딥러닝 언어 모델을 이용한 연구보고서의 참고문헌 자동추출 연구*

Automatic Extraction of References for Research Reports using Deep Learning Language Model

한유경 (Yukyung Han)**
최원석 (Wonsuk Choi)***
이민철 (Minchul Lee)****

초 록

본 연구는 단행본, 학술지, 보고서 등 다양한 종류의 발간물로 구성된 연구보고서의 참고문헌 데이터베이스를 효율적으로 구축하기 위한 것으로 딥러닝 언어 모델을 이용하여 참고문헌의 자동추출 성능을 비교 분석하고자 한다. 연구보고서는 학술지와는 다르게 기관마다 양식이 상이하여 참고문헌 자동추출에 어려움이 있다. 본 연구에서는 참고문헌 자동추출에 널리 사용되는 연구인 메타데이터 추출과 더불어 참고문헌과 참고문헌이 아닌 문구가 섞여 있는 환경에서 참고문헌만을 분리해내는 원문 분리 연구를 통해 이 문제를 해결하였다. 자동 추출 모델을 구축하기 위해 특정 연구기관의 연구보고서 내 참고문헌셋, 학술지 유형의 참고문헌셋, 학술지 참고문헌과 비참고문헌 문구를 병합한 데이터셋을 구성했고, 딥러닝 언어 모델인 RoBERTa+CRF와 ChatGPT를 학습시켜 메타데이터 추출과 자료유형 구분 및 원문 분리 성능을 측정하였다. 그 결과 F1-score 기준 메타데이터 추출 최대 95.41%, 자료유형 구분 및 원문 분리 최대 98.91% 성능을 달성하는 등 유의미한 결과를 얻었다. 이를 통해 비참고문헌 문구가 포함된 연구보고서의 참고문헌 추출에 대한 딥러닝 언어 모델과 데이터셋 유형별 참고문헌 구축 방향을 제안하였다.

ABSTRACT

The purpose of this study is to assess the effectiveness of using deep learning language models to extract references automatically and create a reference database for research reports in an efficient manner. Unlike academic journals, research reports present difficulties in automatically extracting references due to variations in formatting across institutions. In this study, we addressed this issue by introducing the task of separating references from non-reference phrases, in addition to the commonly used metadata extraction task for reference extraction. The study employed datasets that included various types of references, such as those from research reports of a particular institution, academic journals, and a combination of academic journal references and non-reference texts. Two deep learning language models, namely RoBERTa+CRF and ChatGPT, were compared to evaluate their performance in automatic extraction. They were used to extract metadata, categorize data types, and separate original text. The research findings showed that the deep learning language models were highly effective, achieving maximum F1-scores of 95.41% for metadata extraction and 98.91% for categorization of data types and separation of the original text. These results provide valuable insights into the use of deep learning language models and different types of datasets for constructing reference databases for research reports including both reference and non-reference texts.

키워드: 연구보고서, 참고문헌 자동추출, 딥러닝 언어 모델, 자연어 처리, 개체명 인식
research reports, automatic reference extraction, deep learning language model, natural language processing, named entity recognition

* 본 연구는 정보통신정책연구원 2023년도 정보자료 운영사업의 지원을 받아 수행되었음.
** 정보통신정책연구원(yuyu@kisdi.re.kr) (제1저자)
*** 정보통신정책연구원(wschoi@kisdi.re.kr) (공동저자)
**** 카카오엔터프라이즈(bab2min@gmail.com) (교신저자)
■ 논문접수일자 : 2023년 5월 15일 ■ 최초심사일자 : 2023년 6월 3일 ■ 게재확정일자 : 2023년 6월 10일
■ 정보관리학회지, 40(2), 115-135, 2023. http://dx.doi.org/10.3743/KOSIM.2023.40.2.115

1. 서 론

문헌에서 인용되는 참고문헌은 서지 인용 연구의 주요 분석 요소로 활용되고 있다. 참고문헌을 구성하는 저자, 연도, 제목과 같은 메타데이터와 참고문헌 간의 선후 인용 관계를 통해 연구 주제 동향, 문헌 영향 지수, 연구자 네트워크 등을 파악할 수 있다. CiteSpace(Chen, 2006), VOSviewer(Van Eck & Waltman, 2010) 등의 서지 인용 분석 도구에서는 학술지 간 동시 인용되는 참고문헌 관련 연구 주제의 시계열 변화를 시각화하고 있으며, ScienceOn[1]의 논문 타임라인 서비스는 학술지의 참고문헌 인용・피인용 관계를 기반으로 주요 연구 키워드의 시계열 분석 기능을 제공하고 있다.

서지 인용 연구가 좋은 결과를 내기 위해서는 참고문헌 데이터베이스의 구축이 선행되어야 한다. 이강산다정, 이혜인, 현미환(2022)은 참고문헌을 수작업으로 분리하여 서지 정보를 구축하고 있는 국가연구개발보고서(이하 국가R&D보고서) 참고문헌 데이터베이스의 구축 품질 확보 및 참고문헌 활용 서비스 제공을 위해 참고문헌 기술규정 표준화를 연구하였다. 그러나 참고문헌을 일일이 식별하는 수작업은 현실적으로 많은 공수가 소요되며, 국가R&D보고서 데이터베이스에 미구축되는 연구보고서는 서지 정보 관리로부터 사각지대에 있을 수밖에 없다는 한계가 있기 때문에 연구보고서에 대한 보다 빠르고 효율적인 참고문헌 추출 연구가 필요하다.

최근 자연어처리 기술의 발달에 힘입어 참고문헌을 자동으로 추출하는 연구가 활발하게 진행되어 참고문헌 데이터베이스 구축에 크게 기여하고 있다. 다만 이런 참고문헌 자동 추출은 APA, Chicago, MLA 등 통일된 참고문헌 기술양식을 가지고 있는 학술지를 대상으로 주로 연구되었다. 반면 연구보고서는 학술지와 달리 연구기관마다 상이한 참고문헌 기술요소 및 형식을 사용하고 있어 참고문헌 추출이 어려운 상황이다. 기관에 따라 참고문헌을 게재국가(국내/해외)나 자료 유형(웹사이트/법령/데이터베이스 등)과 같은 범주를 자의적으로 분류하여 나열하거나, 부록이나 연구보고서 역대 발간 목록을 참고문헌 영역 뒤에 추가로 배치하기도 한다. 이런 특성 때문에 학술지와는 달리 연구보고서에서는 간단한 규칙을 통해 참고문헌 영역을 기계적으로 추출하는 것은 쉽지 않으며, 추출하여도 그 영역 내부에 참고문헌이 아닌 문구가 섞여 있는 경우가 다수 존재한다. 따라서 연구보고서에서 참고문헌 자동 추출을 수행하기 위해서는 단순히 참고문헌 문구 내에서 정보를 추출하는 연구뿐만 아니라 비참고문헌 유형의 문구들로부터 참고문헌 문구를 구분해내는 방법에 관한 연구가 필요하다.

참고문헌 자동추출 연구는 자연어 언어 처리(Natural Language Processing, 이하 NLP) 분야의 하위 과제(Task)인 개체명 인식(Named Entity Recognition, 이하 NER)과도 긴밀한 관계를 지니고 있다. NER이란 사람, 장소, 기관처럼 분석에 필요한 정보를 개체(Entity)로 구분지어 인식하는 연구방법으로써 이를 활용하여 참고문헌의 메타데이터나 원문 전체를 개체로

1) https://scienceon.kisti.re.kr/

인식하여 해당 개체를 추출할 수 있다. 최근에는 사전 학습 기반의 딥러닝 언어 모델과 접목하여 개체 추출의 정확도를 향상하는 방향으로 NER 연구가 진행되고 있다(Liu, Chen, & Xia, 2022). 또한, 사전 정보 없이 과제를 해결하는 Zero-shot Learning 방식이나 최소 하나 이상의 학습데이터를 소수 입력하는 One/Few-shot Learning 방식의 딥러닝 언어 모델 기반 NER 연구도 등장하고 있다(Lauscher et al., 2020; Wang et al., 2023).

본 연구의 목적은 NER에 활용되는 딥러닝 언어 모델들을 이용하여 연구보고서의 참고문헌 자동추출을 시도하는 것이다. 이를 위해 본 연구는 정부출연 연구기관인 정보통신정책연구원(이하 KISDI)[2]에서 발간된 전체공개 대상 연구보고서를 실험 모델로 하여 중점 분석한다. 딥러닝 언어 모델 중 최근 높은 Zero/One-shot Learning 성능으로 화제가 되고 있는 ChatGPT를 적용해보고, NER에서 높은 성능을 보이는 RoBERTa+CRF의 결과와 비교 분석한다. RoBERTa+CRF는 참고문헌 자동추출의 성능을 다방면으로 비교하기 위해 주요 연구 질문을 아래와 같이 설정한다.

[1] 연구 질문1: 딥러닝 언어 모델을 활용해 연구보고서의 참고문헌을 얼마나 정확히 추출할 수 있는가?
[2] 연구 질문2: 딥러닝 언어 모델에 따라 참고문헌 추출 성능에 차이가 존재하는가?
[3] 연구 질문3: 데이터 유형에 따라 참고문헌 추출 성능에 차이가 존재하는가?

딥러닝 언어 모델과 데이터 유형별 참고문헌 추출 성능은 메타데이터 추출, 자료유형 구분, 원문 분리 세 가지 항목으로 나누어 평가한다. 메타데이터 추출은 참고문헌을 구성하고 있는 요소(저자, 발행연도, 제목 등)들을 정확히 추출할 수 있는지 평가한다. 자료유형 구분은 참고문헌의 유형(학술지, 연구보고서, 뉴스 등) 분류 성능에 대해 평가한다. 마지막으로 원문 분리는 비참고문헌 문구와 참고문헌이 섞인 데이터에서 참고문헌 원문만 추출할 수 있는지 확인한다.

연구 구성은 다음과 같다. 제2장에서는 참고문헌 자동추출 및 딥러닝 모델과 NER 선행 연구들을 설명한다. 제3장은 연구보고서 참고문헌 자동추출의 실험 과정과 데이터셋, 실험 모델에 대한 전반적인 실험 방법을 제시한다. 제4장에서는 제시된 실험 방법을 통해 얻은 연구보고서 참고문헌의 자동추출, 딥러닝 언어 모델 및 데이터 유형별 성능 결과를 각각 비교 분석한다. 마지막으로 제5장에서는 맺음말 및 제언을 제시한다.

2. 참고문헌 추출 관련 연구

2.1 참고문헌 자동추출 연구

참고문헌 자동추출 연구는 학술 문헌에 수록

2) 국무총리 산하 경제·인문사회연구회 소속 연구원. 국내·외 정보화 및 정보통신·방송 분야의 정책·제도·산업 등에 관한 각종 정보를 수집·조사·연구함으로써, 지식정보사회 구현을 위한 국가의 정보통신정책 수립과 국민경제 발전에 공헌하기 위해 설립되었다(정보통신정책연구원 정관 제2조).

된 참고문헌들을 각각 분리하고, 참고문헌별 메타데이터를 추출하는 방식으로 진행되었다. 초기에는 특정한 템플릿을 정하여 해당 템플릿에 일치하는 정보만 추출하는 소위 “템플릿 마이닝”이라는 규칙 기반 연구(Besagni, Belaïd, & Benet, 2003; Hollingsworth, Lewin, & Tidhar, 2005; Huang et al., 2004) 위주로 이루어졌으나, 데이터의 적용 규모가 점차 커지면서 규칙이 모든 데이터를 전부 아우를 수 없는 한계가 나타났다. 이러한 한계를 극복하기 위해 통계 기반 기계학습 모델을 활용한 참고문헌 추출 연구가 등장하기 시작하였다. 이후 Support Vector Machine(SVM)(Zhang et al., 2011), Hidden Markov Model(HMM)(Hetzner, 2008) 등을 활용한 연구가 수행되었으며, 특히 Conditional Random Field(CRF)는 ParsCit(Councill, Giles, & Kan, 2008), GROBID(Lopez, 2009), CERMINE(Tkaczyk et al., 2015)처럼 PDF 형식인 학술지의 참고문헌 추출 도구에 주로 적용되었다. 최근 수행되고 있는 연구들에서는 문맥에 따른 의미 차이를 반영할 수 있는 딥러닝 모델을 기반으로 참고문헌에서 메타데이터를 추출하고 있다. Rodrigues, Colavizza, Kaplan(2018)은 예술과 인류학 분야 학술지의 참고문헌 메타데이터 추출, 원문 분리, 자료 유형 구분을 위해 단어 임베딩 방식, 딥러닝 모델(BiLSTM, CNN), 모델 예측 방식을 번갈아 조합하며 조합별 성능을 비교하였다. 지선영, 최성필(2021)은 PDF 형식인 학술지 내 참고문헌의 메타데이터를 추출하고자 Bi-GRU-CRF와 다국어 BERT 모델들을 병합한 후 비교 및 분석하였다. Voskuil와 Verberne(2021)는 특허의 참고문헌 원문을 분리하기 위해 BERT 기반 모델(BERT, SciBERT, BioBERT)을 활용하였다. Choi et al.(2023)은 다국어로 구성된 학술지의 참고문헌 메타데이터 학습데이터셋을 구축하고, BERT 모델에서 높은 메타데이터 추출 성능을 보임을 제시하였다. 학술지 및 특허와 달리 연구보고서에 관한 참고문헌 자동추출 연구는 아직 미비한 상황이기에 참고문헌의 메타데이터 추출만 아니라 도서, 학술지, 뉴스기사와 같은 자료유형 구분, 문헌 본문에서의 참고문헌 원문 분리까지 다양한 접근으로 살펴볼 필요가 있다.

2.2 딥러닝 언어 모델

딥러닝 언어 모델이란 인간의 신경망처럼 여러 층(Layer)으로 깊이 있게 구성하여 인간의 언어를 이해하고 예측할 수 있는 신경망 모델을 의미하며, 대량의 데이터를 사전에 학습하여 여러 분야의 과제에 적용될 수 있는 사전 학습 기반 딥러닝 언어 모델 연구가 활발히 진행중이다. BERT(Devlin et al., 2018)는 3개 임베딩(Token, Position, Segment)을 이용하여 입력 데이터의 토큰들을 임베딩 벡터로 구성한 후, 임의로 마스킹한 단어를 예측하는 MLM(Masked Language Model)과 문장 연결 여부를 맞히는 NSP(Next Sentence Prediction) 방식으로 학습하는 딥러닝 언어 모델이다. RoBERTa(Liu et al., 2019)는 BERT를 개량한 모델로, 2개의 임베딩(Token, Position)과 MLM 모델만 적용함에도 불구하고 여러 NLP 과제에서 BERT보다 더 높은 성능을 보이고 있다. BERT의 MLM 학습 방식은 마스킹한 토큰의 전후 문맥을 모두 고려하여 토큰을 예측한다면, GPT(Radford et al., 2018)라는 모

델은 마스킹한 토큰의 이전 문맥만을 이용하여 토큰를 예측하는 학습 방식을 택하였다. GPT의 후속 연구 중 하나인 ChatGPT(OpenAI, 2022)는 GPT-3.5-Turbo 버전으로 사람이 직접 학습 데이터별 예측 결과에 상응하는 보상(reward)을 계산하여 학습에 반영했으며, 현재는 OpenAI에서 이를 이용한 대화형 서비스를 제공하고 있다.

2.3 NER

NER의 대표 모델 중 하나인 CRF는 확률적 그래프 모델로, 여러 단어로 구성된 문장에서 각각의 단어에 태그(라벨)를 부착하는 과제를 효과적으로 수행한다. CRF는 이 특징 덕분에 NER에서 널리 사용되어 왔으며 특히 최근에는 딥러닝 언어 모델과 함께 쓰여 NER의 성능을 더욱 향상시키고 있다(Dai et al., 2019; Souza, Noguiera, & Lotufo, 2019; Wu et al., 2021). 딥러닝 언어 모델 및 CRF 모델의 경우 높은 성능을 달성하기 위해 충분한 수의 학습 데이터가 필요하다. 특히 각 개체 종류별로 일정량 이상의 데이터를 확보하는 것이 필수적인데, 이는 비용이 많이 드는 작업이다. 따라서 학습 데이터가 전혀 없는 상황과 1건만 있는 상황에서도 높은 성능을 달성할 수 있는 Zero-shot 및 One-shot learning 연구가 최근 각광받고 있다(Fritzler, Logacheva, & Kretov, 2019; Kim et al., 2021; Yang & Katiyar, 2020). Zero-shot learning은 학습 데이터가 전혀 없는 상황에서도 과제를 해결하는 방식이고, Few-shot learning은 학습 데이터가 소수인 상황에서 과제를 해결하는 방식을 의미한다. 최근 생성형 언어 모델인 ChatGPT의 등장과 함께 이에 대한 연구가 진행되고 있다(González-Gallardo et al., 2023; Hu et al., 2023). ChatGPT는 프롬프트의 설계 방식에 따라 답변의 품질이 크게 달라질 수 있으므로, 과제에 맞는 프롬프트 설계 방식 연구도 함께 등장하고 있다(Wei et al., 2023; White et al., 2023).

3. 참고문헌 추출 실험 방법

3.1 실험 과정

참고문헌 자동추출 분석을 위해 2개의 과제를 설정하였다. 첫 번째 과제는 참고문헌의 메타데이터 추출이다. 딥러닝 언어 모델의 학습을 위해 참고문헌 메타데이터를 개체로 인식할 수 있도록 〈표 1〉처럼 BIO 태깅을 수행하였다. BIO 태깅이란 각 개체에 대해 B/I/O(Begin/Inside/Outside) 태그를 분류하여 개체를 인식하는 작업을 의미한다. 예를 들어, 참고문헌 내 “2015”라는 토큰이 발행연도를 의미하는 경우 2015에 B-year이라는 태그를 붙였다. 두 번째 과제는 원문 분리와 자료유형 구분이다. 참고문헌의 원문과 자료유형(도서, 학술지, 뉴스기사 등)을 같이 구분할 수 있도록 원문에 BIO 태깅을 모두 수행하였다. 즉, A이라는 참고문헌의 자료유형이 학술지인 경우 참고문헌 원문의 시작 토큰에는 B-Journal, 나머지는 I-Journal로 두고, 그 외 해당되지 않는 토큰은 O로 처리하였다.

두 번째 과제는 3개의 데이터셋(연구보고서, 학술지, 학술지+비참고문헌 문구)을 대상으로 2개의

〈표 1〉 BIO 태깅 방식

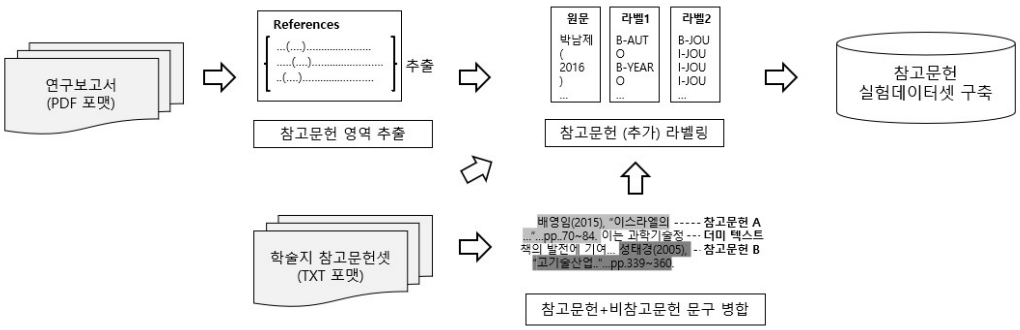
토큰	박남제	(	2016	)	...
메타데이터 유형	B-author	O	B-year	O	...
자료유형	B-journal	I-journal	I-journal	I-journal	...

딥러닝 언어 모델(RoBERTa+CRF, ChatGPT)별 학습과 추론이다. ChatGPT는 연구보고서와 학술지 데이터셋의 경우 참고문헌이 한 문장씩 구성되어 있기에 모두 Zero-shot learning 기반으로 학습하였으며, 학술지+비참고문헌 문구셋에 대해서는 최소 두 문장이 하나의 데이터로 구성되어 있으므로 원문 분리 실험을 감안하여 Zero/One-shot learning 두 가지 방식으로 학습을 진행하였다. 딥러닝 언어 모델 중 RoBERTa+CRF는 KLUE(Park et al., 2021)의 RoBERTa-base를 활용하였으며, ChatGPT는 OpenAI의 ChatGPT API를 이용하였다.

3.2 데이터셋

실험데이터셋은 데이터 유형별 참고문헌 자동추출 성능을 분석하기 위해 연구보고서뿐만 아니라 학술지 및 학술지와 일반 문구가 섞인 데이터셋으로 구성하였다. 전반적인 데이터셋의 구축 과정은 〈그림 1〉과 같다.

〈그림 1〉에서 연구보고서는 KISDI의 전체 공개 범위에 해당하는 발간물 202건으로 삼았다. 실험 대상 발간물 종류는 〈표 2〉와 같으며, 이중 정보통신방송정책, AJIC(Asian Journal of Information and Communications), 우정정보, KISDI Premium Report, ICT Industry Outlook of Korea는 연구원상에서 정기간행물로 분류하고 있으나 국내외 공인학술지에 해당하지 않는 점을 고려하여 모두 연구보고서로 분류하였다. 연구보고서는 모두 PDF 형식 파일이었기에 참고문헌만 별도로 추출해야 하는 작업이 필요하였다. 보고서 원문의 참고문헌 영역의 시작은 “Reference”, “참고문헌” 등 참고문헌 관련 단어를 기준으로 잡고, “부록”이나 “저자 소개”와 같이 참고문헌과 관련이 없는 부분 전까지를 영역의 끝으로 한정한 다음 해당 영역



〈그림 1〉 데이터셋 구축 과정

〈표 2〉 실험 대상 KISDI 발간물 종류

발간물 종류	건수
기본연구	24
정책연구	43
정책자료	16
현안연구	5
정보통신방송정책	51
Asian Journal of Information and Communications	24
우정정보	20
KISDI Primium Report	18
ICT Industry Outlook of Korea	1
합계	202

을 정규표현식으로 추출하였다. 해당 영역 안의 참고문헌들을 수작업으로 분리하는 작업을 통해 총 참고문헌 8,429건을 구축하였다.

〈그림 1〉에서 학술지는 Choi et al.(2023)에서 구축한 DeepData-REFMETA 데이터셋(Korea Institute of Science and Technology Information [KISTI], 2022)을 이용하였다. 수작업과 규칙 기반 자동 라벨링 방식이 포함된 데이터셋이며, 총 3,815,987건 중 수작업으로 라벨링한 평가셋(71,489건)을 학술지 데이터셋으로 삼았다.

학술지+비참고문헌 문구는 학술지 데이터셋과 참고문헌 형식이 아닌 일반 문구로 구성된 텍스트를 병합한 데이터셋이다. 원문 파일에서 참고문헌을 추출하는 경우 참고문헌이 아닌 문구들이 같이 추출될 수 있기에 텍스트에서 참고문헌만을 잘 분리할 수 있는지 알아볼 필요가 있기에 본 데이터셋을 구성하였다. 학술지+비참고문헌 문구셋은 연구보고서셋 및 학술지셋과는 다르게 데이터 1건에 여러 개의 참고문헌이 들어가도록 구성했다. 구체적으로 500자 내로 구성된 문장에 2~9개의 학술지 참고문헌들을 순서와 상관없이 임의로 병합하며 데이터를 증강하였다. 구체적으로 참고문헌 문구들의 집합을 A, 비참고문헌 문구들의 집합을 B라고 할 때 "$b\,a_1\,a_2 \ldots a_n$"($a_i \in A$, $b \in B$, $2 \leq n \leq 9$)의 형태로 텍스트를 결합하여 데이터를 구성하였다. 이러한 작업을 통해 총 26,572건의 데이터를 구축하였다.

데이터셋별 최종 학습/평가데이터셋 건수는 〈표 3〉에서 확인할 수 있다. 라벨링 오류, 문자 인코딩 실패 등으로 인하여 학습과 평가가 불가한 데이터는 제외하였다. 연구보고서는 KISDI 자료 8,429건, 학술지는 KISTI 자료 71,375건, 학술지+비참고문헌 문구는 26,572건이 본 실험에 활용되었다. 전체 데이터셋을 9:1로 나누어 95,737건은 학습데이터로, 10,639건은 평가데이터로 사용했다. RoBERTa+CRF는 모든 학습데이터를 기반으로 학습하되, 평가데이터를 이용하여 데이터 유형별로 평가를 진행한다. ChatGPT는 별도의 학습 없이 평가데이터에 대해서만 평가하였다.

참고문헌의 메타데이터 유형과 건수는 〈표 4〉와 같다. Choi et al.(2023)이 기존에 설정하였던 13개의 유형을 연구보고서에 동일하게 적용

〈표 3〉 학습/평가데이터셋 건수

구분	Train	Test	합계
연구보고서	7,585	844	8,429
학술지	64,237	7,138	71,375
학술지+비참고문헌 문구	23,915	2,657	26,572
합계	95,737	10,639	106,376

〈표 4〉 참고문헌 메타데이터 유형 구성

메타데이터 유형	연구보고서			학술지			학술지+비참고문헌 문구		
	Train	Test	합계	Train	Test	합계	Train	Test	합계
저자	7,280	818	8,098	64,237	7,138	71,375	79341	8,868	88,209
DOI	35	3	38	3,069	332	3,401	3505	429	3,934
호	1,585	182	1,767	37,121	4,170	41,291	43702	4,825	48,527
ISSN	0	0	0	9	3	12	12	1	13
저널	3,163	362	3,525	64,159	7,129	71,288	76092	8,482	84,574
페이지	1,212	140	1,352	63,033	7,011	70,044	73705	8,187	81,892
발행처	316	37	353	44	3	47	46	8	54
발행기관	1,424	168	1,592	444	46	490	543	56	599
발행지	260	39	299	26	0	26	32	2	34
제목	6,627	749	7,376	64,235	7,138	71,373	78348	8,743	87,091
URL	2,003	199	2,202	12,994	1,428	14,422	14960	1,669	16,629
권	1,626	189	1,815	63,230	7,041	70,271	74239	8,261	82,500
발행연도	7,724	843	8,567	64,161	7,128	71,289	76857	8,553	85,410
총계	33,255	3,729	36,984	436,762	48,567	485,329	521382	58,084	579,466

하였으며, 기존에 미포함된 ISBN과 같은 메타데이터 유형은 연구보고서셋에도 역시 존재하지 않아 추가하지 않았다. 다만, 발행기관, 발행지, 발행연도 같은 유형은 포괄적인 의미의 기관이나 장소, 연도로도 적용될 수 있도록 라벨링하였다. 학술지+비참고문헌 문구 중 참고문헌을 제외한 나머지 일반 문구에는 메타데이터 유형을 부여하지 않았다. 연구보고서셋에서는 36,984건, 학술지셋에서는 485,329건, 학술지+비참고문헌 문구셋에서는 579,466건의 메타데이터가 나왔다. 세 데이터 유형 모두 저자, 제목, 발행연도 메타데이터를 대다수 포함하고 있으며, 이에 비해 발행처, 발행지, ISSN은 상대적으로 적게 라벨링되었다.

실험 대상 자료유형은 〈표 5〉와 같이 총 11개로 구성하였다. 학술지셋과 학술지+비참고문헌 문구셋에 등장하는 학술지 참고문헌은 모두 학술지 유형으로 구분하였다. 연구보고서의 자료유형으로는 KISDI의 원고 작성 지침('19.12 개정)에서 제시된 자료유형 9개- 단행본, 학술지, 학위논문, 법률, 뉴스, 프로시딩, 보고서, 웹사이트, 인터뷰-를 기준으로 우선 분류하였다. 다음으로, 미분류된 참고문헌들은 이강산다정, 이혜인, 현미환(2022)의 참고문헌 자료유형를 참고

〈표 5〉 참고문헌 자료유형 구성

자료유형	연구보고서			학술지			학술지+비참고문헌 문구		
	Train	Test	합계	Train	Test	합계	Train	Test	합계
단행본	382	42	424	0	0	0	0	0	0
학술지	1,342	145	1,487	64,237	7,138	71,375	79,341	9,023	88,364
데이터셋	70	10	80	0	0	0	0	0	0
학위논문	33	4	37	0	0	0	0	0	0
가이드라인	52	6	58	0	0	0	0	0	0
법률	118	10	128	0	0	0	0	0	0
뉴스	1,368	144	1,512	0	0	0	0	0	0
프로시딩	154	24	178	0	0	0	0	0	0
보고서	2,903	336	3,239	0	0	0	0	0	0
웹사이트	1,006	105	1,111	0	0	0	0	0	0
인터뷰	4	1	5	0	0	0	0	0	0
총계	7,432	827	8,259	64,237	7,138	71,375	79,341	9,023	88,364

하여 데이터셋, 가이드라인으로 추가 분류하였다. 연구보고서셋의 자료유형 중 보고서, 뉴스, 학술지, 웹사이트는 1,000건 이상 포함되어 있으나 데이터셋, 학위논문, 가이드라인, 인터뷰는 100건 미만으로 구성되어 있었다.

3.3 실험 모델 설계

RoBERTa+CRF의 학습을 위해 Huggingface에 게시된 KLUE(Park et al., 2021)의 RoBERTa-base[3] 토크나이저와 모델을 활용하였다. 참고문헌이 모델에 입력될 수 있도록 토크나이저를 이용해 토큰 단위로 나누었다. 연구보고서셋과 학술지셋의 평균 토큰 수는 각각 42.74개, 46.11개였고, 최대 토큰 수는 249개, 319개였다. 따라서 각 참고문헌 토큰화 결과의 최대 길이는 384로 정하였다. 메타데이터 추출과 원문 분리·자료유형 구분이라는 두 과제를 수행하기 위해 RoBERTa 모델에 각 과제에 맞는 독립적인 CRF 2개를 추가하였다. 모델의 Epoch은 3, Batch Size는 16, Gradient accumulation steps는 2, dropout은 0.1, 학습 알고리즘은 Adam Optimizer, Learning rate는 5e-5로 두었다.

ChatGPT는 Zero-shot learning과 One-shot learning에 맞게 프롬프트를 각각 설계하였다. Zero-shot의 경우 〈표 6〉처럼 System 역할에는 과제에 대한 사전 설명을 입력하고, User 역할에는 참고문헌 문장을 넣어 각 User 질의에 대한 출력을 유도하였다. 이때, ChatGPT 자체에서 원문 변형이 일어나지 않도록 Temperature를 0으로 설정하였다. 프롬프트 질의 설계 작업에서는 1) 대화형으로 질의하되, 2) 원문을 번역하지 않은 채 그대로 출력해야 하고, 3) 메타데이터와 자료유형, 원문을 한꺼번에 추출하도록 하였다. 자료유형의 경우 기존에 설정한 범주 외의 유형까지 나올 가능성이 있기에 자료

3) https://huggingface.co/klue/roberta-base

〈표 6〉 ChatGPT 프롬프트 및 답변 예시

프롬프트		출력 결과
역할	내용	
System	I have a reference set and please extract metadata and data type(journal, report, news, dataset, website, book, guideline, legislation, proceeding, interview, degree paper) for each reference. Don't translate. Your output format should be like this [raw text: author: DOI: issue: ISSN: journal: page: publisher: publishing organization: publishing place: title: url: volume: year: data type:]	[raw text: Mirzaei, M., Eghtesad, M., Alishahi, M.M., 2016. Planing force identification in highvelocity underwater vehicles. J. Vib. Contr. 22, 4176-4191. author: Mirzaei, M., Eghtesad, M., Alishahi, M.M DOI: issue: ISSN: journal: J. Vib. Contr page: 4176-4191 publisher: publishing organization: publishing place: title: Planing force identification in highvelocity underwater vehicles url: volume: 22 year: 2016 data type: journal]
User	Mirzaei, M., Eghtesad, M., Alishahi, M. M., 2016. Planing force identification in highvelocity underwater vehicles. J. Vib. Contr. 22, 4176-4191.	

유형 목록을 질의 안에 같이 제시한 후 유형을 선택할 수 있도록 하였다. 여기서 One-shot의 경우 Zero-shot 프롬프트의 기본 형식은 유지하되 System 역할에 두 참고문헌이 한 문장으로 엮여있는 경우를 input, 각 참고문헌의 추출 결과를 output으로 명시한 예시를 추가 입력하였다.

3.4 평가 방법

RoBERTa+CRF의 추론 결과는 토크나이즈한 개체와 BIO 태그로 구성되어 있으나 ChatGPT는 일반 문장으로 출력하기에 정답 데이터셋과 비교하기 위해서는 출력 결과 형식이 같아야 한다. 이에 따라 RoBERTa+CRF와 정답 데이터셋은 개체의 B를 개체의 시작, 가장 마지막으로 등장하는 I를 개체의 마지막으로 보고 해당 개체에 맞는 부분만 원문에서 가져오고, ChatGPT는 기존 출력 결과를 그대로 가져와 서로 비교하였다. 이와 더불어, ChatGPT에서 개체 추출 시 "N/A"나 "Not available"로 나오는 경우는 개체를 추출하지 않은 것으로 보았다. 모델

성능 평가는 SemEval 2013(Segura-Bedmar, Martínez Fernández, & Herrero-Zazo, 2013)의 평가 방법을 사용하였다. 이는 개체의 일치 여부를 Type, Partial, Strict, Exact으로 구분하고, (Micro) Precision, Recall, F1-score를 각각 계산하는 방식이다. Type은 개체 원문과 상관없이 개체 유형이 같은지, Partial은 개체 유형과 상관없이 개체 원문이 부분일치하는지를 계산한다. Strict은 개체 유형과 상관없이 개체 원문이 완전일치하는지, 마지막으로 Exact은 개체 유형과 개체 원문 모두 완전일치하는지를 계산한다. 이때, Type과 Partial에서는 부분일치한 개체는 0.5, 완전일치한 개체는 1로 산정하여 Precision과 Recall을 구한다. 평가 코드는 davidsbatista의 NER-Evaluation 깃헙 코드[4]를 활용하였다.

4. 실험 결과

참고문헌 자동추출 성능을 평가하기 위해 메타데이터 추출 항목과 자료유형 구분 및 원문 분리 항목으로 나누어 실험 결과를 정리하였다. 이때, 각 항목별로 데이터셋(연구보고서, 학술지, 학술지+비참고문헌 문구)과 딥러닝 언어 모델(RoBERTa+CRF, ChatGPT)들을 종합적으로 비교한다.

참고문헌 메타데이터 추출 결과는 〈표 7〉과 같다. 모든 데이터셋에서 특정 과제 위주로 학습된 모델인 RoBERTa+CRF가 ChatGPT보다 높은 F1-score를 보였다. 또한, 단지 학술지만으로 구성된 학술지셋이 단행본, 학술지, 보고서 등 다양한 종류의 발간물로 구성된 연구보고서셋보다 F1-score가 높았다. 즉, 유사한 학술지 기술형식으로 인해 학술지의 추출 성능이 더 우수함을 볼 수 있다. 학술지+비참고문헌 문구셋의 경우 Zero-shot ChatGPT보다 One-shot ChatGPT의 F1-score가 전반적으로 높았다. 이는 프롬프트에 예시를 추가함으로써 ChatGPT의 메타데이터 추출 성능이 향상함을 알 수 있다.

〈표 8〉 및 〈표 9〉는 메타데이터의 상세 유형별로 메타데이터 원문의 완전일치(Strict)와 부분일치(Type) 결과를 보여준다. 완전일치와 부분일치 성능 결과를 종합적으로 살펴보면, RoBERTa+CRF는 대부분의 메타데이터 유형과 데이터셋에서 ChatGPT보다 성능이 높았다. 하지만, 연구보고서셋의 발행지, 학술지셋의 ISSN/발행처, 학술지+비참고문헌 문구셋의 ISSN에서는 오히려 ChatGPT가 더 좋은 결과를 냈다. 이는 RoBERTa+CRF는 학습을 위해 상대적으로 더 많은 데이터를 필요로하는 반면, ChatGPT는 데이터가 부족한 상황에서도 잘 작동하기 때문이다. 따라서 학습 데이터가 적은 메타데이터 추출에 대해서는 일반적으로 ChatGPT가 더 우수한 성능을 보인다. RoBERTa+CRF의 완전일치 기준에서는 연구보고서셋와 학술지셋 간 DOI의 F1-score가 92.31%p 차이가 나는데, 연구보고서셋은 38건, 학술지셋은 3,401건의 DOI를 가지고 있는 것을 감안할 때 이 격차는 데이터셋 규모 차이에서 기인한 것으로 보인다. Zero-shot ChatGPT

4) https://github.com/davidsbatista/NER-Evaluation

〈표 7〉 메타데이터 추출 성능 비교

데이터셋	모델	평가 척도	Precision	Recall	F1-score
연구보고서	RoBERTa+CRF	Type	**95.01**	**92.85**	**93.92**
		Partial	**96.52**	**94.33**	**95.41**
		Strict	**92.36**	**90.26**	**91.30**
		Exact	**94.52**	**92.37**	**93.43**
	ChatGPT (Zero-shot)	Type	81.41	75.26	78.21
		Partial	88.41	81.73	84.94
		Strict	72.27	66.82	69.44
		Exact	80.41	74.34	77.26
학술지	RoBERTa+CRF	Type	**99.77**	**99.57**	**99.67**
		Partial	**99.74**	**99.54**	**99.64**
		Strict	**99.45**	**99.25**	**99.35**
		Exact	**99.49**	**99.29**	**99.39**
	ChatGPT (Zero-shot)	Type	89.62	86.05	87.80
		Partial	88.50	84.98	86.70
		Strict	75.49	72.49	73.96
		Exact	77.53	74.44	75.95
학술지+ 비참고문헌 문구	RoBERTa+CRF	Type	**97.83**	**95.52**	**96.66**
		Partial	**98.64**	**96.31**	**97.46**
		Strict	**97.28**	**94.98**	**96.12**
		Exact	**97.28**	**94.99**	**96.12**
	ChatGPT (Zero-shot)	Type	78.68	68.52	73.25
		Partial	85.18	74.18	79.30
		Strict	68.89	60.00	64.14
		Exact	70.52	61.41	65.65
	ChatGPT (One-shot)	Type	96.96	95.33	96.14
		Partial	94.53	92.95	93.73
		Strict	92.11	90.57	91.33
		Exact	92.11	90.57	91.33

〈표 8〉 메타데이터 유형별 데이터 원문 완전일치 F1-score 비교

메타데이터 유형	연구보고서		학술지		학술지+비참고문헌 문구		
	RoBERTa+ CRF	ChatGPT (Zero-shot)	RoBERTa+ CRF	ChatGPT (Zero-shot)	RoBERTa+ CRF	ChatGPT (zero-shot)	ChatGPT (one-shot)
저자	**93.87**	45.05	**99.70**	43.67	**98.04**	49.47	56.38
DOI	**6.35**	2.90	**98.66**	62.02	**97.04**	77.58	84.45
호	**77.43**	34.90	**98.66**	63.22	**84.93**	28.95	37.56
ISSN	N/A	N/A	0.00	**1.51**	0.00	2.02	2.47
저널	**77.93**	59.85	**99.30**	73.46	**98.44**	76.38	89.35
페이지	**73.82**	48.15	**99.61**	89.51	**98.92**	87.34	94.22
발행처	**26.67**	23.59	**26.67**	2.26	**57.14**	12.39	16.84
발행기관	**72.19**	36.44	**82.00**	13.45	**92.17**	30.54	26.59
발행지	1.71	**33.01**	N/A	N/A	**22.22**	3.96	4.82
제목	**91.37**	79.93	**99.45**	92.45	**97.48**	90.57	94.35
URL	**84.96**	69.06	**99.37**	64.13	**96.93**	74.61	65.84
권	**78.42**	32.95	**98.67**	60.41	**94.87**	45.66	65.95
발행연도	**91.79**	82.95	**99.42**	84.20	**95.83**	79.81	83.22

〈표 9〉 메타데이터 유형별 데이터 원문 부분일치 F1-score 비교

메타데이터 유형	연구보고서		학술지		학술지+비참고문헌 문구		
	RoBERTa+ CRF	ChatGPT (Zero-shot)	RoBERTa+ CRF	ChatGPT (Zero-shot)	RoBERTa+ CRF	ChatGPT (zero-shot)	ChatGPT (one-shot)
저자	**95.75**	62.09	**98.34**	60.06	**98.73**	81.18	82.88
DOI	**6.35**	2.90	**99.25**	88.26	**97.04**	77.82	84.68
호	**78.86**	42.35	**99.36**	83.85	**86.23**	38.84	42.78
ISSN	N/A	N/A	0.00	**36.36**	0.00	2.02	**2.47**
저널	**82.71**	62.59	**99.52**	85.99	**98.72**	87.89	93.98
페이지	**75.71**	55.03	**99.71**	93.42	**98.95**	89.19	94.82
발행처	**30.00**	26.67	36.36	**50.00**	**57.14**	12.39	16.84
발행기관	**74.87**	36.86	**89.13**	61.54	**92.17**	30.54	26.59
발행지	1.71	**34.95**	N/A	N/A	**22.22**	3.96	4.82
제목	**94.74**	89.03	**99.49**	90.38	**98.00**	93.94	95.40
URL	**87.61**	73.21	**99.65**	75.22	**97.18**	75.56	66.25
권	**80.74**	42.25	**99.46**	77.54	**95.50**	51.44	71.61
발행연도	**93.33**	84.45	**99.66**	83.67	**96.51**	84.41	86.84

의 저자 유형 F1-score는 완전일치에서는 데이터셋별로 45.05%, 43.67%, 49.47%였으나 부분일치에서는 62.09%, 60.06%, 81.18%로 최소 16.39%p, 최대 31.71%p의 차이를 보였다. 이는 ChatGPT가 저자 유형의 결과를 출력할 시 원문 그대로 출력하지 않고 저자 개체들을 원문과 다른 특수기호로 연결하여 부분일치에서는 정답으로 처리되지만, 완전일치에서는 오답으로 처리되었기 때문이다.

자료유형 구분 및 문장 분리 결과는 〈표 10〉과 같다. 여기서 학술지셋과 학술지+비참고문헌 문구셋은 모두 학술지 유형의 참고문헌을 대상으로 구축하였기에 나머지 자료유형은 고려하지 않았다. 전반적으로 RoBERTa+CRF가 ChatGPT에 비해 대부분 성능이 높았다. 학술지셋의 RoBERTa+CRF 결과는 모두 99% 대의 성능이나, 연구보고서셋은 Type 79.98%, Strict 79.49%를 보이며 문장 분리에서 우수하지만 자료유형 구분에서는 다소 낮은 성능을 보이고 있다. 단, 연구보고서셋은 단일 유형인 학술지셋과 다르게 여러 자료유형으로 묶여있는 걸 감안하여 성능을 볼 필요가 있다. 학술지+비참고문헌 문구셋에서는 RoBERTa+CRF는 Type, Zero-shot ChatGPT는 Strict, 그리고 One-shot ChatGPT는 Parital과 Exact에서 상대적으로 높은 성능을 달성했다. 이를 통해 RoBERTa+CRF는 자료 유형 분류에서, ChatGPT는 문장 분리에서 강점을 드러내고 있음을 알 수 있다. ChatGPT의 유형 분류 성능이 낮은 이유는 review나 case reports 등 사전에 제시하지 않은 자료유형을 자체적으로 생성하는 경우가 잦았는데, 이런 경우가 전부 오답으로 처리됐기 때문이다.

〈표 11〉과 〈표 12〉는 자료유형 및 문장 분리별 데이터 원문의 완전일치와 부분일치 성능 결과를 보여준다. 연구보고서셋의 RoBERTa+CRF는 학술지, 뉴스, 보고서, 웹사이트 유형 구분 성능이 ChatGPT보다 높았고, ChatGPT는

〈표 10〉 자료유형 구분 및 문장 분리 성능 비교

데이터셋	모델	평가 척도	Precision	Recall	F1-score
연구보고서	RoBERTa+CRF	Type	**79.78**	**80.17**	**79.98**
		Partial	**98.68**	**99.15**	**98.91**
		Strict	**79.30**	**79.69**	**79.49**
		Exact	**98.44**	**98.91**	**98.67**
	ChatGPT (Zero-shot)	Type	69.48	69.65	69.57
		Partial	98.49	98.73	98.61
		Strict	69.12	69.29	69.20
		Exact	98.19	98.43	98.31
학술지	RoBERTa+CRF	Type	**99.97**	**99.97**	**99.97**
		Partial	**99.95**	**99.95**	**99.95**
		Strict	**99.87**	**99.87**	**99.87**
		Exact	**99.90**	**99.90**	**99.90**
	ChatGPT (Zero-shot)	Type	96.37	96.29	96.33
		Partial	99.89	99.80	99.85
		Strict	96.14	96.06	96.10
		Exact	99.78	99.69	99.73
학술지+ 비참고문헌 문구	RoBERTa+CRF	Type	**96.96**	**95.33**	**96.14**
		Partial	94.53	92.95	93.73
		Strict	92.11	90.57	91.33
		Exact	92.11	90.57	91.33
	ChatGPT (Zero-shot)	Type	94.14	93.24	93.69
		Partial	96.61	95.69	96.15
		Strict	**92.46**	**91.58**	**92.02**
		Exact	95.17	94.26	94.71
	ChatGPT (One-shot)	Type	87.05	88.10	87.57
		Partial	**97.67**	**98.85**	**98.26**
		Strict	86.26	87.30	86.78
		Exact	**97.20**	**98.37**	**97.78**

〈표 11〉 자료유형 구분 및 문장 분리별 데이터 원문 완전일치 F1-score 비교

자료유형	연구보고서		학술지		학술지+비참고문헌 문구		
	RoBERTa+ CRF	ChatGPT (Zero-shot)	RoBERTa+ CRF	ChatGPT (Zero-shot)	RoBERTa+ CRF	ChatGPT (zero-shot)	ChatGPT (one-shot)
단행본	66.67	**71.11**	N/A	N/A	N/A	N/A	N/A
학술지	**86.96**	81.33	**99.87**	96.10	91.33	**92.02**	86.78
데이터셋	0.00	**46.67**	N/A	N/A	N/A	N/A	N/A
학위논문	0.00	**44.44**	N/A	N/A	N/A	N/A	N/A
가이드라인	0.00	**27.27**	N/A	N/A	N/A	N/A	N/A
법률	0.00	**46.67**	N/A	N/A	N/A	N/A	N/A
뉴스	**81.48**	74.50	N/A	N/A	N/A	N/A	N/A
프로시딩	49.12	**51.72**	N/A	N/A	N/A	N/A	N/A
보고서	**84.29**	60.29	N/A	N/A	N/A	N/A	N/A
웹사이트	**70.32**	60.55	N/A	N/A	N/A	N/A	N/A
인터뷰	0.00	**16.67**	N/A	N/A	N/A	N/A	N/A

〈표 12〉 자료유형 구분 및 문장 분리별 데이터 원문 부분일치 F1-score 비교

자료유형	연구보고서		학술지		학술지+비참고문헌 문구		
	RoBERTa+ CRF	ChatGPT (Zero-shot)	RoBERTa+ CRF	ChatGPT (Zero-shot)	RoBERTa+ CRF	ChatGPT (zero-shot)	ChatGPT (one-shot)
단행본	68.89	**71.11**	N/A	N/A	N/A	N/A	N/A
학술지	**86.96**	81.33	**99.97**	96.33	**96.14**	93.69	87.57
데이터셋	0.00	**46.67**	N/A	N/A	N/A	N/A	N/A
학위논문	0.00	**44.44**	N/A	N/A	N/A	N/A	N/A
가이드라인	0.00	**27.27**	N/A	N/A	N/A	N/A	N/A
법률	0.00	**46.67**	N/A	N/A	N/A	N/A	N/A
뉴스	**81.48**	75.17	N/A	N/A	N/A	N/A	N/A
프로시딩	49.12	**51.72**	N/A	N/A	N/A	N/A	N/A
보고서	**85.17**	60.88	N/A	N/A	N/A	N/A	N/A
웹사이트	**70.32**	60.55	N/A	N/A	N/A	N/A	N/A
인터뷰	0.00	**16.67**	N/A	N/A	N/A	N/A	N/A

데이터셋, 학위논문, 가이드라인, 법률, 프로시딩, 인터뷰 유형 구분 성능에서 더 높았다. 이때, 데이터셋, 학위논문, 가이드라인, 법률, 인터뷰 유형의 RoBERTa+CRF 추론 결과는 0이나 ChatGPT는 16% 이상의 성능이 나오는 것으로 보아 이는 메타데이터 추출과 마찬가지로 약 200개 미만의 데이터를 포함한 자료유형에서는 RoBERTa+CRF가 적은 양으로 학습하였기에 ChatGPT보다 오히려 추론 성능이 더 낮아질 수 있음을 알 수 있다. 또한, 연구보고서셋과 학술지셋의 학술지 자료유형 구분은 데이터 원문 일치 범위가 달라져도 각 모델별 0.3%p 내의 성능 차이를 보이지만, 학술지+비참고문헌 문구셋은 RoBERTa+CRF의 경우 최대 4.81%p 성능 차이가 나타나고 있다. 이를 통해 일반 문구의 포함 여부에 따라 데이터 원문의 분리 성능이 달라질 수 있음을 보여준다.

〈표 13〉은 실험 결과에서 발견한 딥러닝 언어 모델의 참고문헌 자동추출에 대한 오류 유형들을 정리한 것이다. RoBERTa+CRF는 참고문헌 원문 분리 시 URL이나 DOI같이 사이트 주소로 구성된 원문 일부만 부분 추출하였다. “https://doi.org/10.12989/aas.2019.6.1.001.”라는 DOI에서 상세 ID인 “10.12989/aas.2019.6.1.001.”를 제외한 나머지 기본 주소만 추출하는 것을 확인하였다. ChatGPT는 권(Volume)이나 호(Issue)를 추출할 때 원문을 가져오는 경향이 있었다. 예를 들어 “제4호”인 경우 “4”만 호에 해당하나, “제4호”라는 원문 그대로를 호 유형으로 인식하였다. 또한, “중소기업청 . 벤처기업협회”와 같이 특수기호가 포함된 원문에서는 “중소기업청, 벤처기업협회”처럼 다른 기호로 변경하여 추출하기도 하였다. 이는 중소기업청과 벤처기업협회를 각각 다른 정보로 인식하고, “ . ”를 이용하여 하나의 유형으로 병합한 것으로 보인다. 또한, 참고문헌 원문 분리 과제에서는 노이즈를 참고문헌의 일부로 보고 같이 분리하거나, 참고문헌은 제외하고 노이즈만 참고문헌으로 출력하거나, 기존 참고문헌과 함께 노이즈도 별도의 참고문헌으로 인식하는 경우가 있었다.

〈표 13〉 자동추출 오류 유형 및 대표 사례

모델	오류 유형	예시
RoBERTa+ CRF	자료유형 구분	Ding, H. (2014). Research on affecting factors of consumers' using willingness to third-party mobile payment. Unpublished master's thesis, Anhui University, Anhui Province, China. - 정답: 학위논문 - 모델 분류: 단행본
	메타데이터 원문 분리	Belmahi, S., Zidour, M. and Meradjah, M. (2019), 'Small-scale effect on the forced vibration of a nano beam embedded an elastic medium using nonlocal elasticity theory' Adv. Aircraft Spacecraft Sci., 6(1), 1-18. https://doi.org/10.12989/aas.2019.6.1.001. (노란 배경: 정답, 밑줄: 모델 출력 결과) - 오류 설명: DOI URL 일부를 누락함
ChatGPT (Zero-shot)	메타데이터 유형 구분	유한영 외, '정신건강 비서 서비스 기술,' 주간기술동향 1904호, IITP, 2019. - 정답: 발행기관 - 모델 분류: 발행처
	메타데이터 원문 분리	ㅇ 호(Issue) 추출 김미례, '기혼여성의 생활스트레스와 우울의 관계에서 자아존중감의 매개효과 및 조절효과 검증,' 한국심리학회지: 건강, 제12권, 제4호, pp.761-777, 2007. - 정답: 4 - 모델 분류: 제4호
		ㅇ 저자 추출 중소기업청, 벤처기업협회 (2016), 「2016년 벤처기업정밀실태조사」. - 정답: 중소기업청, 벤처기업협회 - 모델 분류: 중소기업청, 벤처기업협회
	참고문헌 원문 분리	Thomas, Nancy P. and James M. Nyce. 1998. 'Qualitative Research in LIS: Redux: A Response to a [Re]turn to Positivistic Ethnography.' The Library Quarterly, 68(1): 108-113. 법인인 경우에는 그 명칭 및 정보관리책임자의 연락처를 말한다) 35) http:// (노란 배경: 정답, 밑줄: 모델 출력 결과) - 오류 설명: 참고문헌이 아닌 노이즈도 별도의 참고문헌으로 인식하여 출력함
ChatGPT (One-shot)	자료유형 구분	Sila A, Bougatef A. Antioxidant peptides from marine by-product: isolation, identification and application in food systems. A review. J Funct Foods 2016;21:10-26. https://doi.org/10.1016/j.jff.2015.11.007. 정답: 학술지 모델 분류: 리뷰(review)
	메타데이터 유형 구분	GOLDBERG, David E.; HOLLAND, John H. Genetic algorithms and machine learning, Machine learning, 1988, 3.2: pp.95-99. DOI: http://dx.doi.org/10.1023/a:1022602019183. - 정답: 연도 - 모델 분류: 권
	참고문헌 원문 분리	... 나타낼 수 있는 관련 지표의 발굴P. Desbats, F. Geffard, G. Piolain, A. Coudray, Force-feedback teleoperation of an industrial robot in a nuclear spent fuel reprocessing plant, Ind. Robot Int. J. 33 (2006) 178-186, https://doi.org/10.1108/0143991061070300. (노란 배경: 정답, 밑줄: 모델 출력 결과) - 오류 설명: 앞부분의 노이즈를 인용문구로 함께 선택함
		... 대상국가의 거시적 경제 환경 및 ICT 시장 현 Zhao L, Ye J, Wu GT, Peng XJ, Xia PF, Ren Y. Gentiopicroside prevents interleukin-1 beta induced inflammation response in rat articular chondrocyte. J Ethnopharmacol. 2015 : 172 : 100-7. https://doi.org/10.1016/j.jep.2015.06.031. (노란 배경: 정답, 밑줄: 모델 출력 결과) 오류 설명: 참고문헌이 아닌 노이즈만 출력함

5. 맺음말 및 제언

본 연구는 단행본, 학술지, 보고서 등 다양한 종류의 발간물로 구성된 연구보고서셋의 참고문헌 데이터베이스를 효율적으로 구축하기 위한 것으로 딥러닝 언어 모델을 이용하여 참고문헌의 자동추출 성능을 비교하였다. 자동추출 과제로는 메타데이터 추출과 자료유형 구분 및 원문 분리를 수행하였다. 딥러닝 언어 모델은 NER에서 활용중인 RoBERTa+CRF와 Zero/One-shot learning 방식인 ChatGPT를 이용하였다. 실험 대상은 KISDI의 연구보고서, KISTI(2022)의 학술지 참고문헌, 학술지 참고문헌과 일반 문구를 섞은 데이터셋으로 구축하였다.

연구 수행 결과, 연구보고서셋의 메타데이터 추출 성능은 RoBERTa+CRF 모델에 대하여 F1-score 기준으로 완전일치의 경우 91.30%, 부분일치의 경우 95.41%를 달성하는 등 유의미한 결과를 얻었다. 또한 자료유형 구분 및 원문 분리 성능은 완전일치의 경우 79.49%, 부분일치의 경우 98.91%를 달성하는 등 상대적으로 학술지셋에 비해 낮은 수치이지만 유의미한 결과를 얻었다. 향후 데이터 수량을 증가시키거나 새로운 추출 모델을 적용할 경우 보다 높은 추출 성능값에 도달할 것으로 예상된다.

연구보고서셋 가운데 데이터 수량이 많은 저자, 제목 등의 메타데이터 유형을 추출한 경우 RoBERTa+CRF는 ChatGPT에 비해 Precision, Recall, F1 모두 높은 추출 성능을 기록하였다. 데이터 수량이 적은 메타데이터인 발행지나 ISSN, 자료유형인 데이터셋, 학위논문 등은 ChatGPT의 추출 성능이 RoBERTa+CRF보다 높았다. 결과적으로 실제 활용 측면에서 참고문헌 추출 시, 메타데이터와 자료유형의 학습데이터가 많은 경우에는 RoBERTa+CRF, 적은 경우에는 ChatGPT를 사용하는 것을 제안한다. 단, ChatGPT API를 이용할 때는 모델 입력 토큰 수에 비례하여 과금이 되고, 시간당 API 호출 수와 속도 제한이 있다는 점을 고려해야 한다.

마지막으로 데이터 유형에 따라 참고문헌 추출 성능 비교 결과, RoBERTa+CRF와 ChatGPT 모두 학술지 참고문헌이 들어간 데이터셋의 메타데이터 추출 성능이 연구보고서셋보다 높음을 볼 수 있었다. 한편 자료유형 구분 및 원문 분리에서도 학술지셋의 성능이 연구보고서셋보다 우위를 보였다. 이것은 학술지에 비해 공통 기술 형식이 부족한 연구보고서셋의 참고문헌 추출 난이도가 높다는 것을 보여준다.

참고문헌 추출에 있어서 본 연구는 학술지 및 특허와는 달리 통일된 양식이 없는 연구보고서를 대상으로 참고문헌 추출을 시도하였다는 점에서 기존 연구와 차별성이 있다. 특히 본 연구에서 제안하는 방법은 참고문헌의 메타데이터 추출과 자료유형 구분뿐만 아니라 비참고문헌 문구 속에서 참고문헌 원문을 파악하는 작업까지 수행함으로써 노이즈가 포함된 텍스트에서도 참고문헌을 정확히 추출할 수 있다는 강점이 있다. 또한, 딥러닝 언어 모델 간 자동추출 성능을 비교함으로써 메타데이터 및 자료유형별 추출 모델 적용 방식을 고안할 수 있다. 연구보고서와 학술지 참고문헌 간 자동추출 성능을 고려하였을 때, 연구보고서 참고문헌을 정제하고 추출할 전용 모델을 구축할 필요가 있다. 주요 오류 유형에 대한 학습데이터셋의 수량을 보강하고, 연구기관의 연구보고서를 포

함하여 다양한 자료유형의 참고문헌을 다량 학 로 생각한다.
습한다면 더 좋은 자동추출 성능을 보일 것으

참 고 문 헌

이강산다정, 이혜진, 현미환 (2022). 국가 R&D 보고서 참고문헌 기술항목 구축 개선방안 연구. 한국융합학회논문지, 13(1), 31-42. https://doi.org/10.15207/JKCS.2022.13.01.031

지선영, 최성필 (2021). 사전학습 된 언어 모델 기반의 양방향 게이트 순환 유닛 모델과 조건부 랜덤 필드 모델을 이용한 참고문헌 메타데이터 인식 연구. 정보관리학회지, 38(1), 221-242. https://doi.org/10.3743/KOSIM.2021.38.1.221

Besagni, D., Belaïd, A., & Benet, N. (2003). A segmentation method for bibliographic references by contextual tagging of fields. Seventh International Conference on Document Analysis and Recognition, 384-388. https://doi.org/10.1109/ICDAR.2003.1227694

Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. Journal of the American Society for Information Science and Technology, 57(3), 359-377. https://doi.org/10.1002/asi.20317

Choi, W., Yoon, H. M., Hyun, M. H., Lee, H. J., Seol, J. W., Lee, K. D., Yoon, Y. J., & Kong, H. (2023). Building an annotated corpus for automatic metadata extraction from multilingual journal article references. PloS one, 18(1), e0280637. https://doi.org/10.1371/journal.pone.0280637

Councill, I., Giles, C., & Kan, M. (2008). ParsCit: an Open-source CRF Reference String Parsing Package. LREC, 8, 661-667.

Dai, Z., Wang, X., Ni, P., Li, Y., Li, G., & Bai, X. (2019). Named entity recognition using BERT BiLSTM CRF for Chinese electronic health records. 2019 12th international congress on image and signal processing, biomedical engineering and informatics, 1-5. https://doi.org/10.1109/CISP-BMEI48845.2019.8965823

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. https://doi.org/10.48550/arXiv.1810.04805

Fritzler, A., Logacheva, V., & Kretov, M. (2019). Few-shot classification in named entity recognition task. Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing, 993-1000. https://doi.org/10.1145/3297280.3297378

González-Gallardo, C., Boros, E., Girdhar, N., Hamdi, A., Moreno, J., & Doucet, A. (2023). Yes but.. Can ChatGPT Identify Entities in Historical Documents?. https://doi.org/10.48550/arXiv.2303.17322

Hetzner, E. (2008). A simple method for citation metadata extraction using hidden markov models. Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries, 280-284. https://doi.org/10.1145/1378889.1378937

Hollingsworth, B., Lewin, I., & Tidhar, D. (2005). Retrieving hierarchical text structure from typeset scientific articles: a prerequisite for e-science text mining. Proc. of the 4th UK E-Science All Hands Meeting, 67-273.

Hu, Y., Ameer, I., Zuo, X., Peng, X., Zhou, Y., Li, Z., Li, Y., Li, J., Jiang, X., & Xu, H. (2023). Zero-shot Clinical Entity Recognition using ChatGPT. https://doi.org/10.48550/arXiv.2303.16416

Huang, I.., Ho, J., Kao, H., & Lin, W. (2004). Extracting citation metadata from online publication lists using BLAST. Advances in Knowledge Discovery and Data Mining: 8th Pacific-Asia Conference, 539-548. https://doi.org/10.1007/978-3-540-24775-3_64

Kim, J., Choi, N., Lim, S., Kim, J., Chung, S., Woo, H., Song, M., & Choi, J. D. (2021). Analysis of Zero-Shot Crosslingual Learning between English and Korean for Named Entity Recognition. Proceedings of the 1st Workshop on Multilingual Representation Learning, 224-237. https://doi.org/10.18653/v1/2021.mrl-1.19

Korea Institute of Science and Technology Information (2022). DeepData-REFMETA Version 1.0. http://doi.org/10.23057/47

Lauscher, A., Ravishankar, V., Vulić, I., & Glavaš, G. (2020). From zero to hero: on the limitations of zero-shot cross-lingual transfer with multilingual transformers. https://doi.org/10.48550/arXiv.2005.00633

Liu, X., Chen, H., & Xia, W. (2022). Overview of named entity recognition. Journal of Contemporary Educational Research, 6(5), 65-68. https://doi.org/10.26689/jcer.v6i5.3958

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. https://doi.org/10.48550/arXiv.1907.11692

Lopez, P. (2009). GROBID: Combining automatic bibliographic data recognition and term extraction for scholarship publications. Research and Advanced Technology for Digital Libraries: 13th European Conference, 473-474. https://doi.org/10.1007/978-3-642-04346-8_62

OpenAI (2022). Introducing ChatGPT. Available: https://openai.com/blog/chatgpt/

Park, S., Moon, J., Kim, S., Cho, W. I., Han, J., Park, J., Song, C., Kim, J., Song, Y., Oh, T., Lee, J., Oh, J., Lyu, S., Jeong, Y., Lee, I., Seo, S., Lee, D., Kim, H., Lee, M., Jang, S., Do, S., Kim, S., Lim, K., Lee, J., Park, K., Shin, J., Kim, S., Park, L., Oh, A., Ha, J., & Cho, K. (2021). Klue: Korean Language Understanding Evaluation. https://doi.org/10.48550/arXiv.2105.09680

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.

Rodrigues A. D., Colavizza, G., & Kaplan, F. (2018). Deep reference mining from scholarly literature in the arts and humanities. Frontiers in Research Metrics and Analytics, 21. https://doi.org/10.3389/frma.2018.00021

Segura-Bedmar, I., Martínez Fernández, P., & Herrero-Zazo, M. (2013). Semeval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013). Association for Computational Linguistics, 341-350.

Souza, F., Nogueira, R., & Lotufo, R. (2019). Portuguese named entity recognition using BERT-CRF. https://doi.org/10.48550/arXiv.1909.10649

Tkaczyk, D., 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. https://doi.org/10.1007/s10032-015-0249-8

Van Eck, N. & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics, 84(2), 523-538. https://doi.org/10.1007/s11192-009-0146-3

Voskuil, K. & Verberne, S. (2021). Improving reference mining in patents with BERT. https://doi.org/10.48550/arXiv.2101.01039

Wang, S., Sun, X., Li, X., Ouyang, R., Wu, F., Zhang, T., Li, J., & Wang, G. (2023). GPT-NER: Named Entity Recognition via Large Language Models. https://doi.org/10.48550/arXiv.2304.10428

Wei, X., Cui, X., Cheng, N., Wang, X., Zhang, X., Huang, S., Xie, P., Xu, J., Chen, Y., Zhang, M., Jiang, Y., & Han, W. (2023). Zero-shot information extraction via chatting with chatgpt. https://doi.org/10.48550/arXiv.2302.10205

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. (2023). A prompt pattern catalog to enhance prompt engineering with chatgpt. https://doi.org/10.48550/arXiv.2302.11382

Wu, Y., Huang, J., Xu, C., Zheng, H., Zhang, L., & Wan, J. (2021). Research on named entity recognition of electronic medical records based on roberta and radical-level feature. Wireless Communications and Mobile Computing, 2021, 1-10. https://doi.org/10.1155/2021/2489754

Yang, Y. & Katiyar, A. (2020). Simple and effective few-shot named entity recognition with structured nearest neighbor learning. https://doi.org/10.48550/arXiv.2010.02405

Zhang, X., Zou, J., Le, D. X., & Thoma, G. R. (2011). A structural SVM approach for reference parsing. BMC bioinformatics, 12, 1-7. https://doi.org/10.1186/1471-2105-12-S3-S7

• 국문 참고문헌에 대한 영문 표기
(English translation of references written in Korean)

Ji, Seon-yeong & 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. https://doi.org/10.3743/KOSIM.2021.38.1.221

Lee, Kangsandajeong, Lee, Hyejin, & Hyun, Mihwan (2022). A study on national r&d report reference technological improvement. Journal of the Korea Convergence Society, 13(1), 31-42. https://doi.org/10.15207/JKCS.2022.13.01.031