학술논문 내에서 참고문헌 정보가 포함된 서지 메타데이터 자동 생성 연구

Automatic Generation of Bibliographic Metadata with Reference Information for Academic Journals

정 선 기 (Seonki Jeong)*, 신 현 호 (Hyeonho Shin)**
지 선 영 (Seon-Yeong Ji)***, 최 성 필 (Sungphil Choi)****

목 차

초 록

서지정보는 연구 주제의 최신 동향의 인지와 유용성을 검증하는 데에 참고할 수 있다. 즉, 각자 연구자들이 필요로 하는 문헌에 신속하게 접근하기 위해서는 학술논문에서 저자 정보, 요약, 초록, 참고문헌 등을 쉬운 방법으로 파악해야 한다. 그러나, 현재 출판되는 PDF 형식의 전자 학술논문은 출판 주체별로 고유한 양식을 띄고 있어서, 몇몇 특징에 의한 규칙 기반 추출법으로는 수많은 문헌에서 목표 정보를 추출하여 요약된 서지사항으로 자동 생성하기 어렵다. 이에 본 연구는 학술논문 서지사항 자동 생성에 있어서 양식의 다양성으로 인한 메타데이터 자동 추출의 난점을 극복할 방법을 제안한다. 제안하는 모델은 서지사항이 주로 기술되는 학술논문의 첫 페이지에서 목표 영역과 본문의 시작점을 구분할 수 있는 심층신경망 기반 모델과 앞의 모델로 추출된 서지사항을 상세한 메타데이터로 분류하고 재생성하는 규칙 기반 모델로 구성된다. 제안하는 모델은 참고문헌 요약정보를 생성하는 모델도 포함하는데, 본문의 말미와 참고문헌 시작점의 분리, 그리고 개별 참고문헌 추출을 규칙 기반 방법으로 진행하고, 추출한 각개 참고문헌의 서지정보를 분류하는 데에 심층신경망을 이용하도록 구성하였다. 추가로, 논문 자체의 서지정보를 전후처리 없이 추출/생성하는 모델의 가능성을 확인하기 위하여 참고문헌 영역까지 아우르는 모델을 구축하여 비교 실험을 진행하였다. 실험 결과 본 논문에서 제안하는 방식이 서지정보를 전후처리 하지 않고 진행한 비교 실험에 비하여 더 높은 성능을 보였다.

ABSTRACT

Bibliographic metadata can help researchers effectively utilize essential publications that they need and grasp academic trends of their own fields. With the manual creation of the metadata costly and time-consuming. it is nontrivial to effectively automatize the metadata construction using rule-based methods due to the immoderate variety of the article forms and styles according to publishers and academic societies. Therefore, this study proposes a two-step extraction process based on rules and deep neural networks for generating bibliographic metadata of scientific articlles to overcome the difficulties above. The extraction target areas in articles were identified by using a deep neural network-based model, and then the details in the areas were analyzed and sub-divided into relevant metadata elements. IThe proposed model also includes a model for generating reference summary information, which is able to separate the end of the text and the starting point of a reference, and to extract individual references by essential rule set, and to identify all the bibliographic items in each reference by a deep neural network. In addition, in order to confirm the possibility of a model that generates the bibliographic information of academic papers without pre- and post-processing, we conducted an in-depth comparative experiment with various settings and configurations. As a result of the experiment, the method proposed in this paper showed higher performance.

키워드: 자연어처리, 정보 추출, 참고문헌 추출, 메타데이터 추출, 언어모델
NLP, Information Extraction, Reference Extraction, Metadata Extraction, Language Model

* 경기대학교 문헌정보학과 석사과정(jsk1610@kyonggi.ac.kr) (제1저자)
** 경기대학교 문헌정보학과 석사과정(shinhh9554@gmail.com) (공동저자)
*** 주식회사 보인정보기술(syji@noinit.com) (공동저자)
**** 경기대학교 문헌정보학과 교수(sungpil@gmail.com / ISNI 0000 0004 6772 9269) (교신저자)
논문접수일자: 2022년 7월 17일 최초심사일자: 2022년 8월 1일 게재확정일자: 2022년 8월 15일
한국문헌정보학회지, 56(3): 241-264, 2022. http://dx.doi.org/10.4275/KSLIS.2022.56.3.241

1. 서 론

1.1 연구의 필요성 및 목적

학술논문은 일반적으로 첫 페이지에 제목, 저자, 초록 등으로 해당 논문의 서지정보가, 문헌 말미에는 참고문헌이 기록되어 있다. 서지정보는 논문의 핵심적인 내용과 부가 정보를 나타내는 문헌 대용물(Document Surrogate)로서 정보 이용자의 직접적인 접근 대상이며 원문정보에 대한 상세한 파악을 위한 관문으로 볼 수 있으며(Ziviani et al., 2011), 참고문헌 목록은 저자가 논문을 작성할 때 참고한 문헌을 나열한 것으로, 제목, 저자, 출판연도 등 저자가 인용한 학술 문헌에 대한 일정의 서지정보를 포함하고 있다. 참고문헌은 연구 주제의 최신 동향 파악 및 유용성 검증을 위한 탐색에 주로 활용되며(임수현 외, 2019), 논문의 인용-피인용 관계를 분석하는 정보로도 활용될 수 있다(김재훈 외, 2019; 김지훈, 2020).

대부분의 국내외 학술정보 데이터베이스는 학술논문의 요약정보 혹은 메타데이터를 수작업으로 입력하여 구축된다. PDF 형식으로 기록되는 텍스트 기반 디지털 문헌이 효율적으로 구축되고 상용됨에도 메타데이터 생성이 수작업 위주의 작업으로 진행되는 이유는 학술지와 출판사가 논문을 작성하는 데에 있어 고유의 작성양식을 가지기 때문이다. 고유한 작성양식이란 정해진 기준하에 작성되는 본문의 폭과 글자의 크기, 글꼴의 지정과 같은 형식의 차이, 혹은 그림과 도표의 배치와 같은 위치 정보의 차이 등으로 나타난다. 전 세계에서 가장 큰 생의학 분야 학술정보 데이터베이스인 PubMed Central에서 무작위로 선정한 125,000권의 논문은 약 500곳의 서로 다른 출판사에서 발간되었고, 모두 다른 양식으로 작성되었다(Tkaczyk et al., 2015). 또한, 본 논문에서 연구 데이터 원천으로 사용한 40종의 국내 학술지도 양식면에서 차이를 보였으며, 일부 학술지는 특정 시점에 고유의 양식을 대대적으로 개편한 경우도 있었다.

만일, 모든 학술지가 같은 양식으로 논문을 출판한다면 비교적 단순한 범용 추출 규칙을 이용하여 학술논문의 요약정보를 쉽게 추출할 수 있을 것이다. 그러나, 상기한 경우와 같이 양식의 차이점이 존재하는 상황에서는 범용 추출 규칙의 이용을 통한 방법은 현실적으로 불가하다. 규칙은 특정 학술지에 적합한 방향으로 작성될 것이며, 추출을 위해서 학술지마다 개별 추출 규칙 집합이 각각 구성되어야 하므로 효율적이지 못하다.

1.2 연구 방법

본 논문은 국내 학술지 양식의 다양성에 기인하는 서지정보 요약 자동 생성의 난점을 어느 정도 해결할 수 있는 메타데이터 추출 방법을 제안한다.

우선, 학술지 논문의 첫 번째 면을 대략적인 메타데이터 요소 영역으로 분리하여 추출한다. 이는 “저자 정보” 영역의 경우 저자의 성과 이름이 반드시 포함되어 있고, 학술지별 양식에 따라서 저자의 소속기관과 이메일 등이 포함되며, “날짜” 영역에는 투고일, 수정일, 게재일 등의 세부 정보가 존재하는 등의 특징에 따라서 각각의 요소 영역을 분리하여 구현할 수 있다.

각종 학술지별 고유 양식에 의존하는 메타데이터 요소 영역에 대한 추출은 각기 다른 양식으로 작성된 학술지 논문 표본을 대상으로 학습 집합을 구축하고 심층신경망 기반의 시퀀스 레이블링 모델을 학습하여 수행할 수 있다.

다음으로, 추출된 각개 요소 영역을 기준으로 규칙 기반 분류 방법을 통해 세부 메타데이터를 추출한다. 앞서 예로 들은 "저자 정보" 영역이라면 규칙 기반 분류법을 통하여 저자의 성과 이름, 소속기관, 이메일을 추출할 수 있을 것이다.

제안하는 방법은 위와 같이 요소 영역을 추출한 뒤에 세부 메타데이터를 추출하기 때문에 데이터를 구축하는 데에 두 단계를 거친다. 이 과정을 하나의 단계로 압축하여 학술지 논문의 첫 번째 면에서 서지정보 영역 추출 없이 한 번에 세부 메타데이터 추출이 이루어지는 모델의 실사용 가능성을 확인하기 위하여 추가처리 없이 학술논문에서 세부 메타데이터를 추출하도록 모델을 구성하여 추가 실험을 진행하였다.

참고문헌에서의 메타데이터 추출은 규칙 기반 방법을 통한 참고문헌 영역 인식 및 분리, 심층신경망 기반 모델의 지도학습을 통한 상세 메타데이터 추출로 구성하였다. 학술논문 메타데이터 추출의 정보 원천으로 이용한 상위 40종의 학술지에서 참고문헌 영역의 특징을 파악하여 참고문헌 시작지점을 규칙 기반으로 인식하고 분리했으며, 위의 방법으로 수집한 참고문헌 데이터 집합을 한국과학기술정보연구원이 제공한 참고문헌 메타데이터 파일을 기준으로 지도학습용 데이터 집합으로 자동 구축을 진행하였다. 구축된 데이터 집합을 통해 Bidirectional-GRU-CRF 기반 모델에 사전학습된 언어 모델을 더하여 높은 추출률을 가진 참고문헌 메타데이터 추출 모델을 구성할 수 있다.

논문의 구성은 다음과 같다. 2장은 PDF 원문에서 메타데이터를 추출하기 위해 연구되었던 다양한 사전 연구에 대해서 분석한다. 3장은 본 논문에서 제안하는 2단계 메타데이터 추출 방법론과 참고문헌 추출 방법론에 대해서 세부적으로 설명한다. 제안된 방법론의 효용성 평가를 위한 성능 평가 실험과 결과에 대한 설명은 4장에서 다룬다. 5장에서 결론을 내리고 향후 연구 방향을 소개한다.

2. 관련 연구

1990년 후반부터 학술논문 출판 유통 및 정보 서비스 등을 위한 전자 파일 형식으로 PDF가 널리 사용됨에 따라 PDF 파일 내에서 주요 정보를 추출 및 활용하는 연구가 계속 진행되고 있다(Tkaczyk et al., 2015). 다른 자연어처리 연구와 유사하게 PDF 파일에서 메타데이터와 같은 구조적 정보를 추출하는 연구는 초기에는 추출 규칙, 지식베이스 및 통계 기반 방법(Besagni & Belaïd, 2004; Powley & Dale, 2007)을 중심으로 진행되었고, 최근 들어서는 대부분 학습 데이터 구축을 통한 기계학습 기반 방법에 관한 연구(Tkaczyk et al., 2015; Kovačević et al., 2011; Granitzer et al., 2012; Tkaczyk et al., 2012; Souza, Moreira, & Heuser, 2017; An et al., 2017; Liu et al., 2017)가 중점적으로 진행되고 있다. 전 세계적으로 유관 연구가 활발하게 진행되고 있는 반면에 현재 학술정보 데이터베이스 구축과 서비스가 중요하게 대두됨에도 국내에서는 이와 관련된 연구는 미비한

실정이다(김선우 외, 2019).

Kovačević et al(2011)은 PDF 내의 모든 텍스트 행들에 대해 지지 벡터 기계(Support Vector Machines, SVMs) 기반의 분류를 통해서 각 행이 메타데이터의 어느 요소에 속하는지를 자동으로 식별하는 방법을 사용하였다. 이를 위해서, 개별 행에 대한 서식, 위치, 단어 등과 관련된 자질들을 선정하고 이를 분류 모델에 반영하였다.

Granitzer et al.(2012)에서는 텍스트 자질(구문, 의미, 사전 등)은 물론 PDF에서 제공하는 레이아웃(layout) 자질까지도 활용하여 조건부 랜덤 필드(Conditional Random Fields, CRFs) 및 2단계 SVMs 기반의 시퀀스 레이블링 방법을 사용하여 메타데이터 요소 추출을 시도하였다.

Tkaczyk et al.(2012)은 메타데이터 추출과정에서 최초로 영역(zone) 추출 개념을 도입하였다. 영역 추출을 위해서 은닉 마코프 모델(Hidden Markov Model, HMM)을 사용하였고, 추출된 영역 내에서의 추가적인 추출은 단순한 문자열 분리, 삭제, 하부 문자열 추출 등의 방법을 적용하였다. 영역 추출 모델이 특정 양식에 의존적인 관계로 단일 양식의 논문들에 대한 메타데이터 추출 실험만 진행하였다.

Souza, Moreira, Heuser(2017)는 본 논문에서 제안하는 메타데이터 추출 방법과 유사하게, 2개의 CRFs 모델을 이용한 순차적 추출 기법을 제안하였다. 첫 번째 단계에서는 주로 머리글(header), 제목, 저자 정보, 본문, 바닥글(footnote) 등과 같이 메타데이터 요소들을 포함하는 영역을 추출하고, 두번째 단계에서 세부 요소들을 추출하도록 시스템을 구성하였다. 그러나 이 연구는 특정 양식의 학술지와 학술대회 논문에 대한 정확한 추출에 초점을 맞추고 있었으므로, IEEE, Elsevier, Springer 및 ACM에서 출판된 논문 100건에 대한 성능 실험만 제한적으로 수행하였다. 또한, 2개의 CRFs 모델을 학습하기 위한 학습 데이터를 별도로 구축하고 세부적인 자질 집합을 결정하는데 많은 비용이 발생하는 단점이 존재한다. CERMINE(Content ExtRactor and MINEr)(Tkaczyk et al., 2015)은 과학 분야의 논문에서 메타데이터 및 참고문헌을 추출하고 그 결과를 온라인 및 XML 문서로 구조화하여 제공하는 자바 라이브러리이다. 논문의 구조를 문자, 페이지, 영역 수준으로 분석하고 SVMs 및 규칙 집합 등을 이용하여 요소 정보를 추출한다. 현재 가장 널리 알려지고 활용되는 논문 메타데이터 추출 도구로서 인정받고 있으나 영어 외의 다른 언어를 제대로 지원하지 못하고 있다.

그 외에도 김선우 외(2018)는 PDF의 내용을 이미지로 간주하고 컨볼루셔널 신경망(Convolutional Neural Network, CNN) 모델을 이용한 문자 인식 혹은 이미지 분석을 통해서 추출하는 연구를 시도하였으나 추출 대상 요소나 학술논문 양식 등에 제약이 있다. 최근 국내에서도 한국어 학술논문에 대한 메타데이터 추출 연구가 진행되고 있다. 김선우 외(2018)는 한국과학기술정보연구원(KISTI)이 보유한 상위 10종의 학술지의 학술논문을 XML 형식으로 변환하여 XML 마크업 토큰을 대상으로 학습 집합을 직접 구축하고 대상 요소에 대한 PDF 내에서의 위치 정보, 크기 정보, 텍스트 정보 등을 자질로 활용한 Bidirectional GRU-CRF 모델을 사용하여 메타데이터 추출을 시도하여 비교적 우수한 성능을 보였다. 그러나 대상 학술지의 종류가 10

종이고 실험에 사용된 논문 규모가 총 19,232건으로 제한적인 관계로 제안 모델의 범용성 및 양식 다양성 처리 능력을 검증하기에는 부족한 부분이 있었다.

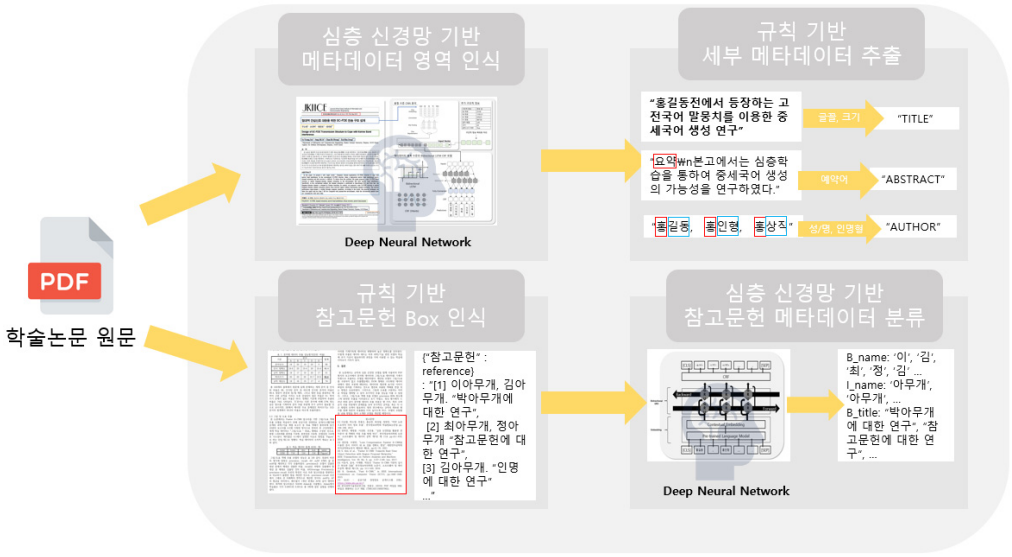
본 논문은 선행연구의 한계점을 보완하기 위해 학술논문 메타데이터 추출이 가능한 대상 학술지의 종류를 최대 40종으로 선정하여 실험에 사용된 논문의 규모를 31,188권으로 확대하였으며 학술논문의 서지정보 영역을 인식하는 모델을 통해 학습 집합을 구축하여 영역 추출 및 좌표 분석 등의 전처리 작업이 필요 없이 텍스트 형식의 학술논문을 입력하면 메타데이터를 추출하는 모델을 제안하였다.

3. 학술논문 메타데이터 추출

본 논문이 제안하는 학술논문 서지 메타데이터 자동 생성 방안은 개개의 학술논문에 대한 메타데이터 추출과 해당 논문의 참고문헌 메타데이터 추출로 구분된다. 〈그림 1〉은 본 논문에서 제안하는 메타데이터 추출 방법론을 단순화한 도식이다.

3.1 학술논문 메타데이터 추출

먼저 학술논문 서지정보 영역 추출은 논문의 첫 번째 면에서 요약정보 영역에 해당하는 부분을 인식하는 모델을 구축・훈련하고 학습이 끝난 모델이 인식한 영역에서 목표 서지 메타데이터 요소를 특정할 수 있는 규칙을 적용하여 세부 메타데이터를 추출하는 두 단계로 구성된다. 본 논문은 임수현 외(2019)에서 구축한 10종의 학술지에 대한 학습 집합을 기반으로 40종의 학술지에서 추출한 데이터를 학습 집합에 추가하여 서지정보 영역 추출 모델의



〈그림 1〉 서지정보 및 참고문헌 요약 생성 모델 작동 도식

훈련에 이용되는 학습 데이터를 구성한다. 학습 데이터 구축의 효율을 위하여 기존에 구축된 메타데이터 집합을 활용하여 영역을 자동으로 매핑하는 모델을 개발하고 이를 수작업으로 검증하였다. 구축된 학습 집합을 활용하여 지도학습 기반의 메타데이터 영역 자동 추출 모델을 구성하였다.

3.1.1 학습집합 가공 및 자질 정보

본 논문은 학습 집합의 구성을 위해 한국과학기술정보연구원(KISTI)으로부터 2010년 이후의 국내 학술지 및 메타데이터 총 144,786건을 지원받았다. 학술지별 논문의 숫자를 파악하여 2018년도 발행 권수 기준으로 상위 40종의 학술지를 선정하였다. 학습 집합 내 학술지 간 논문 숫자의 불균형으로 학습 데이터의 자질이 특정한 학술지의 특징에 과도하게 편향되는 현상을 방지하기 위해 선정한 40종의 학술지에서 출간된 논문의 평균 숫자인 1,399권으로 데이터 원천을 제한하였다. 〈표 1〉과 같이 "한국산학기술학회논문지"를 비롯한 9종의 학술지는 제한량을 초과하는 논문을 출간하였지만, 최신 논문 순서로 제한량만큼을 학습 집합 구축 대상으로 활용하였다. 학습 집합 구축 결과 40종의 학술지에 대한 총 38,337건의 학술논문과 메타데이터 집합이 도출되었다.

〈표 1〉 모델 학습에 투입된 데이터 원천 통계

No.	학술지명	논문수	No.	학술지명	논문수
1	한국산학기술학회논문지	1,399	21	한국식품과학회지	876
2	한국콘텐츠학회논문지	1,399	22	한국수자원학회논문집	850
3	디지털융복합연구	1,399	23	Journal of Power Electronics	841
4	한국정보통신학회논문지	1,399	24	전기학회논문지	798
5	한국컴퓨터정보학회논문지	1,399	25	한국응용과학기술학회지	778
6	아세아태평양축산학회지	1,399	26	Nuclear Engineering and Technology	764
7	한국인터넷방송통신학회논문지	1,399	27	한국과학교육학회지	664
8	대한토목학회논문집	1,399	28	한국축산식품학회지	659
9	한국융합학회논문지	1,399	29	BMB Reports	654
10	KSII Transactions on Internet and Information Systems	1,379	30	응용통계연구	640
11	Journal of Microbiology and Biotechnology	1,352	31	한국건설관리학회논문집	637
12	생명과학회지	1,302	32	한국의류학회지	628
13	한국전자통신학회논문지	1,272	33	디지털콘텐츠학회 논문지	622
14	대한수학회보	1,137	34	Molecules and Cells	600
15	Journal of Korean Neurosurgical Society	1,135	35	인터넷정보학회논문지	592
16	한국식품영양학회지	992	36	Biomolecules & Therapeutics	585
17	한국멀티미디어학회논문지	990	37	공업화학	580
18	정보보호학회논문지	935	38	한국항해항만학회지	575
19	Korean Chemical Engineering Research	910	39	The Korean Journal of Physiology and Pharmacology	548
20	한국재료학회지	907	40	한국문헌정보학회지	544
Total					38,337

학습 데이터 집합 구성 이후, 선택된 논문에서 문단, 문장, 글자 등 각개 텍스트의 값과 위치 정보 등을 추출하기 위해 내부를 분석하였다. 파이썬 공개 라이브러리인 PDFMiner를 활용하여 PDF 문서를 구성하는 문자, 숫자, 기호 등 메타데이터의 구성 요소가 될 수 있는 모든 글자(character)와 이것의 좌표를 추출하였다. 이 좌표는 PDF 내에서 개별 글자의 왼쪽(x^{left}), 오른쪽(x^{right}), 위(y^{top}), 아래(y^{bottom})에 해당하는 위치 정보가 별개로 기록하여 추출하도록 처리하였다. 일부 학술지의 경우 글자의 순서가 눈에 보이는 순서와 다르게 구성되었기 때문에 PDF 내의 모든 글자에 대해서 좌표를 기준으로 재배열하는 작업을 거쳤다. 구체적으로 좌표를 기준으로 해당 글자의 라인(line)을 결정하고 동일한 라인 내에서는 x^{left}의 순서에 따라 문자 정렬 작업을 수행하였다. 그 결과 라인 단위로 분리된 본문 텍스트를 〈그림 2〉와 같이 추출할 수 있었다. 〈그림 2〉는 논문의 첫 번째 면에 존재하는 모든 문자를 PDFMiner를 이용하여 추출하고 각 문자의 좌표를 기준으로 가로 정렬 작업을 수행한 결과이다.

```
00 = {Line} DOI : 10.5392/JKCA.2010.10.11.380\n
01 = {Line} 관광안내서의 스토리텔링적 방문설득 메시지 구조\n
02 = {Line} -대충청 방문의 해 공식 안내서를 중심으로-\n
03 = {Line} Message Structure of Persuasive Storytelling in Travel Guide Booklets of \n
04 = {Line} Great Chungchung Visit Year\n
05 = {Line} 이정현\n
06 = {Line} 우송대학교 호텔관광경영학과 \n
07 = {Line} Jung-Hun Lee(lemucl@naver.com)\n
08 = {Line} 요약\n
09 = {Line} 본 연구는 2010년 대충청 방문의 해 공식 관광안내서에 나타난 방문설득메시지의 구조를 스토리텔링적 \n
10 = {Line} 관점에서 분석했다. 대상은 <2010 대충청방문의 해>, <오서유! 즐거유! 대충청 2010>, 영문판 <It's \n
11 = {Line} Daejeon, Daejeon Tour> 총 3종이었다. 각 안내서의 메시지를 질적내용분석한 결과는 다음과 같다. 첫째, \n
12 = {Line} 메시지는 '~하세요', '~합시다' 등 청유형 어미를 사용, 관광지로의 방문을 권유하고 있다. 이는 독자들에게 \n
13 = {Line} 친근감있게 방문설득하는 감성소구적인 발화이다. 둘째, 메시지는 '즐거운', '아름다운', '해맑은' 등의 형용사\n
14 = {Line} 로 상품의 매력을 수식한다. 셋째, 메시지는 '~습니다', '~합니다'와 같은 상대높임법의 문장을 통해 정중하\n
15 = {Line} 게 방문과 참여를 설득한다. 넷째, 영문 안내서는 비스토리텔링적인 문체로 단순정보전달에만 치중해 있는 \n
16 = {Line} 것으로 나타났다. 결론적으로, 관광안내서의 메시지 구성은 정확한 정보제공과 방문에의 초대를 다양한 형\n
17 = {Line} 용사, 청유형 어미, 상대존대법등을 적절히 배합해 친절하게 스토리텔링함으로써 인간의 감성에 소구하는 \n
18 = {Line} 방문설득메시지로 전달되어질 수 있는 것이다. \n
19 = {Line} ■ 중심어 ::관광안내서::스토리텔링::방문설득메시지\n
20 = {Line} Abstract\n
```

〈그림 2〉 x좌표를 기준으로 정렬된 서지정보

정렬 작업 이후, 같은 x좌표에 존재하는 개별 글자의 글꼴, 크기, x, y 좌표 간의 거리 등의 특성이 같은 줄을 서로 병합한다. 〈그림 2〉의 경우 'Line = 05'에 해당하는 '이정현\n'과 'Line = 06'에 해당하는 '우송대학교 호텔관광경영학과\n'는 동일한 특성을 가지고 있으므로 하나의 영역(area) 혹은 박스(box)로 병합된다. 학술논문은 상-하, 좌-우 순서로 내용이 순차적으로 전개되는 형태로 레이아웃이 구성되어 있고 각각의 요소 정보영역은 사각형으로 정렬되므로 본 논문에서는 이것에 "박스(box)"라는 명칭을 붙였다. 〈그림 2〉에 대한 그룹화 결과는 〈그림 3〉과 같다.

```
00 = {Box} DOI : 10.5392/JKCA.2010.10.11.380\n
01 = {Box} 관광안내서의 스토리텔링적 방문설득 메시지 구조\n-대충청 방문의 해 공식 안내서를 중심으로-\n
02 = {Box} Message Structure of Persuasive Storytelling in Travel Guide Booklets of \nGreat Chungchung Visit Yea
03 = {Box} 이정현\n우송대학교 호텔관광경영학과 \n
04 = {Box} Jung-Hun Lee(lemucl@naver.com)\n
05 = {Box} 요약\n
06 = {Box} 본 연구는 2010년 대충청 방문의 해 공식 관광안내서에 나타난 방문설득메시지의 구조를 스토리텔링적 \n관점에서 분
07 = {Box} ■ 중심어 ::관광안내서::스토리텔링::방문설득메시지\n
08 = {Box} Abstract\n
09 = {Box} This study analyzed the message structure of persuasive storytelling which are expressed in \n
10 = {Box} ■ keyword ::Travel Guide Booklet::Storytelling::Persuasive Message\n
11 = {Box} I. 서 론\n
12 = {Box} 역의 관광정보를 한곳에 수록한 관광정보안내책자이\n다. 이는 한 지역의 특색, 볼거리, 놀거리, 그리고 먹을\n거리 등
13 = {Box} 각 지자체에서 발간하는 다양한 종류의 관광안내서\n는 비인적 관광안내자로서의 임무를 수행하는, 특징지\n
14 = {Box} 접수번호 : #101014-001\n접수일자 : 2010년 10월 14일\n
15 = {Box} 심사완료일 : 2010년 11월 11일\n교신저자 : 이정현, e-mail :\nlemucl@naver.com\n
```

〈그림 3〉 〈그림 2〉의 정렬된 데이터를 "박스" 단위로 재정렬한 결과

〈그림 3〉에서 한 줄로 표시되는 결과는 학술논문상에서 같은 박스에 존재하는 요소 영역을 의미한다. 이것을 활용하여 각각의 박스에 대한 태깅(영역 명칭 부착)을 시도하였다. 해당 박스가 어떤 메타데이터 요소를 나타내고 있는지를 정하기 위해서 비교적 단순한 규칙을 사용하였는데, 예를 들어 해당 박스에 '초록'이라는 용어를 포함하고 있다면 이를 예약어로 삼아 '국문 초록'으로 태깅하고, '키워드' 혹은 '중심어'라는 용어가 포함되어 있다면 '국문키워드'로 지정하는 방식이다.

보통 학술논문의 첫 번째 면은 DOI, 국문 제

목, 영문 제목, 국문 저자, 영문 저자로 이어지는 순서로 전개되나, 학술지별 위치 등 차이가 존재하므로 본 연구에서는 40종의 학술지별 첫 페이지 전개 순서를 분석하였다. 이후 분석 결과를 논문에 적용하여 〈표 2〉에서 제시된 영역 레이블 집합을 기준으로 〈그림 3〉에서 도출된 요소 영역 박스 집합에 대한 순차적 태깅을 시도하였다.

위와 같은 반자동 학습 집합 구축 과정을 통해서 최종적으로 도출되는 메타데이터 영역 태깅 결과는 〈표 3〉과 같다.

〈표 2〉 서지사항 영역에 따른 메타데이터 레이블 목록

Label	Content	Label	Content
ko_title	국문 제목	en_title	영문 제목
ko_author	국문 저자	en_author	영문 저자
ko_org	국문 저자소속기관	en_org	영문 저자소속기관
ko_abstract	국문 초록	en_abstract	영문 초록
ko_keywords	국문 키워드	en_keywords	영문 키워드
doi	DOI	citation	서지정보
journal	학술지명	date	일자
correspond	교신저자	oscp	오픈엑세스/저작권
issn	ISSN	O	기타 영역

〈표 3〉 〈그림 3〉의 요소 영역 박스 집합에 대한 메타데이터 레이블 태깅 결과

Label	Context
O	Regular paper
doi	Journal of the KIECS. pp. 1315-1320, vol. 10, no. 12, Dec. 31 2015, t. 68, pISSN 1975-8170 | eISSN 2288-2189 http://dx.doi.org/10.13067/JKIECS.2015.10.12.1315
ko_title	주파수 공유기법을 적용한 RF 송수신기에 대한 연구
ko_author	김기중*?김종성**?배문관**
en_title	The Study on the RF Transceiver Applied to Cognitive Radio Method
en_author	Ki-Jung Kim*?Jong-Sung Kim**?Moon-Kwan Bae**
O	요 약
ko_abstract	본 논문은 군통신 데이터링크 장비에 대해서 주파수 자원을 효율적으로 이용하기 위하여 고속도약기법, 주파 수환경센싱기법, 수신2채널기법, 수신필터뱅크부 적용 등을 이용하여 동일한 주파수 대역을 사용하는 레이다 등 의 타장비와 공유할 수 있는 기법을 소개한다. …
O	ABSTRACT
en_abstract	In this paper Communication data link equipment on a high speed-hopping method to use the frequency resources efficiently for the frequency environment is introduced such as sensing techniques, using the same frequency band by using the received two-channel technique and the receive filter bank unit applied to be shared with other equipment such as radar and so on. …
ko_keywords	키워드 RF Transceiver, Datalink, CR, Radar, L Band RF 송수신기, 데이터 링크, 주파수 공유 기법, 레이다, L대역
O	비는 거의 포화상태로 주파수 자원의 수요가 공급에 Ⅰ. 서 론 비하여 매우 많기 때문에 주파수 부족 현상이 심각하 국내의 군통신에서 UHF-대역을 사용하는 통신장 게 대두된다. 본 논문은 개발 중인 전술데이 타링크 장...

반자동으로 구축한 학습 집합에는 필연적으로 오류가 포함되어 있을 수 있으므로 이를 수작업으로 검증하는 단계가 필요하다. 이를 위해서 본 연구에서는 검수 작업자가 효과적으로 작업을 수행할 수 있는 후처리 검증/수정 시스템을 개발하여 활용하였다. 시스템의 화면은 〈그림 4〉(지선영, 2021)와 같다.

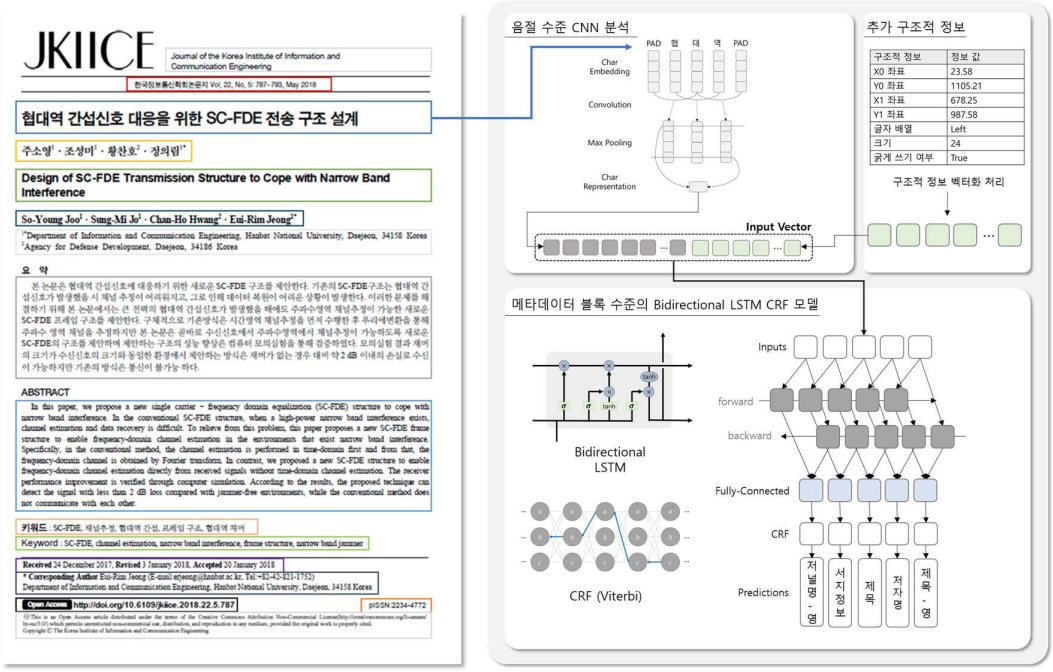
〈그림 4〉 후처리 검증/수정 시스템 화면 (지선영, 2021)

학술논문 파일과 추출한 서지정보 영역 자동 태깅 결과 파일이 쌍으로 입력되면 논문의 첫 페이지가 이미지화 화여 좌측에, 결과 파일이 도표 형태로 정리한 것이 화면 우측에 출력된다. 작업자는 우측 태깅 내용의 정합성을 좌측의 논문 첫 페이지 이미지와 비교하며 검수하고, 검수가 완료된 데이터는 정확성의 정도에 따라 '통과', '수정', '제외'의 3가지 유형으로 데이터를 분류한다. 자동 추출된 결과가 정확한 경우에는 '통과'로 지정하고, 데이터의 추출이 비교적 잘 되었으나 영역 태깅이 매끄럽지 못하여 해당 영역을 대상으로 하는 추출법에 수정이 필요한 경우는 '수정'으로 지정하며, '제외'로 분류된 데이터는 메타데이터 영역의 추출이 제대로 이루어지지 않은 것으로 간주하여 데이터 리스트에서 제외하였다.

학습 집합 구축을 위하여 총 38,337권의 논문 중 PDFMiner 실행 오류 및 PDF 인코딩 오류가 발생한 경우와 이미지 PDF를 제외한 35,942건의 논문에 대한 메타데이터 영역을 자동 태깅하고 수작업 검수 작업을 진행하였다. 검수 작업은 전체 데이터를 적절하게 5등분하여 5인의 석사급 연구원이 각자 7천여 건을 분류하는 작업을 약 2개월간 진행하였고, 주기적 회의를 통하여 작업 도중 발생하는 문제점 및 학술지 간 검증 기준 확립 등에 대한 논의와 '통과' 및 '수정'으로 판정된 데이가 적절하게 분류되었는지 검토하는 과정을 거치며 작업의 일관성을 유지하였다. 그 결과, 35,942건의 작업 대상 데이터 중에서 '통과' 및 '수정'으로 판정된 총 31,188건의 학습 집합을 최종적으로 구축하였다.

3.1.2 Bidirectional-GRU-CRF 기반 메타데이터 영역 추출 모델

본 연구에서는 논문에 대한 메타데이터 영역 추출을 순차적 레이블링 방법론으로 해결하였다. 제안 모델의 성능 향상을 위해서 텍스트 자질 정보를 적극적으로 수용할 수 있는 모델 구조를 채택하였다. 〈그림 5〉(지선영, 2021)는 본 연구에서 제안한 서지정보 영역 인식 및 추출 모델의 구조를 나타낸다. 〈그림 5〉의 좌측에서 박스로 감싼 각개 영역 모두가 합쳐져서 하나의 시퀀스를 구성한다고 가정하고 이 영역의 시퀀스를 구성하는 개별 영역에 대한 레이블을



〈그림 5〉 메타데이터 영역 추출 모델 도식(지선영, 2021)

예측하는 형태로 모델이 구동된다. 〈그림 5〉에서 들 수 있는 예시는, 최상단에 위치한 "Journal of the … Engineering" 박스에서 시작하여 최하단에 위치한 "This is an … Engineering" 박스까지를 순서대로 나열된 하나의 "영역 시퀀스"라고 가정한다.

PDF의 개별 영역 내 텍스트 정보와 해당 영역의 구조적 정보로 특정 영역에 대한 레이블을 지정하고, 이는 곧 하나의 시퀀스를 구성한다. 개별 영역 내의 문장은 음절과 어절단위로 각각 분리되어 CNN 레이어를 거쳐 자질 벡터로 표현된다. 이때, 자질 벡터는 음절과 어절에 대하여 개별적으로 구현하고, 필요에 따라 추가 및 제외하여 실험할 수 있도록 구성하였다. 자질 벡터는 단일 벡터로 병합되어 해당 영역에 대한 입력 자질 벡터로 모델에 삽입된다. 해당 모델은 Bidirectional GRU 모델에 CRF계층을 더한 모델 기반으로 작성하였으며, Bidirectional GRU 모델은 입력 벡터가 게이트를 통과할 때, 이전에 학습된 은닉층 정보의 양을 적절하게 조절하는 모델이다. Bidirectional GRU 셀의 연산은 입력 벡터에 대해 양방향으로 수행되며, 각 셀의 결과를 합쳐서 반환한다. CRF는 Bidirectional GRU 층을 통해 토큰 단위로 계산된 값을 기준으로 레이블을 예측하고, 해당 예측을 기준으로 다음 토큰의 레이블 확률을 예측하는 연산을 수행한다. 본 연구에서는 CRF 층에서 최종 계산한 Log-Likelihood를 최대화하는 방향으로 Bidirectional GRU 모델을 학습하고 CRF 층의 전이 행렬에 Viterbi 알고리즘을 적용하여 레이블을 도출한다.

3.1.3 규칙 기반의 메타데이터 추출

앞서 추출한 메타데이터 17종의 추출 결과를 바탕으로 항목별 특수문자, 키워드 등의 불필요한 데이터를 제거하고 메타데이터를 추출하는 후처리 모델을 통해 메타데이터 16종을 추출하였다. 주된 후처리 과정은 크게 3가지로 구분할 수 있다. 먼저 DOI, 소속기관 등과 같이 일정한 규칙이 있어 정규 표현식을 활용하여 추출이 가능한 경우와 두 번째로 성, 이름 등과 같이 일정한 규칙이 있으나 정규 표현식 사용이 불가한 경우, 마지막으로 키워드 영역에서 '|', ',' 등의 구분자와 같이 불필요한 특수문자 또는 '키워드', '중심어' 등의 단어를 제거하는 경우이다. 메타데이터 추출 모델을 통해 추출할 수 있는 메타데이터 12종 및 후처리 종류는 〈표 4〉와 같다.

〈표 4〉 메타데이터 후처리 방식

	Metadata	Proc type
1	ko_author	규칙기반
2	en_author (first)	
3	en_author (name)	
4	DOI	정규 표현식
5	ko_group	
6	en_group	
7	ko_abstract	특수기호 및 단어 제거
8	en_abstract	
9	ko_keywords	
10	en_keywords	
11	ko_title	-
12	en_title	-

정규 표현식을 사용하여 추출한 메타데이터는 국문 저자소속, DOI 등 25종이다. DOI는 양식이 정해져 있으므로 정규 표현식을 사용하는 것이 적합하며 저자소속기관의 경우 추출 대상 학술지 40종의 소속기관 데이터를 모두 수집하여 규칙을 파악했다. 소속기관 데이터는 기관명 및 부서명의 경우가 대학교, 연구소, 공기업, 사기업 등 타 메타데이터에 비하여 다양하고 표기 방법 또한 자유롭다. 예를 들어 대학교의 경우 기관은 '대학교', '대학', '학교' 등 일정한 형태로 끝나지만, 부서명의 경우는 '학부', '학과', '부', '전공', '과', '교육원', '팀' 등 다양한 형태로 존재한다. 따라서 기관명 및 부서명을 추출하는 정규 표현식의 구현은 되도록 수많은 명칭을 아우를 수 있도록 국문저자소속(기관), 국문저자소속(부서), 영문저자소속(기관), 영문저자소속(부서)을 언어별 1개 정규 표현식을 작성하였고, 각각의 정규 표현식은 국문의 경우 '학교', '연구소', '연구원', '기술원', '센터' 등 20개의 기관명칭과 부서명칭을 구분할 수 있도록 하였다. 영문의 경우 'S/school', 'U/univ of', 'U/univ.', 'C/college', 'T/technology' 등 12개의 기관명칭과 부서명칭을 구분할 수 있도록 하여 되도록 다양한 소속기관명 및 부서명 데이터를 추출하고자 하였다.

규칙 기반 방식으로 추출한 세부 메타데이터는 국문 저자명, 영문 저자명(성), 영문 저자명(이름) 등 총 3종이다. 저자명의 경우 저자 메타데이터 영역이 저자명, 교신저자 기호, 저자 기호 등으로 구성되어 있고 소속기관 또는 이메일 정보와 같은 부가 정보까지 들어있는 경우가 있다. 따라서 소속기관 등의 부가 정보를 제외하고 구분자(Delimiter)를 기준으로 저자 단위로 구분한 뒤 교신저자 기호, 저자 기호 등의 기타정보가 존재하는 경우, 이것을 저자명과 분리한 뒤에 저자의 성과 이름을 추출하였다. 한글 저자명의 경우 저자명 첫 글자를 성으

로, 나머지 글자를 이름으로 간주하였으나 영문 저자명의 경우 구분자에 따라서 성과 이름의 표기 방법이 성 뒤에 이름이 오거나 이름 뒤에 성이 오는 등 일정하지 않고 성과 이름을 구분하는 구분자 또한 다양하게 사용되기 때문에 정규 표현식을 활용하지 않았다.

특수기호 및 단어 제거로 후처리를 진행한 메타데이터는 국/영문초록, 키워드 등 총 4종이다. 국문초록과 영문초록의 경우 초록임을 나타내는 단어('초록', 'Abstract', '요약') 등을 제거하였다. 키워드의 경우 키워드임을 나타내는 단어('키워드', 'Keywords' 등)를 제거하고 키워드 문자열 전체를 구분자를 기준으로 나누어 추출된 텍스트를 개별 키워드 단위로 분리하였다. 구분자는 쉼표, 가운뎃점, 빗금(/) 등이 사용되었다. 국문제목과 영문제목의 경우 추출된 텍스트 전체가 제목을 의미하기 때문에 데이터 후처리를 진행하지 않았다.

3.2 참고문헌 메타데이터 추출

참고문헌 메타데이터 추출은 규칙 기반으로 참고문헌이 나열된 영역을 인식하고, 지도학습 기반의 순차적 레이블링 모델을 통한 참고문헌 메타데이터 추출 단계로 정했다. 순차적 레이블링 모델은 지선영, 최성필(2021)에서 구축한 모델을 기반으로 하는 참고문헌 메타데이터 추출 모델을 구성하였다. 학습 집합 구축의 효율성을 높이기 위하여 학술문헌의 구조 정보를 활용하여 참고문헌 영역을 인식한다. 참고문헌은 한국과학기술정보원이 제공하는 참고문헌 메타데이터 집합을 이용하여 레이블링을 진행하였다. 이렇게 구축된 학습 집합으로 참고문헌 메타데이터 추출 모델을 구성하였다.

3.2.1 규칙 기반 참고문헌 영역 인식 및 추출

본 논문은 국내 학술지에서 발행된 학술논문에 포함된 참고문헌의 메타데이터 추출용 학습 데이터를 구축하기 위하여 3.1과 동일한 상위 40종의 학술지를 이용하였다. 모델에 학습 데이터를 삽입하기 위하여 논문을 텍스트형 정보로 재구성하였다. 이 과정은 학술논문을 페이지 단위로 분석한 뒤, 참고문헌 영역의 특징을 통해 논문 전체에서 참고문헌 영역에 해당하는 부분만 지정할 수 있었다.

참고문헌 영역에 해당되는 부분은 "References", "참 고 문 헌", "참고문헌(REFERENCES)" 등의 예약어를 통하여 인식하였고, 각개 참고문헌의 구분은 흔히 사용되는 "[순서번호]", 혹은 들여쓰기로 구분하였다. 들여쓰기로 참고문헌을 구분하는 경우는 들여쓰기로 시작되는 참고문헌의 box의 x좌표가 일정한 것을 이용하여 해당 부분을 참고문헌 시작점으로 간주하여 개별 참고문헌으로 분리하였다.

3.2.2 참고문헌 메타데이터 자동 추출 학습 집합 구축

학습 집합의 구축을 위하여 53,562권의 학술논문을 분석하여 일정한 양식을 갖춘 총 791,288건의 참고문헌 정보를 획득하였다. 이중 학습 집합으로 사용하기에 적절한 것을 분류하기 위하여 한국과학기술정보연구원에서 제공한 참고문헌 메타데이터 파일을 기준으로 본 연구를 통해 추출한 참고문헌이 해당 메타데이터 파일에 존재하며, 일치율 100%를 보이는 참고문헌만 데이터로 사용하였다. 띄어쓰기 오류와 갖

추어야 할 정보가 부족한 경우, 중복의 제거, 학습 데이터로의 사용이 가능한지에 대한 여부를 따져 추가 선별 작업을 거쳤다. 참고문헌 메타데이터 중 종류에 무관하여 가장 자주 등장하는 메타데이터인 제목, 저자, 발행연도를 기준으로 해당 3종의 메타데이터가 모두 존재하는 참고문헌만 학습 데이터로 가공하였다. 웹 정보원은 정규 표현식을 사용하여 특징을 파악할 수 있으므로, 해당 기준으로 제외하지 않았다.

학습 집합으로 적합한 참고문헌 161,319개와 한국과학기술정보연구원에서 제공한 각 참고문헌의 메타데이터를 매칭하여 자동으로 레이블을 태깅하였다. 언어모델에 삽입가능한 토큰 한계점인 512를 넘는 학습데이터가 최종 제외되어 총 161,315개 길이의 학습데이터를 구축하였다.

〈표 5〉 참고문헌 분류 모델 학습 데이터 집합 통계

	Count
전체 참고문헌	791,288
부적합 데이터 제거	161,319
실험 집단 구축 대상	161,315

3.2.3 Bidirectional-GRU-CRF 기반 참고문헌 메타데이터 자동 추출 모델

본 논문은 참고문헌에 대한 메타데이터 추출을 순차적 레이블 지정 방법론으로 진행하였다. 제안 모델의 성능 향상을 위해서 텍스트 자질 정보를 적극적으로 수용할 수 있는 모델 구조를 채택하였다. 참고문헌은 특수문자 및 띄어쓰기 단위로 토큰화되어 모델에 입력된다. 개별 토큰은 사전 학습된 언어 모델을 통해 문맥적 자질을 포함한 단어 임베딩을 도출한다. 단어 임베딩은 양방향 게이트 순환 유닛의 입력 자질로 활용되며, 단어임베딩 중에서 특정 차례에 입력된 토큰에 대한 Bidirectional-GRU-CRF 모델의 연산 과정은 3.1.2.에 등장한 모델의 연산과정과 같다.

4. 실험 및 분석

4.1 학술논문 서지 메타데이터 추출 모델

4.1.1 데이터 집합 및 실험 환경

모델의 학습 및 실험을 위하여 검수 완료된 31,188권의 학술지를 학습 데이터와 평가 데이터로 구성하였으며 비율은 각각 8:2로 설정하였다. 전체 학습 및 실험을 위해 구축된 데이터 집합에 대한 통계는 〈표 6〉과 같다.

〈표 6〉 서지정보 영역 인식 및 추출 모델 학습 데이터 집합 통계

Type	Count
Train	24,951
Evaluation	6,237
Total	31,188

심층신경망 모델은 모델 학습에 활용하는 다양한 변수 변화에 민감하므로 해당 데이터와 모델 구조에 적합한 변수를 선정하기 위한 성능 최적화 실험을 수행하였다. 학습에는 학습 데이터의 10%가 활용되었고, 성능 측정에는 평가 데이터가 활용되었다. 활용한 평가 방법은 예측된 BIO 토큰이 부착된 메타데이터에 대해 BIO 태그에 대한 청킹(chunking) 작업을 진행한 이후에 정확도와 재현율, F1 성능을 측

정하는 형식을 활용하였다. 다음 〈표 7〉은 성능 최적화 실험에 활용된 하이퍼파라미터의 종류와 범위이며 실험 결과 최종선정된 하이퍼파라미터에는 볼드 및 밑줄 표시로 강조하였다.

〈표 7〉 성능 최적화 실험에 활용된 하이퍼파라미터의 종류와 범위

Hyperparameter	Range of hyperparameter
Number of dimensions in word embedding vectors	**300**, 500
Number of dimensions in char embedding vectors	**300**, 500
Number of CNN cell	8, **16**
Number of dimensions in GRU cell	**300**, 500
Dropout	**0.7**, 0.9
Learning rate	0.01, **0.001**

이후 〈표 7〉의 하이퍼파라미터 세트를 활용하여 전체 모델에 대한 실험을 진행했다. 메타데이터 영역 추출 모델 학습에 사용한 주요 하이퍼파라미터는 〈표 8〉과 같다.

〈표 8〉 실험에 사용된 하이퍼파라미터

Type	Hyperparameter
word embedding	300
char embedding	300
GRU dims	300
Filters	2,3,4
Number of Filter	16
Batch size	8
Dropout	0.7
Learning rate	0.001
Early Stopping	3
Optimizer	Adam

4.1.2 실험 결과

학술논문의 첫 번째 면에서의 메타데이터 영역 추출 모델을 자질별로 실험한 후 결과를 비교하였다. 모든 성능 수치는 소수 셋째 자리에서 반올림한 값이며 실험 결과는 〈표 9〉와 같다. 성능이 가장 높았던 모델은 입력 자질로 음절 및 어절 임베딩을 활용한 모델이며 정확도 97.32%, 재현율 97.23%, F1 점수 97.27의 성능을 보였다. 어절 자질과 음절 자질을 개별적으로 사용한 경우 어절 자질이 음절 자질보다 정확도는 2.29%, 재현율은 2.47 그리고 F1 점수는 2.38 높았다. 이는 어절 및 음절 자질이 메타데이터 영역 추출 모델에 대해 유의미한 영향을 끼친 것을 보여주며 각 자질을 동시에 사용했을 때 더 효과적임을 알 수 있다.

〈표 9〉 서지정보 영역 인식 및 추출 모델 학습 결과

model	Accuracy	Recall	F1
Bidirectional-GRU-CRF + Char embedding	94.23	93.93	94.08
Bidirectional-GRU-CRF + Word embedding	96.52	96.4	96.46
Bidirectional-GRU-CRF + Char + Word	**97.32**	**97.23**	**97.27**

규칙 기반 세부 메타데이터 추출 모델의 성능을 측정하기 위하여 평가 집합의 학술논문 1,767개를 대상으로 메타데이터 추출 실험을 진행하였다. 메타데이터 데이터베이스와 모델이 예측한 세부 메타데이터가 100% 일치하는지 여부를 기준으로 정확도를 측정하였다. 실험 과정에서 추출 결과와 매핑 사전 내의 존재하는 텍스트 값 내의 특수기호, 공백 등이 외관상 일치하나 실제적인 값이 일치하지 않는 등의 문제가 있었다. 이에 실험 과정에서는 특수기호와 공백 정보를 모두 제거한 후에 비교하였다.

〈표 10〉은 메타데이터별 추출 실험 결과이다. 추출이 불가하거나 논문에 해당 메타데이터가 존재하지 않는 경우는 성능을 측정하지 않았다.

〈표 10〉 메타데이터별 추출 성능

Metadata	Accuracy
ko_title	91.73
en_title	87.80
ko_abstract	85.27
en_abstract	89.42
ko_group	27.07
en_group	22.08
ko_author	86.78
en_author (first)	72.64
en_author (name)	74.11
ko_keywords	82.60
en_keywords	87.28
DOI	99.79

실험 결과 40종 학술지의 12종 메타데이터 기준으로 전체 평균 75.8%의 성능을 도출하였다. 대부분의 메타데이터에서 80%가 넘는 정확도를 도출했으나 영문 저자명(성, 이름)은 각각 72.64%, 74.11%로 국문 저자명 86.78%보다 낮았으며, 국문 소속기관 및 영문 소속기관도 각각 27.07%, 22.08%로 낮은 성능을 보였다. 이는 최대한 넓은 범위의 규칙을 적용했음에도 불구하고 논문 1,767건의 모든 소속기관을 추출하기에는 한계가 존재하는 것으로 보인다.

4.1.3 추가 실험 데이터집합 및 실험 환경

본 연구는 학술논문에서 요약된 서지정보를 생성하기 위하여 심층신경망 모델을 사용한 서지정보 영역 추출 후 해당 영역에 알맞은 규칙을 적용하는 방법을 제안하였다. 제안하는 방법의 첫 단계인 영역 추출과정 없이 학술논문에서 메타데이터를 바로 추출하는 모델의 구현 가능성을 확인하기 위하여 서지정보 영역 추출에 사용한 모델에 학술논문의 구조 정보 없이 텍스트 정보만 활용한 학습 데이터를 구축하여 추가 실험을 진행하였다.

추가 실험 모델의 학습 및 실험을 위하여 서지정보 영역 추출 모델의 학습 데이터를 가공하였다. 서지정보 영역 단위로 구성된 학습 데이터와 세부 메타데이터 데이터베이스를 매핑한 뒤 단어 단위로 토큰화 과정을 거쳐 CoNLL 형식으로 학습 데이터를 구성하였다. 〈표 11〉은 서지정보 영역 추출 모델의 학습데이터(Orginal)에 세부 메타데이터 데이터베이스를 매핑한 결과(Result) 중 일부이다.

구축한 학습 데이터는 학습 데이터와 평가 데이터로 구성하였으며 비율은 각각 8:2로 설정하였다. 전체 학습 및 실험을 위해 구축된 데이터집합에 대한 통계는 〈표 12〉와 같다.

이후 〈표 13〉의 하이퍼파라미터를 활용하여 추가 모델에 대한 실험을 진행했다. 추가 모델의 구조는 앞서 사용한 서지정보 영역 추출 모델과 동일하며 입력 자질은 구글에서 공개한 단어 임베딩인 'fastText'를 사용하였다. 영역 추출 모델 학습에 사용한 주요 하이퍼파라미터는 〈표 13〉과 같다.

4.1.4 추가 실험 결과

학술논문 첫 번째 면의 서지정보 영역을 대상으로 세부 메타데이터를 추출한 추가 실험 결과는 〈표 14〉와 같다. 정확도 73.60%, 재현

〈표 11〉 메타데이터 영역 추출 모델의 학습 데이터에 메타데이터 데이터베이스를 매핑한 결과 중 일부

Label	Orginal	Label	Result
O	Regular paper	O	Regular paper
doi	Journal of the KIECS, pp. 1315-1320, vol. 10, no. 12, Dec. 31 2015, t. 68, pISSN 1975-8170 | eISSN 2288-2189 http://dx.doi.org/10.13067/JKIECS.2015.10.12.1315	O	Journal of the KIECS, pp. 1315-1320, vol. 10, no. 12, Dec. 31 2015, t. 68, pISSN 1975-8170 | eISSN 2288-2189 http://dx.doi.org/
		doi	10.13067/JKIECS.2015.10.12.1315
ko_title	주파수 공유기법을 적용한 RF 송수신기에 대한 연구	ko_title	주파수 공유기법을 적용한 RF 송수신기에 대한 연구
ko_author	김기중*?김종성**?배문관**	ko_author	김기중
		O	*?
		ko_author	김종성
		O	**?
		ko_author	배문관
		O	**
en_title	The Study on the RF Transceiver Applied to Cognitive Radio Method	en_title	The Study on the RF Transceiver Applied to Cognitive Radio Method
en_author	Ki-Jung Kim*?Jong-Sung Kim**? Moon-Kwan Bae**	en_author	Ki-Jung Kim
		O	*?
		en_author	Jong-Sung Kim
		O	**?
		en_author	Moon-Kwan Bae
		O	**
O	요 약	O	요 약
ko_abstract	본 논문은 군통신 데이터링크 장비에 대해서 주파수 자원을 …	ko_abstract	본 논문은 군통신 데이터링크 장비에 대해서 주파수 자원을 …
O	ABSTRACT	O	ABSTRACT
en_abstract	In this paper Communication data link equipment on a …	en_abstract	In this paper Communication data link equipment on a …
ko_keywords	키워드 RF Transceiver, Datalink, CR, Radar, L Band RF 송수신기, 데이터 링크, 주파수 공유 기법, 레이다, L대역	O	키워드
		ko_keywords	RF Transcever
		O	,
		ko_keywords	Datalink
		O	,
		ko_keywords	CR
		O	,
		ko_keywords	Radar
		O	,
		ko_keywords	L Band RF 송수신기
		O	,
		ko_keywords	데이터 링크
		O	,
		ko_keywords	주파수 공유 기법
		O	,
		ko_keywords	레이다
이하 생략			

〈표 12〉 추가 실험 데이터집합 통계

Type	Count
Train	24,584
Evaluation	6,146
Total	30,730

〈표 13〉 추가 실험에서 사용한 하이퍼파라미터

Type	Hyperparameter
fastText dims	300
GRU dims	300
Dropout	0.7
Learning rate	0.001
Batch size	8
Early Stopping	3
Optimizer	Adam

〈표 14〉 세부 메타데이터 추출 성능 비교

메타데이터	규칙 기반 추출(제안하는 방법)	영역 추출 생략(추가실험)
ko_title	**91.73**	59.74
en_title	**87.80**	68.56
ko_abstract	**85.27**	61.20
en_abstract	**89.42**	43.76
ko_group	27.07	**39.48**
en_group	22.08	**81.69**
ko_author	**86.78**	80.10
en_author (first)	72.64	**76.98**
en_author (name)	74.11	
ko_keywords	**82.60**	4.41
en_keywords	**87.28**	33.78
DOI	**99.79**	97.26

율 56.61%, F1 점수 63.38%의 성능을 보였다. 보다 상세한 비교를 위하여 규칙 기반으로 추출한 세부 메타데이터의 정확도와 성능을 비교하였다. 그러나 규칙 기반 모델은 1차적으로 추출한 서지정보 영역을 기준으로 해당 영역에 알맞은 규칙을 적용하여 추출하는 것이기 때문에 학술논문 첫 번째 면의 전체 텍스트에서 메타데이터를 추출한 추가 실험 모델과 동일 선상에서 비교하기에는 어려움이 있다.

실험 결과 규칙 기반으로 추출한 성능이 더 높은 메타데이터는 국문 제목, 영문 제목, 국문 초록, 영문 초록, 국문 키워드, 영문 키워드, 국문 저자명이며 추가 실험을 통해 학술논문의 텍스트에서 직접 메타데이터를 추출한 성능이 더 높은 메타데이터는 국문 소속기관, 영문 소속기관, 영문 저자명이었다. 전반적으로 규칙 기반 방법을 통해 메타데이터를 추출하는 방법이 더 높은 성능을 보였으나 소속기관은 텍스트에서 직접 추출하는 것이 국문의 경우 12.41%, 영문의 경우 59.61% 더 높은 성능을 보였으며, 영문 저자명도 텍스트에서 직접 추출하는 것이 약 3.6% 더 높은 성능을 보였다.

4.2 참고문헌 메타데이터 추출 모델

4.2.1 학술논문 내 참고문헌 영역 추출 모델 성능 평가

규칙 기반 참고문헌 영역 추출 실험을 위하여 선정된 실험 대상 데이터는 40종의 학술지에서 참고문헌 리스트를 담은 데이터베이스를 기준으로 PDF 처리에 문제가 없는 논문을 학술지당 50권씩 무작위로 선별하여 총 2,000권으로 이루어진 실험 집합을 구성하였다. 데이터베이스 내에 존재하는 참고문헌 리스트와의 비교를 위해, 공백과 특수기호 제거, 소문자 처리 등의 기본적인 전처리를 거친 이후에 비교를 수행하였다. 학술논문에서 참고문헌 영역을 박스 단위로 추출하는 작업 이후, 참고문헌 영역 내에서 참고문헌 리스트를 추출하였다. 도출한 참고문헌 리스트를 데이터베이스와 비교한 결과는 〈표 15〉와 같다. 2,000권의 논문 내에 포함된 8,872건의 참고문헌 중에 7,603건의 참고문헌이 옳게 추출되었다.

〈표 15〉 규칙 기반 참고문헌 영역 추출 실험 결과

논문 수	2,000
전체 참고문헌 수	8,872
일치 참고문헌 수	7,603
정확도	86.59%

4.2.2 학술논문 내 참고문헌 메타데이터 추출 모델 성능 최적화 실험

심층신경망 기반으로 구현된 모델은 하이퍼파라미터의 조합에 따라 성능의 차이를 보인다. 따라서 본 논문은 성능 평가 이전에 최적의 하이퍼파라미터 조합을 찾기 위한 최적화 실험을 진행하였다. 〈표 16〉은 참고문헌 메타데이터 추출 모델과 관련된 실험에서 사용한 학습 집합의 통계이다.

〈표 16〉 참고문헌 메타데이터 추출 모델 학습 집합 통계

Type	Count
Train	129,044
Evaluation	32,261
Total	161,315

〈표 17〉은 최적화 실험에 사용된 하이퍼파라미터 유형과 각각의 수치 범위와 최적의 파라미터로 선정된 수치를 나타낸다. 최적화 실험 이후, 〈표 18〉의 하이퍼파라미터 값을 활용하여 참고문헌 메타데이터 추출 모델 학습을 진행하였다.

〈표 17〉 참고문헌 영역 추출 성능 최적화 실험에 사용된 하이퍼파라미터의 종류와 범위

Type	Comparison	Hyperparameter
GRU dims	**100**, 200, 300	100
Learning rate	0.01, **0.001**	0.001
Dropout	0.3, **0.5**, 0.7	0.5

〈표 18〉 참고문헌 메타데이터 추출 모델 실험에 사용된 하이퍼파라미터

Type	Hyperparameter
GRU dims	100
Dropout	0.5
Learning rate	0.001
Batch size	16
Early Stopping	3
Optimizer	Adam

4.2.3 실험 결과

본 논문은 참고문헌 메타데이터 추출 모델을 3.2.3에서 설명하였듯 사전 학습된 언어 모델을 통하여 실험한 후, 모델별 결과를 비교하였다. 그리고, 학습데이터의 양이 모델 성능에 끼치는 영향을 알아보기 위하여 성능이 가장 높았던 사전 학습된 언어 모델을 대상으로 학습 데이터의 양을 일정 비율로 조정하며 실험을 진행하였다. 모든 성능 수치는 소수 셋째 자리에서 반올림하였으며, 실험 결과는 〈표 19〉와 같다.

전반적인 성능은 BERT(multilingual base-cased) 모델이 정밀도(Precision) 96.06%, 재현율(Recall) 97.58%, F1 점수 96.82로 다른 모델에 비해 높았다. 국내 학술지에서 발행하는 학술논문의 참고문헌 영역은 국어를 비롯하여 영어, 중국어 등 외국어로도 표기되어 사전 학습된 언어모델 중 다국어를 지원하는 BERT(multilingual base-cased) 모델이 가장 높은 성능을 도출했다고 볼 수 있다.

그러나, 다른 모델과 F1 점수의 차이가 최대 2.7, 최소 0.3 정도로 3미만으로 작아 사용환경에 따라 적합한 언어모델을 사용하는 것이 권장된다. KoBERT의 경우 실험 대상 모델 중 사전 학습된 언어 모델을 구축할 때 추가로 학습된 한국어 데이터의 수가 가장 적은 모델이며, KoELECTRA-Base-v3는 한국어 데이터의 수가 가장 많은 모델이다. 이를 통하여 언어모델을 사전 학습할 때 투입된 한국어 데이터가 많을수록 한국어를 다루는 모델의 전반적인 성능이 증가하는 것으로 볼 수 있다.

모델의 성능이 가장 높았던 BERT(multilingual base-cased)모델을 대상으로 학습 데이터의 규모가 끼치는 영향을 알아보기 위하여 학습 데이터의 규모를 세분화하는 실험을 실시하였다. 학습 조건과 사용한 데이터는 모델별 비교 실험과 같다. 〈표 20〉은 규모별 성능 실험 결과이다.

〈표 19〉 참고문헌 메타데이터 추출 모델 성능 실험 결과

Model	Precision	Recall	F1
Bidirectional-GRU-CRF + KoBERT	93.34	94.91	94.12
Bidirectional-GRU-CRF + HanBERT	94.99	96.60	95.79
Bidirectional-GRU-CRF + KoELECTRA-Base-v3	95.68	97.37	96.52
Bidirectional-GRU-CRF + BERT(multilingual base-cased)	**96.06**	**97.58**	**96.82**

〈표 20〉 학습 데이터 규모별 참고문헌 메타데이터 추출 성능 비교

학습데이터 수	Precision	Recall	F1
6,452	95.69	67.39	96.53
12,904	95.73	97.53	96.62
25,809	95.81	97.62	96.71
51,618	**96.11**	97.48	96.79
103,235	95.98	**97.63**	96.80
129,044	96.06	97.58	**96.82**

정밀도 및 재현율은 모든 학습 데이터를 사용하지 않은 두 경우에 가장 높은 수치를 기록했지만, 다른 경우와의 차이가 미미하며, 총체적인 성능을 표시하는 F1 점수는 모든 학습 데이터를 투입하였을 때에 가장 높다. 그러므로, 학습 집합의 개수가 증가할수록 F1 점수역시 증가하여 더 높은 성능을 내는 것을 알 수 있다.

4.2.4 추가 실험 환경 및 결과

추가 실험은 제안하는 모델과 성능 비교를 위하여 언어모델 2종을 이용하여 학습을 진행하였다. 학습 데이터는 제안하는 모델과 같은 것을 사용하였으며, 모델의 기본 골조인 Bidirectional-GRU-CRF에 사전학습된 언어모델인 KcBERT와 KcELECTRA를 각각 붙여 추가 실험 모델을 준비하였고, 모델 모두 같은 하이퍼파라미터로 실험을 진행하였다.

추가 실험에 투입된 모델은 정확한 비교를 위하여 제안하는 모델의 실험과 완전히 일치하는 환경에서 실시하였다. 각각 언어모델별 성능 수치는 소수점 셋째 자리에서 반올림하였으며, 실험 결과는 〈표 21〉에 기록된 바와 같다.

공개되어있는 대부분의 한국어 BERT가 한국어 위키 등, 뉴스 기사와 책 등의 문어체로 작성되어 편집 과정을 거친, 잘 정제된 줄글을 사전학습을 위한 훈련 데이터 집합의 원천으로 삼는다. 그러나, KcBERT와 KcELECTRA는 일상적으로 사용되는 언어 인식을 위해 새로이 작성된 언어모델로, 공개 이전 사전학습 당시 훈련 데이터 집합으로 신조어를 포함한 구어체 위주, 작성자 자기편집 혹은 한 번도 수정되지 않은 문장을 정보 원천으로 한다. KcBERT의 경우 훈련 데이터 집합을 네이버 뉴스의 댓글에서 수집하였음을, KcELCTRA는 KcBERT에 구어체 데이터를 추가로 삽입하여 학습했음을 밝히고 있다(Lee, 2020). 두 언어모델은 각각 정밀도 면에서 93.47%와 93.52%를, 재현율 면에서 94.65%와 94.88%를 기록하였고, F1 점수의 경우 94.06와 94.19의 성능을 보였다. KcBERT보다 사전학습 데이터 집합이 큰 KcELECTRA가 더

〈표 21〉 추가 실험 결과

Model	Precision	Recall	F1
Bidirectional-GRU-CRF + KcBERT	93.47	94.65	94.06
Bidirectional-GRU-CRF + KcELECTRA	**93.52**	**94.88**	**94.19**

높은 성능을 기록하였다. 두 사전학습 언어 모델을 더한 성능은 4.2.3에서 사용한 사전학습 모델 중 가장 성능이 낮았던 KoBERT와 비슷한 성능을 보여 제안하는 방법과 모델에 가장 적합한 사전학습 언어 모델은 BERT(multilingual base-cased)에서 바뀌지 않았다.

5. 결 론

본 논문은 학술논문 서지 메타데이터 자동 생성을 위하여 디지털 학술지 40종을 분석하여 학술논문에서 추출할 수 있는 메타데이터를 논문 첫 면에 존재하는 논문에 대한 서지정보 영역과 논문 말미에 존재하는 참고문헌에 대한 정보영역으로 나누어 추출하였다. 논문 자체에 대한 메타데이터 추출은 심층신경망 모델을 통한 메타데이터 영역 분류와 규칙 기반으로 세부 메타데이터를 추출하는 2단계 과정을 취하였다. 먼저, 학술논문의 첫 번째 면에서 서지정보 영역을 구별하고 해당 영역에서 활용 가능한 음절, 어절 정보를 자질로 활용하여 Bidirectional-GRU-CRF 모델 구조를 기반으로 메타데이터 추출 모델을 구성하고 실험하였다. 각 자질을 조합하여 실험 및 비교한 결과, 음절 및 어절 자질을 사용한 모델의 F1 점수가 97.27로 가장 높은 결과를 보였으며, 이는 기존 연구인 김선우 외(2019)에서 논문 메타데이터 추출 실험에 사용한 'Bidirectional-GRU CRF' 모델의 F1 점수 84.39에 비하여 높은 수치이다. 세부 메타데이터 추출을 위하여 앞서 추출한 메타데이터 영역에서 규칙 기반으로 제목, 초록, 저자명, 소속기관, 키워드, doi 등 12종의 세부 메타데이터를 정규식, 기타 규칙, 단어 제거 등의 방법을 사용하여 추출하였으며 40종 학술지를 기준으로 평균 정확도 75.8%를 도출했다. 이어서 학술논문의 전체 텍스트에서 메타데이터 추출의 가능성을 확인하기 위하여 서지정보 영역 추출 과정을 생략하고 학술논문의 전체 텍스트를 기준으로 세부 메타데이터를 태깅하는 학습 데이터를 구축하여 Bidirectional-GRU-CRF 모델을 대상으로 실험하였다. 평균 정확도는 73.60%였으며 전반적으로 규칙 기반으로 추출한 메타데이터보다 성능이 저조하였으나 소속기관 및 영문 저자명의 경우 높은 성능을 보여 전체 텍스트 대상 메타데이터 추출의 가능성을 확인하였다.

참고문헌에 대한 메타데이터 추출은 규칙 기반으로 참고문헌 영역을 인식하고 심층신경망 모델을 통한 각개 참고문헌 내의 세부 메타데이터 분류를 진행하는 2단계 과정으로 연구를 진행하였다. 먼저, '참고문헌' 혹은 'REFERENCES' 등 40종의 학술지에서 보편적으로 보이는 참고문헌 시작점을 특징으로 삼아, "[1], [2]" 등 문헌과 문헌의 구분을 돕는 표식, 혹은 들여쓰기 규칙으로 인하여 개개 참고문헌의 시작점과 이어지는 참고문헌 메타데이터 서술의 좌표가 다른 것을 규칙으로 참고문헌 영역을 인식 및 추출하였다. 추출한 참고문헌 영역은 Bidirectional-GRU-CRF 기반으로 작성된 심층 신경망 모델에 사전학습된 언어 모델을 접합하여 4개 모델에서 비교하였고, 실험에 투입된 사전학습 언어모델 중 Bidirectional-GRU-CRF + BERT(multilingual base-cased)가 가장 높은 성능을 보였다. 해당 모델 조합에 학습 데이터 비율을 조정하여 투입한 결과, 학습 데이터를 많이

사용할수록 F1 점수가 높은 수치를 보이는 것을 확인했다. 추가 실험에서 사용한 2개의 사전학습 언어모델을 거친 결과를 비교하는 것으로 제안하는 방법에 가장 적합한 사전학습 언어모델은 정밀도 96.06%, 재현율 97.58, F1 점수 96.82를 각각 기록한 BERT(multilingual base-cased)임을 보였다. 이는 기존 연구인 김선우 외(2018)에서 제시한 F1 점수 97.21보다 낮은 수치이나, 본 논문이 제시한 F1 점수는 김선우 외(2018)에서 실험한 참고문헌 메타데이터 인식뿐만 아니라 추출작업까지 진행한다는 점에서 데이터베이스 생성 자동화 가능성까지 입증할 수 있다.

그러나 본 논문의 실험은 국내에서 출판된 과학기술 분야의 학술지를 대상으로 진행되었기 때문에 모든 분야의 학술지를 대상으로 적용하기에는 한계가 있으며 다양한 자질 및 데이터 가공 방법을 사용한 추가 실험의 가능성이 있다. 향후 더욱 다양한 학술지를 수집하여 처리 가능한 학술지의 종 수를 확장하는 연구와 함께 현재 활발하게 진행 중인 사전 학습된 언어모델을 활용한 언어 자질 보강 등의 연구를 수행할 것이다.

참 고 문 헌

김선우, 지선영, 설재욱, 정희석, 최성필 (2018). Bidirectional GRU-GRU CRF 기반 참고문헌 메타데이터 인식. 한국정보과학회 언어공학연구회: 학술대회 논문집(한글 및 한국어 정보처리), 30, 461-464.

김선우, 지선영, 정희석, 윤화묵, 최성필 (2019). 학술논문 PDF에 대한 딥러닝 기반의 메타데이터 추출 방법 연구. 정보과학회논문지, 46(7), 644-652.

김재훈, 김순영, 임석종, 황혜경 (2019). 학술논문과 참고문헌의 자동매핑 사례 분석. 한국콘텐츠학회논문지, 19(11), 262-269.

김지훈 (2003). 참조연결을 위한 인용정보 자동추출에 관한 연구. 한국문헌정보학회지, 37(1), 247-268.

임수현, 윤태린, 최경철, 조원민, 허재종, 한현우, 이경원 (2019). 학술 문헌 인용 계보 내 피인용지수를 이용한 참고문헌 탐색 인터페이스 제안. 한국HCI학회 학술대회, 526-529.

지선영 (2021). PDF 학술논문 메타데이터 자동 추출 연구. 석사학위논문, 경기대학교.

지선영, 최성필 (2021). 사전학습 된 언어 모델 기반의 양방향 게이트 순환 유닛 모델과 조건부 랜덤 필드 모델을 이용한 참고문헌 메타데이터 인식 연구. 한국정보관리학회지, 38(1), 221-242.

An, D., Gao, L., Jiang, Z., Liu, R., & Tang, Z. (2017). Citation metadata extraction via deep neural network-based segment sequence labeling. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, 1967-1970.

Besagni, D. & Belaïd, A. (2004). Citation recognition for scientific publications in digital libraries.

In First International Workshop on Document Image Analysis for Libraries, 244-252.

Granitzer, M., Hristakeva, M., Knight, R., Jack, K., & Kern, R. (2012), A comparison of layout based bibliographic metadata extraction techniques. In Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics, 2, 1-8.

Kovačević, A., Ivanović, D., Milosavljević, B., Konjović, Z., & Surla, D. (2011). Automatic extraction of metadata from scientific publications for CRIS systems. Program: electronic library and information systems, 45(4), 376-396.

Lee, J. (2020). KcBERT. GitHub. Available: https://github.com/Beomi/KcBERT

Liu, R., Gao, L., An, D., Jiang, Z., & Tang, Z. (2017). Automatic document metadata extraction based on deep networks. In National CCF Conference on Natural Language Processing and Chinese Computing, 305-317.

Powley, B. & Dale, R. (2007). High accuracy citation extraction and named entity recognition for a heterogeneous corpus of academic papers. In 2007 International Conference on Natural Language Processing and Knowledge Engineering, 119-124.

Souza, A., Moreira, V., & Heuser, C. (2017). ARCTIC: metadata extraction from scientific papers in pdf using two-layer CRF. In Proceedings of the 2014 ACM Symposium on Document Engineering, 121-130.

Tkaczyk, D., Bolikowski, L., Czeczko, A., & Rusek, K. (2012) A modular metadata extraction system for born-digital articles. In 2012 10th IAPR International Workshop on Document Analysis Systems, 11-16.

Tkaczyk, D., Szostek, Szostek, P., Fedoryszak, M., Dendek, P. J., & Bolikowski, Ł. (2015). CERMINE: automatic extraction of structured metadata from scientific literature. International Journal on Document Analysis and Recognition, 18, 317-335

Ziviani, N., Gonçalves, M. A., de Moura, E. S., Ribeiro-Neto, B., da Silva, A. S., & Veloso, A. (2011). Information Retrieval Research at UFMG. Journal of Information and Data Management, 2(2), 77-77.

• 국문 참고자료의 영어 표기
(English translation / romanization of references originally written in Korean)

Ji, Seon-Young & Choi, Sung-Pil (2021). A study on recognition of citation metadata using Bidirectional GRU-CRF model based on pre-trained language model. Journal of the Korean Society for information Management, 38(1), 221-242.

Ji, Seon-young (2021). A Study on Automatic Extrqaction of Metadata for papers in PDF format. Master's thesis, Kyonggi University.

Kim, Jae-Hoon, Kim, Soon-Young, Im, Seok-Jong, & Hwang, Hye-Gyung (2019). Case study of journal article and reference mapping. Journal of the Korea Contents Association, 19(11), 262-269.

Kim, Ji-Hoon (2003). A study on automatic extraction of citation information for reference linking. Journal of the Korean Society for Library and Information Science, 37(1), 247-268.

Kim, Seon-Wu, Ji, Seon-Young, Jeong, Hee-Seok, Yoon, Hwa-Mook, & Choi, Sung-Pil (2019). Metadata extraction based on deep learning from academic paper in PDF. Journal of KIISE, 46(7), 644-652.

Kim, Seon-Wu, Ji, Seon-Young, Seol, Jae-Wook, Jeong, Hee-Seok, & Choi, Sung-Pil (2018). Bidirectional GRU-GRU CRF based citation metadata recognition. In Annual Conference on Human and Language Technology, 30, 461-464.

Lim, Su-Hyun, Yoon, Te-Rin, Choi, Gyeong-Cheol, Cho, Won-Min, Heo, Jae-Jong, Han, Heyon-Woo, & Lee Kyung-Won (2019). A proposal for a bibliographic search interface using impact factor in the genealogy of academic literature. in Proceeding of HCI KOREA 2019, 526-529.