IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 26, 2022, accepted May 18, 2022, date of publication May 23, 2022, date of current version June 8, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3176877

AI for Patents: A Novel Yet Effective and Efficient Framework for Patent Analysis

JUNYOUNG SON[1], HYEONSEOK MOON[1], JEONGWOO LEE[2], SEOLHWA LEE[3], CHANJUN PARK[1,4], WONKYUNG JUNG[5], AND HEUISEOK LIM[1,2]

[1]Department of Computer Science and Engineering, Korea University, Seoul 02841, Republic of Korea
[2]Human-Inspired AI Research Institute, Seoul 02841, Republic of Korea
[3]Department of Computer Science, University of Copenhagen, 1017 Copenhagen, Denmark
[4]Upstage, Gyeonggi-do 16942, Republic of Korea
[5]LG Innotek, Seoul 07796, Republic of Korea

Corresponding author: Heuiseok Lim (limhseok@korea.ac.kr)

This work was supported in part by the Ministry of Science and ICT (MSIT), South Korea, under the Information Technology Research Center (ITRC) Support Program, Supervised by the Institute for Information and Communications Technology Planning and Evaluation (IITP), under Grant IITP-2018-0-01405; in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) Grant through the Korea Government (MSIT), a Neural-Symbolic Model for Knowledge Acquisition and Inference Techniques, under Grant 2020-0-00368; and in part by the Basic Science Research Program through the National Research Foundation of Korea (NRF), Ministry of Education, under Grant NRF-2021R1A6A1A03045425.

ABSTRACT Patents provide inventors exclusive rights to their inventions by protecting their intellectual property rights. However, analyzing patent documents generally requires knowledge of various fields, considerable human labor, and expertise. Recent studies to alleviate this problem on patent analysis deal only with the analysis of claims and abstract parts, neglecting the descriptions that contain essential technical cores. Moreover, few studies use a deep learning approach to handle the entire patent analysis process, including preprocessing, summarization, and key-phrase generation. Therefore, we propose a novel multi-stage framework that can aid in analyzing patent documents by using the description part of the patent rather than abstracts or claims with deep learning. The framework comprises two stages: key-sentence extraction and key-phrase generation tasks. These stages are based on the T5 model structure, transformer-based architecture that uses a text-to-text approach. To further improve the framework's performance, we employed two key factors: i) post-training the model with a patent-related raw corpus for encouraging the model's comprehension of the patent domain, and ii) utilizing a text rank algorithm for efficient training based on the priority score of each sentence. We verified that our key-phrase generation method of the framework shows higher performance in both superficial and semantic evaluation than other extraction methods. In addition, we provided the validity and effectiveness of our methods through quantitative and qualitative analysis, demonstrating the practical functionality of our methods. We also provided a practical contribution to the patent analysis by releasing the framework as a demo system.

INDEX TERMS Deep learning, key-sentence extraction, keyword extraction, patent, patent analysis, post training.

I. INTRODUCTION

The importance of the patent never diminishes. Patent grant inventors a monopoly on the economic value of their inventions and motivate them to disclose new technologies and ideas. By protecting intellectual properties, patent ensures that the corresponding inventions are used only with their permission. With the advancement of technology, numerous patents are being filed in various fields. These patents are open to the public and made available by various authorities in many countries or regions around the world, motivating the development of technology worldwide [1], [2]. Accurate analysis of patent implicitly also lead to the understanding of business trends, inspire new industry solutions, and make important investment decisions [3]. Without detracting from the value of the patent document, a careful analysis of the patent is required to evaluate important technical details and potential relationships. [4], [5].

The associate editor coordinating the review of this manuscript and approving it for publication was Wei Wei.

Prior to the registration or the appliance of the patent, several issues are required to be verified, including terms of novelty, inventive step, and industrial applicability. Specifically, investigating whether the targeting patent does not infringe existing patents is the essential step considering that the patent grants the monopoly to the invention [6]. This implies the importance of analyzing previous patents [7], [8]. The patent analysis includes a process wherein prior patents are examined; this examination reveals the novelty and technical values of the invention by analyzing prior related patent documents and evaluating the present patent based on its similarities and differences with prior works [9], [10].

However, patent analysis must have a certain degree of expertise in different research fields such as data mining, information retrieval, and business intelligence [11]. This implies that the patent analysis is a non-trivial task requiring considerable human labor and expertise [7], [12]. Generally, it requires a multi-stage process to analyze the patents, as shown in Fig. 1. Patent documents are lengthy and rich in technical and legal terms, which require tremendous human labor and time to analyze. Hence, it is hard to abstract the patent document, especially in the processing or abstracting stages. This issue highlights the necessity for tools that can be utilized for patent processing and analysis in real-world industrial services, which can effectively alleviate the difficulties and limitations of current patent analysis by reducing human labor and time cost [7].

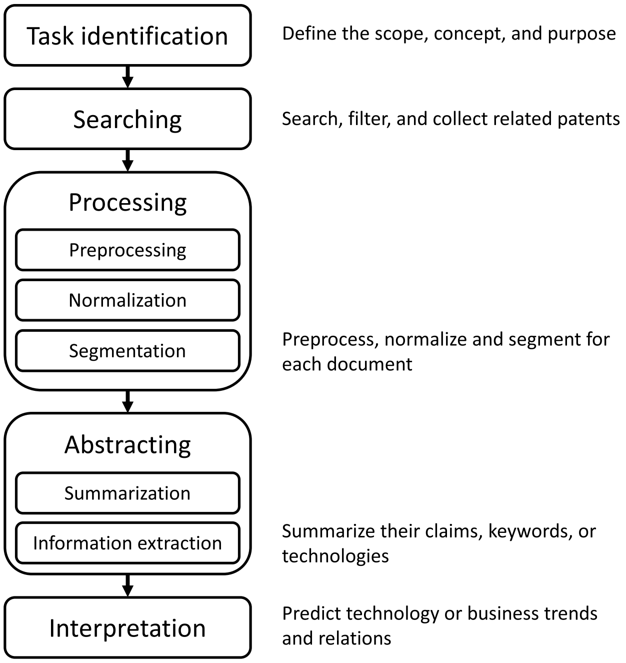

Task identification
Define the scope, concept, and purpose
Searching
Search, filter, and collect related patents
Processing
Preprocessing
Normalization
Segmentation
Preprocess, normalize and segment for each document
Abstracting
Summarization
Information extraction
Summarize their claims, keywords, or technologies
Interpretation
Predict technology or business trends and relations

FIGURE 1. Typical patent analysis process. Because the patent document is described in technical and legal terms and generally has lengthy contents and noises not directly related to the technical details, it requires tremendous human labor and time to generate an in-depth analysis.

Recently, several attempts have been made in applying deep learning techniques for patent analysis that alleviate the difficulties of processing and abstracting patent documents [13]. Although the detailed description contains more technical details than the other sections, most studies mainly focus on the patent analysis using claims or abstracts rather than the description because it often has lengthy and noisy contents, making it difficult to abstract and process [14], [15]. However, analyzing patent documents using only abstracts and claims may be insufficient for providing a detailed technical core because they include only a part of the technical contents of the overall invention in general and legal terms. Therefore, it is crucial to figure out detailed information about the technical core of the invention by utilizing only helpful information in the description section.

To alleviate the limitations of the current automatic patent analysis system, we propose a novel multi-stage framework focusing on analyzing patent documents by handling the lengthy and noisy difficulties of the description section. We focus on reducing the human labor of patent attorneys and analysts required to abstract and process patent documents by designing key-sentence extraction and key-phrase generation tasks with a state-of-the-art language model. Finally, we provide the technical cores of patent documents for future interpretation. Our method overcomes the difficulty of abstracting the detailed description part of the patent, which requires extensive human labor owing to its lengthy and noisy characteristics. Since we provide technical core information as key-sentences and key-phrases, it is easy to interpret and use for real-world patent analysis. To the best of our knowledge, we believe that this is the first work that alleviates the difficulties mentioned earlier in the patent analysis using the multi-stage analysis framework focused on the detailed description and deep learning.

In detail, we employ the Text-To-Text Transfer Transformer (T5) model [16], which shows state-of-the-art performances in many natural language tasks. In addition, as the patent analysis requires a wide range of knowledge, we utilize domain-adaptive post-training to improve comprehension of various fields of our framework. [17].

The first stage's goal is to reduce the time-consuming burden of summarizing the lengthy patent document. It abstracts the patent document using text-rank-based document summarization [18] and key-sentence extraction. This step abstracts the detailed description using the text rank algorithm and trains the key-sentence extraction model using the summarized description based on the T5 encoder. Specifically, we extract essential sentences of the patent document using the priority score of each sentence [19]. As a result, we can effectively increase the training efficiency by reducing the number of sentences in the document based on the text rank algorithm. We validate our intuition about data quality, which directly impacts the performance of our model, by filtering out data that negatively impact the training [20]–[22].

In the second stage, we use a text-to-text-based generation technique using the T5 encoder-decoder architecture. Because we treat the key-phrase extraction task as a text generation task, we can not only extract the phrases directly mentioned in the query but also generate semantically related phrases, considering the latent context that is not directly mentioned in the text. We verify that our generation method is semantically richer and more accurate than the existing key-phrase extraction method by analyzing how many the generative model implies the entire context. We also demonstrate accuracy in terms of surface matching of the key-phrase by confirming its higher surface matching performance than the extractive method, even at the character level.

In summary, our contributions are as follows.

- We propose a two-stage patent analysis framework that can be utilized in real-world industrial services using key-sentence extraction and key-phrase generation based on the T5 model, which has state-of-the-art performance in many NLP tasks. To the best of our knowledge, we believe that this is the first study that tries to resolve the difficulties mentioned above in the patent analysis using the multi-stage analysis framework.
- We figure out that applying the text-rank algorithm to the data pre-processing relieves lengthy and noisy problems of the patent's detailed description section. We also verify that this processing improves the document summarization task's overall performance and training efficiency. To the best of our knowledge, it is the first time to utilize a text-rank algorithm in the data-preprocessing for the deep learning-based patent summarization task.
- We demonstrate the effectiveness of the key-phrase generation approach in the patent domain via the quantitative and qualitative analysis, highlighting the limitation of the extractive key-phrase approach, which cannot extract key-phrases that consider the entire context of the document.
- We make our framework publicly available[1] Accordingly, we hope to contribute to the actual field of patent analysis.

In section II, we provide background knowledge of the structure of the patent document and relevant work of patent analysis. Section III describes our proposed methods: domain-specialized post-training, key-sentence extraction, key-phrase generation, overall framework, and demo system. Section IV-V provides quantitative and qualitative analysis of our proposed methods.

II. BACKGROUND AND RELATED WORK

This section illustrates the critical factors in patent documents and explores previous works, whose points of view in processing and abstracting have been mentioned in Fig. 1.

A typical patent document contains several essential attributes, such as the abstract, claims, and detailed description. The purpose of the abstract section is to provide a descriptive summarization of the invention. The abstract should be written clearly and concisely, not exceeding 150 words. The descriptive details of the patent could be minimal because of the usage of concise words to describe the related techniques in the abstract.

The claims section explicitly states the relevant technical characteristics for which protection is sought. Despite being simple and clear owing to the explicit description, claims are often written in legal terms. Therefore, the information may be abstract, and could therefore lead to a lack of detailed description of the actual purpose of the invention [23], [24].

Finally, the description section aims to explain the subject of the invention; it is written concisely and precisely with expertise over the techniques and sciences regarding the fundamental problems and solutions. Although the description section is specific and contains more technical details than other attributes in the patent documents, the context is very lengthy, and it is possible that noisy text is included.

Most previous studies for patent analysis focused on the abstract and claims because the description was lengthy and noisy. Chen *et al.* [14] categorized the patent documents into information extraction (IE) and named entity recognition (NER) using the Bi-directional Recurrent Neural Networks and Conditional Random fields and Hierarchical Attention Modules consisting of a word-level attention layer and a sentence-level attention. This research compared the performance between NER and IE based on training Word2vec [25] using the abstract and the whole text of the patent documents. Furthermore, they proved that training the model using the abstract outperformed that of the full-text. However, they did not attempt to handle the noise of the text when using full-text or making the summary.

Risch and Krestel [15] analyzed document classification using the title and abstract of the patent. They utilized the UPSTO-5M patent data to train the FastText [26] as word embeddings and utilized the Bi-GRU network to classify the document. Additionally, they truncated the title and abstract by 300 tokens to train the model. However, by only using the title and abstract as features to represent the patent, numerous technical information were omitted, which would have been otherwise necessary for patent analysis.

Unlike the aforementioned studies, some studies have implemented text segmentation to identify the most important parts of a patent using integrated unsupervised and supervised techniques to resolve lengthy and noisy text contained in the description [27]. They predefined the summary and experiment part by following the context flow of the description and then represented the segment using Word2vec and Term Frequency-Inverse Document Frequency (TF-IDF) as sentence features. In the end, the sentence features were used to train the segmentation boundary classification model. Although they utilized valuable information despite their lengthy description, it was oversimplified, in that it only classified the description into two predefined abstract classes. In addition, this study provided well-separated contents by segmenting the description; however, determining which part contained the core of the document remained a challenge.

[1] http://nlplab.iptime.org:9171/

Moreover, Korobkin *et al.* [28] developed the method for extracting the descriptions of the chemical effects and technical functions from the patent texts. They focused on segmenting the patent document by considering morphological features such as dependency trees. Borodin *et al.* [29] also paid attention to parsing the patent texts for the patent analysis. They developed software to extract the semantic subject-action-object (SAO) structures from the USPTO patent documents by parsing them into dependency trees and select potential technology partners based on them.

Patent experts mainly perform a key-phrase-based search when searching the related prior literature. In general, the key-phrase for the search is selected directly by the expert after reading the document [2], [30]. There are some works for key-phrase-based patent text summarization to automatize those processes using deep learning or machine learning. Korobkin *et al.* [31] proposed three steps methodology for patents prior-art. They first utilized multiple pre-trained Latent Dirichlet Allocation (LDA) models to obtain patent topic distributions. They also applied dependency parsing to patent texts for semantic analysis; finally, they measured similarities between the patents based on the parsed trees and re-rank patent documents based on the similarities. Hu *et al.* [2] proposed a method for extracting key-phrases first, and then classified the documentation using the extracted key-phrases in patent documentation. This study achieved better performances than previous statistical approaches, such as the TF-IDF and LDA by using the distributed representation-based key-phrase extraction.

Noh *et al.* [23] analyzed which part of the document was the most representative of the extracted key-phrase based on the concept that the patent document could have representativeness mainly in key-phrase units. Consequently, the TF-IDF method, which was based on abstract, had the best performance; nevertheless, they only considered the statistical key-phrase extraction method, and not the deep contextualized feature. Furthermore, they did not take into account the lengthy and noisy characteristics of the description for making the summary. Therefore, it cannot be confirmed that the key-phrase extracted from the abstract best represents the patent document. In addition, previous studies used a method of extracting key-phrases from documents for patent document analysis. In particular, when performing inter-document analysis and using key-phrases as representatives, this method could only perform a superficial comparison, and could not properly analyze documents composed of other words with similar meanings.

In this study, we focus on the description section, which contains high-quality technical details of the invention. In particular, we propose a multi-stage patent analysis framework based on key-sentence extraction and key-phrase generation to overcome the limitations of the description, which is difficult to analyze because of its lengthy and noisy characteristics, and to also generate semantically deeper contextualized key-phrases.

III. PROPOSED METHOD

A. DOMAIN SPECIALIZED POST TRAINING

1) BASELINE MODEL

Our goal in this study is to present an appropriate model for the generation of key-phrases as well as the extraction of key-sentences. The main approaches of the corresponding models are the auto-regressive generation method and the estimation of the priority score for a specific sentence in the document. By taking this into consideration, we utilize pre-trained T5$_{Large}$ [16] as a baseline model structure to construct a high-performance analysis model.

The corresponding model is trained via a multitask learning strategy that employs multiple Natural Language Processing (NLP) tasks simultaneously in a single training step, which has produced outstanding performances in various NLP fields, including natural language understanding and generation. In the pre-training phase, the model is trained by the multitask learning strategy [32], which includes unsupervised learning through the unlabeled monolingual text.

Specifically, a span corruption strategy [16] is adopted for the unsupervised learning. During this process, several word spans $s_i = \{x_j : x_{j+span_j}\}$ are selected randomly from the original unlabeled text $X = x_1, x_2, \ldots, x_n$, and the corrupted sentence X' is generated by replacing such word spans into the corresponding special tokens $\langle s_i \rangle$. Here, $span_j$ indicates the span length of the j^{th} word span. Given X', the model is trained to generate the replaced span set $S = \{s_1, s_2, \ldots, s_{n_s}\}$ in an auto-regressive process. The training objective of the model θ can be described by equation (1).

$$\max_{\theta} \sum_{(X',S)\in D} \left[\prod_{i=1}^{n} P(s_i|s_{<i}, X', \theta) \right]$$

where

$$S = \langle s_1 \rangle, s_1, \langle s_2 \rangle, s_2, \ldots, \langle s_{n_s} \rangle s_{n_s} \tag{1}$$

In addition to the unsupervised learning, the training of numerous NLP sub-tasks simultaneously occur in the pre-training phase of the T5 model. In this process, all the training processes of the NLP sub-tasks, including the natural language understanding tasks, are standardized into a text-to-text framework. This enables consistent training during pre-training and fine-tuning. Throughout these processes, the model θ is trained to generate the output sequence $Y = \{y_i\}_{i=1}^{m}$ in a sequence-to-sequence [33] based auto-regressive manner by feeding the pre-processed input sequence $X = \{x_i\}_{i=1}^{n}$. This is described in equation (2).

$$P(Y|X) = \prod_{i=1}^{n} p(y_i|X, y_{<i}, \theta) \tag{2}$$

The training objectives of the corresponding process are to minimize the cross-entropy loss between the output embedding representation derived by equation (2) and the reference sequence.

2) POST TRAINING STRATEGY

A post-training strategy is widely utilized in various NLP fields, including domain specialization and task-specific model training, such as sentiment analysis [34], [35]. These studies show that implementing post-training prior to the fine-tuning of the target task can yield considerable improvement in the model performance.

Thus, we proceed with post-training through the patent domain corpus to further improve the model performance. Considering the specificity of the patent domain, numerous entities that seldom exist in the general domain are included in the patent domain corpus. These may restrict the usage of the general-domain pre-trained model(*i.e.* publicly released pre-trained language models) in the patent domain.

We expect that post-training can alleviate such limitations and thus enhance the overall performance in the patent domain [17]. Therefore, we implement post-training in the general T5 model and leverage it in our key-sentence extraction and key-phrase generation models. For the post-training strategy, we adopt span corruption and key-phrase generation methodologies, and the training objectives of the corresponding processes are the same as equation 1 and equation 2.

B. KEY-SENTENCE EXTRACTION

1) NECESSITY FOR KEY-SENTENCE EXTRACTION

The key-sentence extraction model proposed in this study aims to distinguish the most important sentences that represent the main purpose of a patent document from the other sentences and extract them. To achieve this, we leverage the sentence extracting approach [36], which has the advantage of being able to generate key sentences without destroying important information, such as technical terminology.

In particular, the patent document description that this study mainly deals with is usually long and noisy [13]. Generally, one document contains hundreds of sentences wherein less than ten sentences indicate the core contents of the corresponding document. This shows that there are inherent limitations when it comes to the efficiency and effectiveness in dealing with patent documents when utilizing text-to-text generation, as the length of the input sequence could be extremely long [37]. The sentence extracting approach can effectively relieve such limitations by assessing the importance of the respective sentences in a document. We expect that by applying such an approach, long and noisy patent documents can be effectively analyzed, and the model can more efficiently predict the key-sentences of a document.

2) TRAINING STRATEGY

The key-sentence extraction model in our study is trained to quantitatively estimate the priority score of each sentence in a document. Leveraging this model, all the sentences in that document are sorted by their respective priority score, which is derived by the model. Subsequently, we regard high-ranked sentences as key-sentences and extract them.

To implement this process, all the sentences are annotated as either 0 or 1, in that order, that is, 1 if it is a reference sentence(*i.e.* key-sentence in a document) and 0 otherwise. Utilizing this, an extraction model is trained to discriminate the reference sentences in the document. For the training and inference of the model, the encoder hidden state related to the first token of the input sequence $h_0 \in \mathbb{R}^{d_{model}}$ is utilized. Through the pooling layer $W_h \in \mathbb{R}^{d_{model}\times 2}$ and its corresponding row vector W_{h1}, W_{h_0}, the importance of the given input sentence is estimated by the values of each label. Specifically, the priority score $score_X$ of the given sentence can be derived as shown in equation (3).

$$X = (x_0, x_1) = (\frac{\exp(W_{h0}h_0)}{\sum\limits_{i\in\{0,1\}} \exp(W_{hi}h_0)}, \frac{\exp(W_{h1}h_0)}{\sum\limits_{i\in\{0,1\}} \exp(W_{hi}h_0)})$$
$$score_X = x_1 - x_0 \tag{3}$$

During the training process, the main objective is to minimize the cross-entropy loss between the reference label and the model output X, and during the inference, $score_X$ is leveraged as the criterion for judging the importance of a given sentence. Through this, we estimate the priority of a sentence in a given document, and then extract the key-sentences by selecting k sentences with the highest score.

3) TRAINING DATA UP-SCALING

Generally, the detailed description section of the patent document consists of a large number of sentences although only a few sentences are regarded as reference sentences. The major limitation that restricts the performance of the sentence extraction approach is the label imbalance problem. In constructing the training data using such a corpus, an excessively large portion of the training data is labeled as 0, which indicates that the corresponding sentence is a non-reference sentence, whereas only a very small portion is labeled as 1. This raises concerns that the model may be trained to classify most sentences as non-reference sentences without considering their semantic information [38].

To resolve such limitations, we apply the up-scaling method to the training data [39]. This method aims to smooth the label-imbalance of the training data by duplicating specific labeled data the amount of which is relatively small. In this study, we up-scale the reference sentences in the training data, that is, data to be labeled as 1. In particular, we regard the scaling ratio to be a hyperparameter and empirically find the optimal ratio. Thus, we can considerably enhance the model performance and construct a high-performance sentence-extraction model.

4) TEXT-RANK BASED DATA PRE-PROCESSING

To enhance the training efficiency, we adopt a text rank [18] based data pre-processing method. It is a graph-based method that ranks the relative importance of sentences in a document and subsequently summarizes the entire document using a deduced rank. This is shown in equation (4)

$$WS(V_i) = (1-d) + d * \sum_{V_j \in \mathrm{In}(V_i)} \frac{w_{ji}}{\sum_{V_k \in \mathrm{Out}(V_j)} w_{jk}} WS(V_j) \tag{4}$$

In this equation, d is the hyper-parameter which indicates damping factor. Through d, the thresholds for the data selection and the ratio for the summarization are determined. $WS(V_i)$ indicates the quantified importance of sentence V_i, and w_{ji} is a weight calculated based on the sentence similarity between V_j and V_i. $\mathrm{In}(V_i)$ is the set of vertices which point to V_i, and $\mathrm{Out}(V_j)$ is the set of vertices which V_j points to. $\mathrm{In}(V_i)$ and $\mathrm{Out}(V_i)$ are also denoted as predecessors and successors set of V_i, respectively. Following this process, WS for each sentence included in the patent document is accumulated, and all the sentences in the document are sorted by their corresponding WS. Text summarization is performed by selecting sentences that have a high WS value. Through this method, unnecessary data in the document can be effectively trimmed, and the training time can be effectively reduced by cutting down the overall amount of sentences in each document. By applying this method, we can construct a model that effectively extracts the key sentences in the lengthy and noisy description; this can directly affect the performance of the key-phrase generation model that takes the output of the key-sentence extraction model as an input.

C. KEY-PHRASE GENERATION

1) NECESSITY FOR KEY-PHRASE GENERATION

A key-phrase is the core information that describes a sentence effectively. As it provides a concise description that encapsulates the overall information of the sentence, patent experts vigorously leverage key-phrases in patent documents analysis [2], [30]. Hence, for the utilization of key-phrases in a specific sentence, many studies focus on key-phrase extraction from given sentences.

However, extracted key-phrases may lack semantic information and may not wholly capture the information of the given sentence, as it solely consists of the words directly present in that sentence. Therefore, we employ a key-phrase generation method that differs from existing extractive approaches. We expect that the generation method can effectively resolve several limitations present in extractive approaches and can semantically derive rich key-phrases that imply the whole information of the sentence while also preserving the original meaning.

2) TRAINING STRATEGY

For the construction of the key-phrase generation model, the whole architecture of the encoder-decoder structure is utilized for the auto-regressive generation of the key-phrases. We fine-tune the sequence-to-sequence-based key-phrase generation task to our post-trained model. The training objective of the corresponding task is shown in equation (5), which is similar to equation (2).

$$\max_{\theta} \frac{1}{\|D\|} \sum_{(S,K)\in D} \log \left[\prod_{i=1}^{N_k} P(k_i \mid S, k_{<i}, \theta) \right] \tag{5}$$

This equation shows the sequence-to-sequence [33] process of the encoder-decoder model structure, that takes a sentence S as an input of the encoder, and generates key-phrases $K = \{k_i\}_{i=1}^{N_k}$ auto-regressively. Specifically, D indicates the training dataset which components consists of the input sentences and key-pharses pair, and k_i is the i^{th} token in the generated key-phrase K. By feeding sentence S including key-phrase K to the model, the key-phrase generation model is trained to generate key-phrase K auto-regressively in a sequence-to-sequence manner. Generating key-phrases can contain words that are not included in the input sentence, and the generation model can generate relatively fluent and comprehensive key-phrases. Additionally, auto-regressive generation approaches can be robust to error relative to extractive approaches, in that they can generate semantically similar phrases to the input sentence, even if they differ grammatically or lexically, by referring to the whole input sentence.

D. OVERALL FRAMEWORK

We fuse the key-sentence extraction and key-phrase generation models to construct our two-stage framework. This framework can extract key-sentences in a patent document and generate semantically related key-phrases by using the key-sentences. We can significantly reduce the human labor required for patent analysis and contribute to the real-world industrial field. The overall process of our framework is shown in Fig. 2.

The framework comprises two models: the key-sentence extraction and the key-phrase generation models. We summarize the document via the text rank algorithm for the training and utilize the original document for the inference. This process removes unnecessary sentences and extracts technical core information so it can improve training efficiency. We skip the text rank processing for the inference phase because the text rank algorithm does not consider the semantic context and may remove reference sentences.

After the key-sentences are extracted from the document, the key-phrases are generated in a sequence-to-sequence manner by applying the key-phrase generation model using the extracted sentence as an input. Our framework provides both key-sentences and key-phrases for the further flexible utilization by end-users.

E. DEMO SYSTEM

Our framework is publicly available in the form of a demo system for the patent domain industrial field. An example of the actual implementation of the demo system is shown in Fig. 3. As end-users using this demo system feed the patent document description as input, the demo system analyzes the document and provides key-sentences in the document and the key-phrases generated from the key-sentences to users.

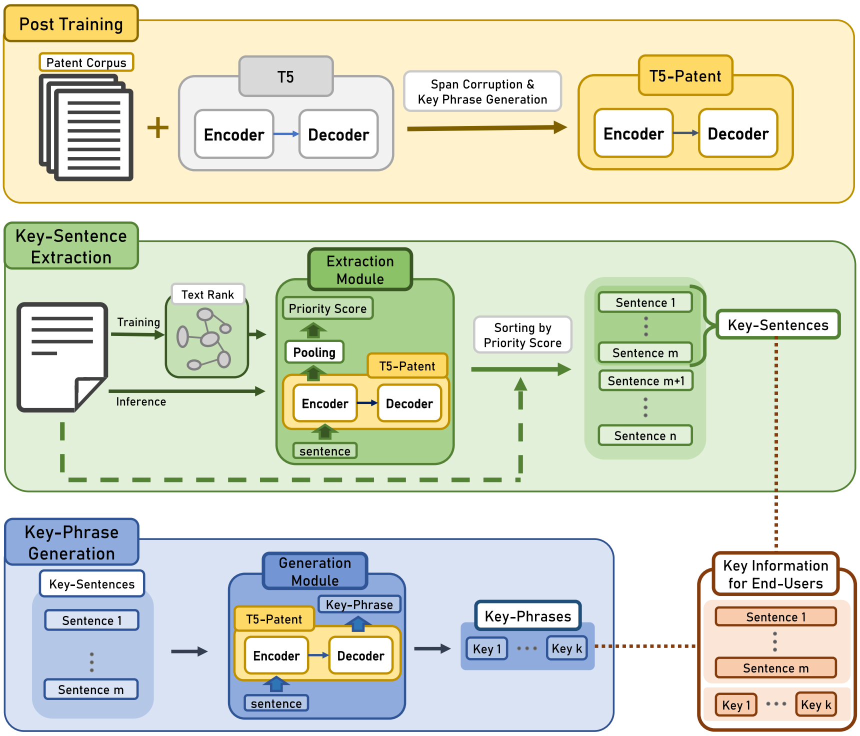


FIGURE 2. Overall process of our proposed model.

TABLE 1. Overview of tasks and datasets.

Task	Domain-specialized post-training		Key-sentence extraction	Key-phrase generation
	Key-phrase generation	Span corruption [16]		
Dataset	KP20K [40]	LG-Innotek patent analysis document		
Domain	Computer science article	Patent in the field of camera lens, sensor, and battery		
Data structure	Article key-phrases pair	Full patent document	Patent document-key-sentence pair	key-sentence key-phrases pair
# of sentence pairs	527,830	28,866	3,956	8,203

This can be of practical use to users who need assistance in patent document analysis, such as a patent attorneys, as the core information contained in the document can be checked easily without a complicated analysis process.

IV. EXPERIMENT SETTINGS

A. DATASET DETAILS

In this study, the patent analysis datasets provided by LG Innotek[2] is used for the key-sentence extraction, key-phrase generation, and post-training tasks to enhance the comprehension of the patent domain knowledge. Table 1 shows the details of the patent datasets provided by LG Innotek; it mainly consists of patent documents with the domains of the camera lens, sensor, and battery. For domain-specialized post-training, key-sentence extraction, and key-phrase generation tasks, we use 28,866 patent documents, 3,956 pairs of patent key-sentences, and 8,203 pairs of key-sentence key-phrases.

In addition to the patent domain-specific corpus provided by LG Innotek, we also utilize the KP20k dataset [40], which consists of training data with 527,830 article key-phrases pairs to further improve the performance of the key-phrase generation model by post-training using this dataset.

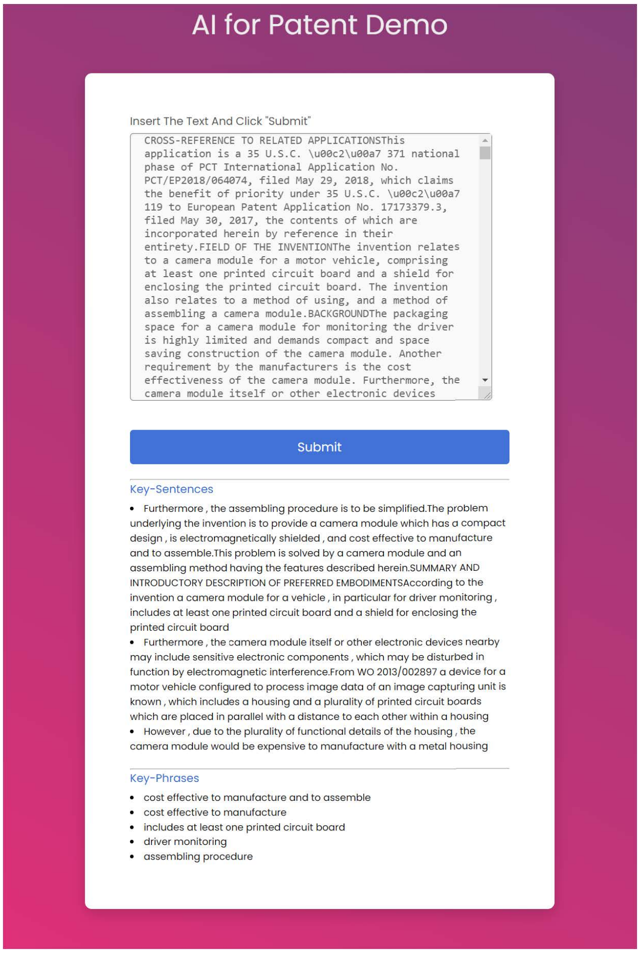
B. DATA PRE-PROCESSING DETAILS

1) KEY-SENTENCE EXTRACTION

We apply the text rank algorithm to the patent document to construct the key-sentence extraction training data efficiently. However, if the document is over-summarized, there may be cases where the reference sentence is not included in the candidate documents. Therefore, as shown in Fig. 4, we verify the minimum ratio by which the reference sentence is not omitted in the summarized candidate documents through an experiment applying the text rank with various thresholds.

The containing ratio denotes the ratio of reference sentences included in the summarized document in the total data when document summarization is applied with a specific threshold.

[2] `https://www.lginnotek.com/`



Insert The Text And Click "Submit"

```
CROSS-REFERENCE TO RELATED APPLICATIONSThis
application is a 35 U.S.C. \u00c2\u00a7 371 national
phase of PCT International Application No.
PCT/EP2018/064074, filed May 29, 2018, which claims
the benefit of priority under 35 U.S.C. \u00c2\u00a7
119 to European Patent Application No. 17173379.3,
filed May 30, 2017, the contents of which are
incorporated herein by reference in their
entirety.FIELD OF THE INVENTIONThe invention relates
to a camera module for a motor vehicle, comprising
at least one printed circuit board and a shield for
enclosing the printed circuit board. The invention
also relates to a method of using, and a method of
assembling a camera module.BACKGROUNDThe packaging
space for a camera module for monitoring the driver
is highly limited and demands compact and space
saving construction of the camera module. Another
requirement by the manufacturers is the cost
effectiveness of the camera module. Furthermore, the
camera module itself or other electronic devices
```

Key-Sentences

- Furthermore , the assembling procedure is to be simplified.The problem underlying the invention is to provide a camera module which has a compact design , is electromagnetically shielded , and cost effective to manufacture and to assemble.This problem is solved by a camera module and an assembling method having the features described herein.SUMMARY AND INTRODUCTORY DESCRIPTION OF PREFERRED EMBODIMENTSAccording to the invention a camera module for a vehicle , in particular for driver monitoring , includes at least one printed circuit board and a shield for enclosing the printed circuit board
- Furthermore , the camera module itself or other electronic devices nearby may include sensitive electronic components , which may be disturbed in function by electromagnetic interference.From WO 2013/002897 a device for a motor vehicle configured to process image data of an image capturing unit is known , which includes a housing and a plurality of printed circuit boards which are placed in parallel with a distance to each other within a housing
- However , due to the plurality of functional details of the housing , the camera module would be expensive to manufacture with a metal housing

Key-Phrases

- cost effective to manufacture and to assemble
- cost effective to manufacture
- includes at least one printed circuit board
- driver monitoring
- assembling procedure

FIGURE 3. Screenshot of our demo system.

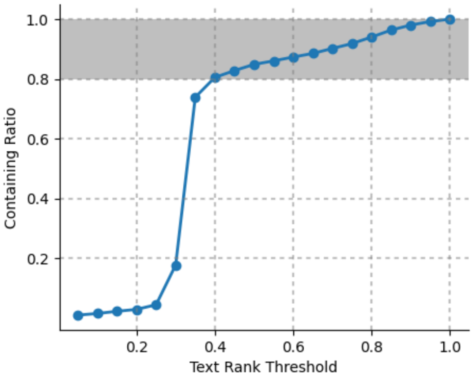


FIGURE 4. Containing ratio of the reference sentences in the summarized document according to text rank threshold. The minimum ratio of not omitting references to 20% or less is approximately 40%.

Through detailed analysis, we empirically confirm that the minimum rate of reducing the omitted case of the reference sentence within the candidate document to 20% or less among all data is approximately 40%. We generate 3,173 summarized data, which reduces the number of sentences in each document by 40% based on the results, and we randomly sample 10% of the total data to construct the validation and test datasets. Table 2 shows the statistics of the configured dataset.

TABLE 2. Statistics of key-sentence extraction dataset.

Dataset		Avg length of sentences	Max length of sentences	Min length of sentences
Train (2,538)	Candidate **w.o.** Text Rank	363.711	3,122	12
	Candidate **w.** Text Rank	144.733	1,248	4
	Reference	1.662	10	1
Valid (317)	Candidate	369.653	5,273	51
	Reference	1.658	7	1
Test (318)	Candidate	376.452	4,867	38
	Reference	1.716	6	1

2) KEY-PHRASE GENERATION

The key-phrase generation dataset consists of 8,203 sentence key-phrases pairs. We exclude the pairs with more than 256 tokens from the pre-processing process for experimental efficiency. We split the dataset 8:1:1 for training, validation, and testing to build 6,562 pairs of training data, 820 pairs of validation data, and 821 pairs of test data, respectively. In particular, we train the key-phrase generation model with one-to-one mapping by dividing each (sentence, n·key-phrases) pair into n·(sentence, single key-phrase) pairs [40] for the training phase. The validation and test data take the original form with a one-to-many structure for evaluation, i.e., n key-phrases are placed for each source input.

C. TRAINING DETAILS

We exploit the T5 model architecture provided by Hugging face, and all training processes are performed using four RTX A6000s.

We post-train the T5 model using span corruption and key-phrase generation tasks for domain-specialized post-training. Especially in the key-phrase post-training, we apply a one-to-one training mechanism that maps a source text to a key-phrase, and we exclude data that have more than 256 tokens to enhance training efficiency [40]. We post-train the model total of 270k steps. In detail, we set the training batch size to 8, targets max length to 256, warm-up ratio to 0.1, and learning rate schedule method to an inverse square root.

For the key-sentence extraction fine-tuning, we only use an encoder of the domain-specialized T5 model. The training details are as follows; training batch size is 32, inputs max length is 512, weight decay is 0.1, and the learning rate is 3e-5. For the key-phrase generation fine-tuning, we employ the full architecture of the domain-specialized T5, which

TABLE 3. Experimental results of key-sentence extraction. Scale {k} indicates the up-scaling ratio of key sentences in the training data. Up-scaling is adopted for the alleviation of class unbalance in the training data.

Model		MRR	Hit @5 Exact	Hit @5 Robust	Hit @5 Rate	Hit @10 Exact	Hit @10 Robust	Hit @10 Rate
T5-patent	No Scale	0.1022	9.12%	13.52%	10.70%	9.43%	14.47%	11.26%
	Scale 10	**0.7080**	71.38%	89.94%	81.02%	83.96%	94.34%	89.44%
	Scale 20	0.6629	70.75%	88.05%	79.70%	83.02%	92.14%	87.98%
	Scale 30	0.6288	**74.53**%	**90.25**%	**82.31**%	87.11%	**94.34**%	90.91%
	Scale 40	0.6402	74.21%	89.62%	82.01%	**87.42**%	**94.34**%	**91.06%**
T5	No Scale	0.0722	5.66%	9.43%	6.89%	6.29%	10.69%	7.76%
	Scale 10	**0.6863**	71.38%	88.36%	80.13%	80.19%	90.88%	85.72%
	Scale 20	0.6671	**73.90**%	**88.99**%	**81.53**%	**85.53**%	**93.71**%	**90.09**%
	Scale 30	0.6417	69.18%	85.53%	77.64%	81.76%	92.77%	87.79%
	Scale 40	0.6177	69.18%	87.42%	78.21%	84.91%	93.40%	89.44%

utilizes the encoder–decoder structure. In detail, we set the training batch size to 32, max length of inputs to 256, max length of targets to 32, weight decay to 0.01, and the learning rate to 0.001.

D. EVALUATION DETAILS

1) KEY-SENTENCE EXTRACTION

We use several evaluation metrics such as the mean reciprocal rank (MRR) [41] to evaluate the performance of the key-sentence extraction model. MRR is a metric that evaluates the confidence in which the model predicts important sentences; it quantifies how important the model predicts the reference sentence to be when the importance of each sentence in the document has a score between zero and one. In this study, as more than one reference sentence can exist in the document, we apply a partially amended metric. This can be represented for the reference sentence set S and document D, as shown in equation 6.

$$\mathrm{MRR} = \frac{1}{|S|} \sum_{(R,D)\in S} \frac{1}{\min \{rank(r|D)\}_{r\in R}} \tag{6}$$

In this equation, $rank(r|D)$ denotes a ranking of the reference sentence r when the sentences in the document are sorted according to the priority order. The higher the importance prediction provided by the model for the reference sentence, the higher the MRR of the model.

To evaluate this model, in addition to MRR, we apply the Hits@k metric. **Hits@k Exact** is the ratio of data that perfectly contains all reference sentences, **Hits@k Robust** is the ratio of data that contains at least one of the reference sentences, and **Hits@k Rate** is the average ratio that contains reference sentences in the top k summary sentences extracted by the model.

2) KEY-PHRASE GENERATION

We evaluate the key-phrase generation model using the F1 score. The F1 score is calculated using equation 7.

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{7}$$

We calculate the F1 score from the following two perspectives. We confirm how well the model generates key-phrases superficially and semantically similar by considering the two perspectives below. First, we use the exact and partial F1 scores to evaluate the superficial matching performance; the exact F1 considers the exact match between prediction and reference as a correct answer. The partial F1 utilizes Jaccard coefficients of the unigram and bigram as the similarity between prediction and reference.

Second, we employ the Bert score [42], which considers semantic similarity using deep contextualized representation from the pre-trained model, such as BERT, rather than the superficial similarity between two key-phrases. In this study, we evaluate the Bert score using RoBERTa$_{Large}$. For more details, given a reference sequence of tokenized tokens $x = \langle x_1, \ldots, x_k \rangle$, the embedding model creates a vector sequence $\langle \mathbf{x}_1, \ldots, \mathbf{x}_k \rangle$. Similarly, a candidate sequence of tokenized token $\hat{x} = \left\langle \hat{x}_1, \ldots, \hat{x}_k \right\rangle$ is mapped to the vector sequence $\left\langle \hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_k \right\rangle$. The Bert score corresponding to the correct answer sequence x and the candidate sequence $\hat{x}$ is calculated by equation 8.

$$\begin{aligned} Recall_{BERT} &= \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} \mathbf{x}_i^T \hat{\mathbf{x}}_j \\ Precision_{BERT} &= \frac{1}{|\hat{x}|} \sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} \mathbf{x}_i^T \hat{\mathbf{x}}_j \\ F1_{BERT} &= 2 * \frac{Precision_{BERT} * Recall_{BERT}}{Precision_{BERT} + Recall_{BERT}} \end{aligned} \tag{8}$$

V. RESULTS AND DISCUSSION

A. KEY-SENTENCE EXTRACTION

Table 3 shows the performance of the key-sentence extraction model proposed in this study. In this experiment, we assess the performance of the model along with different up-scaling ratios to find the optimal scaling ratio for the model training. Up-scaling ratio k indicates that reference sentences in a training data (*i.e.* labeled as 1) is duplicated k times. In addition, to prove the effectiveness of the domain specialized post-training, we compare the performance difference with the model that is not post-trained in the patent domain.

TABLE 4. Experimental results of key-sentence extraction with respect to the text rank. In the case of the model with TextRank, we use the scaling ratio with the best Hit @5 performance in Table 3.

Model		MRR	Hit @5 Exact	Hit @5 Robust	Hit @5 Rate	Hit @10 Exact	Hit @10 Robust	Hit @10 Rate
w. Text Rank	T5-patent	0.6288	**74.53%**	**90.25%**	**82.31%**	**87.11%**	**94.34%**	**90.91%**
	T5	0.6671	73.90%	88.99%	81.53%	85.53%	93.71%	90.09%
w.o. Text Rank	T5-patent	**0.6719**	72.01%	88.36%	80.41%	83.96%	**94.34%**	89.57%
	T5	0.6407	73.58%	88.68%	80.97%	84.59%	93.71%	89.34%

TABLE 5. Experimental results of the key-phrase extraction and generation. The encoder-based extraction method (i.e. BERT) cannot generate the specified number of key-phrases like other approaches. Therefore, we compare only the first and top k cases to ensure a fair experiment.

Approach	Model	F1 @1 Exact	F1 @5 Exact	F1 @k Exact	F1 @1 Partial	F1 @5 Partial	F1 @k Partial	F1 @1 Bert	F1 @5 Bert	F1 @k Bert
Statistical Extraction	TF-IDF [43]	8.06	9.82	9.66	26.98	23.87	31.58	88.56	88.16	89.11
	Text Rank [18]	6.39	10.7	8.44	31.24	29.39	36.75	89.07	86.34	89.61
	RAKE [44]	5.88	11.12	8.05	31.72	28.5	37.98	88.22	86.06	88.96
Encoder-based Extraction	BERT [45]	28.96	-	33.58	46.36	-	52.86	92.56	-	86.8
	SciBERT [46]	29.9	-	34.76	46.18	-	52.37	92.64	-	86.74
Generation	T5	37.86	32.17	41.8	58.28	41.73	65.09	94.79	92.26	94.71
	T5-patent	**60.36**	**45.13**	**67.35**	**72.56**	**49.18**	**81.19**	**97.74**	**93.53**	**97.29**

As shown in Table 3, the overall performance of the model without scaling is relatively low. In this case, the Hit @5 exact is less than 7%, implying that it does not properly grasp the importance of each sentence within the document; moreover, it can be seen that the summary contains only a few reference sentences. This problem can effectively be relieved by applying up-scaling to the reference data. Specifically, when scaling 40 is applied, the Hit@10 rate is 91.09%, implying that the model can generate a summarized document that covers most of the reference sentences. We also find that post-training can consistently enhance the overall performance.

Our experimental results show that generally high up-scaling ratio bears high performance for the hits@k metrics. In terms of the MRR, the model applying scaling 10 shows the best performance; the performance tends to decrease as the up-scaling ratio increases.

This suggests that a proper trade-off should be considered between the exact estimation of the specific reference sentence, and the comprehensive estimation of all the reference sentences. Considering that a single document can have more than one reference sentence, and the priority difference between them is not generally to be taken into account for document analysis, choosing a suitable up-scaling ratio that yields high hits@k metrics may be appropriate.

Additionally, we verify the effectiveness of the text rank algorithm-based training data pre-processing through a comparative analysis between the extraction models trained by the original training data and the pre-processed training data by the text rank algorithm. For training the model, we applied several up-scaling rates that were empirically similar to the previous experiment, and among the trained models, the model with the best performance was selected as the final model. The experimental results are shown in Table 4.

The experimental results show that the model trained with text rank can consistently outperform the model trained with the original data. This shows that the text rank-based pre-processing method applied in this study is practically effective and can be used as an effective strategy to obtain substantial performance improvement while successfully reducing the computing power and GPU resources required for training.

B. KEY-PHRASE GENERATION

To prove the efficiency of our key-phrase generation stage in the framework, we compare two types of F1 scores: superficial (Exact/Partial) and Bert score. The former evaluates surface-level matching performance, including bi-gram similarity; The latter assesses semantic-level matching performance using a pre-trained model's contextualized vector for predicted key-phrases.

As shown in Table 5, encoder-based extraction methods show more outstanding performance than the statistical methods except for F1@k Bert score. This result indicates that although the encoder-based methods can extract a key-phrase with the highest similarity with the context, they struggle to figure out a similar but different form of words. The vanilla T5 model shows the best baseline performances compared to the statistical and encoder-based methods in both superficial and semantic evaluations, showing the highest language comprehension. It is because a training objective employed in pre-training phase of the T5 model tends to perform better in many NLP tasks than a BERT-style objective [16].

TABLE 6. Case study in patent key-sentence extraction. *Rank* refers to the priority rank of reference sentences in a summarized document.

Reference sentences	**Model**	*Rank*
• However, in the prior art, spherical glass lenses are often used which are bulky and costly • Although the use of all-plastic lenses can reduce the volume, however, the thermal expansion coefficient of the plastic material is large, which easily causes the lens to appear focal points offset problem due to temperature changes. In view of the above problems, the disclosure provides an infrared optical imaging lens, having the advantages of miniaturization, low cost and high imaging quality. Embodiments of the disclosure achieve the above object through the following technical solutions. In a first aspect, the disclosure provides an infrared optical imaging lens	T5-patent **w.** text rank	1, 2
	T5 **w.** text rank	3, 4
	T5-patent **w.o.** text rank	2, 7
	T5 **w.o.** text rank	2, 4
• Individual cell voltages in a series battery pack may become unbalanced due to differences in the amount of charge stored on the cells	T5-patent **w.** text rank	1
	T5 **w.** text rank	2
	T5-patent **w.o.** text rank	1
	T5 **w.o.** text rank	3
• This is especially problematic for PKE systems in automotive applications since a larger energy storage capacitor requires a longer time to recharge after discharge, which slows down the ability of the receiving device to process successive incoming signals from the vehicle • It can be readily appreciated that a vehicle designer/engineer does not want an electronic system that experiences an unsatisfactory delay for a car door to unlock or an ignition to fire	T5-patent **w.** text rank	1, 3
	T5 **w.** text rank	1, 9
	T5-patent **w.o.** text rank	4, 7
	T5 **w.o.** text rank	5, 10

In addition, SciBERT, which is pre-trained with a large amount of science-related corpus, shows slightly better performance than vanilla BERT. This result demonstrates the importance of domain adaptation because SciBERT is trained with corpora more related to the patent domain than vanilla BERT. Moreover, the T5 patent model, post-trained with patent-related raw corpus, shows remarkable improvements compared to the vanilla T5 model, demonstrating the best performance in all evaluation metrics, including the exact/partial and bert scores. This result highlights the significance of the domain-adaptive post-training, which makes the model capture technical and legal context in the patent document. In terms of Bert score, which represents semantic matching performance, the T5-patent model shows powerful performance compared to other baselines. Even for Exact and Partial F1 score, which are to evaluate the superficial matching performance, the T5-patent has dominant performances. It is clear that key-phrases generated from our framework are more abundant in terms of semantics and surfaces than those from the existing methods.

C. QUALITATIVE ANALYSIS

1) KEY-SENTENCE EXTRACTION

In this section, we analyze the actual prediction results of our key-sentence extraction model with the various ablations of the model.

As shown in Table 6, the T5-patent model with the text rank grasps the key-sentences as the highest importance compared to other models. We can say that the model with post-training generally has a better understanding than the model without post-training. In other words, these results show that domain-specialized post-training can lead to significant performance improvements. This result is observed both in quantitative analysis and qualitative analysis.

In addition, it can be confirmed that the model trained with the data in which the text rank algorithm is applied extracts the key sentences with better or similar performance than the model trained with the data in which the text rank algorithm is not applied. This result shows that applying the text rank can effectively filter unnecessary noises often included in the description section; it can reduce the training time and ensure efficient utilization of the computing resources, which allows the model to perform better. In other words, we can say that the learning strategies we employ for the key-sentence extraction have a significant effect on the actual performance of the model.

2) KEY-PHRASE GENERATION

We qualitatively analyze the actual results of the key-phrase extraction and generation models such as the SciBERT, T5, and T5-patent model to validate whether our proposed method can more effectively predict than other models.

As shown in Table 7, SciBERT frequently predicts some nouns of the reference as short phrases. In this case, we can observe that SciBERT, which represents the extractive method based on the encoder using the token classification in this study, hard to predict lengthy key-phrases that are otherwise seen as possible in other models. In addition, T5-patent appears to be able to generate semantically similar phrases that consider the entire context, despite some words

TABLE 7. Case study in patent key-phrase extraction/generation for T5-patent, T5, and SciBERT. The highlighted phrases represent the reference key-phrases in the input. In the predicted labels, green fonts represent the exactly matched answers, blue fonts represent partially the matched answers, and pink fonts represent the semantically similar answers.

Input	Model	predicted labels
However, the amplification processing in the signal processing system of the solid-state imaging element has problems of **emphasized noise component**, **emphasized flaw** in the solid-state imaging element, **emphasized fine dust attached** to the solid-state imaging element, **emphasized unevenness** of the solid-state imaging element or optical materials and the like, and these problems lead to **reduction in the yield** of the imaging apparatus.	SciBERT	noise, fine dust, yield
	T5	reduction in the yield, emphasized flaw, emphasized flaw in the solid-state imaging element, reduction in the yield of the imaging apparatus, emphasized fine dust attached to the solid-state imaging element
	T5-patent	emphasized flaw, reduction in the yield, emphasized noise component, emphasized fine dust, emphasized unevenness
However, in the motor with speed reduction mechanism described in Japanese Patent Application Laid-Open Publication No. 2001-294044, the latching claw is disposed on the outside in the diametrical direction than the outer wall surface of the worm wheel housing portion, and the side surfaces and the tip surface of the latching claw are exposed; therefore, **breakage of the latching claw** or **release of the engagement of the latching claw** may be posed, for example, when an object hits the latching claw after assembly of the worm wheel housing portion and the bottom cover.	SciBERT	breakage of the latching claw, release of the engagement of the latching claw
	T5	breakage of the latching claw, release of the latching claw, brokenage of the latching claw, latching claw, release of the engagement
	T5-patent	breakage of the latching claw, breakage, brokenage of the latching claw, release of the engagement, cutage of the latching claw
That is, when the electrode layer is formed using a binder or thickener composed of an organic polymer compound, it has merits that the organic polymer compound can secure **the viscosity of the slurry**, **facilitate coating of a thick film**, and **improve the adhesion** between the electrode active material, the solid electrolyte, other components, and the collector, but on the other hand, it has a demerit that **the resistance** increases	SciBERT	secure the viscosity of the slurry, facilitate coating of a thick film, improve the adhesion, the resistance
	T5	secure the viscosity of the slurry, improve the adhesion, the resistance, the resistance increases, facilitate coating of a thick film
	T5-patent	the viscosity of the slurry, improve the adhesion, secure the viscosity of the slurry, facilitate coating of a thick film, enhance the viscosity of the slurry

not being in the query sentence. For instance, it generated 'brokenage of the latching claw' and 'outage of the latching claw', which is similar to the reference key-phrase 'breakage of the latching claw', and 'enhance the viscosity of the slurry', which is similar to the reference 'secure the viscosity of the slurry'. We can therefore say that the T5-patent can paraphrase some keywords in a phrase with the high comprehensive capacity of the patent domain, which cannot be extracted using the extractive method. Moreover, the T5-patent predicted key-phrases with better accuracy as compared to T5, which proves the effectiveness of the domain-specialized post-training.

VI. CONCLUSION

In this study, a novel multi-stage framework for patent document analysis, consisting of a key-sentence extraction and a key-phrase generation model, was proposed. In particular, in training the key-sentence extraction model, it was confirmed that applying the text rank algorithm as a pre-processing not only increased the learning efficiency but also showed excellent quantitative performance. In addition, it was verified through quantitative and qualitative analysis that the generation method was more effective than the extractive method in extracting key-phrases. We developed a framework by combining models, and implemented it in the form of a demo system for practical use in the patent field. As a future study, we plan to improve the model created through T5 and develop a lightweight framework that allows us to work with smaller-sized models.

ACKNOWLEDGMENT

(Junyoung Son and Hyeonseok Moon contributed equally to this work.)

REFERENCES

[1] H. Lin, H. Wang, D. Du, H. Wu, B. Chang, and E. Chen, "Patent quality valuation with deep learning models," in *Proc. Int. Conf. Database Syst. Adv. Appl.* Cham, Switzerland: Springer, 2018, pp. 474–490.

[2] J. Hu, S. Li, Y. Yao, L. Yu, G. Yang, and J. Hu, "Patent keyword extraction algorithm based on distributed representation for patent classification," *Entropy*, vol. 20, no. 2, p. 104, Feb. 2018.

[3] L. Zhang, L. Li, and T. Li, "Patent mining: A survey," *ACM SIGKDD Explor. Newslett.*, vol. 16, no. 2, pp. 1–19, May 2015.

[4] A. Abbas, L. Zhang, and S. U. Khan, "A literature review on the state-of-the-art in patent analysis," *World Pat. Inf.*, vol. 37, pp. 3–13, Jun. 2014.

[5] D. Bonino, A. Ciaramella, and F. Corno, "Review of the state-of-the-art in patent information and forthcoming evolutions in intelligent patent informatics," *World Pat. Inf.*, vol. 32, no. 1, pp. 30–38, Mar. 2010. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0172219009000465

[6] X. Xue and W. B. Croft, "Automatic query generation for patent search," in *Proc. 18th ACM Conf. Inf. Knowl. Manage. (CIKM)*, 2009, pp. 2037–2040.

[7] Y.-H. Tseng, C.-J. Lin, and Y.-I. Lin, "Text mining techniques for patent analysis," *Inf. Process. Manage.*, vol. 43, no. 5, pp. 1216–1247, Sep. 2007.

[8] S.-H. Liu, H.-L. Liao, S.-M. Pi, and J.-W. Hu, "Development of a patent retrieval and analysis platform—A hybrid approach," *Expert Syst. Appl.*, vol. 38, no. 6, pp. 7864–7868, Jun. 2011.

[9] R. Setchi, I. Spasić, J. Morgan, C. Harrison, and R. Corken, "Artificial intelligence for patent prior art searching," *World Pat. Inf.*, vol. 64, Mar. 2021, Art. no. 102021.

[10] R. Krestel, R. Chikkamath, C. Hewel, and J. Risch, "A survey on deep learning for patent analysis," *World Pat. Inf.*, vol. 65, Jun. 2021, Art. no. 102035.

[11] C. Park, K. Park, H. Moon, S. Eo, and H. Lim, "A study on performance improvement considering the balance between corpus in neural machine translation," *J. Korea Converg. Soc.*, vol. 12, no. 5, pp. 23–29, 2021.

[12] E. D'hondt, S. Verberne, C. Koster, and L. Boves, "Text representations for patent classification," *Comput. Linguistics*, vol. 39, no. 3, pp. 755–775, Sep. 2013.

[13] W. Shalaby and W. Zadrozny, "Toward an interactive patent retrieval framework based on distributed representations," in *Proc. 41st Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jun. 2018, pp. 957–960.

[14] L. Chen, S. Xu, L. Zhu, J. Zhang, X. Lei, and G. Yang, "A deep learning based method for extracting semantic information from patent documents," *Scientometrics*, vol. 125, no. 1, pp. 289–312, Oct. 2020.

[15] J. Risch and R. Krestel, "Learning patent speak: Investigating domain-specific word embeddings," in *Proc. 13th Int. Conf. Digit. Inf. Manage. (ICDIM)*, Sep. 2018, pp. 63–68.

[16] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," 2019, *arXiv:1910.10683*.

[17] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith, "Don't stop pretraining: Adapt language models to domains and tasks," 2020, *arXiv:2004.10964*.

[18] R. Mihalcea and P. Tarau, "TextRank: Bringing order into text," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2004, pp. 404–411.

[19] M. Rajangam and C. Annamalai, "Extractive document summarization using an adaptive, knowledge based cognitive model," *Cogn. Syst. Res.*, vol. 56, pp. 56–71, Aug. 2019.

[20] C. Park, G. Kim, and H. Lim, "Parallel corpus filtering and Korean-optimized subword tokenization for machine translation," in *Proc. 31st Annu. Conf. Hum. Lang. Technol.*, Busan, South Korea, 2019, pp. 11–12.

[21] C. Park, Y. Lee, C. Lee, and H. Lim, "Quality, not quantity?: Effect of parallel corpus quantity and quality on neural machine translation," in *Proc. 32st Annu. Conf. Hum. Lang. Technol.*, 2020, pp. 363–368.

[22] C. Park, J. Seo, S. Lee, C. Lee, H. Moon, S. Eo, and H. Lim, "BTS: Back TranScription for speech-to-text post-processor using text-to-speech-to-text," in *Proc. 8th Workshop Asian Transl. (WAT)*, 2021, pp. 106–116.

[23] H. Noh, Y. Jo, and S. Lee, "Keyword selection and processing strategy for applying text mining to patent analysis," *Expert Syst. Appl.*, vol. 42, no. 9, pp. 4348–4360, Jun. 2015.

[24] T. Roh, Y. Jeong, and B. Yoon, "Developing a methodology of structuring and layering technological information in patent documents through natural language processing," *Sustainability*, vol. 9, no. 11, p. 2117, Nov. 2017.

[25] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, *arXiv:1301.3781*.

[26] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Trans. Assoc. Comput. Linguistics*, vol. 5, pp. 135–146, Dec. 2017.

[27] M. Habibi, A. Rheinlaender, W. Thielemann, R. Adams, P. Fischer, S. Krolkiewicz, D. L. Wiegandt, and U. Leser, "PatSeg: A sequential patent segmentation approach," *Big Data Res.*, vols. 19–20, Mar. 2020, Art. no. 100133.

[28] D. M. Korobkin, S. A. Fomenkov, and A. G. Kravets, "Methods for extracting the descriptions of sci-tech effects and morphological features of technical systems from patents," in *Proc. 9th Int. Conf. Inf., Intell., Syst. Appl. (IISA)*, Jul. 2018, pp. 1–4.

[29] N. Borodin, D. Korobkin, A. Bezruchenko, and S. Fomenkov, "The search for R&D partners based on patent data," *J. Phys., Conf. Ser.*, vol. 2060, no. 1, Oct. 2021, Art. no. 012022.

[30] A. Krishna, Y. Jin, C. Foster, G. Gabel, B. Hanley, and A. Youssef, "Query expansion for patent searching using word embedding and professional crowdsourcing," 2019, *arXiv:1911.11069*.

[31] D. Korobkin, S. Fomenkov, A. Kravets, S. Kolesnikov, and M. Dykov, "Three-steps methodology for patents prior-art retrieval and structured physical knowledge extracting," in *Creativity in Intelligent Technologies and Data Science*, vol. 535, 2015, pp. 124–136.

[32] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proc. 25th Int. Conf. Mach. Learn. (ICML)*, 2008, pp. 160–167.

[33] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 3104–3112.

[34] T. Whang, D. Lee, C. Lee, K. Yang, D. Oh, and H. Lim, "An effective domain adaptive post-training method for BERT in response selection," 2019, *arXiv:1908.04812*.

[35] H. Xu, B. Liu, L. Shu, and P. S. Yu, "BERT post-training for review reading comprehension and aspect-based sentiment analysis," 2019, *arXiv:1904.02232*.

[36] R. Nallapati, F. Zhai, and B. Zhou, "SummaRuNNer: A recurrent neural network based sequence model for extractive summarization of documents," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 1–7.

[37] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," 2020, *arXiv:2004.05150*.

[38] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[39] R. Dewi, W. Bijker, A. Stein, and M. Marfai, "Transferability and upscaling of fuzzy classification for shoreline change over 30 years," *Remote Sens.*, vol. 10, no. 9, p. 1377, Aug. 2018.

[40] R. Meng, S. Zhao, S. Han, D. He, P. Brusilovsky, and Y. Chi, "Deep keyphrase generation," 2017, *arXiv:1704.06879*.

[41] N. Craswell, *Mean Reciprocal Rank*. Boston, MA, USA: Springer, 2009, p. 1703.

[42] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating text generation with BERT," 2019, *arXiv:1904.09675*.

[43] G. G. Chowdhury, *Introduction to Modern Information Retrieval*. London, U.K.: Facet publishing, 2010.

[44] S. Rose, D. Engel, N. Cramer, and W. Cowley, "Automatic keyword extraction from individual documents," *Text Mining, Appl. Theory*, vol. 1, pp. 1–20, Mar. 2010.

[45] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[46] I. Beltagy, K. Lo, and A. Cohan, "SciBERT: A pretrained language model for scientific text," 2019, *arXiv:1903.10676*.

JUNYOUNG SON received the B.S. degree from the Department of Information and Communications Technology, Dongguk University, Seoul, South Korea, in 2021. He is currently pursuing the Integrated master's and Ph.D. degree with the Natural Language Processing and Artificial Intelligence Laboratory. His research interest includes natural language processing, specifically information retrieval.

HYEONSEOK MOON received the B.S. degree from the Department of Science in Mathematics and Engineering, Korea University, Seoul, South Korea, in 2021, where he is currently pursuing the Integrated master's and Ph.D. degree in computer science and engineering, a part of the Natural Language Processing and Artificial Intelligence Laboratory. His research interests include natural language processing, neural machine translation, automatic post-editing, and parallel corpus filtering.

JEONGWOO LEE received the B.S. degree in information and communication engineering from Yeungnam University, Gyeongsan, South Korea, in 2020. He is currently working as a Research Scientist at the Korea University Human-Inspired AI Research Institute. His research interests include natural language processing, conversational AI, dialogue agents, question answering, and deep learning.

SEOLHWA LEE received the Ph.D. degree in computer science and engineering from Korea University, Seoul, South Korea, in August 2021. She is currently a Postdoctoral Fellow at the University of Copenhagen. Her research interests include intersection of human cognitive and natural language processing, dialogue summarization, dialogue agents, text summarization, and deep learning.

CHANJUN PARK received the B.S. degree in natural language processing and creative convergence from the Busan University of Foreign Studies, Busan, South Korea, in 2019. He is currently pursuing the Ph.D. degree with the Department of Computer Science and Engineering, Korea University, Seoul, South Korea. From June 2018 to July 2019, he worked at SYSTRAN as a Research Engineer. He is currently working as an AI Research Engineer at Upstage. His research interests include machine translation, grammar error correction, simultaneous speech translation, and deep learning.

WONKYUNG JUNG received the M.S. degree in electronic engineering from Sogang University, Seoul, South Korea, in 2019. She is currently working at LG Innotek as a Software Engineer. Her research interests include computer vision, natural language processing, and deep learning.

HEUISEOK LIM received the B.S., M.S., and Ph.D. degrees in computer science and engineering from Korea University, Seoul, South Korea, in 1992, 1994, and 1997, respectively. He is currently a Professor with the Department of Computer Science and Engineering, Korea University. His research interests include natural language processing, machine learning, and artificial intelligence.