Hunt for Buried Treasures: Extracting Unclaimed Embodiments from Patent Specifications

Chikara Hashimoto[*] **Gautam Kumar**[*] **Shuichiro Hashimoto**[*] **Jun Suzuki**[§]
[*] Rakuten Institute of Technology, Rakuten Group, Inc.
[§] Center for Data-driven Science and Artificial Intelligence, Tohoku University
chikara.hashimoto@rakuten.com

Abstract

Patent applicants write patent specifications that describe embodiments of inventions. Some embodiments are claimed for a patent, while others may be unclaimed due to strategic considerations. Unclaimed embodiments may be extracted by applicants later and claimed in continuing applications to gain advantages over competitors. Despite being essential for corporate intellectual property (IP) strategies, unclaimed embodiment extraction is conducted manually, and little research has been conducted on its automation. This paper presents a novel task of unclaimed embodiment extraction (UEE) and a novel dataset for the task. Our experiments with Transformer-based models demonstrated that the task was challenging as it required conducting natural language inference on patent specifications, which consisted of technical, long, syntactically and semantically involved sentences. We release the dataset and code to foster this new area of research.[1]

1 Introduction

Patents provide inventors the right to exclude others from using their inventions in exchange for disclosing how to make and use inventions by writing patent specifications. Patents have thus incentivized innovation and benefited industries. Given the increasing number of patent applications even during the COVID-19 pandemic (WIPO, 2022b), it is important to streamline patent application processes with technologies.

A patent specification describes an invention (Figure 1) by specifying one or more ways of embodying the invention, so that people skilled in the art can make and use it. A patent specification also contains claims that specify which embodiment applicants want to patent by stating the technical features necessary for the embodiment. Here, an *invention* refers to a mental construct inside the mind of the inventor, while the *embodiment* of the invention is a physical form of the invention and *claims* protect the embodiments (WIPO, 2022a).

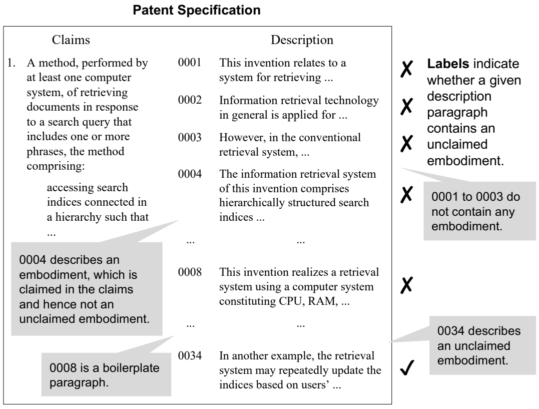

Patent Specification
Claims
Description
1. A method, performed by at least one computer system, of retrieving documents in response to a search query that includes one or more phrases, the method comprising:
accessing search indices connected in a hierarchy such that ...
0001 This invention relates to a system for retrieving ...
0002 Information retrieval technology in general is applied for ...
0003 However, in the conventional retrieval system, ...
0004 The information retrieval system of this invention comprises hierarchically structured search indices ...
... ...
0008 This invention realizes a retrieval system using a computer system constituting CPU, RAM, ...
... ...
0034 In another example, the retrieval system may repeatedly update the indices based on users' ...
✗
✗
✗
✗
✗
✓
Labels indicate whether a given description paragraph contains an unclaimed embodiment.
0001 to 0003 do not contain any embodiment.
0004 describes an embodiment, which is claimed in the claims and hence not an unclaimed embodiment.
0008 is a boilerplate paragraph.
0034 describes an unclaimed embodiment.

Figure 1: Illustration of the unclaimed embodiment extraction (UEE) task. A description paragraph is labeled to indicate if it has an unclaimed embodiment. Our dataset is in Japanese, though this example is written in English for illustration purposes.

A patent specification may describe a variety of embodiments, some of which may be unclaimed because claiming too diverse embodiments in a patent application may violate the *unity of invention*, a requirement for a patent application to relate to one invention only or to a group of inventions so linked as to form a single general inventive concept (USPTO, 2020b). *Continuing application* could be utilized later to claim those unclaimed embodiments in the prior patent application (the *parent* application). A continuing application can claim any embodiments if they are written in its parent's description. Moreover, the filing date of continuing application is the same as its parent's, even if it is filed years after the parent. Applicants can therefore utilize continuing applications strategically by, for instance, writing as many diverse embodiments as possible in the parent application and filing a continuing application to claim unclaimed embodiments in the parent. If the continuing application

[1] https://github.com/rakutentech/UEE_ACL23

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 5: Industry Track, pages 25–36
July 10-12, 2023 ©2023 Association for Computational Linguistics

does not exhaust its parent's embodiments, applicants may have further continuing applications. In so doing, applicants can adapt the claims of continuing application to new products and services of their company, and even new products and services of their competitors, enhancing their industrial competitiveness.

Continuing application requires extracting unclaimed embodiments from a patent specification. This is tedious as it requires understanding a wide variety of embodiments that are strategically arranged in the patent specification, a legal, technical document that may consist of thousands of tokens (Tab 1). **U**nclaimed **E**mbodiment **E**xtraction (**UEE**) has nonetheless been conducted manually without any technological support, and little research has been conducted on UEE.

This paper introduces the novel task of UEE (Figure 1) and the first publicly available dataset for UEE. Besides its practical utility, UEE poses a **new NLP challenge** as it involves two decisions to make (§2), one of which, i.e. *decision (ii)*, requires matching embodiment text in the description with claims to see if the embodiment has been claimed. Decision (ii) can be seen as a real-world natural language inference (NLI) (Bowman et al., 2015), where the hypothesis is a description paragraph and the premise is a set of claims. Although there have been studies on NLI for real-world applications (Holzenberger et al., 2020; Koreeda and Manning, 2021), decision (ii) involves a novel real-world NLI due to the following challenge: The hypothesis and the premise may consist of multiple long sentences which are written in *patentese* and full of technical terms in the target domain and whose syntactic and semantic structures are hard to recognize for non-IP specialists (Ferraro et al., 2014).

Although our UEE dataset has been created based on Japanese patents, extracting unclaimed embodiments from patent specifications is conducted in other countries such as the U.S. This paper gives examples in English for ease of explanation. See Appendices for Japanese examples.

Our **contributions** are as follows:

1. We introduce UEE, a novel, real-world NLP challenge.
2. We create and release the first dataset for UEE.
3. We conducted UEE experiments to demonstrate its difficulty.
4. We release code for reproducibility.

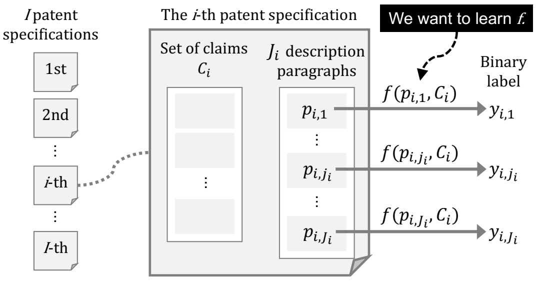


Figure 2: Illustration of the formal definition of UEE.

2 Task

Given a patent specification that comprises a set of claims and a set of description paragraphs about an invention, we want to determine whether each paragraph in the description describes any embodiment of the invention that has not been claimed in the claims (Figure 1). This involves two decisions:

(i) Does a given paragraph describe an embodiment of the invention?

(ii) Has the embodiment in a paragraph, if any, been claimed in the claims already?

We thus need both a paragraph and a set of claims to determine whether the paragraph contains an unclaimed embodiment.

Formally, the task is defined as follows (Figure 2). Suppose we have I patent specifications and the i-th patent specification has J_i description paragraphs. Let p_{i,j_i} be the j_i-th description paragraph in the i-th specification, C_i be a set of claims in the i-th specification, and y_{i,j_i} be a binary label where $y_{i,j_i} = 1$ if p_{i,j_i} describes any embodiment that is not claimed in C_i and $y_{i,j_i} = 0$ otherwise; here, $i = \{1, ..., I\}$ and $j_i = \{1, ..., J_i\}$. Given $N = \sum_i J_i$ training instances, our goal is to learn a function $f(p_{i,j_i}, C_i) \rightarrow y_{i,j_i}$.

The task involves the two decisions (i) and (ii) and a UEE model may make the two decisions separately. Our UEE baseline models in §4.1, nonetheless, make the two decisions in a single step, as it is more straightforward. We will explore different architectures for better utilization of the nature of the task (involving the two decisions) in the future.

In this study, we chose a paragraph as the unit of embodiment description, because in the patent applications, paragraphs are encouraged to be numbered to serve as the unit of work and indeed form a

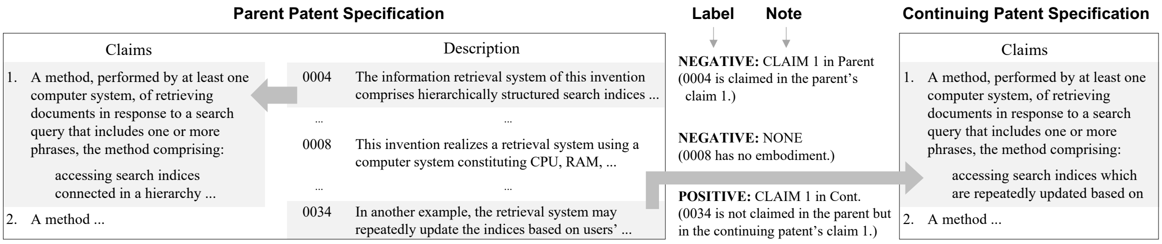

Parent Patent Specification
Label
Note
Continuing Patent Specification
Claims
Description
1. A method, performed by at least one computer system, of retrieving documents in response to a search query that includes one or more phrases, the method comprising:
accessing search indices connected in a hierarchy ...
2. A method ...
0004 The information retrieval system of this invention comprises hierarchically structured search indices ...
...
0008 This invention realizes a retrieval system using a computer system constituting CPU, RAM, ...
...
0034 In another example, the retrieval system may repeatedly update the indices based on users' ...
NEGATIVE: CLAIM 1 in Parent (0004 is claimed in the parent's claim 1.)
NEGATIVE: NONE (0008 has no embodiment.)
POSITIVE: CLAIM 1 in Cont. (0034 is not claimed in the parent but in the continuing patent's claim 1.)
Claims
1. A method, performed by at least one computer system, of retrieving documents in response to a search query that includes one or more phrases, the method comprising:
accessing search indices which are repeatedly updated based on
2. A method ...

Figure 3: Annotation method. Paragraph 0004 of the parent patent is labeled as NEGATIVE because it describes an embodiment claimed in the parent (as indicated by the left arrow). 0008 is NEGATIVE as it has no embodiment. 0034 describes an embodiment that is not claimed by the parent but claimed in the continuing patent from the parent (as the right arrow indicates); 0034 is thus POSITIVE. CLAIM1 and NONE next to the labels are notes given by human annotators to explain their annotation decisions. This example is given in English for illustration purposes.

meaningful unit. Other options would be a phrase, clause, or sentence. We will identify the best unit of embodiment description in the future.

3 Dataset

3.1 Data Source

We acquired source patent data from the Japan Patent Office (JPO) via their web form.[2] The data from JPO contained Japanese patent specifications from 1993 to 2022. We obtained both parent and continuing patent specifications from this data. We created the dataset from patent specifications that had their corresponding continuing patents.

3.2 Annotation Method

As the task involves two decisions, (i) and (ii) in §2, our annotation method is based on the two as illustrated in Figure 3. Specifically, we label a paragraph as negative if it has no embodiment (See paragraph 0008 in Figure 3), or if the embodiment is claimed in the parent patent to which the paragraph belongs (0004 in Figure 3).

For positive annotation, we used the continuing patent generated from the patent to which the target paragraph belonged. If a paragraph describes an embodiment that is claimed in the continuing patent but not in the parent, the paragraph is labeled as positive (0034 in Figure 3). Although we can identify unclaimed embodiments from the parent patent, without relying on the continuing patent, it helps us double-check positive paragraphs.

To use continuing patents, we collected patent specifications with corresponding continuing patents from the JPO data and made pairs of parent and continuing patents as in Figure 3.

We restricted target patents to those with International Patent Classification (IPC) codes[3] that met our business needs; specifically, we mainly chose IPC codes for digital data processing (G06F), information and communication technology (G06Q), and aeroplanes (B64C). IPC is used in over 100 countries to indicate the subject of the invention.[4]

We conducted manual annotation on pairs of parent and continuing patents collected in this way.

On top of positive and negative labels, we leave *notes* that clarify reasons for annotators' labeling decisions (e.g., CLAIM1 and NONE in Figure 3). This is because labeling decisions would be based on patent practitioners' expertise, which may be incomprehensible to researchers, and we expect the notes to improve annotated labels' interpretability. A negative paragraph is given the claim ID of the parent patent as the note if the paragraph's embodiment is claimed in the parent's claims. A negative paragraph is given the note NONE if it has no embodiment. A positive paragraph is given the claim ID of the continuing patent if its embodiment is claimed in the continuing patent's claims.

We use the notes for experiments of decision (i) (§4.2) and decision (ii) (§4.3), too.

3.3 Annotators

Two experienced patent practitioners who were native speakers of Japanese were employed as annotators. We split the 11,951 instances (each consisting of a description paragraph and a set of claims) into two separate sets. Each annotator was assigned to only one set; no instance was annotated by both of them due to our budget constraints.

We nonetheless measured inter-annotator agree-

[2] https://www.jpo.go.jp/system/laws/sesaku/data/download.html (Japanese)

[3] https://ipcpub.wipo.int/

[4] https://en.wikipedia.org/wiki/International_Patent_Classification

Total numbers	
Labeled instances	11,951
Parent patents	971
Continuing patents	1,022
Average numbers	
Tokens per desc. paragraph	88.73 (77.66)
Sentences per desc. paragraph	3.67 (1.96)
Tokens per desc. sentence	23.59 (21.61)
Claims per parent patent	11.37 (6.11)
Tokens per claim	106.44 (95.12)
Label distribution	
POSITIVE:CLAIM	4,564 (38.19%)
NEGATIVE:CLAIM	1,619 (13.55%)
NEGATIVE:NONE	5,768 (48.26%)

Table 1: Statistics of the UEE dataset. In "Average numbers" the figures in parenthesis are standard deviations. The "desc." stands for "description." The "CLAIM" and "NONE" are the notes described in §3.2.

Experiment	Model	F1
UEE Baselines	RoBERTa	**0.8670** (0.0034)
	Longformer	0.7247 (0.0335)
Decision (i)	RoBERTa	0.9259 (0.0057)
Decision (ii)	RoB$_{uee}$	**0.7218** (0.0110)
	RoB$_{jsnli}$ A	0.4029 (0.1434)
	RoB$_{jsnli}$ B	0.4384 (0.2472)

Table 2: F1 of all the models; (Top) the RoBERTa and Longformer baselines for the UEE task reported in §4.1; (Middle) RoBERTa model for decision (i) in §4.2; and (Bottom) the RoBERTa$_{uee}$, RoBERTa$_{jsnli}$ Condition A, and RoBERTa$_{jsnli}$ Condition B models in §4.3 ("RoBERTa" is abbreviated as "RoB"). We report the mean and standard deviation obtained by running each experiment five times. The standard deviation is written in parenthesis. See Appendix D for accuracy, precision and recall for each method.

ment by asking another experienced patent practitioner (a native speaker of Japanese) to annotate 309 instances that the above two annotators worked on after removing their labels and notes. As a result, Cohen's kappa (Cohen, 1960) was 0.465, indicating moderate agreement (Landis and Koch, 1977). According to the kappa score, experts may disagree occasionally. The question is then how much experts' disagreement affects the performance, as pointed out by an anonymous reviewer. We will explore this question in the future.

3.4 The Dataset

The resulting UEE dataset has 11,951 instances. We use 60% for training, 20% for development, and 20% for testing. Table 1 shows the statistics of the dataset. We used the tokenizer of our RoBERTa in §4.1 to count tokens. In "Label distribution" in Table 1, "**POSITIVE**:CLAIM"[5] refers to a positive instance. The instance with paragraph 0034 in Figure 3 has this label. "**NEGATIVE**:CLAIM" means a negative instance with an embodiment that has already been claimed in the parent patent and hence with a note of the corresponding claim ID. "**NEGATIVE**:NONE" is a negative instance without embodiment. The instances with paragraphs 0004 and 0008 in Figure 3 are examples of these two types of negatives, respectively.

In Appendix A, we show an example data instance of the UEE dataset in Japanese and English for illustration purposes.

4 Experiments

We conducted experiments of UEE with baseline models based on Transformer (Vaswani et al., 2017) to see the difficulty of the task (§4.1). We also conducted experiments of making the decisions (i) and (ii) in §2 as independent tasks for a better understanding of the task (§4.2 and §4.3).

These experimental results show the following:

(A) Baseline Transformer-based models deliver mediocre performances (§4.1).

(B) Despite patent specifications' being long, UEE models do not necessarily have to deal with long documents (§4.1).

(C) The bottleneck in UEE is decision (ii) (§4.3).

4.1 UEE Baselines

We evaluated RoBERTa (Liu et al., 2019) and Longformer (Beltagy et al., 2020) for the UEE task, because these are ones of standard Transformer-based models, and Longformer is known to be able to deal with long documents such as patent specifications. However, we do not claim these are optimal models for the task; we will explore better models in the future. Our RoBERTa was built from a base-sized one which we call Rinna RoBERTa[6] and had been pre-trained on Japanese CC-100 (Conneau et al.,

[5] We may omit claim IDs of the notes, e.g. "1", hereafter.

[6] https://huggingface.co/rinna/japanese-roberta-base

2020) and Japanese Wikipedia. The maximum sequence length was 512. We fine-tuned Rinna RoBERTa on the UEE training set for ten epochs with the training batch size, the warm-up steps, and the learning rate being set to 128, 100, and 5e-5, respectively. We used AdamW (Loshchilov and Hutter, 2019) for optimization. We will describe the hyper-parameter settings of models we experimented with in Appendix B, hereafter.

Our Longformer was converted from Rinna RoBERTa following Beltagy et al. (2020).[7] See Appendix B.2 for its hyper-parameter setting.

The first two rows in Tab 2 labeled with §4.1 show F1 of the two baseline models on the UEE test set. We fine-tuned and evaluated each model five times. The reported figures are the mean and standard deviation obtained from the five runs.

The result indicates that our baselines have non-negligible room for improvement. Given Transformers' successes in many tasks (e.g., a base-sized RoBERTa fine-tuned and evaluated on JSNLI (Yoshikoshi et al., 2020), a Japanese NLI dataset, delivers the F1 of 0.93 (Yanaka and Mineshima, 2022)), we think that UEE is challenging.

The result also indicates that RoBERTa outperforms Longformer. Actually, we expected the opposite result, because the input to UEE models, i.e. a pair of a description paragraph and a set of claims, tends to be long; the average number of tokens in a description paragraph is 88.73 and that of tokens in a set of claims is 1210.22 (= 106.44 × 11.37), as Table 1 shows.

We suspect that this unexpected result is due to the fact that, in the UEE dataset, more than 70% of embodiments in description paragraphs with **NEGATIVE**:CLAIM are claimed in the first three claims. Models then do not always have to read through all the claims. This is probably because of the preferred order of claims: Claims should preferably be arranged in order of scope so that the first claim presented is the least restrictive (USPTO, 2020a); i.e. the most general claims should come first.

4.2 Decision (i)

We conducted experiments of making only the decisions (i), i.e. whether a paragraph described any embodiment, to see how difficult it was.

To train and test a model for decision (i), we created training, development, and test sets for decision (i) from the corresponding set of the UEE dataset as follows.[8] We regarded the instances in the UEE dataset whose note is NONE as negative and the rest as positive, because note NONE indicates the corresponding paragraph has no embodiment as described in §3.2. The positive-negative ratio was then about 52:48 as Tab 1 indicates.

We built a model from Rinna RoBERTa (§4.1) again for this experiment. The experimental protocol was the same as our RoBERTa in §4.1. See Appendix B.3 for its hyper-parameter setting.

The third row in Tab 2 labeled with §4.2 shows F1 of the model on the test set. The reported figures are the mean and standard deviation obtained from five runs of fine-tuning and evaluation. This result indicates that decision (i) is a modest task.

4.3 Decision (ii)

We also conducted experiments to make the decision (ii), i.e. whether the embodiment of a paragraph has been claimed, as an independent task.

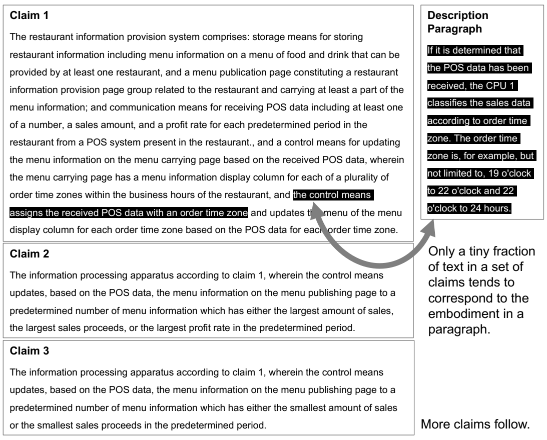
As discussed in §1, decision (ii) can be seen as an NLI task where the hypothesis is a paragraph and the premise is a set of claims. For training and test of models for decision (ii), in order to focus on its NLI aspect, we ignored UEE dataset instances with **NEGATIVE**:NONE. This is because it is obvious for a paragraph without any embodiment to be unclaimed, i.e. not entailed by a set of claims. Besides, we used not only parent patents but also their continuing patents in the UEE dataset, as it is straightforward to use them for decision (ii).

Accordingly, we created training, development, and test sets for decision (ii) from the corresponding set of the UEE dataset as follows. We generated positive instances for decision (ii) from the UEE dataset by pairing a paragraph of **POSITIVE**:CLAIM and a set of claims in the *continuing patent*; e.g. the pair of paragraph 0034 and the continuing patent's claims in Figure 3. We also generated decision (ii) positives by pairing a paragraph of **NEGATIVE**:CLAIM and a set of claims in the *parent patent*; e.g. paragraph 0004 and the parent patent's claims in Figure 3.

Likewise, decision (ii) negatives were generated by pairing a paragraph of **POSITIVE**:CLAIM and a set of claims in the *parent patent* and also by pairing a paragraph of **NEGATIVE**:CLAIM and a set of claims in the *continuing patent*. In Figure 3, pair of

[7] See `convert_model_to_long.py` in the supplementary material for implementation.

[8] For dataset creation for (i) and (ii), see: `Decision1/src/dataset.py` and `Decision2/src/dataset.py` in `github.com/rakutentech/UEE_ACL23`.



Claim 1

The restaurant information provision system comprises: storage means for storing restaurant information including menu information on a menu of food and drink that can be provided by at least one restaurant, and a menu publication page constituting a restaurant information provision page group related to the restaurant and carrying at least a part of the menu information; and communication means for receiving POS data including at least one of a number, a sales amount, and a profit rate for each predetermined period in the restaurant from a POS system present in the restaurant., and a control means for updating the menu information on the menu carrying page based on the received POS data, wherein the menu carrying page has a menu information display column for each of a plurality of order time zones within the business hours of the restaurant, and the control means assigns the received POS data with an order time zone and updates the menu of the menu display column for each order time zone based on the POS data for each order time zone.

Claim 2

The information processing apparatus according to claim 1, wherein the control means updates, based on the POS data, the menu information on the menu publishing page to a predetermined number of menu information which has either the largest amount of sales, the largest sales proceeds, or the largest profit rate in the predetermined period.

Claim 3

The information processing apparatus according to claim 1, wherein the control means updates, based on the POS data, the menu information on the menu publishing page to a predetermined number of menu information which has either the smallest amount of sales or the smallest sales proceeds in the predetermined period.

Description Paragraph

If it is determined that the POS data has been received, the CPU 1 classifies the sales data according to order time zone. The order time zone is, for example, but not limited to, 19 o'clock to 22 o'clock and 22 o'clock to 24 hours.

Only a tiny fraction of text in a set of claims tends to correspond to the embodiment in a paragraph.

More claims follow.

Figure 4: Example of a claimed embodiment, which is translated to English for illustration purposes. See Appendix C for the original Japanese example.

paragraph 0034 and the parent patent's claims and that of paragraph 0004 and the continuing patent's claims are decision (ii) negatives.

The positive-negative ratio was then 50:50.

We fine-tuned Rinna RoBERTa with this training set. We call the resulting model RoBERTa$_{uee}$. The experimental protocol was the same as the previous experiments (§4.1 and §4.2).

Since decision (ii) is an NLI task, we also fine-tuned Rinna RoBERTa using the JSNLI dataset[9] for comparison. We call it RoBERTa$_{jsnli}$. We used the same test set as RoBERTa$_{uee}$ for evaluation. RoBERTa$_{jsnli}$'s high performance for Japanese NLI tasks has been shown in the literature (Yanaka and Mineshima, 2022).

Note that while JSNLI is a ternary classification task, i.e. *entailment*, *contradiction*, and *neutral*, decision (ii) is binary, i.e. *positive* and *negative*. We, therefore, need to align JSNLI's labels with our binary labels. We experimented with two label alignment conditions: Condition A was to align *entailment* with *positive* and *contradiction* and *neutral* with *negative*. Condition B was the same as Condition A, except that we ignored *contradiction*; only *neutral* was aligned with *negative*. This is reasonable because, even if an embodiment is not claimed in a set of claims, it does not necessarily imply that the two pieces of text are contradictory.

For the hyper-parameter setting and fine-tuning of RoBERTa$_{jsnli}$, refer to Appendix B.5.

The last three rows in Tab 2 labeled with §4.3 show F1 of the models on the test set. Although RoBERTa$_{uee}$ was the best among them, it has a large room for improvement. This indicates that decision (ii) is difficult and is the bottleneck for UEE. Looking closely at the data revealed that a tiny fraction of text in a set of claims, which usually is a long document, tended to correspond to an embodiment (Figure 4), because each claim may consist of various technical features from more than one description paragraph. This would make decision (ii) challenging, together with the other factors discussed in §1; i.e. patent specifications consisting of technical, long, syntactically and semantically involved sentences written in *patentese*.

RoBERTa$_{jsnli}$ delivered low performances under both conditions, probably because of the domain discrepancy between the NLI task in JSNLI and UEE. We think this result shows the necessity of a dedicated dataset for UEE.

4.4 Discussion

The baselines delivered mediocre performances for UEE. We observed that decision (ii) makes UEE difficult. Nevertheless, we believe UEE is a worthy challenge, as it would eventually contribute to the industry by streamlining patent application processes. We also believe that, to this end, utilizing the outcomes of the current study would help.

5 Related Work

Patent Document Processing NLP systems for real-world applications in, for instance, e-commerce (Malmasi et al., 2021, 2022), medical (Rumshisky et al., 2020; Naumann et al., 2022), and legal areas (Aletras et al., 2021; Preotiuc-Pietro et al., 2022) has gained attention, probably because it has become more feasible to serve practical needs thanks to the success of Transformer-based models and pre-training methods (Devlin et al., 2019).

Patent document processing has also been studied extensively (Aras et al., 2019; Krestel et al., 2021, 2022). Its most studied areas are machine translation (Tsujii and Yokoyama, 2007; Utsuro et al., 2019; Nakazawa et al., 2021, 2022) and information retrieval (Tait et al., 2008, 2009, 2010; Risch et al., 2020).

There have recently been studies that would directly facilitate patent applications. Sharma et al. (2019) created a dataset for summarizing patents and proposed baselines. Tonguz et al. (2021) proposed a method for claim generation formulated as text summarization. Aslanyan and Wetherbee (2022) created a dataset for phrasal matching for

[9] We used `train_w_filtering.tsv` of JSNLI 1.1.

better patent similarity measurement. Gao et al. (2022) proposed a method for predicting whether a given patent application would be approved.

However, to the best of our knowledge, no study has addressed UEE; we are the first to do that.

Natural Language Inference NLI has showcased the comprehension ability of NLP systems (Wang et al., 2019; Nie et al., 2020; Poliak, 2020) and provided datasets for their training (Conneau et al., 2017; Reimers and Gurevych, 2019). Introducing diverse NLI tasks would then push the boundary of NLP. Recent studies have introduced new NLI tasks targeting real-world applications (Romanov and Shivade, 2018; Holzenberger et al., 2020; Koreeda and Manning, 2021; Sadat and Caragea, 2022). Our decision (ii) is a novel real-world NLI for patents that poses a new challenge.

6 Conclusion

We introduced UEE, a novel NLP challenge, and created a corresponding dataset. Our experiments showed that UEE was challenging due to the difficulty of making the decision (ii). We hope that the research community will address this challenge by utilizing the UEE dataset and code that we created and released.

Future Work We have not explored better architectures for the task extensively. Although RoBERTa performed reasonably well, more capable, human-instruction-aligned architectures have been developed recently (Bahrini et al., 2023; OpenAI, 2023). We will explore the capability of these more recent large language models for the task.

7 Ethics Statement

The scope of this work is to introduce NLP technologies to the continuing patent application process to make it more efficient. The outcomes from our work would therefore have an industrial impact through enabling organizations to file more continuing patents with less time. There would then be a risk that, if our technologies were available to only particular organizations, fair competitions could not be ensured. We therefore decided to release the dataset and code to the public.

This work was intended to be beneficial to patent-related processes and studies in artificial intelligence, machine learning, and NLP. The outcomes from this work should therefore be used only for these purposes.

The coverage of the UEE dataset in terms of the IPC subclass, language, and countries and regions are limited. Care must be taken when using this dataset, accordingly.

The dataset does not contain any personal information, as it has been created from publicly available patent specifications. We nonetheless took special care to check if any personal information was included in the dataset by accident when creating the dataset.

All the data we used in this work are publicly available. The pre-trained language model that we used, i.e. RoBERTa, is also publicly available. Our Longformer was converted from the RoBERTa with a method that was also known to the public. Besides, since we have released all the necessary code and dataset along with the paper, all the experimental results in the paper are reproducible.

Regarding the hiring of the annotators (the expert patent practitioners), we negotiated with their company in advance to fairly determine the charge, which was the equivalent of the cost of hiring expert patent practitioners for patent search. We explained to the human annotators about the purpose of the data annotation and how it would be used in advance of the annotation.

Regarding the compute in our experiments, we executed 30 fine-tuning processes, which took 57 hours in a single Nvidia A100 GPU in total.

8 Acknowledgment

We are deeply grateful to all those who played a role in the success of this project. We would like to thank Yoshimasa Utsumi, Kazutaka Ohara, Takeshi Imamura, Hsuan-Yu Kuo, Takako Yoshida, Vijay Daultani, Colin Ian Rowat, and Aztec Co., Ltd. for their invaluable input and support throughout the research process. Their insights and expertise were instrumental in shaping the direction of this project.

References

Nikolaos Aletras, Ion Androutsopoulos, Leslie Barrett, Catalina Goanta, and Daniel Preotiuc-Pietro, editors. 2021. *Proceedings of the Natural Legal Language Processing Workshop 2021*. Association for Computational Linguistics, Punta Cana, Dominican Republic.

Hidir Aras, Linda Andersson, Florina Piroi, Jianan Hou, Juan Carlos Gomez, Tobias Fink, Allan Hanbury, Benjamin Meindl, Ingrid Ott, Ulrich Theodor Zierahn, Tatyana Skripnikova, Anna Weihaar, Sebastian Blank, Hans-Peter Zorn, Farag Saad, Stefan

Helfrich, and Mustafa Sofean. 2019. 1st Workshop on Patent Text Mining and Semantic Technologies (PatentSemTech 2019).

Grigor Aslanyan and Ian Wetherbee. 2022. Patents Phrase to Phrase Semantic Matching Dataset. In *3rd Workshop on Patent Text Mining and Semantic Technologies*, PatentSemTech.

Aram Bahrini, Mohammadsadra Khamoshifar, Hossein Abbasimehr, Robert J. Riggs, Maryam Esmaeili, Rastin Mastali Majdabadkohne, and Morteza Pasehvar. 2023. ChatGPT: Applications, Opportunities, and Threats.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer. *arXiv:2004.05150*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.

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.

Gabriela Ferraro, Hanna Suominen, and Jaume Nualart. 2014. Segmentation of patent claims for improving their readability. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, pages 66–73, Gothenburg, Sweden. Association for Computational Linguistics.

Xiaochen Gao, Zhaoyi Hou, Yifei Ning, Kewen Zhao, Beilei He, Jingbo Shang, and Vish Krishnan. 2022. Towards comprehensive patent approval predictions:beyond traditional document classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 349–372, Dublin, Ireland. Association for Computational Linguistics.

Nils Holzenberger, Andrew Blair-Stanek, and Benjamin Van Durme. 2020. A Dataset for Statutory Reasoning in Tax Law Entailment and Question Answering. In *Proceedings of the 2020 Natural Legal Language Processing (NLLP) Workshop*.

Yuta Koreeda and Christopher Manning. 2021. ContractNLI: A dataset for document-level natural language inference for contracts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1907–1919, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ralf Krestel, Hidir Aras, Linda Andersson, Florina Piroi, Allan Hanbury, and Dean Alderucci. 2021. 2nd Workshop on Patent Text Mining and Semantic Technologies (PatentSemTech2021). In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 2693–2696.

Ralf Krestel, Hidir Aras, Allan Hanbury, Linda Andersson, Florina Piroi, and Dean Alderucci. 2022. 3rd Workshop on Patent Text Mining and Semantic Technologies (PatentSemTech2022). In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*, ICLR.

Shervin Malmasi, Surya Kallumadi, Nicola Ueffing, Oleg Rokhlenko, Eugene Agichtein, and Ido Guy, editors. 2021. *Proceedings of the 4th Workshop on e-Commerce and NLP*. Association for Computational Linguistics, Online.

Shervin Malmasi, Oleg Rokhlenko, Nicola Ueffing, Ido Guy, Eugene Agichtein, and Surya Kallumadi, editors. 2022. *Proceedings of the Fifth Workshop on e-Commerce and NLP (ECNLP 5)*. Association for Computational Linguistics, Dublin, Ireland.

Toshiaki Nakazawa, Hideya Mino, Isao Goto, Raj Dabre, Shohei Higashiyama, Shantipriya Parida, Anoop Kunchukuttan, Makoto Morishita, Ondřej Bojar, Chenhui Chu, Akiko Eriguchi, Kaori Abe, Yusuke Oda, and Sadao Kurohashi. 2022. Overview of the 9th workshop on Asian translation. In *Proceedings of the 9th Workshop on Asian Translation*, pages 1–36, Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Toshiaki Nakazawa, Hideki Nakayama, Isao Goto, Hideya Mino, Chenchen Ding, Raj Dabre, Anoop Kunchukuttan, Shohei Higashiyama, Hiroshi Manabe, Win Pa Pa, Shantipriya Parida, Ondřej Bojar, Chenhui Chu, Akiko Eriguchi, Kaori Abe, Yusuke Oda, Katsuhito Sudoh, Sadao Kurohashi, and Pushpak Bhattacharyya, editors. 2021. *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*. Association for Computational Linguistics, Online.

Tristan Naumann, Steven Bethard, Kirk Roberts, and Anna Rumshisky, editors. 2022. *Proceedings of the 4th Clinical Natural Language Processing Workshop*. Association for Computational Linguistics, Seattle, WA.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 Technical Report.

Adam Poliak. 2020. A survey on recognizing textual entailment as an NLP evaluation. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 92–109, Online. Association for Computational Linguistics.

Daniel Preotiuc-Pietro, Ilias Chalkidis, Catalina Goanta, Leslie Barrett, and Nikolaos Aletras, editors. 2022. *Proceedings of the Natural Legal Language Processing Workshop 2022*. Association for Computational Linguistics, Abu Dhabi, UAE.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Julian Risch, Nicolas Alder, Christoph Hewel, and Ralf Krestel. 2020. PatentMatch: A Dataset for Matching Patent Claims & Prior Art. *ArXiv*, abs/2012.13919.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1586–1596, Brussels, Belgium. Association for Computational Linguistics.

Anna Rumshisky, Kirk Roberts, Steven Bethard, and Tristan Naumann, editors. 2020. *Proceedings of the 3rd Clinical Natural Language Processing Workshop*. Association for Computational Linguistics, Online.

Mobashir Sadat and Cornelia Caragea. 2022. SciNLI: A corpus for natural language inference on scientific text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7399–7409, Dublin, Ireland. Association for Computational Linguistics.

Eva Sharma, Chen Li, and Lu Wang. 2019. BIGPATENT: A large-scale dataset for abstractive and coherent summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2204–2213, Florence, Italy. Association for Computational Linguistics.

John Tait, Helmut Berger, Michael Dittenbach, and Mihai Lupu, editors. 2009. *Proceedings of the 2nd International Workshop on Patent Information Retrieval, PaIR '09, Hong Kong, SAR, China, November 6, 2009*. Association for Computing Machinery.

John Tait, Helmut Berger, Michael Dittenbach, and Georg Sommer, editors. 2008. *PaIR '08: Proceedings of the 1st ACM Workshop on Patent Information Retrieval*. Association for Computing Machinery.

John Tait, Christopher G. Harris, and Mihai Lupu, editors. 2010. *Proceedings of the 3rd international workshop on Patent information retrieval*. Association for Computing Machinery.

Ozan Tonguz, Yiwei Qin, Yimeng Gu, and Hyun Hannah Moon. 2021. Automating claim construction in patent applications: The CMUmine dataset. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 205–209, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jun'ichi Tsujii and Shoichi Yokoyama, editors. 2007. *Proceedings of the Workshop on Patent translation*. Copenhagen, Denmark.

The United States Patent & Trademark Office USPTO. 2020a. Parts, Form, and Content of Application. In *Manual of Patent Examining Procedure*, chapter 600.

The United States Patent & Trademark Office USPTO. 2020b. PCT Rule 13 Unity of Invention.

Takehito Utsuro, Katsuhito Sudoh, and Takashi Tsunakawa, editors. 2019. *Proceedings of the 8th Workshop on Patent and Scientific Literature Translation*. European Association for Machine Translation, Dublin, Ireland.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. In *Advances in Neural Information Processing Systems*, NeurIPS, pages 5998–6008.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*.

World Intellectual Property Organization WIPO. 2022a. WIPO Patent Drafting Manual, Second Edition.

World Intellectual Property Organization WIPO. 2022b. Worldwide IP Filings Reached New All-Time Highs in 2021, Asia Drives Growth.

Hitomi Yanaka and Koji Mineshima. 2022. Compositional Evaluation on Japanese Textual Entailment and Similarity. *Transactions of the Association for Computational Linguistics*, 10:1266–1284.

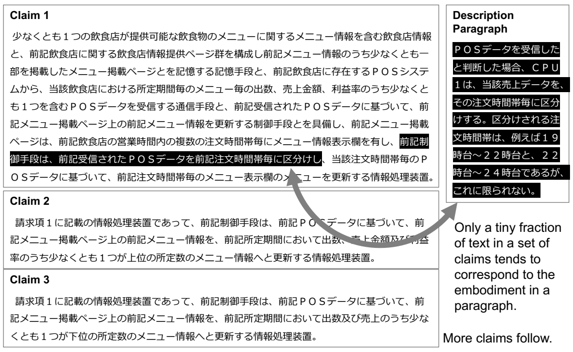
Takumi Yoshikoshi, Daisuke Kawahara, and Sadao Kurohashi. 2020. Multilingualization of natural language inference datasets using machine translation (in Japanese). In *The 244th Meeting of Natural Language Processing*.

```
{
  "appNum": "2021xxxxxx",
  "paraNum": "0052",
  "paraTxt": "The first data may be data available for
    learning of the first model M1, and is not limited ..."
  "claims": [
    {
      "claimNum": "1",
      "claimTxt": [
        "A processing execution system including:",
        "a second classification information acquisition
         unit for acquiring second classification ..."
      ]
    },
    {
      "claimNum": "2",
      "claimTxt": [
        "The processing execution system of claim 1,",
        "wherein the estimator estimates validity ..."
      ]
    }
  ],
  "label": "positive",
  "note": [ "1" ],
  "contAppNum": "2021yyyyyy",
  "contClaims": [
    {
      "claimNum": "1",
      "claimTxt": [
        "A processing execution system including...",
      ]
    },
}
```

Figure 5: Example data. Paragraph 0052 from the patent specification for application 2021xxxxxx contains an unclaimed embodiment (labeled as positive). This is translated to English for illustration purposes.

```
{
  "appNum": "2021xxxxxx",
  "paraNum": "0052",
  "paraTxt": "第１データは、第１モデルＭ１の学習に利用可能なデータ
    であればよく、ウェブページのタイトルに限られない。例えば、第１デ
    ータはウェブページから作成された要約であってもよい。"
  "claims": [
    {
      "claimNum": "1",
      "claimTxt": [
        "第１データと、当該第１データの分類に関する第１分類情報と、
        の関係が学習された第１モデルに基づいて、第２データの分類に
        関する第２分類情報を取得する第２分類情報取得部と、",
        "前記有効性の推定結果に基づいて処理を実行する実行部と、",
        "を含む処理実行システム。"
      ]
    },
    {
      "claimNum": "2",
      "claimTxt": [
        "前記推定部は、前記第２モデルに基づいて有効性を推定する、",
        "請求項１に記載の処理実行システム。"
      ]
    }
  ],
  "label": "positive",
  "note": [ "1" ],
  "contAppNum": "2021yyyyyy",
  "contClaims": [
    {
      "claimNum": "1",
      "claimTxt": [
        "第１データと、当該第１データの...",
      ]
    },
}
```

Figure 6: Example data in Japanese



Claim 1

少なくとも１つの飲食店が提供可能な飲食物のメニューに関するメニュー情報を含む飲食店情報と、前記飲食店に関する飲食店情報提供ページ群を構成し前記メニュー情報のうち少なくとも一部を掲載したメニュー掲載ページとを記憶する記憶手段と、前記飲食店に存在するＰＯＳシステムから、当該飲食店における所定期間毎のメニュー毎の出数、売上金額、利益率のうち少なくとも１つを含むＰＯＳデータを受信する通信手段と、前記受信されたＰＯＳデータに基づいて、前記メニュー掲載ページ上の前記メニュー情報を更新する制御手段とを具備し、前記メニュー掲載ページは、前記飲食店の営業時間内の複数の注文時間帯毎にメニュー情報表示欄を有し、前記制御手段は、前記受信されたＰＯＳデータを前記注文時間帯毎に区分けし、当該注文時間帯毎のＰＯＳデータに基づいて、前記注文時間帯毎のメニュー表示欄のメニューを更新する情報処理装置。

Claim 2

請求項１に記載の情報処理装置であって、前記制御手段は、前記ＰＯＳデータに基づいて、前記メニュー掲載ページ上の前記メニュー情報を、前記所定期間において出数、売上金額及び利益率のうち少なくとも１つが上位の所定数のメニュー情報へと更新する情報処理装置。

Claim 3

請求項１に記載の情報処理装置であって、前記制御手段は、前記ＰＯＳデータに基づいて、前記メニュー掲載ページ上の前記メニュー情報を、前記所定期間において出数及び売上のうち少なくとも１つが下位の所定数のメニュー情報へと更新する情報処理装置。

Description Paragraph

ＰＯＳデータを受信したと判断した場合、ＣＰＵ１は、当該売上データを、その注文時間帯毎に区分けする。区分けされる注文時間帯は、例えば１９時台～２２時台と、２２時台～２４時台であるが、これに限られない。

Only a tiny fraction of text in a set of claims tends to correspond to the embodiment in a paragraph.

More claims follow.

Figure 7: Claimed embodiment example in Japanese

A Example of the UEE Dataset

Figure 5 illustrates an entry of the dataset. Each entry is a JSON object and consists of the application number of the patent from which the target paragraph is extracted (appNum), the identifier of the target paragraph (paraNum), the paragraph text (paraTxt), a set of claims from the same application (claims), the label (label), the note (note), the application number of the continuing patent (contAppNum), and the continuing patent's claims (contClaims). claims and contClaims consist of individual claims' identifier (claimNum) and text (claimTxt).[10] Here, paraTxt, a set of claimTxts, and label correspond to p_{i,j_i}, C_i, and y_{i,j_i} in §2, respectively. Figure 6 shows the Japanese version of Figure 5.

B Hyper-parameter Setting

B.1 UEE Baseline Rinna RoBERTa

Although we described Rinna RoBERTa's hyper-parameter setting for the baseline experiment in §4.1, we repeat it here for the sake of completeness.

The maximum sequence length was 512. We fine-tuned Rinna RoBERTa on the UEE training set for ten epochs with the training batch size, the warm-up steps, and the learning rate being set to 128, 100, and 5e-5, respectively. We used AdamW (Loshchilov and Hutter, 2019) for optimization.

B.2 UEE Baseline Longformer

The maximum sequence length was 4,096. We fine-tuned our Longformer on the UEE training set for ten epochs with the training batch size, the gradient accumulation steps, the warm-up steps, and the learning rate being set to 16, 2, 200, and 2e-5, respectively, based on Beltagy et al. (2020). We used the AdamW optimizer.

B.3 Decision (i) RoBERTa

The hyper-parameter setting for this model is the same as the UEE Baseline Rinna RoBERTa in B.1.

[10] claimTxt is a list of text. This is because a claim is usually long and split into segments for readability. We keep this structure in JSON format.

Sec.	Model	Accuracy	Precision	Recall	F1
§4.1	RoBERTa	**0.8979** (0.0025)	0.8453 (0.0047)	**0.8899** (0.0065)	**0.8670** (0.0034)
	Longformer	0.8258 (0.0147)	**0.8839** (0.0072)	0.6157 (0.0508)	0.7247 (0.0335)
§4.2	RoBERTa	0.9261 (0.0050)	0.9539 (0.0112)	0.8998 (0.0178)	0.9259 (0.0057)
§4.3	RoBERTa$_{uee}$	**0.7238** (0.0047)	**0.7274** (0.0152)	**0.7175** (0.0339)	**0.7218** (0.0110)
	RoBERTa$_{jsnli}$ A	0.5067 (0.0044)	0.5146 (0.0123)	0.3874 (0.2847)	0.4029 (0.1434)
	RoBERTa$_{jsnli}$ B	0.5098 (0.0072)	0.4088 (0.2286)	0.4776 (0.2779)	0.4384 (0.2472)

Table 3: Performances of all the evaluated models. In the last two rows, "A" and "B" stand for Condition A and B, respectively. The figures are the mean and standard deviation from five runs.

B.4 Decision (ii) RoBERTa$_{uee}$

The hyper-parameter setting for this model is the same as the UEE Baseline Rinna RoBERTa in B.1.

B.5 Decision (ii) RoBERTa$_{jsnli}$

We fine-tuned RoBERTa$_{jsnli}$ for ten epochs with the training batch size, the warm-up steps, and the learning rate being 128, 500, and 3e-5, respectively. We used the AdamW optimizer. Although the instances in the JSNLI dataset have already been tokenized, we re-tokenized them with Rinna RoBERTa's tokenizer.

C Japanese Example of a Claimed Embodiment

Figure 7 shows the Japanese version of Figure 4.

D Full Evaluation Results

Table 3 shows accuracy, precision, recall, and F1 of all the evaluated models.