Coordinate Structure Extraction for Patent Claims using Multilingual LLMs

Tsukasa Ishimaru[1], **Takehito Utsuro**[1], **Masaaki Nagata**[2]
[1] University of Tsukuba, Natural Language Processing Laboratory
s2210161_@_u.tsukuba.ac.jp, utsuro_@_iit.tsukuba.ac.jp
[2] NTT Corporation, NTT Communication Science Laboratories
masaaki.nagata_@_ntt.com

Abstract

This study proposes a simple, one-stage approach to coordinate structure extraction using multilingual Large Language Models (LLMs) with Translation between Augmented Natural Languages (TANL) to develop an error detection system for coordinate structure translation. Unlike conventional multi-component methods such as CoRec, our method employs an end-to-end Transformer decoder (LLM) trained via Continual Pre-Traning (CPT) and/or Supervised Fine-Tuning (SFT) on English and Japanese datasets obtained from parsed treebanks that includes coordinate structures. We evaluated the proposed models on 100 English and Japanese patent claims manually annotated with coordinate structure tags. The proposed method using open-weight models such as Llama-3.2-8B or gemma-3-4b-it significantly outperformed GPT-5 and CoRec by approximately 0.02-0.03 in F1 score for the English task. The proposed method using open-weight models such as llama-3-youko-8b and Llama-3-swallow-8B-0.1v significantly outperformed GPT-5 by approximately 0.02-0.05 in F1 score for the Japanese task. In addition, models using both English and Japanese training data significantly outperform those using monolingual training data only.

Keywords: coordinate structure extraction, multilingual large language models, patent claim translation

1. Introduction

Coordinate structures frequently appear in many documents. In particular, patent claims often include numerous coordinate structures within a document. According to our analysis, approximately 64% of the 1,035,219 patent claims in the 2020 JaParaPat (Nagata et al., 2024) dataset contain at least one coordinate conjunction, indicating the potential presence of coordinate structures. We consider such frequently occurring coordinate structures to cause many problems in patent claim translation, including omissions and inconsistencies between the coordinate structures before and after translation.

The goal of our research is to develop an error detection system for coordinate structure translation, as shown in Figure 1, which consists of coordinate structure extraction of the source and target languages and their alignment. As the first step in this approach, we focus on coordinate structure extraction in this paper.

To extract coordinate structure in a sentence, Wang et al. (2023) proposed CoReC[1] , the previous state-of-the-art method, which uses a transformer encoder (BERT (Devlin et al., 2019)) and performs two tasks in sequence: identifying coordinators and detecting conjunct boundaries. However, this method suffers from several drawbacks, as its multi-component design increases training complexity and prolongs training time. Each component processes coordinators independently, which prevents the model from capturing their interdependencies.

We propose a coordinate structure extraction method using multilingual Large Language Models (LLMs). The proposed method is an approach to end-to-end one-stage analysis using a transformer decoder (LLM) based on Translation between Augmented Natural Languages (TANL) (Paolini et al., 2021). TANL is a method that performs structure prediction as a translation from an input sentence to one with structure annotations. In the first step of the proposed method, we constructed datasets from English and Japanese parsed treebanks by annotating tags to coordinate structures. Next, using these datasets (input: source sentence, output: tagged sentence), we applied continual pre-training (CPT) (Ke et al., 2023) and/or supervised fine-tuning (SFT) (Zhang et al., 2023) to the LLMs as a TANL task, enabling them to learn coordinate structure extraction.

In the experiment, we evaluated the performance of coordinate structure extraction on 100 patent claims that we manually annotated with coordinate structure tags, measuring performance by the F1 score. The results show that applying CPT and/or SFT to LLMs significantly improves the F1 score for coordinate structure extraction compared to LLMs without CPT nor SFT. Parts of trained models, despite having only a single component, outperform the previous state-of-the-art method (CoRec). In addition, training on both English and Japanese data improved F1 score over

[1] https://github.com/qingwang-isu/CoRec.git

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 4931–4941
11–16 May 2026.

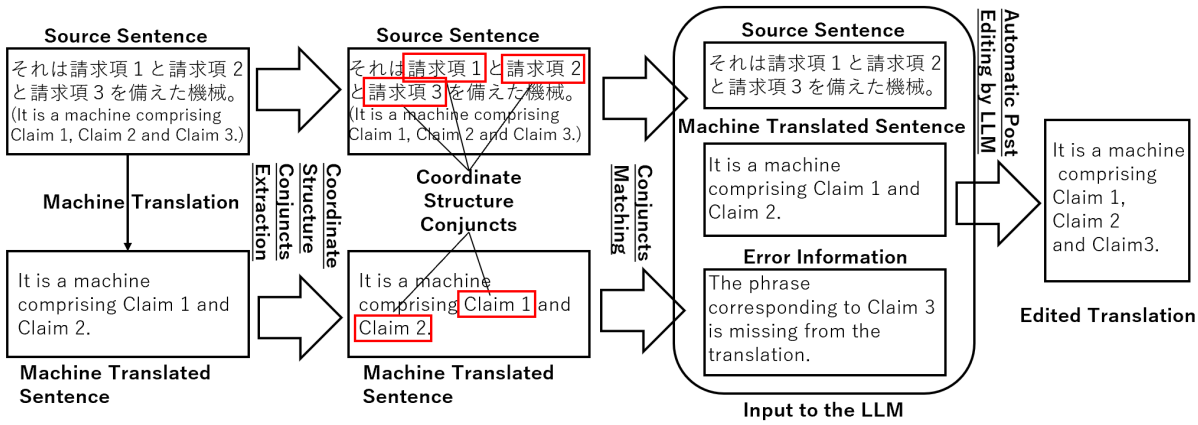


Figure 1: Automatic Post Editing of Patent Claim MT based on Coordinate Structure Conjuncts Matching

monolingual data training. In English extraction, SFT and CPT+SFT, while in Japanese extraction, CPT, SFT, and CPT+SFT significantly improved. In summary, our contributions are threefold:

- We demonstrated that end-to-end coordinate structure analysis can be easily achieved using TANL.
- We demonstrated that parts of the proposed models trained with CPT and SFT outperformed the previous state-of-the-art method (CoRec), proprietary models, and the models without CPT nor SFT.
- We demonstrated that extending the training data to English and Japanese improved the F1 score of coordinate structure extraction compared with using monolingual data.

2. Related Works

Many studies have explored the task of coordinate structure analysis. Ficler and Goldberg (2016b) proposed an LSTM-based neural network model for coordination boundary detection, centering on the concepts of similarity between conjuncts and coherence. However, this method only handles conjunct pairs, which naturally limits their ability to manage coordinate structures containing three or more conjuncts.

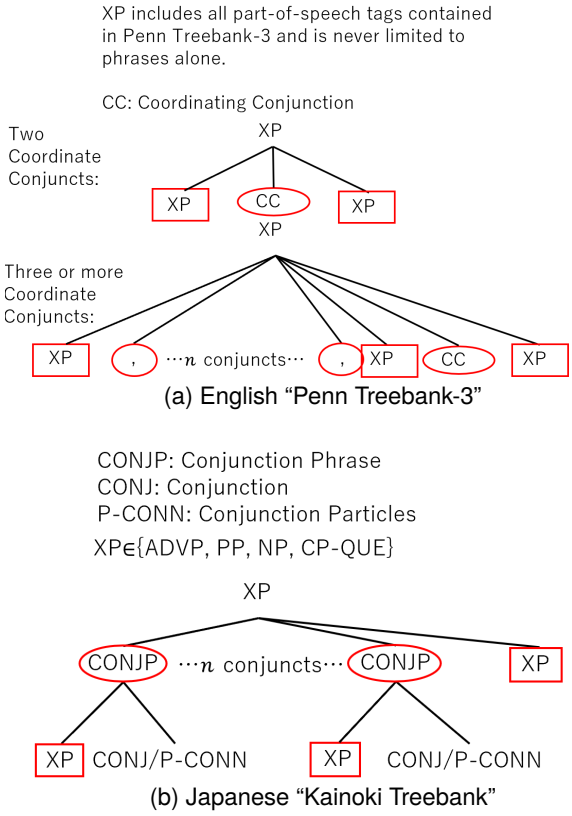
Teranishi et al. (2019) proposed a method that decomposes coordination identification into three local tasks, learns each task, and reconstructs a consistent structure using the CKY parsing algorithm (Younger, 1967). Wang et al. (2023) proposed CoRec, which uses BERT, consisting of two components: coordinator identification and conjunct boundary detection. This is the previous state-of-the-art method. However, these approaches still face challenges, including increased training complexity and a decline in F1 scores caused by the inability to capture the interdependency between those two components.

We propose an end-to-end, one-stage coordinate structure extraction model using an LLM based on TANL, unlike those previous approaches.

Paolini et al. (2021) proposed Translation between Augmented Natural Languages (TANL). They used TANL for structure prediction tasks such as named entity recognition, semantic role labeling, and coreference resolution. While conventional structural prediction tasks require training a dedicated task discriminator on top of BERT, TANL unifies all structural prediction tasks into text-to-text translation, enabling them to be processed using the same architecture.

Kawarada et al. (2024) proposed applying the TANL framework to argument structure analysis. They formulate argument mining (AM) as a text generation task. It adopts an approach to handling three stages of argument span identification, component classification, and relation classification, where unifying them within a single encoder-decoder models (T5 (Raffel et al., 2020)/FLAN-T5 (Chung et al., 2022), completing structural prediction through text generation.

Relying on those research findings, we propose an approach to extracting coordinate structures based on the TANL framework. Unlike previous studies, we implemented TANL using a transformer decoder (LLM) instead of an encoder–decoder model such as T5/FLAN-T5, incorporating distinct training techniques such as CPT and/or SFT.



(a) English "Penn Treebank-3"

(b) Japanese "Kainoki Treebank"

Figure 2: Coordinate Structures of English and Japanese Treebanks with n Conjuncts

3. Monolingual Coordinate Structure Conjuncts Extraction from Patent Claim

3.1. Overview

We aim to extract the coordinate structure contained within English and Japanese patent claims. We propose a method involving the following steps: (1) We determine the coordinate structures based on English and Japanese treebanks, as shown in Figure 2, then construct English and Japanese datasets containing prompt, source sentence and sentence with coordinate structure tags, as shown in the center of Figure 3. (2) We train several LLMs on the English dataset and similarly train several LLMs on the Japanese dataset, based on TANL. (3) We train a multilingual LLM using a mixture of English and Japanese datasets, based on TANL.

3.2. Developing Training Data from Monolingual Treebanks

Table 1 shows the number of sentences in the training and evaluation datasets constructed. We prepared separate datasets of 7,569 sentences each for English and Japanese. To train the multilingual model, we combined both datasets, resulting in a total of 15,138 sentences. In addition, to change the data used for CPT and SFT, we created one additional dataset for each. In section 3.4, we performed both CPT and SFT twice on each dataset, and we report the average scores as the final results in section 4.

	Training		Evaluation	
	En	Ja	En	Ja
Sentence with Coordinate Structures	5,329	5,329	80	80
Sentence without Coordinate Structures	2,240	2,240	20	20
Total	7,569	7,569	100	100

Table 1: # Sentences of Training and Evaluation Datasets

3.2.1. English Training Data

We used "Penn Treebank-3 (Marcus et al., 1999)" as English training data resource. Figure 2a shows the coordinate structures we extracted from Penn Treebank-3. The details of the coordinate structures are as follows. We define the minimal coordinate structure when sibling nodes XP before and after a CC (Coordinating Conjunction) tag share the same label. If the preceding sibling node uses ‘’,’‘, we also treat the node XP before it as a coordinate conjunct, increasing the number of coordinate conjuncts according to the number of ‘’,’‘ used. The XP label includes all part-of-speech labels from the Penn Treebank-3, while the CC label contains only the words "and" and "or".

This definition differs from that of Ficler and Goldberg (2016a), who adopted a broader view of coordination. Their definition includes not only conjuncts and coordinators but also additional elements such as shared modifiers, shared arguments, markers (e.g., both, either), and parenthetical phrases. In contrast, we focus on the minimal parsed coordinate structure determined by identical constituent labels surrounding a CC tag, providing a more rule-based and structurally consistent criterion for identifying coordinated elements.

Next, we insert tags into the coordinate structures. Figures 3 and 4a illustrate our tagging methods for non-nested and nested coordinate structures, respectively. In this way, we construct a dataset containing a prompt, a source sentence without tags, and an English sentence with coordinate structure tags for training LLMs. Figure 3 shows the prompt we use.

3.2.2. Japanese Training Data

We used "Kainoki Treebank (Kainoki, 2022)" as a Japanese training data resource. Figure 2b shows the coordinate structures we extracted from Kainoki Treebank. The details of the coordinate structures are as follows. We define the minimal configuration of a coordinate structure such that the

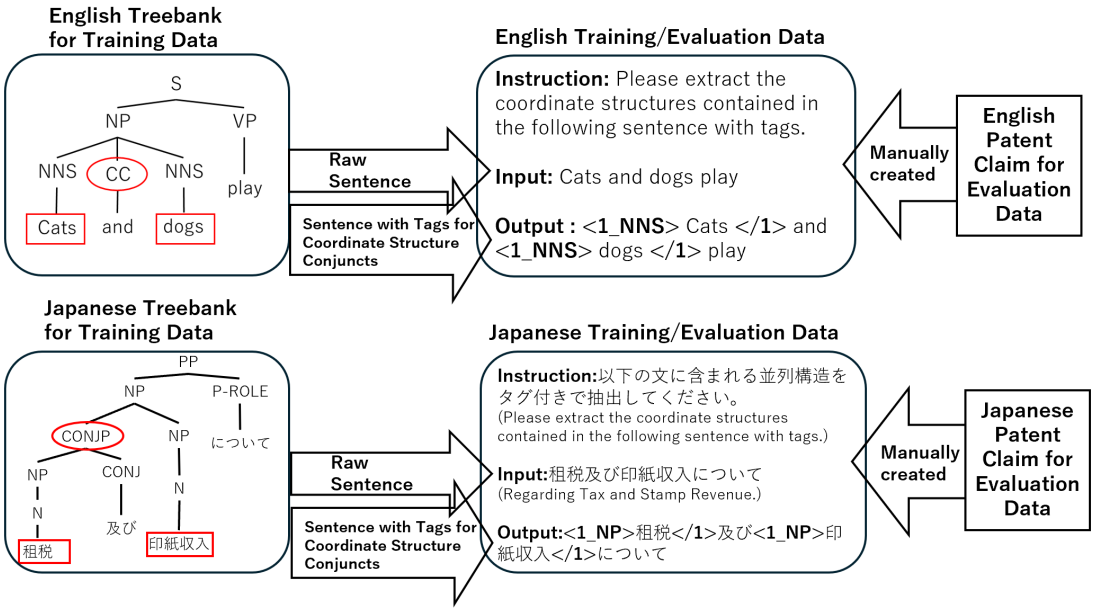


Figure 3: Overview of Developing English and Japanese Training/Evaluation Datasets of LLMs for Coordinate Structure Conjuncts Extraction (examples with a non-nested coordinate structure)

child node XP of a CONJP (Conjunction Phrase) node and the sibling node XP of the CONJP form a coordinate structure, and we increase the number of coordinate conjuncts according to the number of CONJP nodes contained in the sibling node. We limit the labels included in XP to ADVP (Adverbial Phrases), PP (Partitive Phrases), NP (Noun Phrases), and CP-QUE (Interrogative Clauses).

Next, we insert tags into the coordinate structures. Figures 3 and 4b illustrate our tagging methods for non-nested and nested coordinate structures, respectively. In this way, we construct a dataset containing a prompt, a source sentence without tags, and a Japanese sentence with coordinate structure tags for training LLMs. Figure 3 shows the prompt we use.

3.2.3. Bilingual Training Data

We also create a bilingual coordinate structure extraction model. To train the multilingual model, We prepare datasets of 7,569 English sentences and 7,569 Japanese sentences, totaling 15,138 sentences, as mentioned above. In this training data, we randomize the order of the English and Japanese sentences so that we randomize the order of languages in bilingual training.

3.3. Evaluation Data from Patent Claims

Figure 3 illustrates how we create the evaluation dataset. We construct the evaluation data using patent claims published in WIPO (World Intellectual Property Organization (WIPO), 2025) (English patents, 2024-2025) and J-PlatPat (for Patent Information , J-PlatPat) (Japanese patents, 2020-2025).

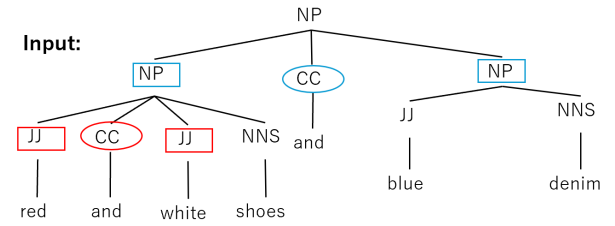
We manually create all evaluation data and apply the same coordinate structure tagging method used for the training data in each language. We also use prompts in the same format as the training data during model input. The evaluation datasets contain 100 sentences each for English and Japanese, including sentences with/without coordinate structures. Table 1 shows the composition of the English and Japanese evaluation datasets.

To ensure the objectivity of the manual evaluation, we conducted an independent annotation study with a second annotator in addition to the primary annotator. We randomly selected a total of 100 sentences from the evaluation dataset (50 Japanese and 50 English) and measured agreement. The results showed high agreement, with F1 scores of 0.8029 for English and 0.8657 for Japanese. These findings indicate that the annotators formed strong consensus regarding the identification criteria for coordinate structures defined in this study, thereby supporting the robustness of the evaluation dataset.

3.4. Training

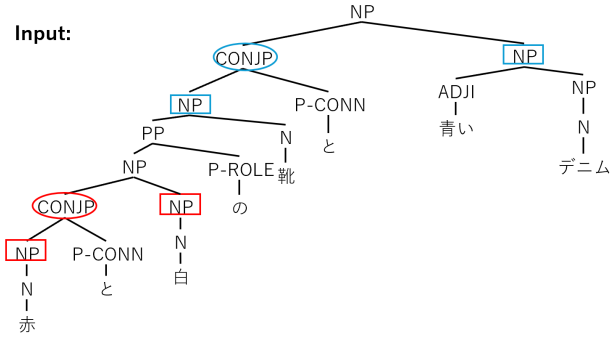
3.4.1. Continual Pre-Training

We first performed Continual Pre-Training (CPT) (Ke et al., 2023) as the model training method. We apply CPT to adapt LLMs to coordinate structures as a structural prediction task



Output: <1_NP> <2_NP> red </2> and <2_NP> white </2> shoes </1> and <1_NP> blue denim </1>

(a) English



Output: <1_NP><2_NP>赤</2>と<2_NP>白</2>の靴</1>と <1_NP>青いデニム</1>

(b) Japanese

Figure 4: Nested Coordinate Structures of English and Japanese Treebanks with n conjuncts (an input from a treebank and an output of a sentence with coordinate structure conjunct tags)

framed as translation between augmented natural languages (TANL (Paolini et al., 2021)). The purpose of CPT is to strengthens the model's ability to understand the long and complex cooperative structures frequently found in patent claims, enabling it to more accurately capture the relationships between the conjuncts that constitute these coordinate structures. We adopt QLoRA (Dettmers et al., 2023), following the approach of Kawarada et al. (2024), to reduce GPU memory usage during training of large-parameter models. QLoRA quantizes the model and applies a Low-Rank Adapter (LoRA) (Hu et al., 2021), reducing the number of parameters the model trains while maintaining performance comparable to full fine-tuning. We perform CPT on the following models in this experiment:

- We used Meta-Llama-3-8B[2], llama-3.2-3B[3], and Mistral-7B-v0.1[4] to extract English coordinate structures.
- We used llama-3-youko-8b[5], Llama-3-Swallow-8B-v0.1[6], and Llama-3-ELYZA-JP-8B[7] to extract Japanese coordinate structures.
- We used gemma-3-4b-it[8] to extract both English and Japanese coordinate structures.

We selected the above models for their optimized performance and parsing capabilities across languages, prioritizing open-source reproducibility. We used the 7,569 sentences English and Japanese training datasets shown in Table 1 for CPT in the monolingual extraction models. For gemma-3-4b-it, we combined the English and Japanese datasets, each containing 7,569 sentences, resulting in a total of 15,138 sentences, as described in section 3.2.3.

3.4.2. Supervised Fine-Tuning

We performed Supervised Fine-Tuning (SFT) (Zhang et al., 2023) as the model training method. We use SFT to frame coordinate structure prediction as TANL, optimizing the general knowledge LLMs typically possess for the specific task of coordinate structure extraction. The purpose of SFT is to adapt the model to generate responses in the expected output format for given input sentences according to provided instructions. In SFT, we adopt QLoRA as described in section 3.4.1.

We used the same models for SFT as those in section 3.4.1. We performed SFT on both models without CPT and models with CPT to verify whether combining domain adaptation via CPT with task adaptation via SFT produces a synergistic effect on coordinate structure extraction performance. We used different dataset for SFT than those used for CPT. For the monolingual extraction models, we use the 7,569 sentences English and Japanese training datasets shown in Table 1. For gemma-3-4b-it, we combine the English and Japanese datasets, each containing 7,569 sentences, totaling 15,138 sentences, as described in section 3.2.3.

3.4.3. Model Architecture

In this study, we adopted a decoder-only LLM within the TANL framework for two reasons.

[2] https://huggingface.co/meta-llama/Meta-Llama-3-8B

[3] https://huggingface.co/meta-llama/Llama-3.2-3B

[4] https://huggingface.co/mistralai/Mistral-7B-v0.1

[5] https://huggingface.co/rinna/llama-3-youko-8b

[6] https://huggingface.co/tokyotech-llm/Llama-3-Swallow-8B-v0.1

[7] https://huggingface.co/elyza/Llama-3-ELYZA-JP-8B

[8] https://huggingface.co/google/gemma-3-4b-it

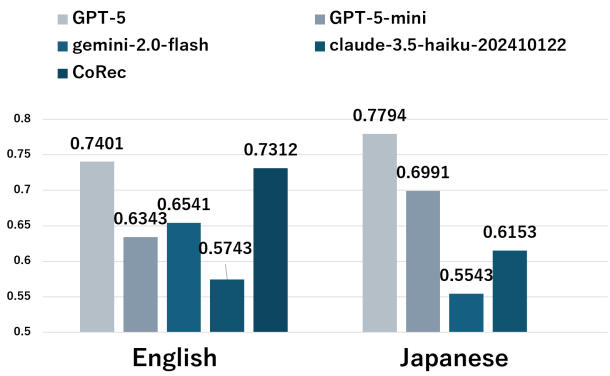


Figure 5: F1 Scores of Coordinate Structure Extraction by Proprietary Models and CoRec (Wang et al., 2023)

First, modern LLMs possess extensive multilingual knowledge and contextual understanding, making them ideal for processing the complex, long-form syntax of patent claims.

Second, the TANL approach "translating raw text into structured format" aligns perfectly with the instruction-following capabilities of current LLMs, ensuring high-precision extraction of coordinate structures.

4. Experiment and Discussion

4.1. The Procedure

We evaluated various models with and without CPT and SFT, as well as baseline models using F1 score based on 100 patent claims manually annotated with coordinate structure tags in English and Japanese, as described in section 3.3. We calculated the F1 score using spans corresponding to coordinate conjuncts as the unit, ignoring labels. To verify whether the proposed model significantly outperforms the baseline, we performed a t-test (Student, 1908) at significance levels of 0.05 and 0.01.

As baseline models for English, we used several proprietary models and CoRec, which is the previous state-of-the-art model by Wang et al. (2023)[9]. Both CoRec and the proposed method are trained on datasets outside the patent domain. Although we evaluate the models on patent claims, this setting applies equally to both approaches, and therefore the comparison remains fair. Since CoRec only supports English, and we could not find any publicly available model for Japanese coordinate structure extraction, we used only proprietary models as the Japanese baseline models in this study.

We applied a 3-shot prompting setup for the models without CPT or SFT (denoted as "w/o training") and for the baselines, while we used a 0-shot prompting setup for the models with CPT and/or SFT. Each prompt contained three representative sentence types for both English and Japanese: (1) A sentence containing a nested coordinate structure. (2) A sentence containing a non-nested coordinate structure. (3) A sentence without any coordinate structure. This configuration allows us to evaluate how well the model generalizes to the structural complexity of coordinate conjuncts. We excerpt the English and Japanese prompts to avoid excessive length, showing only a single English prompt as an example.

Example of 3-shot Prompt of Baselines and w/o Training Models

"**instruction**": "Please extract the coordinate structures contained in the following sentence with tags."
"**input**": "IBM's visits to the debt market are closely watched by treasurers at other corporations and by credit market analysts."
"**output**": "IBM's visits to the debt market are closely watched <1_PP> by treasurers at other corporations </1> and <1_PP> by credit market analysts </1>."
. . .
"**output**":

As proprietary models, we used GPT-5/-mini[10], gemini-2.0-flash[11], and claude-3-5-haiku-20241022[12]. Figure 5 shows the F1 scores of proprietary models, where those of the proprietary models are averaged over five inferences each. For the proposed models and CoRec, we performed a single inference to obtain evaluation results, since repeated inferences produce identical results. In the next section, among those proprietary models and CoRec, the one with the highest F1 score is adopted as the baseline and compared with the F1 scores of the proposed models.

4.2. Monolingual Training

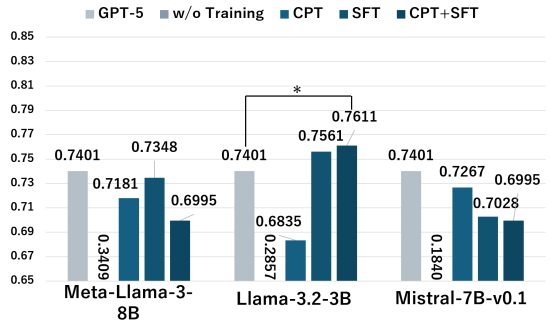
Figure 6a compares the F1 scores for the English coordinate structure extraction task among the baseline, models without CPT nor SFT (denoted as "w/o training"), and models trained with CPT and/or SFT. We observed significant improvements in the models that applied CPT+SFT to

[9] https://github.com/qingwang-isu/CoRec.git

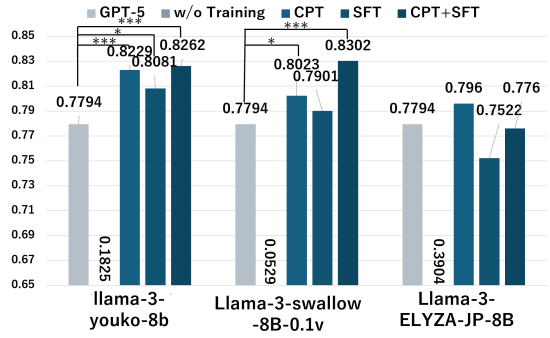
[10] https://openai.com/gpt-5

[11] https://blog.google/technology/google-deepmind/gemini-model-updates-february-2025/

[12] https://claude.ai



(a) English



(b) Japanese

Figure 6: F1 Scores of Coordinate Structure Extraction by Monolingual Models (baselines as GPT-5 in English and Japanese. $***$ as significance level at $p < 0.01$ and $*$ as at $p < 0.05$.)

Llama-3.2-3B compared to baseline. The Llama-3.2-3B model achieved the highest performance with CPT+SFT, attaining an F1 score of 0.7611. While CPT+SFT produced a synergistic effect in Llama-3.2-3B, we did not observe such an effect in the Meta-Llama-3-8B and Mistral-7B-0.1v models.

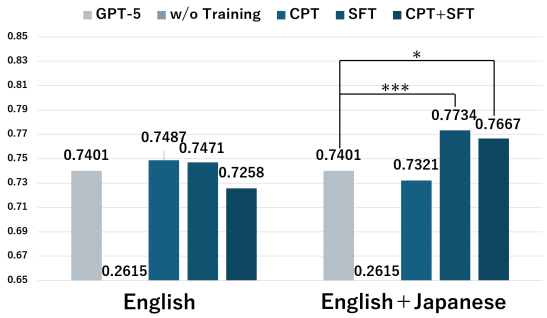
These results suggest that the presence or absence of knowledge related to coordinate structures and the potential for synergy from CPT+SFT vary depending on the LLM used in the experiment.

Figure 6b compares the F1 scores for Japanese coordinate structure extraction task among the baseline, models without CPT nor SFT (denoted as "w/o training"), and models trained with CPT and/or SFT. We observed significant improvements in the models that applied CPT alone, SFT alone and CPT + SFT to Llama-3-youko-8b, and in the models that applied CPT alone and CPT+SFT to Llama-3-swallow-8B-0.1v, compared to baseline. The Llama-3-swallow-8B-0.1v model achieved the highest performance with CPT+SFT, attaining an F1 score of 0.8302. While CPT+SFT produced a synergistic effect in Llama-3-youko-8b and Llama-3-swallow-8B-0.1v, the Llama-3-ELYZA-JP-8B model performs better with CPT alone.

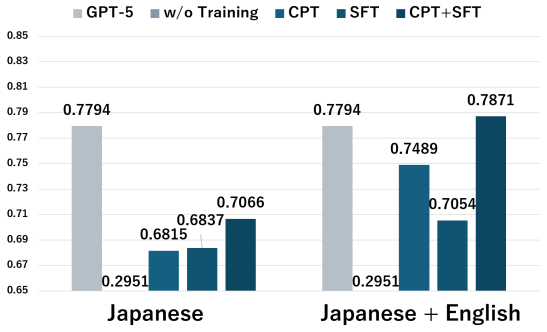
For all models, those trained only with SFT record the lowest F1 scores among the training methods applied to the same model. These results suggest that the LLMs in our experiments could not acquire sufficient knowledge about coordinate structures through SFT alone.

In both the English and Japanese tasks, training methods produce performance differences that depend on the model, indicating that we must tailor the optimal training strategy to the individual properties of each LLM. For all proposed models, the majority of extraction errors were either omissions or errors in the boundaries of modifiers.

4.3. Bilingual Training



(a) English

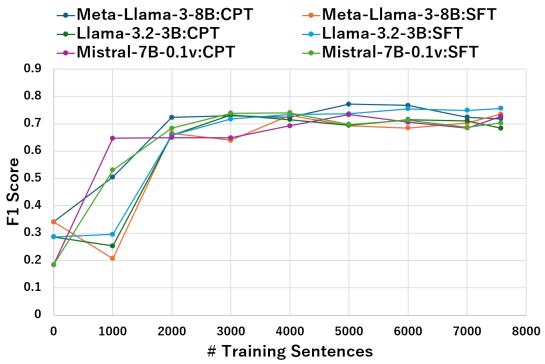


(b) Japanese

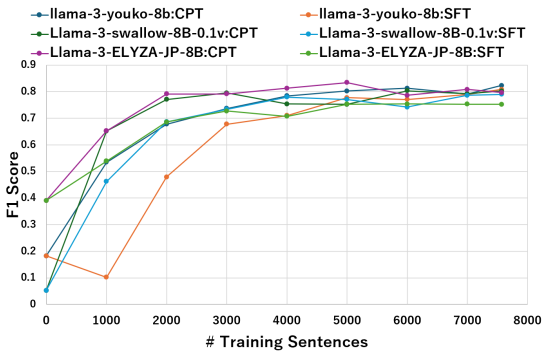
Figure 7: F1 Scores of Coordinate Structure Extraction by gemma-3-4b-it (baselines as GPT-5 in English and Japanese. $***$ as significance level at $p < 0.01$ and $*$ as at $p < 0.05$.)

Figure 7a compares the F1 scores for English coordinate structure extraction task among the baseline, models (gemma-3-4b-it) without CPT nor SFT (denoted as "w/o training"), and models trained with CPT and/or SFT. We observed significant improvements in models that applied SFT alone and CPT+SFT on the bilingual training dataset compared to baseline. The model achieved the highest performance when we applied SFT alone on the bilingual training dataset, attaining an F1 score of 0.7734.

To compare with models trained on the English dataset, models trained on the bilingual dataset

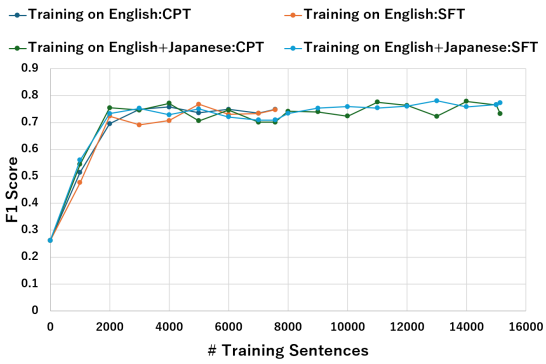


(a) English

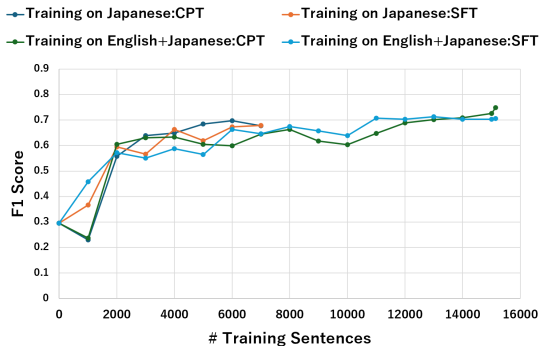


(b) Japanese

Figure 8: Changes in F1 Scores per # Training Sentences in Monolingual Extraction Models



(a) English



(b) Japanese

Figure 9: Changes in F1 Scores per # Training Sentences in gemma-3-4b-it

show significant improvements for both SFT alone and CPT+SFT. These results suggest that training models on the bilingual dataset allows them to learn parsed knowledge and expression patterns from Japanese coordinate structures, potentially enhancing performance on English tasks for some models.

The models that applied CPT alone show decreased F1 scores. We attribute this decrease to the fact that models already pre-trained on much more English data treat the newly added Japanese coordinate structure knowledge (without SFT) as noise.

Figure 7b compares the F1 scores for Japanese coordinate structure extraction task among baseline, models (gemma-3-4b-it) without CPT nor SFT (denoted as "w/o training"), and models trained with CPT and/or SFT. We did not observe any significant difference compared to baseline. The model achieved the highest performance when we applied CPT+SFT on the bilingual training dataset, attaining an F1 score of 0.7871.

We observed significant improvements in the models that applied CPT alone and CPT+SFT on the bilingual dataset compared to models trained on the Japanese dataset. These results suggest that gemma-3-4b-it initially lacks sufficient knowledge of Japanese coordinate structures, and training with bilingual data helps it generalize that knowledge. We did not observe any significant improvement in models that apply SFT alone on the bilingual training dataset. We attribute this result to gemma-3-4b-it lacking sufficient knowledge of Japanese coordinate structures, so simply adapting the output format using bilingual dataset (without CPT) cannot help it acquire this knowledge.

4.4. Learning Curve

Figure 8 shows the learning curves illustrating changes in F1 scores for the English and Japanese coordinate structure extraction tasks performed by monolingual models as we increased the number of training sentences in increments of 1,000. Comparing Figures 8a and 8b , we observe that English tends to reach saturation earlier than Japanese, and the number of training sentences required for convergence varies across models. In both English and Japanese, CPT raised the F1 score more rapidly than SFT. We attribute this trend to the models' limited prior knowledge of coordinate structures; when the training data are small in quantity, providing knowledge about coordinate structures helps the models extract them more effectively than merely adapting the output format. The final F1 scores remain

nearly constant, indicating that the amount of training data used in the experiments was sufficient.

Figure 9 shows the learning curves illustrating changes in F1 scores for the English and Japanese coordinate structure extraction tasks performed by gemma-3-4b-it as we increased the number of sentences in the monolingual and multilingual training datasets in increments of 1,000. Comparing Figures 9a and 9b, we observe that English tends to reach saturation earlier than Japanese, and the number of training sentences required for convergence varies depending on the training method. While the F1 scores for the English task remain almost constant at the final stage, those for the Japanese task show a slight improvement when trained on multilingual data, suggesting that the number of data used in the experiments may not have been sufficient.

4.5. Error Analysis

For all proposed models, the majority of extraction errors were either omissions or errors in the boundaries of modifiers. While specific countermeasures for deep nesting remain a subject for future research, we argue that a certain degree of error in modifier boundaries is acceptable in the context of our study. This is because the subsequent step toward our goal involves aligning coordinate phrases between the source and translated texts; therefore, boundary errors are tolerable as long as they do not preclude successful alignment.

The primary reason why the English model's F1 score falls below 0.8 lies in the difficulty of boundary decisions, particularly regarding the presence or absence of articles such as "the." In the context of our objective—aligning coordinate structure conjunction—the impact of such minor boundary discrepancies is limited. These small differences do not hinder semantic alignment and can therefore be regarded as acceptable errors. In many cases, the model correctly identified the core components of the coordinate structure, but subtle boundary shifts prevented the output from satisfying the strict "Exact Match" criterion.

One possible reason why CPT+SFT did not always yield the best results lies in our use of QLoRA during training. In this study, we adopted QLoRA due to resource constraints and to remain consistent with the experimental setting of TANL. However, because QLoRA updates only specific low-rank matrices, it allows far fewer trainable parameters than full-parameter training. As a result, the CPT+SFT training process may have been less stable.

5. Conclusion

In conclusion, we propose an effective approach for coordinate structure extraction. We implement TANL using an LLM and demonstrate that it can perform coordinate structure extraction. We showed that models trained with CPT and/or SFT achieve higher F1 scores than previous state-of-the-art methods (CoRec (Wang et al., 2023)), proprietary models, and open-weight instruction-tuned models trained without CPT or SFT. In addition, we showed that extending the training data to English and Japanese languages improves the F1 scores of coordinate structure extraction compared with using monolingual data.

6. Ethics Statement

This study carefully adheres to ethical research standards regarding data usage, model handling, and reproducibility. All datasets used in the experiments were derived from publicly available patent data and treebanks. The Penn Treebank-3 was officially obtained through the authorized distribution process and used in compliance with its license agreement. None of the data used in this study contains any personally identifiable information.

The large language models (LLMs) used in the experiments were employed solely for the purpose of this research and not for any applications beyond the stated objectives. We have taken appropriate precautions to prevent the misuse of data, trained models, and outputs derived from this study. All experimental procedures were conducted with respect for ethical standards in data handling, ensuring transparency and reproducibility while minimizing potential risks associated with LLM utilization.

We plan to release the dataset and the trained models at the camera-ready stage to promote transparency and reproducibility. No personally identifiable or sensitive information is included in the data.

7. Limitations

This study focuses on patent claim data in English and Japanese, and experiments on other languages or document genres have not yet been conducted.

The proposed model occasionally fails to extract deeply nested coordinated structures and may misidentify the boundaries of modifiers.

Currently, no large-scale dataset explicitly annotates coordinate structures in patent claims. Therefore, we constructed an evaluation dataset consisting of only 100 manually annotated sentences.

This limited size may not sufficiently reflect the true performance of the model.

In future work, we plan to conduct experiments on additional languages to further examine the generalization capability of our method.

8. Bibliographical References

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling Instruction-Finetuned Language Models. *arXiv preprint arxiv:2210.11416*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient Finetuning of Quantized LLMs. *arxiv preprint arxiv:2305.14314*, 37.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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. Association for Computational Linguistics.

Jessica Ficler and Yoav Goldberg. 2016a. Coordination annotation extension in the Penn Tree Bank. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 834–842, Berlin, Germany. Association for Computational Linguistics.

Jessica Ficler and Yoav Goldberg. 2016b. A neural network for coordination boundary prediction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 23–32, Austin, Texas. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv preprint arXiv:2106.09685*.

Masayuki Kawarada, Tsutomu Hirao, Wataru Uchida, and Masaaki Nagata. 2024. Argument Mining as a Text-to-Text Generation Task. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2002–2014, St. Julian' s, Malta. Association for Computational Linguistics.

Zixuan Ke, Yijia Shao, Haowei Lin, Tatsuya Konishi, Gyuhak Kim, and Bing Liu. 2023. Continual Pre-training of Language Models. *arXiv preprint arxiv:2106.09685*.

Masaaki Nagata, Makoto Morishita, Katsuki Chousa, and Norihito Yasuda. 2024. JaParaPat: A large-scale Japanese-English parallel patent application corpus. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9452–9462, Torino, Italia. ELRA and ICCL.

Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, Rishita Anubhai, Cicero Nogueira dos Santos, Bing Xiang, and Stefano Soatto. 2021. Structured Prediction as Translation between Augmented Natural Languages. *arXiv preptint arxiv:2101.05779*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Student. 1908. The probable error of a mean. *Biometrika*, 6(1):1–25.

Hiroki Teranishi, Hiroyuki Shindo, and Yuji Matsumoto. 2019. Decomposed local models for coordinate structure parsing. 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 3394–3403, Minneapolis, Minnesota. Association for Computational Linguistics.

Qing Wang, Haojie Jia, Wenfei Song, and Qi Li. 2023. CoRec: An easy approach for coordination recognition. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15112–15120, Singapore. Association for Computational Linguistics.

Daniel H Younger. 1967. Recognition and parsing of context-free languages in time n^3. *Information and Control*, 10(2):189–208.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, and Fei and Wu. 2023. Instruction Tuning for Large Language Models: A Survey. *arXiv preprint arXiv:2308.10792*.

9. Language Resource References

Japan Platform for Patent Information (J-PlatPat). 2025. J-platpat database: Japanese patent and utility model information. Accessed: 10 October 2025.

Ed Kainoki. 2022. Kainoki, ed (2022) "The Kainoki Treebank -a parsed corpus of contemporary Japanese" https://kainoki.github.io (accessed 9 octover 2025). Accessed: 9 October 2025.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1999. Penn treebank-3. Distributed via LDC: LDC99T42, Accessed: 9 October 2025.

World Intellectual Property Organization (WIPO). 2025. Patentscope database: International patent applications. Accessed: 9 October 2025.