prospectuses, including US public filings, are part of the annual FinToc Shared Task: "Financial Document Structure Extraction" competition at Financial Narrative Processing Workshops [19, 3, 25, 1]. While the section titles of the prospectuses in .pdf were annotated in the FinToc training and test sets, these datasets do not appear to be publicly available; also the models used in the competitions do not appear to be open sourced. The FinToc section title extraction task is closest to our section title extraction task. The principal difference is that in the context of FinToc each section title in the body of a prospectus needs to be extracted, while in our case, we also need to determine the level of hierarchy of each section title.[4]. The best F1 scores on the FinToc section title extraction task for prospectuses in English were achieved in 2021-22 by an XGBoost-based classifier: specifically the score of 81.8% was achieved using manually engineered visual and structural features of the documents [5, 25], and the score of 90% was achieved by additionally using text features [4, 1].

Section identification broadly falls under the rubric of document layout analysis. Most layout analysis datasets are comprised of images of documents, and we are not aware of layout analysis datasets for documents in HTML or other markup languages. Moreover, the documents in standard layout datasets, like PubLayNet and Docbank, are not similar to contracts [34, 22]. A more recent DocLayNet dataset includes legal statutes and financial disclosures, which are somewhat similar to contracts [27], but we are not aware of any layout analysis datasets that specifically include contracts.

3 MarkupMnA Dataset and Section Title Extraction Task

As the above discussion suggests, contextually coherent segmentation of long documents is an important task in the context of legal NLP. However, there exists no publicly available datasets or open source models for segmentation of contracts. We approach this problem by identifying section titles and numbers, which are

[4] The separate FinToc task of extracting the TOC, which is a single table apart from section titles located across the body of the document, requires classification of the hierarchy of the TOC entries.

commonly present in long contracts. In particular, public M&A agreements on EDGAR, which is our focus, are invariably organized into sections and subsections.

Accordingly, we release the MARKUPMNA dataset containing 151 public M&A agreements with annotated section titles and certain other related text elements, as discussed below. Since visual cues, such as formatting information helps identify sections titles, MARKUPMNA includes visually rich HTML versions of these agreements.

Although the analysis of visual cues and layout has been traditionally used with respect to .pdf version of contracts and other documents, we focus on HTML's for the following reasons.

- Visually rich documents are currently disclosed on EDGAR primarily in the HTML format, and our choice avoids errors and computational overhead associated with using an OCR software to extract text from .pdf and mapping the labels to the geometric coordinates of the corresponding pages.
- There appears to be relatively limited research on multimodal models using layout information in long HTML documents, and harnessing documents on the EDGAR database provides a compelling motivation for developing such datasets and models. There are many long documents, other than M&A agreements, in HTML on EDGAR and elsewhere on the Internet.
- The representation of an HTML document as a DOM tree, as described in Section 4.1 leads to a natural hierarchical representation of the underlying text when the sections and its titles/numbers are marked by special formatting patterns.
- We expect it will be possible to use our dataset to automatically annotate the .pdf versions of the corresponding contracts. On the other hand, it would not be as straightforward to go in the other direction, i.e., automatically label text in HTML nodes using labels associated with the geometric coordinates of the images of that text.

3.1 Data Source

EDGAR holds over 15 million filings by companies required to make public disclosures under US securities laws; it is the largest public repository of US legal documents. Among other things, public filers disclose material contracts and material events or corporate changes, which typically include significant mergers and acquisitions. Thus, EDGAR holds a large number of contracts, including M&A agreements. Their typical length is tens of thousands of words, which exceeds the maximum token size of most language models. EDGAR filings are mostly available in .htm and .txt, while some older filings are available in .pdf. Most recent M&A agreements are only available in .htm. Although many agreements contain a section hierarchy, none of the document format sources on EDGAR provides this segmentation in an explicit machine-readable format.

Legal and business professionals use contracts on EDGAR as “precedents” to draft and/or negotiate new contracts. Non-public contracts, which are often similar to those available on EDGAR, are analyzed by attorneys and business professions to understand potential sources of legal and commercial risk in M&A transactions and securities issuances; this process is referred to as “due diligence". Normally professionals identify the clauses relevant to their task (using the table of contents and/or section titles whenever they are available) and then interpret the meaning of each relevant clause.

The MARKUPMNA dataset contains M&A agreements sourced from EDGAR in HTML. We chose the same agreements as those that were included (in plain text) in the MAUD dataset in order to facilitate future work on understanding the effects of segmentation on reading comprehension and information extraction tasks.

3.2 Section Extraction Task

Given a contract in HTML, the task entails identifying each piece of displayable text as belonging to one of the following categories: (a) document number or title (an HTML file may contain multiple documents, e.g., several exhibits); (b) (sub) section number or title at varying depth; (c) page number; (d) or none of the above. We note that page numbers may be used to segment an HTML file into individual pages, e.g., for use with visio-linguistic models, instead of markup-based models benchmarked in this work.

3.3 Annotation and Statistics of the Dataset

We first investigated creating a rule-based workflow to label different categories of text in a given HTML document, similar to the approaches used to annotate the PubLayNet and DocBank datasets. The XML code of scientific papers and the LaTeXcode of papers on arXiv used in those datasets are essentially data structures containing layout features, which led to automated annotation workflows. On the other hand, HTML is a displaying format, which is not consistent across the M&A agreements and not easily amenable to a rule-based layout annotation.

Therefore, we annotated the dataset manually. This was done by a subset of authors without involving third parties. To expedite the labeling process, we have created a Javascript-based tagging tool, which enables an annotator to highlight a segment of text within a browser and assign an appropriate label.

The statistics of the dataset set is provided in Table 2. We use the BEIOS tagging scheme typically used in named entity recognition (NER)-style problems. We label *individual* xpath nodes as beginning, end, inside, or outside of each segment of interest. For example, if a given section title spans multiple nodes then the first node is labeled as B_ST, the last node is labeled E_ST, and all tokens between would be labeled I_ST. If instead a given section title spans a single token we label it as S_ST, where S stands for single. All other categories of text are labeled with the OUTSIDE label. See Appendix A for further details. Figure 6 shows an excerpt from a CSV file in the dataset corresponding to labels displayed in Figure 4. It illustrates the xpath identifiers and corresponding annotations provided by an annotator.

4 Benchmark and Experiments

As the multimodal baselines, we use the MarkupLM and XDoc models. MarkupLM is a transformer-based model pretrained in a self-supervised fashion jointly on text and HTML information of visually rich documents [21]. It models an HTML document as a document object model (DOM) tree where each string of text is associated with a so-called *xpath* identifier containing the sequence of HTML tag paths from the root of the document to the leaf that contains that string. This model adds the xpath embeddings to the BERT backbone architecture initialized with RoBERTa. During pre-training, jointly with the text embeddings, the model learns the embeddings of xpaths, which represent visual cues of the text, including its hierarchical position relative to other sections of text and style information, such as bold face font or italics.

XDoc is another recent multimodal transformer-based model. It was pretrained on text in three different formats: plain text, images of documents, and HTMLs [10]. This model also uses the BERT backbone initialized with RoBERTA, and shares the backbone parameters (the word embedding and transformer layers) across different input modalities. For our experiments, we only use the HTML modality.

We chose MarkupLM and XDoc as the multimodal base-lines because they are publicly available models jointly pretrained on text and markup information; as such they appear to be suitable for task-specific fine-tuning using a relatively small manually annotated dataset, like MARKUPMNA. Let us briefly mention a few other multimodal models that we did not benchmark. DOM-LM is a recent mulitmodal transformer-based model for htm documents that was pretrained in a self-supervised fashion similar to MarkupLM [13]. However, the original model does not appear to be open sourced.[5] A number of other multimodal neural models for HTML documents have been trained for specific tasks. For example, Webformer was trained for structured information extraction, such as information about events, products and movies [30]. This model does not appear to have been pre-trained jointly on markup information and text.

Finally, as the text-only ablation baseline, we use RoBERTa, a classic pretrained language representation models based on Bidirectional Encoder Representations from Transformers (BERT) architecture [24, 14].

4.1 Training

We randomly chose a test set of 20 agreements and an evaluation set of 10 agreements, and used the remaining 121 agreements as the test set. Using Adam, we fine-tune pre-trained MarkupLM and XDoc models on

[5] An implementation of DOM-LM at `https://github.com/Misterion777/DOM-LM` is unofficial.

Table 1: Results on MARKUPMNA section extraction task for various models. The outside labels are included in AUPRC but excluded from the F1, Precision and Recall metrics

Model	Macro Average				Weighted Avg			
	AUPRC	F1	Precision	Recall	AUPRC	F1	Precision	Recall
MarkupLM-base	0.5618	87.88%	81.20%	95.77%	0.9914	91.22%	87.08%	95.77%
MarkupLM-large	0.4662	80.42%	69.88%	94.71%	0.9864	86.14%	79.00%	94.71%
RoBERTa	0.4796	82.87%	74.14%	93.93%	0.9857	88.22%	83.17%	93.93%
XDoc	0.5423	89.02%	83.00%	95.98%	0.9903	91.76%	87.90%	95.98%
MarkupLM-xpath-masking	0.4199	80.78%	88.92%	93.70%	0.9754	83.30%	78.35%	88.92%
MarkupLM-xpath-shuffle	0.3319	76.23%	67.16%	88.14%	0.9648	79.14%	71.81%	88.14%
MarkupLM-text-masking	0.5695	86.80%	79.53%	95.54%	0.9900	90.67%	86.27%	95.54%

the section identification task posed as a multi-class classification problem (the RoBERTA model is also fine-tuned for text-only ablation experiments). These models are available on HuggingFace [32]. We use the BEIOS tagging scheme described above in the context of xpath node-based annotations for token-level predictions as well. For the hidden state associated with each token that we obtain during the forward pass, we have a classification head followed by a softmax function resulting in a probability distribution over all BEIOS labels across all categories. Since most of the tokens belong to the outside category, we use class weights in the softmax to address the class imbalance. This approach is similar to the training of NER models. We use a limited hyperparameter search described in Appendix B.1. For validation purposes, we use macro-average F1 score, which excludes the outside labels, as described below.

4.2 Metrics

Due to the class imbalance present in our dataset, we use F1, precision, recall and area under the precision-recall curve (AUPRC) to evaluate model performance (the latter metric is primarily used in the MAUD contract dataset [31]). Moreover, we exclude the *outside* label, which represents the largest class, for purposes of computing F1, precision and recall (but not AUPRC).

Macro Average vs Weighted Average To compute F1, precision, recall and AUPRC in a multi-class setting, we compute each of these metrics for each class and then average it across all classes (macro-average approach). We additionally compute our metrics by taking a weighted average over all classes rather than the macro average. The latter approach ensures that we do not underreport model performance due to poor performance on classes with fewer instances.

4.3 Results

In Table 1, we report the performance of the multimodal benchmarks MarkupLM-base and -large and XDoc as well as the xpath-only benchmark MarkupLM-text-masking. We also report the results of the text-only ablation models RoBERTa, MarkupLM-xpath-masking and -xpath-masking. These results demonstrate that MarkupLM and XDoc, as well as MarkupLM-text-masking, achieve a promising performance comparable to the performance of other models on legal NLP tasks in the literature, e.g., the AURPC metrics in [31], and outperform the text-only ablation models.

Macro-average over BEIS labels of individual categories We also report the F1, recall and precision metrics of MarkupLM-base for weighted average over the BEIS labels of individual classes in Figure 2. As the examination of this figure, and the confusion matrices in Figure 8 and 10 in the Appendix suggest, most of the errors involve (a) misclassification of the depth of section hierarchy, e.g., section titles of depth 3 (SSST) classified as those of depth 4 (SSSST), and (b) misclassification of labels that are relatively infrequent in the dataset (such as document titles (T), document title numbers (TN) and Section number and titles of depth 4 (SSSSN and SSSST).

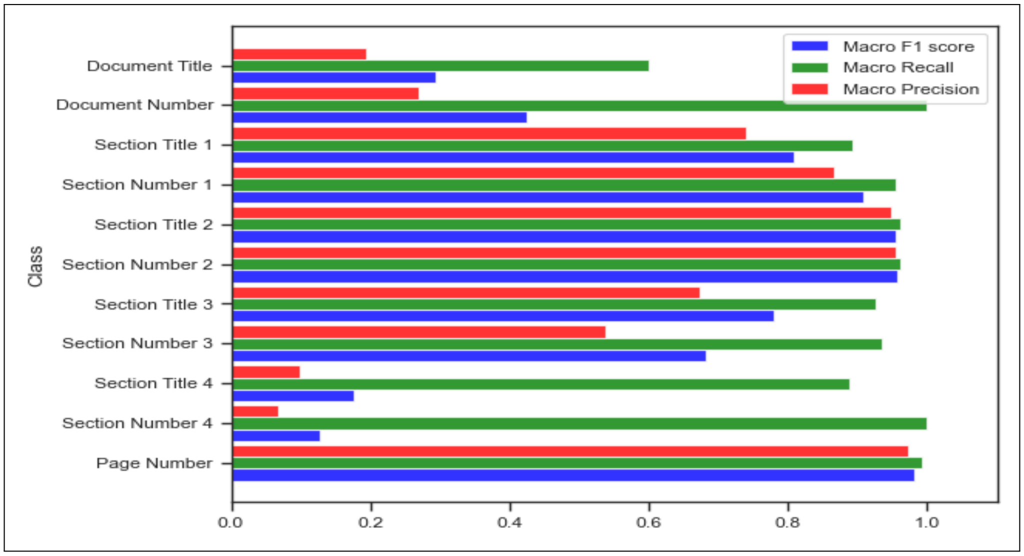


Figure 2: Category-wise F1, recall and precision of MarkupLM-base (using macro average over BEIS labels)

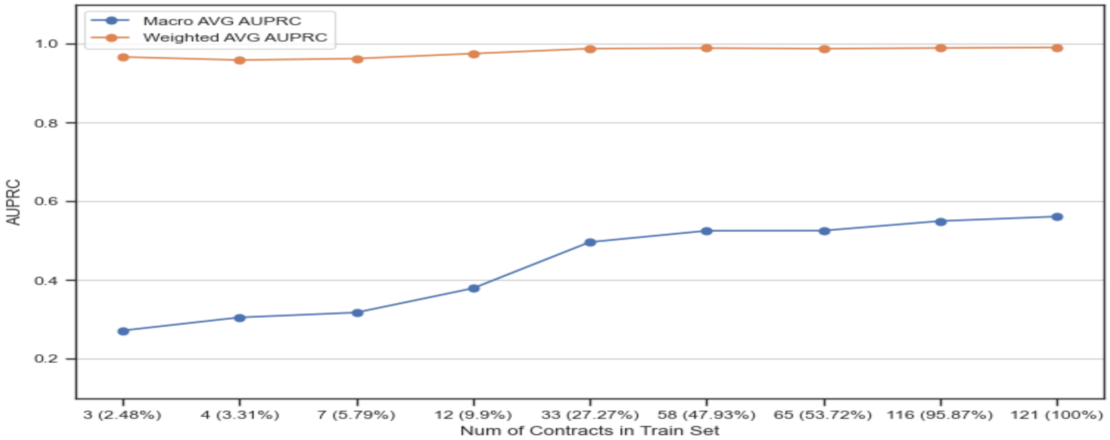


Figure 3: Macro and weighted AUPRC of MarkupLM-base as a function of the size of training dataset

Size of Training Data Figure 3 shows the effect of the number of training examples on MarkupLM-base performance. It can be seen that the performance increases rapidly when the model is trained on $\sim 3\%$ versus $\sim 50\%$ of the training data, and does not significantly improve further when the model is trained on 100% of the training data. We set the same number of max epochs for each running (50) using a patience of (5) for each to enable early stopping based on performance metrics on the validation set. Moreover, based on the confusion matrices in Appendix B.3, we see that when the training set is small (12 or fewer contracts), MarkupLM-base often misclassifies section titles as regular text (outside labels). When the training sets get larger (33 and 65 contracts) the model classifies sections titles as such, but often misclassifies their depth of hierarchy. Finally, when MarkupLM is trained on the full training set (121 contracts), the errors of the latter type subside as well.

xpath only model - MarkupLM-text-masking We also trained and tested MarkupLM (MarkupLM-text-masking) where masked the text information, so that the model only used the HTML information – *xpath* identifiers of the document object model (DOM) tree. As shown in Table 1 and Figure 11, the performance of MarkupLM-text-masking is comparable to the models jointly trained and tested on the text and markup information (MarkupLM-base and XDoc). This is because the DOM tree captures the section title hierarchy when the section titles and other text are individually formatted with rich visual cues, which is generally the case with US public M&A agreements. We conjecture that for agreements that are not as well formatted as the US public M&A agreements, such as agreements involving smaller and/or different transactions, these trees will be "noisier" and the model will need to rely on the text along with the *xpath* identifiers.

Text-only ablation To investigate the effect of encoding the DOM tree structure on model performance, we also perform ablation experiments using MarkupLM-base and RoBERTA. In the experiments using MarkupLM-base, we nullify the contribution of the xpath embeddings, first, by masking the xpath embedding (MarkupLM-xpath-masking) and second, by randomly shuffling the xpath embeddings in order to break the structure of the DOM tree (MarkupLM-xpath-shuffle). In each case, we follow this approach during both training and test time. These results can be found in Table 1. In both cases, the performance of the models reduces, pointing to the value of encoding the structure of the DOM tree in this prediction task. In Appendix B.2, we also report results from our experiments using constrained decoding during test time. This somewhat improves the performance of MarkupLM-xpath-masking, but this improvement is not sufficient to compensate for the loss of performance due to the loss of the xpath information.

As an another ablation study, we benchmark RoBERTa, a transformer-based model, which is the backbone of MarkupLM and XDoc, using the plain text of the MARKUPMNA agreements. As shown in Table 1, the HTML-based models outperform RoBERTa. As the confusion matrix in Figure 9 shows, RoBERTAa misindentifies the depth of the section hierarchy more often than the HTML-based models (MarkupLM-base, XDoc and MarkupLM-text-masking).

5 Conclusion and Future Work

We introduced a dataset and benchmark for identifying section titles and certain other related items in public M&A agreements. Our results are promising and point to the potential to apply our pipeline to a great variety of long documents within and beyond the legal domain. This work is a necessary intermediate step before we can evaluate the utility of segmentation in a variety of downstream tasks in legal and non-legal NLP, such as the MAUD deal point extraction task that is currently benchmarked only at 20% AUPRC [31], determination of spans / candidate extraction regions in [20]; we also hope that segmentation based on section titles would provide an alternative over the sliding window segmentation in [17]. One possible framework for evaluating the utility of segmentation is Rissanen Data Analysis, which has been used in a variety of settings in NLP, such as evaluating the utility of generating subquestions before answering a question and analyzing the value of rationales and explanations [26].

We have only fine-tuned pre-trained models on section extraction tasks using MARKUPMNA and have not conducted extensive in-domain pre-training since we achieve relatively high scores using fine-tuning alone. A fruitful direction for future research would be to understand whether pretraining on a much larger than MARKUPMNA corpora of unlabelled contracts from EDGAR and/or fine-tuning on weakly annotated data through the use of heuristics-based labelling functions (for example automatically annotated sections of contracts contained in the LEDGAR dataset) would improve the title extraction benchmark. In this setting the entire MARKUPMNA could be treated as a test set.

The broader layout analysis problem entails capturing all formatting and layout information, beyond the special formatting of sections and subsections, such as paragraphs, headers, footnotes, tables (most importantly the table of contents), etc. Solving this problem requires a more extensive labeling scheme to capture all the relevant layout and formatting information. We leave this potential extension to future work.

While a number of current models, like GPT-4, are multimodal, they typically use visual information, which in our setting would require rendering the HTMLs into images. We expect that an automated pipeline could be used to convert the HTML versions of the contracts in MARKUPMNA into a PDF and associate each annotation with the page number and geometric coordinates of the corresponding PDF page. Therefore, we expect that our dataset could also be used to benchmark visio-linguistic models on segmentation tasks.

Acknowledgments and Disclosure of Funding

V.A.K. acknowledges support by NSF grant DMS-1937254. S.R., P.I., R.B., S.K., and V.K. thank Julia Kempe for advising on an earlier project that led to this work. Also many thanks to Clifton King and David Chen for assisting with engineering efforts.

References

[1] A. Ait Azzi, S. Bellato, B. Carbajo Coronado, M. El-Haj, I. El Maarouf, M. Gan, A. Gisbert, J. Kang, and A. Moreno Sandoval. The financial document structure extraction shared task (FinTOC 2022). In *Proceedings of the 4th Financial Narrative Processing Workshop @LREC2022*, pages 83–88. European Language Resources Association, 2022.

[2] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

[3] N.-I. Bentabet, R. Juge, I. El Maarouf, V. Mouilleron, D. Valsamou-Stanislawski, and M. El-Haj. The financial document structure extraction shared task (FinToc 2020). In *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*, pages 13–22. COLING, 2020.

[4] A. Bogatenkova, O. Belyaeva, A. Perminov, and I. Kozlov. ISPRAS@FinTOC-2022 shared task: Two-stage TOC generation model. In *Proceedings of the 4th Financial Narrative Processing Workshop @LREC2022*, pages 89–94. European Language Resources Association, 2022.

[5] C. Bourez. FinTOC2021 - document structure understanding. In *Proceedings of the 3rd Financial Narrative Processing Workshop*, pages 89–93. Association for Computational Linguistics, 2021.

[6] I. Chalkidis and I. Androutsopoulos. A deep learning approach to contract element extraction. In *Proceedings of the 30th International Conference on Legal Knowledge and Information Systems (JURIX)*, pages 155–164, 2017.

[7] I. Chalkidis, I. Androutsopoulos, and A. Michos. Extracting contract elements. In *Proceedings of the 16th Edition of the International Conference on Articial Intelligence and Law*, ICAIL '17, page 19–28. Association for Computing Machinery, 2017.

[8] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, and I. Androutsopoulos. Neural contract element extraction revisited. In *Workshop on Document Intelligence at NeurIPS 2019*, 2019. URL `https://openreview.net/forum?id=B1x6fa95UH`.

[9] I. Chalkidis, A. Jana, D. Hartung, M. Bommarito, I. Androutsopoulos, D. Katz, and N. Aletras. LexGLUE: A benchmark dataset for legal language understanding in English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330. Association for Computational Linguistics, 2022.

[10] J. Chen, T. Lv, L. Cui, C. Zhang, and F. Wei. XDoc: Unified pre-training for cross-format document understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1006–1016, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.71. URL `https://aclanthology.org/2022.findings-emnlp.71`.

[11] M. Curtotti and E. McCreath. Corpus based classification of text in Australian contracts. In *Proceedings of the Australasian Language Technology Association Workshop 2010*, pages 18–26, 2010.

[12] X.-H. Dang, R. Akella, S. Bahrami, V. Sheinin, and P. Zerfos. Unsupervised threshold autoencoder to analyze and understand sentence elements. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 3267–3276, 2018.

[13] X. Deng, P. Shiralkar, C. Lockard, B. Huang, and H. Sun. DOM-LM: Learning generalizable representations for HTML documents, 2021. URL `https://arxiv.org/pdf/2201.10608.pdf`.

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL `https://aclanthology.org/N19-1423`.

[15] European Organization For Nuclear Research and OpenAIRE. Zenodo, 2013. URL `https://www.zenodo.org/`.

[16] A. Hegel, M. Shah, G. Peaslee, B. Roof, and E. Elwany. The law of large documents: Understanding the structure of legal contracts using visual cues. In *Document Intelligence Workshop at KDD*, 2021.

[17] D. Hendrycks, C. Burns, A. Chen, and S. Ball. Cuad: An expert-annotated nlp dataset for legal contract review. In J. Vanschoren and S. Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1. Curran, 2021.

[18] S. Holland, A. Hosny, S. Newman, J. Joseph, and K. Chmielinski. The dataset nutrition label: A framework to drive higher data quality standards. *arXiv preprint arXiv:1805.03677*, 2018.

[19] R. Juge, I. Bentabet, and S. Ferradans. The FinTOC-2019 shared task: Financial document structure extraction. In *Proceedings of the Second Financial Narrative Processing Workshop (FNP 2019)*, pages 51–57. Linköping University Electronic Press, 2019.

[20] Y. Koreeda and C. Manning. ContractNLI: A dataset for document-level natural language inference for contracts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1907–1919, 2021.

[21] J. Li, Y. Xu, L. Cui, and F. Wei. MarkupLM: Pre-training of text and markup language for visually rich document understanding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6078–6087, 2022.

[22] M. Li, Y. Xu, L. Cui, S. Huang, F. Wei, Z. Li, and M. Zhou. DocBank: A benchmark dataset for document layout analysis. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 949–960. International Committee on Computational Linguistics, 2020.

[23] R. Liaw, E. Liang, R. Nishihara, P. Moritz, J. E. Gonzalez, and I. Stoica. Tune: A research platform for distributed model selection and training. *arXiv preprint arXiv:1807.05118*, 2018.

[24] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach, 2019. URL `https://arxiv.org/abs/1907.11692`.

[25] I. E. Maarouf, J. Kang, A. A. Azzi, S. Bellato, M. Gan, and M. El-Haj. The financial document structure extraction shared task (FinTOC2021). In *Proceedings of the 3rd Financial Narrative Processing Workshop*, pages 111–119. Association for Computational Linguistics, 2021.

[26] E. Perez, D. Kiela, and K. Cho. Rissanen Data Analysis: Examining dataset characteristics via description length. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8500–8513. PMLR, 18–24 Jul 2021. URL `https://proceedings.mlr.press/v139/perez21a.html`.

[27] B. Pfitzmann, C. Auer, M. Dolfi, A. S. Nassar, and P. Staar. Doclaynet: A large human-annotated dataset for document-layout segmentation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 3743–3751, 2022.

[28] H. Pham, G. Wang, Y. Lu, D. Florencio, and C. Zhang. Understanding long documents with different position-aware attentions, 2022. URL `https://arxiv.org/abs/2208.08201`.

[29] D. Tuggener, P. von Däniken, T. Peetz, and M. Cieliebak. LEDGAR: A large-scale multi-label corpus for text classification of legal provisions in contracts. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1235–1241. European Language Resources Association, 2020.

[30] Q. Wang, Y. Fang, A. Ravula, F. Feng, X. Quan, and D. Liu. Webformer: The web-page transformer for structure information extraction. In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 3124–3133, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450390965. doi: 10.1145/3485447.3512032. URL `https://doi.org/10.1145/3485447.3512032`.

[31] S. H. Wang, A. Scardigli, L. Tang, W. Chen, D. Levkin, A. Chen, S. Ball, T. Woodside, O. Zhang, and D. Hendrycks. MAUD: An expert-annotated legal NLP dataset for merger agreement understanding, 2023. URL `https://arxiv.org/abs/2301.00876`.

[32] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, Oct. 2020. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/2020.emnlp-demos.6`.

[33] J. Wu, L. Ouyang, D. M. Ziegler, N. Stiennon, R. Lowe, J. Leike, and P. Christiano. Recursively summarizing books with human feedback, 2021. URL `https://arxiv.org/pdf/2109.10862.pdf`.

[34] X. Zhong, J. Tang, and A. J. Yepes. Publaynet: Largest dataset ever for document layout analysis. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1015–1022. IEEE Computer Society, 2019.

A Dataset details

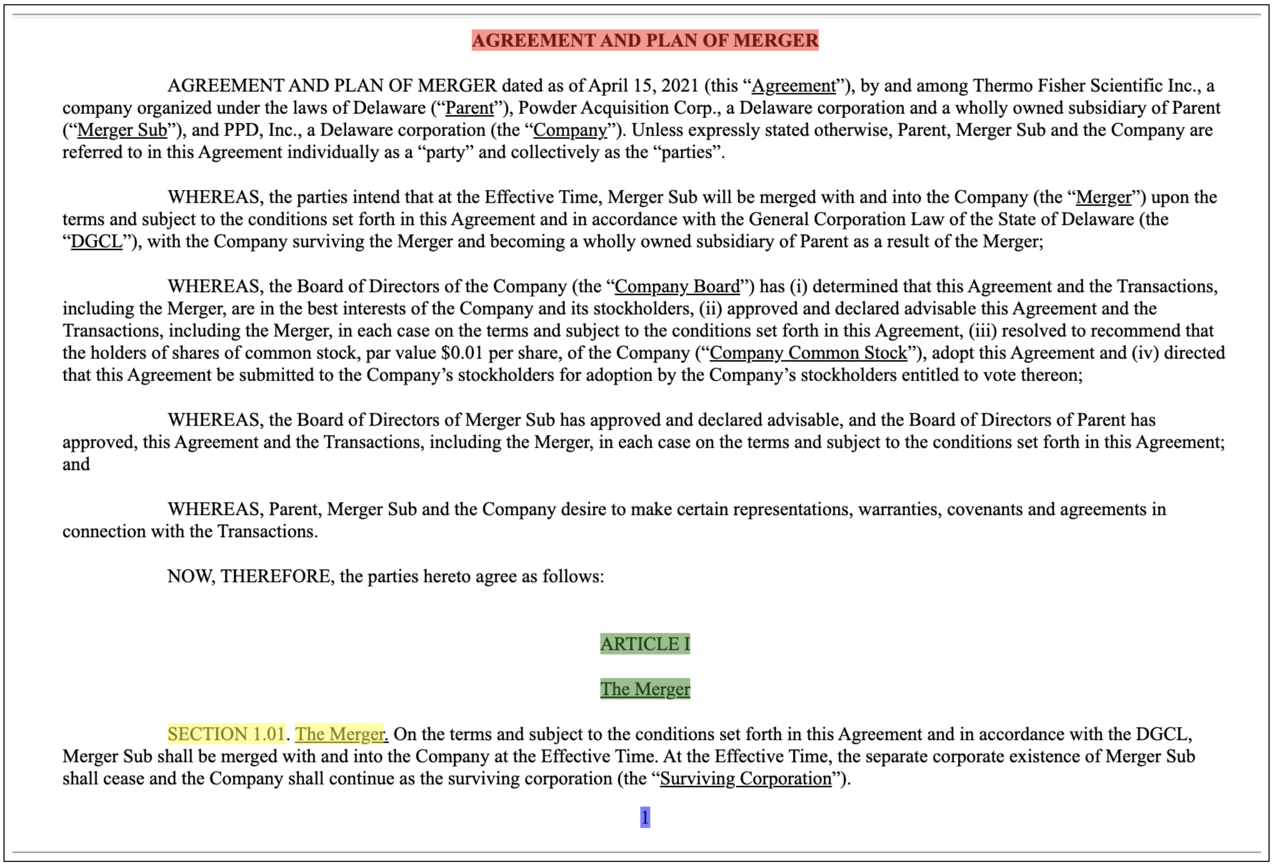
We obtained the HTML versions of the agreements corresponding to the .txt versions included in the MAUD dataset by manually searching for them in the SEC EDGAR database. Next, to create section information annotations, we created a custom labeling tool that ran in an annotator's web browser. Using this tool, we highlighted nodes corresponding to page numbers and section titles at varying depths. The output of this tool included two JSON files, one containing all the nodes in a given contract - both the ones highlighted by a user and the ones not highlighted, and a second containing only the nodes highlighted by a user. In the final step, we ran a custom Python script that merged the highlighted node JSON with the all nodes JSON and created the BEIOS labels for a given category resulting in the CSV files that comprise this dataset. See Figure 6 as an example. The statistics of the dataset set is provided in Table 2.

Figures 4 and 5 show labels of a sample MARKUPMNA agreement, and Figure 6 shows an excerpt from a CSV file containing the annotations. Figure 7 shows the corresponding data schema of these CSV files.[6]

[6] Our schema format is a simplified version of the format described in `https://pypi.org/project/csv-schema/`.

Table 2: Statistics of the 151 documents in MarkupMnA. The per category figures refer to the number of each type of label category found in a single document. (1)-(4) refer to the depth of the respective (sub)section in the section hierarchy. Total tokens refer to the total word pieces in a document.

Category	Legend	Average	Min	Max
Document number	**TN**	1.22	0	11
Document title	**T**	12.49	6	61
Section number (1)	**SN**	12.57	6	81
Section title (1)	**ST**	20.46	9	202
Section number (2)	**SSN**	108.59	0	350
Section title (2)	**SST**	110.01	54	347
Section number (3)	**SSSN**	53.88	0	237
Section title (3)	**SSST**	34.88	0	235
Section number (4)	**SSSSN**	3.42	0	96
Section title (4)	**SSSST**	1.43	0	30
Page number	**N**	109.59	0	325
None of the above	**O**	2612.92	1177	8478
Total tokens		74693.53	43020	220795



AGREEMENT AND PLAN OF MERGER

AGREEMENT AND PLAN OF MERGER dated as of April 15, 2021 (this "Agreement"), by and among Thermo Fisher Scientific Inc., a company organized under the laws of Delaware ("Parent"), Powder Acquisition Corp., a Delaware corporation and a wholly owned subsidiary of Parent ("Merger Sub"), and PPD, Inc., a Delaware corporation (the "Company"). Unless expressly stated otherwise, Parent, Merger Sub and the Company are referred to in this Agreement individually as a "party" and collectively as the "parties".

WHEREAS, the parties intend that at the Effective Time, Merger Sub will be merged with and into the Company (the "Merger") upon the terms and subject to the conditions set forth in this Agreement and in accordance with the General Corporation Law of the State of Delaware (the "DGCL"), with the Company surviving the Merger and becoming a wholly owned subsidiary of Parent as a result of the Merger;

WHEREAS, the Board of Directors of the Company (the "Company Board") has (i) determined that this Agreement and the Transactions, including the Merger, are in the best interests of the Company and its stockholders, (ii) approved and declared advisable this Agreement and the Transactions, including the Merger, in each case on the terms and subject to the conditions set forth in this Agreement, (iii) resolved to recommend that the holders of shares of common stock, par value $0.01 per share, of the Company ("Company Common Stock"), adopt this Agreement and (iv) directed that this Agreement be submitted to the Company's stockholders for adoption by the Company's stockholders entitled to vote thereon;

WHEREAS, the Board of Directors of Merger Sub has approved and declared advisable, and the Board of Directors of Parent has approved, this Agreement and the Transactions, including the Merger, in each case on the terms and subject to the conditions set forth in this Agreement; and

WHEREAS, Parent, Merger Sub and the Company desire to make certain representations, warranties, covenants and agreements in connection with the Transactions.

NOW, THEREFORE, the parties hereto agree as follows:

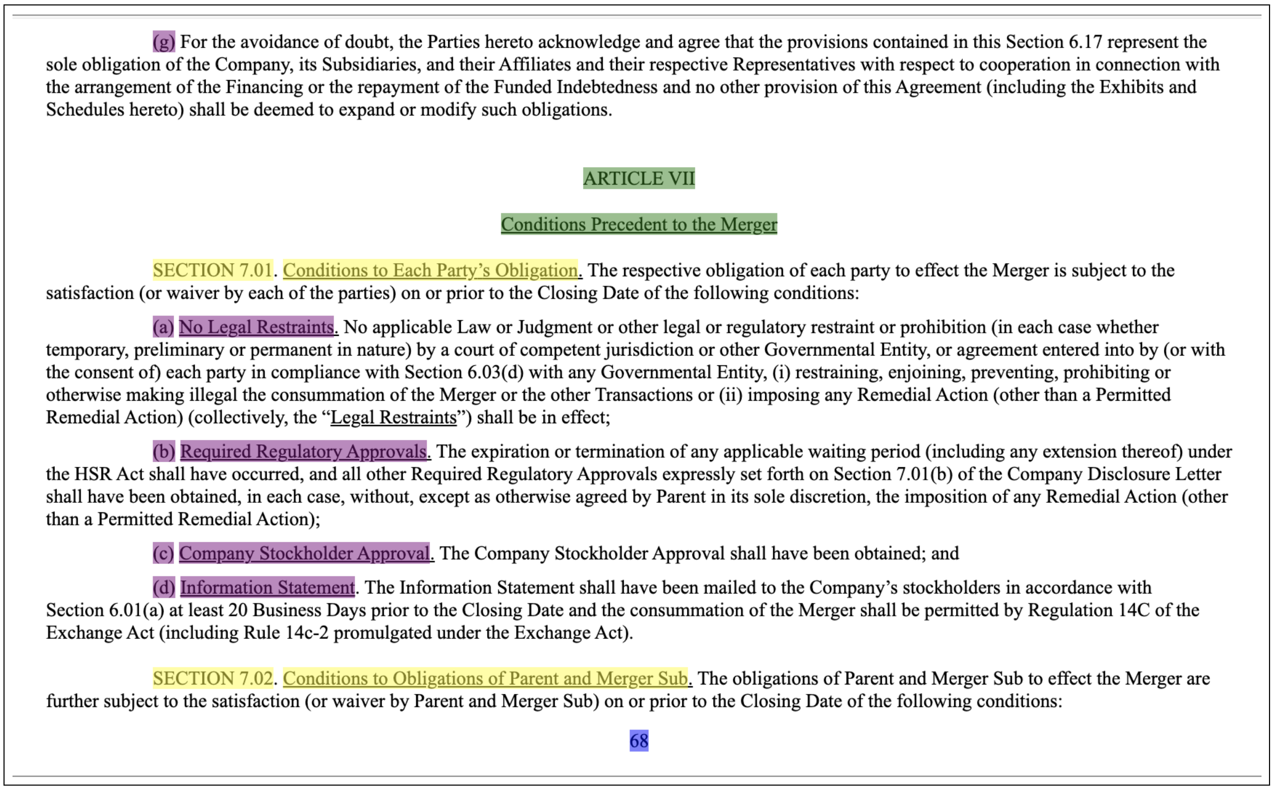
ARTICLE I

The Merger

SECTION 1.01. The Merger. On the terms and subject to the conditions set forth in this Agreement and in accordance with the DGCL, Merger Sub shall be merged with and into the Company at the Effective Time. At the Effective Time, the separate corporate existence of Merger Sub shall cease and the Company shall continue as the surviving corporation (the "Surviving Corporation").

1

Figure 4: The first page of a sample MARKUPMNA agreement with ground truth labels for document titles, section titles of depth 1 and 2 contained `contract_104.csv` displayed by red, green and yellow highlight, respectively.



(g) For the avoidance of doubt, the Parties hereto acknowledge and agree that the provisions contained in this Section 6.17 represent the sole obligation of the Company, its Subsidiaries, and their Affiliates and their respective Representatives with respect to cooperation in connection with the arrangement of the Financing or the repayment of the Funded Indebtedness and no other provision of this Agreement (including the Exhibits and Schedules hereto) shall be deemed to expand or modify such obligations.

ARTICLE VII

Conditions Precedent to the Merger

SECTION 7.01. Conditions to Each Party's Obligation. The respective obligation of each party to effect the Merger is subject to the satisfaction (or waiver by each of the parties) on or prior to the Closing Date of the following conditions:

(a) No Legal Restraints. No applicable Law or Judgment or other legal or regulatory restraint or prohibition (in each case whether temporary, preliminary or permanent in nature) by a court of competent jurisdiction or other Governmental Entity, or agreement entered into by (or with the consent of) each party in compliance with Section 6.03(d) with any Governmental Entity, (i) restraining, enjoining, preventing, prohibiting or otherwise making illegal the consummation of the Merger or the other Transactions or (ii) imposing any Remedial Action (other than a Permitted Remedial Action) (collectively, the "Legal Restraints") shall be in effect;

(b) Required Regulatory Approvals. The expiration or termination of any applicable waiting period (including any extension thereof) under the HSR Act shall have occurred, and all other Required Regulatory Approvals expressly set forth on Section 7.01(b) of the Company Disclosure Letter shall have been obtained, in each case, without, except as otherwise agreed by Parent in its sole discretion, the imposition of any Remedial Action (other than a Permitted Remedial Action);

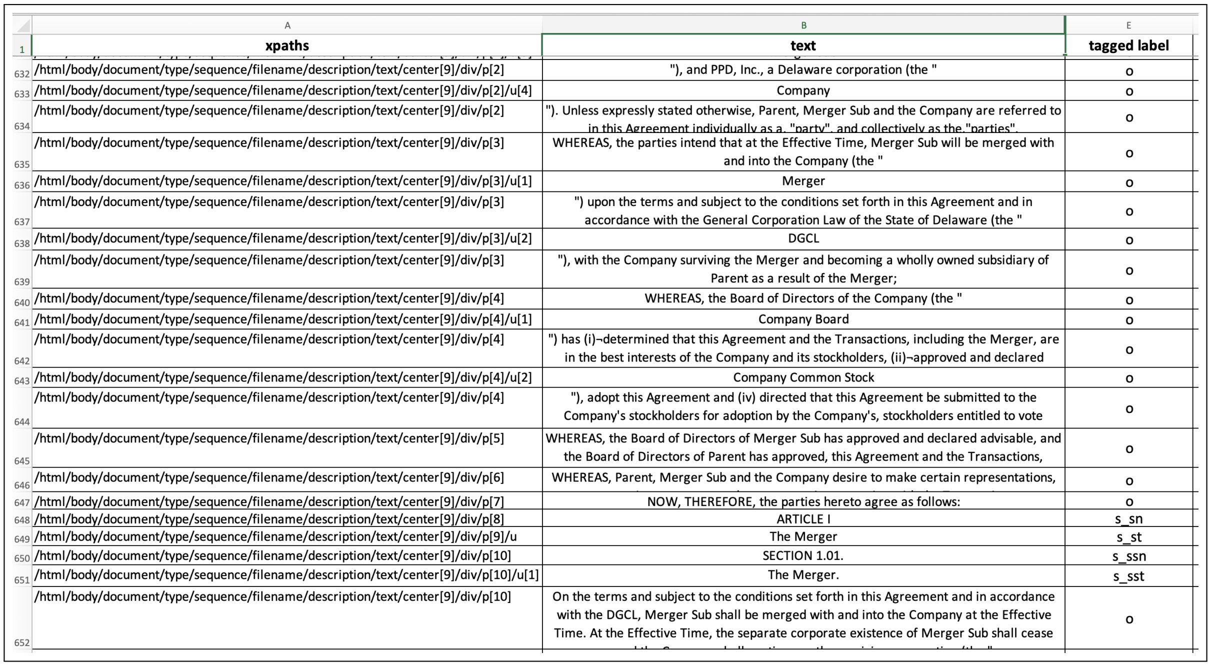
(c) Company Stockholder Approval. The Company Stockholder Approval shall have been obtained; and

(d) Information Statement. The Information Statement shall have been mailed to the Company's stockholders in accordance with Section 6.01(a) at least 20 Business Days prior to the Closing Date and the consummation of the Merger shall be permitted by Regulation 14C of the Exchange Act (including Rule 14c-2 promulgated under the Exchange Act).

SECTION 7.02. Conditions to Obligations of Parent and Merger Sub. The obligations of Parent and Merger Sub to effect the Merger are further subject to the satisfaction (or waiver by Parent and Merger Sub) on or prior to the Closing Date of the following conditions:

68

Figure 5: Another page from the sample agreement excerpted in Figure 4 with ground truth labels representing the section hierarchy of depth 1, 2 and 3 shown by green, yellow and purple highlight, respectively.



	A	B	E
1	xpaths	text	tagged label
632	/html/body/document/type/sequence/filename/description/text/center[9]/div/p[2]	"), and PPD, Inc., a Delaware corporation (the "	o
633	/html/body/document/type/sequence/filename/description/text/center[9]/div/p[2]/u[4]	Company	o
634	/html/body/document/type/sequence/filename/description/text/center[9]/div/p[2]	"). Unless expressly stated otherwise, Parent, Merger Sub and the Company are referred to in this Agreement individually as a "party", and collectively as the "parties".	o
635	/html/body/document/type/sequence/filename/description/text/center[9]/div/p[3]	WHEREAS, the parties intend that at the Effective Time, Merger Sub will be merged with and into the Company (the "	o
636	/html/body/document/type/sequence/filename/description/text/center[9]/div/p[3]/u[1]	Merger	o
637	/html/body/document/type/sequence/filename/description/text/center[9]/div/p[3]	") upon the terms and subject to the conditions set forth in this Agreement and in accordance with the General Corporation Law of the State of Delaware (the "	o
638	/html/body/document/type/sequence/filename/description/text/center[9]/div/p[3]/u[2]	DGCL	o
639	/html/body/document/type/sequence/filename/description/text/center[9]/div/p[3]	"), with the Company surviving the Merger and becoming a wholly owned subsidiary of Parent as a result of the Merger;	o
640	/html/body/document/type/sequence/filename/description/text/center[9]/div/p[4]	WHEREAS, the Board of Directors of the Company (the "	o
641	/html/body/document/type/sequence/filename/description/text/center[9]/div/p[4]/u[1]	Company Board	o
642	/html/body/document/type/sequence/filename/description/text/center[9]/div/p[4]	") has (i)¬determined that this Agreement and the Transactions, including the Merger, are in the best interests of the Company and its stockholders, (ii)¬approved and declared	o
643	/html/body/document/type/sequence/filename/description/text/center[9]/div/p[4]/u[2]	Company Common Stock	o
644	/html/body/document/type/sequence/filename/description/text/center[9]/div/p[4]	"), adopt this Agreement and (iv) directed that this Agreement be submitted to the Company's stockholders for adoption by the Company's, stockholders entitled to vote	o
645	/html/body/document/type/sequence/filename/description/text/center[9]/div/p[5]	WHEREAS, the Board of Directors of Merger Sub has approved and declared advisable, and the Board of Directors of Parent has approved, this Agreement and the Transactions,	o
646	/html/body/document/type/sequence/filename/description/text/center[9]/div/p[6]	WHEREAS, Parent, Merger Sub and the Company desire to make certain representations,	o
647	/html/body/document/type/sequence/filename/description/text/center[9]/div/p[7]	NOW, THEREFORE, the parties hereto agree as follows:	o
648	/html/body/document/type/sequence/filename/description/text/center[9]/div/p[8]	ARTICLE I	s_sn
649	/html/body/document/type/sequence/filename/description/text/center[9]/div/p[9]/u	The Merger	s_st
650	/html/body/document/type/sequence/filename/description/text/center[9]/div/p[10]	SECTION 1.01.	s_ssn
651	/html/body/document/type/sequence/filename/description/text/center[9]/div/p[10]/u[1]	The Merger.	s_sst
652	/html/body/document/type/sequence/filename/description/text/center[9]/div/p[10]	On the terms and subject to the conditions set forth in this Agreement and in accordance with the DGCL, Merger Sub shall be merged with and into the Company at the Effective Time. At the Effective Time, the separate corporate existence of Merger Sub shall cease	o

Figure 6: Excerpt from MARKUPMNA file `contract_104.csv` with ground truth labels corresponding to Figure 4.

```
{
    "columns": [
        {"name": "xpaths", "type": "string"},
            // xpath of each node, includes both highlighted nodes (corresponding to section information
            // labels) and nodes not highlighted by the annotator (corresponding to "outside" labels)

        {"name": "text", "type": "string"},
            // text contained in each node, includes both highlighted nodes (corresponding to section
            // information labels) and nodes not highlighted by the annotator (outside labels)

        {"name": "highlighted_xpaths", "type": "string"},
            // the xpath corresponding to the nodes if it is highlighted by the annotator, i.e.,
            // corresponds to a section information label

        {"name": "highlighted_segmented_text", "type": "string"},
            // the text contained in a node highlighted by the annotator

        {"name": "tagged_sequence", "type": "string"}
            // the label provided by the annotator (or outside label if the node was not manually annotated)
    ],
    "delimiter": ","
}
```

Figure 7: CSV schema of MarkupMnA

B Experimental details

B.1 Hyperpameter search

Due to time constraints, we performed a limited hyperparameter search over the learning rate (loguniform sampling over range (5e-5 to 5e-2)) and the number of epochs (2, 4, 6, 8, 10). The experiments were implemented using the Ray framework[23] using the Optuna search algorithm [2]. Based on the results we proceeded with using a learning rate of 2e-5 for our experiments. We leave an extensive hyperparameter search to future work.

Table 3: Training hyperparameters for all models.
**Other models* refers to MarkupLM Base, Xpath-Masking, Xpath-Shuffling and Text Masking, and RoBERTA **Effective batch size achieved using gradient accumulation of 4

Hyperparam	**MarkupLM** Large	**Other models***
Learning Rate	2e-5	2e-5
Batch Size**	32	8
Max Epochs	50	50
Patience	5	5
Weight Decay	0.01	0.01
Adam β_2	0.999	0.999
Adam ϵ	1e-8	1e-8

B.2 Constrained Decoding

Since we are dealing with a sequence labeling task, a priori, we know that there are certain predicted sequences that are not allowed. For example, if the model classifies a given token as the beginning of a section title (B_SEC1), we know that the next token must be either an inside or end of a section title (I_SEC1, E_SEC1). We apply these constraints during test time alone and report the corresponding performance metrics in Table 4 and 5. We observe that the performance metrics seem to drop when applying this logic. This could be due to the propagation of errors when the model makes a false positive prediction. For example if the model incorrectly predicts an outside class as a beginning of title (b_t) label, then we force the model to predict the next label as either inside or end of a title (i.e i_t, e_t). However, the next label may very well be an outside label itself.

Table 4: Comparison of results obtained using normal decoding versus results obtained using constrained decoding where we applied a constraint on the possible output classes depending on the predicted class from the previous timestep. These results were obtained using a macro-average over all classes.

Macro Avg	**Normal Decoding**				**Constrained Decoding**			
	AUPRC	F1	Precision	Recall	AUPRC	F1	Precision	Recall
MarkupLM-base	0.5618	87.88%	81.20%	95.77%	0.5317	83.73%	77.44%	91.12%
MarkupLM-large	0.4662	80.42%	69.88%	94.71%	0.4208	72.23%	63.56%	83.66%
MarkupLM-xpath-masking	0.4199	80.78%	74.00%	88.92%	0.3041	83.07%	78.68%	87.98%
MarkupLM-xpath-shuffle	0.3319	76.23%	67.16%	88.14%	0.2822	68.49%	62.19%	76.20%
MarkupLM-text-masking	0.5695	86.80%	79.53%	95.54%	0.5336	82.69%	76.21%	90.38%

Table 5: Comparison of results obtained using normal decoding versus results obtained using constrained decoding where we applied a constraint on the possible output classes depending on the predicted class from the previous timestep. These results were obtained using a weighted-average over all classes.

Weighted Avg	**Normal Decoding**				**Constrained Decoding**			
	AUPRC	F1	Precision	Recall	AUPRC	F1	Precision	Recall
MarkupLM-base	0.9914	91.22%	87.08%	95.77%	0.9907	91.29%	87.45%	95.48%
MarkupLM-large	0.9864	86.14%	79.00%	94.71%	0.9823	86.00%	80.08%	92.86%
MarkupLM-xpath-masking	0.9754	83.30%	78.35%	88.92%	0.9654	88.40%	85.65%	91.33%
MarkupLM-xpath-shuffle	0.9648	79.14%	71.81%	88.14%	0.9525	73.76%	82.28%	77.79%
MarkupLM-text-masking	0.9900	90.67%	86.27%	95.54%	0.9889	90.52%	86.46%	94.99%

B.3 Confusion matrices - BEIS labels corresponding to a single category are treated as one

As the statistics of the dataset indicates, certain classes, like document titles and section titles of hierarchy depth 4, are relatively infrequent in the training data (relative to section titles of smaller depth and page numbers). Therefore, the model may have a hard time distinguishing between different BEIS labels within that category (which are treated as different categories for purposes of Table 1, Figure 2, Figure 3 and Tables 4 and 5). In this section, we report confusion matrices using a slightly different error metric: the individual BEIS labels within a given category, e.g., section titles of give hierarchy depth, are treated as a single label. Therefore, the misidentification of the B vs E vs I vs S labels during inference on the test set does not contribute to error if the labels belong to the same category.

Full training set The first set of confusion matrices reports the inference results for MarkupLM-base (Figure 8), RoBERTAa (Figure 9), XDoc (Figure 10) and MarkupLM-text-masking model (Figure 11), each trained on the entire training set of 121 contracts. These matrices are discussed in Section 4.3.

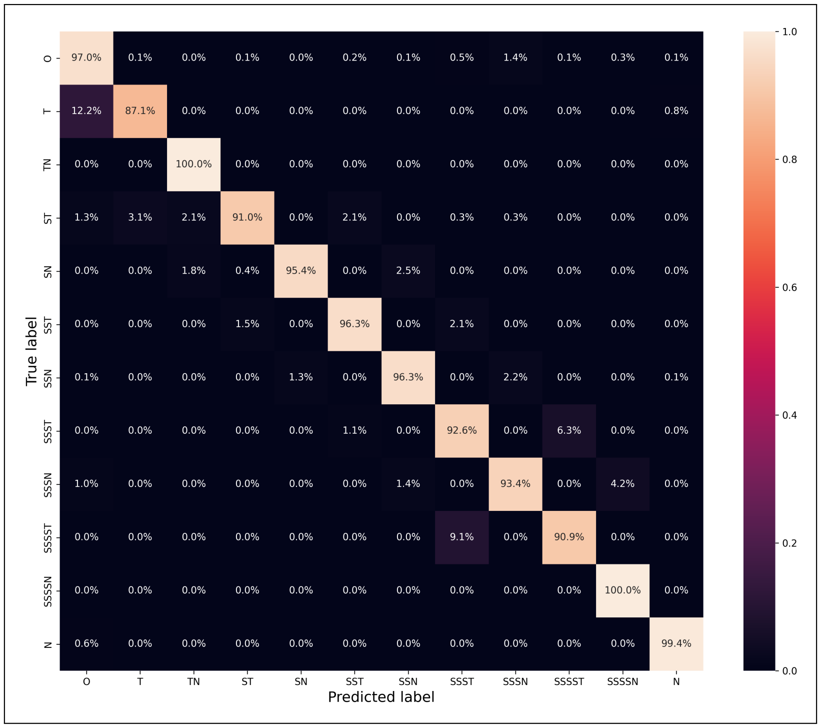


Figure 8: Confusion matrix for MarkupLM-base model trained on 121 contracts (entire training set).

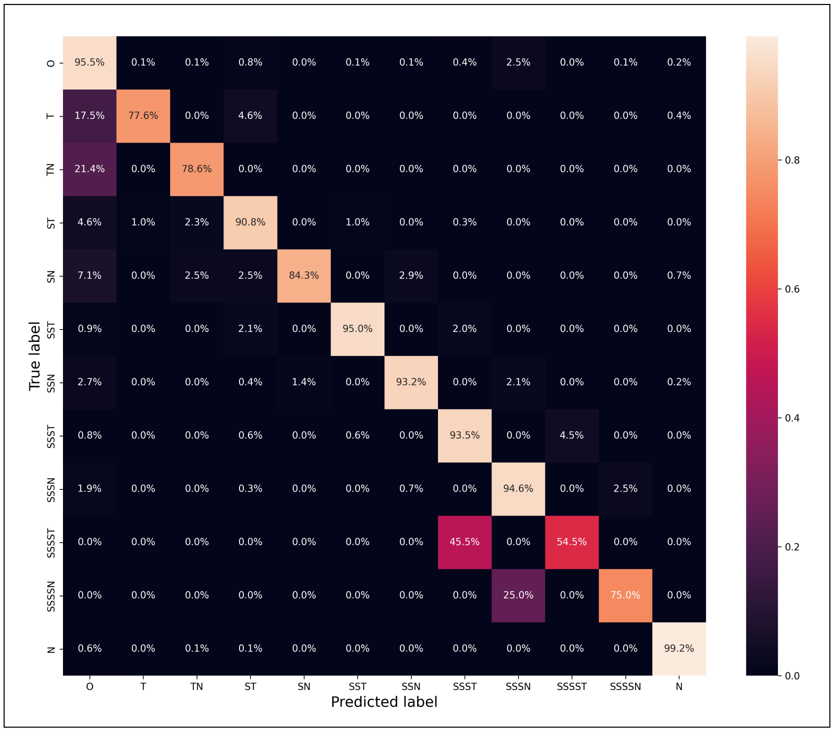
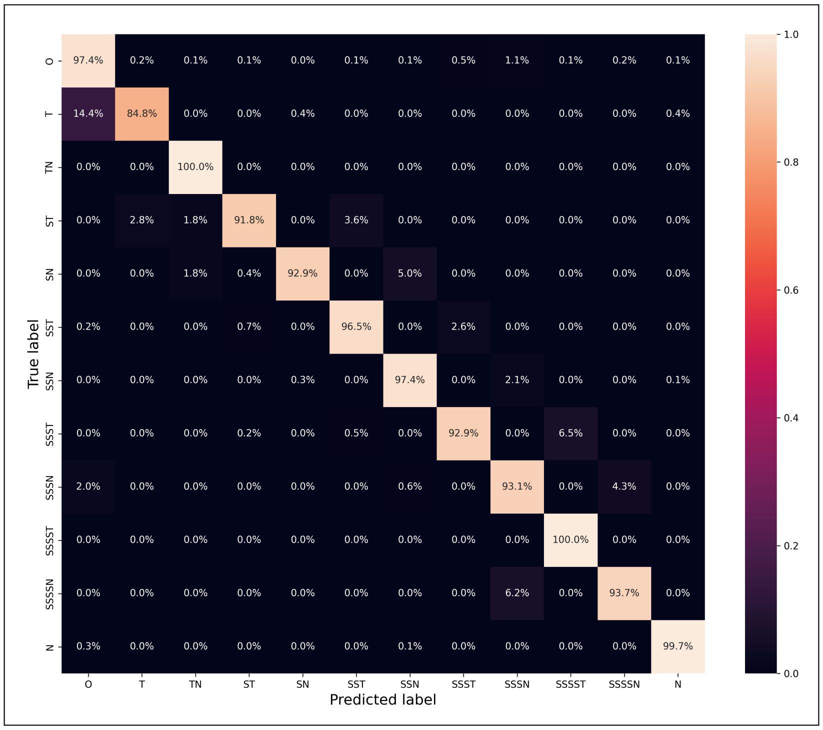
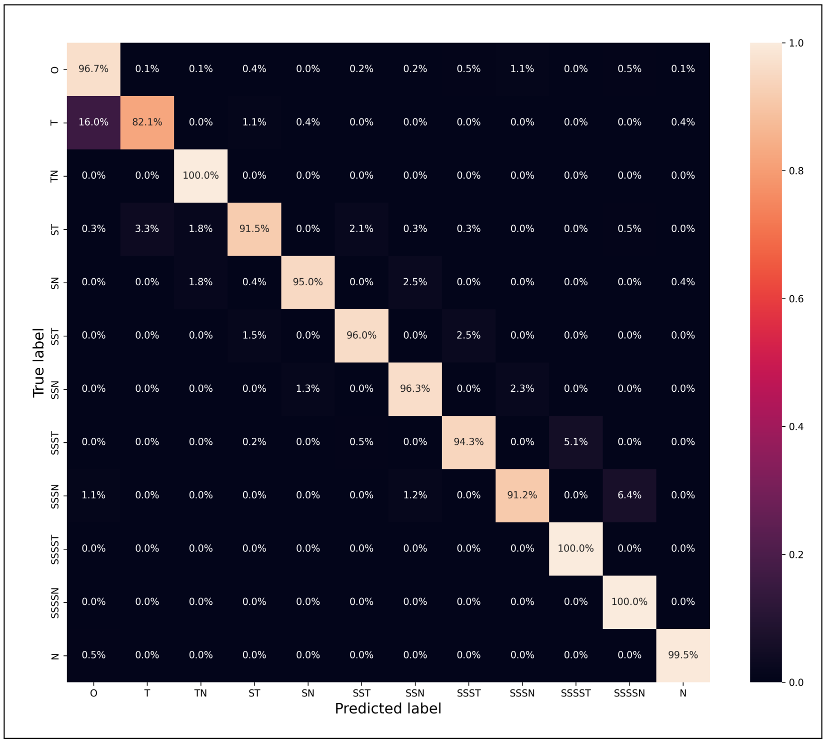


Figure 9: Confusion matrix for RoBERTa model trained on 121 contracts (entire training set).



	O	T	TN	ST	SN	SST	SSN	SSST	SSSN	SSSST	SSSSN	N
O	97.4%	0.2%	0.1%	0.1%	0.0%	0.1%	0.1%	0.5%	1.1%	0.1%	0.2%	0.1%
T	14.4%	84.8%	0.0%	0.0%	0.4%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.4%
TN	0.0%	0.0%	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
ST	0.0%	2.8%	1.8%	91.8%	0.0%	3.6%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
SN	0.0%	0.0%	1.8%	0.4%	92.9%	0.0%	5.0%	0.0%	0.0%	0.0%	0.0%	0.0%
SST	0.2%	0.0%	0.0%	0.7%	0.0%	96.5%	0.0%	2.6%	0.0%	0.0%	0.0%	0.0%
SSN	0.0%	0.0%	0.0%	0.0%	0.3%	0.0%	97.4%	0.0%	2.1%	0.0%	0.0%	0.1%
SSST	0.0%	0.0%	0.0%	0.2%	0.0%	0.5%	0.0%	92.9%	0.0%	6.5%	0.0%	0.0%
SSSN	2.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.6%	0.0%	93.1%	0.0%	4.3%	0.0%
SSSST	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%
SSSSN	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	6.2%	0.0%	93.7%	0.0%
N	0.3%	0.0%	0.0%	0.0%	0.0%	0.0%	0.1%	0.0%	0.0%	0.0%	0.0%	99.7%

Figure 10: Confusion matrix for XDoc model trained on 121 contracts (entire training set).



	O	T	TN	ST	SN	SST	SSN	SSST	SSSN	SSSST	SSSSN	N
O	96.7%	0.1%	0.1%	0.4%	0.0%	0.2%	0.2%	0.5%	1.1%	0.0%	0.5%	0.1%
T	16.0%	82.1%	0.0%	1.1%	0.4%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.4%
TN	0.0%	0.0%	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
ST	0.3%	3.3%	1.8%	91.5%	0.0%	2.1%	0.3%	0.3%	0.0%	0.0%	0.5%	0.0%
SN	0.0%	0.0%	1.8%	0.4%	95.0%	0.0%	2.5%	0.0%	0.0%	0.0%	0.0%	0.4%
SST	0.0%	0.0%	0.0%	1.5%	0.0%	96.0%	0.0%	2.5%	0.0%	0.0%	0.0%	0.0%
SSN	0.0%	0.0%	0.0%	0.0%	1.3%	0.0%	96.3%	0.0%	2.3%	0.0%	0.0%	0.0%
SSST	0.0%	0.0%	0.0%	0.2%	0.0%	0.5%	0.0%	94.3%	0.0%	5.1%	0.0%	0.0%
SSSN	1.1%	0.0%	0.0%	0.0%	0.0%	0.0%	1.2%	0.0%	91.2%	0.0%	6.4%	0.0%
SSSST	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%
SSSSN	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
N	0.5%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	99.5%

Figure 11: Confusion matrix for xpath only inference experiment using MarkupLM-text-masking model trained on 121 contracts (entire training set).

Training dynamics In this section, we report confusion matrices evaluating the inference of MarkupML-base on the test set when this model is trained on 4 contracts (Figure 12), 12 contracts (Figure 13), 33 contracts (Figure 14) and 65 contracts (Figure 15). These matrices are also discussed in Section 4.3.

C Supplementary materials

Dataset documentation and intended uses. We provide dataset documentation in accordance with the dataset nutrition labels framework of [18]. For the Metadata, Provenance, Variables module, are provided in Table 6. For the Statistics module see Table 2.

Metadata	
Filename	`contract_###.csv`
Format	csv
Url	`https://doi.org/10.5281/zenodo.8034852`
Domain	natural language processing
Keywords	law, merger and acquisition agreements, document AI, legal NLP, EDGAR, hierarchical segmentation
Type	
Rows	Varies
Columns	5
Missing	none
License	CC BY 4.0 (`data`) MIT License (`code`)
Released	June 2023
Range	N/A
Description	This dataset is a collection of 151 public merger acquisition agreements with annotated section titles and certain other related information. These agreements were sourced in HTML format from Securities and Exchange Commission's EDGAR database. The dataset contains document object model (DOM) representations of these HTML together with labels corresponding to each DOM node.

Variables	
xpaths	Unique path used to identify nodes corresponding to the outside label as well as categories of interest
text	Text contained in each node of a contract, (both outside label and categories of interest)
highlighted_xpaths	Unique path used to identify nodes that contain a category of interest
highlighted_segmented_xpaths	Text contained in nodes that contain a category of interest
tagged_sequence	String label characterizing the category of interest contained in the node.

Provenance	
Source	
EDGAR	(`https://www.sec.gov/edgar`)
Author	
Name	S. Rao, P. Islam, R. Bollineni, S. Khosla, T. Fei, Q. Wu, K. Cho, V.A.Kobzar
Email	`vak2116@columbia.edu`

Table 6: Dataset Nutrition Labels for MarkupMnA.

The MarkupMnA dataset is intended to be used by researchers to build or evaluate algorithms for predicting section titles in M&A agreements and other related tasks.

Dataset and Benchmark URLs The MarkupMnA dataset can be downloaded from `https://doi.org/10.5281/zenodo.8034852`. The code and benchmark models are available at `https://github.com/MarkupMnA/MarkupMnA-Markup-Based-Segmentation-of-MnA-Agreements`.

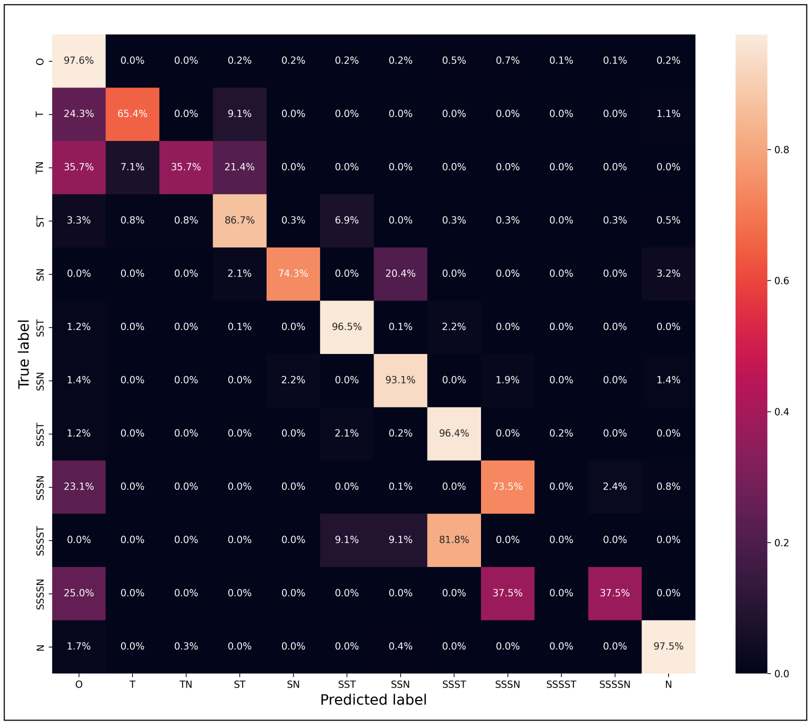


Figure 12: Confusion matrix evaluating the test inference of MarkupLM-base trained on 4 contracts.

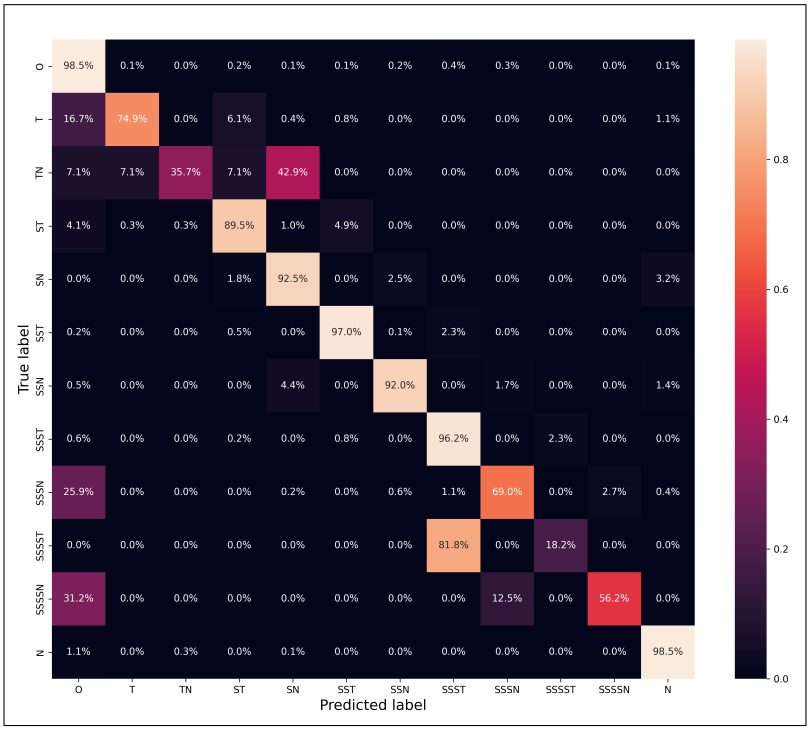


Figure 13: Confusion matrix evaluating test inference of MarkupLM-base trained on 12 contracts.

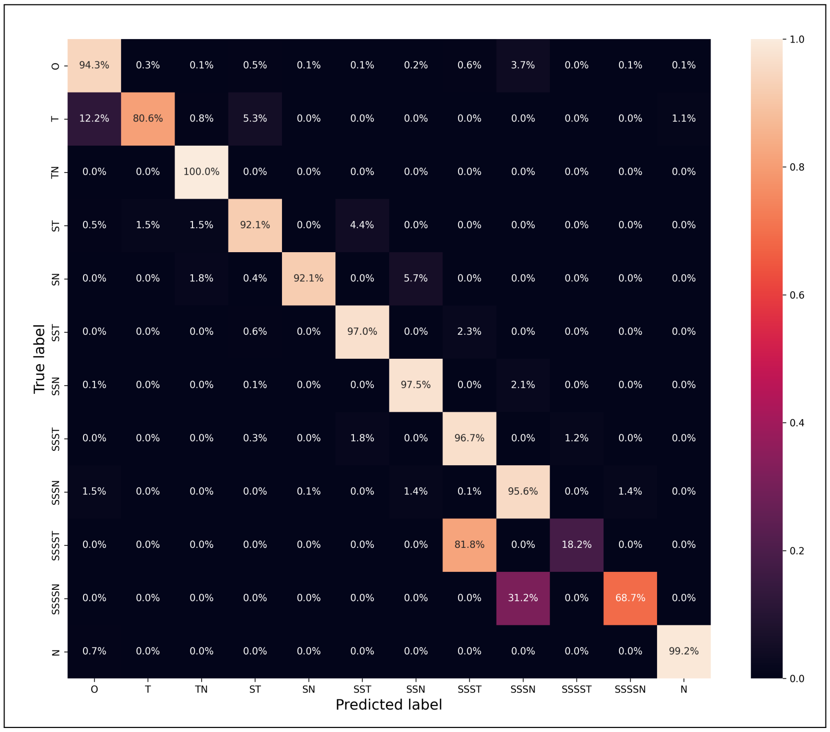


Figure 14: Confusion matrix evaluating the test inference of MarkupLM-base trained on 33 contracts.

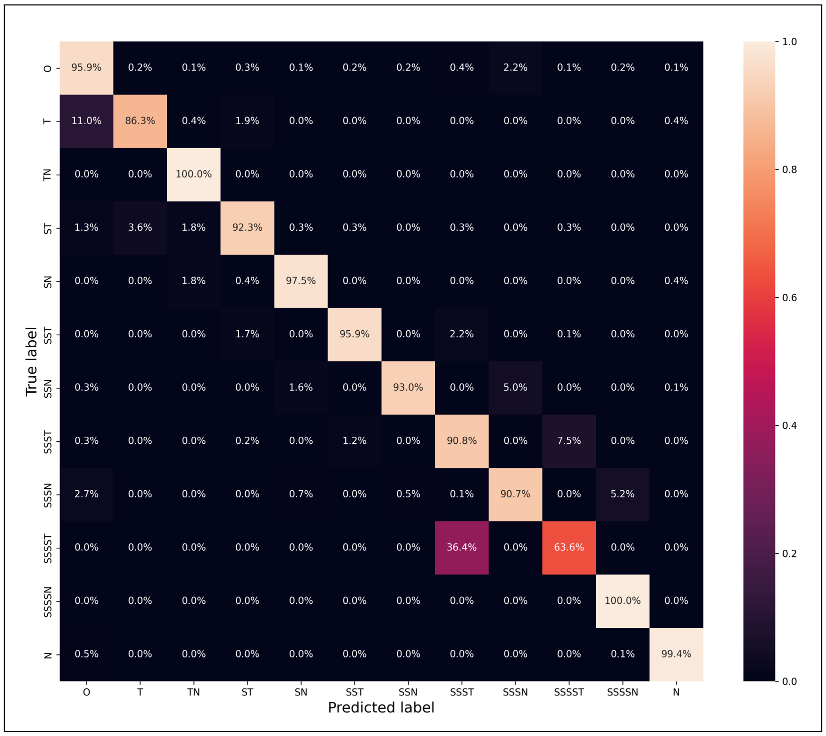


Figure 15: Confusion matrix evaluating the test inference of MarkupLM-base trained on 65 contracts.

Disclaimer of liability. The dataset and all related software are provided "as is", without warranty of any kind, express or implied, including but not limited to the warranties of merchantability, fitness for a particular purpose and noninfringement, and in no event shall the authors or copyright holders be liable for any claim, damages or other liability, whether in an action of contract, tort or otherwise, arising from, out of or in connection with the dataset, the software or the use or other dealings in the dataset or the software.

Hosting, licensing, and maintenance plan. The dataset is hosted and maintained on Zenodo [15], and the code is hosted on GitHub. The dataset is released under the CC BY 4.0 license and the code is released under the MIT License. Zenodo metadata is openly available under CC0 licence, and all open content is openly accessible through open APIs.[7]

Links to access the dataset and its metadata. The MARKUPMNA dataset is available at `https://doi.org/10.5281/zenodo.8034852`. The code and the benchmark models are available at `https://github.com/MarkupMnA/MarkupMnA-Markup-Based-Segmentation-of-MnA-Agreements`.

Data format The dataset is provided as csv files, which can be read using standard libraries. We provide the schema in Figure 7 [8].

Long-term preservation. We arrange for the long-term preservation of the dataset by uploading it to Zenodo.

Explicit license The dataset is released under the CC BY 4.0 license and the code is released under the MIT License.

Structured metadata We release the metadata along with the dataset on Zenodo. Zenodo metadata is openly available under CC0 licence, and all open content is openly accessible through open APIs.[9]

Persistent dereferenceable identifier. `https://doi.org/10.5281/zenodo.8034852`

Reproducibility. We release on Github the code and documentation necessary to reproduce the results in this paper.

[7] `https://about.zenodo.org`

[8] `https://pypi.org/project/csv-schema/`

[9] `https://about.zenodo.org`