Legal Knowledge and Information Systems
G. Sileno et al. (Eds.)

doi:10.3233/FAIA230963

Information Extraction from Lengthy Legal Contracts: Leveraging Query-Based Summarization and GPT-3.5

May Myo ZIN [a,1], Ha Thanh NGUYEN [a] , Ken SATOH [a] , Saku SUGAWARA [a] , and Fumihito NISHINO [a]
[a] *National Institute of Informatics, Tokyo, Japan.*
ORCiD ID: May Myo Zin https://orcid.org/0000-0003-1315-7704 , Ha Thanh Nguyen https://orcid.org/0000-0003-2794-7010 , Ken Satoh https://orcid.org/0000-0002-9309-4602, Saku Sugawara https://orcid.org/0000-0002-0061-0680 , Fumihito Nishino https://orcid.org/0000-0001-7368-4923

Abstract. In the legal domain, extracting information from contracts poses significant challenges, primarily due to the scarcity of annotated data. In such situations, leveraging large language models (LLMs), such as the Generative Pre-trained Transformer (GPT) models, offers a promising solution. However, the inherent token limitations of these models can be a bottleneck for processing lengthy legal contracts. This paper presents an unsupervised two-step approach to address these challenges. First, we propose a query-based summarization model that extracts sentences pertinent to predefined queries, concisely representing lengthy contracts. This summarization ensures that the core information remains intact while simultaneously addressing the token limitation issue. Subsequently, the generated summary is fed to GPT-3.5 for precise information extraction. Our approach effectively overcomes the challenges of token limitations and zero resources, enabling efficient and scalable information extraction from legal contracts. We compare our results with those obtained from supervised models that have been fine-tuned on domain-specific annotated data. Experimental results demonstrate the remarkable effectiveness of our approach, as it achieves state-of-the-art performance without the need for domain-specific training data.

Keywords. Information extraction, text summarization, lengthy legal contracts, zero-resource, large language models, unsupervised approach

1. Introduction

In recent years, the field of information extraction has seen significant progress with three notable approaches. First, pre-trained Question Answering (QA) models, fine-tuned for specific tasks, have excelled in answering questions based on text [1], making them useful for automating information extraction through targeted queries. Second, pre-trained Named Entity Recognition (NER) models have proven effective in automatically identifying and categorizing entities within the text, simplifying the extraction of specific information from unstructured documents [2-4]. Lastly, prompt engineering techniques have been employed with Generative Pre-trained Transformer (GPT) [5] models,

[1] Corresponding Author: May Myo Zin, maymyozin@nii.ac.jp

leveraging their natural language understanding to extract information by guiding their responses to tailored prompts or queries [6]. These approaches, empowered by large language models (LLMs), offer the potential to significantly streamline document review processes, saving time and resources while enhancing the accuracy of information extraction.

In the legal domain, extracting information from contracts presents three primary challenges for existing state-of-the-art models. The initial challenge is the scarcity of data required to train or fine-tune models for achieving the high accuracy. Another challenge is the substantial size of many contracts, which often exceeds the processing capacity of the current transformer architectures. Transformer-based models have a maximum sequence length they can handle. If a contract exceeds the limit, it may need to be divided into smaller segments, which can complicate the analysis process. The third challenge is that contracts contain a mixture of short entities, such as names and dates, and lengthy ones, such as entire clauses like non-compete and audit rights. This raises the question of selecting the most appropriate approach, as the NER approach is better suited for extracting short entities, whereas the QA approach is more effective for longer entity extraction [7].

In this paper, we present an unsupervised two-step approach for overcoming the challenges posed by zero-resource data, lengthy contracts, and the extraction of both short and long entities. Considering the zero-resource problem, we leverage the power of LLMs such as GPT models without requiring any training or fine-tuning data. However, to overcome the maximum token limitation of these models, we first develop a query-based contract summarization approach to reduce the text size of the contract. Then, as the second step, we utilize GPT-3.5 to extract information from the contract summaries. Through prompt engineering with GPT-3.5, our approach can effectively extract both short and lengthy entities from the contract summaries. Subsequently, we conduct a comparative evaluation of our approach by benchmarking it against state-of-the-art QA models, using the CUAD [8] test set as our performance test data.

2. Related Work

Cutting-edge transformer models, used for tasks like information extraction, often require substantial training data for good accuracy. However, the legal field often lacks abundant labeled data. This shortage is a major challenge for developing accurate information extraction models, especially for legal contracts. Creating annotated training data is crucial but achieving error-free annotations is tough and costly. For example, annotating a dataset like CUAD (510 contracts) can cost up to $2 million [8]. This highlights the need for alternative information extraction methods, especially in low-resource scenarios where extensive annotation is impractical. In our study, we will leverage a transfer-learning approach to tackle the zero-resource problem. Utilizing LLMs like GPT models can alleviate the lack of annotated data for some specific tasks. Instead of relying solely on annotated training data, we can customize these models for the legal domain by incorporating domain-specific prompts.

Legal contracts are notorious for their lengthy nature, often surpassing the maximum token limits of state-of-the-art transformer architectures. While LLMs have excelled in various NLP tasks, their constraints in processing substantial documents impede effective analysis and information extraction from extensive legal contracts, which typically span many pages of text. Even the powerful LLMs, such as GPT-3.5 models, have token limitations that prevent the entire contract from being fed as input. To tackle these challenges, existing literature outlines three primary strategies for preprocessing long documents for use with the LLMs. The simplest approach involves truncating the lengthy input text into a shorter sequence within a predefined maximum length [9 , 10]. While this allows the use of off-the-shelf LLMs, it is heavily influenced by lead bias and can lead to significant information loss as the document length increases [11, 12, 13]. Another approach, text chunking, involves breaking down a long document into smaller, semantically similar segments and processing each segment independently before aggregation [8, 14]. While this method preserves the information of the entire document, it can disrupt long-range dependencies. This poses challenges for answer extraction. When answers are extracted independently from each segment, the challenge arises of comparing the scores for each segment's answer to determine the best choice as the final output [15, 16]. The third approach is selecting salient texts. This approach, based on the assumption that vital information occupies a small portion of a lengthy document, employs content selection to concatenate relevant segments for processing by LLMs [17]. While it can enhance downstream tasks, the quality of extracted texts relies heavily on the selection of snippets throughout the document [11, 18]. Moreover, Legal contracts encompass a wide range of entities, from short and discrete elements like party names and specific dates to extremely long entities, such as non-compete and audit rights which are entire contractual clauses embedded within lengthy contracts. Traditional token classification models (i.e., NER models), often trained with the IOB tagging scheme, face difficulties in accurately extracting long entities. In contrast, QA models have demonstrated promise in effectively extracting long entities, indicating the potential for alternative approaches to address this issue [7].

3. Our Approach

We propose a simple and effective approach to improve information extraction from lengthy legal contracts, even in the absence of training data. Our approach, illustrated in Figure 1, comprises two key phases: Query-based Summarization and Information Extraction.

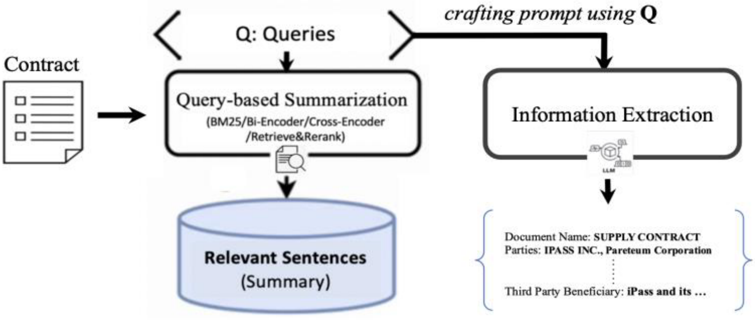

Q: Queries
crafting prompt using Q
Contract
Query-based Summarization
(BM25/Bi-Encoder/Cross-Encoder /Retrieve&Rerank)
Relevant Sentences
(Summary)
Information Extraction
LLM
Document Name: SUPPLY CONTRACT
Parties: IPASS INC., Pareteum Corporation
Third Party Beneficiary: iPass and its ...

Figure 1. Overall architecture of the proposed approach.

3.1. Query-based Summarization

We design a query-based summarization model in an unsupervised setting. Given the specific queries, this model extracts only the relevant sentences from the contract. These extracted sentences form a summary that contains pertinent information. Reducing the text size through summarization is especially crucial when integrating with LLMs, which have input token limitations, for the information extraction process.

3.1.1. Methodologies for Sentence Extraction

In this section, we describe the predominant approaches to sentence extraction. Our study encompasses both traditional and neural network-based approaches. The details for each are outlined below:

- **BM25** [19]: Originates from the probabilistic retrieval model and assesses relevance using term frequencies, inverse document frequency, and a standard tokenization method. Sentences are ranked by their BM25 score to measure relevance to a query.
- **Bi-Encoder** [20]: Encodes the query and document separately, determining relevance through the cosine similarity of the embeddings.
- **Cross-Encoder** [20]: Combines the encoding of the query and sentence, generating a relevance score for each query-sentence pair to rank the sentences.
- **Retrieve & Re-rank**[2]: Uses Bi-encoder for initial sentence retrieval, then Cross-Encoder for re-ranking to enhance accuracy.

3.1.2. Summarization Workflow

3.1.2.1. Sentence Extraction for Multiple Queries

Given a document D and a set of queries $Q_1, Q_2, \dots, Q_N$, we extract top-k relevant sentences from the document for each query Q_i using one of the aforementioned methodologies. This process can be represented as follows:

$$\Sigma_i = Extract(D, Q_i, k) \tag{1}$$

where Σ_i is the set of top-k sentences from the document D relevant to the query Q_i. The exact number, k, can be adjusted depending on the desired length and depth of the summary.

3.1.2.2. Combination of Extracted Sentences

All the extracted sentences from the various queries are then combined to create a combined set of sentences, C.

$$C = \Sigma_1 \cup \Sigma_2 \cup \dots \cup \Sigma_N \tag{2}$$

3.1.2.3. Redundancy Removal and Ordering

From the combined set C, redundant sentences are removed to create a summary S. This summary is then ordered based on the original sentence IDs (i.e., the ID of the first sentence of the contract is 1, the second is 2, and so on) to maintain logical coherence. This process is represented as follows:

$$S = Order(RemoveRedundancy(C)) \tag{3}$$

[2] https://www.sbert.net/examples/applications/retrieve_rerank/README.html

Ordering sentences in contract preserves the original context, ensuring a logical flow and readability. Maintaining this order also reduces the risk of misinterpretation. Thus, it is crucial to keep the original sequence when summarizing.

3.2. Information Extraction

After the process of contract summarization, the ordered summary S is subjected to information extraction via GPT-3.5. This phase is represented as follows:

$$I = ExtractInfo(S, Prompt) \tag{4}$$

GPT-3.5 is one of the advanced models developed by OpenAI[3]. Its design allows for understanding context, discerning intricate details, and producing relevant outputs tailored to the user's prompts. This capability makes it an invaluable tool for tasks such as information extraction.Using this model, we can extract both short and long entities with high accuracy. The success of the extraction largely depends on optimally crafted prompts. The prompts themselves are carefully designed to provide specific guidance to the GPT-3.5 model for extracting information from the summarized contracts, rather than generating its own answer text.

4. Experiments

4.1. Data

We use CUAD [8] test dataset to evaluate our models. CUAD is a comprehensive QA dataset comprising 510 legal contracts. This dataset encompasses a wide range of contract types, totaling 25 different varieties. The contracts in this dataset exhibit considerable variation in terms of length, spanning from just a few pages to well over one hundred pages. The dataset's questions are designed to extract information related to 41 different types of entities. These entities can range from brief items such as party names and dates to much lengthier ones like non-compete and audit rights clauses, which may require retrieval of several sentences or even entire paragraphs. Table 1 presents the data statistics of the CUAD dataset.

Table 1. Data statistics of the CUAD dataset.

Total Contracts	Type of Contracts	Type of Entities	Training Samples	Test Samples
510	25	41	16,728	4,182

4.2. Evaluation Metrics

For the summarization task, we do not have gold standard summaries to directly evaluate our methods. To address this challenge, we have formulated an evaluation metric as follows:

$$Accuracy = \frac{Number\ of\ Gold\ Answers\ found\ in\ Summaries}{Total\ Number\ of\ Gold\ Answers} \tag{5}$$

[3] https://openai.com/

To evaluate the performance of the information extraction models, we calculated and compared (Macro-averaged) F1 and Exact Match (EM) scores for each model. These metrics, originally proposed by Rajpurkar et.,al in the original SQuAD [21] paper, are widely recognized in the QA task evaluation.

4.3. Query-based Summarization

4.3.1. Implementation Details

As described in section 3.1.1, we employed not only traditional model but also neural network-based models for query-based summarization tasks. For the traditional approach, we used BM25 algorithm which is implemented in the rank_bm25[4] python package. For the neural network-based approaches, we utilized a Bi-Encoder, specifically "multi-qa-MiniLM-L6-cos-v1", and a Cross-Encoder, “mmarco-mMiniLMv2-L12-H384-v1”. We leveraged the SentenceTransformers[5] python framework for both the Bi-Encoder and Cross-Encoder. We created 41 distinct queries, each corresponding to a specific entity, for the purpose of sentence extraction. We conducted comprehensive experiments, exploring over 20 distinct query variations for each entity type, and determined which one yielded the best results. The first 5 queries used in this study are displayed in Table 2. The remaining queries are modified versions of CUAD's questions with the exclusion of the phrases “*Highlight the parts (if any) of this contract related to*” and “*that should be reviewed by a lawyer. Details*”. These segments were affecting the accuracy of the similarity calculation between the query and the sentence from the contract.

Table 2. Query Examples.

Entity Name	Query
Document Name	XXX AGREEMENT. Which agreement is being made or entered into?
Parties	Party A and Party B. The agreement is entered into by and between which Party A and Party B?"
Agreement Date	Agreement date/signed date: what is the date of the agreement?
Effective Date	Effective date: when the agreement will be affective?
Expiration Date	On what date will the contract's initial term expire?

4.3.2. Results and Discussions

We present the summarization results of four different models in Table 3, evaluated on the CUAD test set. The accuracy score is calculated using Eq. (5). For the Retrieve & Re-Rank approach, we first employ a Bi-Encoder to retrieve 32 potential sentences that answer the input query. Subsequently, we use a more advanced Cross-Encoder to score the query and all retrieved sentences based on their relevance. The cross-encoder significantly boosts performance, especially when searching over a corpus that the bi-encoder was not trained on. We then sort the results by the Cross-Encoder scores and select only the top-k sentences as our final choices. In these experiments, we set the value of k to 3, 5, and 10. A larger k value ensures the inclusion of relevant sentences in the summary. When k is set to a smaller value (i.e., k = 3), some relevant sentences are missing in the BM25’s extracted summary compared to other approaches. Yet, when we set a larger k value, the BM25 model outperformed the rest. BM25 is a heuristic-based method, and its performance is more consistent without requiring re-training. In contrast, the Bi-Encoder and Cross-Encoder need large amounts of labeled data for training or

[4] https://pypi.org/project/rank-bm25/

[5] https://www.sbert.net/

fine-tuning to grasp the semantic relevance between queries and sentences for this specific task. Hence, although these models possess the ability to understand deeper semantic relationships, being based on large-scale language models, achieving higher accuracy in this specialized task necessitates fine-tuning. Based on these results, we selected BM25 as our primary approach for the contract summarization.

We also investigated these models using the original CUAD's questions as the desired queries for this task. However, due to the inclusion of some extra and irrelevant parts in the queries, there was a negative impact on the similarity calculation process between the queries and the sentences.

Table 3. Summarization results on CUAD's test set.

Model	Accuracy (top-3)	Accuracy (top-5)	Accuracy (top-10)
BM25	0.58	**0.80**	**0.88**
Bi-Encoder	0.61	0.69	0.76
Cross-Encoder	**0.69**	0.73	0.86
Retrieve & Re-Rank	0.65	0.71	0.85

4.4. Information Extraction

4.4.1. Implementation Details

This research focuses on the zero-resource problem; hence we did not use any training or fine-tuning data for this task. We leverage a GPT-3.5 model, gpt-3.5-turbo-16k, in unsupervised scenarios. This model has been trained on a vast and diverse corpus of text data, providing it with robust language processing capabilities. By employing gpt-3.5-turbo-16k, we aimed to harness its pre-existing knowledge in a zero-shot learning setting and accommodate larger maximum token counts. To ensure optimal results, we invested significant effort in crafting appropriate prompts that guide the GPT3.5 to produce the desired outputs. Our experimentation involved exploring over 50 different prompt variations, with the most successful prompt presented in Figure 2.

We have conducted a comparative analysis of our unsupervised approach against existing supervised QA methodologies. To perform this evaluation, we employed two QA models: RoBERTa-based QA and DeBERTa-based QA, utilizing the best CUAD's fine-tuned models[6] available, namely, roberta-large and deberta-v2-xlarge. It's worth noting that these models have already been fine-tuned on the CUAD dataset.

4.4.2. Results and Discussion

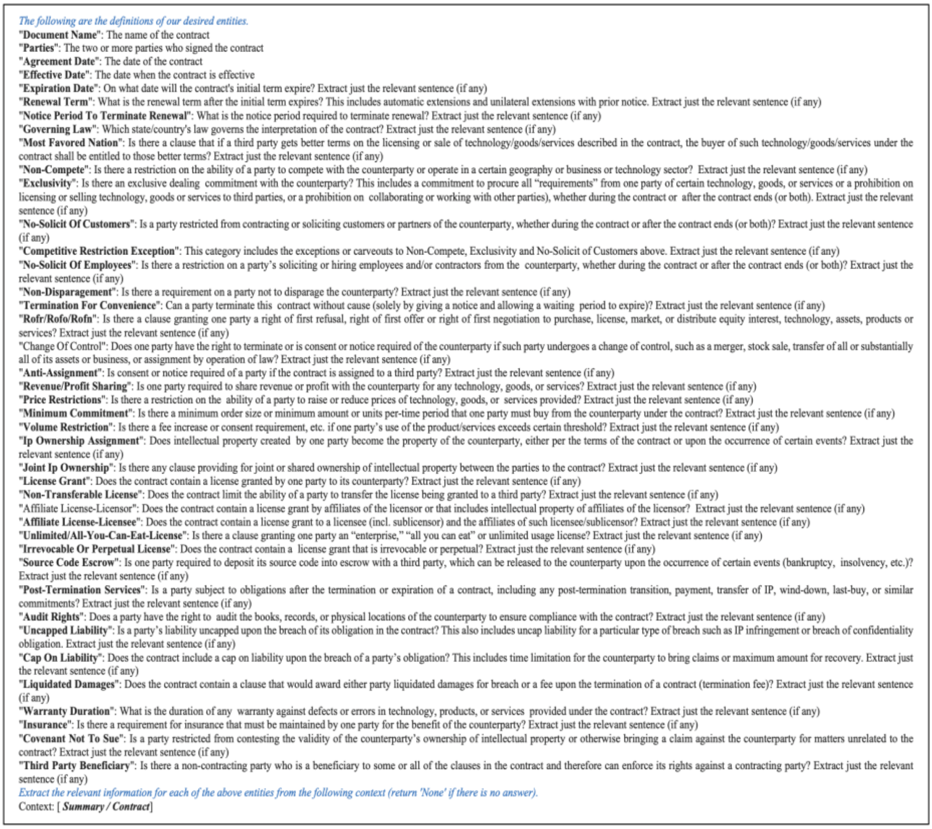
We evaluated the performance of the models using the CUAD test dataset, which comprises 4,182 test samples from 102 contracts. These 102 contracts include both short and long contracts. Notably, out of the 102 contracts, 20 contain tokens that exceed the maximum token limit of the gpt-3.5-turbo-16k model. Since the CUAD's fine-tuned QA models employ a chunking mechanism to handle lengthy contracts, entire contracts can be fed as input to these models. In contrast, due to the max token limitation of GPT-3.5 in our approach, we input the contract summaries. The contract summaries are created using a BM25-based summarization approach. To avoid exceeding the max token limitations of gpt-3.5-turbo-16k (i.e., 16,385 tokens), we use a setting of k = 5 in the summarization process.

[6] https://zenodo.org/record/4599830/

The experimental results are shown in Table 4. Notably, our unsupervised approach demonstrated superior performance compared to the supervised models, even though these supervised models had already been fine-tuned on the CUAD dataset. The results indicate that our summarization method effectively captures the desired information in the summary while also addressing the token limitation issue of LLMs in lengthy contracts reviews.

Table 4. Experimental results comparing the unsupervised approach with two supervised QA models for information extraction on CUAD's test dataset.

Model	Input Data	Exact Match	Precision	Recall	F1
RoBERTa QA [8]	Entire Contract	0.69	0.76	0.70	0.73
DeBERTa QA [8]		0.70	0.76	0.71	0.73
GPT 3.5 (Ours)	Prompt + Contract Summary	**0.72**	**0.77**	**0.76**	**0.76**



The following are the definitions of our desired entities.

"Document Name": The name of the contract

"Parties": The two or more parties who signed the contract

"Agreement Date": The date of the contract

"Effective Date": The date when the contract is effective

"Expiration Date": On what date will the contract's initial term expire? Extract just the relevant sentence (if any)

"Renewal Term": What is the renewal term after the initial term expires? This includes automatic extensions and unilateral extensions with prior notice. Extract just the relevant sentence (if any)

"Notice Period To Terminate Renewal": What is the notice period required to terminate renewal? Extract just the relevant sentence (if any)

"Governing Law": Which state/country's law governs the interpretation of the contract? Extract just the relevant sentence (if any)

"Most Favored Nation": Is there a clause that if a third party gets better terms on the licensing or sale of technology/goods/services described in the contract, the buyer of such technology/goods/services under the contract shall be entitled to those better terms? Extract just the relevant sentence (if any)

"Non-Compete": Is there a restriction on the ability of a party to compete with the counterparty or operate in a certain geography or business or technology sector? Extract just the relevant sentence (if any)

"Exclusivity": Is there an exclusive dealing commitment with the counterparty? This includes a commitment to procure all "requirements" from one party of certain technology, goods, or services or a prohibition on licensing or selling technology, goods or services to third parties, or a prohibition on collaborating or working with other parties), whether during the contract or after the contract ends (or both). Extract just the relevant sentence (if any)

"No-Solicit Of Customers": Is a party restricted from contracting or soliciting customers or partners of the counterparty, whether during the contract or after the contract ends (or both)? Extract just the relevant sentence (if any)

"Competitive Restriction Exception": This category includes the exceptions or carveouts to Non-Compete, Exclusivity and No-Solicit of Customers above. Extract just the relevant sentence (if any)

"No-Solicit Of Employees": Is there a restriction on a party's soliciting or hiring employees and/or contractors from the counterparty, whether during the contract or after the contract ends (or both)? Extract just the relevant sentence (if any)

"Non-Disparagement": Is there a requirement on a party not to disparage the counterparty? Extract just the relevant sentence (if any)

"Termination For Convenience": Can a party terminate this contract without cause (solely by giving a notice and allowing a waiting period to expire)? Extract just the relevant sentence (if any)

"Rofr/Rofo/Rofn": Is there a clause granting one party a right of first refusal, right of first offer or right of first negotiation to purchase, license, market, or distribute equity interest, technology, assets, products or services? Extract just the relevant sentence (if any)

"Change Of Control": Does one party have the right to terminate or is consent or notice required of the counterparty if such party undergoes a change of control, such as a merger, stock sale, transfer of all or substantially all of its assets or business, or assignment by operation of law? Extract just the relevant sentence (if any)

"Anti-Assignment": Is consent or notice required of a party if the contract is assigned to a third party? Extract just the relevant sentence (if any)

"Revenue/Profit Sharing": Is one party required to share revenue or profit with the counterparty for any technology, goods, or services? Extract just the relevant sentence (if any)

"Price Restrictions": Is there a restriction on the ability of a party to raise or reduce prices of technology, goods, or services provided? Extract just the relevant sentence (if any)

"Minimum Commitment": Is there a minimum order size or minimum amount or units per-time period that one party must buy from the counterparty under the contract? Extract just the relevant sentence (if any)

"Volume Restriction": Is there a fee increase or consent requirement, etc. if one party's use of the product/services exceeds certain threshold? Extract just the relevant sentence (if any)

"Ip Ownership Assignment": Does intellectual property created by one party become the property of the counterparty, either per the terms of the contract or upon the occurrence of certain events? Extract just the relevant sentence (if any)

"Joint Ip Ownership": Is there any clause providing for joint or shared ownership of intellectual property between the parties to the contract? Extract just the relevant sentence (if any)

"License Grant": Does the contract contain a license granted by one party to its counterparty? Extract just the relevant sentence (if any)

"Non-Transferable License": Does the contract limit the ability of a party to transfer the license being granted to a third party? Extract just the relevant sentence (if any)

"Affiliate License-Licensor": Does the contract contain a license grant by affiliates of the licensor or that includes intellectual property of affiliates of the licensor? Extract just the relevant sentence (if any)

"Affiliate License-Licensee": Does the contract contain a license grant to a licensee (incl. sublicensor) and the affiliates of such licensee/sublicensor? Extract just the relevant sentence (if any)

"Unlimited/All-You-Can-Eat-License": Is there a clause granting one party an "enterprise," "all you can eat" or unlimited usage license? Extract just the relevant sentence (if any)

"Irrevocable Or Perpetual License": Does the contract contain a license grant that is irrevocable or perpetual? Extract just the relevant sentence (if any)

"Source Code Escrow": Is one party required to deposit its source code into escrow with a third party, which can be released to the counterparty upon the occurrence of certain events (bankruptcy, insolvency, etc.)? Extract just the relevant sentence (if any)

"Post-Termination Services": Is a party subject to obligations after the termination or expiration of a contract, including any post-termination transition, payment, transfer of IP, wind-down, last-buy, or similar commitments? Extract just the relevant sentence (if any)

"Audit Rights": Does a party have the right to audit the books, records, or physical locations of the counterparty to ensure compliance with the contract? Extract just the relevant sentence (if any)

"Uncapped Liability": Is a party's liability uncapped upon the breach of its obligation in the contract? This also includes uncap liability for a particular type of breach such as IP infringement or breach of confidentiality obligation. Extract just the relevant sentence (if any)

"Cap On Liability": Does the contract include a cap on liability upon the breach of a party's obligation? This includes time limitation for the counterparty to bring claims or maximum amount for recovery. Extract just the relevant sentence (if any)

"Liquidated Damages": Does the contract contain a clause that would award either party liquidated damages for breach or a fee upon the termination of a contract (termination fee)? Extract just the relevant sentence (if any)

"Warranty Duration": What is the duration of any warranty against defects or errors in technology, products, or services provided under the contract? Extract just the relevant sentence (if any)

"Insurance": Is there a requirement for insurance that must be maintained by one party for the benefit of the counterparty? Extract just the relevant sentence (if any)

"Covenant Not To Sue": Is a party restricted from contesting the validity of the counterparty's ownership of intellectual property or otherwise bringing a claim against the counterparty for matters unrelated to the contract? Extract just the relevant sentence (if any)

"Third Party Beneficiary": Is there a non-contracting party who is a beneficiary to some or all of the clauses in the contract and therefore can enforce its rights against a contracting party? Extract just the relevant sentence (if any)

Extract the relevant information for each of the above entities from the following context (return 'None' if there is no answer).

Context: [***Summary / Contract***]

Figure 2. Illustration of GPT-3.5 prompt for information extraction.

4.5. Ablation Study and Analysis

In practice, contracts can be quite lengthy. For instance, in the CUAD dataset, out of a total of 510 contracts, only 390 can fit within the token limits of the gpt-3.5-turbo-16k model. This means that 120 contracts exceed the model's capacity. A potential solution is to divide these lengthy contracts into two or three sections, then input each section separately into the model. However, this method may result in the model extracting multiple answers for a single entity. This presents challenges both in terms of splitting the contracts effectively and in determining which extracted answer is the most accurate.

In this section, we performed a qualitative analysis to evaluate the reliability of our approach when dealing exclusively with shorter contracts. Our aim was to determine whether the process of summarization remains dependable in such scenarios and to assess if GPT-3.5, combined with our optimized prompt, can extract information more effectively than the current state-of-the-art DeBERTa QA. In the CUAD test dataset, only 82 contracts are short contracts that have fewer than 16,385 tokens. Our results, presented in Table 5, indicate that GPT-3.5 significantly outperforms DeBERTa QA when processing these short 82 contracts. This evidence suggests that, with a meticulously crafted prompt, GPT-3.5 can efficiently extract information from legal contracts, even without additional resources. We also examined the performance of the models on our constructed summaries of short contracts. Notably, the performance of DeBERTa QA remained consistent whether applied to the summaries or the entire contracts. Similarly, the accuracy of GPT-3.5 on the summaries was almost identical to its performance on the full contracts. Our summarization approach, therefore, did not introduce any drawback in this task. Instead, it maintained the integrity and informational content of the original documents, ensuring that critical legal details were not omitted or misrepresented. This indicates the robustness of our models and underscores its potential for practical application in legal document processing and information extraction.

Table 5. Experimental results comparing our unsupervised approach with DeBERTa QA for information extraction form short contracts of CUAD's test dataset.

Model	Input Data	Exact Match	F1
DeBERTa QA [8]	Entire Contract	0.75	0.77
DeBERTa QA [8]	Contract Summary	0.75	0.77
GPT 3.5 (Ours)	Prompt + Entire Contract	0.78	0.82
GPT 3.5 (Ours)	Prompt + Contract Summary	0.77	0.81

5. Conclusion

In the complex field of legal contract analysis, our research introduces a groundbreaking unsupervised pipeline that effectively addresses three key challenges. Firstly, the scarcity of labeled training data, a significant hindrance in the legal domain, is overcome by leveraging the capabilities of GPT-3.5. This model, with its extensive pre-existing knowledge, eliminates the need for fine-tuning, making our approach particularly valuable in zero-resource scenarios. Secondly, we tackle the challenge presented by the extensive length of legal contracts through our innovative summarization techniques. By utilizing traditional methods like BM25, the model ensure that the essence of contracts, regardless of their length, is captured without the loss of critical information. Lastly, our pipeline excels at handling the diverse nature of contract information, from brief entities to lengthy clauses. Through carefully crafted prompts and the power of GPT-3.5, we ensure precise extraction of both short and long entities with the commendable accuracy. In essence, our research offers a comprehensive, scalable, and efficient solution, setting a new benchmark in the field of legal contract processing.

For future work, we plan to conduct more extensive experiments involving various language models (LLMs) on this dataset and explore their performance on other datasets.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 22H00543 and JST, AIP Trilateral AI Research, Grant Number JPMJCR20G4.

References

[1] Dong K, Sun A, Kim JJ, Li X. Shall We Trust All Relational Tuples by Open Information Extraction? A Study on Speculation Detection. arXiv preprint arXiv:2305.04181. 2023 May 7.

[2] Jofche N, Mishev K, Stojanov R, Jovanovik M, Zdravevski E, Trajanov D. Named Entity Recognition and Knowledge Extraction from Pharmaceutical Texts using Transfer Learning. Procedia Computer Science. 2022 Jan 1;203:721-6.

[3] Jofche N, Mishev K, Stojanov R, Jovanovik M, Zdravevski E, Trajanov D. Pharmke: Knowledge extraction platform for pharmaceutical texts using transfer learning. Computers. 2023 Jan 9;12(1):17.

[4] Dorrn T, Dambier N, Müller A, Kuwertz A. A Textual Information Extraction Application based on XML Data Models and a Multidimensional Natural Language Processing Pipeline Approach. In2022 International Conference on INnovations in Intelligent SysTems and Applications (INISTA) 2022 Aug 8 (pp. 1-6). IEEE.

[5] Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training. 2018.

[6] May MZ, Ha-Thanh N, Ken S, Saku S, Fumihito N. Improving translation of case descriptions into logical fact formulas using LegalCaseNER. InNineteenth International Conference on Artificial Intelligence and Law (ICAIL 2023) 2023.

[7] Lam L, Ratnamogan P, Tang J, Vanhuffel W, Caspani F. (2023). Information Extraction from Documents: Question Answering Vs Token Classification in Real-World Setups. In: Fink, G.A., Jain, R., Kise, K., Zanibbi, R. (eds) Document Analysis and Recognition - ICDAR 2023. ICDAR 2023. Lecture Notes in Computer Science, vol 14188. Springer, Cham. https://doi.org/10.1007/978-3-031-41679-8_12

[8] Hendrycks D, Burns C, Chen A, Ball S. Cuad: An expert-annotated nlp dataset for legal contract review. arXiv preprint arXiv:2103.06268. 2021 M

[9] Mike L, Yinhan L, Naman G, Marjan G, Abdelrahman M, Omer L, Ves S, and Luke Z. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv:1910.13461, 2019.

[10] Hyunji P, Yogarshi V, and Kashif S. Efficient classification of long documents using transformers. In ACL, 2022.

[11] Zhang Y, Ni A, Mao Z, Wu CH, Zhu C, Deb B, Awadallah AH, Radev D, Zhang R. Summ^n: A multi-stage summarization framework for long input dialogues and documents. arXiv preprint arXiv:2110.10150. 2021 Oct 16.

[12] Koh HY, Ju J, Liu M, Pan S. An empirical survey on long document summarization: Datasets, models, and metrics. ACM computing surveys. 2022 Dec 23;55(8):1-35.

[13] Dong Z, Tang T, Li L, Zhao WX. A survey on long text modeling with transformers. arXiv preprint arXiv:2302.14502. 2023 Feb 28.

[14] Liu Y, Ni A, Nan L, Deb B, Zhu C, Awadallah AH, Radev D. Leveraging locality in abstractive text summarization. arXiv preprint arXiv:2205.12476. 2022 May 25.

[15] Wang Z, Ng P, Ma X, Nallapati R, Xiang B. Multi-passage bert: A globally normalized bert model for open-domain question answering. arXiv preprint arXiv:1908.08167. 2019 Aug 22.

[16] Gong H, Shen Y, Yu D, Chen J, Yu D. Recurrent chunking mechanisms for long-text machine reading comprehension. arXiv preprint arXiv:2005.08056. 2020 May 16.

[17] Ding M, Zhou C, Yang H, Tang J. Cogltx: Applying bert to long texts. Advances in Neural Information Processing Systems. 2020;33:12792-804.

[18] Nie Y, Huang H, Wei W, Mao XL. Capturing Global Structural Information in Long Document Question Answering with Compressive Graph Selector Network. arXiv preprint arXiv:2210.05499. 2022 Oct 11.

[19] Robertson SE, Walker S. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. InSIGIR'94: Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, organised by Dublin City University 1994 (pp. 232-241). Springer London.

[20] Reimers N, Gurevych I. Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084. 2019 Aug 27.

[21] Rajpurkar P, Zhang J, Lopyrev K, Liang P. Squad: 100,000+ questions for machine comprehension of text. arXiv preprint arXiv:1606.05250. 2016 Jun 16.