Analyzing the Structure of U.S. Patents Using Patent Families

Jun Nakamitsu
Graduate School of Science and Engineering
Chuo University
Tokyo, Japan
a17.8sca@g.chuo-u.ac.jp

Satoshi Fukuda
Faculty of Science and Engineering
Chuo University
Tokyo, Japan
fukuda.satoshi.3238@kc.chuo-u.ac.jp

Hidetsugu Nanba
Faculty of Science and Engineering
Chuo University
Tokyo, Japan
nanba@kc.chuo-u.ac.jp

***Abstract*—Researchers and developers search for patents in fields related to their own research to obtain information on issues and effective technologies in those fields for use in their research. However, it is impossible to read through the full text of many patents, so a method that enables patent information to be grasped briefly is needed. In this study, we analyze the structure of U.S. patents with the aim of extracting important information. Using Japanese patents with structural tags such as "field", "problem", "solution", and "effect", and corresponding U.S. patents (patent families), we automatically created a dataset of 81,405 U.S. patents with structural tags. Furthermore, using this dataset, we conduct an experiment to assign structural tags to each sentence in the U.S. patents automatically. For the embedding layer, we use a language representation model, Bidirectional Encoder Representations from Transformer, pretrained on patent documents and construct a multi-label classifier that classifies a given sentence into one of four categories: "field", "problem", "solution", or "effect". Using a loss function that considers the unbalanced amount of data for each structural tag, we are able to classify sentences related to "field", "problem", "solution", and "effect" with precision of 0.6994, recall of 0.8291, and F-measure of 0.7426.**

Keywords—patent, document structure analysis, machine translation, machine learning

I. Introduction

When researchers and company engineers consider new research or development, utilizing patent information is important for grasping the latest technical trends. On the other hand, it is difficult to read through all the patents published around the world. Under such circumstances, a method that enables an efficient overview of technical trends is needed. To overview technical trends, it is effective to classify patents according to the viewpoints of technologies and problems, etc. However, to do so, it is necessary to extract the description part of technologies and problems from each patent. Therefore, this study aims to analyze the structure of U.S. patents.

Unlike U.S. patents, Japanese patents have explicit items such as "Field of Technology" (hereinafter referred to as "field"), "Problem to Be Solved by the Invention" (hereinafter referred to as "problem"), "Solution for Solving the Problem" (hereinafter referred to as "solution"), and "Effect of the Invention" (hereinafter referred to as "effect"). As a result, researchers and company engineers have to spend more time reading U.S. patents. Therefore, in this study, we perform a structural analysis of U.S. patents and automatically extract sentences that provide clues for classification.

To achieve this, we analyze the structure of U.S. patents by using patent families. We first find from U.S. patents the bilingual sentences described in the "field", "problem", "solution", and "effect" sections in Japanese patents. Then, we construct a U.S. patent dataset with a clear structure by assigning the same structural tags as those of the Japanese patent to the found sentences. Finally, by applying machine learning using the created dataset, we construct a system that can automatically extract sentences related to "field", "problem", "solution", and "effect", even for U.S. patents that do not have patent families.

By extracting key sentences about technology trends from U.S. patents, we can obtain clues for clustering each patent. This also allows researchers and company engineers to understand the current technology of their competitors and to conduct efficient research and development that meets global demand.

II. Related Work

A. Structural Analysis of Technical Documents

Although it is important to analyze technical documents such as patents and research papers, it is difficult to analyze the entire text. In such cases, it is useful to clarify initially the structure of the document to narrow down the target sentences before conducting the analysis. In this section, we introduce a study that used machine learning to analyze the structure of technical documents.

Prabhakaran et al. [1] analyzed the structure of abstracts with the aim of predicting the growth and decline of scientific topics. They constructed a classifier that applies seven different labels to sentences ("background", "objective", "data", "design", "method", "result," and "conclusion") using manually labeled abstracts, in which the authors assigned labels to sentences, as training data. They applied Conditional Random Field to these data and parsed approximately 2.4 million abstracts to investigate the relationship between labels and topics. The results showed that the technologies discussed in "conclusion" sentences tended to decline, while the technologies discussed in "method" sentences were in the early stage of growth.

Li et al. [2] assumed that evidence plays an important role in biomedical research and extracted evidential descriptions of the figures and tables from biomedical articles. They constructed a model consisting of embedding, attention, and tagging layers. For embedding, they used BioGloVe [3], BioBERT [4], and SciBERT [5], which were pretrained on biomedical texts, and Recurrent Neural Network (RNN) and Long Short Term Memory (LSTM) for attention. Experiments were conducted on

two datasets, PubMed-RCT [6] and SciDT [7], and the models using SciBERT and LSTM performed the best.

Given this background, the purpose of this study is to analyze the structure of U.S. patents with the aim of classifying patents. To achieve this, we analyze the structure of U.S. patents by assigning four types of structural tags, "field", "problem", "solution", and "effect", to each sentence in a U.S. patent. We then construct training data with the Japanese–U.S. patent family and use Bidirectional Encoder Representations from Transformer (BERT) [8] for embedding.

B. Extracting Information in Units Shorter than a Sentence

Techniques for extracting important information from units shorter than a sentence are said to be useful for document retrieval and trend analysis. In the patent classification, which is the purpose of this study, it is necessary to extract the names of classification axes from important sentences in patents, so this research is very relevant for the future.

Gupta et al. [9] proposed a method for automatically extracting words belonging to three categories, "focus", "technique", and "domain", from paper abstracts by pattern matching. In this method, for example, the object appearing immediately after the verb "propose" is considered to belong to the "focus" category.

Heffernan et al. [10] created a classifier to determine whether a phrase in a paper indicates a problem or a solution to improve the effectiveness of technical document retrieval and compare similar papers. They first created a list of phrases similar to "problem" and "solution" and collected sentences containing these phrases from papers. The collected sentences were then parsed, and phrases that could be said to indicate a problem or solution were selected as positive samples. The phrases that had the same sentence structure as the positive sample but did not indicate a problem or solution were considered as the negative sample. Using the collected samples as training data, three types of classifiers were trained: Naive Bayes, Logistic Regression, and SVM. The experimental results showed that the SVM-based classifier had the highest accuracy.

Weston et al. [11] constructed a system to analyze sentences in materials science articles with the aim of assisting in obtaining information about materials science. First, they manually assigned seven types of tags ("material", "phase", "sample descriptor", "property", "application", "synthesis method", and "characterization") to each word in the paper's abstract. Then, they used those data as training data for information extraction by LSTM.

In this study, sentences with structural tags are extracted from patents. In the future, it will be necessary to determine the axis of technology analysis using information extraction methods based on units shorter than sentences, such as those presented in this section.

C. Translation of Patent Documents

In this study, we create a dataset using Japanese and U.S. patent families. In the process of creating the dataset, translation from Japanese to English is required. In this section, we introduce efforts required for the translation of patent documents.

One of the tasks in NTCIR-10 [12] is the Patent Machine Translation Task. This task provides a large test collection containing training, development, and test data for Chinese/English and Japanese/English patent machine translation. The collection contains a bilingual Japanese–English patent translation corpus of about 3.2 million pairs. We constructed a Japanese–English machine translation system based on state-of-the-art Transformer architecture [13].

III. Analyzing the Structure of U.S. Patents

A. Analyzing the Structure of U.S. Patents Using Patent Families

As described in Section I, U.S. patents do not explicitly include items describing "field", "problem", "solution", and "effect". Therefore, we propose a method to construct a structural analysis system of U.S. patents by creating a dataset with structural tags using patent families and performing machine learning using the dataset as training data.

Generally, patent rights are granted independently in each country. To obtain patent rights in each country, the applicant needs to apply for patents for the same invention in several countries. Such a group of patent documents with the same content is called a patent family. Although the language and structure of patents in the patent family differ, the texts that compose the application documents closely correspond to each other. In this study, we analyze the structure of U.S. patents by assigning the same structural tags to sentences in U.S. patents that share the same meaning as sentences in Japanese patents to which the structural tags were manually assigned. Then, using the structured U.S. patents as training data, we conduct machine learning to construct a structural analysis system for U.S. patents.

The procedure for the structural analysis of U.S. patents is shown below.

1) Translate each sentence with a structure tag in the Japanese patent into English.
2) Represent all the sentences in the U.S. patent and the translated sentences in *1)* as vectors. The sentence in the U.S. patent that has the highest cosine similarity with the translated sentence in *1)* is assigned the same structure tag as that of the Japanese patent.
3) Using the data in *2)* as training data, perform machine learning to extract automatically sentences related to "field", "problem", "solution", and "effect" from U.S. patents.

Steps *1)* and *2)* are described in detail in Section III.*B*, and step *3)* is described in Section III.*C*.

B. Dataset Creation

First, a machine translator built using the NTCIR-10 Patent Machine Translation Test Collection and the sequence modeling tool FAIRSEQ [14] was used to translate sentences with structural tags in Japanese patents into English. FAIRSEQ is a tool that can be used to train text generation models such as machine translation. The translator built in this experiment uses Transformer and achieved a BLEU score of 44.11.

Next, we extracted sentences from the U.S. patent that share the same meaning as the translated sentences. First, the

Identify applicable funding agency here. If none, delete this text box.

sentences in the translation result and the full text of the U.S. patent are vectorized using PatentSBERTa [15], and then the cosine similarity between the sentences in the translation result and the full text of the U.S. patent, which is a patent family of the Japanese patent, is calculated. The sentence in the U.S. patent with the highest score is considered to have the same meaning as that in the Japanese patent, and is therefore assigned the same structure tag as the Japanese patent. Using this method, a dataset of U.S. patents with structural tags was created. Since the patent families were faithfully translated, the sentence with the highest score was expected to be almost a bilingual sentence. Note that a sentence may be assigned more than one structure tag because it may have the highest score for more than one resulting translation. In addition, not all structure tags are necessarily present in Japanese patents. For example, some Japanese patents do not include "effect".

Steps *1)* and *2)* were used to assign structure tags to 81,405 U.S. patents that had Japanese patents as their patent families. Of the total 22,016,132 sentences, 1,366,165 sentences were assigned at least one of the four types of structure tags. In this experiment, we did not use sentences that had not been assigned any structure tags; we only classified sentences that had been assigned one or more structure tags. If the classifier is used for the full text of patents, there is a high possibility that structure tags will be assigned to ordinary sentences as well, but considering that patent classification will be performed in the future, we believe that the presence of some noise is not a significant problem.

C. Analyzing the Structure of U.S. Patents by Machine Learning

Machine learning is performed using the dataset created in the previous section as training data. We used BERT to assign structural tags to sentences. BERT is a Transformer-based pretrained model that can be applied to any task in natural language processing. By fine-tuning BERT to the structural tag classification task, we built multi-label classifiers that automatically classified a given sentence into one of four categories: "field", "problem", "solution", or "effect".

We pretrained the BERT model using 3.5 million sentences in the detailed description sections and claims of U.S. patents. We then constructed a classifier using the patent-specific BERT based on the pretrained BERT model. Due to the unbalanced number of data for each structural tag in the dataset, we also tested undersampling and weighted loss function methods.

IV. EXPERIMENT

To confirm the validity of the structural analysis methods for U.S. patents proposed in Section III, we conducted experiments under various conditions.

A. Experimental Setup

Experimental Data

Of the 1,366,165 sentences that were automatically assigned structural tags according to Section III.*B*, 60% were used for training, 20% for validation, and 20% for testing. Table I shows a breakdown of the structural tags assigned.

TABLE I. BREAKDOWN OF STRUCTURE-TAGGED SENTENCES

	Number of Sentences			
	Training	***Validation***	***Testing***	***Total***
Field	56,486	18,969	18,727	94,182
Problem	243,606	81,320	81,336	406,262
Solution	464,662	155,017	154,825	774,504
Effect	106,648	35,324	35,706	177,678

Method

We examined the following three methods: patent-specific BERT, undersampling, and weighted loss function. Furthermore, to confirm the effectiveness of our methods, we compared them with a classifier using BERT-base-uncased, which is a standard BERT model for English texts. For all classifiers, the sigmoid function and binary cross-entropy loss were used to calculate the loss. The training parameters were: maximum number of tokens = 128, batch size = 256, and number of epochs = 10. The details of each method are described below.

- **Patent-specific BERT** (our method): BERT was pretrained using 3.5 million sentences from U.S. patents. The classifiers were trained using the dataset in Table I. The learning rate was 1e-6.
- **Undersampling** (our method): Due to the disproportionate number of data for each structural tag, we undersampled the training data. We matched the number of sentences in each structural tag to the number of training sentences in the "field" with the lowest number of sentences (56,486). The validation and testing numbers are shown in Table I. Patent-specific BERT was used for embedding. The learning rate was 1e-10.
- **Weighted loss function** (our method): The dataset had an unbalanced number of positive and negative examples for each structural tag. Therefore, we weighted the losses when calculating them. We increased the weight of positive examples by multiplying the loss by the ratio of negative to positive examples. Patent-specific BERT was used for embedding. The classifiers were trained using the dataset in Table I. The learning rate was 1e-6.
- **BERT-base-uncased** (baseline method): As a baseline method, we used BERT-base-uncased instead of patent-specific BERT.

B. Results

We evaluated the classification of each structural tag. As shown in Table II, the results of the experiment showed that one of our methods, "patent-specific BERT", obtained an F-measure of 0.7426, which outperformed the others.

C. Discussion

In this experiment, we estimated bilingual sentences based on the assumption that sentences in a patent family are translated in a one-to-one relationship. However, there are cases in which a single sentence in a Japanese patent is divided into two sentences in a U.S. patent. Conversely, two sentences in a Japanese patent can be combined into a single sentence in a U.S.

TABLE II. CLASSIFICATION RESULTS

		Precision	*Recall*	*F-measure*
Patent-specific BERT	Field	0.9457	0.8494	**0.8920**
	Problem	**0.8654**	0.7548	0.8052
	Solution	0.8297	**0.8877**	**0.8573**
	Effect	0.6757	0.3056	**0.4158**
	Average	0.8291	0.6994	**0.7426**
Under-sampling	Field	0.0838	**0.9084**	0.1528
	Problem	0.2976	**0.9962**	0.4575
	Solution	0.5603	0.7666	0.6467
	Effect	0.1308	**0.9998**	0.2307
	Average	0.2681	**0.9178**	0.3719
Weighted loss function	Field	0.6506	0.9378	0.7636
	Problem	0.7644	0.8341	0.7967
	Solution	0.8517	0.8359	0.8433
	Effect	0.3387	0.7276	0.4601
	Average	0.6514	0.8339	0.7159
BERT-base-uncased (baseline)	Field	**0.9525**	0.8434	0.8915
	Problem	0.8464	0.7714	**0.8061**
	Solution	**0.8346**	0.8758	0.8543
	Effect	**0.7045**	0.2766	0.3921
	Average	**0.8345**	0.6918	0.7360

patent. Although the method used in this experiment produced highly accurate translations, we believe that some issues need to be addressed in the subsequent discovery of bilingual sentences. In the future, it will be necessary to consider the possibility of determining multiple bilingual sentences depending on the similarity score.

The undersampling method did not provide sufficient training because the number of sentences used for training was greatly reduced. This resulted in the assignment of structural tags to the majority of sentences, which led to a high recall but extremely low precision.

In this experiment, sentences without any structural tags were not used as training data. The reason for this is that untagged sentences are far more numerous than tagged sentences. We believe that eliminating untagged sentences would increase the likelihood that structural tags would be assigned to sentences that should not be assigned structural tags. It is therefore necessary to investigate the accuracy of classification for sentences that should not be assigned structural tags in the future.

V. CONCLUSION

In this study, we created a dataset containing 81,405 U.S. patents with structural tags. Using this dataset, we conducted an experiment to classify automatically sentences from the U.S. patents that are important for classifying each patent onto a technical analysis axis. The experimental results showed that one of our methods, patent-specific BERT, obtained an F-measure of 0.7426, which outperformed the others.

REFERENCES

[1] Vinodkumar Prabhakaran, William L. Hamilton, Dan McFarland, and Dan Jurafsky, "Predicting the rise and fall of scientific topics from trends in their rhetorical framing," In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pp.1170–1180, 2016.

[2] Xiangci Li, Gully Burns, and Nanyun Peng, "Scientific discourse tagging for evidence extraction," In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pp.2550–2562, 2021.

[3] Gully A Burns, Xiangci Li, and Nanyun Peng, "Building deep learning models for evidence classification from the open access biomedical literature," Database, 2019.

[4] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," Bioinformatics, pp.1234–1240, 2020.

[5] Iz Beltagy, Arman Cohan, and Kyle Lo, "SciBERT: pretrained contextualized embeddings for scientific text," In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pp.3615–3620, 2019.

[6] Franck Dernoncourt and Ji Young Lee, "Pubmed 200k RCT: a dataset for sequential sentence classification in medical abstracts," In Proceedings of the 8th International Joint Conference on Natural Language Processing, pp.308–313, 2017.

[7] Gully APC Burns, Pradeep Dasigi, Anita de Waard, and Eduard H Hovy, "Automated detection of discourse segment and experimental types from the text of cancer pathway results sections," Database, 2016.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2017.

[9] Sonal Gupta and Christopher D. Manning, "Analyzing the dynamics of research by extracting key aspects of scientific papers," In Proceedings of the 5th International Joint Conference on Natural Language Processing, pp.1–9, 2011.

[10] Kevin Heffernan and Simone Teufel, "Identifying problems and solutions in scientific text," Scientometrics, pp.1367–1382, 2018.

[11] Leigh Weston, Vahe Tshitoyan, John M. Dagdelen, Olga Kononova, Kristin Aslaug Persson, Gerbrand Ceder, and Anubhav Jain, "Named entity recognition and normalization applied to large-scale information extraction from the materials science literature," Chemical Information and Modeling, pp.3692–3702, 2019.

[12] Isao Goto, Ka Po Chow, Bin Lu, Eiichiro Sumita, and Benjamin K. Tsou, "Overview of the patent machine translation task at the NTCIR-10 workshop," In Proceedings of the 10th NTCIR Conference, pp.260–286, 2013.

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is all you need," In Proceedings of Advances in Neural Information Processing Systems, pp.6000–6010, 2017.

[14] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli, "FAIRSEQ: a fast, extensible toolkit for sequence modeling," In Proceedings of North American Association for Computational Linguistics (NAACL): System Demonstrations, pp.48–53, 2019.

[15] Hamid Bekamiri, Daniel S. Hain, and Roman Jurowetzki, "PatentSBERTa: a deep NLP based hybrid model for patent distance and classification using augmented SBERT," arXiv preprint arXiv:2103.11933, 2021.