AI-based Approach to Efficient Information Extraction for Supply Chain Contracts

Hong Ping, Yap
Infineon Technologies AP
Singapore
HongPing.Yap@infineon.com

Wee Ling, Ong
Infineon Technologies AP
Singapore
WeeLing.Ong@infineon.com

Jonathan, Koh
Infineon Technologies AP
Singapore
WenJieJonathan.Koh@infineon.com

Liu Haoran
Institute of System Science
National University of Singapore
Singapore
haoran.h.liu@u.nus.edu

Yang Tiancheng
Institute of System Science
National University of Singapore
Singapore
tiancheng.yang@u.nus.edu

Chen Zihao
Institute of System Science
National University of Singapore
Singapore
zihao.chen@u.nus.edu

***Abstract*—Global semiconductor supply chain management has played a critical role to improve if not maintain our service and value in this uncertain era. Contracts between suppliers and businesses constitute the very foundation of the global supply chain. Supply chain contracts are regularly being updated, owing to the volatile environment. It is crucial for supply chain professionals to be able to retrieve the right information quickly and accurately from the lengthy clauses in the contract. Hence, we present the application of pre-trained model (PTM), CUAD-RoBERTa with few-shot learning PERFECT framework to extract the salient portions of a supply chain contract, enable supply chain professionals to retrieve information and generate deep insights with high efficiency. The decent results open the possibility of future applications in streamlining contracts and other documents crucial to business operations.**

***Index Terms*—Natural Language Processing, Contract Information Extraction, Few-shot tuning**

I. Introduction

In our modern society where millions of transactions occur daily between businesses themselves and with consumers, it is imperative to have the proper contracts in place to document the legal obligations for the accountable parties. One of the main challenges for professionals in today's business world is the ability to retrieve the relevant terms and conditions (T&Cs) efficiently and accurately from these contracts, given the extensive legal jargon and multiple pages of clauses within them. Information retrieval from legal documents has historically been a manual undertaking. It is a very tedious task to read large amounts of text only to look for a specific piece of information.

Efficient supply chain management is crucial to achieving faster time-to-market in the semiconductor industry. It is essential to have the ability to retrieve relevant information from supply chain contracts easily. Infineon Technologies AG, a global semiconductor leader in power systems and IoT, has taken several steps towards digitalization to enable this capability. One such step is the introduction of a dedicated application, which provides centralized and digitalized storage of volume-related customer agreements. The aim of this solution is to holistically plan capacity and align operations with our contractual obligations in real-time. Additionally, in our current work, we have gone the extra mile to facilitate key contractual information extraction to eliminate the pain-point of scanning through contracts to check for topics of interest, such as delivery windows, or amended information in a set of contracts, which can be highly time-consuming.

This paper presents an approach to ease key information retrieval in supply chain contracts by using CUAD-RoBERTa [1] as a pre-trained language model in self-supervised and few-shot learning PERFECT [2] framework scenarios. Our experiment showed that few-shot learning using PERFECT framework yielded reasonable results in classifying topics of interest in our contracts. Such application eases the retrieval of key information and enables users to gain valuable insights into the contracts, potentially reducing costs and improving operations. The same approach can be extended to other contract applications, potentially streamlining, and improving the contract information retrieval process in the organization to be visible, searchable, and easy to use.

II. Related Work

Information extraction from various documents is getting more prevalent in today's business world, with many AI cloud services, such as Google AI Cloud and Baidu AI cloud, providing the capabilities for information extraction. One example is TextMind, a PTM-based platform under Baidu, which enables users to analyse a wide variety of documents, including receipts for claim reimbursements, financial statements and resumes [3]. Additionally, many startups, targeting niche specialist domain areas in legal technology, have emerged in recent years. These include Kira Systems, Luminance and ThoughtRiver [4]. Despite the many legal technology solutions available in the market, significant effort is still required to customize the system for our contracts' context. Furthermore, with the emergence of a series of effective PTMs for NLP,

979-8-3503-5409-6/24/$31.00 ©2024 IEEE
DOI 10.1109/CAI59869.2024.00034

such as GPT, BERT, RoBERTa, and T5, which have proven their effectiveness in extracting rich language knowledge from large-scale unlabelled corpora [5], an in-house solution is being explored instead.

Transfer learning using PTMs has demonstrated to be a highly effective approach in many NLP tasks. Researchers have been exploring a variety of transfer learning techniques, which include feature-based transfer learning [6], fine-tuning transfer learning, adapter tuning [7], and prompt-based fine-tuning [8], [9]. In the legal NLP domain, researchers have investigated several tasks, including legal entity recognition, document classification, legal question answering and legal summarization [10]–[12]. However, there is little prior work on contract review in the supply chain environment. The most related work to ours is that in [1] which covers some topics of interest related to supply chain contracts. It is a large-scale dataset for contract review based on 510 contracts and 13,101 labelled clauses, spanning across 41 label categories used to fine-tune on several PTMs for NLP.

Given the highly imbalanced and scarce labelled samples, we have chosen to adopt PERFECT as our transfer learning framework with CUAD-RoBERTa.

III. Dataset and Preprocessing

In this project, we mainly focus on Infineon's active Basic Supply Agreement (BSA) contracts from Asia Pacific dated as early as 1990. There are 2 types of contracts being considered: Main Contract and Contract Addendum & Amendment (CAA). The Main Contract is the initial supply contract document outlining the fundamental supply chain-related terms agreed between Infineon and the customer, while CAA refers to the subsequent contractual documents which amend existing terms or add new clauses to the original version. In total, there are 104 main contracts and 34 CAA contracts used.

The contracts have a variety of layout formats and quality in pdf form. Due to contract confidentiality, some samples have been extracted online for illustration purposes in Figure 1. While some of them are easy to read and have a clear format, there are others, particularly the older ones, with lower quality which could have been scanned copies of paper documents. Additionally, there are other complications to consider, such as multi-lingual, watermark, header and footer, company stamps, signatures, and handwriting. These can further complicate the process of extracting clear information from these contracts. In this paper, removal of Chinese in Chinese-English bilingual contents, headers and footers, and watermarks are in focus as these elements contribute to the bulk of the contracts. Those regions of non-interest are masked out using common image pre-processing techniques such as edge detection, morphological and Chinese language detection.

In order to segment the remaining contents into paragraphs, a hybrid approach is adopted, utilizing both Tesseract [13] and easyOCR [14] techniques. Afterwards, a final cleaning step is applied to eliminate gibberish, as some stamps and seals may be converted into random characters. The resulting output is stored in JSON format, illustrated in Figure 2, and used as input for the language model.

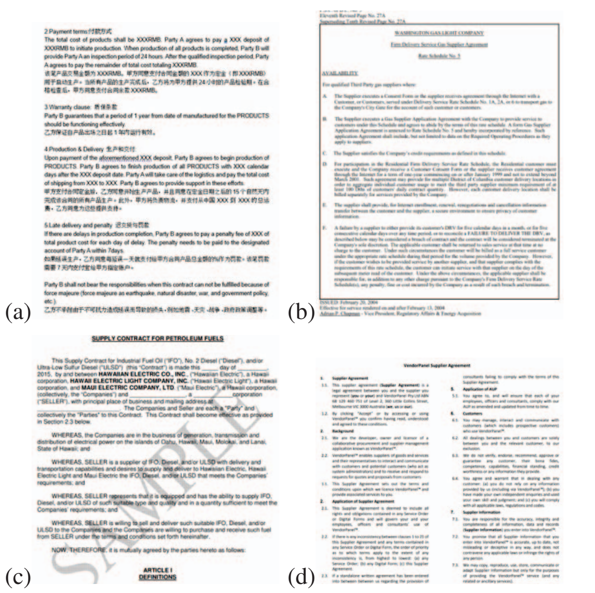
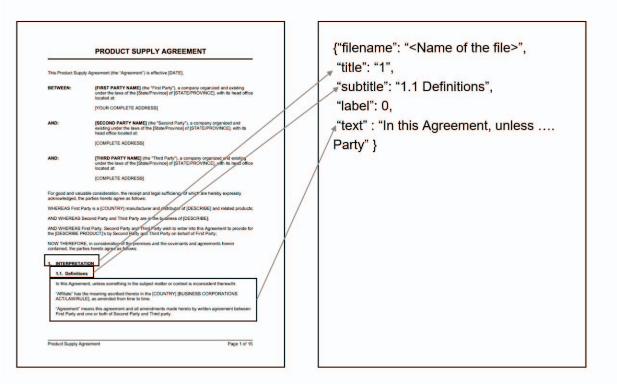


Fig. 1. Arbitrary samples of Contracts with (a) multilingual (b) header and footer (c) watermark (d) different format



{"filename": "<Name of the file>",
"title": "1",
"subtitle": "1.1 Definitions",
"label": 0,
"text" : "In this Agreement, unless ....
Party" }

Fig. 2. Illustration of the preprocessing from contracts (left) to Json format (right)

Each paragraph is annotated using regular expressions that are reviewed and moderated manually by domain experts. From the various topics being discussed in the contracts, we have selected seven critical topics that are of business priority and outlined them in Table I. Note that the contract may not necessary contain all the topics.

IV. Methodology

PERFECT (Prompt-free and Efficient paradigm for FEw-shot Cloze-based fine-Tuning) is a verbalizer and pattern free few-shot learning method that uses soft prompts and task-specific adapters [7] to enable sample-efficient fine-tuning.

Contract Type	Total supports
Payment Term	81
Warranty Period	81
Incoterm	81
Product Change Notification	74
Incoming Inspection	69
Order Response	29
Quantity tolerance	14

TABLE I
NUMBER OF TOPICS FOUND AT PARAGRAPH LEVELS

It leverages prototypical networks [15] for inference, which helps reduce the impact of imbalanced data. This approach allows for efficient utilization of computational resources while still attaining high performance in the target task. By selectively fine-tuning specific adapters in the pre-trained model, PERFECT effectively incorporates domain-specific knowledge from the target task, enhancing the model's capabilities to handle new data while minimizing the risk of catastrophic forgetting or overfitting. The structure of PERFECT is shown as Figure 3.

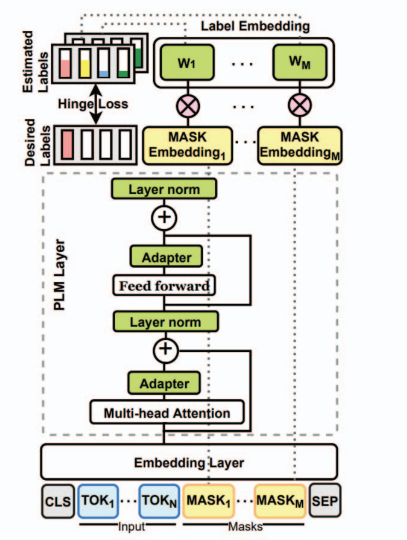


Fig. 3. PERFECT architecture

PERFECT is designed to be similar to the pre-training phase, while replacing handcrafted patterns and verbalizers with new components, namely task-specific adapters and design label embeddings for the classes, to describe the task and learn the labels. The underlying PLM model is fixed and only required to optimize the new parameters from label embeddings, adapters and layer norms (the green blocks), as illustrated in Figure 3.

Multi-class classification PERFECT [2] is implemented on top of the RoBERTa-base-CUAD, a RoBERTa based model with 125 parameters fined-tuned with CUAD [1] dataset. The main advantage of this approach is that the PTM is pre-trained based on legal domain data that is closest to supply chain contracts relevancy and trained with few-shot technique with low samples to customized further into our domain context.

We attempted to experiment with topic classification at the sentence level. Unfortunately, we encountered difficulties because some topics had content that spanned across multiple sentences. Breaking the paragraph into sentences caused the semantic meaning of the paragraph to be lost, making it challenging for the model to accurately capture the complete context. Ultimately, we determined to classify topics at paragraph level. This allowed the model to fully capture the meaning of the text and achieve reasonable results.

For our model training methodology, we closely followed the recommended hyperparameters and inference strategy from PERFECT [2]. However, due to the limited resources available for this work, we reduced the training batch size to 24. To improve the model's performance, we experimented with a learning rate higher than 10^{-4} as suggested by Mahabadi et al. [2]. Multiple train-test-split percentages were evaluated, and we ultimately chose to use a 60% training and 40% test set, stratified by each topic. This split was used to ensure the model could perform well in the real world, especially given the low number of samples for some of the topics.

V. RESULT AND DISCUSSION

For each paragraph, we measured the performance in term of precision, recall and F_1 score. Here, true positives (TP) are paragraphs correctly classified as the intended topics, false positives (FP) are paragraphs incorrectly classified and false negatives (FN) are paragraph incorrectly classified as not part of the topics of interest. F_1 (harmonic mean) is commonly used to combine precision and recall.

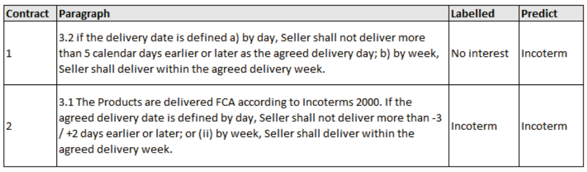
Contract Type	Precision	Recall	F_1-Score
Payment Term	0.90	0.81	0.85
Warranty Period	0.86	1.00	0.93
Incoterm	0.55	0.97	0.70
Product Change Notification	0.52	0.97	0.67
Incoming Inspection	0.91	0.71	0.80
Order Response	0.33	1.00	0.50
Quantity tolerance	0.08	0.17	0.11

TABLE II
PRECISION, RECALL, F_1-SCORE, MEASURED PER PARAGRAPHS ACROSS BSA CONTRACTS INSTANCES

The results are shown in Table II. Overall, the paragraphs are correctly classified to the intended topics with reasonable f1-score. The model accurately predicts incoming inspection, payment term, and warranty period. These topics have a strong correlation to specific sentence structures and word choices, with little ambiguity. For instance, the warranty period is typically indicated by the phrase "the warranty period shall be" , payment term by "full payment shall be received within", and incoming inspection by "products shall be inspected within". It's worth noting that despite the strong pattern, regex performance is still not on par with the model.

Although this approach has shown some initial promise, it doesn't work very well for certain selected topics. Specifically, in long paragraphs, the model tends to classify the paragraph as "no interest", as observed in cases such as Quantity Tolerance. This limitation may be due to the PTM's inability to handle long tokens, as the BERT model can only handle up to 512 tokens. Additionally, transformers with full attention, such as BERT, can be very costly to process long sequences [16].

Additionally, we have observed high recall but low precision in the Incoterm, Order Response, and Product Change Notification (PCN) topics. We have found that some contract paragraphs discussing Incoterm often mention delivery window as well. Depending on how these paragraphs are structured, the Incoterm topic may consider the semantic structure of the delivery window and misclassify paragraphs that only mention delivery window. As a result, the Incoterm topic captures some of the delivery window paragraphs, leading to low precision, as illustrated in Table III.



Contract	Paragraph	Labelled	Predict
1	3.2 if the delivery date is defined a) by day, Seller shall not deliver more than 5 calendar days earlier or later as the agreed delivery day; b) by week, Seller shall deliver within the agreed delivery week.	No interest	Incoterm
2	3.1 The Products are delivered FCA according to Incoterms 2000. If the agreed delivery date is defined by day, Seller shall not deliver more than -3 / +2 days earlier or later; or (ii) by week, Seller shall deliver within the agreed delivery week.	Incoterm	Incoterm

TABLE III
MODEL ACCURATELY CLASSIFIED INCOTERM FOR CONTRACT 2 BUT NOT FOR CONTRACT 1

A major component of business operations for business-to-business (B2B) companies is the receipt of orders from buyers and sending back of order responses on a regular basis. The order response topic in our context refers to the duration for the B2B company to reply to the customer with a delivery confirmation date. Unfortunately, the semantic structures of the paragraphs for order response and other topics are quite similar. As a result, the model has captured many false positive cases, leading to low precision. This is also observed for Product Change Notification (PCN) where the semantics captured might also be discovered for Product Discontinuation as both these topics are closely linked to product lifecycle processes and involve early customer communication.

VI. Conclusion and Future Work

We have shown the practical application of transfer learning with PTM to handle information extraction tasks in supply chain contracts under low-data regime scenario. Moving forward, we plan to further improve the results by experimenting with sparse attention mechanisms in PTMs such as BigBird and Longformer, to overcome the limitation of BERT's 512-word input. Additionally, a multi-label approach could be employed to tackle paragraphs that cover multiple topics. Further work is also needed to explore better state-of-the-art generalized approaches to preprocess contracts and handle the varied quality and layout, with the goal of expanding the application of these techniques to other contract types.

This approach could be expanded to extract more topics from the same contracts, scale up to include other contracts of similar application and improve overall business operations.

Acknowledgment

We would like to express our gratitude to Dr. Wang Aobo, a Lecturer and Consultant in the Artificial Intelligence Practice at the Institute of Systems Science at NUS, and Mr. Chen Chang, a Master of Science student in Data Science and Machine Learning at NUS, for their contributions to this work.

References

[1] Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. CUAD: an expert-annotated NLP dataset for legal contract review. *CoRR*, abs/2103.06268, 2021.

[2] Rabeeh Karimi Mahabadi, Luke Zettlemoyer, James Henderson, Lambert Mathias, Marzieh Saeidi, Veselin Stoyanov, and Majid Yazdani. Prompt-free and efficient few-shot learning with language models. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3638–3652, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[3] Haifeng Wang, Jiwei Li, Hua Wu, Eduard Hovy, and Yu Sun. Pre-trained language models and their applications. *Engineering*, 25:51–65, 2023.

[4] IDEX Consulting. 5 of the best lawtech start-ups, July 2019.

[5] Jiajia Duan, Hui Zhao, Qian Zhou, Meikang Qiu, and Meiqin Liu. A study of pre-trained language models in natural language processing. In *2020 IEEE International Conference on Smart Cloud (SmartCloud)*, pages 116–121, 2020.

[6] Xiaofeng Zhong, Shize Guo, Hong Shan, Liang Gao, Di Xue, and Nan Zhao. Feature-based transfer learning based on distribution similarity. *IEEE Access*, 6:35551–35557, 2018.

[7] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. *CoRR*, abs/1902.00751, 2019.

[8] Yuxian Gu, Xu Han, Zhiyuan Liu, and Minlie Huang. PPT: pre-trained prompt tuning for few-shot learning. *CoRR*, abs/2109.04332, 2021.

[9] Ning Ding, Yulin Chen, Xu Han, Guangwei Xu, Xiaobin Wang, Pengjun Xie, Haitao Zheng, Zhiyuan Liu, Juanzi Li, and Hong-Gee Kim. Prompt-learning for fine-grained entity typing. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6888–6901, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[10] Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. How does NLP benefit legal system: A summary of legal artificial intelligence. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5218–5230, Online, July 2020. Association for Computational Linguistics.

[11] Abhay Shukla, Paheli Bhattacharya, Soham Poddar, Rajdeep Mukherjee, Kripabandhu Ghosh, Pawan Goyal, and Saptarshi Ghosh. Legal case document summarization: Extractive and abstractive methods and their evaluation. In Yulan He, Heng Ji, Sujian Li, Yang Liu, and Chua-Hui Chang, editors, *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1048–1064, Online only, November 2022. Association for Computational Linguistics.

[12] Sarthak Dalal, Amit Singhal, and Brejesh Lall. Lexrank and pegasus transformer for summarization of legal documents. In Dilip Singh Sisodia, Lalit Garg, Ram Bilas Pachori, and M. Tanveer, editors, *Machine Intelligence Techniques for Data Analysis and Signal Processing*, pages 569–577, Singapore, 2023. Springer Nature Singapore.

[13] R. Smith. An overview of the tesseract ocr engine. In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, volume 2, pages 629–633, 2007.

[14] D.R. Vedhaviyassh, R. Sudhan, G. Saranya, M. Safa, and D. Arun. Comparative analysis of easyocr and tesseractocr for automatic license plate recognition using deep learning algorithm. In *2022 6th International Conference on Electronics, Communication and Aerospace Technology*, pages 966–971, 2022.

[15] Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning. *CoRR*, abs/1703.05175, 2017.

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.