http://dx.doi.org/10.16874/jslckc.2022..65.002

고서 한자 인식 OCR의 데이터 수집과 활용 방안 고찰*

구현아**

목 차

국문초록

OCR(Optical Character Recognition)이란 사람이 쓰거나 기계로 인쇄한 문자를 컴퓨터가 읽을 수 있는 문자로 변환하는 기술로, 오늘날 우편사업, 교육, 금융 및 물류, 의료업 등에 쓰이고 있으며, 최근 우리나라 전통 고서나 고문서 안의 한자를 대상으로 범위를 확장하고 있다. 국내 고서 한자 OCR의 개발은 정부에 의해 2009년 최초로 시도된 이후, 2020년 본격적으로 1,000만자라는 대규모 단위의 체계적인 사업이 시작된 바 있다. 그러나 2020년 시행된 고서 한자 OCR 사업의 경우 다양한 한자를 수집하지 않았으며 대부분 해서만을 수집하였다는 한계점을 가지고 있다. 다양한 한자의 여러 서체를 수집하는 것은 OCR의 성능을 결정짓는 핵심적인 요소이다. 이에, 본

* 본 논문은 2022년 5월 28일 열린 춘계 중국어문학회에서 발표한 원고를 수정, 보완한 것이다. 토론을 맡아주신 문준혜 선생님 덕분에 논의를 심화, 확대할 수 있었다. 세 분의 익명의 심사위원께도 감사의 말씀을 드린다.

** 용인대학교 교양교육원 부교수

연구는 고서 한자 OCR구축을 위한 데이터 수집의 방안과 고서 한자 OCR의 활용 방안을 탐색해보았다. 먼저, 정확도가 높은 고서 한자 OCR을 구축하기 위해 고서 이외에도 서화, 예술 작품, 생활 용품 등 원천 데이터 종류를 확대할 필요가 있다. 또, 다양한 서체를 수집하기 위해 금속활자, 목활자, 목판본 등 인쇄 도구를 기준으로 할 수도 있다. 또한, 서로 다른 많은 한자를 포함하기 위해 운서나 옥편, 자전을 필수적으로 수집할 필요가 있다. 고서 한자 OCR의 결과물은 번역, 디지털 아카이브의 구축, 글꼴 개발, 관광 산업, 서체 인식을 통한 저자 및 년대 추정, 보존학에 활용될 수 있으며, 이는 사업 계획 단계에서 각 연구 기관, 교육 기관, 지역 박물관이나 역사관 등의 수요 기관과의 논의를 통해 구체적인 목표를 설정하고 진행해야할 필요가 있다.

고서 한자 OCR은 우리 문화를 담고있는 매우 중요한 기록유산으로 이에 대한 접근성을 높이는 것은 우리나라 인문학의 발전을 앞당기며, 이를 기반한 새로운 콘텐츠의 제작으로 학문, 산업 분야의 발전과 다양한 일자리를 창출하는데 기여할 것이다. 본 연구의 성과가 향후 더 높은 정확성과 사용성을 갖춘 고서 한자 인식 OCR을 개발하는데 일조할 수 있기를 기대한다.

키워드: 광학문자인식, 한자, 고서, 서체, 금속활자, 목판본, 번역, 아카이브, 글꼴

1. 서론

OCR(Optical Character Recognition)이란 ‘광학 문자 인식’, 혹은 ‘문자 인식’으로도 칭하며, 사람이 쓰거나 기계로 인쇄한 문자를 컴퓨터가 읽을 수 있는 문자로 변환하는 기술을 의미한다. 이렇게 텍스트 형태로 변환된 문자는 새롭게 편집되거나 활용될 수 있기 때문에, 오늘날 우편사업, 교육, 금융 및 물류, 의료업 등 다양한 산업 분야에서 활용되고 있다.[1] 한편, 최근 OCR은

1) OCR 제품으로는 네이버 클라우드 플랫폼, ABBYY, Adobe, Nuance, Readiris, Grooper 등이 유료로 제공되고 있으며, 구글드라이브, 네이버 웨일, MORT, 알PDF, Capture2Text, Capture2OCR, New OCP Free Online OCR이 무료로 제공되고 있다. 기타 모바일로 제공되는 서비스로 CamScanner, Adobe Scan, Office Lens, TextGrabber 등을 들 수 있다. “나무위키(namu.wiki)”, ‘OCR’검색 결과. (검색 일자: 2022년 6월 5일)

현대 인쇄물이나 손필기 뿐만 아닌 우리나라 전통 고서나 고문서의 한자를 대상으로 범위를 확장하고 있다. 이렇게 우리나라 전통 고서 안의 한자를 인식하는 OCR은 우리 문헌을 효과적으로 전산화하여 보존 가치를 높이고, 이를 기반한 각 세부 분야의 아카이브 구축, 관광산업에 응용할 수 있어 높은 활용 가치를 지닌다.

고서 한자 인식 OCR은 가장 먼저 한자 텍스트가 있는 원문 이미지의 수집, 한자 인식률을 높이기 위한 원문이미지 데이터 정제, 데이터라벨링과 텍스트의 대응으로 이루어지는 데이터 가공, OCR모듈의 구축의 과정을 거친다. 이를 간단히 그림으로 나타내면 다음과 같다.

[그림 1] 고서 한자 인식 OCR 개발 과정

보통의 OCR엔진은 흰색 바탕에 검은색 글씨로 쓰여진 이미지, 즉 배경과 텍스트의 구분이 뚜렷한 이미지에 대해서는 높은 인식률을 보이지만, 이 구분이 뚜렷하지 않은 이미지에 대해서는 저조한 인식률을 보이는데,[2] 이는 우리

2) 이로 인해 텍스트를 효과적으로 추출하기 위한 전처리 과정에서 복잡도가 높은 이미지에 대한 개선된 텍스트 영역을 검출하거나, 이미지를 전처리하는 과정을 통해 OCR의 인식도를 높이기도 한다. 박정은, 주경돈, 김철연, 「이미지 내의 텍스트 데이터 인식 정확도 향상을 위한 멀티 모달 이미지 처리 프로세스」, 『데이터베이스연구』 제34권 제3호, 2018, pp. 149-152 참조.

나라 전통 고서가 공통적으로 가진 특징으로 고서 한자 OCR을 만드는데 있어 가장 큰 난제로 작용한다. 뿐만 아니라, 원본의 노후화, 오염, 다양한 서체의 존재, 이체자의 다양함, 한자의 경계 처리의 어려움 등으로 인해 일반의 OCR에서 낮은 인식률을 보인다.[3] 이러한 이유로 고서 한자 OCR을 개발하는 일은 상당한 어려움이 따른다. 따라서 국내에서 기존에 이루어진 고서의 전산화는 OCR에 의한 자동 입력보다는 수작업에 의한 방법으로 이루어진 경우가 대부분이었다.[4]

한자를 자국의 문자로 하고 있는 중국의 경우 고서 한자 OCR에 대한 개발과 이를 통한 데이터 구축 작업이 90년대 후반 이미 시작되었다. 대표적으로 1997년 文淵閣의 『四庫全書』 電子版의 구축이 OCR기술을 활용하였고, 이외에도『中國基本古籍庫』의 구축에도 이 기술이 활용되었다.[5] 또한, SCUT Tripitaka OCR,[6] 北京如是AI研究院OCR,[7] i-慧眼OCR[8]등 다양한 고서 한자 OCR이 개발되었으며,[9] 데이터 정제, 가공에 대한 활발한 연구가 이루어지고 있다.[10] 이외에도, 웹상으로 한자로 된 고서 데이터 이미지와 디지털 텍스트

3) 장만대, 김민수, 이택헌, 김진형, 곽희규, 「필기 한자 고문서의 디지털 라이브러리화를 위한 입력 시스템」, 『2003년도 한국정보학회 가을 학술발표논문집』 Vol.30. No.2, 2005, p. 535 참조.

4) 조선시대 자료를 수작업 입력한 대표적인 결과물로 “한국어 역사 자료 말뭉치(https://kohico.kr/)”, “조선시대 외국어 학습서 DB (http://waks.aks.ac.kr/rsh/?rshID=AKS-2011-AAA-2101)” 및 “위키문헌(https://ko.wikisource.org/wiki/)”등을 들 수 있다. 원문 이미지를 제공하는 데이터베이스로는 대표적으로 “규장각한국학연구원(https://kyu.snu.ac.kr/)”, “디지털장서각(https://jsg.aks.ac.kr)”, “국립중앙도서관(https://www.nl.go.kr/)” , “디지털한글박물관(https://archives.hangeul.go.kr/)” 등을 들 수 있다.

5) 『四庫全書』의 최초 작업 시 3,452권이 구축되었고, 2014년에 이에 더하여 9,000권이 구축되었다.『中國基本古籍庫(v.8版)』에는 12,000만여권이 구축되어있다.

6) https://47.101.165.49/textv2/lineRec.html/

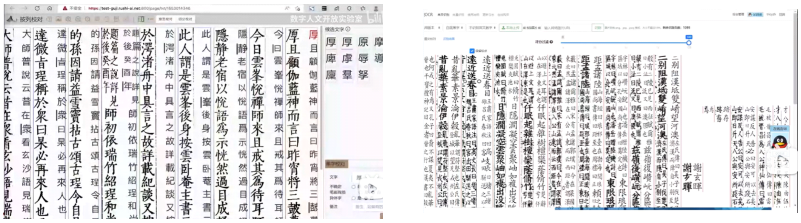
7) https://guji.rushi-ai.net:800/

8) https://dzcj.unihan.com.cn/

9) 이들 OCR은 모두 개인 정보 기기를 통한 인증을 통해 회원 가입 및 이용이 가능하다.

10) 현재 중국의 “書同文古籍數據庫(http://guji.unihan.com.cn/Trials)”에서 OCR로 인한 『四部叢刊』, 『十通』, 『大明會典』, 『大明實錄』 등 51권의 고적의 데이터베이스를 제

의 데이터셋이 이미 공개되어 있기도 하며,[11] 족보 데이터셋[12], 갑골문 데이터셋[13]이 학술 기관에 의해 구축되어 있다.



[그림 2] (좌) 北京如是AI研究院OCR, (우) i-慧眼OCR

반면, 국내에서는 매우 최근에 들어서야 정부 주도 하에 국내 고서의 한자를 기반으로 한 OCR 개발을 시작하였다. 첫 번째는 2008년 중소기업기술정보진흥원에서 진행한 “OCR기반의 한자전문인식모듈 및 인식오류처리시스템 개발”이다. 이는 정부 기관을 통해 진행한 최초의 한자인식 OCR개발 프로젝트로, 우리나라 고서를 기반으로 한 OCR엔진 및 학습엔진, 오류 처리 공정을 개발하였다는 의의를 갖는다.[14] 다음은 10여년이 지난 이후에야 재개되었다. 이는 과학기술부와 한국정보사회진흥원이 주최하는 인공지능 학습용 데이터셋 사업 중 하나인 “고서 한자 인식”사업이다. 최근 정부는 데이터를 기반으로 하여 국가 데이터경제를 활성화하기 위하여 “인공지능 학습용 데이터셋 구축”이라는 이름으로 다양한 분야의 데이터 구축 사업을 진행하고 있는데, 그 중에서도 고서 한자 인식 사업은 2020년 최초로 실시된 이래 매년 실시되

공하고 있으나, 웹사이트 상으로 개인 혹은 기관이 요청해야 열람이 가능하다. 이외에도 데이터 가공 기술인 YOLO, SSD, FasterR-CNN, Mask R-CNN, EAST, BDN 등이 개발되었다. 金連文(2022), 「古籍OCR—數據, 方法及應用」, 張弛宜(2022), 「古籍OCR二十五年工程應用力程」, “古籍智能信息處理 第四講” (https://www.bilibili.com/video/BV19S4y1B7ru) 참조.

11) http://github.com/HCIILAB/MTHv2_DATAsets_Release

12) “中文古籍族譜數據集(HDRC-Chinese)”

13) “Oracle-20K”, “Oracle AYNU”, “OBC306”등이 있다.

14) ㈜디티아이, 「OCR기반의 한자전문인식모듈 및 인식오류처리시스템 개발 최종보고서」, https://scienceon.kisti.re.kr/, 2010.

고 있다. 이는 주데이터인 고서 이미지에 포함된 1,000만자의 세그먼트 정보(바운딩박스)와 유니코드 텍스트 데이터와 이에 대한 보조데이터인 세그멘테이션 학습 데이터셋과 한자 클러스터링 학습 데이터셋을 구축하는 것을 사업 내용으로 하며, 주데이터는 고서 한자 인식 OCR모델을 도출시키는 학습에 사용될 수 있고, 보조데이터는 고서 한자 데이터 저작도구의 성능을 개선시키는 데 사용될 수 있다. 이 사업은 기본적으로 고서한자 이미지 수집, 데이터의 정제와 가공, AI모델의 개발을 주요 내용으로 하고 있지만, [그림 1]에 나타난 것처럼 실제 AI모델에 대한 응용서비스를 목표로 하여 OCR을 개발하고 있다. 이외에도, 한국연구재단 STEAM사업으로 경북대학교 주관으로 이루어진 "디지털 라키비움 구축을 위한 기계학습 기반 전통기록물 해독"사업의 결과물인 「한국 초서체 고문헌 인공지능 번역 검색 시스템(Digital Larchiveum)」이 있는데,[15] 이는 초서체 문헌에 대한 번역을 최종 결과물로 하지만, 그 과정으로서 초서체 자동 인식 인공지능 소프트웨어인 OCR도 함께 개발되었다. 그러나 이 사업에 수집된 원문이미지는 3,000장 정도이며 이 안의 18,000자를 대상으로 하고 있어,[16] 인공지능 학습용 데이터셋의 고서 한자 인식 OCR사업에 비하면 데이터 수집 규모가 매우 작다.

고서 한자 인식 OCR에 대한 기존의 연구 성과는 주로 컴퓨터공학 분야를 중심으로 데이터 정제 및 가공에 대한 방법이 논의되었다. 문자 인식을 위한 전처리 과정에서 한자 인식을 높이는 방안 (김의정 외 1996, 안성옥 외 1997, 원남식 1999, 조규태 외 2005 등), 데이터 교정 및 가공 방안 (이병희 1997, Zhao 2003, 장만대 외 2004, 장만대 외 2005, Zhong 2015, Yang 2015, 김태환 외 2017, 박정은 외 2018 등) 이 고찰 된 바 있다. 그러나, 인식률이 높은 고서 한자 OCR을 위해 원천데이터, 즉 원문이미지를 수집하는 방안이나 OCR의 활용 방안에 대해서는 거의 논의가 이루어진 바가 없다. 고서 한자 인식

15) http://dila.co.kr/minfor.php/

16) 경북대학교, 「디지털 라키비움 구축을 위한 기계학습 기반 전통기록물 해독 최종보고서」, https://scienceon.kisti.re.kr/, 2020, p. 10 참조.

OCR에 있어서 데이터 수집이나 활용 방안에 대해서 연구가 필요한 이유는 다음과 같다. 첫 번째, 다양한 데이터를 수집하는 것은 인공지능 모델의 성능을 결정짓는 핵심적인 요소이다.[17] 따라서 다양한 고서 한자의 이미지를 모으는 것이 관건이 되며, 어떤 기준으로 어떻게 모을 것인가에 대해 연구가 필요하다. 이때, 우리나라에서 쓰여진 금속활자, 목활자, 목판본과 같은 인쇄 도구는 다양한 자형을 모을 수 있는 근거가 되며, 인쇄 도구와 이를 기반으로 인출된 고서에 대한 파악을 통해 효율적으로 다양한 데이터를 수집할 수 있다. 두 번째로, 楷書, 行書, 隸書, 篆書 등 다양한 서체를 어떻게 수집할 것인가는 원천데이터의 종류와 밀접한 관계가 있다. 대부분의 해서 및 행서는 서적류에서 쉽게 볼 수 있지만, 전서, 예서, 초서 등은 예술 작품이나 생활 용품에 쓰인 경우가 많다. 따라서 한자가 쓰여진 대상, 즉 원천데이터의 종류에 대한 고찰이 선행되어야 한다. 마지막으로, OCR의 응용은 현재 우편이나 물류, 교육 분야에 쓰이는데, 고서 한자 인식 OCR의 경우 관련된 직접적인 응용 서비스가 존재하지 않으며, 학계, 교육, 문화관광 사업 등에 쓰일 수 있는 높은 가치를 갖음에도 불구하고 그 활용 방안에 대해 탐색이 이루어진 바가 없다. 특히, 도서관이나 기록관, 박물관과 같이 문화유산을 소장하고 있는 기관은 공통적으로 디지털시대에 이를 효과적으로 정리하고 전승하며, 대중에게 서비스를 제공해야 하는 과제를 안고 있다. 따라서 이러한 기관 안에서의 고서 한자 인식 OCR의 활용은 매우 중요한 기술이 아닐 수가 없다. 이외에도, 다양한 활용 가능성을 염두에 둔 고서 한자 OCR의 개발은 여러 분야에 있어서 데이터경제의 진작과 관련 일자리를 창출하는데 기여할 수 있을 것이다.

따라서, 본 연구에서는 정확도 높은 국내 고서 한자 OCR를 위한 기반으로써 다양한 서체를 어떻게 수집할 것인지에 대한 데이터 수집 방안을 고찰하

17) 인공지능 학습용 데이터 품질 확보를 위하여 한국정보통신기술협회(TTA)에서는 적합성, 정확성, 유효성을 핵심지표로 정의하고 있는데, 이 중에서도 적합성이란 원시데이터의 대표성, 포괄성, 다양성, 사실성 등 AI학습용으로 데이터셋이 적합하게 구성되어 있는지를 확인하는 지표가 된다. 이창수, 김선호, 이진우, 「빅데이터 품질 관리 표준화 현황」, 『한국정보통신기술협회 special report』, 2019, pp. 1-6 참조.

고, 이 OCR이 어떤 분야에 쓰일 수 있는지에 대한 활용 방안을 논의해보고자 한다. 이를 위해서는 2020년도에 실시된 인공지능 학습용 데이터셋 고서 한자 인식 OCR사업의 한계점을 주요 대상으로 분석하고자 한다. 기존에 선행된 OCR사업 중 이를 대상으로 하는 이유는 이 사업이 데이터 수집, 사업 규모면에서 볼 때 가장 크고, AIHUB 뿐만 아니라 각종 매체 자료를 통해서 사업 진행 상황을 상세하게 파악할 수 있기 때문이다.[18] 따라서 이 사업의 내용과 한계점을 2장에서 논의해보도록 하겠다. 3장에서는 고서 한자 OCR데이터의 원천데이터를 수집하는 방안을 크게 인쇄 도구, 서체, 고서의 세 가지 면으로 나누어 고찰해보도록 한다. 해당 사업이 조선시대 고서 한자를 대상으로 하였기에, 본 연구도 조선시대로 범위를 제한하여 데이터 수집 방안을 고찰해보고자 한다. 4장에서는 고서 한자 텍스트를 직접적 혹은 간접적으로 학계, 각 교육기관이나 연구기관, 관광사업 등에 어떻게 활용될 수 있는지, 그 응용 방안을 탐색할 것이다.

2. 현행 고서 한자 인식 OCR의 한계점

본 장에서는 2020년 인공지능 학습용 데이터셋 사업 중 '고서 한자 인식 OCR' 사업의 내용을 대상으로 이 사업의 특징과 한계점을 논의해보고자 한다.[19] 원천데이터의 유형으로 보았을 때 이 사업은 한국국학진흥원에 소장된 고문헌 중 98%를 고서를 이용하였고, 해서체 98.6%, 행서체 1.4%를 수집하

18) AIHUB(https://aihub.or.kr/)에서 고서 한자 인식 OCR에 대한 보고서, 데이터 설명서, 구축활용 가이드, 샘플데이터 등에 관한 정보를 열람할 수 있다. 이외에도, 유튜브에서도 "인공지능 학습용 데이터 교육 영상"을 시청할 수 있다. 이외에도 안동MBC NEWS "유교책판, 한자 자동 변환 프로그램 개발", 경북일보 "한국국학진흥원, 고서 속 한자 자동 번역하는 인공지능 개발"등과 같은 매체 뉴스를 참고 할 수 있다.

19) 이 사업은 2020년에 최초 시작되어 2021년에도 행해졌으나, AIHUB 사이트에는 2020년에 수행한 사업에 대한 보고서만이 열람 가능하다.

였다. 이 사업은 1000만자를 수집하는 것을 목표로 하고 있으므로 해서체가 980만자 분량, 행서체가 20만자 분량을 차지한다. 이를 정리하면 다음과 같다.[20]

구분	내용		비율 및 자수
원천데이터 유형	고서		98%
	고문서		2%
서체 및 자수	해서		98.6%, 980만자
	행서		1.4%, 20만자
판본	인쇄본	목판본	86.5%
		목활자, 연활자본, 석인본	11%
	필사본		2.5%

[표 1] 2020년 고서 한자 인식 사업 데이터 수집 상황

2008년 행해진 「OCR기반의 한자전문인식모듈 및 인식오류처리시스템 개발」이 70,000자를 수집한 것과 비교해볼 때, 위 사업은 1,000만자라는 매우 방대한 수량의 데이터를 구축했다는 점에서 크게 차이가 난다. 즉, 이는 정부 단위로 행해진 고서 한자 인식 분야의 최초의 대규모 사업이라 할 수 있다. 그러나, 기존의 체계적인 사업 성과가 부재한 만큼 다음과 같은 두 가지 면에서 중요한 한계점을 갖고 있다.

첫 번째는 원천데이터의 유형에 대한 것이다. 이들 중 대부분은 고서에 해당하는데, 이 사업의 보고서에 의하면 문집류와 경전류를 사용하였다고 언급되어 있다. 또한, 이 문헌에 주로 출현하는 한자의 수가 5천자 이내에 지나지 않으며, 문집에 출현하는 한자와 유니코드 한자를 비교했을 때 1만여자는 1회도 출현하지 않았고, 0.0001%이상 사용된 한자는 7.013자, 0.0005%이상 사용된 한자도 4,817자에 불과하다고 지적하였다. 이는 원천데이터를 주로

20) 누리IDT, 「인공지능 학습용 데이터 구축, 활용 가이드라인-고서 한자 인식-」, https://aihub.or.kr/, 2021, pp. 7-8 참조.

문집에 국한시킨 결과이다. 고서는 크게 經, 史, 子, 集으로 나뉠 수 있는데 이 중 가장 다양한 한자를 포함하고 있는 것이 經部에 있는 小學類 중에서도 자전류이다. 자전류의 대표인 韻書에는 통상 10,000자가 넘는 한자가 포함되어 있다. 고서 한자 인식 OCR는 상용하지 않는 한자에 대해서도 인식할 수 있어야 높은 성능을 지녔다고 평가할 수 있을 것이다. 따라서, 운서와 같이 다양한 한자를 포함하고 있는 자전류의 수집은 데이터의 다양성을 확보하기 위해 필수적이다.

두 번째는 서체에 대한 부분이다. 이 사업에서 수집된 서체의 98%이상이 해서이고, 1.4%가 행서에 해당한다. 즉, 대부분의 데이터가 해서에 해당하는 것이다. 행서의 비율이 매우 낮은 것은 물론이고, 초서, 예서, 전서와 같은 다양한 서체에 대한 이미지를 전혀 수집하지 않은 것은 이 사업의 가장 큰 맹점이다. 이 보고서에서는 원천 데이터 수집기관 자료의 10% 밖에 디지털화가 진행되지 않은 상태에서 각 글자별 다양한 서체를 탐색, 추적하는 것이 불가능하다고 그 이유를 언급하였다. 이는 해서를 제외한 나머지 서체를 해석하고 데이터를 구축할 수 있는 전문적 인력 및 시간의 확보에 문제가 있었던 것으로 추측된다. 우리 문헌 중 필사본의 많은 부분이 행서, 초서로 기록되어 있고, 서화, 탁본, 도장 등이 예서나 전서로 기록되어 있기 때문에 이 서체들에 대해서도 인식할 수 있도록 다양한 원천 자료의 종류를 확보하는 것은 필수불가결하다. 앞서 언급한 「한국 초서체 고문헌 인공지능 번역 검색 시스템」은 우리 문헌의 다수가 초서로 쓰여져있기 때문에 이를 인식하고 번역하기 위해 만들어진 것이다. 초서체에 특화한 인식, 번역 시스템이 생길 정도로 초서에 대한 디지털화는 매우 중요하다. 따라서, 사업 규모 안에서 일정 부분을 필수적으로 다른 서체의 번역에 따른 비용으로 책정하고, 이에 대한 데이터를 구축하지 못한 점은 매우 아쉬운 점으로 남는다.

서체, 원본의 형태, 자료 분류는 서로 연계된 요소이다. 본 사업의 데이터가 편향성을 갖는 이유는 원천데이터 수집 기관인 한국국학진흥원에서 대부분의 자료를 '유교책판'으로 수집하였기 때문이다. 이는 대부분 목판본으

로 되어있고 해서로 쓰여져 있다. 따라서, 수집 대상을 유교책판에 국한시키지 않고, 다양한 서체를 반영하고 있는 고서, 고문서를 수집하였더라면 데이터의 편향성을 없애고, 고서 한자 인식 OCR의 사용성을 높일 수 있었을 것이다.

위에 언급한 데이터 수집에 있어서의 문제 이외에도 다른 몇 가지를 언급해보고자 한다. 첫 번째는 사업 결과물의 게시, 공유에 대한 문제이다. 현재 해당 사업에 대한 결과물은 AIHUB에 관련된 보고서, 프로그램 및 가이드를 다운받을 수 있게 되어있으나, 프로그램 설치부터 오류가 발생한다. 고서 한자 인식 OCR이 클라우드 기반의 서비스를 제공하지 않고, AIHUB사이트에 들어가야만이 이 프로그램에 대한 정보를 알 수 있다는 점도 큰 한계점으로 작용한다. 이는 근본적으로 우리나라에 디지털화 결과물을 쉽게 공유하고 쓸 수 있게 해야한다는 사회적 인식과 기반이 부족한 것에 기인한다. 두 번째는 고서 한자 인식 OCR을 통한 학계, 산업 분야에의 응용 방안에 대한 탐색이 부족한 것이다. 보고서에는 한자 디지털 텍스트를 통한 다양한 분야의 콘텐츠 생산을 유도하고 이와 관련된 산업 일자리를 창출하는 것을 사업의 효과로 언급하고 있으나, 구체적인 결과물이 제시되지 않고 있다.[21] 사실, 이 사업은 이 사업의 결과물에 대한 수요 기관이 함께 참여할 수 있는 컨소시엄 형태로 이루어진다. 따라서 고서 한자 인식 OCR의 결과가 필요한 교육 기관, 연구 기관, 박물관, 역사관 등을 수요 기관으로 구성하고 이들이 필요로 하는 구체적인 응용 서비스가 어떤 것이 있는지 목표로 설정하고, 사업의 일환으로 그 응용 서비스가 함께 개발되어야 할 필요가 있다.

21) 누리IDT, 「인공지능 학습용 데이터 구축, 활용 가이드라인-고서 한자 인식-」, https://aihub.or.kr/, 2021, p. 9 참조.

3. 고서 한자 인식 OCR의 데이터 수집 방안

다양한 데이터의 수집은 최상의 OCR을 만들기 위한 가장 기본적인 조건이 된다. 다양한 데이터란 두 가지를 의미하는데, 첫 번째로는 해서, 행서, 초서, 전서, 예서 등 다양한 서체를 반영한 데이터이다. 앞서 2020년 행해진 고서 한자 OCR사업의 데이터가 대부분 해서였음을 언급한 바 있다. 특정 한자에 대해서 해서의 자형만을 인식하고 행서, 초서, 전서 등 나머지 자형에 대한 인식률이 떨어진다면 정확도 높은 OCR이라 할 수 없을 것이다. 두 번째는 서로 다른 최대한 많은 자수를 수집하는 것이다. OCR이 상용 한자만을 인식하고 상용 한자가 아닌 글자에 대해서 전혀 인식하지 못한다면 역시 고성능의 OCR이라 할 수 없을 것이다. 따라서, 본 장에서는 다양한 서체와 다양한 글자를 수집하는 방안에 대해서 논해보도록 하겠다.

1) 다양한 서체의 수집

한자 서체는 크게 해서, 행서, 초서, 예서, 전서로 나뉠 수 있으며 이 서체들이 두루 포함되도록 수집해야 사용성과 정확성이 높은 OCR을 구축할 수 있을 것이다. 주지하듯, 해서나 행서는 고서나 고문서에서 흔히 볼 수 있으나, 초서, 예서, 전서 자료는 그렇지 않다. 따라서 이번 절에서는 크게 다양한 서체를 수집하는 방안을 원천 데이터의 종류와 인쇄 도구의 면으로 나누어 살펴보고자 한다.

(1) 원천데이터의 종류

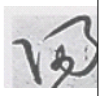
고서의 대부분은 해서 혹은 행서로 쓰여져 있다. 초서는 필사본 고서, 필첩, 간독에 많이 등장하고, 예서는 고서의 제목 표기, 서화, 서예작품, 전서는 비

석의 탁본, 도장에 많이 등장한다. 초서의 필사본은 「규장각한국학연구원」, 「한국학중앙연구원」과 같은 연구기관의 데이터베이스에 공개된 원문이미지를 이용할 수도 있으며, 도서관, 각 지역 박물관, 역사관의 소장 자료를 이용할 수도 있다. 예서의 조선시대 고서에 많이 나타나지 않으나 국립중앙박물관이나 경기도박물관, 서예박물관, 각 교육 기관이나 기타 지역 박물관 안의 서예, 서화 작품의 촬영을 통해 이미지를 수집할 수 있다. 전서 역시 이들 기관 안에 소장된 탁본, 도장 자료를 활용할 수 있다. 즉, 다양한 서체를 많은 수량 확보하기 위해서는 원천자료를 고서에만 국한시키지 않고 예술, 생활용품 등으로 분야를 확대하여 수집하고, 수집 기관 역시 다양한 곳으로 확대할 필요가 있다.[22] 예를 들면 다음과 같다.

[그림 3] (좌부터) 칠곡 선봉사 대국국사비 탁본(불교중앙박물관), 송시열 초상화(경기도박물관), 이건창 서간문(경기도박물관), 김수증 오언절구(수원박물관)

이를 통해 다음과 같이 여러 서체의 데이터를 수집할 수 있다. 아래와 같다.

22) 필자가 「규장각한국학연구원」에 소장된 자료를 대상으로 파악한 결과, 초서는 『荷齋日記』, 『槐院啓達』, 『城役及各公廨重修記』, 『公文日錄』과 같은 고서, 『秋史簡帖』, 『李三晩筆帖』 등의 간첩, 필첩 등 자료를 통해 100만자 이상 확보가 가능하며, 전서는 『印藪』, 『寶蘇堂印存』, 圓嶠眞本』등과 같은 자료를 통해 2000자이상 확보가 가능했다.

	소장처	내용	花	上	愁	歸	色
초서	성균관대학교	黃耆老 작품23)					
	제주추사관	金正喜 작품24)					
예서	경남대학교	兪漢芝 隸書 綺園帖	朝 	日 	春 	盡 	題 
	한국학중앙연구원	初搨攀雲閣					
전서	한국학중앙연구원	鄭孝俊 神道碑銘 탁본	公 	神 	道 	碑 	銘 
		鄭眉壽 神道碑銘 탁본					

[표 2] 초서, 예서, 전서 소장처 및 작품, 글자 예시

23) https://blog.daum.net/taopia1/580 (검색일자: 2022년 6월 9일)
24) 김두한, 「추사의 여정을 따라 삼각산기행시축」, 서울: 북한산국립공원사무소, 2021, pp.2-15 참조.

⑵ 인쇄 도구

다양한 서체를 수집할 수 있는 또 하나의 효과적인 방법은 다양한 인쇄 도구를 통한 수집이다. 서사자가 다른 다양한 서체를 수집하는 것이 다양한 데이터를 수집할 수 있는 방안이겠지만, 방대한 양의 자료를 어디서 어떻게 수집할 지는 여간 쉽지 않은 일이다. 따라서, 기존에 서체 혹은 서사자에 대한 연구가 이루어진 활자, 판본을 근거로 할 수 있다. 예를 들어 금속활자는 癸未字, 丙辰字, 甲寅字 등이 모두 다른 자형을 기반으로 하고 있으므로, 조선시대 제작된 금속활자로 인출된 고서를 두루 포함되게 하면 다양한 서체 데이터를 확보할 수 있는 것이다.

조선시대 고서의 인쇄 도구는 크게 활자와 판본으로 나뉠 수 있다. 활자는 금속활자, 목활자, 연활자, 도활자 등이 있고, 판본은 목판본, 석인본 등이 있다. 본 절에서는 조선시대의 주류를 차지하는 금속활자, 목활자, 목판본 인쇄물의 수집을 통해 데이터 다양성을 높이는 방안에 대해서 살펴보도록 하겠다.

① 금속활자

고려시대에는 국초부터 목판인쇄가 성행하였으나 이는 목판이 쉽게 불에 타 없어지거나 비용과 시간이 많이 걸리는 단점이 있었다. 이에, 여러 분야의 책을 수시로 찍어낼 수 있는 금속활자가 조선시대때 본격적으로 쓰이게 되었다.[25] 우리나라에서는 고려시대에도 이미 활자가 주조되었는데, 그 예로 證道歌字, 詳定禮文字, 興德寺字를 들 수 있고, 조선시대 사용된 활자는 癸未字, 庚子字, 初鑄甲寅字, 丙辰字, 庚午字, 乙亥字, 丁丑字, 戊寅字 등 36여개 금속활자가 존재한다. 을해자는 강희안 글씨, 을유자는 정난종 글씨, 숙종자는 숙종어필을 기반으로 하는 등 개인 서체를 바탕으로 한 것도 있고, 인력자, 낙동계자, 현종실록자처럼 어떤 서체를 기반하였는지 특정할 수 없는 것도 있

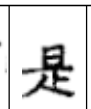
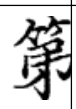
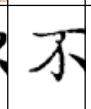
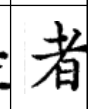
25) 천혜봉, 『한국금속활자본』, 서울:범우사, 1993, p. 11 참조.

다. 또한 경진자, 무오자, 무신자, 정유자는 갑인자를, 재주한구자와 삼주한구자는 한구자를 본 딴 것처럼, 이전에 주조된 활자 서체를 기반으로 재주조 한 것도 있다.[26] 따라서 고서를 선정할 때 위 금속활자로 인쇄된 고서가 골고루 포함하도록 한다면 데이터의 다양성을 높일 수 있다. 아래 각각의 금속활자로 인쇄된 서적과 글자의 예시를 정리하였다.[27]

종류	서명	第	有	得	之	以	不	是	者
癸未字 (1403)	『十七史纂古今通要』 『宋朝表箋總類』								
庚子字 (1420)	『資治通鑑綱目』 『重新校正入註附音通鑑外記』								
甲寅字 (1434)	『東國正韻』 『四餘纏度通軌』 『釋譜詳節』								
庚午字 (1450)	『詳說古文眞寶大全前集』 『詳說古文眞寶大全後集』								
乙亥字 (1455)	『高麗史』 『分類杜工部詩』 『唐詩正音輯註』								

26) 청주고인쇄박물관, 『直指와 金屬活字의 발자취』, 청주: 우리기획, 2002, pp. 455-458 참조.

27) 위에 나열한 것은 모두 관에 의해 제작된 관주활자에 속한다. 이외에도 민간, 사찰 등에서 조선 후기부터 금속활자에 의한 인쇄물이 나오기 시작했다. 이 경우 동일한 글자라 하더라도 글자 모양이 다르고, 획의 굵기 또한 일정하지 않으므로 논의에서 제외한다. 천혜봉, 『한국금속활자본』, 서울:범우사, 1993, pp. 175 참조.

丁丑字(1457)	『金剛般若波羅密經』								
	『金剛經五家解說誼』								
戊寅字(1458)	『易學啓蒙要解』								
乙酉字(1455)	『大方廣圓覺修多羅了義經』								
	『奎章閣志』								
甲辰字(1484)	『文翰類選大成』								
	『佛果圜悟禪師碧巖錄』								
	『左傳句讀直解』								
	『西山先生眞文忠公文章正宗』								
癸丑字(1493)	『新增東國輿地勝覽』								
庚辰字(1580)	『纂註分類杜詩』								
	『書傳大典』								
	『朱子語類』								
乙亥字體經書字(1587)	『小學諺解』								
	『大學諺解』								
	『中庸諺解』								
	『孟子諺解』								
	『孝經諺解』								
	『蘭雪軒集』								
戊午字(1618)	『詩傳大全』								
	『老乞大諺解 平壤版』								

戊申字(1668)	『三韻通考』 『御製常訓諺解』 『排字禮部韻略』								
顯宗實錄字(1677)	『顯宗實錄』								
韓構字(1677)	『行軍須知』 『曆事明原』								
元宗字(1693)	『孟子諺解』 『孟子大文』	-							
壬辰字(1772)	『大學章句大全』 『唐陸宣公奏議』								
丁酉字(1777)	『唐宋八字百選』								
壬寅字,再鑄韓構字(1782)	『奎章閣志』								
整理字(1795)	『杜律分韻』 『進饌儀軌』								
全史字(1816)	『士小節』 『華音啓蒙諺解』								

[표 3] 조선시대 금속활자의 종류와 서명, 글자 예시

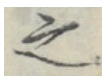
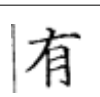
② 목활자

목활자는 금속활자의 보자(補字)로 쓰이기도 했고, 긴급하게 인쇄할 때나

특정 체재로 인쇄하고 싶은 경우 만들어져 사용되었다. 또한, 관서, 왕실, 사찰, 서원, 개인들이 모두 목활자를 만들어 금속활자와 병용하여 서적 인쇄를 사용할 정도로 발달해왔다.[28] 조선시대 초기에는 사회, 경제 질서가 완전히 회복되지 않은 상태에서 긴요한 자료를 손쉽게 목활자로 찍을 수 있었는데, 그 예로 1395년 서적원 인출의 『大明律直解』나 1395-1397년 공신도감 인출의 『開國願從功臣錄券』을 들 수 있다. 목활자는 세종 이후부터 정교하게 발달하기 시작하였는데, 1448년 반포된 『東國正韻』의 본문 큰 글자나 1455년 찍어낸 『洪武正韻譯訓』을 들 수 있다. 목활자는 개인의 서체를 반영한 경우도 있고, 기존의 금속활자의 서체를 모방하여 제작된 것도 있다. 천혜봉(2001)에 의하면 금성자는 을해자계, 추향당자는 갑진자체, 훈련도감자의 갑인자, 경오자, 을해자, 갑진자, 병자자체는 각각 금속활자의 해당 자체를 모방한 것이고, 선조실록자는 갑인자 및 을해자, 인조실록자는 경오자체를 모방한 것이다. 한편, 동국정운자와 홍무정운자는 진양대군 서체, 생생자는 사고전서 취진판 강희자전자, 춘추강자는 조윤형과 황운조의 글씨, 문계박자나 교서관필 서체자(숙종 14년경), 기영필서체자, 성천자 등은 필서체를 기반으로 하였다. 이외에도, 이 책에서 언급하지 않은 다양한 개인 서체를 반영한 목활자가 존재한다. 따라서, 데이터의 다양성을 확보하기 위해서는 금속활자와 겹치는 서체를 제외한 개인 서체를 반영한 목활자를 최대한 반영하는 것이 관건이 될 것이다. 그 예를 들면 다음과 같다.[29]

28) “우리역사넷” (http://contents.history.go.kr/front), ‘목활자’ 검색 결과. (검색일자: 2022년 6월 5일)

29) 천혜봉(2001)에 나타난 ‘한국 목활자연표’에서 금속활자를 본따 만든 목활자를 제외한 나머지 활자의 인본에 나타난 몇 글자의 예시를 든 것이다.

종류	서명	文	者	以	之	有
호음자 (16세기 후반)	『湖陰雜稿』					
효종실록자 (1660)	『孝宗實錄』					
문계박자 (1621)	『虛庵遺稿』					
교서관필체자1 (1648이후)	『纂圖互註周禮』					
교서관필체자2 (1688이후)	『箕雅』					
기영필서체자 (1791)	『五山集』					
생생자 (1792)	『御定人瑞錄』					
춘추강자 (1797)	『春秋左氏傳』					–
성천자 (1798)	『竇巖先生文集』					

[표 4] 조선시대 목활자 종류와 서명, 글자 예시

조선시대 목활자가 얼마나 많은 종류가 존재했는지 파악할 길은 없다. 다만, 장서각에서 판본이 목활자본으로 나온 서적만 해도 2800건이 나오며,[30]

30) "디지털 장서각"(https://jsg.aks.ac.kr/) '목활자' 검색 결과 참조. (검색일자: 2022년 6월 5일)

규장각에서 목활자로 간행된 서적 역시 200여건이 검색된다.[31] 또한, 이들이 기존에 연구된 어느 활자에 속하는지, 혹은 새로운 목활자로 분류할 수 있는지에 대해서도 상세한 연구가 선행되어야 할 것이다.

③ 목판본

목판본은 조선시대 인쇄 서적 가운데 가장 큰 비중을 차지한다. 목판 간행은 대량 출판이 가능하여 국가 통치에 필요한 법률서, 농업서, 의학서, 천문서, 과학기술서, 경서, 역사서 등 많은 서적을 인출, 반포하였고, 각 지역별 특색있는 판본을 새겨 민간의 서책 수요를 충족하기도 하였다. 특히 조선시대에는 문집의 간행이나 성리학적 저술, 족보등이 주를 이룬다. 조선시대에 활자 인쇄가 있었음에도 불구하고 계속해서 목판본이 서적 출판을 주도하게 된 것은 언제든지 원하는 만큼 찍어낼 수 있는 인출의 용이성과 보존의 용이성을 가지고 있었기 때문이다.[32] 목판본은 달필가에게 원고를 출판된 모양과 같도록 정서하게 한 다음, 각수가 판을 새겨 먹을 바른 뒤 찍어내기 때문에 서사자의 필체가 그대로 담겨있다. 보통 서사자는 한 명이 담당하는 것이 보통이지만, 권수가 많은 관판과 같은 경우 여러 명의 서사자가 참여하기도 하고, 신분은 관리, 승려에서 일반인까지 다양했다.[33] 이러한 이유로 볼 때, 목판본은 가장 다양한 서체를 반영하고 있는 인쇄 형태라고 볼 수 있다. 그러나, 조선시대에 간행된 목판본의 서적 수량이 워낙 방대하고, 본 세기 들어서야 각 지역, 기관에 소장된 목판들의 소재와 서지를 파악하기 시작하여 이들의 차이점을 조사하여 데이터에 포함시키는 것은 한계가 있다. 2020년도 고서 한자 인식 OCR데이터 사업은 한국국학진흥원에 소장된 목판을 대상로 작업한 것

31) “규장각한국학연구원”(https://kyudb.snu.ac.kr/) 해제에서 ‘목활자’ 검색 결과 참조. (검색일자: 2022년 6월 5일)
32) 남권희 외, 『목판의 행간에서 조선의 지식문화를 읽다』, 서울: 글항아리, 2014, pp. 12-23 참조.
33) 위의 책, p. 15, p. 45.

이므로, 향후 고서 한자 데이터는 규장각, 장서각과 같은 다른 연구 기관의 목판본이나 지역 박물관, 사찰, 각 대학 도서관 등에 소장된 목판을 대상으로 데이터를 수집할 수 있다. 예를 들어, 현재 조사가 완료된 全南 書院의 목판, 소수 박물관, 경기 강원지역 불교 경판, 전국 사찰 소장 목판 등을 활용할 수 있다.[34] 다른 서사자에 의해 제작된 목판은 모두 다른 종류의 데이터라고 할 수 있기 때문에, 서로 다른 서사자가 쓴 목판이 많으면 많을수록 데이터의 다양성도 높아지는 것이지만 현실적으로는 진행되는 사업의 규모에 맞게 수량을 제한하거나, 장기적인 계획 아래 다양한 소장처의 목판본을 수집하는 것이 가능한 대안이라 할 수 있다.

소장처	내용	有	之	以	爲	得
平昌 上院寺	木彫文殊童子坐像 腹藏遺物[35]					
독립기념관	『東經大全』[36]					-
국립중앙도서관	『北溪文集』[37]					-
원주 고관화 박물관	『佛頂心陀羅尼經』[38]					
	『五倫行實圖』[39]					

[표 5] 조선시대 목판본 소장처 및 서명, 글자 예시

34) 위의 책, pp. 28-29.
35) "문화재청(http://www.heritage.go.kr/)", '평창 상원사 목조문수동자좌상 복장유물' 검색. (검색일자: 2022년 6월 5일)
36) 윤석산, 「새로 발견된 목판본 동경대전에 관하여」, 『동학학보』 제20호, 2010, pp. 201-230 참조.
37) "한국민족문화대백과사전(http://encykorea.aks.ac.kr/)", '북계문집' 검색. (검색일자: 2022년 6월 5일)
38) "문화재청(http://www.heritage.go.kr/)", '불정심다라니경' 검색. (검색일자: 2022년 6월 5일)

이 밖에도, 다양한 서체를 수집함에 있어 몇 가지 언급할 사항이 있다. 사실, 국내에 특정 한자에 대한 다양한 서체의 용례를 모아놓은 웹사이트로 「고문서서체용례사전」가 존재한다.[40] 이는 장서각에 소장된 여러 고서의 글자를 바탕으로 한 것으로, 행서가 56.84%, 초서가 25.17%, 해서가 15.26%, 나머지 2.73%가 기타 서체로 이루어져있고, 각 유니코드 한자에 대한 각 글자의 이미지를 다운로드가 가능하다.

[그림 4] 한국학 자료포털「고문서서체용례사전」'手'의 예시

이 자료는 「한국 초서체 고문헌 인공지능 번역 검색 시스템」의 원천 자료로 이미 사용된 바 있다.[41] 그러나, 이 자료만을 갖고 고서 한자 OCR을 만든다는 것은 다음과 같은 이유로 불가능하다. 첫 번째로 가공되지 않은 그대로의 원문 이미지를 기반으로 하여 텍스트 인식률을 높이는 방법, 한자의 경계 처리 방법, 이로 인한 바운딩박스에 대한 기술이 적용되어 고서 한자 OCR

39) 위 사이트, '오륜행실도' 검색. (검색일자: 2022년 6월 5일)
40) https://kostma.aks.ac.kr/segment/segmentList.aspx/
41) Lee. Min ho, 「Traditional Archive Decoding based on Deep Learning for Digital LARCHIVEUM」, 『世界漢字學會 第8會 論文集』, 2021, p. 52 참조.

이 구축되는 것인데, 이미 바운딩박스로 나뉘어진 이미지를 통해서는 데이터 정제나 가공에 대한 AI기술을 도출할 수 없다. 고서 한자 OCR은 훈련에 의해 만들어지는 인공지능 프로그램인데, 그 원천 기술이 없기 때문에 OCR도 존재할 수 없다. 두 번째로, 위 사전 사이트로는 각 유니코드에 대한 여러 자형은 확인할 수 있어도, 고서 전체에 대한 연속적인 텍스트를 결과물로 얻을 수가 없다. 고서 한자 OCR은 원문 텍스트에 대한 데이터베이스, 번역과 같이 확장된 응용 서비스를 염두에 두고 진행되는 사업인데, 위와 같은 단순한 대응 이미지로 이와 같은 활용이 불가능하다.

이외에, 기타 서체의 데이터 정제 및 가공에 있어서의 주의 사항에 대해 언급해보겠다. 초서, 예서, 전서 자료는 이를 해독할 수 있는 전문 인력을 확보해야 한다. 초서, 예서, 전서에 대한 고서 인식 한자 원천데이터 자체가 수집된 바 없기 때문에, 이 부분은 인력에 의해 한 글자 한 글자 해독을 거쳐 수작업을 통해 입력하는 수 밖에 없을 것이다. 따라서 중국문자학 관련 전공자 중에서도 해당 서체를 전문적으로 연구한 경험이 있는 전공자를 확보해야한다. 특히 초서는 국내에 전문적으로 탈초할 수 있는 인력 자체가 매우 드물기 때문에 학계, 연구 기관, 지자체 등의 사회적 커넥션을 이용하여 전문 인력을 확보해야할 필요성이 있다. 특히, 전문 인력에 대한 인건비는 사업비와도 연계된 요소로, 전체 사업비 안에서 초서, 예서, 전서에 대한 비율을 얼마나 수집하느냐가 인건비 지출과 직접적인 연관이 있다. 따라서 이들 서체의 수집량을 잘 고려하여 전문 인력 인원을 결정할 필요가 있다. 또한, 일반적으로 사업이 제한된 시간 안에 진행되기 때문에 사전부터 원천 데이터 이미지 안의 텍스트의 양을 잘 계산하여 이에 따른 일정을 구체적으로 정할 필요가 있다.[42]

42) 특히 과학기술부와 NIA주최의 “인공지능 학습용 데이터셋” 사업의 경우 사업 기간이 7개월 남짓되는데 이 기간 안에 5천만장의 이미지와 천만자를 수집해야하므로, 상당히 제한된 기간 안에 행해지는 노동집약적인 사업이라 할 수 있다. 해서를 제외한 나머지 글자, 특히 초서는 해독에 전문성을 요하므로 사업 사전에 최대한 많은 인력을 확보하고, 사업 초기부터 집중적인 해독 작업을 해야 사업 기간에 장애를 받지 않고 사업을 진

2) 다양한 글자의 수집

고서는 크게 經, 史, 子, 集으로 나뉘는데 그 중에서도 가장 많은 종류의 한자를 두루 보여주는 것은 經部의 小學類로 분리되는 자전류라고 할 수 있으며, 이들을 우선적으로 포함하여야 데이터의 다양성을 확보할 수 있다. 정재철(2013)에 의하면 우리나라 자전류는 韻書, 類書, 玉篇, 辭典으로 나뉠 수 있는데, 이중 조선시대에 간행된 것은 辭典을 제외한 나머지 韻書, 類書, 玉篇이다. 먼저, 채영순(1986)에 의하면 조선시대 운서는 15종이 편찬되었는데, 현존하고 있는 운서로는 『東國正韻』, 『洪武正韻譯訓』, 『四聲通解』, 『續添洪武正韻(上)』[43], 『三韻通考』, 『三韻補遺』, 『增補三韻通考』, 『華東正音通釋韻考』, 『三韻聲彙』, 『華東叶音通釋』, 『奎章全韻』 11종 운서가 있다. 다음으로 類書는 한자어휘 사전으로 『訓蒙字會』 및 『類合』, 『新增類合』을 들 수 있지만, 이들의 자수는 제한적이므로 이들이 데이터의 다양성을 높여준다고 보긴 어렵다.[44] 마지막은 玉篇으로, 조선시대 『韻會玉篇』, 『新刊排字禮部韻略玉篇』, 『三韻聲彙補玉篇』, 『全韻玉篇』을 들 수 있다. 또한, 대상을 조선시대 문헌으로 국한시킨다면 근대 자전 『國漢文新玉篇』, 『字典釋要』도 추가할 수 있다. 따라서, 데이터의 다양성을 확보하기 위해서는 운서와 옥편(혹은 자전)을 우선적으로 수집해야 한다.

자전 중에서는 여러 서체를 수록한 자전도 존재한다. 1815년 현재덕이 편찬한 『草彙』를 들 수 있다. 이 책에는 각 글자에 대한 해서 및 여러 초서 자체를 수록하였다. 따라서 이를 활용하면 해당 글자에 대한 초서 자체를 더욱 쉽게 수집할 수 있게 된다. 위에 예로 든 각 고서의 종류별 예시를 들면 아래

행할 수 있을 것이다.

43) 하권은 실전되었다.

44) 『삼운성휘』가 12,965자 (을축 완영 개판본은 12,971자), 『화동정음통석운고』가 11,377자, 『규장전운』이 13345자인 것에 비하면 『유합』은 1,512자, 『신증유합』은 3,000자로, 수록자의 수량에 있어 현저한 차이가 난다.

와 같다.

[그림 5] 『삼운성휘』, 『전운옥편』, 『초휘』

이상으로 데이터의 다양성을 확보하기 위한 방법으로 다양한 서체와 글자를 수집하는 방안에 대해서 알아보았다. 데이터 수집의 내용적 측면만큼 중요한 것이 수행적 측면에 관한 것이다. 다양한 서체를 해독하고 입력할 수 있는 전문 인력의 인원, 인건비, 작업 시간에 대한 사전 조사와 계획이 있어야만이 현실적으로 좋은 성능의 고서 한자 OCR 개발로 이어질 것이다.[45]

4. 고서 한자 OCR의 활용 방안

본 장에서는 고서 한자 OCR의 각 분야에서의 활용 방안을 구체적으로 논의해보고자 한다. 먼저, OCR로 도출된 텍스트 데이터는 그 자체만으로도 해당 분야의 지식을 얻고 싶어하는 누구에게나 원문에 대한 정보를 제공해줄 수 있다. 앞에서 예로 들었던 중국의 한자 인식 OCR인 「i-慧眼」을 통해 데이터베이스화된 51종의 고적의 디지털 텍스트를 열람할 수 있다. 이외에도

45) 국내에서 수행되는 대부분의 연구 사업이 충분한 사전 조사 기간 없이 단기간 안에 노동집약적인 형태로 행해지는 것도 완성도 높은 인공지능 모델 개발에 장애가 되는 요소라고 할 수 있다. 공고의 게시와 지원서 제출의 시간적 간격이 짧아 이러한 사전 조사 작업이 이루어기 어렵다. 따라서, 연구 사업의 완성도를 높이기 위해서는 사업 기간의 연장이 필요하며, 충분한 사전 조사 기간을 통한 선정 이전 중간 평가 시스템을 도입하는 것도 하나의 방안이 될 수 있다.

중국 고서 아카이브인 「CTEXT」에서는,[46] 해당 서적에 따라 원문 이미지, 텍스트, 번역문을 제공하는데 여기서도 OCR기술이 활용되고 있다.[47] 해당 서적의 검색 결과에서 '共'이라 표기된 부분에서는 해당 원문 이미지에서 OCR을 거친 텍스트를 제공하며, '像'에서는 원문 이미지를 주고 제공하고, 때에 따라서는 텍스트 데이터도 함께 제공한다. 예를 들면 다음과 같다.

[그림 6] (좌) i-慧眼OCR , (우)「CTEXT」예시

좌측은 「i-慧眼」에서 제공하는 고서 데이터베이스이다. 이 웹사이트에서는 해당 고서에 대한 디지털 텍스트 뿐만 아니라 표점 부호가 있는 텍스트로의 전환 기능, 필기체 인식기와 같은 기능도 함께 제공하고 있다. 이는 고서 열람 시 필요한 매우 편리한 기능으로 우리나라 고서 한자 OCR개발에 있어 중요한 참고점이 되기도 한다. 우측은 청대 후기 지어진 음운 저작인 『音韻逢源』의 OCR을 적용한 이후의 결과이다. 그러나 한자에 대한 인식 결과에 적지 않은 오류가 있다.

「CTEXT」의 경우와 같이 OCR을 통한 한자 인식에 오류가 있기는 하지만, 중국은 이미 웹사이트 상에서 이를 즉각적으로 사용하는 단계에 있다. 그

46) https://ctext.org/

47) 이 사이트의 'System Statistics'부분을 참조하면 秦漢이전 문헌의 5,687,074자가 텍스트 파일로 구현되었으며, 漢代이후 문헌의 20,385,868자가 텍스트 파일로 구현되었다고 소개되어 있다.

러나, 서론에서 언급한 바와 같이 우리나라에서는 OCR프로그램 자체가 개발 초기 단계에 있기 때문에, 고서에 대한 데이터베이스는 주로 입력에 의한 수작업으로 진행되었다. 주요한 예로는 다음 몇 가지를 들어보도록 하겠다. 첫 번째로, 「한국어 역사자료 말뭉치」는 15세기부터 20세기 이르는 방대한 양의 국내 고서에 대해 텍스트 문서로 정리한 것으로, 국내 고서에 대한 텍스트를 가장 많이 제공하는 사이트로 꼽을 수 있다. 두 번째로, 「조선시대 외국어학습서 DB」는 조선시대 사역원에서 간행된 20종의 중국어, 만주어, 몽골어, 일본어 학습서의 언문(諺文) 부분을 정리한 사이트로, 텍스트와 해당 부분에 대한 이미지를 동시에 제공하고, 불완전하지만 일정 단어로 검색할 수 있는 기능도 있다. 세 번째로, 「규장각한국학연구원」을 들 수 있다. 이 사이트에서는 일반적으로 원문이미지를 제공하지만, 일부 문헌에 대해서는 텍스트를 제공하고 있다. 이 사이트는 원문이미지 열람 페이지에 '이미지+원문'의 탭이 존재하지만 여기에 텍스트를 올려놓지 않고 해당 서명의 서지 정보 부분에 텍스트를 올려놓아 사용성이 다소 떨어진다. 그 예를 들면 다음과 같다.

[그림 7] (좌)「조선시대 외국어 학습서 DB」, (우)「규장각한국학연구원」 예시

「CTEXT」와 비교해 볼 때, 국내에는 아직 고서에 대한 대부분의 데이터 가공이 스캔된 이미지만을 제공하는 제1차 데이터가공 수준에 머물러 있고, 아직 일부 서적에 대한 이미지만을 제공한다.[48] 향후 높은 정확성을 갖춘

OCR이 개발된다면 수작업을 거치지 않고도 손쉽게 고서를 데이터베이스화 할 수 있다. 최희수(2013)는 한국국학진흥원에서 진행하는 목판 아카이브 사업의 발전 방향을 제시하면서, 여기에 참여 인물이나 집단의 정보, 학맥 정보와 같은 메타데이터를 통해 아카이브를 구성할 필요가 있다고 언급한 바 있다. 효율적인 지식의 생산과 유통의 관점에서 볼 때 고서 한자 OCR의 데이터를 단순히 백과사전식으로 나열하지 않고 시기, 인물, 분야, 서지 유형과 같은 메타 데이터를 기준으로 효과적으로 정보를 얻을 수 있도록 데이터베이스를 구성하는 것도 향후 데이터베이스가 나아가야할 방향이기도 하다.

다음은 OCR로 도출된 텍스트를 번역에 활용하는 방안이다. 고서 한자 이미지에서 자동 번역 결과를 얻는 것은 2020년 고서 한자 인식 OCR사업의 보고서에서도 목표로 제기되었던 것이지만, 이 사업의 결과물은 고서 한자 인식 OCR프로그램을 개발하는 것에 그쳤다.[49] 국내에서 자동 번역이 가능한 사이트의 예 중 첫 번째는, 「한국고전번역원」의 '한문고전 자동번역'이다.[50]

[그림 8] 한국고전번역원 한문고전 자동번역 중 『승정원일기』 번역의 예

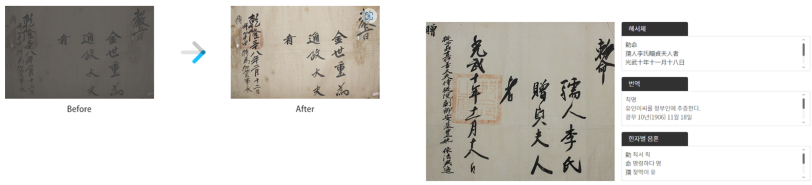
48) 이외에도 한국국학진흥원에서는 "유교넷(https://www.ugyo.net/)"에서 다양한 문집에 대한 텍스트 데이터를 제공하고 있는데, 여기에는 2020년 고서 한자 OCR 사업으로 인한 결과물도 포함되어 있을 것으로 추측된다.

49) 누리IDT, 「인공지능 학습용 데이터 구축, 활용 가이드라인-고서 한자 인식-」, https://aihub.or.kr/, 2021, p. 1 참조.

50) http://aitr.itkc.or.kr/

이는 2017년 한국고전번역원과 충남대학교가 ICT기반 공공서비스 촉진 사업의 일환으로 추진한 '인공지능 기반 고전문헌 자동번역시스템 구축'사업의 결과물이다. 2017년부터 2021년까지 5년동안의 사업으로 진행되었으며, 『승정원일기』의 자동번역 모델을 개발하였을 뿐 아니라 자동번역을 지원하는 공유 플랫폼을 구축하였다. 그러나, 『승정원일기』와 『천문고전』을 대상으로 한 자동번역 결과만을 제공하며, 아직 시험판 단계에 있어 번역 품질이 매우 낮아, 위 두 문헌을 제외한 나머지 고전 원문에 대한 자동 번역의 정확도는 매우 낮다. 또한, 아직 양질의 코퍼스를 구축하기 위한 입력 데이터의 효과, 가장 효과적인 코퍼스를 구성하기 위한 원문의 구성, 자연어 처리를 위한 정보의 구축, 번역 모델의 개선과 같은 여러 가지 과제를 안고 있다.[51]

두 번째는 앞서 언급한 「한국 초서체 고문헌 인공지능 번역 검색 시스템」이다. 이은 우리 고서의 많은 수가 초서체 (혹은 행초서)로 되어 있는 것에 착안하여 초서체로 된 고서, 고문서 등을 AI를 이용하여 판독 및 번역문을 제공하는 시스템이다.[52] 그 예를 들면 다음과 같다.



51) 김우정, 「古典文言文 기계번역의 현황과 과제」, 『중국문학』 109호, 2021, pp. 21-49 참조.

52) 이 시스템은 한국학자료센터(kostna.ask.kr), 고문서서체용례사전, 유교넷, 고문서자료관에 소장된 것과 연구팀이 모은 약 3,000종의 고서, 고문서 안에 기록된 18,000자를 대상으로 하였다. 5,779의 글자를 다시 ㄱ-ㅎ에 이르는 자군(字群)으로 분리하였다. Lee. Min ho, 「Traditional Archive Decoding based on Deep Learning for Digital LARCHIVEUM」, 『世界漢字學會 第8會 論文集』, 2021, pp. 51-52 참조.

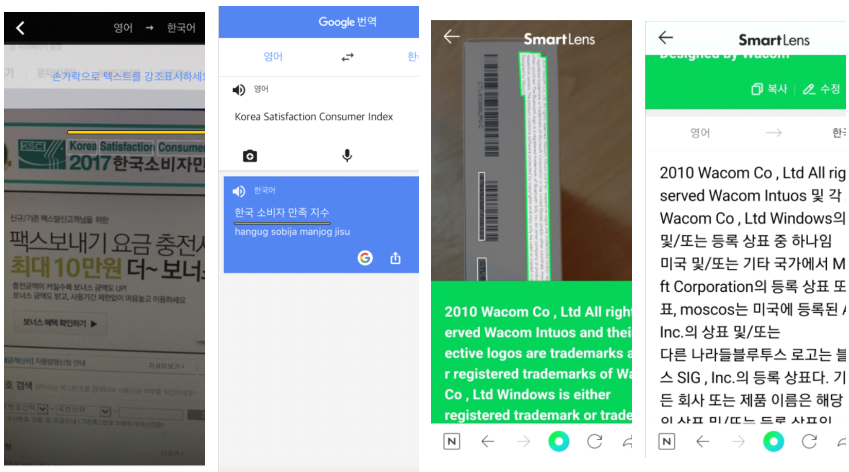
[그림 9] (좌측 위) 이미지 보정 단계, (좌측 아래) 데이터 정제 단계, (우측 위) 탈초 및 번역문 제공 단계, (우측 아래) 새로운 이미지에 대한 번역 결과 예시

이 번역시스템은 먼저 원본이미지의 화질, 밝기, 명암을 개선하는 이미지 보정단계, 원본 이미지 글자를 번역하기 용이하도록 데이터를 정제하는 단계, 문자 인식 후 이를 해서로 제시하고 번역문과 음, 훈을 보여주는 단계로 나뉜다. 이 시스템은 본 연구가 주제로 하고 있는 고서 한자 인식 OCR의 성격과 가장 유사한데, 번역에 대한 서비스를 제공하는 것을 최종 결과물로 하고, 초서를 대상으로 하고 있다는 점과 인터넷 상에서 서비스를 제공하고 있다는 점이 상이하다. 필자가 시험적으로 행초서 이미지를 본 시스템을 통해 번역을 신청한 결과, 우측 하단과 같이 이미지 분할에 문제가 많았고, 따라서 번역의 정확도도 매우 낮았다. 보고서에 따르면 이 소프트웨어의 검출기 성능은 93%, 오검출률(FPPC) 3개(50자/페이지 기준), 문자 인식률은 89%에 달하며, 번역은 BLEU-4 점수가 0.3942로, 기준 목표치를 초과하여 달성하였다고 언급되어 있으나,[53] 실제 이용 시 문자 인식과 번역의 정확도는 매우 낮았다. 따라서, 이 시스템 역시 데이터 정제와 가공 면에 있어 향후 많은 연구가 필요할 것으로 보인다.[54]

53) 경북대학교, 「디지털 라키비움 구축을 위한 기계학습 기반 전통기록물 해독 최종보고서」, https://scienceon.kisti.re.kr/, 2020, p. 4 참조.

54) 이 시스템의 데이터 정제 및 가공 과정에 대한 연구 성과로는 이장원, 장길진, 「YOLO 검출기를 이용한 한자의 강건한 위치 검출을 위한 학습자료 자동 생성 인용」, 『Journal of The Institute of Electronics and Information Engineers』, Vol55 No.7,

번역은 해당 문헌과 관계된 여러 분야의 학문적 연구에도 편리하게 사용될 수 있을 뿐 아니라, 관광산업의 진작에도 활용될 수 있을 것이다. 일반 대중이 여러 고적지나 사찰, 문화재를 접하는데 있어 가장 장애물로 작용하는 것이 한자로 된 기록물의 뜻을 파악하지 못하는 것이다. 편액, 현판, 주련, 비석 등에 나타난 한자를 바로 찍어 번역까지 바로 제공받을 수 있는 프로그램이 있다면, 이는 관광산업에 매우 획기적으로 사용될 수 있을 것이다. 현재 사진 촬영 후 즉각적으로 번역을 제공해주는 어플리케이션으로는 '구글 번역'이나 '네이버 스마트렌즈'를 들 수 있다. 그 예시를 들면 다음과 같다.[55]



[그림 10] (왼쪽에서 첫 번째, 두 번째) 구글 번역의 예시, (세번째, 네 번째) 네이버 스마트 렌즈의 예시

위 두 어플리케이션은 즉각적으로 이미지를 촬영하여 해당 텍스트에 대한 번역을 제공받을 수 있게 되어있는데, 텍스트로 인식하는 과정에서 오타가 있으며, 번역 또한 오류가 많다.[56]

2018 참조.

55) 구글 번역의 예시는 https://m.blog.naver.com/no1_hanafax/221202981945, 네이버 스마트렌즈의 예시는 https://lifenourish.tistory.com/913 참조.

이 밖에도 현재 시중에는 「한자사전 사전인식」이라는 한자를 촬영한 이후 이를 텍스트로 출력하고, 각 한자에 대한 음, 뜻을 제공하는 어플리케이션이 있다.[57] 이 어플리케이션에서는 텍스트를 읽어주기, 노트 저장, 사전 검색, 복사하기 등과 같은 부가 기능들도 제공한다. 그러나, 각 글자에 대한 정보는 알 수 있어도 전체 텍스트에 대한 번역은 제공하지 않는다. 이와 같이 이미지에 대한 글자 인식, 번역 프로그램은 아직 많은 한계점을 갖고 있다. 따라서, 명승지나 문화재 안의 한자를 찍어 바로 번역을 제공해주는 어플리케이션을 개발하기까지는 장기적인 연구가 필요할 것으로 사료된다.

다음은, 다양한 개인 서체를 기반으로 한 글꼴의 개발이다. 국내에는 현재까지 유명인의 한글 글꼴을 개발한 사례가 있다. 예를 들어, GS칼텍스가 2019년 대한민국 임시정부 수립 100주년을 맞이해 윤봉길, 윤동주, 김구, 한용운 등 독립운동가의 손글씨를 이용하여 '독립서체'를 개발한 바 있으며,[58] 국립한글박물관에서 덕온공주체를 개발하였고,[59] 뮤지컬 명성황후 25주년을 기념하여 제작사 에이콤과 폰트 제작업체 다온폰트가 명성황후체를 개발한 바 있으며,[60] 산돌구름에서 숙종의 어필로 숙종 서체를 개발하였다.[61]

56) 사진 촬영에 대한 즉각적인 번역을 제공해주는 서비스에 대한 사용성, 효용성 등에 대한 연구결과는 많지 않다. 다만, 구글, 네이버, 다음과 같은 인터넷 포털 사이트에서 '구글', '사진', '번역', '스마트렌즈' 등으로 검색했을 때 등장하는 여러 인터넷 게시글에서 두 서비스에 대해 공통적으로 오타와 번역의 오류를 지적하여 정확성이 낮다는 것을 알 수 있다. 한편, 네이버 스마트렌즈에 대한 분석은 송지성, 정다희(2020)의 「인공지능 기반 네이버 앱 검색 서비스 사용성 연구」를 들 수 있는데, 이 글에서는 스마트렌즈에 등장하는 오타에 대한 예방책을 마련할 것과 사용자에 맞는 메뉴를 구성할 것을 주장하였다.

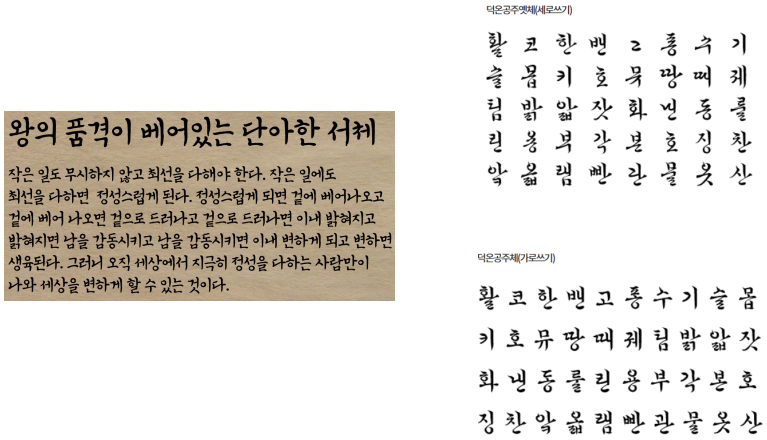
57) 이는 2019년 11월에서 출시된 어플리케이션으로, 유료로 제공되고 있다.

58) https://gscaltexmediahub.com/ (검색일자: 2022년 6월 6일)

59) https://hanfont.hangeul.go.kr/relaxfont/font/deokon.do (검색일자: 2022년 6월 6일)

60) https://www.daonfont.com/pages/shop/view.php?prd_id=10967&cate_id=2&m= (검색일자: 2022년 6월 6일)

61) https://www.sandollcloud.com/font/740.html (검색일자: 2022년 6월 6일)



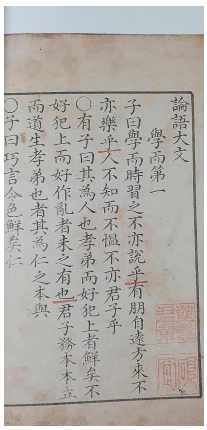
[그림 11] (좌) 숙종 글꼴, (우) 덕온공주 글꼴

만약, 한자를 기반으로 글꼴을 개발한다면 앞에서 예로 든 활자 안에 반영된 여러 개인서체를 이용할 수 있다. 예를 들어, 경오자의 안평대군 글씨, 을해자의 강희안 글씨, 정축자와 무인자에 반영된 세조 글씨 등을 서체로 개발할 수 있다. 이외에도, 특정 지역 박물관이나 역사관에 많이 소장된 자료를 이용하여 해당 인물의 서체를 개발하여 전시나 홍보, 체험프로그램에 사용할 수도 있다. 예를 들어, 경기도박물관에는 과거 경기 사대부 들의 지석이 많이 매장되어 있는데,[62] 그 중 대표적이라 할 수 있는 병자호란의 무장이었던 이완의 지석(17세기 말), 영의정 심지원의 지석(17세기 말)을 이용하여 이들의 서체를 개발, 박물관의 독자 자체로 이용하여 홍보나 전시장 내 여러 표지에 사용할 수 있다. 이 밖에도 서사자가 확실한 자료를 이용하여 해당 서사자의 글꼴을 만들어 낼 수 있다.

한자 인식 OCR은 고고학에 방면의 해석, 진위여부 판정, 검증을 위한 증거 자료로 쓰일 수 있다. 다음 자료는 개인 수집가 소장 자료인 『논어대문』인

62) 경기도박물관 학예운영실 조준호실장 서신 교류 (일자: 2022년 2월 8일)

데, 작자가 기록되어 있지 않으나 『조선왕조실록』을 기록한 이우가 기록한 것으로 추정된다.63)


論語大文
學而第一
子曰學而時習之不亦說乎有朋自遠方來不
亦樂乎人不知而不慍不亦君子乎
○有子曰其爲人也孝弟而好犯上者鮮矣不
好犯上而好作亂者未之有也君子務本本立
而道生孝弟也者其爲仁之本與
○子曰巧言令色鮮矣仁

[그림 12] 『논어대문』筆寫本

만약, 고서의 원문이미지를 수집할 때 기록자에 대한 정보도 함께 데이터로 구축한다면, 그 서체의 특징을 기반으로 새로운 자료를 발굴했을 때 작자를 쉽게 고증할 수 있을 것이다. 또한, 간행 년대 미상인 발굴 자료에 대해서도 서체를 기반으로 작자를 고증할 수 있다면 대략적인 간행 년대도 추측할 수 있다. 뿐만 아니라, 원천데이터 이미지 자체는 보존학 분야에서 문서의 변색, 손상 등의 정보를 정량적으로 평가하고 기록하는데 쓰일 수 있다.64)

이번 장에서는 고서 한자 OCR의 활용에 대해서 논의해보았다. 고서 한자 OCR은 크게 고서의 번역, 디지털 아카이브의 구축, 글꼴 개발, 서체 인식을 통한 저자 및 년대 추정에 활용될 수 있다. 위에 서술한 여러 가지 활용 방안은 무엇보다 기본적으로 고서 한자 OCR의 사용성과 정확도가 전제가 되어야 것이다. 이를 위해서는 고서 한자 OCR의 데이터 정제나 가공에 대해서 더 많

63) 개인 수집가 석한남씨 서신 교류 (일자: 2022년 2월 13일)
64) 김하영, 유우식, 「옛한글 문서의 전자문서화와 정보공유 방법 제안」, 『보존과학회지』, Vol37, No.3, 2021, pp. 267-278 참조.

은 연구가 진행되어야 함은 물론이며, 위에 열거한 학문 연구, 고고학, 관광산업 등의 소용을 미리 염두하고 고서 한자 OCR이 필요한 각 교육기관, 지역 박물관이나 역사관을 수요기관으로 지정하거나 긴밀한 협조를 하여 구체적인 활용 방안을 목표로 설정하고 사업을 추진해야할 것이다.

5. 결론

정보화 시대 속에서 많은 분야에 디지털화가 도입되고 사용되고 있으나, 우리나라에서 고서 한자에 대한 디지털화는 아직 초보적인 단계에 머물러있다. 국내 고서 한자 OCR의 개발은 정부에 의해 2009년 최초로 시도된 이후, 2020년 본격적으로 1,000만자라는 대규모 단위의 사업이 시작되었으나 이 사업 역시 대부분 목판본으로 간행된 문집 안의 해서를 대상으로 했다는 한계점을 가지고 있다. 이에, 본 연구는 고서 한자 OCR구축을 위한 데이터 수집의 방안과 고서 한자 OCR의 활용 방안을 탐색해보았다. 먼저, 정확도가 높은 고서 한자 OCR을 구축하기 위해서는 다양한 서체 자료를 수집해야 하는데, 고서는 해서나 행서로 쓰여진 경우가 대부분이므로 서화, 예술 작품, 생활용품 등 원천 데이터 종류를 확대할 필요가 있다. 다양한 서체를 확보하기 위해 금속활자, 목활자, 목판본 등 인쇄 도구를 근거로 이미지를 수집할 수도 있다. 또한, 서로 다른 많은 한자를 포함하기 위해서는 운서나 옥편, 자전을 필수적으로 수집해야한다. 고서 한자 OCR의 결과물은 번역, 디지털 아카이브의 구축, 글꼴 개발, 관광 산업, 서체 인식을 통한 저자 및 년대 추정, 보존학에 활용될 수 있으며, 이는 사업 계획 시 각 연구 기관, 교육 기관, 지역 박물관이나 역사관 등의 수요 기관과의 논의를 통해 구체적인 목표를 설정하고 진행해야할 필요가 있다.

고서 한자 OCR은 우리 문화를 담고있는 매우 중요한 기록유산으로 이에 대한 접근성을 높이는 것은 우리나라 인문학의 발전을 앞당기고 디지털 경제

를 진작시킬 수 있을 것이다. 본 연구의 성과가 향후 더 높은 정확성과 사용성을 갖춘 고서 한자 인식 OCR을 개발하는데 일조할 수 있기를 기대한다.

參考文獻

구현아, 김바로, 신수영, 엄지, 「조선시대 중국어 역학서 데이터베이스 구축 연구」, 『중국학』 78호, 2022.

김두한, 『추사의 여정을 따라 삼각산기행시축』, 서울: 북한산국립공원사무소, 2021.

남권희 외, 『목판의 행간에서 조선의 지식문화를 읽다』, 서울: 글항아리, 2014.

천혜봉, 『한국금속활자본』, 서울:범우사, 1993.

천혜봉, 『한국목활자본』, 서울:범우사, 2001.

청주고인쇄박물관, 『直指와 金屬活字의 발자취』, 청주: 우리기획, 2002.

김우정, 「古典文言文 기계번역의 현황과 과제」, 『중국문학』 109호, 2021.

김하영, 유우식, 「옛한글 문서의 전자문서화와 정보공유 방법 제안」, 『보존과학회지』 Vol37, No.3, 2021.

박정은, 주경돈, 김철연, 「이미지 내의 텍스트 데이터 인식 정확도 향상을 위한 멀티 모달 이미지 처리 프로세스」, 『데이터베이스연구』 제34권 제3호, 2018.

이남희, 「고문헌 디지털 아카이브 구축과 한자 처리 문제」, 『嶺南學』 제17호, 2010.

이소연, 「함께 만드는 미래: 디지털 융합과 문화유산기관의 협력」, 『정보관리학회지』 제29권 제3호, 2012.

이장원, 장길진, 「YOLO 검출기를 이용한 한자의 강건한 위치 검출을 위한 학습 자료 자동 생성 인용」, 『Journal of The Institute of Electronics and Information Engineers』 Vol55, No.7, 2018.

이창수, 김선호, 이진우, 「빅데이터 품질 관리 표준화 현황」, 『한국정보통신기술협회 special report』, 2019.

장만대, 김민수, 이택헌, 김진형, 곽희규, 「필기 한자 고문서의 디지털 라이브러리화를 위한 입력 시스템」, 『2003년도 한국정보학회 가을 학술발표논문집』 Vol.30. No.2, 2005.

정다희, 「인공지능 기반 네이버 앱 검색 서비스 사용성 연구」, 『한국디자인문화학회지』 vo.27, no,1, 2020.

최동빈, 강윤희, 조인수, 박용범, 「한자 이미지 분할 기법 및 Mask R-CNN 성능 평가」, 『Journal of Platform Technology』 vol.7, no.3, 2019.

최희수, 「한국국학진흥원 목판 아카이브 구축 및 서비스 방향에 대한 연구-메타데이터의 검토 및 재구성 방안을 중심으로」, 『국학연구』, 2013

Lee, Min ho, 「Traditional Archive Decoding based on Deep Learning for Digital LARCHIVEUM」, 『世界漢字學會 第8會 論文集』, 2021.

S. Zhao, 「Two-stage segmentation of unconstrained handwritten Chinese characters」, 『Pattern Recognition』 36 (1) 2003.

W. Yang, 「Improved deep convolutional neural network for online handwritten Chinese character recognition using domain-specific knowledge」, 『Document Analysis and Recognition (ICDAR)』, 13th International Conference on, 2015.

Z. Zhong, 「High performance offline handwritten Chinese character recognition using GoogLeNet and directional feature map」, 『Document Analysis and Recognition(ICDAR)』, 13th International Conference on IEEE, 2015.

㈜디티아이, 「OCR기반의 한자전문인식모듈 및 인식오류처리시스템 개발 최종보고서」, 2010.

누리IDT, 「인공지능 학습용 데이터 구축, 활용 가이드라인-고서 한자 인식-」, 2021.

경북대학교, 「디지털 라키비움 구축을 위한 기계학습 기반 전통기록물 해독 최종보고서」, 2020.

“규장각한국학연구원(https://kyu.snu.ac.kr/)”

“국립중앙도서관(https://www.nl.go.kr/)”

“나무위키(namu.wiki/)”

“디지털장서각(https://jsg.aks.ac.kr)”

“디지털한글박물관(https://archives.hangeul.go.kr/)”

“문화재청(http://www.heritage.go.kr/)”
“우리역사넷 (http://contents.history.go.kr/fron/t)”
“유교넷(http://ugyo.net/)”
“조선시대 외국어 학습서 DB(http://waks.aks.ac.kr/rsh /?rshID=AKS-2011-AAA-2101/)”
“한국민족문화대백과사전(http://encykorea.aks.ac.kr/)”
“한국어 역사 자료 말뭉치(https://kohico.kr/)”
“한국 초서체 문헌 인공지능 번역 검색 시스템(http://dila.co.kr/minfor.php/)”
“한문고전자동번역(http://aitr.itkc.or.kr/)”
“北京如是AI硏究院OCR(https://guji.rushi-ai.net:800/)”
“i-慧眼OCR(https://dzcj.unihan.com.cn/)”
“AIHUB(https://aihub.or.kr/)”
“Chinese Text Project(https://ctext.org/)”
“SCUT Tripitaka OCR(https://47.101.165.49/textv2/lineRec.html/)”

Abstract

A Study on methods of Data Collection and application for Optical Character Recognition of Chinese Character in ancient book

Khoo, Hyun-ah

Digitalization has been introduced and used in many fields in the information age. But digitalization of ancient Chinese characters in Korea is still at an elementary stage. The development of OCR for ancient Chinese characters in Korea was first attempted by the government in 2009. After that, a large-scale systematic project of 10 million character in 2020, but this project also has a limitation in that most of it collected books published in woodblock prints. Therefore, this study explored data collection methods for the establishment of OCR for ancient Chinese characters and the use of OCR of old Chinese characters. First, in order to build OCR of ancient Chinese characters with high accuracy, various typeface data must be collected. And since old books are mostly written in the square style of Chinese handwriting or the semicursive style of writing, they must expand the types of source data. Such as calligraphy, art works, and household goods. Various fonts can be collected based on printing tools. Such as metal type, wood type, and woodblock print. The diversity of data can be secured by allowing the font to include different types and blco books. In addition, it is essential to collect Chinese rhyme book, Okpyeon, and dictionaries to include many different Chinese characters. For example, 『Hongmu Jeongwun yeokhun』, 『Sasungtonghae』, 『Samwunsunghui』, 『Gyujangjeonwun』, etc.

The results of ancient Chinese characters OCR can be used in translation, digital archive construction, font development, tourism industry. Author and publication period can be ascertained through font recognition. And it can be used in preservation studies, for example, the degree of damage can be proven by the original image.

OCR, an ancient Chinese character, is a very important record heritage

containing Korean culture. Digitalization of heritage will accelerate the development of Korean humanities and contribute to the development of academic and industrial fields. And it can create various jobs through the production of new contents. It is hoped that the results of this study will help develop ancient Chinese character recognition OCR with higher accuracy and usability in the future.

Key words : OCR, Chinese character, ancient book, the style of writing Chinese characters, metal type, wood block, translation, archive, font

투 고 일 : 2022. 7. 10. / 심 사 일 : 2022. 7. 15.~ 2022. 8. 15. / 게재확정일 : 2022. 8. 20.