Improving Patent Classification using AI-Generated Summaries

Naoya Yoshikawa[1,2], Ralf Krestel[1]

[1]ZBW – Leibniz Information Centre for Economics & Kiel University, Düsternbrooker Weg 120, 24105, Kiel, Germany
[2]Japan Patent Office, 3-4-3 Kasumigaseki, Chiyoda-ku, 100-8915, Tokyo, Japan

Abstract
This study investigates the effectiveness of using summaries generated by large language models (AI-generated summaries) to improve the performance of automatic patent classification. We propose a novel approach to use AI-generated summaries of patent text fields (abstract, claims and detailed description) as training data for classification models: using two patent datasets, USPTO-70k and CLEF-IP 2011, we perform experiments focused on both subclass-level multi-label classification and subgroup-level multi-class classification tasks. The results show that models trained on AI-generated summaries of claims and detailed descriptions achieve significantly higher scores than models trained on the original text. This result suggests that AI-generated summaries effectively extract information relevant to patent classification and contributes to the development of automatic patent classification technology.

Keywords
Patent Classification, Large Language Model, Patent Summarization, Subgroup Classification

1. Introduction

Patent classifications play an important role in the accurate and efficient management of patent information. In addition to the International Patent Classification (IPC) maintained by the World Intellectual Property Organization (WIPO), patent offices around the world maintain their own patent classification systems, including the Cooperative Patent Classification (CPC) of the European Patent Office (EPO) and the United States Patent and Trademark Office (USPTO), and the File Index (FI) and File Forming Term (F-term) of the Japan Patent Office (JPO). The IPC is an internationally uniform classification with a hierarchical structure of sections (e.g. "A"), classes (e.g. "A61"), subclasses (e.g. "A61B"), main groups (e.g. "A61B17") and subgroups (e.g. "A61B17/29"). Patent professionals assign IPC codes to patent applications at the subgroup-level according to IPC guidelines [1]. This assignment process requires expertise and is complex and costly. Recent advances in natural language processing and deep learning have shown excellent results in various classification tasks. However, the diversity of patent classifications, evolving technology levels, inconsistent field-specific terms in patent documents, and the multi-label nature in which multiple labels can be assigned to a single application still make it challenging to automate patent classification accurately.

There are a number of approaches to the patent classification task, and supervised machine learning models using large amounts of classified patent text data published by WIPO and national patent offices have been widely adopted. One of the elements that can improve the performance of these models is the availability of high-quality and sufficient training data. Patent documents have a variety of text fields, such as title, abstract, claims, and detailed description, which contain rich textual information, but also much information that is not relevant to the classification task. Since it is difficult to manually extract high-quality textual information from a large number of patent documents, training data is usually created by mechanically extracting textual information from the text fields of patent documents. In previous studies, textual information such as title and abstract, claims, and the first few hundred words of the detailed description have been used [2, 3, 4, 5].

Moreover, mamy previous studies have focused on classification tasks at subclass-level, rather than at the subgroup-level for the IPC and CPC. Because subgroups are subdivided into approximately 70,000 labels for IPC and 240,000 labels for CPC, and patent classification is a multi-label classification task where multiple labels can be assigned, accurate automatic classification at the subgroup-level is very difficult. Subgroups are characterized by specific parts of the subject matter covered by the upper main group or subclass [1]. As a result, the title and abstract of a patent document, for example, may not contain sufficient information about the subgroups.

Since the introduction of the transformer architecture, one of the main approaches has been fine-tuning pretrained models using task-specific training data. Many previous studies have shown good performance using transformer-based models. However, these studies used

5th Workshop on Patent Text Mining and Semantic Technologies (PatentSemTech) 2024
yoshikawa-naoya@jpo.go.jp (N. Yoshikawa);
rkr@informatik.uni-kiel.de (R. Krestel)
0000-0002-5036-8589 (R. Krestel)

training data that was simply extracted, such as the first few hundred words of each text field, and the quality of the training data is debatable. Ideally, the input text should be no more than a few hundred words, containing important information relevant to the assigned classification.

Large Language Models (LLMs), which are pre-trained on large amounts of text data, have shown high performance in a variety of natural language processing tasks. Using LLMs to summarize text has the potential to extract information that is highly relevant to classification tasks.

The objective of this study is to compare the performance of models trained on summaries of patent documents generated by LLMs with the performance of models trained on the original patent text. We also evaluate the performance of subclass-level and subgroup-level classification tasks; text summarization with LLMs can be a promising approach to improve the quality of training data for patent classification tasks. By generating concise summaries from lengthy text of patent documents that contain important information relevant to classification, the training efficiency and performance of the model can be improved. Furthermore, the summarized text would be of a suitable length to serve as input for the transformer-based model, thus capturing the meaning of the entire text while reducing computational cost.

2. Related Work

Our goal is to classify patents based on automatically generated summaries (AI-generated summaries). Therefore, we explore in the following related work on patent classification on the one hand, and patent summarization on the other hand.

2.1. Patent Classification

Recent research results that are highly relevant to this study are presented. In recent years, deep learning methods have attracted a great deal of attention in patent classification tasks. Li et al. [4] proposed a deep learning algorithm DeepPatent, which combines word embedding and convolutional neural networks, and achieved 73.88% precision at the IPC subclass-level using the USPTO-2M dataset and 83.98% using the CLEF-IP 2011 dataset. They compared various combinations of text fields as training data and concluded that using the first 100 words of the title and abstract was optimal. Lee and Hsiang [3] fine-tuned a pre-trained BERT model for patent classification and reported that they achieved better performance than DeepPatent. They also showed that using only claims as training data instead of title and abstract produced comparable performance. Roudsari et al. [2] fine-tuned pre-trained language models to investigate multi-label patent classification performance for patent text. The first 128 words of each text field were compared as input text, and the combination of title and abstract showed good performance. They also noted that longer sentences should be considered when using detailed patent descriptions and claims. Pujari et al. [6] addressed the issue of limited input text size for neural models based on pre-trained transformers. They proposed a new approach to effectively integrate information obtained from multiple text fields. They published the USPTO-70k dataset extended to include claims, detailed description, brief-summary, and figure description, and in particular, they found that the brief-summary text field in the US patent document is the most useful for CPC subclass-level classification. Yadrintsev et al. [7] compared KNN (k-nearest neighbors) and fastText as IPC subclass-level classifiers using CLEF-IP competition data. In the CLEF English test sample (1000 documents), the micro-averaged F1-score were 71.0 for KNN and 70.4 for fastText.

All of the above studies were conducted at the IPC or CPC subclass-level; there are not many reports of studies on automatic classification at the IPC subgroup-level. Hoshino et al. [8] proposed a new decoder architecture that takes into account the hierarchical structure of IPC and a model that considers the content of all claims by extracting important information from the claims. The model showed a significant improvement in accuracy compared to previous methods, especially in subgroup-level prediction. They extracted nouns and their proportions from the claims as input text, but suggest that other methods of information compression should be considered. Zuo et al. [5] compared different approaches to automatically classify French patent documents at the IPC main group and subgroup-levels. Their experiments showed the need for more sophisticated techniques such as data augmentation, clustering, and negative sampling at deeper levels such as subgroups. Chen and Chang [9] proposed a three-step classification (TPC) algorithm that achieved 36.07% accuracy at the IPC subgroup-level classification. D'hondt et al. [10] demonstrated the effectiveness of combining words and PoS-filtered skipgrams. They showed that extending the textual representation from traditional word-only based features to more fine-grained phrase-based features significantly improves the performance of automatic classification at the subgroup-level.

2.2. Patent Document Summarization and LLM-based Summary Generation

Sharma et al. [11] proposed BIGPATENT, a large dataset containing 1.3 million U.S. patent documents and abstract and coherent summaries written by humans. Experiments with BIGPATENT suggest that summarization tasks for specialized texts such as patent documents require deeper understanding and abstraction than simply extracting phrases from the original text. Ding et al. [12] generated summaries for 1630 patent document abstract and claims combinations using several text summarization models. They concluded that the GPT-3.5-turbo model can summarize patent documents better than other models, and they stated that prompting strategy is the key to the success of patent document summarization. Yang et al. [13] evaluated ChatGPT's performance on various text summarization tasks using a benchmark dataset and found that the summaries generated by ChatGPT were comparable to those generated by traditional methods, suggesting that ChatGPT is a promising powerful tool for text summarization. As in the above study, there have been reports of using LLMs to generate summaries of long texts, including patent documents, and evaluating their quality. However, to the best of my knowledge, there are no reported cases of using LLMs to generate a summary of a patent document and using the generated summary to perform a patent classification task.

3. Datasets

For our experiments, we used two popular patent datasets: one from USPTO and one from CLEF.

USPTO-70k Dataset. The USPTO-70k [14] is a dataset for CPC subclass-level patent classification tasks shown in the upper part of Table 1. It consists of training data from 50250 USPTO patent documents from 2006 to 2017, validation data from 10,000 documents in 2018, and testing data from 10,000 documents in 2019. This dataset contains various text fields of patent documents such as title, abstract, claims, and detailed description.

Since the USPTO-70k does not contain subgroup-level classification information, we augmented the dataset with subgroups of the main IPC (the IPC that best represents the technical field to which the patent belongs) by referring to USPTO Bulk Data Storage System.[1] As in the previous studies [9, 10], only subgroup labels with at least seven training documents were retained for the subgroup-level classification task. After data cleaning, the patent documents shown in the middle part of Table 1 were obtained.

CLEF-IP 2011 Subset. The CLEF-IP 2011 dataset [15] consists of over 2.6 million patent documents from the EPO and 0.4 million patent documents from the WIPO, filed between 1978 and 2009. We extracted EPO English patent documents from 2000 to 2009 that contain IPC labels, title, abstract, claims, and detailed description from the CLEF-IP 2011 dataset. Training data and validation data were collected according to the data size of USPTO-70k dataset, as shown in the lower part of Table 1. We used the CLEF-IP test sample (1000 documents) as test data. The word-piece token distributions in the different text fields of the CLEF-IP 2011 subset shown in Figure 1 show a similar distribution to the token distribution for the USPTO-70k dataset presented by Pujari et al. [14]

While the USPTO has adopted the concept of "main IPC" and uniquely determines the IPC corresponding to every patent document, the EPO does not follow the main IPC rule. Therefore, the subgroup-level multi-class classification task was performed only on the USPTO-70k dataset.

Table 1
Description of datasets

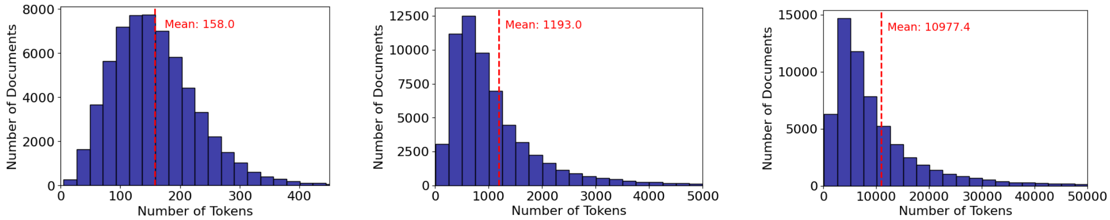
	Train	Valid	Test
USPTO-70k			
Subclass Multi-label			
# of documents	50,625	10,000	10,000
Avg. labels per patent	1.98	2.25	2.32
Total labels	630	573	585
Subgroup Multi-class			
# of documents	30,251	4,784	4,288
Total labels	1282	952	860
CLEF-IP 2011 subset			
Subclass Multi-label			
# of documents	50,000	10,000	1,000
Avg. labels per patent	2.00	2.01	1.87
Total labels	605	554	323

4. Experimental Setup

In this study, the following procedure was used for the experiments:

1. **Summary Generation.** Using each text field (abstract, claims, and detailed description) of the patent document as input, the LLM generated a summary for each text field.
2. **Patent Classification.** AI-generated summaries were used to fine-tune the pre-trained models to adapt them to a multi-label or multi-class classification task. Multi-label classification is a problem where each sample may belong to more than one label, and multi-class classification is a problem where each sample belongs to one of the classes. The fine-tuned model predicts the classification using the generated summary as input.
3. **Evaluation.** To evaluate the performance of the fine-tuned model, the prediction results were an-

[1] https://bulkdata.uspto.gov/



(a) Abstract (b) Claims (c) Detailed description

Figure 1: Token distributions for each different text field on CLEF-IP 2011 subset

Table 2
Three types of system role messages for summary generation

Prompt type	System role message
Patent	You are a patent expert. Summarize the patent document given by the user in 100 words. Output only the summary.
Simple	Summarize the document given by the user in 100 words. Output only the summary.
Elaborate	You are a patent expert. Summarize the patent document given by the user in 100 words focusing on an addition to the state of the art. "Addition to the state of the art" means the difference between the subject matter in a patent document and the collection of all technical subject matter that has already been placed within public knowledge. Output only the summary.

alyzed using hierarchical precision, recall, and F1-score (See section 4.3.). For comparison, the same evaluation was performed on the results of model training and classification prediction with the original texts instead of AI-generated summaries.

Each procedure is further described in detail in the following subsections.

4.1. Summary Generation

We used the gpt-3.5-turbo-0125 model to generate summaries from patent documents. The prompts given to the model consisted of one of the three types of system role messages shown in Table 2 and a user role message containing the patent text. The "Elaborate" prompt in Table 2 is based on the description in VIII. PRINCIPLES OF THE CLASSIFICATION of the IPC guidelines [1].[2]

To ensure as much reproducibility of the experiment as possible, the temperature value of the model was set to 0.0, and a seed was specified (seed=42). With the seed set, the system will do its best to sample deterministically, and repeated requests with the same seed and parameters should return the same results. Note, however, that this is currently a beta feature and does not always produce exactly the same output.[3]

Prompts are created so that the sum of system and user messages does not exceed the maximum context window size (16,385 tokens). If the maximum context window size is exceeded, the first portion of text up to the maximum context window size is used and the remainder is truncated.

4.2. Patent Classification

In this study, we applied the RoBERTa (Robustly Optimized BERT Pretraining Approach) model [16] to multi-label/multi-class classification tasks for patent documents. RoBERTa has been reported to require less time for fine-tuning than other transformer-based models [2]. We used the pre-trained RoBERTa-Base model and adapted it to the patent classification task by adding a dropout layer and a linear layer of size equal to the number of classification classes. The model was optimized with the Adam optimizer using the hyperparameters shown in Table 3.

Furthermore, based on the results of preliminary experiments conducted with reference to Merchant et al. [17],

[2] The "Elaborate" prompt requires a timestamp as the basis for the "collection of all technical subject matter already in the public domain" to be followed faithfully. However, since it would be unfair to give metadata for individual patent documents only to the "Elaborate" prompt and not to the other prompts ("Patent", "Simple"), we decided not to give their time stamps. Detailed descriptions usually include a description of the background technology, so it would be possible to infer "state of the art" based on that information.

[3] https://platform.openai.com/docs/api-reference/chat/create

Table 3
Hyperparameters for fine-tuning RoBERTa

Hyperparameter	Value
Batch size	64
Learning rate	5e-5
Number of epochs	15
Max sequence length	512
Dropout	0.2

we found that freezing all layers except the linear layer, the pooler layer, and the last three layers of RoBERTa had little effect on performance. Therefore, we unfroze these layers and performed fine-tuning for training efficiency.

For the multi-label classification task, we trained the model with the prediction threshold set to 0.5, based on the report by Giczy et al. [18]. After training, the prediction threshold of the model was varied from 0.1 to 0.9 in 0.1 increments, label predictions were made on the validation data, and the prediction threshold with the highest micro-average hierarchical F1-score (See section 4.3.) was selected as the final model. As shown in Figure 2, the maximum micro-averaged F1-score was obtained with a prediction threshold of 0.2 and 0.3 when using the USPTO-70k dataset and the CLEF-IP 2011 subset, respectively.

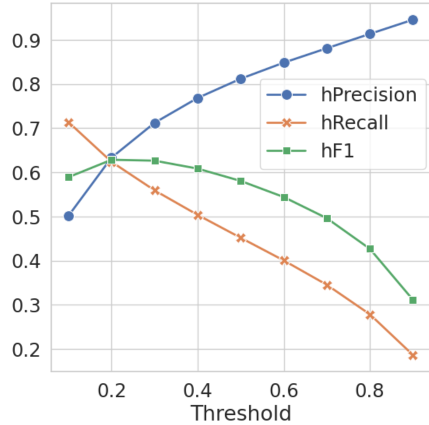
4.3. Evaluation

Evaluating AI-generated Summaries. In this study, two automatic evaluation metrics, ROUGE [19] and BERTScore [20], were used to evaluate the AI-generated summaries. For both indicators, scores were calculated using the original text of the patent document abstract as the reference summary.

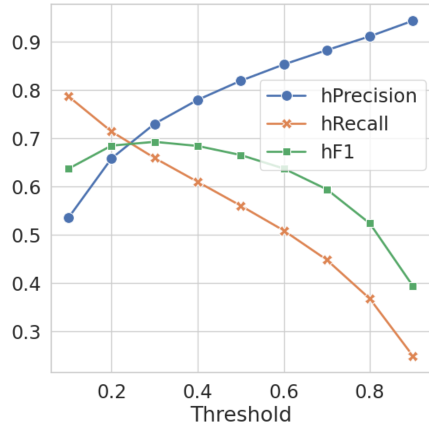
The ROUGE score is a measure that evaluates the degree of n-gram overlap between the generated summary and the reference summary. ROUGE-1 and ROUGE-L were used in this study; ROUGE-1 measures unigram overlap, while ROUGE-L evaluates summary similarity based on the longest common subsequence.

The BERTScore is a measure that evaluates the similarity between generated and reference summaries using the pre-trained language model BERT. AI-generated summaries are likely to use different vocabulary than the original abstracts, but by using BERTScore, word semantic similarity can be taken into account.

Evaluating Patent Classification Performance. Following Pujari et al. [14, 6], we used hierarchical precision, recall, and F1-score defined as $hP = \frac{\sum |P_i \cap T_i|}{\sum |P_i|}$, $hR = \frac{\sum |P_i \cap T_i|}{\sum |T_i|}$ and $hF1 = \frac{2 \cdot hP \cdot hR}{hP + hR}$ and proposed by Kiritchenko et al. [21] to evaluate subclass-level multi-label classification. For each patent document i in the test data, the predicted label set P_i consists of all predicted labels and their ancestors. Similarly, the true label set T_i consists of true labels and their ancestors. These evaluation metrics consider the degree of agreement between predicted and true labels in hierarchically structured classifications, even at higher levels of the hierarchy. For example, if the true label is "G06F", the predicted label of Model A is "G06K", and the predicted label of Model B is "H04L", the prediction of Model A, which is consistent with the class-level "G06", is rated better than that of Model B. The subgroup-level multi-class classification was evaluated using accuracy, following Chen and Chang [9] and D'hondt et al. [10]. Since this is a multi-class classification task, the accuracy is consistent with the micro-averaged F1-score.



(a) USPTO-70k dataset



(b) CLEF-IP 2011 subset

Figure 2: Comparision of various threshold values for RoBERTa model (subclass-level classification, abstract original text)

Table 4
Examples of original patent document texts and AI-generated summaries on USPTO-70k dataset

Text field	Text
Abstract Original text	"An exterior mirror assembly including an attachment member for supporting an approach light. The attachment member interconnects a mirror housing to a vehicle and includes an opening for receiving a lens. Light projects through the lens from the attachment member in order to illuminate a predetermined area in proximity to the vehicle. A light source may be housed within the support member or, alternatively, a light source may be housed interior to the vehicle and a light path transport light from the light source to the lens for projection from the support member."
Claims Original text (first 100 words)	"1. An exterior rear view mirror assembly comprising: a housing; a reflective mirror supported by the housing and arranged in a rearwardly facing direction; a support member having an outboard end attached to the housing and an inboard end secured to an outer surface of a motor vehicle, the support member including an opening; and a light transmitting lens disposed in the opening of the support member, the lens enabling light to project from an interior of the support member to an exterior of the support member to illuminate a predetermined area in proximity to the motor vehicle. 2. The"
Detailed description Original text (first 100 words)	"DESCRIPTION OF THE PREFERRED EMBODIMENTS Referring now to the figures, and in particular FIGS. 1–3, a mirror assembly10includes a housing12which houses and supports a reflective element14, such as mirror glass. Mirror assembly10of the present invention will be described with respect to an exterior driver side rear view mirror which enables the vehicle operator to view an area beside and rearward with respect to the vehicle. Housing12connects to an attachment member16including a sail portion18and an integral, laterally projecting support arm or bracket portion20. The interconnection between housing12and attachment member16enables housing12to pivot with respect to the support arm portion20. Particularly, housing12pivots rearward and"
Abstract AI-generated summary	"The patent describes an exterior mirror assembly with an attachment member supporting an approach light. The attachment member connects the mirror housing to a vehicle and has an opening for a lens. Light passes through the lens to illuminate an area near the vehicle. The light source can be within the attachment member or inside the vehicle, with a light path transporting light from the source to the lens for projection."
Claims AI-generated summary	"The patent document describes an exterior rear view mirror assembly for a motor vehicle. It includes a housing, a reflective mirror, a support member with a light transmitting lens for illuminating an area near the vehicle, and optional features like a light assembly with adjustable light source, reflective elements, environmental seal, sail for securing the support member, and a cover for the lens. The system can also include a timer, speed sensor, switch, and control circuit for activating the light source based on vehicle speed or various sensors."
Detailed description AI-generated summary	"The patent document describes a mirror assembly for vehicles with an integrated approach light system. The mirror assembly includes a housing supporting a reflective element and an attachment member with a sail portion and a support arm portion. The approach light can be located in the support arm portion or sail portion, providing illumination for areas adjacent to the vehicle. Various configurations for the approach light placement, activation, and control are discussed, including adjustable reflectors, variable timeout functions, and integration with vehicle systems. The invention aims to improve visibility, safety, and serviceability of the approach light system in vehicle mirror assemblies."

5. Results and Discussion

5.1. Summary Generation

Table 4 shows examples of original patent document text and AI-generated summaries on the USPTO-70k dataset. Both the original text and the AI-generated summary of the abstract describe the basic content of the invention. The original text of the claims is written in a manner specific to patent claims. In addition, both the original text and the AI-generated summary of the claims specifically describe the components of the invention. This is because patent claims must clearly indicate the scope of the patent right. The original text of the detailed description describes a specific embodiment of the invention. The AI-generated summary of the detailed description describes the most comprehensive content of the invention, such as its structure, function, and purpose, compared to the other texts.

Table 5 shows the metrics for the summary generated

Table 5
Comparison of summary evaluation metrics for each AI-generated summary

Dataset	Text field	Avg. # of words	ROUGE-1	ROUGE-L	BERTScore-F1
USPTO-70k	AB (AI)	59.4 ± 16.7	0.548	0.469	0.910
	CL (AI)	86.5 ± 11.8	0.399	0.261	0.868
	DD (AI)	97.9 ± 12.8	0.372	0.236	0.862
CLEF-IP 2011	AB (AI)	63.4 ± 17.7	0.520	0.423	0.894
	CL (AI)	85.3 ± 12.6	0.379	0.248	0.859
	DD (AI)	96.1 ± 7.4	0.357	0.225	0.855

AB: Abstract, CL: Claims, DD: Detailed description, AI: AI-generated summary

Table 6
Performance comparison on USPTO-70k dataset for subclass-level multi-label classification

Classifier	Text field	Micro-avg.						Macro-avg.	
		hP	hR	hF1	P	R	F1	hF1	F1
RoBERTa	AB (OT)	62.9	61.1	62.0	53.6	52.8	53.2	30.2	21.4
	CL (OT)	61.6*	60.9	61.2**	52.1*	52.6	52.3**	30.3	21.3
	DD (OT)	57.1**	47.9**	52.1**	46.8**	39.3**	42.7**	22.0**	14.8**
	AB (AI)	63.4	61.1	62.2	53.9	52.8	53.4	30.5	21.4
	CL (AI)	64.7*	63.7**	64.2**	55.6**	55.6**	55.6**	33.4**	24.0**
	DD (AI)	**65.3***	**64.3****	**64.8****	**56.2****	**56.1****	**56.1****	**33.5****	**24.4****
THMM [6]	CL (OT)	67.5	58.0	62.4	-	-	-	31.0	-
	DD (OT)	66.9	54.1	59.8	-	-	-	29.9	-
	SC (OT)	70.2	65.0	67.5	-	-	-	38.9	-

hP: hierarchical Precision, hR: hierarchical Recall, hF1: hierarchical F1-score, P: Precision, R: Recall, F1: F1-score, AB: Abstract, CL: Claims, DD: Detailed description, SC: Title, Abstract, Claims, Detailed description, Brief-Summary, and Figure description, OT: Original text, AI: AI-generated summary
*p<0.05, **p<0.01

Table 7
Performance comparison on USPTO-70k dataset for subgroup-level multi-class classification

Text field	hAcc	Acc	Macro-avg.	
			hF1	F1
AB (OT)	52.2	22.3	16.0	11.0
CL (OT)	52.0	21.9	16.0	11.1
DD (OT)	41.2**	14.3**	9.8**	6.5**
AB (AI)	52.0	22.0	16.4	11.3
CL (AI)	**53.9****	**23.9****	17.7**	12.9**
DD (AI)	**53.9****	23.5**	**18.2****	**13.1****

hAcc: hierarchical Accuracy, Acc: Accuracy, AB: Abstract, CL: Claims, DD: Detailed description, OT: Original text, AI: AI-generated summary
*p<0.05, **p<0.01

from each text field in the patent document. The ROUGE score represents the score of the respective AI-generated summary of the abstract, claims, and detailed description when the original text of the abstract is used as the reference; the BERT score is similar. In both datasets, each score decreased as text moved from abstract to claims to detailed description. This means that summaries generated from claims and detailed description contain more words and meanings that are different from the original text of the abstract. As the text moved from abstract to claims to detailed description, the decrease in BERTScore was smaller than the decrease in ROUGE score. This indicates that although matching at the word level is decreasing, semantic similarity is relatively maintained. This indicates that the gpt-3.5-turbo-0125 model tends to generate contextually relevant summaries for longer and more complex texts, such as claims and detailed description.

5.2. Patent Classification

USPTO-70k Dataset. Tables 6 and 7 show the results of the patent classification task using AI-generated summaries as training data for the "Patent" prompt (shown in Table 2) on the USPTO-70k dataset. According to Welch's t-test with a sample size of 5, scores were significantly higher for both subclass-level multi-label classification and subgroup-level multi-class classification tasks when using AI-generated summaries of claims or detailed de-

Table 8
Performance comparison on CLEF-IP 2011 subset for subclass-level multi-label classification

Classifier	Text field	Micro-avg.						Macro-avg.	
		hP	**hR**	**hF1**	**P**	**R**	**F1**	**hF1**	**F1**
RoBERTa	AB (OT)	69.4	65.0	67.1	59.5	57.6	58.5	22.9	18.7
	CL (OT)	69.5	62.5*	65.8**	59.8	55.2*	57.4**	22.2	18.2
	DD (OT)	70.1	68.0**	69.0**	60.3	60.3**	60.3**	25.1**	20.9**
	AB (AI)	70.2	64.6	67.3	60.7	57.1	58.8	22.9	18.8
	CL (AI)	71.5**	67.1*	69.2**	62.1**	59.6*	60.8**	24.5*	20.2**
	DD (AI)	**72.5****	**70.6****	**71.5****	**63.0****	**63.3****	**63.1****	**27.2****	**22.8****

hP: hierarchical Precision, hR: hierarchical Recall, hF1: hierarchical F1-score, P: Precision, R: Recall, F1: F1-score, AB: Abstract, CL: Claims, DD: Detailed description, OT: Original text, AI: AI-generated summary
*p<0.05, **p<0.01

Table 9
Performance comparison of AI-generated summaries with three different prompts (patent, simple, and detailed) for subclass-level multi-label classification

Dataset	Text field	Prompt	Micro-avg.						Macro-avg.	
			hP	**hR**	**hF1**	**P**	**R**	**F1**	**hF1**	**F1**
USPTO-70k	DD (AI)	Patent	**65.3**	**64.3**	**64.8**	**56.2**	**56.1**	**56.1**	33.5	24.4
		Simple	64.9	64.3	64.6	56.0	**56.1**	56.0	**34.4**	**24.8**
		Elaborate	64.2	64.2	64.2*	55.3	**56.1**	55.7*	33.9	24.7
CLEF-IP 2011	DD (AI)	Patent	72.5	70.6	71.5	63.0	63.3	63.1	27.2	22.8
		Simple	**72.6**	**71.4***	**72.0**	**63.5**	**64.2***	**63.9***	**27.9**	**23.6**
		Elaborate	72.3	69.2*	70.7	62.8	62.1	62.4	26.2	21.9

hP: hierarchical Precision, hR: hierarchical Recall, hF1: hierarchical F1-score, P: Precision, R: Recall, F1: F1-score, DD: Detailed description, AI: AI-generated summary
*p<0.05, **p<0.01

Table 10
Performance comparison of AI-generated summaries with three different prompts on USPTO-70k dataset for subgroup-level multi-class classification

Prompt	hAcc	Acc	Macro-avg.	
			hF1	**F1**
Patent	54.2	23.6	18.2	13.1
Simple	54.0	23.5	18.0	12.9
Elaborate	**54.3**	**24.4***	**18.5**	**13.4**

hAcc: hierarchical Accuracy, Acc: Accuracy, hF1: hierarchical F1-score, F1: F1-score, DD: Detailed description, AI: AI-generated summary
*p<0.05, **p<0.01

scriptions than when using the original abstract text. However, using AI-generated summaries of abstracts did not result in significant differences in scores. This suggests that the generative model can provide effective text for classification tasks by summarizing important information from claims or detailed description.

The scores were significantly lower when using the original text of the detailed description than when using the original text of the abstract. This may be because the detailed description on the USPTO-70k dataset often begins with figure captions or notes on the scope of the invention's disclosure, and in many cases, the technical features of the invention are not included in the first 100 words.

Comparing the results with Pujari et al.'s THMM [6], the combination of RoBERTa and the AI-generated summary was superior to the combination of THMM and the original text when claims or detailed descriptions were used. In particular, the micro-averaged hierarchical F1-score was 5.0 points higher, and the macro-averaged hierarchical F1-score was 3.6 points higher when detailed descriptions were used, indicating that AI-generated summaries effectively extract important information from long texts such as detailed descriptions.

CLEF-IP 2011 Subset. Table 8 shows the results on the CLEF-IP 2011 subset. each score was significantly higher when using the AI-generated summary of claims and detailed descriptions than when using the original abstract text.

In particular, the micro-averaged hierarchical F1-score improved by 4.4 points when using the AI-generated summaries of the detailed description compared to when using the original abstract text. This indicates that the AI-generated summaries effectively extracts classification-relevant information from the detailed description.

The scores were also significantly higher when the original text of the detailed description was used. This may be because EPO patent documents often include the technical field to which the invention belongs (e.g., "The present invention relates to a sound and heatinsulating material.) at the beginning of the detailed description.

In all cases, the results are below Yadrintsev et al.'s KNN [7]. This is likely due to the fact that this study uses only a single text field and about one-tenth the size of the CLEF-IP 2011 dataset for training data.

The best (non-hierarchical) F1-score for this method is 63.1 points when using AI-generated summaries of detailed description, which is lower than the scores from previous studies such as those from DeepPatent by Li et al. [4] (83.98 points) and KNN by Yadrintsev et al. [7] (71.0 points). This is likely due to the fact that this method uses only a single text field and only about one-tenth the size of the CLEF-IP 2011 dataset as training data.

Prompt Comparison. Tables 9 and 10 shows a comparison of classification performance when using AI-generated summaries by each of the three prompts (Patent, Simple, and Elaborate).

For the subclass-level classification task, the performance with the "Patent" prompt and with the "Simple" prompt is competitive on the USPTO-70k dataset, while the performance with the "Simple" prompt is slightly higher on the CLEF-IP 2011 subset. On the other hand, performance with the "Elaborate" prompt tends to be slightly poorer on both datasets. These results suggest that overly detailed summaries are not necessarily effective for subclass-level patent classification tasks but rather that even summaries generated with the "Simple" prompt are effective enough.

For the subgroup-level classification task, the performance of the "Elaborate" prompt was slightly better than that of the "Patent" and "Simple" prompts. The results suggest that the summary generated by the "Elaborate" prompt may be useful for more detailed classification tasks.

From the above, we believe that in the patent classification task, it is important to use prompts with a reasonable level of detail depending on the classification level, although the improvement in performance obtained by optimizing the prompts is limited. However, since only three prompts were compared in this experiment, a comprehensive investigation of the effects of the various prompts on performance on the patent classification task is a topic for future work.

6. Conclusion

In this study, we investigate the effect of using summaries generated by LLMs to improve the performance of automatic patent classification. Our experiments on the USPTO-70k dataset and the CLEF-IP 2011 subset demonstrated that models trained on AI-generated summaries of claims and detailed descriptions achieve significantly higher scores compared to those trained on original abstract text in both subclass-level multi-label classification and subgroup-level multi-class classification tasks. These results suggest that AI-generated summaries adequately capture information relevant to patent classification.

The proposed approach improves automatic patent classification techniques by utilizing LLMs to generate high-quality summaries. We believe that this research builds new possibilities for improving the accuracy and efficiency of patent classification, which is important for managing the ever-increasing amount of patent information.

Future research directions include exploring the optimal prompts for generating summaries of patent documents by LLMs and investigating the applicability of the proposed approach to other patent classification models. Although this study experimented with a flat approach that does not consider the hierarchical structure of patent classification, the performance of an automatic patent classification system could be further enhanced by incorporating the interdependence of each hierarchical label into the automatic classification process.

References

[1] World Intellectual Property Organization, Guide to the international patent classification (2023), https://www.wipo.int/edocs/pubdocs/en/wipo-guide-ipc-2023-en-guide-to-the-international-patent-classification-2023.pdf, 2023. Accessed: 2024-4-16.

[2] A. Haghighian Roudsari, J. Afshar, W. Lee, S. Lee, Patentnet: multi-label classification of patent documents using deep learning based language understanding, Scientometrics 127 (2022) 207–231.

[3] J.-S. Lee, J. Hsiang, PatentBERT: Patent classification with fine-tuning a pre-trained BERT model, World Patent Information 61 (2020).

[4] S. Li, J. Hu, Y. Cui, J. Hu, DeepPatent: patent classification with convolutional neural networks and word embedding, Scientometrics 117 (2018) 721–744.

[5] Y. Zuo, B. Sagot, K. Gerdes, H. Mouzoun, S. Ghamri Doudane, Exploring data-centric strategies for French patent classification: A baseline and comparisons, in: Actes de CORIA-TALN 2023. Actes de la 30e Conférence sur le Traitement Au-

tomatique des Langues Naturelles (TALN), volume 1 : travaux de recherche originaux – articles longs, ATALA, 2023, pp. 349–365.

[6] S. C. Pujari, F. Mantiuk, M. Giereth, J. Strötgen, A. Friedrich, Evaluating neural multi-field document representations for patent classification, in: BIR 2022: 12th International Workshop on Bibliometric-enhanced Information Retrieval at ECIR 2022, April 10, 2022, hybrid, 2022.

[7] V. Yadrintsev, A. Bakarov, R. Suvorov, I. Sochenkov, Fast and accurate patent classification in search engines, Journal of Physics: Conference Series 1117 (2018) 012004. doi:10.1088/1742-6596/1117/1/012004.

[8] Y. Hoshino, Y. Utsumi, Y. Matsuda, Y. Tanaka, K. Nakata, IPC prediction of patent documents using neural network with attention for hierarchical structure, PLoS One 18 (2023) e0282361. doi:https://doi.org/10.1371/journal.pone.0282361.

[9] Y.-L. Chen, Y.-C. Chang, A three-phase method for patent classification, Information Processing & Management 48 (2012) 1017–1030. doi:https://doi.org/10.1016/j.ipm.2011.11.001.

[10] E. D'hondt, S. Verberne, N. Oostdijk, L. Boves, Patent Classification on Subgroup Level Using Balanced Winnow, Springer Berlin Heidelberg, 2017, pp. 299–324. doi:10.1007/978-3-662-53817-3_11.

[11] E. Sharma, C. Li, L. Wang, BIGPATENT: A large-scale dataset for abstractive and coherent summarization, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2019, pp. 2204–2213. doi:10.18653/v1/P19-1212.

[12] J. Ding, H. Chen, S. Kolapudi, L. Pobbathi, H. Nguyen, Quality evaluation of summarization models for patent documents, in: 2023 IEEE 23rd International Conference on Software Quality, Reliability, and Security (QRS), 2023, pp. 250–259. doi:10.1109/QRS60937.2023.00033.

[13] X. Yang, Y. Li, X. Zhang, H. Chen, W. Cheng, Exploring the limits of chatgpt for query or aspect-based text summarization, 2023. arXiv:2302.08081.

[14] S. C. Pujari, A. Friedrich, J. Strötgen, A multi-task approach to neural multi-label hierarchical patent classification using transformers, in: Proceedings of the 43rd EUROPEAN CONFERENCE ON INFORMATION RETRIEVAL, Online, 2021.

[15] F. Piroi, M. Lupu, A. Hanbury, V. Zenz, CLEF-IP 2011: Retrieval in the intellectual property domain, in: CLEF 2011 Labs and Workshop, Notebook Papers, 19-22 September 2011, Amsterdam, The Netherlands, volume 1177 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2011.

[16] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, 2019. arXiv:1907.11692.

[17] A. Merchant, E. Rahimtoroghi, E. Pavlick, I. Tenney, What happens to BERT embeddings during fine-tuning?, in: Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, Association for Computational Linguistics, 2020, pp. 33–44. doi:10.18653/v1/2020.blackboxnlp-1.4.

[18] A. V. Giczy, N. A. Pairolero, A. A. Toole, Identifying artificial intelligence (ai) invention: A novel ai patent dataset, The Journal of Technology Transfer 47 (2022) 476–505.

[19] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, Association for Computational Linguistics, 2004, pp. 74–81.

[20] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert, in: International Conference on Learning Representations, 2020.

[21] S. Kiritchenko, S. Matwin, A. F. Famili, et al., Functional annotation of genes using hierarchical text categorization, in: Proc. of the ACL Workshop on Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics, 2005.