AI 데이터 구축 가이드라인

03
한국어 글자체 이미지 AI 데이터 구축 가이드라인

Superb AI

	성명	소속	직위
가이드라인 제안	이현동	Superb AI	이사
가이드라인 작성	이현동	Superb AI	이사
	이종혁	Superb AI	이사
	이정권	Superb AI	이사
전문가 검토 위원	구영현	세종대학교	교수
	배동석	TTA	책임
	김송이	TTA	선임

※ 주관: 키니앤파트너스/ 참여: 슈퍼브에이아이

서 문

1. 가이드라인의 목적

이 가이드라인의 목적은 한글 관련 인공지능 기술개발을 위해, 한글을 Labeling하기 위한 절차와 세부지침에 제공하기 위해 작성되었다. 이를 통해 구축Task 진행하는 과정에서 발생되는 비용(커뮤니케이션 코스트 등)을 축소하고 향후 연구 개발자간에 실제 기술 개발에 필요한 가이드라인을 제공하기 위해 작성되었다.

한글 글자체 이미지 데이터는 자연어 처리 분야에서 가장 기초적인 언어 자원으로 중요성이 높아지고 있다. 반면, 우리나라에는 인공지능 기술 벤처들의 수도 많지 않으며, 관련 기술 개발을 위한 데이터 기반들도 부족한 상황이다.

2. 주요 내용 요약

한국정보화진흥원의 한국어 글자체 이미지 AI데이터의 경우, 손글씨 및 인쇄체 총 500만자 및 Text in the wild 10만장의 이미지로 구축되었다. 본 구축가이드에서는 해당 데이터 구축에 활용된 기준과 수집, 가공 및 검수절차까지의 내용을 요약한다.

3. 인용 문서와의 비교

3.1 인용 문서와의 관련성

- 해당 사항 없음

3.2 인용 표준 또는 문서와 본 가이드라인의 비교표

- 해당 사항 없음

한국어 글자체 이미지 AI데이터 구축

Guideline for Korean Character Dataset

03

1. 적용 범위

4차 산업혁명의 핵심 기반인 산업별 실제 데이터, 인공지능(AI) 학습데이터를 전방위적으로 구축하고, 공공데이터의 원칙적 개방 등 공공·민간 데이터의 획기적 개방과 데이터 거래 활성화가 실행되는 상황임.

그럼에도 불구하고 한글에 대한 대규모 학습용 데이터는 부재한 상황으로, 이에 당사는 한국 글자체 인식 처리 기술의 표준화 및 교환·공유할 수 있는 체계를 마련하고 한글 인식기술 개발의 토대를 구축하기 위해 한글형 글씨체 DB 확보하고자 함

본 표준은 한국어 글자체 이미지 AI데이터 구축을 위한 절차와 Labeling 지침을 제시하며, 지식베이스 구축 및 효용성 검증으로 제한함. 효과적인 한글 기반 인공지능 기술개발을 위해 활용 가능한 이미지 및 Labeling 정보 등을 지식베이스로 구축·공개하고자 함

※ 본 과제에서 활용되는 국어 표준은 국립국어원이 지정한 기준을 차용해 사용됨

2. 인용 표준

• 해당 사항 없음

3. 용어 정의

● 3.1 구축 한글 글자체 정의

— 음절(syllable) : 자음과 모음으로 합성된 하나의 말소리 단위. 음절은 몇 개의 음소로 이루어지며, 모음은 단독으로 음절이 될 수 있음

— 단어(word) : 분리하여 자립적으로 쓸 수 있는 말 혹은 이에 준하는 말

— 어절(syntactic word) : 문장을 구성하고 있는 각각의 마디로 문장 성분의 최소 단위로서 띄어쓰기 단위

— 문장(sentence) : 생각 및 감정을 말로 표현할 때 완결되는 내용을 나타내는 최소 단위

● 3.2 Labeling 용어 정의

— Annotation : 해당 데이터 내의 이미지 및 텍스트, 숫자 등의 정보 값을 입력하는 과정을 뜻하며, 해당 정보 값은 프로그래밍 언어로 구성되어 있음

— bounding box : 좌상단, 우하단의 좌표로 구성되어 있는 4point polygon 직사각형

— Rectangle : 4point polygon으로 Center, Width, height, angle의 자유도 5를 가지는 직사각형

● 3.2 기술 용어 정의

— 데이터 증강 (augmentation) : Overfitting의 문제를 해결하기 위해 필요한 학습용 데이터의 양을 늘리는 방법으로 구축된 데이터에 대한 특징을 인위적으로 변경하여 학습용 데이터의 양을 증대하는 기술

4. 디지털 파일 명명 규칙

● 4.1 파일구조

① 구축데이터의 원본 이미지는 아래의 폴더로 구성

가) db1_hw : 손글씨체

나) db1_pr : 인쇄체

다) db2 : Text-in-the-wild

② 각각의 폴더는 아래의 형태로 저장

가) Image의 경우, 다운로드가 용이하도록 10gb 단위로 분할 압축 예정

나) 압축형식 : zip, tar

```
- dataset_info.json
- images
        - *.png or   *.jpg
```

③ 이미지 파일에 대한 정보와 각 이미지 파일에 대한 annotation이 담겨있는 JSON 형식의 파일 구조 : 내부 구조의 경우, 각 데이터 별 "5. Labeling 정보 구조" 부분 참고

가) Image 파일 형식 : PNG (혹은 JPG)의 이미지 파일로 구성되며 해상도는 일정하지 않음

5. 손글씨 데이터 구축

● 5.1 구축 체계

구축단계	세부절차	필수여부	세부설명
수집	수집 대상 선정	필수	수집이 필요한 글자, 단어, 문장 정의 - 글자 : 현대식 한글 11,172자 - 음절(단어) : 국립국어원 한국어 학습용 어휘 6,000 낱말 활용 (약 한글 14,000자) - 문장 : AI Hub내, 기계독해 데이터 활용 (뉴스기사로 구성, 약 한글 30,000자)
	수집 Material 제작	필수	글자/단어/문장으로 구성된 각각의 Work sheet (작업 Material) 구성 후 출력
	데이터수집자 채용	필수	17세 이상~60세 이하, 남여 성비 1:1로 구성
	데이터 수집	필수	Work sheet 작성
정제	데이터 변환	필수	Work sheet의 Text를 Digital file형태로 변환
	데이터 정제	필수	Image file 형태 데이터 유효성 자체 검증
가공	추출 (Cropping)	필수	전체 Image 파일에서 한글 영역 추출
	가공 (Labeling 입력)	필수	추출된 개별 Image 파일의 한글 정보 입력
검수	전수 검사	필수	Image 와 입력된 Label 값의 일치 여부 확인
	품질 인증	필수	외부 자문단을 통한 품질 인증

① 수집

가) 수집 대상 선정

현대 한글의 모든 글자인 11,172자를 기본으로 데이터를 수집,
이후, 실생활 활용도가 높은 단어(혹은 어절), 문장 등을 선별하여 한글 데이터 수집 진행

1) 글자(음절) : 국립국어원에서 공개한 “현대 한글 11,172 글자” 활용
2) 단어(어절) : 국립국어원에서 공개한 “한국어 학습용 어휘 6,000 낱말” 활용
3) 문장 : 말뭉치 자료 활용 예정 (AI Hub내 AI데이터-기계독해 자료 활용)
4) 수집 양 : 전체 구축량 500만자 중, 손글씨 데이터 250만자 확보*

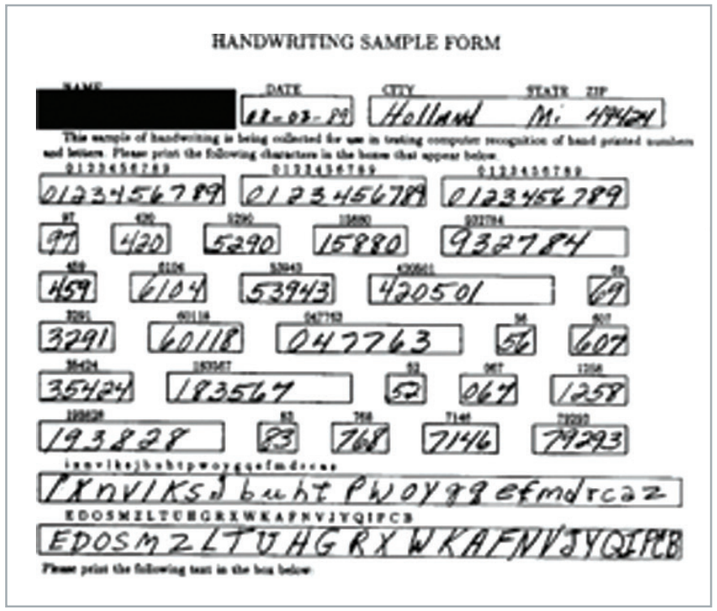
* 한국어 인식기술 관련 개발사 Interview 결과, 손글씨 데이터 양 확보 필요성에 대한 의견이 많았음 이에, 손글씨 데이터를 250만자로 확대하여 수집

나) 수집 Material 제작(Worksheet) : 손글씨 데이터 확보를 위해 글자, 단어, 문장이 포함된 Worksheet 제작

1) 지정된 글자(음절), 단어(어절), 문장으로 구성된 Worksheet*를 데이터 입력원(수집자)가 직접 손글씨로 작성하여 손글씨체 데이터를 구축

* 해당 Worksheet를 기반으로 인쇄체 구축 글자, 단어, 문장도 제작

[그림 5-1] 미 NIST에서 시행한 손글씨/단어별 데이터구축 Worksheet 예시



2) Worksheet 구성

구분	구성 글자 양	비고
글자 (음절)	11,172자	
단어 (어절)	최대 14,000자	필요 시, 추가 단어(어절 편성)
문장	약 30,000자	필요 시, 추가 문장 편성

다) 데이터 수집자 채용

1) 선정기준 : 다양한 손글씨를 확보하기 위해 성비, 나이대에 대한 기준을 수립하여 데이터 입력원 선발
2) 입력원 구성
- 다양한 성별/연령대별 고려하여 Worksheet 작성자 선발
- 약 200여명의 표본 집단을 선정하여 채용

라) 데이터 수집 : 데이터 수집자에게 수집 Material(Work Sheet) 배포 후, 작성 요청

1) 아래 형태로, 안내 문구를 아래칸에 그대로 작성하는 형태로 진행

2) 데이터 수집에 활용되는 필기구는 실생활에서 가장 많이 사용되는 볼펜으로 제한

[그림 5-2] 예시1 : 제시된 글자(음절)을 하단의 박스 안에 기입하도록 안내

가	각	갂	갃	간	갅	갆
갇	갈	갉	갊	갋	갌	갍
갎	갏	감	갑	값	갓	갔
강	갖	갗	갘	같	갚	갛
개	객	갞	갟	갠	갡	갢
갣	갤	갥	갦	갧	갨	갩
갪	갫	갬	갭	갮	갯	갰
갱	갲	갳	갴	갵	갶	갷
가	갹	갺	갻	갼	갽	갾

[그림 5-3] 예시1 : 제시된 단어(어절)을 하단의 박스 안에 기입하도록 안내

참다	크기	고기	남기다	서양
주요	냄새	여기다	공연	남녀
내놓다	떼다	속다	준비	구월
맑다	소년	소식	유월	작용
허리	놓치다	다기	독립	또다시
머릿속	쇠고기	위반	카드	평생
간부	관념	굉장히	단어	덮다
몰다	배우	비추다	신발	앞서다
자격	통제	계단	김치	낯설다

3) 수집 Guide (예시)

Work Sheet 작성 Guide

1. 활동 소개

본 활동는 한국정보화진흥원주관사업인, 인공지능 학습용 데이터 구축사업의 한국어 글자체 이미지 데이터 구축을 위한 데이터 수집 과제의 일환으로 수행되는 데이터 수집 활동입니다. 향후 연구용 데이터로 공개 예정입니다.

2. 작성 Guide

- 본 Work sheet는 글자 / 단어(어절) / 문장으로 구성되어 있습니다.
- 구성은 아래와 같습니다.

1) 글자 : 11,172자

2) 단어 : 약 14,000자 : 평소처럼 연속적으로 작성해도 됨

3) 문장 : 약 30,000자 : 평소처럼 연속적으로 작성해도 됨

- 각 페이지에 작성된 글을 아래 지정된 칸에 기입해주시면 됩니다.
- 오탈자가 발생할 경우, 화이트를 통해 수정 후 작업을 해주시기 바랍니다.

3. 제출 기한

- 본 과제 제출은 과제 발행일로부터 2주입니다.

4. 저작권 판매/양도 규정

- 작업 후, 본 데이터는 Superb AI에 양도/귀속됩니다.

② 정제

가) 데이터 변환 : 수집된 Work Sheet는 수집하여 대규모 네트워크 스캔 작업을 통해 디지털 데이터로 변환

1) 수집 Work Sheet 디지털화 시행

· 파일 형식 : jpg, jpeg, png 등

2) 디지털화된 이미지 데이터 분류 : 글자(음절), 단어(어절), 문장

[그림 5-4] 대규모 네트워크 스캔 시행 방법

나) 데이터 정제 : 고품질 Raw Data 확보를 위해 디지털화된 Work Sheet의 품질검사

– 검증 작업자를 통한 품질 검사 시행 : 디지털 파일 내, 글자 이미지가 계획대로 확보되었는지 육안으로 판단

1) 스캔품질 확인 : 흔들림(육안으로 식별 불가한 흔들림) 및 기타 노이즈(오탈자, 글자 외 오염물질, 구겨짐 등) 확인
2) 오탈자 검수 : 제시된 음절/어절/문장과 수집자의 작성분 일치 여부 확인

③ 가공

가공자 선발

1) 그래픽/디자인 전공 혹은 관련 경력 2년 이상의 후보자 모집
2) 입사 Test (QA이해도) 및 2주간 업무 Test(작업정확도) 시행
3) 작업정확도 기반 우수 작업자 선발 후 가공 작업 수행

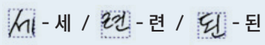
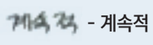
가) 추출 : 1차 가공 output인 글자별 데이터 확보를 위한 이미지 cropping

구분	수집 양	형태
글자 (음절)	11,172자	글자 단위로 11,172개 이미지 추출
단어 (어절)	최대 14,000자(약 6,000 단어)	단어 단위로 약 6,000개 이미지 추출
문장	최대 30,000자 (약 1,000 문장)	문장 단위로 약 1,000개 이미지 추출

나) 가공 : 각 이미지에 표시된 글자를 json 파일 형태로 입력 Annotation 정보 기입

1) Cropping된 이미지는 Annotation 도구에 1개씩 가공됨
2) 작업자가 이미지의 글자를 입력할 경우, 사전에 입력된 형식에 맞춰 Annotation 정보가 생성됨

[그림 5-5] 참고 : 가공 이미지

 - 세 / - 련 / - 된	 - 계속적
글자 단위의 가공	단어 단위의 가공

④ 검증

가) 전수검사 Process : 구축해야 할 Rawdata 및 Annotation 정보와 정확하게 일치하는지 확인

1) 시스템 셋팅 : 서버에 구축된 데이터 “Rawdata, Json파일” 업로드
2) 파일 불러오기 : 데이터 검수인원이 Paring 된 Rawdata, Json 파일 로드
3) 검수하기 : 데이터 검수인원은 아래 2가지 확인

· Rawdata가 유효한가 : 흔들리거나 깨지지 않았는가?
· Json 파일 정보가 정확한가 : 위의 지시문과 Annotation 정보가 일치한가

4) 저장하기 : 검수가 완료되어, 정확하게 작업된 데이터는 저장, 외 데이터는 폐기
5) A/S진행 : 구축목표량에 미달할 경우, 추가 작업 계획 수립 후 실행

나) 품질 인증* : 외부 자문단을 통한 데이터 품질 검증

* 첨부자료#3 품질인증 참조

다) 법률 이슈 검증

1) 지적재산권 법률 검토
· 손글씨의 경우, 데이터 입력원(수집자)을 통해 데이터 활용 및 공개에 대한 동의를 득하여 구입
· 저작권협회를 통해 확인한 결과, “개인 동의” 및 “구입”을 통해 확보한 데이터는 지적재산권 이슈가 없는 것으로 확인

● 5.2 Labeling 정보 구조

① Labeling Structure

JSON 형식

```
{
   "info": info,
   "images": [image],
   "annotations": [annotation],
   "licenses": [license]
}

info{
   "name": str, // 데이터 셋 이름
   "description": str, // 데이터 셋 설명 (optional)
   "url": str, // 데이터 셋 홈페이지 (optional)
   "date_created": datetime, // JSON 파일이 만들어진 날짜/시간
}

images{
   "id": str, // 이미지 아이디
   "width": int, // 이미지 width
   "height": int, // 이미지 height
   "file_name": str, // 이미지 파일 이름
   "license": str, // 라이센스 이름 (optional)
   "date_captured": datetime // 사진 찍힌  날짜/시간
(optional)
}

annotations{
   "id": str, // annotation 아이디
   "image_id": str, // 이미지 아이디
   "text": str, // 이미지에서 보이는 문장  내용
   "attributes": object // annotation 특이사항 (optional, 손글
씨의 경우 작성자의 성별, 연령대 등의 정보가 들어가며 인쇄체 데이
터의 경우 font의  종류와 크기 등 생성에 사용된 변수들이 기록됨)
}

licenses{
   "name": str, // 라이센스 이름 (license안에서 unique 해야 함)
```

// JSON 예시

```
{
   "info": {
    "name": “한글  OCR용 손글씨 데이터 셋”,
    "date_created": "2000-01-01T00:00:00+00:00"
  },
   "images": [
   {
      "id": "12345678-1234-5678-1234-567812345678",
      "width": 1920,
      "height": 1080,
      "file_name": "test.jpg",
      "license": "Attribution License",
      "date_captured": "2000-01-01T00:00:00+00:00"
   }
  ],
   "annotations": [
   {
      "id": "87654321-8765-4321-8765-432187654321",
      "image_id":  "12345678-1234-5678-1234-
567812345678",
      "text": "안녕"
   }
  ],
   "licenses": [
   {
      "name": "Attribution License",
      "url":  "http://creativecommons.org/licenses/
by/2.0/"
   }
  ]
}
```

Label Structure

프로젝트	한국어 글자체 이미지 AI데이터 구축	단계	설계
문서번호	EDK_AM_A_13 버전 1.2	작성일자	6/28/2019

인터페이스ID		Structure Name	구축DB1 Label Structure
Label 구조	JSON	개발유형	

NO	항목명		길이	타입	코드 여부	필수 여부	Name/Value	비고
	한글명	영문명						
	데이터셋정보	info						JSON Object
1	데이터셋명	name	100	String		Y		
2	데이터셋상세설명	description	1000	String				
3	데이터셋URL	url	200	String				
4	데이터셋생성일자	date_created	100	String		Y		
	이미지정보	images		List				List of JSON Object
5-1	이미지식별자	id	100	String		Y		
5-2	이미지너비	width	4	Number		Y		
5-3	이미지높이	height	4	Number		Y		
5-4	이미지파일명	file_name	100	String		Y		
5-5	이미지라이선스	license	100	String				
5-6	이미지촬영일자	date_captured	100	String				
	어노테이션정보	annotations		List				List of JSON Object
6-1	어노테이션식별자	id	100	String		Y		
6-2	연관이미지식별자	image_id	100	String		Y		
6-3	어노테이션텍스트	text	1000	String		Y		
	어노테이션속성	attributes		Object				JSON Object
6-4-1	종류	type	50	String				공통 : 글자(음절), 단어(어절), 문장
6-4-2	나이	age	3	String				손글씨 : 정보 수집자의 나이
6-4-3	성별	gender	1	String				손글씨 : 남, 여
6-4-4	직업	job	50	String				손글씨 : 정보 수집자의 직업
6-4-5	글꼴	font	50	String				인쇄체 : 수집된 글꼴
	라이선스	licenses		List				List of JSON Object
7-1	라이선스명	name	100	String		Y		
7-2	라이선스URL	url	200	String		Y		
참고사항								

② 고려사항

가) 모든 이미지의 한글 글자 정보를 하나의 파일(JSON 형식)에 저장하는 구조

나) "Text'에 들어가는 내용은 최소 한 글자에서 최대 한 줄의 문장

6. 인쇄체 데이터 구축

● 6.1 구축 체계

구축단계	세부절차	필수여부	세부설명
수집	수집 대상 선정	필수	수집이 필요한 글자, 단어, 문장 정의 - 글자 : 현대식 한글 11,172자 - 음절(단어) : 국립국어원 한국어 학습용 어휘 6,000 낱말 활용 (약 한글 14,000자) - 문장 : AI Hub내, 기계독해 데이터 활용 (뉴스기사로 구성, 약 한글 30,000자)
	수집 Material 정의	필수	한글/단어/문장으로 구성된 약 5.5만장의 Work sheet 를 하나의 구축 set로 지정 후, 다른 폰트로 구축가공
	Material 다양성 확보	선택	글자 폰트, 배경색, 노이즈 등 case 적용
	데이터 수집	필수	화면 캡쳐
정제	데이터 정제	필수	Image file 형태 데이터 유효성 자체 검증
가공	추출 (Cropping)	필수	전체 Image 파일에서 한글 영역 추출
	가공 (Labeling 입력)	필수	추출된 개별 Image 파일의 한글 정보 입력
검수	전수 검사	필수	Image 와 입력된 Label 값의 일치 여부 확인
	품질 인증	필수	외부 자문단을 통한 품질 인증

① 수집

가) 수집 대상 선정 : "손글씨 데이터 구축"의 1.1. 가 항과 동일

나) 수집 Material 정의 : "손글씨 데이터 구축" 1.1. 나 항과 동일 (글자수 250만자)

다) Material 다양성 확보

1) 다양한 폰트 확보*: 시장 수요 기반으로 활용 빈도수가 높은 폰트 선정

* 첨부자료#2 구축 폰트 참조

라) 데이터 수집 : 인쇄물 생성 프로그램을 통한 인쇄체 Material 생성

1) 인쇄물 자동 생성 프로그램 개발 : 손글씨 데이터 구축과 동일한 데이터를 다양한 변수를 반영하여 출력할 수 있는 시스템 개발

※ 글씨체마다 적용되어 있는 고유한 자간 값을 활용

2) 선정된 인쇄체 폰트를 활용하여 자동으로 글자, 단어, 문자 생성 진행

② 정제 : “손글씨 데이터 구축” 1.2 항과 동일

③ 가공 : “손글씨 데이터 구축” 1.3 항과 동일

④ 검수 : “손글씨 데이터 구축” 1.4 항과 동일

※ 데이터 증강 : 기본 구축량 외 기타 변수를 적용하여 증강 데이터 구축 (선택)

• 데이터 증강 변수 (랜덤 적용하여 3배수 구축)

	노이즈	선명도	왜곡	배경 변경	Flip	크기 변경	회전	글자 색상변경
내용	배경 얼룩	blur 10%	구겨짐	회색, 노랑색	상하, 좌우	12pt, 16pt	정각도, 좌우 20	검정, 빨강, 파랑, 흰색 등

• 변형 가이드 : AI모형 활용을 위한 변형 가이드 안내

— 노이즈 적용 : Gaussian Noise 및 Salt & Pepper 노이즈 적용
— 선명도 변경 : 3x3 Blur kernel 및 Gaussian blur kernel을 이용하여 blur 효과 적용
— 왜곡 : Radial distortion 및 Homography를 이용하여 이미지 왜곡
— 배경 변경 : 배경 영역을 추출하여 임의의 이미지로 대치 Flip: 상하 및 좌우
— Flip을 임의로 적용 크기변경 : Area resize method를 적용하여 2배 이내에서 크기 변경
— 회전 : 텍스트의 맨 왼쪽 지점을 기준으로 90도 이내에서 회전 적용
— 색상 변경 : 전체 이미지의 색상을 임의의 색상으로 변경 (이미지 내 동일한 color transform 적용)

● 6.2 Labeling 정보 구조 : 손글씨 데이터 구축 : 2항과 동일

7. Text-in-the-wild 데이터 구축

● 7.1 구축 체계

구축단계	세부절차	필수여부	세부설명
수집	수집 대상 선정	필수	수집 데이터 정의 - 도로교통표지판 / 상품 / 간판 / 도서
	수집 Guide-line 작성	필수	유효한 데이터 수집을 위한 가이드 제시
	수집 담당자/기관 선정	필수	수집전문 인력 채용 및 크라우드 소싱 업체
	데이터 수집	필수	이미지 데이터 촬영/수집
정제	데이터 정제	필수	촬영된 데이터의 유효성 자체 검증
가공	1차 가공 (영역 검출)	필수	이미지 데이터 내, 한글 글자 영역 표시
	2차 가공 (한글 입력)	필수	표시된 영역 내, 한글정보 입력
검수	전수 검사	필수	Image 와 입력된 Label 값의 일치 여부 확인
	품질 인증	필수	외부 자문단을 통한 품질 인증

① 수집

가) 수집 대상 : 한글인식기술 사용빈도가 높은 일상생활의 한글 데이터 확보를 원칙으로 함 (관련 기술 개발사 의견 수렴)

1) 도로교통표지판 : 자율주행 관련 산업 적용
2) 상품 : 이커머스 관련 산업 적용
3) 간판 : 자율주행 관련 산업 적용
4) 도서 : 출판 및 이커머스 관련 산업 적용
5) 수집 양 : 총 10만장

구분	수집 양	비고
도로교통표지판	동일한 비율을 목표로 구축 * 단, 수집 상황에 맞춰 수집양이 변동 될 수 있음	도로교통표지판 및 이정표 등
상품		상표, 특정 브랜드 상품 등
간판		상시 외부 노출되어 있는 자영업자 및 기업 간판
도서		책 외부 표지

나) 수집 Guide-line 작성 : 아래 내용을 수행할 데이터 수집자 모집 (예시)

1) 과제 소개 (수집목적 전달)

일상생활에서 발견할 수 있는 간판/광고판 내에 있는 한글, 영문, 숫자 사진을 촬영하여 업로드 해주시는 공모전입니다.

※ 본 공모전은 한국정보화진흥원 주관 사업인, 인공지능 학습용 데이터 구축사업의 한국어 글자체 이미지 데이터 구축을 위한 데이터 수집 공모전입니다. 향후 연구용 데이터로 공개 예정입니다.

2) 촬영 가이드

- 한글이 사진의 10~80%를 차지해야 함
- 촬영된 이미지 내 문자는 최대 7자를 넘지 않아야 함(작은 글자는 촬영되면 안 됨)
- 가로사진 O / 세로사진 X
- 권장 해상도 : FHD 이상 (1920 X 1080)
- 파일 형식 : JPG / JPEG
- 올려보거나 내려보며 촬영한 사진이 아닌 정면 시선으로 촬영한 사진이어야 함 (가능 시선 각도 상하 15도)
- 피사체가 명확해야 함 (흔들린 사진 X / 노이즈가 심한 사진 X)

※ 목표 피사체 외의 한글, 영문, 숫자 등이 최대한 나오지 않게, 목표 피사체 앵글을 크게 촬영하여 주시기 바랍니다.

3) 저작권 판매/양도 규정

- 공모 요강에 맞는 사진은 실시간으로 판매되며, 판매 완료된 사진의 저작권(2차적 저작물 작성권 포함)은 Superb AI에 양도/귀속됩니다.

다) 수집 담당자/기관 선정 :

1) 전문 사진 촬영 기사 채용 : 전문적인 사진 촬영을 위한 프리랜서 작가 고용
2) 크라우드소싱 업체 선정 : 다양한 환경의 이미지 수집을 위한 크라우드소싱 방법 선택

라) 데이터 수집 : 지정된 수집 Guide에 맞춰 데이터 수집

② 정제 : "손글씨 데이터 구축" 1.2 항과 동일

③ 가공 : Image 당 최대 15자로 제한

가) 1차 가공 (영역 추출) : 1차 가공 output인 이미지 내 Text 영역 검출

1) Text가 있는 위치의 좌표값을 Bounding Box 형태로 좌측 상단의 좌표 (x1,y1)값과 우측 하단의 좌표 (x2,y2)의 값을 json 파일 형태로 기록

[그림 7-1] 예시 : 이미지 내 한글 글자 영역 표시

2) 2차 가공(한글정보 입력) : bounding box가 표시된 영역의 글자를 받아쓰기(Transcription) 하여 기입

[그림 7-2] 예시 : box로 표시된 영역의 Text를 입력

찾을 수 있는 항목

1. PROFESSIONAL
2. FOR
3. HAND
4. DRIP
5. COFFEE
6. DELPINO
7. GOOD
8. CAFE
9. ROASTERY
10. 핸드드립
11. 로스터리
12. 카페
13. 델피노

④ 검수

가) 전수 검사 : 구축해야 할 Rawdata 및 Annotation 정보와 일치하는지 확인

나) 품질 인증* : 외부 자문단을 통한 데이터 품질 검증

* 첨부파일 3 참조

● 7.2 Labeling 정보 구조

① Labeling Structure

```
// JSON 형식
{
   "info": info,
   "images": [image],
   "annotations": [annotation],
   "licenses": [license]
}

info{
   "name": str, // 데이터 셋 이름
   "description": str, // 데이터 셋 설명 (optional)
   "url": str, // 데이터 셋 홈페이지 (optional)
   "date_created": datetime, // JSON 파일이 만들어진 날짜/시간
}

images{
   "id": str, // 이미지 아이디
   "width": int, // 이미지 width
   "height": int, // 이미지 height
   "file_name": str, // 이미지 파일 이름
   "license": str, // 라이센스 이름 (optional)
 "date_captured":   datetime // 사진 찍힌 날짜/시간 (optional)
}

annotations{
   "id": str, // annotation 아이디
   "image_id": str, // 이미지 아이디
   "bbox": [x, y, width, height], // text의 위치 정보
   "text": str, // 이미지에서 보이는 문장  내용
   "attributes": object // annotation 특이사항
(optional, 글자가 기울어지거나 뒤집어졌는지, 난이도 등이 기입될 수 있음)
}

licenses{
   "name": str, // 라이센스 이름 (license안에서 unique 해야 함)
   "url": str // 라이센스 내용이 들어  있는 링크
}
```

```
// JSON 예시
{
   "info": {
    "name": "한글 OCR용 text-in-the-wild 데이터 셋",
    "date_created": "2000-01-01T00:00:00+00:00"
 },
   "images": [
   {
      "id": "12345678-1234-5678-1234-567812345678",
      "width": 1920,
      "height": 1080,
      "file_name": "test.jpg",
      "license": "Attribution License",
      "date_captured": "2000-01-01T00:00:00+00:00"
   }
 ],
   "annotations": [
   {
      "id": "87654321-8765-4321-8765-432187654321",
      "image_id":   "12345678-1234-5678-1234-
567812345678",
      "bbox": [5, 7, 20, 10],
      "text": "안녕"
   }
 ],
   "licenses": [
   {
      "name": "Attribution License",
      "url":   "http://creativecommons.org/licenses/by/2.0/"
   }
 ]
}
```

Label Structure

프로젝트	한국어 글자체 이미지 AI데이터 구축	단계	설계
문서번호	EDK_AM_A_13 버전 1.2	작성일자	6/28/2019

인터페이스ID		Structure Name	구축DB2 Label Structure
Label 구조	JSON	개발유형	

NO	항목명		길이	타입	코드 여부	필수 여부	Name/Value	비고
	한글명	영문명						
	데이터셋정보	info		Object				JSON Object
1	데이터셋명	name	100	String		Y		
2	데이터셋상세설명	description	1000	String				
3	데이터셋URL	url	200	String				
4	데이터셋생성일자	date_created	100	String		Y		
	이미지정보	images		List				List of JSON Object
5-1	이미지식별자	id	100	String		Y		
5-2	이미지너비	width	4	Number		Y		
5-3	이미지높이	height	4	Number		Y		
5-4	이미지파일명	file_name	100	String		Y		
5-5	이미지라이센스	license	100	String				
5-6	이미지촬영일자	date_captured	100	String				
	어노테이션정보	annotations		List				List of JSON Object
6-1	어노테이션식별자	id	100	String		Y		
6-2	연관이미지식별자	image_id	100	String		Y		
6-3	어노테이션텍스트	text	1000	String		Y		
6-4	어노테이션속성	attributes		Object				
6-5	어노테이션바운딩박스	bbox	4	List		Y		List of [x, y, width, height]
	라이센스	licenses		List				List of JSON Object
7-1	라이센스명	name	100	String		Y		
7-2	라이센스URL	url	200	String		Y		
참고사항								

② 고려사항

가) 모든 이미지의 한글 글자 정보를 하나의 파일(JSON 형식)에 저장하는 구조

나) bbox(bounding box)는 해당 이미지를 벗어나지 않음

8. 구축 데이터셋 정보

● 8.1 구축 데이터셋 정보

구분	글자수	데이터 출처	비고
손글씨	2,500,000	글자(음절) : 국립국어원 “현대 한글 11,172자” 단어(어절) : 국립국어원 “한국어 학습용 어휘 6,000낱말”, 약 14,000자 문장 : AI Hub 내 기계독해 데이터, 약 30,000자	-
인쇄체	2,500,000		증강변수를 적용후, 추가 구축
Text-in-the wild	이미지 10만장	서울 시내 간판, 도서, 상품	-

● 8.2 구축목표 수량

구분	구축 내용			운영 인력 (명)	운영 기간	비고
	분류	소분류	구축량			
최종 500만자 + 약 50만자 (이미지10만장)	손글씨 250만자	글자	60만자	50	2019.07.01 ~ 2019.11.30	주간 단위 수집량 약 5,000자 수집과 병렬로 정제, 가공 진행
		단어	80만자	50 (글자 40면 포함)	2019.07.01 ~ 2019.11.30	주간 단위 수집량 약 10,000자 수집과 병렬로 정제, 가공 진행
		문장	110만자	50 (단어 70명 포함)	2019.07.01 ~ 2019.11.30	주간 단위 수집량 약 10,000자 수집과 병렬로 정제, 가공 진행
	인쇄체 250만자	글자	60만자	3	2019.07.01 ~ 2019.11.30	프로그래밍을 통한 작업 수행 후 출력/스캔&가공 진행
		단어	80만자	3	2019.07.01 ~ 2019.11.30	
		문장	110만자	3	2019.07.01 ~ 2019.11.30	
	Text in the Wild (약 50만자)	이미지	10만장	10	2019.06.17 ~ 2019.11.30	크라우드 소싱 병행, 동일한 비율을 목표로 구축하되, 수집 상황에 맞춰 수집양이 변동 될 수 있음

8.3 Ground Truth 정의

① 손글씨 / 인쇄체 / Text-in-the-wild의 Ground Truths 정의

가) Multi Voting 후, 응답자 80%가 같은 판단을 한 경우, Ground Truths로 판정

② 시행방법

가) 제시된 Task와, 데이터 수집자가 만든 이미지 준비 (수집자가 작성한 글자)

나) 5명 이상의 인원이 해당 데이터를 보고 바르게 작성했는지 여부에 대한 true / false를 판단

다) 응답자 중 80%이상이 ture를 응답할 경우, Ground Truths로 판정

부속서A

구축 폰트

A.1 인쇄체 폰트 선정 기준

- 시장 수요조사 : 설문 진행 (총 272명 참여)
- 대한민국 정부 블로그 : 기업 및 공공기관에서 사용하는 무료 글꼴 (링크)
- 실증기관 interview 조사

A.2 폰트 선정 목록

- Survey 결과, 2회 이상 언급된 폰트 46종 (특정 기관 내에서만 사용되는 폰트 제외) 공공기관에서 사용되는 폰트 11종
- 실증기관의견 반영 3종 선정
- 이 중, 저작권 이슈가 발생가능한 폰트 및 현대한글 11,172자를 구현하지 못하는 글자는 제외

번호	폰트명	소유사	선정근거		
			설문	공공기관	실증기관
1	바탕	-	●		
2	돋움	-	●		
3	굴림	-	●		
4	맑은고딕	-	●		
5	궁서	-	●		
6	휴먼명조	휴먼컴퓨터	●		
7	휴먼고딕	휴먼컴퓨터	●		
8	중고딕	-	●		
9	나눔고딕	네이버	●		
10	나눔바른고딕	네이버	●		
11	나눔고딕 코딩	네이버	●		
12	나눔명조	네이버	●		
13	나눔손글씨펜	네이버	●		
14	나눔손글씨붓	네이버	●		
15	노토산스	구글&어도비	●		
16	본고딕	구글&어도비	●		
17	노토세리프	구글&어도비	●		
18	대한민국정부 상징	문체부	●	●	
19	호국	국방부	●	●	
20	서울남산	서울시청	●	●	

번호	폰트명	소유사	선정근거		
			설문	공공기관	실증기관
21	서울한강	서울시청	●	●	
22	제주 한라산	제주시		●	
23	제주 명조	제주시		●	
24	제주 고딕	제주시	●	●	
25	부산	부산시	●	●	
26	고양	고양시	●	●	
27	전북	전북시		●	
28	푸른전남	전라남도		●	
29	아리따부리	아모레퍼시픽	●		
30	SKT 뫼비우스	SKT	●		
31	빙그레따옴	빙그레	●		
32	티몬소리	티몬	●		
33	미생	다음	●		
34	스포카한산스	스포카	●		
35	함초롱돋움	한글과컴퓨터	●		
36	함초롱바탕	한글과컴퓨터	●		
37	농협희망	농협	●		●
38	하나	하나금융그룹	●		●
39	조선명조	조선일보	●		●
40	한겨레	한겨레	●		
41	대한	윤디자인	●		
42	가는안상수체	-	●		
43	수화명조	정수화폰트연구소	●		
44	태룡고딕	정수화폰트연구소	●		
45	EBS훈민정음새론	EBS	●		
46	EBS훈민정음	EBS	●		
47	KBIZ 한마음고딕	중소기업중앙회	●		
48	KBIZ 한마음명조	중소기업중앙회	●		
49	만화진흥원	(재)한국만화영상진흥원	●		
50	이롭게바탕	이롭게	●		

※ 폰트의 "비영리 사용여부" 기준 지속 확인 및 업데이트 예정. 저작권 이슈가 발생할 경우 해당 폰트는 삭제 혹은 교체 예정

A.3 폰트 선정 근거

- 폰트관련 전문기관 Survey를 통해 선정
- 응답자 분포

구분	공공기관	기업	교육기관	기타
분포	16.54%	58.82%	8.82%	15.81%

- 설문 결과 요약

구분	맑은고딕	나눔고딕	돋움체
분포	37.04%	9.63%	3.70%

A.4 수요조사 및 실증기관 Survey를 통해 추가 선정

- 폰트다운로드 기관 참고 (네이버/다음)
- 한국전자문서협회 114개 회원사를 대상으로 2019년 06월 10일부터 2019년 06월 14일까지 온라인 설문으로 총 7개 항목의 설문으로 진행
  (실증기관 또는 실증기관의 고객사에서 주로 사용하는 한국어 글자체는 맑은체가 52%, 나눔체 16%, 굴림체 12%, 중고딕 4%, 돋움체 4%, 기타로는 기본 윈도우 폰트 외에 기관 고유의 폰트 사용이 4%로 응답)

A.5 고려 사항

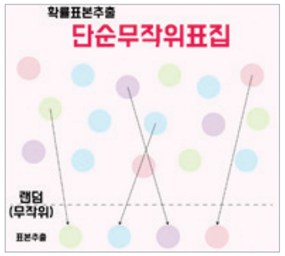
- 폰트 관련 지적재산권 해결 방안 고려 : 비영리 목적의 배포 가능여부 확인
- 다양한 글자 자간에 대한 데이터 분류 제시 : 폰트 고유의 자간을 활용할 예정
- Material다양성 확보를 위해 다양한 폰트 확보, 구겨짐, 오염환경 이외에 글자 크기, 회전, Random Crop, 커브, Low Contrast 등의 효과 추가
  (https://github.com/Sanster/text_renderer 참조)
- 저작권 관련 확인 결과, 민간/공공기관에서 활용되는 폰트의 경우, "비영리" 목적에 한해 대외공개 관련 지적재산권 이슈가 없는 것으로 파악됨
  (현재 추가 확인 중이며, 문제가 예상될 경우 유료 구매 및 폰트 대체 작업 진행)

부속서B

품질검사 (외부자문단을 통한 품질 검사)

● B.1 무작위 표본 추출 (simple Random Sampling)

본 검증 방법은 통계학에서 사용하는, 모집단(population)의 각각의 요소 또는 사례들이 표본(sample)으로 선택될 가능성이 같게 되는 표본 추출법임. 유한모집단에서 n개의 추출단위로 구성된 모든 부분 집합이 표본으로 선택될 확률이 같도록 설계된 표본추출방법을 뜻함



● B.2 외부 자문단을 통한 데이터 품질 검사 계획

가) 슈퍼브에이아이 구축 데이터셋에서 sample 5,000개를 추출 (전체 구축량의 0.1%수준)

나) 복수의 자문단(추후 확정)이 해당 데이터셋의 정확도를 검증한 후, 의견을 제출

다) Multivoting을 통해 제출된 의견이 “정확하게 구축되었음” 로 일치할 경우, 목표 수준에 충족된 데이터셋으로 판단

● B.3 검사 기준은 아래와 같음 (평가기준 : 일치율 95%이상)

가) 손글씨/인쇄체 :

1) 제시된 음절/어절/문장의 구축 수량이 일치하는가
2) 각각의 이미지는 흔들림 및 노이즈에 대한 문제가 없는가
3) 각각의 한글은 오탈자 없이 기입되었는가

나) Text-in-the-wild

1) 제시된 구축 수량과 일치하는가
2) 각각의 이미지는 흔들림 및 노이즈에 대한 문제가 없는가
3) 각각의 한글은 오탈자 없이 기입되었는가
4)이미지 내 한글이 정확하게 boundingbox로 표시되었는가?
5) 표시된 box 안의 한글이 정확하게 기입되었는가?

다) 추후 추가 문항 개발 필요성 확인 및 추가

1) 자문단 의견 반영 및 타 데이터셋 구축 Guide 참고

부속서C

Label Structure

● C.1 손글씨 및 인쇄체 데이터 Structure

Label Structure

프로젝트	한국어 글자체 이미지 AI데이터 구축	단계	설계
문서번호	EDK_AM_A_13 버전 1.2	작성일자	6/28/2019

인터페이스ID		Structure Name	구축DB1 Label Structure
Label 구조	JSON	개발유형	

NO	항목명 한글명	항목명 영문명	길이	타입	코드 여부	필수 여부	Name/Value	비고
	데이터셋정보	info						JSON Object
1	데이터셋명	name	100	String		Y		
2	데이터셋상세설명	description	1000	String				
3	데이터셋URL	url	200	String				
4	데이터셋생성일자	date_created	100	String		Y		
	이미지정보	images		List				List of JSON Object
5-1	이미지식별자	id	100	String		Y		
5-2	이미지너비	width	4	Number		Y		
5-3	이미지높이	height	4	Number		Y		
5-4	이미지파일명	file_name	100	String		Y		
5-5	이미지라이선스	license	100	String				
5-6	이미지촬영일자	date_captured	100	String				
	어노테이션정보	annotations		List				List of JSON Object
6-1	어노테이션식별자	id	100	String		Y		
6-2	연관이미지식별자	image_id	100	String		Y		
6-3	어노테이션텍스트	text	1000	String		Y		
	어노테이션속성	attributes		Object				JSON Object
6-4-1	종류	type	50	String				공통 : 글자유형, 단어(어절) 등장
6-4-2	나이	age	3	String				손글씨 : 정보 수집자의 나이
6-4-3	성별	gender	1	String				손글씨 : 남 여
6-4-4	직업	job	50	String				손글씨 : 정보 수집자의 직업
6-4-5	글꼴	font	50	String				인쇄체 : 수집된 글꼴
	라이선스	licenses		List				List of JSON Object
7-1	라이선스명	name	100	String		Y		
7-2	라이선스URL	url	200	String		Y		
참고사항								

● C.2 Text-in-the-wild 데이터 Structure

Label Structure

프로젝트	한국어 글자체 이미지 AI데이터 구축	단계	설계
문서번호	EDK_AM_A_13 버전 1.2	작성일자	6/28/2019

인터페이스ID		Structure Name	구축DB2 Label Structure
Label 구조	JSON	개발유형	

NO	항목명 한글명	항목명 영문명	길이	타입	코드 여부	필수 여부	Name/Value	비고
	데이터셋정보	info		Object				JSON Object
1	데이터셋명	name	100	String		Y		
2	데이터셋상세설명	description	1000	String				
3	데이터셋URL	url	200	String				
4	데이터셋생성일자	date_created	100	String		Y		
	이미지정보	images		List				List of JSON Object
5-1	이미지식별자	id	100	String		Y		
5-2	이미지너비	width	4	Number		Y		
5-3	이미지높이	height	4	Number		Y		
5-4	이미지파일명	file_name	100	String		Y		
5-5	이미지라이선스	license	100	String				
5-6	이미지촬영일자	date_captured	100	String				
	어노테이션정보	annotations		List				List of JSON Object
6-1	어노테이션식별자	id	100	String		Y		
6-2	연관이미지식별자	image_id	100	String		Y		
6-3	어노테이션텍스트	text	1000	String		Y		
6-4	어노테이션속성	attributes		Object				
6-5	어노테이션바운딩박스	bbox	4	List		Y		List of [x, y, width, height]
	라이선스	licenses		List				List of JSON Object
7-1	라이선스명	name	100	String		Y		
7-2	라이선스URL	url	200	String		Y		
참고사항								