

LAME: Layout-Aware Metadata Extraction Approach for Research Articles

Jongyun Choi¹, Hyesoo Kong², Hwamook Yoon², Heungseon Oh³ and Yuchul Jung^{1,*}

¹Department of Computer Engineering, Kumoh National Institute of Technology (KIT), Gumi, Korea ²Korea Institute of Science and Technology Information (KISTI), Daejeon, Korea ³School of Computer Science and Engineering, Korea University of Technology and Education (KOREATECH), Cheonan, Korea *Corresponding Author: Yuchul Jung. Email: jyc@kumoh.ac.kr Received: 02 December 2021; Accepted: 27 January 2022

> Abstract: The volume of academic literature, such as academic conference papers and journals, has increased rapidly worldwide, and research on metadata extraction is ongoing. However, high-performing metadata extraction is still challenging due to diverse layout formats according to journal publishers. To accommodate the diversity of the layouts of academic journals, we propose a novel LAyout-aware Metadata Extraction (LAME) framework equipped with the three characteristics (e.g., design of automatic layout analysis, construction of a large meta-data training set, and implementation of metadata extractor). In the framework, we designed an automatic layout analysis using PDFMiner. Based on the layout analysis, a large volume of metadata-separated training data, including the title, abstract, author name, author affiliated organization, and keywords, were automatically extracted. Moreover, we constructed a pre-trained model, Lavout-MetaBERT, to extract the metadata from academic journals with varying layout formats. The experimental results with our metadata extractor exhibited robust performance (Macro-F1, 93.27%) in metadata extraction for unseen journals with different layout formats.

> **Keywords:** Automatic layout analysis; layout-MetaBERT; metadata extraction; research article

1 Introduction

With the development of science and technology, the number of related academic papers distributed periodically worldwide has reached more than several hundred thousand. However, their layout styles are as diverse as their subjects and publishers, although the portable document format (PDF) is widely used globally as a standardized text-based document provision format. For example, the information order is inconsistent when converting such a document to text because no layout information separating the document content is provided. Thus, extracting meaningful information such as metadata, including title, author names, affiliations, abstracts, and keywords, from a document is quite challenging.



This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Research on extracting metadata or document objects from PDF documents using machine learning has increased [1–7]. In aspects of natural language processing (NLP) approach, Open-source software, such as Content ExtRactor and MINEr (CERMINE) [4] and GeneRation of Bibliographic Data (GROBID) [5], automatically extract metadata using the sequential labeling technique but generally do not take the layouts into account in detail. Therefore, they do not show reasonable metadata extraction performances for every research article due to their diverse (and sometimes bizarre) layout formats.

Unlike existing NLP based metadata extraction approaches, PubLayNet [1], LayoutLM [7], and DocBank [3] employ object detection models, such as Mask region-based convolutional neural network (Mask R-CNN) [8] and Faster R-CNN [9], to detect the layout of academic literature and extract document objects (e.g., text areas, figures, tables, titles, lists, and so on). For example, the PubLayNet-based model detects the layouts of PubMed papers well. However, when other documents that do not appear in training data are given, it omits some information fields or fails to extract the correct regions of documents objects, as shown in Fig. 1. In addition, Fig. 1a shows inconsistent extraction results, such as missing author information or detecting a two-column layout as a one-column layout. In the case of Fig. 1b, the model successfully extracts the layout of the English title but fails to detect that of the Korean title. Moreover, it generates strange detections in the middle of the English abstraction layout. To mitigate the symptoms in the layout detection task, we decided to construct a pre-trained model from various documents with various layouts.

In terms of training data and its coverage, PubLayNet and LayoutLM automatically construct the training data using the metadata provided by PubMed Central Open Access-eXtensible Markup Language (PMCOA-XML) or LaTex. Nevertheless, these are primarily for extracting figures and tables; they do not cover all the necessary metadata, such as the abstract, author name, keyword, or other data [1]. Moreover, to the best of our knowledge, the PMCOA-XML data of papers are only limited to biomedical journals, and small numbers of LaTex data are available in the public domain. Recently, some training data for metadata extraction with consideration of the layout for the selected 40 Korean scientific journals were manually crafted [6]. However, its layout-aware data quality is not so satisfactory due to inconsistent and noisy annotations.

To guarantee consistent annotation quality in constructing layout-aware training data and a more sophisticated language model for advanced metadata extraction, we propose a LAyout-aware MEtadata extraction (LAME) framework composed of three key components. First, an automatic layout analysis for metadata is designed with PDF information extracted by python library PDFMiner¹. Second, a large amount of layout-aware metadata is automatically constructed by analyzing the first page of papers in selected journals. Finally, we implemented a metadata extractor that contains Layout-aware Bidirectional Encoder Representations from Transformers for Metadata (Layout-MetaBERT) models that follow the BERT architecture [10].

In addition, to show the effectiveness of the Layout-MetaBERT models, we performed a set of experiments with other existing pre-trained models and compared them with the state-of-the-art (SOTA) model (i.e., bidirectional gated recurrent units and conditional random field (Bi-GRU-CRF)) for metadata extraction.

Our main contributions are as follows:

• We proposed an automatic layout analysis method that doesn't require PMCOA-XML (or Latex) data for metadata extraction.

¹https://github.com/pdfminer/pdfminer.six

- We automatically generated training data for the layout-aware metadata from 70 research journals (65,007 PDF documents).
- We constructed a new pre-trained language model, Layout-MetaBERT, to deal with the metadata of research articles having varying layouts.
- We demonstrated the effectiveness of Layout-MetaBERT on ten unseen research journals (13,300 PDF documents) with diverse layouts compared with the existing SOTA model (i.e., Bi-GRU-CRF).

	과학 기술 연구 동향 파악을 위한 연구방법론의 문장 분류 [Sentence Classification for Scientific Research Trends Analysis]		Journal of the Korea Academia-Industrial cooperation Society Vol. X., No. X, pp, 2019
	최중윤	경유철	
 Approx 100 minute of the field of the field	CHOIJONG YUN		국가 과학기술 표준분류 체계 기반 연구보고서 문서의 자동 분류 연구
<text></text>	Rumoh National Institute of Technology	Kumph National Institute of Technology	최종윤', 정유철², 한혁³
<text></text>	Sumpor Huld Braver come	wellkumph as k	'금오공과대학교 컴퓨터공학과, '금오공과대학교 컴퓨터공학과, '한국과학기술정보연구원
<text></text>	요약문	0 97% 법을 원구방법론 추출도메인에서의 적용을	Research on Text Classification of Research Reports using Korea
	0 100% 충판되는 과학기승논문에서 연구방법론	시작으로, 딥러닝 계열의 다양한 실험조합을 실험	National Science and Technology Standards Classification Codes
<text></text>	을 자동을 파악하는 것은 매우 중요한 태스크이	하였다. 이미 족들게 활용되고 있는 열정 주문의 성능을 낸다고 알려진 SVM(Support Vector	Jong-Vun Choi ¹ Vu-Chul Jung ⁴ Huck Han
<text><text><text><text><text><text><text><text><text></text></text></text></text></text></text></text></text></text>	다. 본 연구에서는 한글논문 초록에 존재하는 연 구방법로을 관려 문장들(목적 문제, 실험방법, 실	Machine)계열의 알고리즘인 SGD(Stochastic	2032
Right Signed Read Big Name Const Citik Const Ci	험결과, 결론, 기대효과 등)을 자동으로 분류하기	Gradient Descent) [1]에서는 69.21%의 정확도 를 확인하였다. 전반적으로 딥러닝 계열의 알고리	26003 Computer engineering, Kumoh national institute of technology
Image: Ima	위해 텍스트 분류에 사용되는 대표적인 5개의 알 고리중등(SGD, CNN, LSTM, CNN + Bir	름이 좋은 성능을 보였으며, CNN+GRU기법 에서	Korea Institute of Science and Technology Information
법률 여도 (이용 등 이용 전·아이용 (유용으로) 정말 데스트 (이용 등 이용 전·아이용 (유용으로) 정말 네스트 (이용 등 이용 전·아이용 (무료) (유용으로) (유용으로) (유용 (무료)) (유용 (무료))	directional GRU, 그리고 CNN+GRU)의 실험을	은 78.83%의 정확도를 확인하였다.	0 100% 화기술 분야의 연구개발 결과는 연구보고서 형태로 국가과학기술정보서비스(NTIS)에 제출된다. 각 연구보고
환환 실험에서 78.83%의 정확도를 확인해있다. 문화 실험에서 78.83%의 정확도를 확인해있다. 무화 성치 기용 무화 성치 기용 무화 성치 기용 무화 성치 기용 무화 성치 기용 무화 성치 기용 무화 성치 기용 무화 성치 기용 무화 성치 기용 무화 성치 기용 무화 성치 기용 무화 성치 기용 <tr< th=""><th>진행하였다. 특히, IT 및 화학분야의 초록으로 구 성된 테스트 데이터를 이용하 CNN+GRU의 원책</th><th>2 관려 연구</th><th>저는 국가과막기울 표준 분류체계 (K-NSCC)에 따른 분류코드를 가지고 있는데, 보고서 작성자가 제울 시에 수동으로 입력하게끔 되어있다. 하지만 2000며 개가 넘는 세분류를 가지고 있기에, 분류체계에 대한 정확한 이해가 없이는 부정</th></tr<>	진행하였다. 특히, IT 및 화학분야의 초록으로 구 성된 테스트 데이터를 이용하 CNN+GRU의 원책	2 관려 연구	저는 국가과막기울 표준 분류체계 (K-NSCC)에 따른 분류코드를 가지고 있는데, 보고서 작성자가 제울 시에 수동으로 입력하게끔 되어있다. 하지만 2000며 개가 넘는 세분류를 가지고 있기에, 분류체계에 대한 정확한 이해가 없이는 부정
Particle 관리 이 문 전 문 가 가지 않는 것이 수 있는 것이 이 가 수 있는 것이 수 있는 것이 수 있는 것이 이 가 수 있는 것이 수 있 수 있는 것이 수 있는	조합 실험에서 78.83%의 정확도를 확인하였다/	80년 0.99월 해외논문에 대해서는 연구개발의 동형	확한 분류코드를 선택하기 십상이다. 새로이 수집되는 연구보고서의 양과 다양성을 고려해 볼 때, 이들을 기계적으로 보
방법 전화 시도가 없었다. 또한 논문 뿐만 아니며 확실 가방법은, 고학가술는은, 초록, 문상분류, 기지 14/9/44 또한 핵심 가위트를 추용가기 위해 가방법은, 고학가술는은, 초록, 문상분류, 기지 14/9/44 또한 핵심 가위트를 추용가기 위해 가방법은, 고학가술는은, 초록, 문상분류, 기지 14/9/44 또한 핵심 가위트를 추용가기 위해 가방법은, 고학가술는은, 초록, 문상분류, 기지 14/9/44 또한 핵심 가위트를 추용하기 위해 가방법은, 고학가술는은, 초록, 가능한, 기지 14/9/44 또한 핵심 가위트를 통한 효율적인 분류를 위해 가방법은, 기술은 가방법은 추용하기 위해 19/9/10/10/10/10/10/10/10/10/10/10/10/10/10/		을 파악하기 위하여 초록에서 추출하기 위해 다	다 정확하게 분류할 수 있다면 보고서 제출자의 수고를 덜어줄 수 있을 뿐만 아니라, 다른 부가 가지적인 분석 서비스 들과의 연계가 수월할 것이다. 하지만 국내에서 과학기술표준 분류체계에 기반을 둔 문서 자동 분류 연구 사례는 거의
전부방법은, 관측, 운방분류, 가자 방문, 요리님 지원 ····································	52.0	양한 시도가 있었다. 또한 논문 뿐만 아니라 특허 나 저녁에서 또한 해식 키워드를 추축하기 위해	없으며 공개된 학습데이터도 전무하다. 본 연구는 KISTI가 보유하고 있는 최근 5년간 (2013년~2017년) NTIS 연구보고
비원이었다 비원이 있다 비원이었다 비원이 있다 비용 비원이 있다 비용	면구방법론, 과학기술논문, 초록, 문장분류, 기계	면구가 진행되었으나 대부분의 연구는 영어권에	서 메타정보를 활용한 최초의 시도로써, 방대한 과학기술표준 분류체계를 기반으로 하는 국내 연구보고서들을 대상으로 높은 서누은 보이는 문서 자도 보르기번은 도출하는 여구를 지해하여다. 이를 의해 과하기술 표준보로 체계에서 과하기
서로 비용의 적용 및 문의 재명(재용)과 키위드 및 용의 및 위도 및 유명 및 위도 및 유명 및 유	학습, 딥러닝	서 진행되었으며 한글에서는 아직까지 연구가 횁 박하게 지해되지 않고 있다.	을 분야의 연구보고서를 분류하기에 적합한 중분류 210여 개를 선별하였으며, 연구보고서 메타 데이터의 특성을 고려한
 비용의 지수의 유럽 사용의 취직 지수 지수		·····································	전처리를 진행하였다. 특히, 가장 영향력 있는 필드인 과제영(제목)과 키워드만을 이용한 TK_CNN 기반의 답러닝 기법을 해야하다. 제아, 모래은 태스트, 분류에서 좋은 성능은 날이고, 있는 기계하수면든 100 iterate OVC CNN, CPU들)과 비고
Example 다양한 연구방법은의 지려자기 위한 지 Example 다양한 연구방법은의 지려자기 위한 지 PC 매일 사용이 출판되는 가락기을 문화들다 APS 관계 시용하는 연구가 건경되었으며, [3]에서는 백스트를 처리할 미 PNA국 CNN을 함께 사용하는 연구가 건경되었으며, [3]에서는 백스트를 처리할 미 PNA국 CNN을 함께 사용하는 연구가 건경되었으며, [3]에서는 백스트를 처리할 미 PNA국 CNN을 함께 사용하는 연구가 건경되었으며, [3]에서는 백스트를 처리할 미 PNA국 CNN을 함께 사용하는 연구가 건경되었으며, [3]에서는 백스트를 처리할 미 PNA국 CNN을 함께 사용하는 연구가 건경되었으며, [3]에서는 백스트를 처리할 미 PNA국 CNN을 함께 사용하는 연구가 건경되었으며, [3]에서는 백스트를 처리할 미 PNA국 CNN을 함께 사용하는 연구가 건경되었으며, [3]에서는 백스트를 처리할 미 PNA국 CNN을 함께 사용하는 연구가 건경되었으며, [3]에서는 백스트를 처리할 미 PAS 한 연구방법은의 자용하는 경우, [3]에서는 백스트를 추운 한 위법 - 전용 또 아이지	KIPE HINE COLODAR	해 다양한 연구가 진행되었다 [2]에서는 텍스트	하였으며, Top-3 F1점수 기준으로 1~7%에 이르는 성능 우위를 확인하였다.
지하는 입구방법을 것을 이 관련 법 연구 이 가장 2 분들 등 이 적으 표 (NNA CNNE 함께 사용하는 연구가 진행되었으며 등 대체에는 데 나는 National Science and Technology Information Server (NTES) In the form of research regores that a classification codes according to Korea national science and technology stand classification codes (K-NSCC). In addition, there are few cases of another the Science and Technology stand classification codes without class under the NSCC in addition, there are few cases of another the Science and Technology stand classification codes without class under the NSCC in addition, there are few cases of another the Science and Technology stand classification codes without class under the NSCC in addition, there are few cases of another the Science and Technology stand classification codes without class under the NSCC in addition, there are few cases of another the Science and Technology stand classification codes without class under the NSCC in addition, there are few cases of another the Science and Technology stand classification codes without classification codes withou	생각본전에서는 한글로 기술된 과학기술 텍스트에 존재하는 다양화 여구방법류를 최러하기 위하 기	를 분류할 때 RNN이 아닌 CNN을 통한 연구가 지해되었으며 [3]에서는 텍스트를 최리하 때	0 93% In South Korea, the results of R&D (Research and Development) in the field of science and technology
보다. 매일 사로이 출연되는 계약기술 문화들과 수동으로 도위신 전문가에 위해 본석하는 것은 비용이 매우 크며, 개용으로 분석하는 것은 비용이 매우 크며, 가용으로 분석하는 것을 받 수술을 위한 여기 비용이 매우 크며, 가용으로 분석하는 것을 받 수술을 위한 여기 비용이 매우 크며, 가용으로 분석하는 것을 받 보다 분수에 있어서 어떤 기법이 가 도 성용되다. 매리서, 크릭가 유용으로 분석하는 것을 받 수술을 위한 여기 지하는 연구방법문과 자용으로 분석하며 있어서 여편 기법이 가 가용 연구방법문과 자용으로 분석하며 있어서 이편, 가장이 있다. 이가 한 가용 수술을 위한 여기 가용 연구방법문과 자용으로 분수에 있어서 이편, 가장이 있다. (5-7) 가용 연구방법문 관측을 위해 여기 가용 연구방법문 관측을 위해 여기 가용 연구방법문 가용	계학습기반의 문장분류기술에 관련된 연구를 다	RNN과 CNN을 함께 사용하는 연구가 진행되었	are submined to the National Science and Technology information Service (NTIS) in the form of research reports,
 비용이 팩우 크며, 지종으로 분석하는 것도 한 변수를 운전하며, 가격은 것도 한 변수를 분류에 있어서 다양한 가법 등 통해 변수로 분류에 있어서 다양한 가법 등 분류에 있어서 다양한 가법 등 통해 변수로 분류에 있어서 다양한 가법 등 방법 등 분해 물 통해 한 기법 수 관점 등 분위 한 약을 통해 한 기법 등 관점 등 통해 기원 등 관점 등 통해 기원 등 분류를 위한 약을 가장해 있다. 15-27 보 전 구성 법원 · 관점 및 하는 예구가 법원 · 관점 및 하는 예구가 법원 · 관점 및 하는 예구가 법원 · 관점 및 하는 예구 법원 · 관점 및 · 관점	문다. 매일 새로이 출판되는 과학기술 문헌들을 수도으로 도미이 권무가에 의해 부석하는 것은	으며 함께 사용했을 때 독립적으로 사용하는 것 보다 좋은 효율을 나타내는 것을 보여주었으며	classification codes (K-NSCC). However, considering more than 2000 sub-categories, it is non-trivial to choose correct
한 별 관광 형식의 다양정으로 모이망이 그 선이 도는 상당하다 미리세, 격학기술 문화들 에서 적 시하는 연구방법문과 사용으로 추용하기 위부 수 가방법문의 타그 경계, 연구 방법문 ·추용을 위한 안동 또 나주 진행하나 있다. [5-7] 관 입장 바 하는 연구방법문 관람 정 하는 데이터를 구축, 그리고 연구방법문 전문 정부를 정 하는데 의력하에 이트는 연구금 신행되었다. 관 연구에서는 기존에 잘 알려진 SVM제열 기 (a)	그 비용이 매우 크며, 자동으로 분석하는 것도 문	[4]에서 또한 텍스트 분류에 있어서 다양한 기법	classification codes without clear understanding of the K-NSCC. In addition, there are few cases of automatic
시체는 연구방법론을 치용으로 추용하기 위해 연 구방법문의 태그 경제, 연구방법은 추용을 위한 달광 법 학습 데이터를 구축, 그리고 연구방법은 관람 관람 취상에 이르는 연구 전문 전문은 문문을 하는 연구 전문 전문은 문문을 하는 연구 전문 전문은 문문을 하는 연구 전문 전문을 하는 연구 전문 전문 전문 전문 전문 전문 전문 전문 전문 전문 전	한 별 문장 형식의 다양성으로 말미맘아 그 난이 도는 상당하다 따라서 과학기술 문화들 에서 제	들을 통해 텍스트 분류에 있어서 어떤 기법이 가 장 효율적이가에 대해 여구하 기로이 있다. 이외	accument classification research based on the K-NSCC, and there are no training data in public domain. To our best show ledge, this study is the first attempt to build highly performing K-NSCC classification system based on the NTIS
구분법원의 태그 설립, 연구 방법은 추출을 위한 달경 별 학습 데이터를 구축, 그리고 연구방법은 관련 공방류 성능의 최적하에 이르는 연구 전별하였다. 본 연구에서는 기존에 잘 알려진 SVM계열 기 (a) (b)	시하는 연구방법론을 자동으로 추출하기 위해 연	에도 텍스트 분류를 통한 키워드 추출을 위한 연	reports meta-information of the last five years (2013-2017). To this ends, about 210 mid-level categories were selected
관련 문경문류 성용계 최적화에 이르는 연구를 사망하였다. 3. 제안 방법 as compared with several machine learning methods (e.g., Linear SVC, CNN, GRU, etc.) that show good performa to text classification, and has a performance advantage of 1-7% based on the Top-3 FL score 본 연구에서는 기존에 잘 알려진 SVM계열 기	구방법론의 태그 경의, 연구 방법론 추출을 위한 문장 별 한숨 데이터를 구축 그리고 연구방법로	구들도 다수 진행된바 있다. [5-7]	and conducted preprocessing considering the characteristics of research report metadata. More specifically, we propose a CNN technique using only task names and keywords, which are the most influential fields. The proposed model
전 행정 전 · · · · · · · · · · · · · · · · · ·	관련 문장분류 성능의 최적화에 이르는 연구를		is compared with several machine learning methods (e.g., Linear SVC, CNN, GRU, etc.) that show good performance
은 면구에서는 기존에 잘 알려진 SVM제절 기 (a) (b)		3. 제안 방법	Kewords ' Deer Jamma Text classification Research report Promotions NTR-
(a) (b)	본 연구에서는 기존에 잘 알려진 SVM계열 기	3.1. 사용한 소프트웨어들	0.66%
	((a)	(b)

Figure 1: Layout analysis results of PubLayNet [1]

2 Related Work

2.1 Metadata Extraction

Various attempts have been made to analyze and extract information from documents and classify them into specific categories. Studies on text classification have been continuous since 1990, and the performance of text classification has gradually improved with the employment of sophisticated machine learning algorithms, such as the support vector machine (SVM) [11], conditional random fields (CRF) [12], convolutional neural network (CNN) [13], and bidirectional long short-term memory (BiLSTM) [14]. Afterward, various successful cases using Bidirectional Encoder Representations

from Transformers (BERT) [10] pre-trained with a large-scale corpus were introduced in the field of NLP. In the studies by [15,16], the pre-trained BERT model was fine-tuned on the text classification task, and it showed results close to or superior to the SOTA result for the target data. BERT-based pre-trained models became popular due to their high performances in various NLP fields, and more advanced pre-trained models [17–19] were introduced according to various research purposes.

As a previous SOTA model of our metadata extraction task, a Bi-GRU-CRF model trained more than 20,000 human-annotated pages of layout boxes for metadata [6] from research articles showed an 82.46% of F1-score. However, accurately detecting and extracting regions for each type of metadata in documents is still a nontrivial task because of the various layout formats.

2.2 Document Layout Analysis

Document layout analysis (DLA) [7] and several PDF handling efforts [6,11,20] have been conducted to understand the structure of documents. The DLA aims to identify the layout of text and nontext objects on the page and detect the layout function and format. Recently, the LayoutLM model [7] employed three different information elements for BERT pre-training to identify layouts: 1) layout coordinates, 2) text extracted using optical character recognition software, and 3) image embedding by understanding the layout structure through image processing. Moreover, NLP-based DLA research on various web documents [21], Layout detections and layout creation methods to find text information and location [8,22,23] have been studied. In [2,24] applied the object detection technique to text region detection. Interestingly, widely used object detection techniques (e.g., Mask R-CNN [8] and Faster R-CNN [9]) have been applied to the metadata extraction field [1,3].

Due to the high cost of training data construction for DLA, many studies have attempted to build datasets automatically. For example, the PubMed Central website, which includes academic documents in the biomedical field, provides a PMCOA-XML file for each document, enabling an analysis of the document structure. In the case of PubLayNet [1], which utilizes the PubMed dataset, the XML and PDFMiner's TextBoxes were matched to construct about 1 million training data. However, this is generally possible only when accurate coordinates are provided to separate each layout and the text information elements for each.

3 Proposed Framework

Fig. 2 depicts our LAME framework consisting of three major components: automatic layout analysis, layout-aware training data construction, and metadata extractor. Stage 1 analyzes the given PDF's first-page layout by using PDFMiner. However, due to the incompleteness of the parsing results of the PDFMiner, it undergoes a set of reconstruction, refinement, and adjustment procedures for identifying several metadata that appeared on the first page. In stage 2, we build many training data used in stage 3. The building process matches the identified metadata in stage 1 with the previously existing correct metadata values. Finally, we implement a novel metadata extractor by pre-training our Layout-MetaBERT model with the training data of stage 2 and fine-tuning it for target corpus.

3.1 Automatic Layout Analysis

To understand the layout that separates each metadata element in the given PDF file, we must observe the text and coordinate information on the document's first page. To this end, we employ the open-source software, PDFMiner, to extract meaningful information surrounding the text in the PDF files. For example, if we parse a PDF document with the software, we obtain information on the page, TextBox, TextLine, and character (Char) hierarchically, as illustrated in Fig. 3. These include

various text information, such as coordinates, text, font size, and font for each object. For example, text coordinates appear in the form of (x, y) coordinates along with the height and width of the page.



Figure 2: Layout-Aware Metadata Extraction (LAME) framework



Figure 3: TextBox reconstruction based on the results of PDFMiner

3.1.1 Textbox Reconstruction

It is possible to extract document layout by utilizing PDFminer, but this conveys erroneous information, which is hard to use directly. Therefore, to reduce existing errors in TextBox recognition, as depicted in Fig. 3, TextBoxes were reconstructed starting from the Char unit with the information obtained from the PDFMiner. The detailed steps are as follows:

- 1) Extract Token Information: First, we extract all characters that appear on the first page of the PDF document and their information.
- 2) Characters to TextLine: Second, the spacing between characters is analyzed using the coordinate information for each Char. Generally, each token's x-coordinate distance (i.e., character spacing) appears the same, but the distance is slightly different depending on alignment methods or language. Therefore, after collecting characters in the same y coordinate, the corresponding characters were sorted based on the x-coordinate value. Sometimes academic papers have two columns, so different lines may exist in the same Y coordinate. If the distance between two characters is smaller than the font size, we regarded them as one line.
- 3) TextLines to TextBox: Finally, after aligning the TextLines based on the y-axis, if the distance between each y-coordinate is smaller than the height of each TextLine, the two different TextLines are regarded as the same TextBox. However, this method cannot create a TextBox accurately by separating paragraphs from paragraphs. For more elaborate TextBox composition, it needs to decide whether to configure the TextBox by considering the left x-coordinate x₀, the right x-coordinate x₁, and the width (W) of each TextLine. For example, for sentences like those in Fig. 4, we can think of two cases composing a TextBox by comparing each TextLine. First, the beginning of a paragraph is usually indented. Therefore, if the difference between the x₀ values of L_i and L_{i-1} is greater than the font size of the Chars existing in each TextLine, the two TextLines should be included in different TextBoxes. Second, a TextLine that appears at the end of a paragraph has a shorter width because it has fewer Chars on average. Therefore, when the width of L_{i-1} is smaller than the width of L_i, L_{i-1} and L_i should be assigned different TextBoxes.



Figure 4: Example of TextBox separation

3.1.2 Refinement with Font Information

PDFMiner can produce various pieces of information in terms of font information, such as the font name and style (e.g., bold, italic, etc.) as listed in Tab. 1.

However, as in Fig. 5, English is frequently used for Korean abstracts in some journals published in Korea. In particular, abstracts written in Korean and English appear together on the first page of some research articles. In addition, certain strings are often treated as bold or italic and often have different fonts and sizes, such as section titles. Considering this problem, when composing a TextBox using the coordinate information described above, if the font information displayed on each line is different, it is not simply judged as a different line. After analyzing the font information of different languages that appear, the TextBox was determined by considering the number of the appearing fonts (e.g., bold and italic).

Text	PDFMiner output font name		
Abstract	ELNFKM+KoreanGD-Bold-KSCpc-EUC-H		
Abstract	INPILL+Gulim		
Abstract	Arial-BoldItalicMT		
ABSTRACT	GPJCIE+YDIYGO130		

Table 1: Example of font information when PDFMiner is applied

요 약

최근, 비대면 경험 및 서비스에 관한 관심이 증가하면서 스마트폰이나 태블릿과 같은 모바일 기기를 이용하여 손 쉽게 이용할 수 있는 웹 동영상 콘텐츠에 대한 수요가 급격히 증가하고 있다. 이와 같은 요구사항에 대응하기 위하여, 본 논문에서는 애니메이션이나 영화에 등장하는 명소를 방문하는 무대 탐방 경험을 제공할 수 있는 영상 콘텐츠를 보다 효율적으로 제작하기 위한 기법을 제안한다. 이를 위하여, Google Maps와 Google Street View API를 이용하여 무대탐방 지역에 해당하는 이미지를 수집하여 이미지 데이터셋을 구축하였다. 그 후, 딥러닝 기반의 style transfer 기 술을 접목시켜 애니메이션의 독특한 화풍을 실사 이미지에 적용한 후 동영상화하기 위한 방법을 제시하였다. 마지막 으로, 다양한 실험을 통해 제안하는 기법을 이용하여 보다 재미있고 흥미로운 형태의 무대탐방 영상 콘텐츠를 생성 할 수 있음을 보였다.

ABSTRACT

Recently, as interest in non-face-to-face experiences and services increases, the demand for web video contents that can be easily consumed using mobile devices such as smartphones or tablets is rapidly increasing. To cope with these requirements, in this paper we propose a technique to efficiently produce video contents that can provide experience of visiting famous places (i.e., stage tour) in animation or movies. To this end, an image dataset was established by collecting images of stage areas using Google Maps and Google Street View APIs. Afterwards, a deep learning-based style transfer method to apply the unique style of animation videos to the collected street view images and generate the video contents from the style-transferred images was presented. Finally, we showed that the proposed method could produce more interesting stage-tour video contents through various experiments.

Figure 5: Example of when a Korean abstract and an English abstract exist together

Although font information helps the layout composition, it is still confusing when the same font information is used for individual information marking or bold processing for emphasis or different metadata. Additional processing is required to correctly connect individual fonts to make a layout using the font information. Therefore, we compared only texts described in Korean and English and used only the fonts of the same language to determine the layout.

3.1.3 Adjustment of Text Box Order

Academic papers may consist of one or two columns depending on the format for each journal. In some cases, only the main body consists of two columns, and the title, abstract, and author name are displayed in one column. For example, in Fig. 3, such information as the title and author name was arranged in the center, but the document object identifier (DOI) information or academic journal names appeared separately on the left and right sides. To effectively identify metadata consistently from varied layout formats, we sorted the textboxes extracted from the first pages of the research articles sequentially from top to bottom based on the y-axis.

3.2 Automatic Training Data Construction

We compared the content from the layout text that extracted stage.1 with the metadata prepared in advance to construct the layout-aware metadata automatically. If no metadata is available for the given research article, metadata can be automatically obtained through the DOI lookup. Therefore, this technique can be extended to all journal types where the registered DOI exists.

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.

The Levenshtein distance was calculated using Python's fuzzywuzzy². The scores calculated using the BLEU [25] measure were summed to determine whether the given metadata displays a degree of agreement of 80% or more. Nevertheless, some post-processing is required in the process. In analyzing the text after extraction, some problems occur when dealing with expression substitutions (e.g., "<TEX>," cid:0000). Encoding errors reduced the portion of mathematical expressions that can be removed as much as possible, and we excluded the text with encoding problems to avoid these errors.

3.3 Metadata Extractor

To implement our metadata extractor, we newly pre-trained a layout-aware language model, socalled, Layout-MetaBERT, that can effectively deal with metadata from research articles. Although pre-training a BERT model requires a large corpus and a long training time, a fine-tuning step can make a difference in performance depending on the characteristics of the data used for pre-training. For example, when pre-trained with specific domain data, such as SciBERT [19] and BioBERT [26], they performed better than Google's BERT model [10] in downstream tasks of science and technology or medical fields. However, to our best knowledge, there is no pre-trained model designed to extract metadata based on research article data.

Fig. 6 describes how the previously constructed training data are used for pre-training and finetuning the Layout-MetaBERT. The pre-training stage is to construct a general-purpose layout-aware metadata language model. Meanwhile, the fine-tuning stage aims to build an optimized metadata classification model for targeting corpus. Unlike the Google BERT model [10], in our Layout-MetaBERT pre-training, each document layout was considered a sequence. Thus, each layout was classified by the [SEP] token to prepare the training data, but other pre-training procedures and hyperparameters are the same as BERT [10]. At this time, pre-training loss is the sum of Next Sentence Prediction (NSP) loss and Masked Language Modeling (LM) loss. The Masked LM Loss computes

²https://github.com/seatgeek/fuzzywuzzy

the cross-entropy loss value of the predicted token y_{ii} for the masked token \hat{y}_{ij} :

$$L_{MLM} = -\frac{1}{N} \sum_{j=1}^{N} \sum_{i=1}^{C_{vocabsize}} y_{ij} log(\hat{y}_{ij})$$
(1)



Figure 6: Pretraining layout-MetaBERT and fine-tuning for classification downstream task

Previously, the NSP was used to predict whether two input sentences are consecutive sentences or not. However, in this work, the NSP predicts whether two inputted layouts are consecutive or not. The NSP loss computes the binary cross-entropy loss value of the predicted \hat{y}_i for the y value indicating whether it is a pair or not.

$$L_{NSP} = -\frac{1}{N} \sum_{i=1}^{N} y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)$$
Final loss is:
(2)

$$L = L_{MLM} + L_{NSP} \tag{3}$$

In pretraining Layout-MetaBERT models, we followed three size models of the Google BERT: base (L = 12, H = 768, A = 12), small (L = 4, H = 512, A = 8), and tiny (L = 2, H = 128, A = 2), where L is the number transformer blocks, H is the hidden size, and A is the number of self-attention heads. We used a dictionary of 10,000 words built through the WordPiece [27] mechanism. We automatically generated training data extracted from the first page of 60 research journals among the 70 journals for the pre-training.

In the Fine-tuning step, each Layout text information is used as input. For the classification task, we feed the final LayoutMeta-BERT vector for the [CLS] token into a Fully connected layer, and it

computed softmax we can obtain meta classification result. The prediction result loss computes the cross-entropy, and the prediction accuracy computes Macro and Micro F1-score.

4 Experiments

We summarize the results of three major components to examine the applicability of the LAME framework. First, we compare the proposed automatic layout analysis results with other layout analysis techniques. Second, we describe the statistics of the training data constructed according to the results of the automatic layout analysis. Finally, we compare metadata extraction performances of our constructed Layout-MetaBERT models with other deep learning and machine learning techniques after fine-tuning.

4.1 Comparison with Other Layout Analysis Methods

No correct answers exist for the target research articles; thus, we compared the generated layout boxes from PDFMiner, PubLayNet, and the proposed layout analysis method for two randomly selected documents (e.g., A and B) as depicted in Fig. 7.

In Fig. 7, when PDFMiner performs a layout analysis, it generates multiple boxes for abstract and each section (A) and extracts empty layouts between sections as errors (B). Sometimes, it cannot separate keywords from the abstract area (B). Therefore, it is challenging to build training data if we use PDFMiner directly without further modifications.

In the case of Detectron2 based model, which is trained with PubLayNet data, it extracts document objects using object detection technique. So, building training data is impossible because text information cannot be obtained if optical character recognition software is not provided. Although it seems almost similar to ours, it easily fails to detect some metadata elements (A) because PubLayNet does not have metadata fields (e.g., author, affiliation, keywords) as training data. They only cover PubMed papers with almost no citations similar layout formats. Thus, when other documents that do not appear in training data are given, it omits some information fields or fails to extract the correct regions of documents objects (B), and shows inconsistent extraction results, such as missing author information or detecting a two-column layout as a one-column layout (B).

The proposed method could generate a good enough layout analysis for the first page of the research articles through the comparisons. Comparing all three layout analysis results manually for each layout box to calculate the accuracy requires too much human labor and is beyond the scope of this paper. The performance of the constructed Layout-MetaBERT indirectly measures the quality of the layout analysis.

4.2 Training Data Construction

To reflect various kinds of layout formats, we used 70 research journals (Appendix 1) provided by the Korea Institute of Science and Technology Information (KISTI) to extract 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 journals, two journals were written in only Korean, 23 journals in only English, and 45 in Korean and English.

For each layout that separates metadata on the first page of the 70 journals (65,007 PDF documents), automatic labeling with ten labels was performed, and other layouts not included in the relevant information were labeled O. The statistics of automatically generated training data are presented in Tab. 2.



Figure 7: Layout analysis comparisons with PDFMiner and PubLayNet

4.3 Experimental Results

To check the performance of the proposed Layout-MetaBERT, 70 research journals (65,007 documents) were divided into 60 (51,676 documents) for pre-training (and fine-tuning) and 10 (13,331 documents) for testing, respectively. Tab. 3 lists the training and testing performances of the three Layout-MetaBERT models with widely used metadata extraction techniques. Finally, Tab. 4 describes the Macro-F1 and Micro-F1 scores for metadata classification comparisons with existing pre-trained models.

4.3.1 Fine-tuning and Hyperparameters

In fine-tuning with various pre-trained language models (e.g., three different sized models of Layout-MetaBERT, KoALBERT, KoELECTRA, and KoBERT), all experiments were conducted under the same configurations with an epoch of 5, batch size of 32, learning rate of 2e-5, and maximum

sequence length of 256. In addition, we used the Nvidia RTX Titan 4-way system and Google's TensorFlow framework in Python 3.6.9 for pre-training and fine-tuning.

Metadata field	Label (i.e., layout)	Count
Out of boundary	0	637,856
Title (in Korean)	title_ko	46,056
Title (in English)	title_en	64,414
Affiliation (in Korean)	org_ko	39,233
Affiliation (in English)	org_en	63,434
Abstract (in Korean)	abstract_ko	31,885
Abstract (in English)	abstract_en	55,318
Keywords (in Korean)	keywords_ko	21,685
Keywords (in English)	keywords_en	61,221
Author name (in Korean)	author_name_ko	56,306
Author name (in English)	author_name_en	35,631

Table 2: Statistics for automatically generated training data

Table 3: Train and test performances of metadata extraction

Models	Micro-F1 (train)	Micro-F1 (test)
Layout-MetaBERT (base)	0.9559	0.936
Layout-MetaBERT (small)	0.9595	0.9333
Layout-MetaBERT (tiny)	0.9432	0.9293
KoBERT ³	0.8086	0.7901
KoalBERT ⁴	0.9014	0.8978
KoELECTRA ⁵	0.9354	0.9204
Bi-GRU-CRF [6] (without position)	0.8610	0.8912
Bi-GRU-CRF [6] (with position)	0.9442	0.0985
CNN [13]	0.9425	0.824
SVM [11]	0.9411	0.8114

4.3.2 Stable Performances of Layout-MetaBERT

The proposed Layout-MetaBERT models can effectively extract metadata, as listed in Tab. 3. In particular, Layout-MetaBERT models make significant differences compared to the existing SOTA (i.e., Bi-GRU-CRF) model.

Even the tiny model with the fewest parameters among the Layout-MetaBERT models has higher performance than other pre-trained models in Macro-F1 and Micro-F1 scores, as displayed in Tab. 4.

³https://github.com/SKTBrain/KoBERT

⁴https://huggingface.co/kykim/albert-kor-base

⁵https://github.com/monologg/KoELECTRA

Moreover, three Layout-MetaBERT models have only minor differences between the Micro-F1 and Macro-F1 scores compared to other pre-trained models. Moreover, the Layout-MetaBERT models exhibit 90% or more robustness in metadata extraction, confirming that pre-training the layout units with the BERT schemes is feasible in the metadata extraction task.

	Layout- MetaBERT (tiny)	Layout- MetaBERT (small)	Layout- MetaBERT (base)	KoBERT	KoALBERT	'KoELECTRA
Model size	5 M	16 M	110 M	110 M	12 M	110 M
0	0.94	0.94	0.95	0.85	0.92	0.94
title_ko	0.92	0.94	0.93	0.59	0.89	0.91
title_en	0.92	0.91	0.92	0.76	0.8	0.87
org_ko	0.96	0.96	0.96	0.36	0.96	0.94
org_en	0.86	0.87	0.9	0.64	0.79	0.9
abstract_ko	0.92	0.93	0.93	0.84	0.9	0.92
abstract_en	0.92	0.93	0.94	0.84	0.91	0.91
keywords_ko	0.94	0.95	0.95	0.86	0.94	0.95
keywords_en	0.87	0.89	0.9	0.51	0.91	0.73
author_name_ko	0.97	0.97	0.96	0.75	0.95	0.92
author_name_en	0.56	0.62	0.92	0.3	0.41	0.57
Micro f1	0.9293	0.9333	0.9360	0.7901	0.8978	0.9204
Macro f1	0.8891	0.9009	0.9327	0.6636	0.8527	0.8691

Table 4: Metadata extraction performances of primary BERTmodels for each label

4.3.3 Experiments with Position Information

Unlike other models, the Bi-GRU-CRF model used the absolute coordinates of metadata with other textual features. However, the model failed to discriminate unseen layouts from unseen journals when using the coordinate information for training various journal layout formats. Therefore, to determine the validity of the coordinate information, we performed additional experiments with the Bi-GRU-CRF (with position) and Bi-GRU-CRF (without position) models. Although Bi-GRU-CRF (with position) model demonstrated high performance in the training stage, it failed to recognize metadata-related layouts in unseen journals (less than 10% as F1 score). However, the performance of the Bi-GRU-CRF (without position) model had somewhat lower performance in the training stage compared to the other models. The model performed well, similar to that of KoALBERT. Thus, we confirmed that using absolute coordinate information can only be applied under the premise that the journals used in training also are used in testing.

5 Discussion

5.1 Additional Performance Improvements

The proposed Layout-MetaBERT exhibited higher results than the existing SOTA model [6]. However, absolute coordinate information could obtain poor results for documents in a format not learned. In addition, the proposed layout analysis method separates the metadata well from the first page of the academic documents of various layouts.

However, the accuracy of the automatically generated training data is not perfect. There may be errors due to the difference between the metadata format of the document and the metadata written in advance. As mentioned, encoding errors also occur in extracting text from mathematical formulas or PDF documents. Generating the correct layout significantly affects extracting metadata and is an essential factor in automatically generating data. Therefore, if more sophisticated training data can be generated, the performance of Layout-MetaBERT can be further improved.

5.2 Restrictions of Layout-MetaBERT

Much research has been conducted on automatically extracting layouts from PDF documents. Creating accurate layouts has a significant influence on meta-extraction. This study attempted to compose the layout of the first page of an academic document using text information. Based on this, we trained the Layout-MetaBERT and confirmed the positive results for the applicability to the meta classification module. However, the proposed technique cannot be applied to all documents. For example, an image-type PDF cannot be used unless the text is extracted. In this case, the extraction must be performed using a high-performance optical character recognition module.

5.3 Expansion to Other Metadata Types

This study focused on extracting five major metadata elements (i.e., titles, abstracts, keywords, author names, and author information). Considering that the target research articles contain elements written in English, Korean, or both, the number of metadata becomes 10. However, other metadata (e.g., publication year, start page, end page, DOI, volume number, journal title, etc.) can be extracted further by applying highly refined regular expressions in the post-processing step.

6 Conclusion

This paper proposes the LAME framework to extract metadata from PDFs of research articles with high performance. First, the automatic layout analysis detects the layout regions where metadata exists regardless of the journal formats based on text features, text coordinates, and font information. Second, by constructing automatic training data, we built high-quality metadata-separated training data for 70 journals (65,007 documents). In addition, our fine-tuned Layout-MetaBERT (base) demonstrated excellent metadata extraction performance (F1 = 94.6%) for even unseen journals with diverse layouts. Moreover, Layout-MetaBERT (tiny) with the fewest parameters exhibited superior performance than other pre-training models, implying that well-separated layouts induce effective metadata extraction when they meet appropriate language models.

In future work, we plan to conduct experiments to determine whether the proposed model applies to the more than 500 other journals not used in this study. Moreover, resolving potential errors in the automatically generated training data is a concern to create layouts that separate each metadata element in an advanced way. Furthermore, extending the number of metadata items extracted without post-processing is an exciting but challenging task to resolve as future work. **Funding Statement:** This work was supported by the Korea Institute of Science and Technology Information (KISTI) through Construction on Science & Technology Content Curation Program (K-20-L01-C01), 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.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

- X. Zhong, J. Tang and A. J. Yepes, "Publaynet: Largest dataset ever for document layout analysis," in *Int. Conf. on Document Analysis and Recognition (ICDAR)*, Sydney, NSW, Australia, IEEE, pp. 1015–1022, 2019.
- [2] L. Melinda, R. Ghanapuram and C. Bhagvati, "Document layout analysis using multigaussian fitting," in *14th IAPR Int. Conf. on Document Analysis and Recognition (ICDAR)*, Kyoto, Japan, vol. 1, pp. 747–752, 2017.
- [3] L. Minghao, X. Yiheng, C. Lei, H. Shaohan, W. Furu *et al.*, "DocBank: A benchmark dataset for document layout analysis," in *28th Int. Conf. on Computational Linguistics*, Barcelona, Spain, pp. 949–960, 2020.
- [4] D. Tkaczyk, P. Szostek, M. Fedoryszak and P. Jan, "CERMINE: Automatic extraction of structured metadata from scientific literature," *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 18, no. 4, pp. 317–335, 2015.
- [5] P. Lopez, "GROBID: Combining automatic bibliographic data recognition and term extraction for scholarship publications," in *Int. Conf. on Theory and Practice of Digital Libraries*, Berlin, Heidelberg, Springer, pp. 473–474, 2009.
- [6] S. Kim, S. Ji, H. Jeong, H. Yoon and S. Choi, "Metadata extraction based on deep learning from academic paper in pdf," *Journal of KIISE*, vol. 46, no. 7, pp. 644–652, 2019.
- [7] X. Yiheng, L. Minghao, C. Lei, H. Shaohan, W. Furu *et al.*, "Layoutlm: Pre-training of text and layout for document image understanding," in *Proc. 26th ACM SIGKDD Int. Conf. on Knowledge Discovery & Data Mining*, New York, NY, USA, pp. 1192–1200, 2020.
- [8] K. He, G. Gkioxari, P. Dollár and R. Girshick, "Mask r-cnn," in Proc. IEEE Int. Conf. on Computer Vision, Venice, Italy, pp. 2961–2969, 2017.
- [9] R. Girshick, "Fast r-cnn," in *Proc. IEEE Int. Conf. on Computer Vision*, Santiago, Chile, pp. 1440–1448, 2015.
- [10] J. Devlin, M. W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, Minnesota, USA, vol. 1, pp. 4171– 4186, 2019.
- [11] H. Han, C. L. Giles, E. Manavoglu, H. Zha and E. A. Fox, "Automatic document metadata extraction using support vector machines," in *Joint Conf. on Digital Libraries*, Houston, TX, USA, pp. 37–48, 2003.
- [12] M. Abramson, "Sequence classification with neural conditional random fields," in *IEEE 14th Int. Conf. on Machine Learning and Applications (ICMLA)*, Miami, Florida, USA, pp. 799–804, 2015.
- [13] Y. Kim, "Convolutional neural networks for sentence classification," MS thesis, University of Waterloo, 2015.
- [14] P. Zhou, Z. Qi, S. Zheng, J. Xu, H. Bao et al., "Text classification improved by integrating bidirectional LSTM with two-dimensional max pooling," in Proc. COLING 2016, the 26th Int. Conf. on Computational Linguistics: Technical Papers, Osaka, Japan, pp. 3485–3495, 2016.
- [15] A. Adhikari, A. Ram, R. Tang and J. Lin, "DocBERT: BERT for document classification," arXiv preprint arXiv:1904.08398, 2019. [Online]. Available: http://arxiv.org/abs/1904.08398.

- [16] S. Yu, J. Su and D. Luo, "Improving bert-based text classification with auxiliary sentence and domain knowledge," *IEEE Access*, vol. 7, pp. 176600–176612, 2019.
- [17] X. Gu, K. M. Yoo and J. Ha, "DialogBERT: Discourse-aware response generation via learning to recover and rank utterances," arXiv preprint arXiv:2012.01775v1, 2021. [Online]. Available: https://arxiv.org/ abs/2012.01775.
- [18] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma et al., "ALBERT: A lite BERT for self-supervised learning of language representations," in *Int. Conf. on Learning Representations*, New Orleans, USA, 2019.
- [19] I. Beltagy, K. Lo and A. Cohan, "SciBERT: A pretrained language model for scientific text," in Proc. 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th Int. Joint Conf. on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, pp. 3615–3620, 2019.
- [20] A. R. Katti, C. Reisswig, C. Guder, S. Brarda, S. Bickel et al., "Chargrid: Towards understanding 2d documents," in Proc. 2018 Conf. on Empirical Methods in Natural Language Processing, Brussels, Belgium, pp. 4459–4469, 2018.
- [21] Ł Garncarek, R. Powalski and T. Stanisławek, "LAMBERT: Layout-aware (language) modeling for information extraction," in arXiv preprint arXiv:2002.08087, 2020. [Online]. Available: https://arxiv.org/ abs/2002.08087.
- [22] Z. -Q. Zhao, P. Zheng, S. Xu and X. Wu, "Object detection with deep learning: A review," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 11, pp. 3212–3232, 2019.
- [23] Y. Liu, L. Jin, Z. Xie, C. Luo, S. Zhang et al., "Tightness-aware evaluation protocol for scene text detection," in Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition, Long Beach, CA, USA, pp. 9612– 9620, 2019.
- [24] A. Simon, J.-C. Pret and A. P. Johnson, "A fast algorithm for bottom-up document layout analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 3, pp. 273–277, 1997.
- [25] K. Papineni, S. Roukos, T. Ward and W. J. Zhu, "Bleu: A method for automatic evaluation of machine translation," in *Proc. 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, Pennsylvania, USA, pp. 311–318, 2002.
- [26] L. Jinhyuk, Y. Wonjin, K. Sungdong, K. Donghyeon, K. Sunkyu *et al.*, "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.
- [27] Y. Wu, M. Schuster, Z. Chen, Q. Le, M. Norouzi *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," arXiv preprint arXiv:1609.08144, 2016. [Online]. Available: https://arxiv.org/abs/1609.08144v2.

Appendix A

Journals	NUMBERS OF SELECTED PAPERS
Train Set	
JOURNAL OF THE KOREAN CLEFT	248
PALATE-CRANIOFACIAL ASSOCIATION	
Korean Journal of Construction	575
Engineering and Management	
JOURNAL OF INTERNET COMPUTING AND SERVICES	513
JOURNAL OF THE KOREAN SOCIETY OF RADIOLOGY	726

JOURNAL OF THE KOREA INSTITUTE OF 2477 INFORMATION AND COMMUNICATION ENGINEERING KOREAN JOURNAL OF MATERIALS RESEARCH 892 FOOD SCIENCE OF ANIMAL RESOURCES 636 KOREAN JOURNAL OF PEDIATRICS 882ARCH 761 JOURNAL OF INFORMATION PROCESSING SYSTEMS 97 JOURNAL OF INFORMATION PROCESSING SYSTEMS 97 JOURNAL OF INFORMATION PROCESSING SYSTEMS 97 JOURNAL OF THE KOREA CONVERGENCE SOCIETY 1411 JOURNAL OF THE KOREA CONVERGENCE SOCIETY 1411 JOURNAL OF THE KOREA NOCIETY OF CLOTHING 621 AND TEXTILES 76 MOLECULES AND CELLS 760 JOURNAL OF THE KOREAN SOCIETY OF NURSING 772 JOURNAL OF THE KOREAN SOCIETY OF 310 INTEGRATIVE MEDICINE 772 JOURNAL OF THE KOREAN SOCIETY OF 310 INTEGRATIVE MEDICINE 772 JOURNAL OF THE INSTITUTE OF INTERNET, 1558 BROADCASTING AND COMMUNICATION 727 RESOURCES ASSOCIATION 7448 ELECTRONIC COMMUNICATION 741 JOURNAL OF THE KOREA INSTITUTE OF 852 ELECTRONIC COMMUNICATION 7448 ELECTRONIC SUBGERY 1135 THE JOURNAL OF THE KOREA INSTITUTE OF 852 ELECTRONIC COMMUNICATION 7448 ELECTRONIC SUBGERY 737 JOURNAL OF KOREA INSTITUTE OF 852 ELECTRONIC COMMUNICATION 7448 ELECTRONICS, AND COM	Continued	1	
INFORMATION AND COMMUNICATION ENGINEERING KOREAN JOURNAL OF MATERIALS RESEARCH SOEAN JOURNAL OF MATERIALS RESEARCH FOOD SCIENCE OF ANIMAL RESOURCES SOEAN OLINAL OF PEDIATRICS KOREAN OLINAL OF PEDIATRICS KOREAN CHEMICAL ENGINEERING RESEARCH JOURNAL OF INFORMATION PROCESSING SYSTEMS JOURNAL OF DIGITAL CONVERGENCE SOCIETY JOURNAL OF THE KOREA CONVERGENCE SOCIETY JOURNAL OF THE KOREA CONVERGENCE SOCIETY MOLECULES AND CELLS MOLECULES AND CELLS MOLECULES AND CELLS MOLECULES AND CELLS JOURNAL OF THE KOREAN SOCIETY OF JOURNAL OF KOREAN ACADEMY OF NURSING THE JOURNAL OF THE KOREAN SOCIETY OF THE JOURNAL OF THE KOREAN SOCIETY OF THE JOURNAL OF THE KOREAN SOCIETY OF JOURNAL OF KOREAN MATER JOURNAL OF KOREAW WATER JOURNAL OF KOREA WATER JOURNAL OF THE KOREA INSTITUTE OF BULETIN OF THE KOREA INSTITUTE OF BULETIN OF THE KOREA INSTITUTE OF SOCIATION JOURNAL OF FUNCES ARCHIVES OF PLASTIC SURGERY THE JOURNAL OF ADVANCED SMART ARCHIVES OF PLASTIC SURGERY APPLIED CHEMISTRY FOR ENGINEERING JOURNAL OF KOREA INSTITUTE OF INFORMATION, TECHNOLOGY FEDIATRIC GASTROENTEROLOGY, HEPATOLOGY & PEDIATRIC GASTROENTEROLOGY, HEPATOLOGY & TECHNOLOGY FEDIATRIC GASTROENTEROLOGY, HEPATOLOGY & TECHNOLOGY TECHNOLOGY THE JOURNAL OF THE KOREA CONTENTS SIZ4 ASSOCIATION THE JOURNAL OF KOREAN ORIENTAL INTERNAL MEDICINE KSII TRANSACTIONS ON INTERNET AND JZ56 INFORMATION SYSTEMS (TIIIS) JOURNAL OF KOREAN MOLICINE REHABILITATION SIGN JOURNAL OF KOREAN MOLICINE REHABILITATION SIGN JOURNAL OF PHOSICIONE REHABILITATION JOURNAL OF PHOSICIONE REHABILITATION JOURNAL OF PHOSICIONE REHABILITATION JOURNAL OF PHORENAL OF PHYSIOLOGY AND SIB	JOURNAL OF THE KOREA INSTITUTE OF	2477	
ENGINEERING KOREAN JOURNAL OF MATERIALS RESEARCH KOREAN JOURNAL OF MATERIALS RESEARCH FOOD SCIENCE OF ANIMAL RESOURCES KOREAN JOURNAL OF PEDIATRICS KOREAN GENERAL RESEARCH KOREAN RESEARCH KORE	Information and Communication		
KOREAN JOURNAL OF MATERIALS RESEARCH892FOOD SCIENCE OF ANIMAL RESOURCES636KOREAN JOURNAL OF PEDIATRICS812KOREAN CHEMICAL ENGINEERING RESEARCH761JOURNAL OF INFORMATION PROCESSING SYSTEMS97JOURNAL OF INFORMATION PROCESSING SYSTEMS97JOURNAL OF INFORMATION PROCESSING SYSTEMS97JOURNAL OF DIGITAL CONVERGENCE3351JOURNAL OF THE KOREA CONVERGENCE SOCIETY1411JOURNAL OF THE KOREA NOCETY OF CLOTHING621AND TEXTILES360JOURNAL OF KOREAN ACADEMY OF NURSING772JOURNAL OF KOREAN ACADEMY OF NURSING772JOURNAL OF KOREAN ACADEMY OF NURSING772JOURNAL OF THE KOREAN SOCIETY OF310INTEGRATIVE MEDICINE1558THE JOURNAL OF THE KOREAN SOCIETY OF310JOURNAL OF KOREAN ACADEMY OF NURSING727RESOURCES ASSOCIATIONJOURNAL OF KOREA WATERJOURNAL OF KOREA WATER727RESOURCES ASSOCIATIONJOURNAL OF ADVANCED SMARTJOURNAL OF THE KMS1133INTERNATIONAL JOURNAL OF ADVANCED SMART341CONVERGENCE448ELECTRONIC COMMUNICATION SCIENCES537JOURNAL OF THE KOREA INSTITUTE OF852ELECTRONIC COMMUNICATION SCIENCES537JOURNAL OF THE KOREA CONTENTS5524ASSOCIATION448ELECTRONIC GASTRO ORIENTAL INTERNAL603MUEDICINETHE JOURNAL OF THE KOREA CONTENTS5524ASSOCIATION5181250THE JOURNAL OF KOREAN ORIENTAL INTERN	Engineering		
Food Science of Animal Resources636Korean Journal of Pediatrics812Korean Chemical Engineering Research761Journal of Information Processing Systems97Journal of Digital Convergence3351Journal of The Korea Convergence Society1411Journal of The Korean Society of Clothing621and Textilles360Molecules and Cells360Journal of Korean Academy of Nursing772Journal of Korean Academy of Nursing772Journal of Korean Academy of Nursing772Journal of The Institute of Internet,1558Broadcasting and Communication100JOURNAL OF Korean Academy of Nursing727RESOURCES ASSOCIATION1133Bulletin of the KMS1133International Journal of advanced Smart341convergence337Archives of Plastic Surgery1135The Journal of The Korea Institute of Information,448Electronic Communication Sciences537Journal of Korea Institute of Information,448Electronics, and Communication537Technology537Pediatric Gastroenterology, Hepatology & 373Nutrition5524Association5524Association5524Massociation518	Korean Journal of Materials Research	892	
KOREAN JOURNAL OF PEDIATRICS812KOREAN CHEMICAL ENGINEERING RESEARCH761JOURNAL OF INFORMATION PROCESSING SYSTEMS97JOURNAL OF DIGITAL CONVERGENCE3351JOURNAL OF THE KOREA CONVERGENCE SOCIETY1411JOURNAL OF THE KOREAN SOCIETY OF CLOTHING621AND TEXTILES360MOLECULES AND CELLS360JOURNAL OF KOREAN ACADEMY OF NURSING772JOURNAL OF THE KOREAN SOCIETY OF310INTEGRATIVE MEDICINETHE JOURNAL OF THE INSTITUTE OF INTERNET,THE JOURNAL OF THE INSTITUTE OF INTERNET,1558BROADCASTING AND COMMUNICATIONJOURNAL OF KOREA WATERJOURNAL OF THE KMS1133INTERNATIONAL JOURNAL OF ADVANCED SMART341CONVERGENCE337ARCHIVES OF PLASTIC SURGERY1135THE JOURNAL OF THE KOREA INSTITUTE OF852ELECTRONIC COMMUNICATION SCIENCES373Applied CHEMISTRY FOR ENGINEERING537JOURNAL OF KOREA INSTITUTE OF INFORMATION,448ELECTRONICS, AND COMMUNICATION448ELECTRONICS, AND COMMUNICATION5524ASSOCIATIONTHE JOURNAL OF THE KOREA CONTENTS5524ASSOCIATIONTHE JOURNAL OF KOREAN ORIENTAL INTERNAL603MEDICINEKSII TRANSACTIONS ON INTERNET AND1256INFORMATION SYSTEMS (THIS)JOURNAL OF KOREAN MEDICINE REHABILITATION250THE KOREAN JOURNAL OF PHYSIOLOGY AND518	FOOD SCIENCE OF ANIMAL RESOURCES	636	
KOREAN CHEMICAL ENGINEERING RESEARCH761JOURNAL OF INFORMATION PROCESSING SYSTEMS97JOURNAL OF DIGITAL CONVERGENCE SOCIETY3351JOURNAL OF THE KOREA CONVERGENCE SOCIETY1411JOURNAL OF THE KOREAN SOCIETY OF CLOTHING621AND TEXTILES360MOLECULES AND CELLS360JOURNAL OF KOREAN ACADEMY OF NURSING772JOURNAL OF KOREAN ACADEMY OF NURSING772JOURNAL OF THE KOREAN SOCIETY OF310INTEGRATIVE MEDICINE1558THE JOURNAL OF THE INSTITUTE OF INTERNET,1558BROADCASTING AND COMMUNICATION100JOURNAL OF KOREA WATER727RESOURCES ASSOCIATION1133INTERNATIONAL JOURNAL OF ADVANCED SMART341CONVERGENCE200ARCHIVES OF PLASTIC SURGERY1135THE JOURNAL OF THE KOREA INSTITUTE OF852ELECTRONIC COMMUNICATION SCIENCES448ELECTRONIC COMMUNICATION SCIENCES373JOURNAL OF KOREA INSTITUTE OF INFORMATION,448ELECTRONICS, AND COMMUNICATION537JOURNAL OF KOREA INSTITUTE OF INFORMATION,448ELECTRONICS, AND COMMUNICATION773THE JOURNAL OF THE KOREA CONTENTS5524ASSOCIATION5524MUTRITION1256INFORMATION SYSTEMS (TIIS)500JOURNAL OF KOREAN ORIENTAL INTERNAL603MEDICINE550THE JOURNAL OF KOREAN ORIENTAL INTERNAL603MEDICINE550INFORMATION SYSTEMS (TIIS)550JOURNAL OF KOREAN	Korean Journal of Pediatrics	812	
JOURNAL OF INFORMATION PROCESSING SYSTEMS 97 JOURNAL OF DIGITAL CONVERGENCE 3351 JOURNAL OF THE KOREA CONVERGENCE SOCIETY 1411 JOURNAL OF THE KOREAN SOCIETY OF CLOTHING 621 AND TEXTILES 621 MOLECULES AND CELLS 360 JOURNAL OF KOREAN ACADEMY OF NURSING 772 JOURNAL OF KOREAN ACADEMY OF NURSING 772 JOURNAL OF THE KOREAN SOCIETY OF 310 INTEGRATIVE MEDICINE THE JOURNAL OF THE INSTITUTE OF INTERNET, 1558 BROADCASTING AND COMMUNICATION JOURNAL OF KOREAN WATER 727 RESOURCES ASSOCIATION BULLETIN OF THE KMS 1133 INTERNATIONAL JOURNAL OF ADVANCED SMART 341 CONVERGENCE ARCHIVES OF PLASTIC SURGERY 1135 THE JOURNAL OF THE KOREA INSTITUTE OF 852 ELECTRONIC COMMUNICATION SCIENCES APPLIED CHEMISTRY FOR ENGINEERING 537 JOURNAL OF KOREA INSTITUTE OF INFORMATION, 448 ELECTRONIC GASTROENTEROLOGY, HEPATOLOGY & 373 NUTRITION THE JOURNAL OF THE KOREA CONTENTS 5524 ASSOCIATION THE JOURNAL OF KOREA ORIENTAL INTERNAL 603 MEDICINE KSII TRANSACTIONS ON INTERNET AND 1256 INFORMATION SYSTEMS (TIIS) JOURNAL OF KOREAN MEDICINE REHABILITATION 250 THE JOURNAL OF KOREAN MEDICINE REHABILITATION 250 THE KOREAN JOURNAL OF PHYSOLOGY AND 518	Korean Chemical Engineering Research	761	
JOURNAL OF DIGITAL CONVERGENCE 33351 JOURNAL OF THE KOREA CONVERGENCE SOCIETY 1411 JOURNAL OF THE KOREA CONVERGENCE SOCIETY 1411 JOURNAL OF THE KOREAN SOCIETY OF CLOTHING 621 AND TEXTILES 360 JOURNAL OF THE KOREAN ACADEMY OF NURSING 772 JOURNAL OF KOREAN ACADEMY OF NURSING 772 JOURNAL OF THE KOREAN SOCIETY OF 310 INTEGRATIVE MEDICINE THE JOURNAL OF INE INSTITUTE OF INTERNET, 1558 BROADCASTING AND COMMUNICATION 727 RESOURCES ASSOCIATION 727 RESOURCES ASSOCIATION 727 RESOURCES ASSOCIATION 741 SULLETIN OF THE KMS 1133 INTERNATIONAL JOURNAL OF ADVANCED SMART 341 CONVERGENCE 74 ARCHIVES OF PLASTIC SURGERY 1135 THE JOURNAL OF THE KOREA INSTITUTE OF 852 ELECTRONIC COMMUNICATION SCIENCES 74 APPLIED CHEMISTRY FOR ENGINEERING 537 JOURNAL OF KOREA INSTITUTE OF INFORMATION, 448 ELECTRONICS, AND COMMUNICATION 75 TECHNOLOGY 75 PEDIATRIC GASTROENTEROLOGY, HEPATOLOGY & 373 NUTRITION 75 THE JOURNAL OF THE KOREA CONTENTS 5524 ASSOCIATION 75 THE JOURNAL OF KOREAN ORIENTAL INTERNAL 603 MEDICINE 75 KSII TRANSACTIONS ON INTERNET AND 1256 INFORMATION SYSTEMS (TIIS) JOURNAL OF KOREAN MEDICINE REHABILITATION 250 THE KOREAN MEDICINE REHABILITATION 250 THE JOURNAL OF KOREAN MEDICINE REHABILITATION 250 THE JOURNAL OF KOREAN MEDICINE REHABILITATION 250 THE KOR	JOURNAL OF INFORMATION PROCESSING SYSTEMS	97	
JOURNAL OF THE KOREA CONVERGENCE SOCIETY 1411 JOURNAL OF THE KOREAN SOCIETY OF CLOTHING 621 AND TEXTILES 621 MOLECULES AND CELLS 360 JOURNAL OF KOREAN ACADEMY OF NURSING 772 JOURNAL OF KOREAN ACADEMY OF NURSING 772 JOURNAL OF THE KOREAN SOCIETY OF 310 INTEGRATIVE MEDICINE 727 THE JOURNAL OF THE INSTITUTE OF INTERNET, 1558 BROADCASTING AND COMMUNICATION 727 RESOURCES ASSOCIATION 727 RESOURCES ASSOCIATION 727 BULLETIN OF THE KMS 1133 INTERNATIONAL JOURNAL OF ADVANCED SMART 341 CONVERGENCE 747 ARCHIVES OF PLASTIC SURGERY 1135 THE JOURNAL OF THE KOREA INSTITUTE OF S22 ELECTRONIC COMMUNICATION SCIENCES 74 APPLIED CHEMISTRY FOR ENGINEERING 537 JOURNAL OF KOREA INSTITUTE OF INFORMATION, 448 ELECTRONICS, AND COMMUNICATION 7448 ELECTRONICS, AND COMMUNICATION 7448 ELECTRONICS, AND COMMUNICATION 7448 TECHNOLOGY 74 PEDIATRIC GASTROENTEROLOGY, HEPATOLOGY & 373 NUTRITION 748 THE JOURNAL OF THE KOREA CONTENTS 5524 ASSOCIATION 75124 ASSOCIATION 75124 ASSOCIATION 1125 THE JOURNAL OF KOREAN ORIENTAL INTERNAL 603 MEDICINE 7514 KSII TRANSACTIONS ON INTERNET AND 1256 INFORMATION SYSTEMS (TIIS) JOURNAL OF PHYSIOLOGY AND 518	JOURNAL OF DIGITAL CONVERGENCE	3351	
JOURNAL OF THE KOREAN SOCIETY OF CLOTHING 621 AND TEXTILES 360 JOURNAL OF KOREAN ACADEMY OF NURSING 772 JOURNAL OF KOREAN ACADEMY OF NURSING 772 JOURNAL OF THE KOREAN SOCIETY OF 310 INTEGRATIVE MEDICINE THE JOURNAL OF THE INSTITUTE OF INTERNET, 1558 BROADCASTING AND COMMUNICATION 727 RESOURCES ASSOCIATION 31133 INTERNATIONAL JOURNAL OF ADVANCED SMART 341 CONVERGENCE ARCHIVES OF PLASTIC SURGERY 1135 THE JOURNAL OF THE KOREA INSTITUTE OF 852 ELECTRONIC COMMUNICATION SCIENCES APPLIED CHEMISTRY FOR ENGINEERING 537 JOURNAL OF KOREA INSTITUTE OF INFORMATION, 448 ELECTRONICS, AND COMMUNICATION TECHNOLOGY PEDIATRIC GASTROENTEROLOGY, HEPATOLOGY & 373 NUTRITION THE JOURNAL OF KOREA CONTENTS 5524 ASSOCIATION THE JOURNAL OF KOREA CONTENTS 5524 ASSOCIATION THE JOURNAL OF KOREA CONTENTS 5524 ASSOCIATION THE JOURNAL OF KOREA CONTENTS 252 HEDITIC GASTROENTEROLOGY, HEPATOLOGY & 373 NUTRITION THE JOURNAL OF KOREA CONTENTS 5524 ASSOCIATION THE JOURNAL OF KOREA CONTENTS 252 HEDITIC GASTROENTEROLOGY, HEPATOLOGY & 373 NUTRITION THE JOURNAL OF KOREAN ORIENTAL INTERNAL 603 MEDICINE KSII TRANSACTIONS ON INTERNET AND 1256 INFORMATION SYSTEMS (TIIS) JOURNAL OF KOREAN MEDICINE REHABILITATION 250 THE KOREAN JOURNAL OF PHYSIOLOGY AND 518	JOURNAL OF THE KOREA CONVERGENCE SOCIETY	1411	
AND TEXTILES MOLECULES AND CELLS MOLECULES AND CELLS MOLECULES AND CELLS MOLECULES AND CELLS MOLECULS	JOURNAL OF THE KOREAN SOCIETY OF CLOTHING	621	
Molecules and Cells360JOURNAL OF KOREAN ACADEMY OF NURSING772JOURNAL OF KOREAN ACADEMY OF NURSING772JOURNAL OF THE KOREAN SOCIETY OF310Integrative Medicine1558Broadcasting and CommunicationJOURNAL OF The Institute of Internet,JOURNAL OF KOREA WATER727RESOURCES ASSOCIATIONJULLETIN OF THE KMSBulletin of the KMS1133INTERNATIONAL JOURNAL OF ADVANCED SMART341CONVERGENCEArchives of Plastic SurgeryArchives of Plastic Surgery1135The JOURNAL OF THE KOREA INSTITUTE OF852Electronic Communication Sciences537JOURNAL OF KOREA INSTITUTE OF INFORMATION,448Electronics, and Communication448Electronics, and Communication731TrechnologyTechnologyPediatric Gastroenterology, Hepatology &373NUTRITIONThe JOURNAL OF KOREA ONTENTS5524Association5524KSII Transactions on Internet and1256Information Systems (TIIS)JOURNAL OF KOREAN MEDICINE REHABILITATION250The Korean Journal of Physiology and518	and Textiles		
JOURNAL OF KOREAN ACADEMY OF NURSING 772 JOURNAL OF THE KOREAN SOCIETY OF 310 INTEGRATIVE MEDICINE 758 THE JOURNAL OF THE INSTITUTE OF INTERNET, 1558 BROADCASTING AND COMMUNICATION 727 RESOURCES ASSOCIATION 727 RESOURCES ASSOCIATION 741 BULLETIN OF THE KMS 1133 INTERNATIONAL JOURNAL OF ADVANCED SMART 341 CONVERGENCE 758 ARCHIVES OF PLASTIC SURGERY 1135 THE JOURNAL OF THE KOREA INSTITUTE OF 852 ELECTRONIC COMMUNICATION SCIENCES 7537 JOURNAL OF KOREA INSTITUTE OF INFORMATION, 448 ELECTRONICS, AND COMMUNICATION 759 PEDIATRIC GASTROENTEROLOGY, HEPATOLOGY & 373 NUTRITION 75524 ASSOCIATION 755524 ASSOCIATION 75524 ASSOCIATION 755524 ASSOCIATION 75524 ASSOCIATION 755524 ASSOCIATION 7555757777777777777777777777777777777	Molecules and Cells	360	
JOURNAL OF THE KOREAN SOCIETY OF 310 INTEGRATIVE MEDICINE 310 THE JOURNAL OF THE INSTITUTE OF INTERNET, 1558 BROADCASTING AND COMMUNICATION 727 RESOURCES ASSOCIATION 727 RESOURCES ASSOCIATION 341 CONVERGENCE 341 CONVERGENCE 341 ARCHIVES OF PLASTIC SURGERY 1135 THE JOURNAL OF THE KOREA INSTITUTE OF 852 ELECTRONIC COMMUNICATION SCIENCES 357 JOURNAL OF KOREA INSTITUTE OF INFORMATION, 448 ELECTRONICS, AND COMMUNICATION 52 ELECTRONICS, AND COMMUNICATION 448 ELECTRONICS, AND COMMUNICATION 52 FECHNOLOGY 752 PEDIATRIC GASTROENTEROLOGY, HEPATOLOGY & 373 NUTRITION 5524 ASSOCIATION 5524 ASSOCIATION 5524 ASSOCIATION 5524 ASSOCIATION 1256 THE JOURNAL OF KOREAN ORIENTAL INTERNAL 603 MEDICINE 603 MEDICINE 5524 ASSOCIATION 1256 NFORMATION SYSTEMS (TIIS) 1256 INFORMATION SYSTEMS (TIIS) 1256 INFORMATION SYSTEMS (TIIS) 250 THE KOREAN MEDICINE REHABILITATION 250 THE KOREAN JOURNAL OF PHYSIOLOGY AND 518	JOURNAL OF KOREAN ACADEMY OF NURSING	772	
INTEGRATIVE MEDICINE1558The JOURNAL OF THE INSTITUTE OF INTERNET,1558BROADCASTING AND COMMUNICATION727JOURNAL OF KOREA WATER727RESOURCES ASSOCIATION1133BULLETIN OF THE KMS1133INTERNATIONAL JOURNAL OF ADVANCED SMART341CONVERGENCE341ARCHIVES OF PLASTIC SURGERY1135THE JOURNAL OF THE KOREA INSTITUTE OF852ELECTRONIC COMMUNICATION SCIENCES448ELECTRONIC COMMUNICATION SCIENCES537JOURNAL OF KOREA INSTITUTE OF INFORMATION,448ELECTRONICS, AND COMMUNICATION448ELECTRONICS, AND COMMUNICATION73VUTRITION73THE JOURNAL OF THE KOREA CONTENTS5524ASSOCIATION5524THE JOURNAL OF KOREAN ORIENTAL INTERNAL603MEDICINE1256INFORMATION SYSTEMS (TIIS)1256JOURNAL OF KOREAN MEDICINE REHABILITATION250THE KOREAN JOURNAL OF PHYSIOLOGY AND518	JOURNAL OF THE KOREAN SOCIETY OF	310	
The Journal of The Institute of Internet,1558Broadcasting and Communication727Besources Association727Resources Association1133Bulletin of the KMS1133International journal of advanced smart341convergence1135Archives of Plastic Surgery1135The Journal of The Korea Institute of852Electronic Communication Sciences537Applied Chemistry for Engineering537Journal of Korea Institute of Information,448Electronics, and Communication537Technology73Pediatric Gastroenterology, Hepatology &373Nutrition5524Association603Medicine5524KSII Transactions on Internet and1256Information Systems (TIIS)250Journal of Korean Medicine Rehabilitation250The Korean Journal of Physiology and518	INTEGRATIVE MEDICINE		
BROADCASTING AND COMMUNICATION727JOURNAL OF KOREA WATER727RESOURCES ASSOCIATION1133BULLETIN OF THE KMS1133INTERNATIONAL JOURNAL OF ADVANCED SMART341CONVERGENCE341ARCHIVES OF PLASTIC SURGERY1135THE JOURNAL OF THE KOREA INSTITUTE OF852ELECTRONIC COMMUNICATION SCIENCES537APPLIED CHEMISTRY FOR ENGINEERING537JOURNAL OF KOREA INSTITUTE OF INFORMATION,448ELECTRONICS, AND COMMUNICATION448ELECTRONICS, AND COMMUNICATION7TECHNOLOGY73PEDIATRIC GASTROENTEROLOGY, HEPATOLOGY &373NUTRITION7THE JOURNAL OF THE KOREA CONTENTS5524Association403MEDICINE403KSII TRANSACTIONS ON INTERNET AND1256INFORMATION SYSTEMS (TIIS)250JOURNAL OF KOREAN MEDICINE REHABILITATION250THE KOREAN JOURNAL OF PHYSIOLOGY AND518	THE JOURNAL OF THE INSTITUTE OF INTERNET,	1558	
JOURNAL OF KOREA WATER 727 RESOURCES ASSOCIATION 727 BULLETIN OF THE KMS 1133 INTERNATIONAL JOURNAL OF ADVANCED SMART 341 CONVERGENCE 741 ARCHIVES OF PLASTIC SURGERY 1135 THE JOURNAL OF THE KOREA INSTITUTE OF 852 ELECTRONIC COMMUNICATION SCIENCES 7537 JOURNAL OF KOREA INSTITUTE OF INFORMATION, 448 ELECTRONICS, AND COMMUNICATION 751 TECHNOLOGY 751 PEDIATRIC GASTROENTEROLOGY, HEPATOLOGY & 373 NUTRITION 751 THE JOURNAL OF THE KOREA CONTENTS 5524 ASSOCIATION 7524 ASSOCIATION 7524 KSII TRANSACTIONS ON INTERNET AND 1256 INFORMATION SYSTEMS (TIIS) 7518	BROADCASTING AND COMMUNICATION		
RESOURCES ASSOCIATIONBULLETIN OF THE KMS1133INTERNATIONAL JOURNAL OF ADVANCED SMART341CONVERGENCE341ARCHIVES OF PLASTIC SURGERY1135THE JOURNAL OF THE KOREA INSTITUTE OF852ELECTRONIC COMMUNICATION SCIENCES452APPLIED CHEMISTRY FOR ENGINEERING537JOURNAL OF KOREA INSTITUTE OF INFORMATION,448ELECTRONICS, AND COMMUNICATION448ELECTRONICS, AND COMMUNICATION773TECHNOLOGY773PEDIATRIC GASTROENTEROLOGY, HEPATOLOGY &373NUTRITION714THE JOURNAL OF THE KOREA CONTENTS5524ASSOCIATION603MEDICINE1256INFORMATION SYSTEMS (TIIS)250JOURNAL OF KOREAN MEDICINE REHABILITATION250THE KOREAN JOURNAL OF PHYSIOLOGY AND518	JOURNAL OF KOREA WATER	727	
Bulletin of the KMS1133International journal of advanced smart341convergence1135Archives of Plastic Surgery1135The Journal of The Korea Institute of852Electronic Communication Sciences537Applied Chemistry for Engineering537Journal of Korea Institute of Information,448Electronics, and Communication448Electronics, and Communication7Technology73Pediatric Gastroenterology, Hepatology &373Nutrition5524Association603Medicine1256Information Systems (TIIS)250Journal of Korean Medicine Rehabilitation250The Korean Journal of Physiology and518	RESOURCES ASSOCIATION		
INTERNATIONAL JOURNAL OF ADVANCED SMART341CONVERGENCE1135ARCHIVES OF PLASTIC SURGERY1135THE JOURNAL OF THE KOREA INSTITUTE OF852ELECTRONIC COMMUNICATION SCIENCES537APPLIED CHEMISTRY FOR ENGINEERING537JOURNAL OF KOREA INSTITUTE OF INFORMATION,448ELECTRONICS, AND COMMUNICATION448PECINOLOGY73PEDIATRIC GASTROENTEROLOGY, HEPATOLOGY &373NUTRITION5524Association5524Association603MEDICINE1256Information Systems (TIIS)250JOURNAL OF KOREAN MEDICINE REHABILITATION250THE KOREAN JOURNAL OF PHYSIOLOGY AND518	BULLETIN OF THE KMS	1133	
CONVERGENCE1135ARCHIVES OF PLASTIC SURGERY1135THE JOURNAL OF THE KOREA INSTITUTE OF852ELECTRONIC COMMUNICATION SCIENCES537APPLIED CHEMISTRY FOR ENGINEERING537JOURNAL OF KOREA INSTITUTE OF INFORMATION,448ELECTRONICS, AND COMMUNICATION448ELECTRONICS, AND COMMUNICATION7TECHNOLOGY733PEDIATRIC GASTROENTEROLOGY, HEPATOLOGY &373NUTRITION7THE JOURNAL OF THE KOREA CONTENTS5524ASSOCIATION5524MEDICINE603MEDICINE1256INFORMATION SYSTEMS (TIIS)1256JOURNAL OF KOREAN MEDICINE REHABILITATION250THE KOREAN JOURNAL OF PHYSIOLOGY AND518	INTERNATIONAL JOURNAL OF ADVANCED SMART	341	
ARCHIVES OF PLASTIC SURGERY1135THE JOURNAL OF THE KOREA INSTITUTE OF852ELECTRONIC COMMUNICATION SCIENCES537APPLIED CHEMISTRY FOR ENGINEERING537JOURNAL OF KOREA INSTITUTE OF INFORMATION,448ELECTRONICS, AND COMMUNICATION448ELECTRONICS, AND COMMUNICATION7TECHNOLOGY7PEDIATRIC GASTROENTEROLOGY, HEPATOLOGY &373NUTRITION7THE JOURNAL OF THE KOREA CONTENTS5524ASSOCIATION603MEDICINE7KSII TRANSACTIONS ON INTERNET AND1256INFORMATION SYSTEMS (TIIS)250JOURNAL OF KOREAN MEDICINE REHABILITATION250THE KOREAN JOURNAL OF PHYSIOLOGY AND518	CONVERGENCE		
The Journal of The Korea Institute of852Electronic Communication Sciences537Applied Chemistry for Engineering537Journal of Korea Institute of Information,448Electronics, and Communication448Electronics, and Communication7Technology7Pediatric Gastroenterology, Hepatology &373Nutrition7The Journal of the Korea Contents5524Association603Medicine7KSII Transactions on Internet and1256Information Systems (TIIS)250Journal of Korean Medicine Rehabilitation250The Korean Journal of Physiology and518	Archives of Plastic Surgery	1135	
Electronic Communication Sciences537Applied Chemistry for Engineering537Journal of Korea Institute of Information,448Electronics, and Communication448Technology7Pediatric Gastroenterology, Hepatology &373Nutrition5524Association5524Association603Medicine1256Information Systems (TIIS)250Journal of Korean Medicine Rehabilitation250The Korean Journal of Physiology and518	THE JOURNAL OF THE KOREA INSTITUTE OF	852	
Applied Chemistry for Engineering537Journal of Korea Institute of Information,448Electronics, and Communication7Technology7Pediatric Gastroenterology, Hepatology &373Nutrition7The Journal of the Korea Contents5524Association603Medicine7KSII Transactions on Internet and1256Information Systems (TIIS)250Journal of Korean Medicine Rehabilitation250The Korean Journal of Physiology and518	ELECTRONIC COMMUNICATION SCIENCES		
JOURNAL OF KOREA INSTITUTE OF INFORMATION,448ELECTRONICS, AND COMMUNICATION	APPLIED CHEMISTRY FOR ENGINEERING	537	
Electronics, and CommunicationTechnologyPediatric Gastroenterology, Hepatology &373NutritionThe Journal of the Korea ContentsAssociationThe Journal of Korean Oriental Internal603MedicineKSII Transactions on Internet and1256Information Systems (TIIS)Journal of Korean Medicine Rehabilitation250The Korean Journal of Physiology and518	JOURNAL OF KOREA INSTITUTE OF INFORMATION.	448	
TECHNOLOGY PEDIATRIC GASTROENTEROLOGY, HEPATOLOGY & 373 NUTRITION THE JOURNAL OF THE KOREA CONTENTS 5524 Association THE JOURNAL OF KOREAN ORIENTAL INTERNAL 603 MEDICINE KSII TRANSACTIONS ON INTERNET AND 1256 INFORMATION SYSTEMS (THS) JOURNAL OF KOREAN MEDICINE REHABILITATION 250 THE KOREAN JOURNAL OF PHYSIOLOGY AND 518	ELECTRONICS, AND COMMUNICATION		
Pediatric Gastroenterology, Hepatology &373Nutrition5524The Journal of the Korea Contents5524Association603The Journal of Korean Oriental Internal603Medicine1256Information Systems (TIIS)1256Journal of Korean Medicine Rehabilitation250The Korean Journal of Physiology and518	TECHNOLOGY		
NUTRITIONTHE JOURNAL OF THE KOREA CONTENTS5524Association5524THE JOURNAL OF KOREAN ORIENTAL INTERNAL603MEDICINE1256KSII TRANSACTIONS ON INTERNET AND1256INFORMATION SYSTEMS (TIIS)1000000000000000000000000000000000000	Pediatric Gastroenterology, Hepatology &	373	
The Journal of the Korea Contents5524Association	NUTRITION		
Association The Journal of Korean Oriental Internal 603 Medicine KSII Transactions on Internet and 1256 Information Systems (TIIS) Journal of Korean Medicine Rehabilitation 250 The Korean Journal of Physiology and 518	The Journal of the Korea Contents	5524	
The Journal of Korean Oriental Internal603Medicine1256KSII Transactions on Internet and1256Information Systems (TIIS)518	Association		
MEDICINEKSII TRANSACTIONS ON INTERNET AND1256Information Systems (TIIS)250JOURNAL OF KOREAN MEDICINE REHABILITATION250THE KOREAN JOURNAL OF PHYSIOLOGY AND518	THE JOURNAL OF KOREAN ORIENTAL INTERNAL	603	
KSII Transactions on Internet and1256Information Systems (TIIS)250Journal of Korean Medicine Rehabilitation250The Korean Journal of Physiology and518	Medicine		
INFORMATION SYSTEMS (THS)JOURNAL OF KOREAN MEDICINE REHABILITATION250THE KOREAN JOURNAL OF PHYSIOLOGY AND518	KSII Transactions on Internet and	1256	
JOURNAL OF KOREAN MEDICINE REHABILITATION250THE KOREAN JOURNAL OF PHYSIOLOGY AND518	INFORMATION SYSTEMS (TIIS)		
THE KOREAN JOURNAL OF PHYSIOLOGY AND 518	JOURNAL OF KOREAN MEDICINE REHABILITATION	250	
	THE KOREAN JOURNAL OF PHYSIOLOGY AND	518	
Pharmacology	Pharmacology		
THE JOURNAL OF KOREAN PHYSICAL THERAPY 561	THE JOURNAL OF KOREAN PHYSICAL THERAPY	561	
JOURNAL OF PREVENTIVE MEDICINE AND PUBLIC 420	JOURNAL OF PREVENTIVE MEDICINE AND PUBLIC	420	
Health	Health		

(Continued)

Conti	nued	
THE KOREAN JOURNAL OF THORACIC AND	779	
Cardiovascular Surgery		
CHILD HEALTH NURSING RESEARCH	393	
BIOMOLECULES & THERAPEUTICS	580	
Asian-Australasian Journal of Animal	417	
Sciences		
JOURNAL OF ENVIRONMENTAL HEALTH SCIENCES	533	
THE KOREAN JOURNAL OF PARASITOLOGY	705	
Journal of the Korean Society of Physical Medicine	547	
Journal of The Korea Institute of	919	
INFORMATION SECURITY AND CRYPTOLOGY		
JOURNAL OF THE KOREAN ASSOCIATION FOR	638	
Science Education		
Journal of the Korean Applied Science and Technology	747	
JOURNAL OF THE KOREAN SOCIETY OF CIVIL	1226	
Engineers		
JOURNAL OF DIGITAL CONTENTS SOCIETY	580	
THE KOREAN JOURNAL OF FOOD AND NUTRITION	967	
JOURNAL OF KOREAN NEUROSURGICAL SOCIETY	1132	
JOURNAL OF MICROBIOLOGY AND	1281	
BIOTECHNOLOGY		
THE KOREAN JOURNAL OF APPLIED STATISTICS	591	
JOURNAL OF POWER ELECTRONICS	154	
MICROBIOLOGY AND BIOTECHNOLOGY LETTERS	463	
THE JOURNAL OF KOREAN MEDICINE	418	
Ophthalmology & Otorhinolaryngology &		
Dermatology		
JOURNAL OF THE KOREAN LIBRARY AND	477	
INFORMATION SCIENCE SOCIETY		
JOURNAL OF THE KOREA SOCIETY OF COMPUTER	1912	
AND INFORMATION THE JOURNAL OF OPENTAL OPERATION \mathcal{X}	404	
Grand of Oriental Obstetrics &	404	
THE TRANSPORTIONS OF THE KOREAN INSTITUTE	797	
THE TRANSACTIONS OF THE KOREAN INSTITUTE	/8/	
IOURNAL OF LEE SOURCE	1202	
JOURNAL OF LIFE SCIENCE	1203	
EIKI JOURNAL	1100	
NIPS I RANSACTIONS ON SOFTWARE AND DATA	525	
ENGINEERING	462	
JOURNAL OF KOREAN INAVIGATION AND PORT RESEARCH	403	

	Continued
JOURNAL OF KOREA MULTIMEDIA SOCIETY	939
Sub Total	51676
	Test Set
THE JOURNAL OF KOREAN ACADEMY OF	444
Prosthodontics	
Korean Journal of Food Science and	869
Technology	
NUCLEAR ENGINEERING AND TECHNOLOGY	755
BMB REPORTS	647
JOURNAL OF KOREAN SOCIETY OF DENTAL	873
Hygiene	
JOURNAL OF KOREA ACADEMIA-INDUSTRIAL	7109
COOPERATION SOCIETY	
JOURNAL OF BROADCAST	667
ENGINEERING	
JOURNAL OF THE KOREA SOCIETY INDUSTRIAL	568
INFORMATION SYSTEM	
JOURNAL OF IKEEE	779
JOURNAL OF CONVERGENCE FOR INFORMATION	620
Technology	
SUB TOTAL	13331
TOTAL	65007