●○ OCR 과제

다양한 형태의 한글 문자 이미지 인식 데이터

●○ 개요: 다양한 형태의 한글 문자 이미지 인식 데이터셋이란?

자연어처리(NLP, Natural Language Processing) 분야에서 한글의 광학글자인식(OCR, Optical Character Recognition)기술 개발에 활용할 수 있는 학습 데이터셋으로 주식회사 엔티정보에서 구축했으며, 약 50만건이상의 다양한 형태의 한글 문자 이미지 학습데이터로 구성되어 있다.

OCR문자 인식은 주로 이미지 파일로 된 문자를 따로 타이핑 할 필요없이 텍스트 파일(Word, Excel 등)으로 변환하는 기술이며, AI와 OCR을 융합하여 다양한 유형의 문서 이미지를 분석/분류하고 디지털화해야 하는 로봇 프로세스 자동화(RPA)에서 핵심 기반 기술로 평가받고 있으며 인공지능 딥러닝으로 향상된 인식 정확도화 빠른 처리속도로 사람이 반복적으로 수행해야되 하는 단순 업무를 자동화하는데 OCR 을 활용하기 위해 인식율을 높이기 위한 연구 및 문서불류 서비스나 보안솔루션 등 다양한 상용화서비스를 진행하고 있다.

●○ 사례

| 사례 1 |

- K 기관의 거래명세표 인식을 통한 중도매 식재료 통계 분석
  - 거래명세는 서식표준화가 안되있으며, 특히 영세 중도매 상들이 자체 서식으로 각각 제작해 발급하고 있어 기존 일반 OCR로는 인식이 어려우며, 또한 현장 중도매 상인 들이 직접 스마트폰으로 촬영해 해당 식자재 품목의 거래 내역(수량, 금액) 등을 전송하기가 어려움
  - AI 기반의 인식기술을 적용, 기존 인소체 및 손글씨 등의 학습을 활용하여 거래명세서의 삭자재 품목과 금액, 수량을 인식, 한글 및 문자로 전환하고, 일부 인식이 안되거나, 잘못 인식된 데이터는 촬영자가 현장에서 수정하혀 K기관 서버로 전송, K 기관에서는 해당 텍스트와 데이터를 활용 주기적인 식자재의 중도매 흐름과 통계 분석을 할 수 있도록 함

| 사례2 |

- SK텔레콤의 음악플랫폼 플로(FLO)가 OCR 기술을 이용해 선보인 '캡쳐 이미지로 플레이리스트 만들기'에 대한 사용자 반응이 뜨겁다. 서비스 출시 두 달여만에 총 490만 곡이 플로로 이동되고 약 16만 개의 플레이리스트가 생성된 것으로 나타났다.
- 플로는 지난 8월20일부터 OCR(Optical Character Recognition, 광학 문자 인식) 기술을 적용한 '캡쳐 이미지로 플레이리스트 만들기' 기능을 도입하며 신규 고객 확보에 나선 바 있다.
- 캡쳐 이미지로 플레이리스트 만들기는 OCR 기술을 활용, 이미지에서 텍스트를 인식해 추출하는 기술로 스크린 캡쳐만으로도 간편하게 음악 플레이리스트를 자동 생성해 준다. 음악 플랫폼 변경 시 이용자들이 겪는 플레이리스트 이동의 불편함을 OCR 기술로 개선했다.

●○ 데이터셋의 구성

본 데이터셋은 국립국어원의 완성형 한글 2,350자와 실생활에서 활용도가 높은 단어, 각종 신청서, 계약서, 청약서, 고지서, 발주서, 검수서 등에서 사용되는 글자 단어를 선별하여 데이터를 수집하였으며 인쇄체 70,500장과 손글씨 453,600장의 총 524,100장 이미지와 각 이미지의 Json파일로 구성 되어있다.

해당 데이터셋을 통해 글자인식 알고리즘과, 단어인식 알고리즘을 개발할 수 있도록 했다.

표1 | 데이터셋의 구성

데이터 종류	포함 내용	구축량	제공방식
인쇄체	• 완성형 글자 2,350자 • 서로 다른 3개 사이즈(28*28, 54*54, 128*128) • 유·무료 폰트 10개	70,500	*.jpg *.json 파일세트
손글씨	• 완성형 글자 2,350자 • 한국어 학습용 어휘 목록(6,000단어) (동음이의어와 한글자 단어를 제외한 5,210단어) • 성별 연령이 다른 작성자 120명(모든비율은 1:1)	453,600	*.jpg *.json 파일세트

●○ 데이터셋의 설계 기준과 분포

기본적으로 완성형 2,350자와 실생활에 활용도가 높고 각종 신청서와 계약서, 청약서, 고지서, 발주서, 검수서 등에서 사용되는 단어등을 선별하여 한글 글자체 이미지 데이터 수집을 진행하였다.
인쇄체 폰트선정에 대해 관공서의 서식에 사용된 폰트, 한국저작권위원회의 2020년 상반기 폰트 저작물 인기순위를 활용하여 인쇄체 한글 이미지획득을 위한 폰트를 선정하였다.

손글씨의 경우 다양한 손글씨를 수집하기 위해 연령층과 성비균형을 통하여 수집하였다.

그림1 | 데이터셋 구성 개요

- 인쇄체 : 폰트선정은 주로 관공서의 민원신청서류 통해 조사하였으며 과거 법정서식의 경우 바탕, 바탕체, 신명조, 한컴바탕 등으로 작성되어있으며 개정을 통해 최근에 정비된 법정서식은 HY헤드라인M, 견고딕, 돋움, 돋움체로 등이 사용되었으며 한국저작권위원회의 2020년 상반기 폰트 저작물 인기순위중 상위 5위의 폰트를 사용하여 폰트를 선정하였다.
- 손글씨 : 손글씨 작성자에 대한 성비와 연령은 균등한 수집을 원칙으로 수집되었다.

표2 | 손글씨 데이터 수집 분포

글자	작성자 분포					
	20대 이하	30대	40대	50대 이상	합계	성비
완성형 2,350자 5,210단어	30	30	30	30	120	1:1

●○ 데이터 구조

데이터셋에 따른 항목과 해당 값은 아래 테이블과 같다.

No.	항목		길이	타입	필수 여부	비고
	한글명	영문명				
1	데이터셋 정보	Info				
1-1	데이터셋명	Name	100	String	Y	
1-2	데이터셋설명	Description	100	String	Y	
1-3	데이터셋생성일자	Data_created	100	String	Y	
2	이미지 정보	Image				
2-1	이미지파일명	File_Name	100	String	Y	
2-2	이미지파일경로	File_URL	100	String	Y	
2-3	이미지너비	Width	100	Number	Y	
2-4	이미지높이	Height	100	Number	Y	
2-5	이미지해상도	Dpi	100	Number	Y	
2-6	이미지 컬러	Bit	100	Number	Y	
3	텍스트정보	Text				
3-1	텍스트 타입	Type	100	String	Y	
3-2	출력형태	Output	100	String	Y	
3-3	문자	letter	100	String		Type=letter
3-3-1	문자 값	Value	100	String		Type=letter
3-4	단어	Word	100	String		Type=Char
3-4-1	문자박스	Charbox	100	String		Type=Char
3-4-1	문자 값	Value	100	String		Type=Char
4	라이선스 값	License				
4-1	출력형태	Output	100	String	Y	
4-2	폰트종류	Font	100	String		Output=Prient
4-3	폰트번호	Font_No	100	Number		Output=Prient
4-4	폰트라이선스업체명	Font_License	100	String		Output=Prient
4-5	폰트라이선스주소	Font_URL	100	String		Output=Prient
4-6	수기작성자 번호	Writer_NO	100	Number		Output=Handwriter
4-7	수기작성자 성별	Writer_Gender	100	String		Output=Handwriter
4-8	수기작성자 연령	Writer_Age	100	Number		Output=Handwriter

그림2 | 데이터셋 구성 개요

●○ 데이터 예시

이 데이터는 인쇄체와 손글씨의 각각 문자와 단어에 대한 데이터이며 인쇄체일 경우 license 항목 중 Writer에 대한 항목이 없으며 손글씨의 경우 license 항목 중 Font에 대한 항목이 없다.
문자와 단어 데이터셋의 차이는 Text 항목의 Type에 따라 결정되며 단어의 경우 charbox에 대한 항목이 없다.

표3 | 문자 데이터셋 구성 개요


```
{
        "info": {
                        "name": "Korean
OCR Data Set (Sample)",
                        "desciption": "Korean
OCR Data Set (Printed Text, Sample)",
                        "date_created":
"2019-08-24 20:06:55"
        },
        "image":{
                "file_name":"10001003.jpg",
"file_url":"c:\date\prient\10001003.jpg",
                "width":125,
                "height":125,
                "dpi":300,
                "bit":24
        },
        "text":{
                "type":"letter",
                "output":"print",
                "letter":{
                                "value":"간"
                }
        },
        "license":{
                "output":"print",
                "font":"굴림",
                "font_no":001,
                "font_license":"폰트나라",
                "font_url":"www.license.com",
                "writer_no":,
                "gender":"",
                "age":""
        }

}
```


인쇄체 문자


```
{
        "info": {
                        "name": "Korean
OCR Data Set (Sample)",
                        "desciption": "Korean
OCR Data Set (Handwrite Text, Sample)",
                        "date_created":
"2019-08-24 20:06:55"
        },
        "image":{
                "file_name":"30001003.jpg",
"file_url":"c:\date\handwrite\30001003.jpg",
                "width":127,
                "height":112,
                "dpi":300,
                "bit":24
        },
        "text":{
                "type":"letter",
                "output":"handwrite",
                "letter":{
                                "value":"간"
                }
        },
        "license":{
                "output":"handwrite",
                "font":"",
                "font_no":
                "font_license":"",
                "font_url":"",
                "writer_no":001,
                "gender":"male",
                "age":"32"
        }

}
```


손글씨 문자

표4 | 단어 데이터셋 구성 개요


```
{
        "info": {
                        "name": "Korean
OCR Data Set (Sample)",
                        "desciption": "Korean
OCR Data Set (Printed Text, Sample)",
                        "date_created":
"2019-08-24 20:06:57"
        },
        "image":{
                "file_name":"20001009.jpg",
"file_url":"c:\date\prient\20001009.jpg",
                "width":259,
                "height":111,
                "dpi":300,
                "bit":24
        },
        "text":{
                "type":"word",
                "output":"print",
                "word":[
                        {

"charbox":[20,4,109,104],
                                        "value":"예"

                        },
                        {

"charbox":[150,5,246,100],
                                        "value":"상"
                        }
                ]
        },
        "license":{
                "output":"print",
                "font":"굴림",
                "font_no":001,
                "font_license":"폰트나라",
                "font_url":"www.license.com",
                "writer_no":,
                "gender":"",
                "age":""
        }

}
```


인쇄체 단어


```
{
        "info": {
                        "name": "Korean
OCR Data Set (Sample)",
                        "desciption": "Korean
OCR Data Set (Handwrite Text, Sample)",
                        "date_created":
"2019-08-24 20:06:55"
        },
        "image":{
                "file_name":"40001009.jpg",
"file_url":"c:\date\handwrite\40001009.jpg"
                "width":245,
                "height":112,
                "dpi":300,
                "bit":24
        },
        "text":{
                "type":"word",
                "output":"handwrite",
                "word":[
                        {

"charbox":[17,18,111,93],
                                        "value":"예"

                        },
                        {

"charbox":[149,12,227,102],
                                        "value":"상"
                        }
                ]
        },
        "license":{
                "output":"handwrite",
                "font":"",
                "font_no":
                "font_license":"",
                "font_url":"",
                "writer_no":001,
                "gender":"male",
                "age":"32"
        }
```


손글씨 단어

●○ 데이터 구축 과정

기본적으로 한글 단어를 여러 가지 데이터를 확보하기 위해 인쇄체의 경우 다양한 폰트를 사용 출력 후 스캔하여 구축하였으며 손글씨의 경우 성별/나이등의 인원을 고려하여 120명의 수집자를 모집하여 글자와 단어 총 118장으로 구성된 WorkSheet를 작성한 뒤 스캔하여 데이터를 구축하였다.
스캔한 손글씨는 1차적으로 육안을 검사를 통해 오탈자와 작성지침에 벗어난 데이터를 정제하며 정제 과정 중 사람이 발생시킬 수 있는 추가적인 오류는 검증하기위해 이미지의 자동Croping과 바코드를 활용하여 바코드 정책에 따라 각 Croping이미지의 파일명과 메타데이터를 자동으로 입력하여 생성한다.

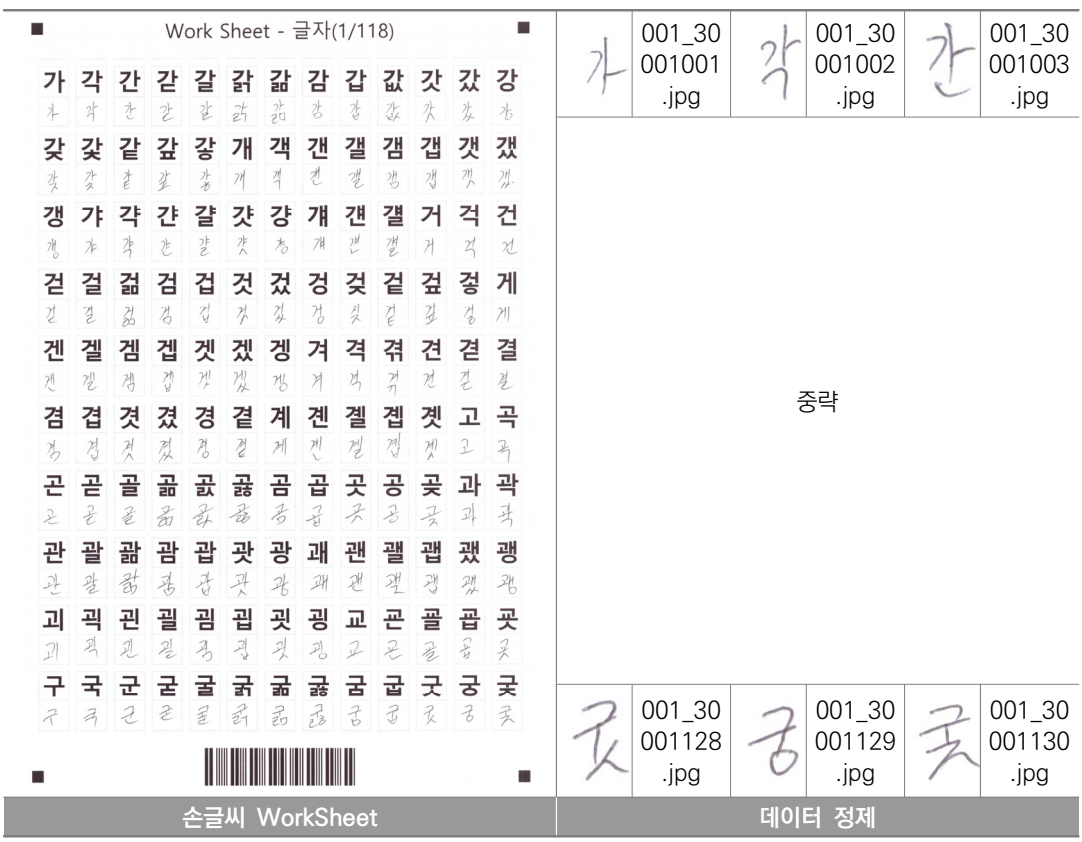

그림3 | 데이터 획득 WorkSheet와 데이터 정제

이렇게 정제된 이미지는 효율적인 데이터 제작을 위한 웹 기반 가공툴에 업로드되며 정제된 이미지와 메타데이터가 동일 여부를 2차 점검함과 동시 글자 단위의 라벨링 작업과 단어의 라벨링 및 바운딩 박스 작업을 진행하게 된다.

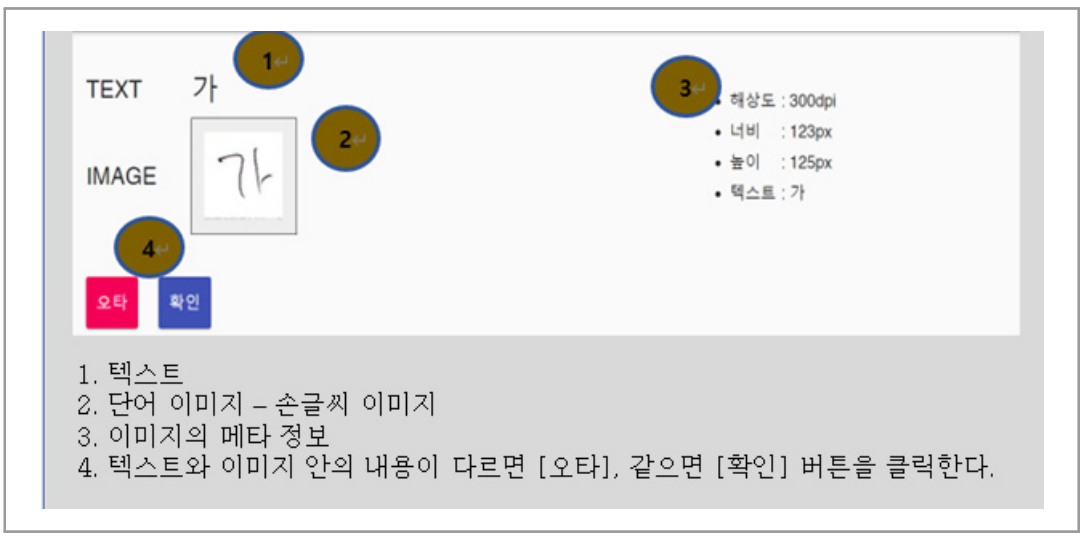

그림4 | 효율적인 데이터 제작을 위한 웹 기반 툴

구분	글자	단어
인쇄체	갓 – 갓	– 연출하다
손글씨	갓 – 갓	– 오래간만

그림5 | 가공 이미지

●○ 검수와 품질 확보

대량의 데이터를 높은 품질로 생성하기 위하여, 단가는 낮지만 품질 관리가 어려운 크라우드소싱 방식의 데이터생성 작업을 보완하기 위한 검수 프로세스의 정립은 데이터셋 구축에 매우 중요한 의미를 갖는다. 이 데이터셋에서는 3단계 검수 체계를 구축했는데, 크라우드 워커들이 작업한 결과물을 가이드라인에서 제시한 형식에 맞는지 체크리스트를 활용하여(글자 – 라벨링정보 등, 단어

라벨링정보, 바운딩박스정보 등) 크라우드 워커의 1차 교차검수가 이루어지고, 교차검수 완료된 결과물에 대해서 유효한지 검수하는 데이터 구축 참여업체 직원으로 구성된 재검수자가 팀이 있으며 이렇게 만들어진 데이터셋을 전체적으로 들여다보며 확인하는 수행업체의 전문가검수팀 배치하여 최종적인 데이터셋의 품질을 담보할 수 있었다.

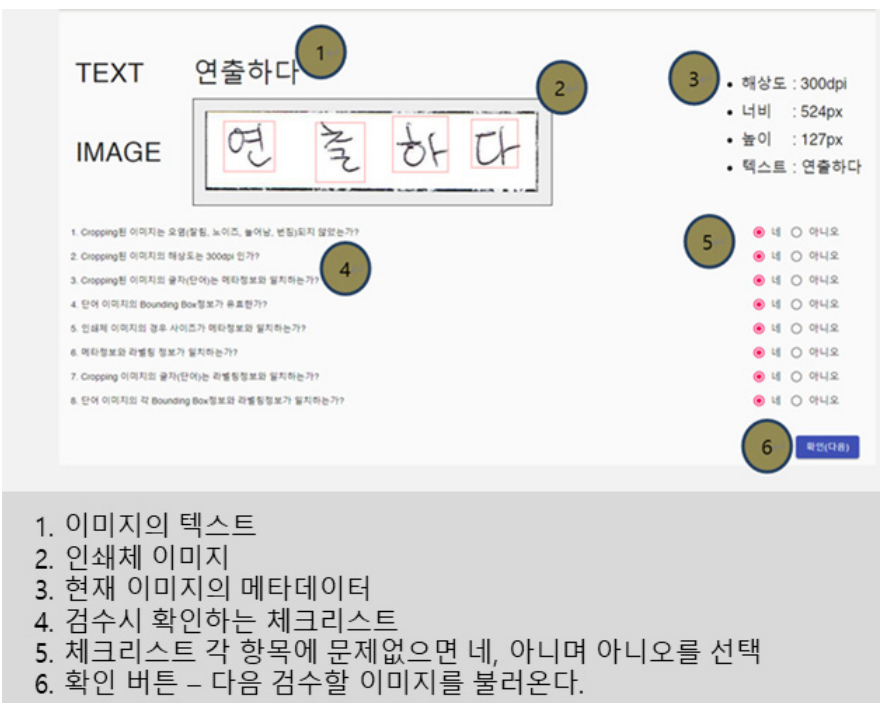

그림6 | 품질 확보를 위한 3단계 품질 검수 체계

●○ 데이터 구축 담당자

수행기관(참여) : 주식회사 엔티정보
(전화: 043-225-7890, 이메일 : csy78990@gmail.com)