sustainability

MDPI

Article

Layout Aware Semantic Element Extraction for Sustainable Science & Technology Decision Support

Hyuntae Kim, Jongyun Choi, Soyoung Park and Yuchul Jung *

Department of Computer Engineering, Kumoh National Institute of Technology, Gumi 39177, Korea; 20216035@kumoh.ac.kr (H.K.); 20216093@kumoh.ac.kr (J.C.); haluna8836@kumoh.ac.kr (S.P.)
* Correspondence: jyc@kumoh.ac.kr; Tel.: +82-54-478-7536

Abstract: New scientific and technological (S&T) knowledge is being introduced rapidly, and hence, analysis efforts to understand and analyze new published S&T documents are increasing daily. Automated text mining and vision recognition techniques alleviate the burden somewhat, but the various document layout formats and knowledge content granularities across the S&T field make it challenging. Therefore, this paper proposes LA-SEE (LAME and Vi-SEE), a knowledge graph construction framework that simultaneously extracts meta-information and useful image objects from S&T documents in various layout formats. We adopt Layout-aware Metadata Extraction (LAME), which can accurately extract metadata from various layout formats, and implement a transformer-based instance segmentation (i.e., Vision based Semantic Elements Extraction (Vi-SEE)) to maximize the vision-based semantic element recognition. Moreover, to constructing a scientific knowledge graph consisting of multiple S&T documents, we newly defined an extensible Semantic Elements Knowledge Graph (SEKG) structure. For now, we succeeded in extracting about 6 million semantic elements from 49,649 PDFs. In addition, to illustrate the potential power of our SEKG, we provide two promising application scenarios, such as a scientific knowledge guide across multiple S&T documents and questions and answering over scientific tables.

Keywords: multi-modal; document layout analysis; metadata; document structure; document object; semantic elements; knowledge graph; transformer; decision support

Citation: Kim, H.; Choi, J.; Park, S.; Jung, Y. Layout Aware Semantic Element Extraction for Sustainable Science & Technology Decision Support. *Sustainability* **2022**, *14*, 2802. https://doi.org/10.3390/su14052802

Academic Editor: Hamid Khayyam

Received: 13 January 2022
Accepted: 22 February 2022
Published: 28 February 2022

1. Introduction

Decision support systems or specific methods for science and technology (S&T) problems or social issues can be employed effectively across various domain user types related to policymaking, research topic search, research method survey, comparing experimental results, emerging technology trend analyses, etc.

Junior researchers (or novice users) may have difficulty collecting target information due to lacking domain knowledge. However, even domain experts usually feel burdened considering the vast and rapidly growing body of scientific literature, expert blogs, commercial technical reports, and patents. Search engines are common tools for information seeking, allowing users to access related documents or paragraphs containing search queries on the premise that full document texts have been indexed. An alternative method to find relevant information for research topic or job is to visit online Q&A communities, such as Knowledge iN (of Naver), Reddit, and/or Quora. However, although most deliver substantial information, they can sometimes contain prejudiced opinions or commercial references that are untrustworthy.

Suppose these S&T documents were well separated and re-organized as reusable knowledge. Then, users could selectively access only relevant knowledge and utilize it in decision-making processes. Unfortunately, although the requirement is becoming critical, few decision support systems are available due to many technical implementation

limitations. Therefore, in order to resolve these limitations, this study aims to enable a sophisticated decision support system by extracting semantic elements from S&T documents and constructing a knowledge graph with the semantic elements.

Knowledge graphs (KGs) are promising enablers for effective decision support systems [1]. Most KGs comprise large quantities of triple sentences, representing vast knowledge, but building a well-equipped KG is challenging due to technological limitations and expensive human evaluation. The KG mainly extracts triples by implementing domain-specific entity name recognizers and relationship extractors to extract concepts (or entities) and identify relevant relationships between the concepts. Internal KG structures and hence their construction performance differ significantly depending on target entity types and relationship granularities [2–5]. For example, PubMed KG [6] connects bio-entities, authors, articles, affiliations, and funding from approximately 29 million PubMed abstracts. Although BioBERT [7] based entity extraction outperformed the previous state-of-the-art (SOTA) models, it only achieved F1-score = 51%, which is far from ideal. Mondal et al. [8] recently proposed SciNLP-KG, an end-to-end natural language processing (NLP) KG construction with 30,000 NLP papers focusing on four extracted relationship types among tasks, datasets, and evaluation metrics. However, their relationship extraction modules still only achieved an F1-score < 80%. Liu et al. [9] defined a metaknowledge architecture to construct structural knowledge with documents, in contrast with previous KGs but similar to the present paper's approach. They employed a multi-modal metaknowledge extraction model to extract and organize metaknowledge elements (e.g., titles, authors, abstracts, and sections) from a government policy document dataset and DocBank [10] dataset. However, due to the computer-annotated data quality, the experiments on the DocBank are just performed with image features rather than multi-modal.

To accommodate the high-end needs of sophisticated decision support systems, we aimed to construct a reusable scientific KG in this paper. To do so, we propose a layout-aware semantic element extraction (LA-SEE) framework that can extract meta and semantic knowledge from S&T documents and construct a KG with the extracted semantic elements. In particular, we combine text-based and vision-based techniques internally to deal with textual and image features. More specifically, compared with existing multi-modal studies [10,11], this paper employs a BERT based language model (i.e., layout-aware metadata extraction (LAME)) [12] to tackle metadata extraction and instance segmentation with transformers (ISTR) (i.e., vision-based semantic element extraction (Vi-SEE)) to achieve vision based object detection. Combining the two models makes semantic element extraction insensitive to layout format. Post-processing procedures extract more accurately captions of figure and table, references, and paragraphs, as well as texts to organize reusable knowledge. Finally, we map those post-processed semantic elements onto the defined semantic elements knowledge graph (SEKG).

We performed semantic element extraction across 70 journals with different layouts to verify the proposed LA-SEE feasibility. Two extraction mechanisms (i.e., LAME and Vi-SEE) were robustly implemented to outperform recent SOTA techniques. Semantic element extraction from 49,649 input PDFs provided 6,782,685 semantic elements for 11 types. Significant contributions from this paper can be summarized as follows:

(1) We implement the Vi-SEE model with a SOTA instance segmentation object detection algorithm.
(2) We define a new scientific KG with 11 different semantic elements (i.e., SEKG).
(3) We define a new LA-SEE framework for textual and visual semantic element extraction and knowledge organization; combining our previously built LAME framework [13] for metadata extraction, Vi-SEE for visual object detection, and SEKG structure for knowledge organization.
(4) We propose two user scenarios based on the proposed SEKG to confirm promising applications.

2. Related Work

2.1. Metadata Extraction from Articles

Research on document structure and information extraction has been steadily ongoing. Primary research directions for metadata extraction can be categorized into rule based, textual feature based machine learning, and vision based object detection. For example, SVM [14], CNN [15], and CRF [16,17] algorithms are popular techniques used with textual features. Pre-training approaches based on large scale text corpora have shown significant successes in several NLP tasks recently, including text classification and sequential labeling [12,18–22].

Sufficient high-quality training datasets annotated with target labels are essential to implement a modest metadata extraction model. Each dataset may have a different annotation level, depending on the research purpose. For example, reference [14] had sentence-level metadata annotations. Reference [16] applied BIO tagging for tokenized words to train a Bi-LSTM-CRF model for metadata extraction, and reference [23] used paragraph-level (or clustered text) annotations. Other studies considered font, font size, and location information to re-organize text chunks to detect layout and extract metadata [24,25]. In contrast, reference [26] automatically annotated document layout elements (i.e., text, titles, lists, tables, and figures) to apply object detection techniques [27,28] for document layout analysis, which is related to metadata extraction.

2.2. Vision-Based Document Analysis

Several studies introduced transformers into object detection tasks, motivated by recent successes for transformers in NLP [29]. A detection transformer (DETR) [30] reconstructed complex object detection components by employing a simple transformer encoder and decoder architecture, providing a neck component to bridge the CNN body for feature extraction and a detector head for prediction. However, although DETR achieved a high detection performance, it suffered from slow convergence, e.g., DETR required 500 epochs, whereas conventional Faster R-CNN [27] training required less than 50 epochs [31]. Recent studies have confirmed the great potential for end-to-end object detection [30,32,33]. Hence, bipartite matching cost has become an essential component for achieving end-to-end object detection. For example, in contrast to [34,35], segmentation explored end-to-end mechanisms with recurrent neural networks, and end-to-end ISTR [36] used the similarity metric for mask embeddings as bipartite matching cost for masks and incorporated transformers [29] to improve end-to-end instance segmentation. We use ISTR in the proposed vision-based semantic element detection task because it showed SOTA level performance even with approximation based suboptimal embeddings.

Document layout analysis is an essential task in automatic document understanding. Its main goal is to identify regions of interest in unstructured documents and recognize each region's roles. However, the task is non-trivial due to document layout diversity and complexity. Many deep learning models have been proposed for this task in computer vision (CV) and NLP fields. Most consider either only visual features [26], only textual features [12], or both modalities [11]. Visual features can identify some regions (e.g., figures, tables), whereas textual features are critical to discriminate visually similar regions (e.g., keywords, abstract, affiliation, author names, etc.). However, single modality models have insufficient capability for layout modeling, hence multi-modal approaches have recently become more popular [9,10,37]. However, they typically contain only hundreds of labeled pages due to prohibitive labeling costs to annotate many layout objects per page, which is insufficient to train and evaluate deep learning based models [27]. Although some multi-modal approaches use automatic data construction methods [10,11,38], they are not interoperable because they employ fundamentally different layout object types and training data formats.

2.3. Scientific Knowledge Extraction

The NLP community includes considerable research on extracting information or knowledge from the scientific literature. Earlier studies focused on identifying citation contexts [39] and extracting key concepts [40] or phrases [41,42]. Most approaches attempted to construct knowledge bases by defining scientific entities and extracting semantic relationships between the entities [2–4]. More recently, reference [8] constructed task-dataset-metric triples from NLP papers by extracting entities and their relationships within and across different sentences/documents.

2.4. Document Modeling

Ronzano and Saggion [43] proposed a platform to extract vast amounts of structural and semantic information from scientific publications, represented as Resource Description Framework (RDF) datasets. Yang et al. [44] designed a weakly-supervised text-to-graph neural network to provide concise, structured representations for documents, by generating concept maps connecting important concepts and interaction links. Zheng et al. [45] introduced four granularity levels for document modeling: documents, paragraphs, sentences, and tokens, reflecting the natural hierarchical document structure. More recently, reference [9] defined a document structure tree model to organize knowledge element extraction from documents and determine their relationships, such as juxtaposition and inclusive, between sections at different levels.

The above works motivated us to extract key semantic elements within the document and derive critical links across multiple documents using the proposed document network structure. Unlike existing knowledge graph construction research, S&T documents exist at the center of reusable knowledge extraction in this study. Therefore, general metadata of documents and their figures, tables, and references were considered semantic elements of knowledge construction. Section 3.3 defines the semantic element knowledge graph (SEKG) because a large number of documents can be interconnected to build vast S&T knowledge. It can be linked to the knowledge graph based on the triple sentences (e.g., relation-entity1-entity2), but we focus on extracting and connecting the document's metadata and the figures and tables of the detailed section or page within the document. There is currently no pre-secured multi-modal training data for semantic elements of different levels. Therefore, text feature-based model (i.e., LAME) is in charge of metadata extraction, and the vision-based object detection model (i.e., Vi-SEE) is responsible for the remaining semantic elements. Moreover, our post-processing delineates the realms of ambiguous semantic elements for more accurate semantic elements identification.

3. LA-SEE Framework

This study proposed a LA-SEE framework to extract meta-information, text, sub-titles, references, figures, tables, and captions from scientific PDFs. Figure 1 shows that proposed LA-SEE framework comprises three major components.

(1) The LAME [13] model extracts five metadata types (title, author, affiliation, keywords, and abstract) from the first PDF page.
(2) Vi-SEE performs object detection for the remaining pages to extract other semantic elements (paragraphs, figures, tables, captions, and references) and post-processing to obtain texts of the elements. Figures and tables are saved as image files, whereas other metadata and semantic elements are converted to JavaScript Object Notation (JSON) format.
(3) Extracted semantic elements go through knowledge organizing/mapping under our SEKG structure defined in Section 3.3. The metadata from LAME and document objects from Vi-SEE are collectively referred to as semantic elements.

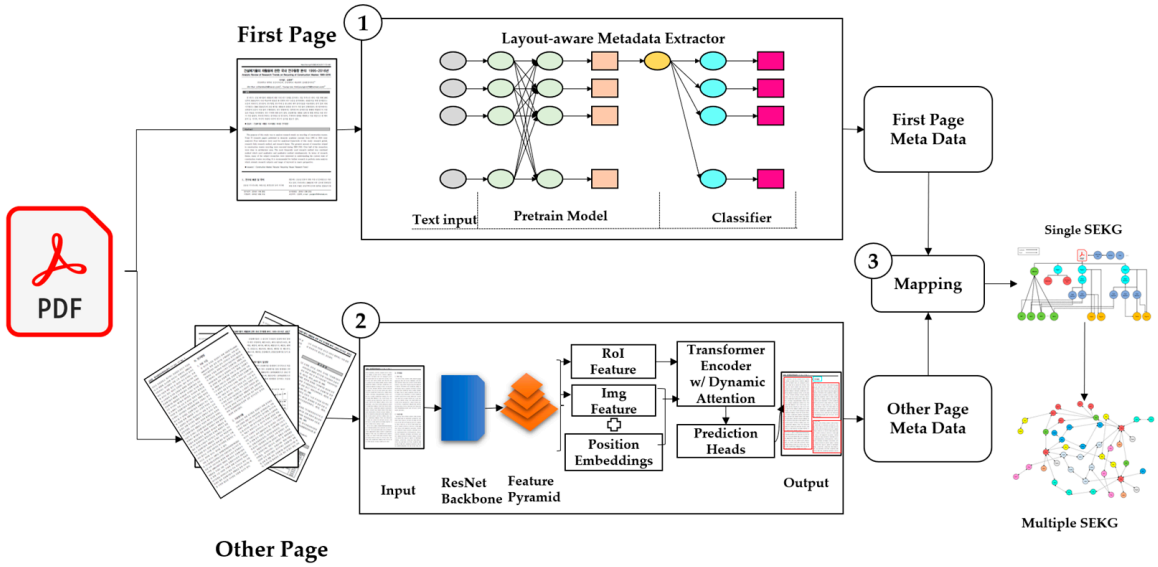


Figure 1. Proposed system architecture.

3.1. LAME

We adopted our prior work, LAME framework to discriminate metadata elements in the first document page, considering text block characteristics in the heterogeneous meta-information layouts [13]. Figure 2 shows that the LAME framework comprises three major components: automatic layout analysis, layout-aware training data construction, and metadata extraction. Stage 1 analyzes the PDF's first-page by using PDFMiner, then is subject to reconstruction, refinement, and adjustment procedures to identify the various metadata on the first page due to incomplete PDFMiner parsing results. Stage 2 builds the many training datasets used in Stage 3. The building process matches identified metadata from Stage 1 with previous correct metadata values. However, the compared textual content is not always precisely matched. Therefore, to determine the extent of the match, we allowed only fields with almost identical (or high similarity) matches for each layout text information element automatically acquired in the previous step as training data. We used a mixed textual-similarity measure for efficient computation based on the Levenshtein distance and bilingual evaluation understudy (BLEU) score.

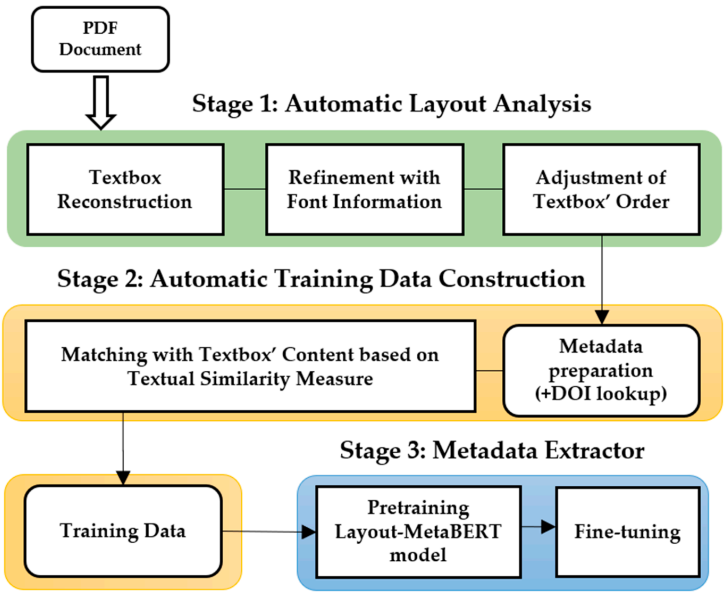


Figure 2. LAME (excerpted and revised from [13] (p. 4)).

The created dataset have not correct answer dataset for comparing results, and manual comparison spend much time and resource. Thus, to determine the accuracy of the training data generated through Stage 2, we indirectly evaluated the data quality through the metadata extraction in Stage 3. Finally, a novel metadata extractor is defined by pre-training the Layout-MetaBERT model with the Stage 2 training data and fine-tuning it for the target corpus.

We chose a fine-tuned Layout-MetaBERT (base) with robust metadata extraction performance (F1 = 94.6%) even for unseen journals with diverse layouts by referring to various experimental results for the LAME framework [13].

3.2. Vi-SEE

Figure 3 describes the proposed Vi-SEE model, which utilizes ISTR [36] to detect objects in pages in the PDF document except for the first page. Images from the PDF pass through the ISTR based detection model to identify candidate bounding boxes (BBoxes) for text, titles, lists, figures, and tables. Input image passes through the convolution natural network based on the reset backbone, produces a feature pyramid. RoI (Region of Interest) feature and image feature are separated from the feature pyramid, and image feature and position feature are concatenated. Moreover, transformer encoder with dynamic attention fuses the image + position and RoI features for prediction head. Each detected area is converted into actual data through a set of post-processing procedures using the detected BBoxes corresponding categorical labels: (1) text extraction for text, lists, and titles, (2) figure/table extraction, and (3) caption extraction. Previous studies have only performed this at area-level detection, whereas the proposed modules include detailed techniques to extract precise regions for semantic element areas and related texts.

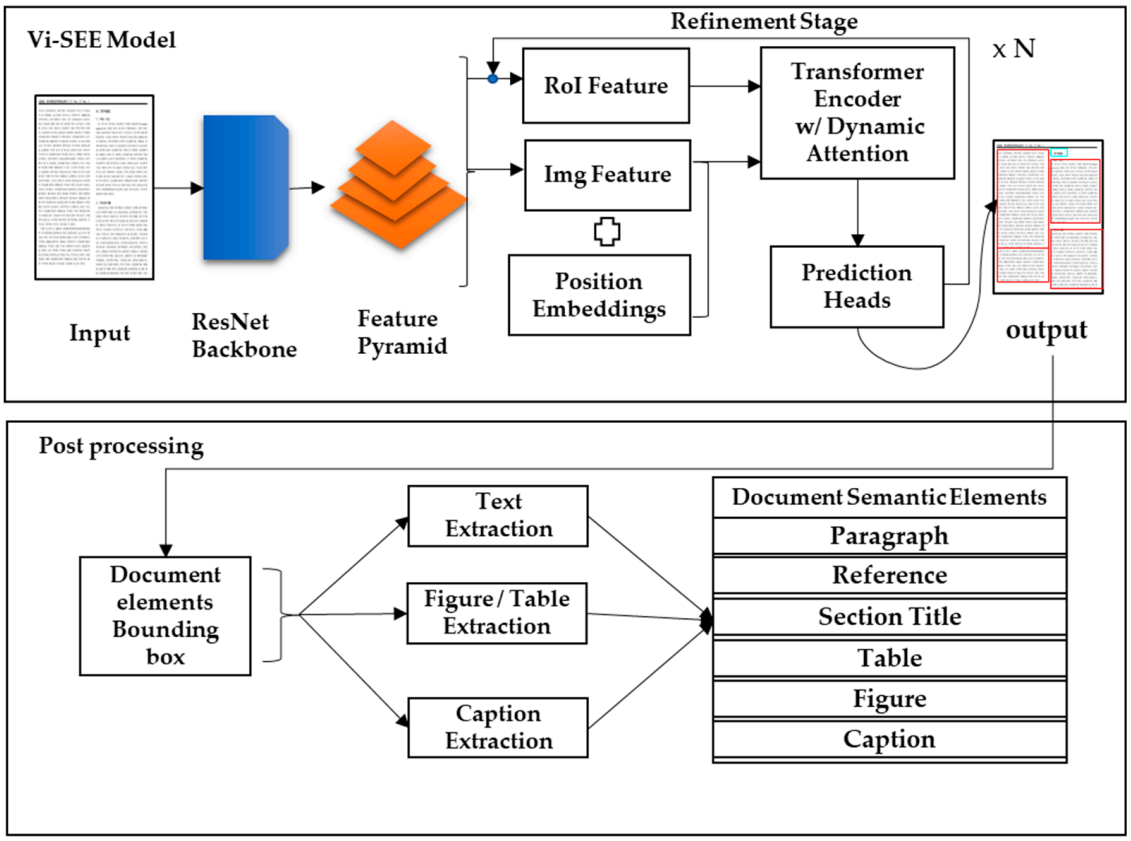


Figure 3. Proposed Vi-SEE details.

3.2.1. ISTR Selection

Before selecting the ISTR [36], we compared three popular object detection models, Mask R-CNN [28], DETR [30], and ISTR, for accurate semantic element extraction from the document. They all use the ResNet [46] backbone. The Mask R-CNN model is a derived image segmentation model after Faster R-CNN [27]. It has a similar structure to Faster R-CNN except its object mask branch, RoI alignment, and decoupling mask prediction and class prediction. However, the Mask R-CNN model suffers from low detection speed due to the detection pipeline's non-maximum suppression (NMS) stage.

On the other hand, DETR omits NMS from the detection pipeline while improving speed similarly to Faster R-CNN. It predicts all objects at once and only has simple pipelines that do not require NMS or anchors and is good for finding large objects, but fails to find small/middle sized objects. The ISTR algorithm provides end-to-end instance segmentation by regressing low-dimensional embeddings rather than raw masks, which enables training to be effectively conducted with a small number of matched samples. Regressing with the embeddings allows a recurrent refinement strategy that can process detection and segmentation concurrently, boosting performance. It updates query boxes and refines the prediction sets. We chose ISTR as the main Vi-SEE algorithm because there are many medium and large objects in our target documents. The primary training method of ISTR learning follows DETR [30]. A key point in ISTR learning is that there is a refinement stage. The basic formula for self-attention of ISTR is as follows:

$$Output = softmax\left(\frac{Query \times Key^{Transpose}}{\sqrt{d}}\right) \times Value$$

In Multi-Head Attention, a dynamic attention module is added so that RoI and image features can be well fused, and it is summarized as follows.

$$Feature^i = RoI^i \times fully - connection(Output)$$

Furthermore, the refinement stages can improve the performance of the predicted bounding boxes, classes, and masks by updating the query boxes.

When the page of the document enters as the input of the model, object detection is performed through the ISTR model. The object detection task is to detect instances of objects of a certain class within an image by considering the bounding box area, segmentation area, and candidate labels.

3.2.2. Post-Processing Identified Semantic Elements

- Text extraction: BBox areas are converted into texts using PDFMiner [47] parsing results for text, lists, and titles extracted from the ISTR based model. PDFMiner returns parsed texts with position information for PDF document. We extract texts using the left-top and right-bottom positions for the detected areas. The extracted semantic elements are references, paragraphs, and section titles.
- Figure/table extraction: We take screenshots encompassing the BBoxes and save them as images for detected areas such as figures and tables.
- Caption extraction: Rather than using BBox coordinates for the detected semantic elements (text, figures, and tables), we find candidate areas for captions based on a distance measure and change the areas into texts using PDFMiner's parsed results. The closest text BBox is resolved as a caption. We compute the distance between the midpoint for the detected figure (or table) BBoxes and the midpoint for text BBoxes. FTmid refers to the midpoint found using the bounding box of figure or table. Tmid refers to the midpoint found using the bounding box of text. Thus, midpoint for figure (or table) BBox can be expressed as

$$FTmid(x_1, y_1) = (|x_2 - x_1|, |y_2 - y_1|) \quad (1)$$

and text BBox as

$$Tmid(x_1, y_1) = (|x_2 - x_1|, |y_2 - y_1|) \quad (2)$$

Distance for each midpoint can be expressed as

$$Dist = \sqrt{(FTmid(x_1) - Tmid(x_1))^2 + (FTmid(y_1) - Tmid(y_1))^2} \quad (3)$$

and midpoint for a caption as

$$Caption\ mid = Min(Dist) \quad (4)$$

Caption text is extracted in the same way as for normal text extraction.

Figure 4 shows examples of objects extracted through Vi-SEE as well as the semantic elements and their labels.

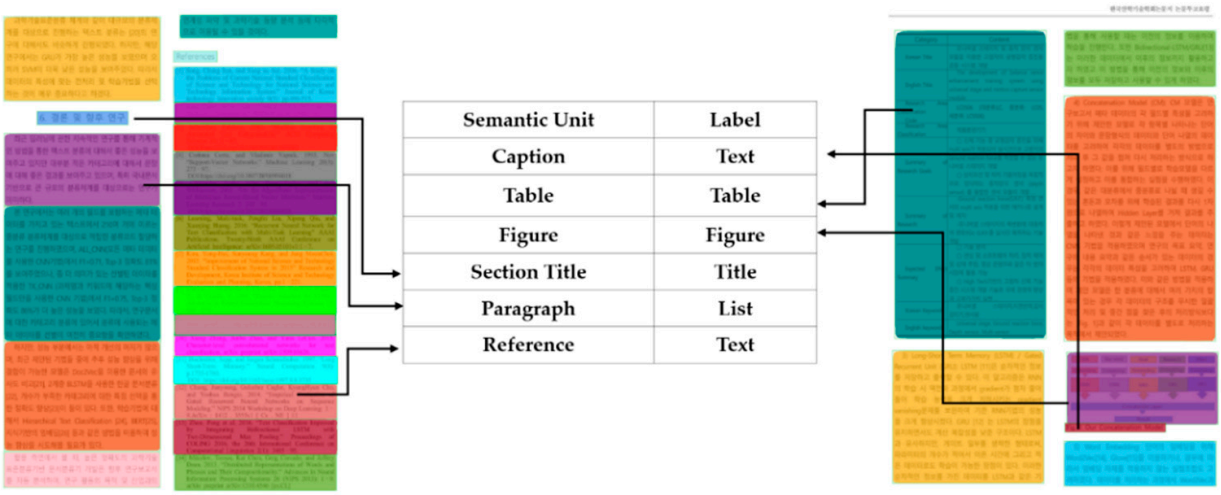


Figure 4. Extracted semantic element examples from Vi-SEE.

3.3. Organizing Knowledge with SEKG for Multiple Documents

Many applications that require analyzing a large amount of knowledge from various angles become possible once the knowledge relationships in S&T documents are identified, and if knowledge from different documents are interconnected. Suppose those semantic elements representing knowledge across a considerable number of documents are well organized. Then, researchers (or policymakers) can expedite their decision-making by streamlining information/knowledge collection and analysis. For example, reference [48] performed a behavioral study on citations, reference [3] extracted tasks, datasets, metrics, and scores from NLP papers to automatically construct a leaderboard, and reference [9] suggested a metaknowledge construction framework and document structure tree model to reduce gaps between human knowledge perception and entity-relationship triplets.

Influenced by those studies, we defined an SEKG for multiple document structures that can connect multiple semantic elements in a single document, or across multiple documents, as shown in Figure 5. Relationships are identified between 11 semantic elements types extracted from documents using the proposed LAME and Vi-SEE modules, and mapped under the SEKG structure. The first page of most documents includes significant metadata, including author name(s) and affiliation(s), publisher, abstract, and introduction. We regarded these metadata separately from document's contents that were not included in a specific page or section. These metadata elements provide an essential reasoning link when several documents are linked, as shown in Figure 6.

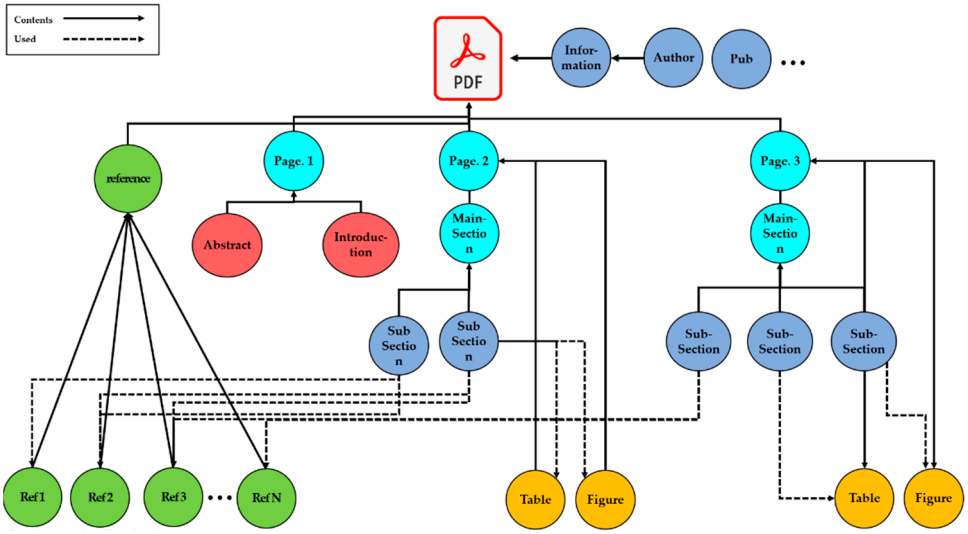


Figure 5. SEKG definition within a document.

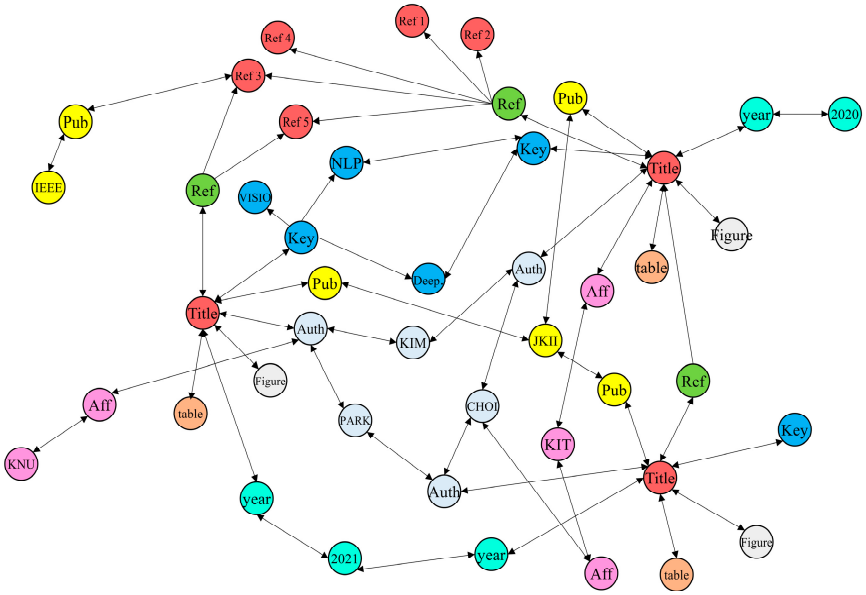


Figure 6. SEKG example across multiple documents.

Semantic elements extracted from a document have a hierarchical structure from the main section to the sub-section. However, it is essential to consider when figures (or tables) located on different pages can be cited more than once from different pages. Therefore, the proposed SEKG structure maps extracted semantic elements to a network node rather than the hierarchical structure, considering various relationships connecting figures, tables, and references.

4. Experiments

4.1. Datasets

4.1.1. Data for Metadata Extraction

We use the first pages of 65,007 PDF documents from 70 S&T journal articles to reflect various document layout formats for the metadata extraction task. It is the same dataset used in our prior work [13]. We extracted major metadata elements, such as titles, author names, author affiliations, keywords, and abstracts, in Korean and English based on the automatic layout analysis in Section 3.1. Among the 70 journal articles, two were only in Korean, 23 were only in English, and 45 were Korean and English. Automatic labeling

was applied with ten labels for each layout that separated metadata on the first page of articles with other layouts not included in the relevant information labeled as O. Table 1 summarizes the automatically generated training data.

Table 1. Automatically generated training data [13] (p. 9).

Metadata Field	Label (i.e., Layout)	Count
Out of boundary	O	637,856
Title (in Korean)	title_ko	46,056
Title (in English)	title_en	64,414
Affiliation (in Korean)	aff_ko	39,233
Affiliation (in English)	aff_en	63,434
Abstract (in Korean)	abstract_ko	31,885
Abstract (in English)	abstract_en	55,318
Keywords (in Korean)	keywords_ko	21,685

4.1.2. Dataset for Vi-SEE

Large high-quality annotated training datasets are essential to creating a robust object detection model. However, accurately detecting target semantic elements from PDF documents is still not guaranteed even if similar datasets exist [10,26,49] due to varying layout formats across journals. Therefore, we constructed the proposed Vi-SEE module training dataset with the following steps.

(1) Five major semantic elements (i.e., section title, paragraph, reference, table, and figure) were pseudo-labeled for the 70 scientific journal articles using the Mask-RCNN [28] model trained with the PubLayNet [26] dataset following COCO data format [50].

(2) The coco-annotator API was used to modify the mask parts that were not properly labeled, as shown in Figure 7. Five paid annotators performed a cross-check on each other's work to guarantee annotation quality. Revised pages that focused on error-prone cases amounted to 20,079, summarized in Table 2. The data were randomly divided into training (i.e., fine-tuning) and testing sets at 80:20, respectively.

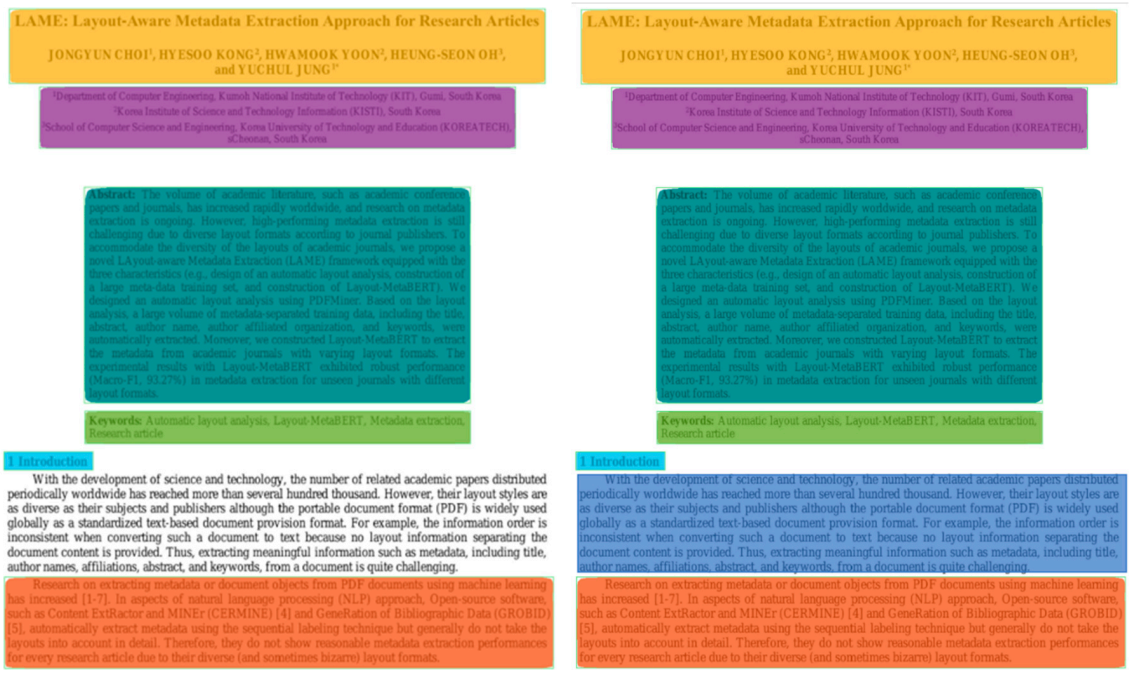


Figure 7. Data construction example: (**a**) Data extracted using Mask R-CNN model trained on PubLayNet data and (**b**) modified data.

Table 2. Constructed dataset summary.

Semantic Elements	Training Set	Test Set	Total
Section title	32,284	5724	38,008
Paragraph	111,253	19,990	131,243
Reference	57,813	10,197	68,010
Table	4433	782	5215
Figure	12,937	2303	15,240
Total num. of pages	17,024	3055	20,079

4.2. Proposed LA-SEE Performance

Table 3 shows the device information and the version of cuda we used in the experiments. We used i9-10900 CPU and two Tesla v100 GPUs to fine-tune comparison targets with LAME and Vi-SEE models.

Table 3. Details of the used system settings for experiments.

System Settings	Specification
CPU	Intel(R) Core(TM) i9-10900X CPU @ 3.70GHz
GPU	Tesla V100-PCIE-32GB × 2
RAM	256 G
Cuda Version	Cuda 10.1

Table 4 shows that the proposed LAME model effectively extracted metadata, achieving F1-score $\geq$ 90% for all extractions and average F1-score = 93%, confirming that pre-training the layout units with BERT schemes is feasible. Similarly, the proposed Vi-SEE model effectively detected semantic elements using vision, achieving average mAP = 85%.

Table 4. Experimental results for LAME and Vi-SEE compared with other SOTA models.

Element	LAME [13] (F1-Score)	KoELECTRA [51] (F1-Score)
Title	0.92	0.87
Abstract	0.9	0.9
Keywords	0.94	0.91
Author	0.9	0.73
Affiliation	0.92	0.57
Average	0.93	0.87
Element	**Vi-SEE (mAP)**	**Mask R-CNN Trained with PubLayNet and Fine-Tuned with Our Data (mAP)**
Section title	0.6841	0.6445
Paragraph	0.8456	0.8018
Table	0.9323	0.8989
Figure	0.8975	0.7144
Reference	0.8959	0.8398
Average	0.8493	0.7798

We performed a set of transfer learning for the constructed data before building the Vi-SEE module, based on three pre-trained models (as shown in Table 5):

(1) Mask R-CNN model pre-trained with PubLayNet data,
(2) DETR model pre-trained with ImageNet data, and
(3) ISTR model pre-trained with ImageNet data.

Table 5. Algorithm performances under different configurations.

Model	AP	AP50	AP75	APm	APl
A	77.99%	93.46%	86.98%	41.61%	80.23%
B	81.5%	97.1%	90.1%	59.3%	82.1%
C	85.11%	98.16%	93.33%	65.44%	85.60%

Notes: A: Detectron2 with PubLayNet and Our Data, B: DETR with Our Data, C: ISTR with Our Data.

We used the Mask R-CNN model trained with PubLayNet data based on the Detectron2 framework for our fine-tuning task. Both DETR and ISTR used the pre-trained ResNet-101 model [52,53] as the backbone in their fine-tuning stage. We follow the default configurations of each model.

The fine-tuned models achieved overall modest performance on AP50, whereas the ISTR based model achieved highest mAP on AP50. Semantic elements in the documents were primarily large and medium scale, but small scale when Common Object in Context (COCO) metrics were applied [54]. The ISTR based model, detects medium and large objects well, achieving superior results to DETR, whose strength lies in detecting large objects. Looking at Table 5, ISTR is about 23% higher than Mask R-CNN in average precision medium (APm) and about 6% higher than DETR. In average precision large (APl), it is about 5% higher than Mask R-CNN and about 3% higher than DETR, showing the best performance. Therefore, it better detects the area of the semantic element than others.

4.3. Constructed Semantic Element Statistics

Table 6 shows the statistics for 6,782,685 semantic elements extracted from 49,649 PDF documents using the proposed LA-SEE framework. Although semantic element counts for each type differ, this statistic is useful for estimating the number of knowledge instances acquired considering the number of input documents.

Table 6. Automatically constructed SEKG summary.

Semantic Element Types	Extracted Elements (Count)
Title	49,094
Abstract	60,526
Keywords	56,634
Author	52,216
Affiliation	50,951
Section title	1,019,749
Paragraph	2,700,498
Table	138,613
Figure	388,416
Caption	527,029
Reference	1,711,959
Total PDF documents	49,649
Total extracted semantic elements	6,782,685

5. Decision Support Applications in Science and Technology Domain

When users search for desired information using search engines, such as Google or Naver, users employ relevant keywords to search for desired information using search engines, such as Google or Naver, and check search results (title, snippets, summary, and document) one by one to determine if they are relevant for their information needs. However, users often perform very repetitive searching and checking processes to access sufficient suitable information. The proposed SEKG framework provides relational information access that supports quick decision-making while reducing laborious information searches. For example, users can perform numerous reasoning types over the relationships among research data, relationships between semantic elements across multiple documents,

related keywords through directly (or indirectly) linked documents, and large KG comprising triple sentences.

The SEKG can be applied to several real-world S&T applications in various fields, including, but not limited to, science knowledge guides, question answers over a large number of figures and tables, and generating textual explanations for scientific issues. We describe how SEKG satisfies scientific requirements with two applications below.

5.1. Scientific Knowledge Guide

Researchers commonly find and compare academic documents and research reports, time costs for information-seeking are rapidly increasing due to continuously increasing number of documents.

The proposed SEKG framework offers an elegant solution for the problem, providing relevant figures, tables, and captions that satisfy user requirements. A semantic query is sent to the SEKG for knowledge discovery, and the SEKG delivers a group of figures that meet the query conditions.

Figure 8 shows SEKG results differ significantly from general search engine results. For example, suppose an NLP beginner examines pre-trained models published in recent studies with the query, "Text Pretrained Model". The SEKG enables easy and quick access to pretrained model pictures (e.g., BERT, TinyBERT, BART, ELECTRA, and DialogBERT) mentioned in various research papers, affiliations for authors that developed these models, and related paper titles. New knowledge, such as research trends for major research institutions, can also be summarized as required by modifying user queries over the SEKG.

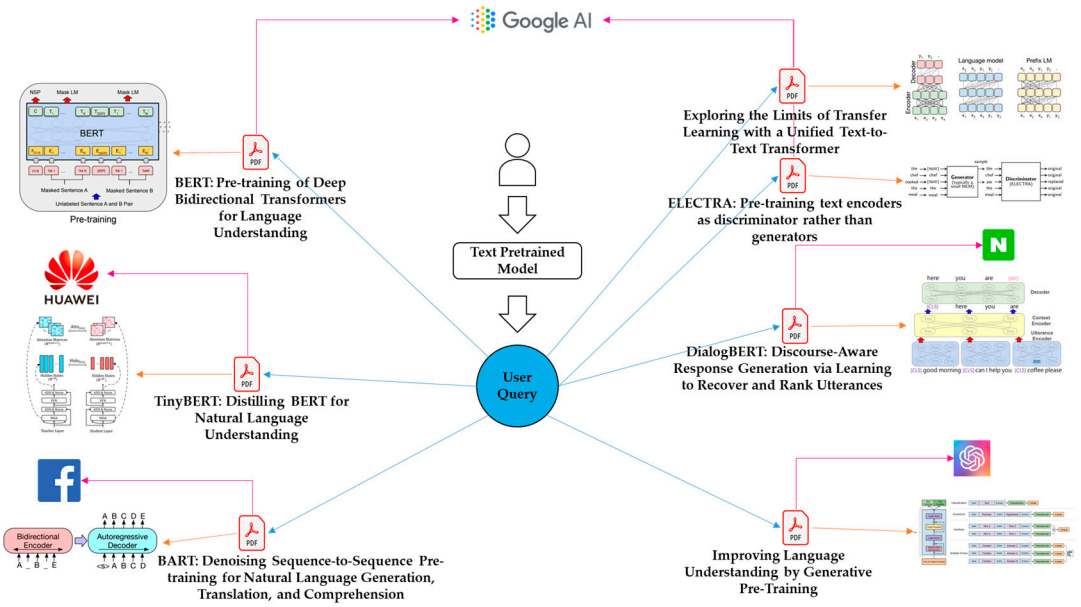


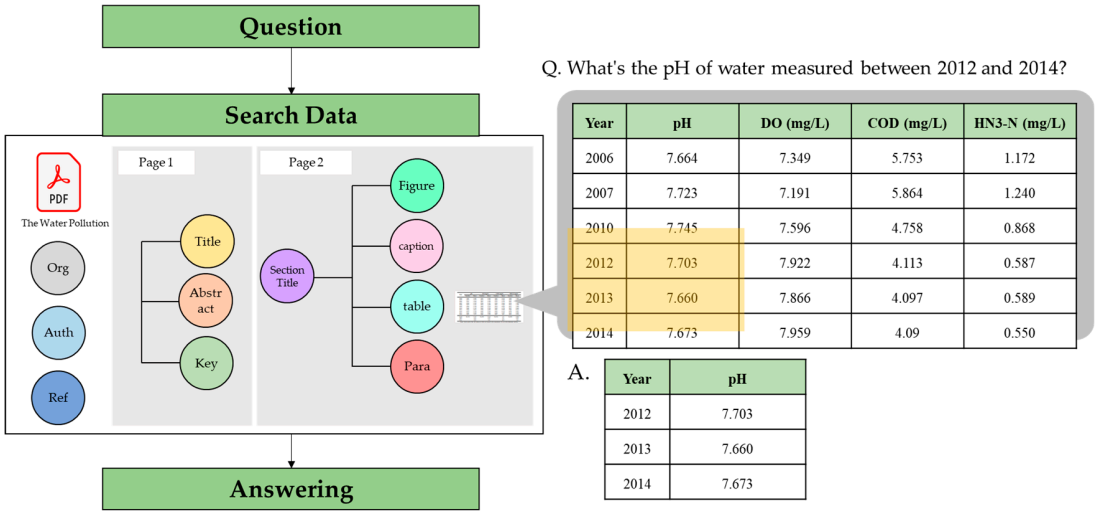
Figure 8. Scientific knowledge guide: Figure search.

5.2. Questions and Answering over Tables

Tables in papers and reports deliver condensed information, commonly employing numerical values to represent actual experimental performance or statistical results. Therefore, accessing the values provides many benefits to researchers. Suppose a table search is performed to satisfy the user's information request. For example, search for tables that contain captions and descriptions that match user keywords, but still require a selection process is within them. In this case, the SEKG can directly access exact values in the tables while minimizing the selection process or inferring new values based on these values.

For example, suppose a user is interested in water pollution content in the environmental field and wants to know the mean annual pH for 2012–2014 water measurements.

SEKG will select the table containing pH values from 2012, 2013, and 2014 among several water pollution documents, as shown in Figure 9.



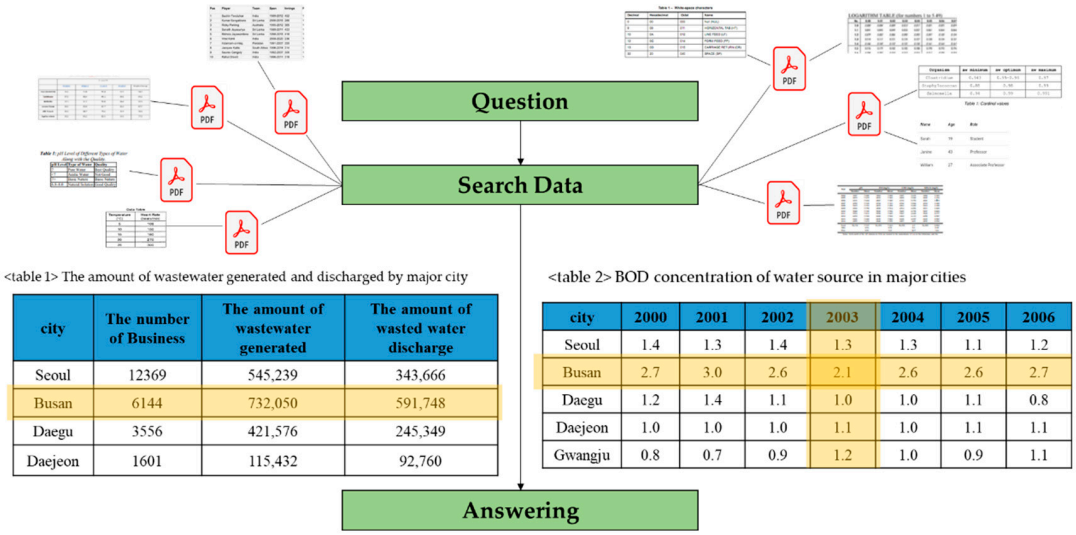
Year	pH	DO (mg/L)	COD (mg/L)	HN3-N (mg/L)
2006	7.664	7.349	5.753	1.172
2007	7.723	7.191	5.864	1.240
2010	7.745	7.596	4.758	0.868
2012	7.703	7.922	4.113	0.587
2013	7.660	7.866	4.097	0.589
2014	7.673	7.959	4.09	0.550

Year	pH
2012	7.703
2013	7.660
2014	7.673

Figure 9. Accessing specific table values.

Employing a table QA (Question-Answer) module [55] and mathematical reasoning injection [56] allows the user to obtain average pH between 2012 and 2014. Rather than simply providing the information, the average value is computed using the language model's mathematical reasoning capability. Furthermore, several tables extracted from two or more document types in the same field can be processed and provided to suit the user's requirements.

For example, suppose a researcher wants to know Busan's BOD (the degree of contamination by organic substances) in 2003 and wonders if the distribution of pollutants affects BOD. In this case, SEKG finds a table of BOD content in water pollution-related documents, as shown in Figure 10, Table 2, which contains the content for Busan in 2003. For more complex questions, SEKG may search the pollutant distribution table (Figure 10, Table 1) and provide pollutant distribution for Busan (Figure 10, Table 2). Thus, various information comprising tables, figures, and statistics can be provided to suit user requirements regardless of specific fields and data quantities, by analyzing, processing, and combining data beyond simple information provisioning.



city	The number of Business	The amount of wastewater generated	The amount of wasted water discharge
Seoul	12369	545,239	343,666
Busan	6144	732,050	591,748
Daegu	3556	421,576	245,349
Daejeon	1601	115,432	92,760

city	2000	2001	2002	2003	2004	2005	2006
Seoul	1.4	1.3	1.4	1.3	1.3	1.1	1.2
Busan	2.7	3.0	2.6	2.1	2.6	2.6	2.7
Daegu	1.2	1.4	1.1	1.0	1.0	1.1	0.8
Daejeon	1.0	1.0	1.0	1.1	1.0	1.1	1.1
Gwangju	0.8	0.7	0.9	1.2	1.0	0.9	1.1

Figure 10. Accessing Multiple Tables Obtained from Different Documents.

6. Conclusions and Future Work

This paper proposed the LA-SEE framework to build a reusable SEKG from various documents. In particular, 11 semantic element types were defined and extracted from various S&T journals using LAME and Vi-SEE. LA-SEE uses BERT based metadata extraction with textual features and ISTR based object detection to achieve SOTA performance with textual and image features. As results, we established a large scale SEKG comprising 6 million semantic elements using LAME and Vi-SEE and discussed two usage scenarios (i.e., scientific knowledge guide and QA over tables) to highlight the proposed SEKG framework applicability and extensibility. In the first scenario, it was possible to find and present figures of similar architectures belonging to semantically similar topics in several different documents through SEKG. Furthermore, in the second scenario, we showed that it is possible to present values that satisfy user needs by accessing joinable tables' values in different documents.

The limitation of this study is that the training data of 11 semantic elements of SEKG do not have consistency. Each of the LAME and Vi-SEE training data has different levels of annotation (i.e., text or vision), and multi-modal features are not considered yet. Therefore, it is necessary to construct a dataset and apply an advanced training algorithm to consider multi-modality in our future research. In addition, although the currently constructed data sets are composed of about 40 journals in different formats, there is still a limit to accurately processing S&T documents in various subject domains. Therefore, when an S&T document of a new subject domain is an input, it may be challenging to extract semantic elements, so to apply it to documents in other domains, a process of generating a new dataset and training a new model is required.

Moreover, further work remains to better handle various exceptions and errors naturally occurring due to formatting and related differences among documents. For example, LAME does not always correctly identify target elements and Vi-SEE fails to distinguish figure regions comprising complex images. We plan to employ multi-modal transformer techniques to address these issues, rather than single-modal approaches, which will require high-quality Optical Character Recognition (OCR) module(s) to convert document data into multi-modal training sets containing massive documents numbers.

We will also investigate accurately extracting related figure and table descriptions and add them as new semantic elements to SEKG. Although figures and tables are primary information in S&T documents, their corresponding descriptions are not currently considered. If SEKG were empowered with explanatory texts for figures and tables, it would be possible to build new scientific, conversational AI applications by enabling table-to-text (or figure-to-text) functionality.

Author Contributions: Conceptualization by H.K. and Y.J.; Development of Software and Evaluation by H.K. and J.C.; Data Analytics and Writing-Original Draft Preparation by H.K., S.P., J.C. and Y.J.; Writing-Review and Editing by Y.J. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the Korea Institute of Science and Technology Information (KISTI) through Construction on Science & Technology Content Curation Program (K20-L01-C01) and the National Research Foundation of Korea (NRF) under a grant funded by the Korean Government (MSIT) (No. NRF-2018R1C1B5031408). In addition, this research is the result of a study on the HPC Support project supported by the Ministry of Science and ICT and NIPA.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Knowledge Graph. Available online: https://en.wikipedia.org/wiki/Knowledge_graph (accessed on 10 December 2021).
2. Augenstein, I.; Das, M.; Riedel, S.; Vikraman, L.; McCallum, A. SemEval 2017 Task 10: ScienceIE-Extracting Keyphrases and Relations from Scientific Publications. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Vancouver, BC, Canada, 3–4 August 2017; pp. 546–555. [CrossRef]
3. Hou, Y.; Jochim, C.; Gleize, M.; Bonin, F.; Ganguly, D. Identification of Tasks, Datasets, Evaluation Metrics, and Numeric Scores for Scientific Leaderboards Construction. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 5203–5213. [CrossRef]
4. Jain, S.; van Zuylen, M.; Hajishirzi, H.; Beltagy, I. SciREX: A Challenge Dataset for Document-Level Information Extraction. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 7506–7516. [CrossRef]
5. Gábor, K.; Buscaldi, D.; Schumann, A.K.; QasemiZadeh, B.; Zargayouna, H.; Charnois, T. Semeval-2018 task 7: Semantic relation extraction and classification in scientific papers. In Proceedings of the 12th International Workshop on Semantic Evaluation, New Orleans, LA, USA, 5–6 June 2018; pp. 679–688. [CrossRef]
6. Xu, J.; Kim, S.; Song, M.; Jeong, M.; Kim, D.; Kang, J.; Rousseau, J.F.; Li, X.; Xu, W.; Torvik, V.I.; et al. Building a PubMed knowledge graph. *Sci. Data* **2020**, *7*, 205. [CrossRef] [PubMed]
7. Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C.H.; Kang, J. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **2020**, *36*, 1234–1240. [CrossRef] [PubMed]
8. Mondal, I.; Hou, Y.; Jochim, C. End-to-End NLP Knowledge Graph Construction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 1885–1895. [CrossRef]
9. Liu, S.K.; Xu, R.L.; Geng, B.Y.; Sun, Q.; Duan, L.; Liu, Y.M. Metaknowledge Extraction Based on Multi-Modal Documents. *IEEE Access* **2021**, *9*, 50050–50060. [CrossRef]
10. Li, M.; Xu, Y.; Cui, L.; Huang, S.; Wei, F.; Li, Z.; Zhou, M. DocBank: A Benchmark Dataset for Document Layout Analysis. In Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain, 13–18 September 2020; pp. 949–960. [CrossRef]
11. Xu, Y.; Xu, Y.; Lv, T.; Cui, L.; Wei, F.; Wang, G.; Lu, Y.; Florencio, D.; Zhang, C.; Che, W.; et al. LayoutLMv2: Multi-modal pre-training for visually-rich document understanding. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Virtual Event, 1–6 August 2021; Volume 1, pp. 2579–2591. [CrossRef]
12. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019; Volume 1, pp. 4171–4186. [CrossRef]
13. Choi, J.; Kong, H.; Yoon, H.; Oh, H.-S.; Jung, Y. LAME: Layout Aware Metadata Extraction Approach for Research Articles. *arXiv* **2021**, arXiv:2112.12353.
14. Han, H.; Giles, C.L.; Manavoglu, E.; Zha, H.; Zhang, Z.; Fox, E.A. Automatic document metadata extraction using support vector machines. In Proceedings of the 2003 Joint Conference on Digital Libraries, Houston, TX, USA, 27–31 May 2003; pp. 37–48. [CrossRef]
15. Kim, Y. Convolutional Neural Networks for Sentence Classification. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, Doha, Qatar, 25–29 October 2014. [CrossRef]
16. Kim, S.; Ji, S.; Jeong, H.; Yoon, H.; Choi, S. Metadata Extraction based on Deep Learning from Academic Paper in PDF. *J. KIISE* **2019**, *46*, 644–652. [CrossRef]
17. Luong, M.-T.; Nguyen, T.D.; Kan, M.-Y. Logical structure recovery in scholarly articles with rich document features. *Int. J. Digit. Libr. Syst.* **2010**, *1*, 1–23. [CrossRef]
18. Adhikari, A.; Ram, A.; Tang, R.; Lin, J. DocBERT: BERT for Document Classification. *arXiv* **2019**, arXiv:1904.08398.
19. Yu, S.; Su, J.; Luo, D. Improving bert-based text classification with auxiliary sentence and domain knowledge. *IEEE Access* **2019**, *7*, 176600–176612. [CrossRef]
20. Gu, X.; Yoo, K.M.; Ha, J.-W. Dialogbert: Discourse-Aware Response Generation via Learning to Recover and Rank Utterances. *arXiv* **2021**, arXiv:2012.01775.
21. Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; Soricut, R. Albert: A Lite Bert for Self-Supervised Learning of Language Representations. *arXiv* **2020**, arXiv:1909.11942.
22. Beltagy, I.; Lo, K.; Cohan, A. Scibert: A Pretrained Language Model for Scientific Text. *arXiv* **2019**, arXiv:1903.10676.
23. Garncarek, Ł.; Powalski, R.; Stanisławek, T.; Topolski, B.; Halama, P.; Graliński, F. LAMBERT: Layout-Aware Language Modeling for Information Extraction. In Proceedings of the International Conference on Document Analysis and Recognition, Lausanne, Switzerland, 5–10 September 2021; pp. 532–547. [CrossRef]
24. Constantin, A.; Pettifer, S.; Voronkov, A. PDFX: Fully-automated PDF-to-XML conversion of scientific literature. In Proceedings of the 2013 ACM Symposium on Document Engineering, Florence, Italy, 10–13 September 2013; pp. 177–180. [CrossRef]
25. Ahmed, M.W.; Afzal, M.T. FLAG-PDFe: Features oriented metadata extraction framework for scientific publications. *IEEE Access* **2020**, *8*, 99458–99469. [CrossRef]

26. Zhong, X.; Tang, J.; Yepes, A.J. Publaynet: Largest dataset ever for document layout analysis. In Proceedings of the 2019 International Conference on Document Analysis and Recognition, Sydney, Australia, 20–25 September 2019; pp. 1015–1022. [CrossRef]
27. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
28. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969. [CrossRef]
29. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
30. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 213–229. [CrossRef]
31. Detectron2. Available online: https://github.com/facebookresearch/detectron2 (accessed on 7 December 2021).
32. Sun, P.; Zhang, R.; Jiang, Y.; Kong, T.; Xu, C.; Zhan, W.; Tomizuka, M.; Li, L.; Yuan, Z.; Wang, C.; et al. Sparse R-CNN: End-to-end object detection with learnable proposals. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 14454–14463. [CrossRef]
33. Wang, J.; Song, L.; Li, Z.; Sun, H.; Sun, J.; Zheng, N. End-to-end object detection with fully convolutional network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 15849–15858. [CrossRef]
34. Ren, M.; Zemel, R.S. End-to-end instance segmentation with recurrent attention. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6656–6664. [CrossRef]
35. Shen, Y.; Ji, R.; Wang, Y.; Wu, Y.; Cao, L. Cyclic guidance for weakly supervised joint detection and segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 697–707. [CrossRef]
36. Hu, J.; Cao, L.; Lu, Y.; Zhang, S.; Wang, Y.; Li, K.; Huang, F.; Shao, L.; Ji, R. ISTR: End-to-End Instance Segmentation with Transformers. *arXiv* **2021**, arXiv:2105.00637.
37. Kaplan, F.; Oliveira, S.A.; Clematide, S.; Ehrmann, M.; Barman, R. Combining visual and textual features for semantic segmentation of historical newspapers. *J. Data Min. Digit. Humanit.* **2021**. [CrossRef]
38. Xu, Y.; Lv, T.; Cui, L.; Wang, G.; Lu, Y.; Florencio, D.; Zhang, C.; Wei, F. LayoutXLM: Multi-Modal Pre-Training for Multilingual Visually-Rich Document Understanding. *arXiv* **2021**, arXiv:2104.08836.
39. Teufel, S.; Siddharthan, A.; Tidhar, D. Automatic classification of citation function. In Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, Sydney, Australia, 22–23 July 2006; pp. 103–110. [CrossRef]
40. Tsai, C.T.; Kundu, G.; Roth, D. Concept-based analysis of scientific literature. In Proceedings of the 22nd ACM International Conference on Information & Knowledge Management, San Francisco, CA, USA, 27 October–1 November 2013; pp. 1733–1738. [CrossRef]
41. Kim, S.N.; Medelyan, O.; Kan, M.Y.; Baldwin, T. Automatic keyphrase extraction from scientific articles. *Lang. Resour. Eval.* **2013**, *47*, 723–742. [CrossRef]
42. Hasan, K.S.; Ng, V. Automatic keyphrase extraction: A survey of the state of the art. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, MD, USA, 22–27 June 2014; Volume 1, pp. 1262–1273. [CrossRef]
43. Ronzano, F.; Saggion, H. Knowledge extraction and modeling from scientific publications. In Proceedings of the International Workshop on Semantic, Analytics, Visualization, Montreal, QC, Canada, 11 April 2016; pp. 11–25. [CrossRef]
44. Yang, C.; Zhang, J.; Wang, H.; Li, B.; Han, J. Neural concept map generation for effective document classification with interpretable structured summarization. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, Xi'an, China, 25–30 July 2020; pp. 1629–1632. [CrossRef]
45. Zheng, B.; Wen, H.; Liang, Y.; Duan, N.; Che, W.; Jiang, D.; Zhou, M.; Liu, T. Document Modeling with Graph Attention Networks for Multi-grained Machine Reading Comprehension. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 6708–6718. [CrossRef]
46. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
47. PDFMiner: Python PDF Parser and Analyzer. Available online: http://www.unixuser.org/~{}euske/python/pdfminer/ (accessed on 20 November 2021).
48. Jurgens, D.; Kumar, S.; Hoover, R.; McFarland, D.; Jurafsky, D. Measuring the evolution of a scientific field through citation frames. *Trans. Assoc. Comput. Linguist.* **2018**, *6*, 391–406. [CrossRef]
49. Li, M.; Cui, L.; Huang, S.; Wei, F.; Zhou, M.; Li, Z. Tablebank: Table benchmark for image-based table detection and recognition. In Proceedings of the 12th Language Resources and Evaluation Conference, Marseille, France, 11–16 May 2020; pp. 1918–1925.
50. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Doll, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 740–755. [CrossRef]

51. KoELECTRA: Pretrained ELECTRA Model for Korean. Available online: https://github.com/monologg/KoELECTRA (accessed on 17 December 2021).
52. DETR GitHub. Available online: //github.com/facebookresearch/detr (accessed on 10 December 2021).
53. ISTR GitHub. Available online: https://github.com/hujiecpp/ISTR (accessed on 10 December 2021).
54. Coco Dataset Detection Eval. Available online: https://cocodataset.org/#detection-eval (accessed on 21 December 2021).
55. Chen, W.; Chang, M.; Schlinger, E.; Wang, W.Y.; Cohen, W.W. Open Question Answering over Tables and Text. In Proceedings of the International Conference on Learning Representations, Vienna, Austria, 4–8 May 2021.
56. Geva, M.; Gupta, A.; Berant, J. Injecting Numerical Reasoning Skills into Language Models. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 946–958. [CrossRef]