●○ 영어 번역 말뭉치 과제

한국어-영어 번역 말뭉치(사회과학)

●○ 개요: 사회과학 한영 말뭉치란?

사회과학 분야의 한국어 문장을 영어 문장으로 번역하는 AI 기반 번역기 기술 개발에 활용할 수 있는 학습 데이터셋으로 ㈜트위그팜에서 구축했으며, 총 150만 건의 한국어 영어 문장이 쌍으로 구성되어 있다.

구축된 말뭉치는 사회과학에 특화된 AI 모델의 학습을 위한 목적을 가지고 있으며, 사회과학 분야의 전문 번역기, 사회과학 분야의 용어사전 개발 등에 활용될 수 있다.

사회과학 한영 말뭉치로 학습된 번역기를 사용한 사례는 아래를 참고할 수 있다.

표1 | 번역기를 활용한 번역의 사례

구분	원문(한국어)	번역문(영어)
1	언어교육의 측면에서 볼 때, 문학은 문화교육을 통한 읽기와 쓰기 교육에 적합한 장르로 볼 수 있다.	From the perspective of language education, literature, through cultural education, is a genre suitable for reading and writing education.
2	단편 소설 〈메밀꽃 필 무렵〉을 읽는 과정에서 이해하는 데 어려움을 겪었던 어휘 또는 표현을 정리한다.	While reading the short story entitled "When Buckwheat Flowers Bloom," list down the words or expressions that were difficult to understand.
3	수많은 데이터들은 마치 감성을 가진 것처럼 자유자재로 움직이면서 인간의 움직임을 분석하고 있다.	Numerous data analyze human movements, allowing free movement as if they had emotions.

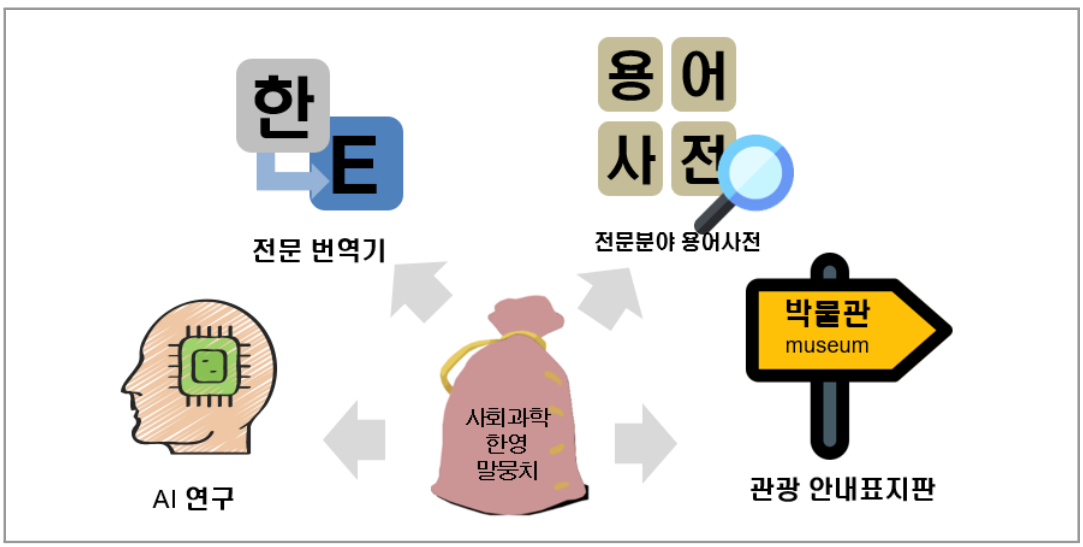

그림1 | 사회과학 한영 말뭉치 활용 사례

●○ 데이터셋의 구성

본 데이터셋은 사회과학 분류에 속한 법률, 교육, 경제, 문화, 관광, 예술의 6개 대분야와 그 아래의 22개 소분야로 구성되어 있다. 해당 분야들은 번역 수요가 높고 전문 지식과 고급 번역기술을 필요로 하여 높은 활용도를 가진다.

각 분야별로 평균 10만 건의 고품질 데이터 셋을 보유하고 있어 해당 분야 번역 연구에 활용하거나 각 분야별 상용 번역기를 학습시키기에 충분한 양으로 볼 수 있다.

그림2 | 데이터셋 구성도(6개 대분야, 22개 소분야)

●○ 데이터 포맷

본 데이터는 AI 번역 모델의 학습을 주목적으로 하며 다양한 분야에 활용될 것을 가정한다. 이에 여러 분야에서 활용되기 위해 포맷 변환이 용이하며, 데이터의 사이즈가 작아서 웹 통신에서 활용성이 높은 JSON 포맷을 사용한다. 그리고 다양한 플랫폼에서 데이터 상호 호환이 가능하도록 unicode로 인코딩 하였다. 하나의 데이터는 원문(한국어 문장), 번역문(영어 문장), 출처 등의 정보를 포함한다.


```
{
    "properties": {
        "data": {
            "properties": {
                "properties": {
                    "type": {
                        "type": "string"
                    },
                    "ko": {
                        "type": "string"
                    },
                    "en": {
                        "type": "string"
                    },
                    "domain": {
                        "type": "string"
                    },
                    "license": {
                        "type": "string"
                    },
                    "style": {
                        "type": "string"
                    }
                },
                "type": "object"
            },
            "type": "array"
        }
    },
    "type": "object"
}
```


그림3 | JSON 포맷의 스키마

●○ 데이터 구조

데이터셋에 따른 항목과 해당 값은 아래 테이블과 같다.

한국어 문장과 영어문장은 각각 'ko', 'en' properties 에 포함되고, 나머지 properties 는 데이터의 label 역할을 한다. 문체와 법률개정정보 properties 는 해당되는 경우만 선택적으로 기재된다.

표2 | 사회과학 한영 말뭉치 데이터 구조표

항목	property	필수 여부	설명	예시
한국어 문장	ko	Y	한국어문장 포함	"이러한 두 목적 간의 상충관계를.."
영어문장	en	Y	한국어문장 포함	"It is required to analyze the.."
분야 (도메인)	source	Y	법률, 교육 등 6개 대분야 중 택일	"법률"
출처	domain	Y	한국학술정보, 비교형사법연구, 조세재정연구원 등 원문 출처 기재	"한국빅데이터학회"
라이선스	license	Y	원문의 라이선스, 명시적 라이선스 또는 상용 사용 여부 명시	"open"
문체	style	N	문어체와 구어체를 구분하여 원문의 문제 명시	"문어체"
법률개정 정보	law_history	N	법령 데이터 경우 개정 날짜 기재	"2018-10-05"

●○ 데이터 예시

이 데이터는 사회 분야의 말뭉치 데이터 예시이다.


```
{
    "data":
        [
            {
                "ko": "중학교 2학년 시절의 여가만족도가 상대적으로 가장 큰 영향력을 미치는 것으로 나타났다.",
                "en": "The level of leisure satisfaction in the second year of middle school was found to have relatively the greatest influence.",
                "source": "경기대학교 관광종합연구소",
```

```
                "domain": "관광",
                "license": "open",
                "style": "문어체"
            },
            {
                "ko": "직무만족은 조직몰입에 긍정적인 영향을 미치는 것으로 조직성과로 나타날 수 있음을 보여준다.",
                "en": "Job satisfaction shows that organizational performance can be shown as having a positive effect on organizational immersion.",
                "source": "경기대학교 관광종합연구소",
                "domain": "관광",
                "license": "open",
                "style": "문어체"
            },
            {
                "ko": "관광객들이 급증함으로 인해 고창학원관광농원이 훼손되고,생태적 회복 능력을 완전히 상실하게 될 위기에 처해있는 상황이다.",
                "en": "The surge in tourists is threatening to damage the Gochang Academy's tourism farm and completely lose its ecological resilience.",
                "source": "경기대학교 관광종합연구소",
                "domain": "관광",
                "license": "open",
                "style": "문어체"
            },
            {
                "ko": "한국 노동패널 조사에서 연구목적에 알맞은 지표로서 조직몰입도 직무만족도 요인별 직무만족도 일 적합도 이직의도를 이용하였다.",
                "en": "In the Korean Labor Panel Survey, the organization's immersion, job satisfaction, and job satisfaction by factors were used as indicators suitable for research purposes.",
                "source": "경기대학교 관광종합연구소",
                "domain": "관광",
                "license": "open",
                "style": "문어체"
            }
        ]
}
```


그림4 | 데이터 예시

●○ 데이터 구축 과정

데이터 올리기, 말뭉치 정제, 기계 번역, 전문 번역가의 재번역, 번역 품질 평가, 재번역, 기술 검수, 품질검증, 말뭉치 취합 및 공급으로 이루어진 프로세스를 통하여 300만 건의 말뭉치 데이터를 구축하였다. 이를 위해 데이터 수집 담당, 데이터 정제 담당, 데이터 가공(번역) 담당, 데이터 감수(전문가 리뷰) 담당, 데이터 품질 담당의 역할을 구분하였다.

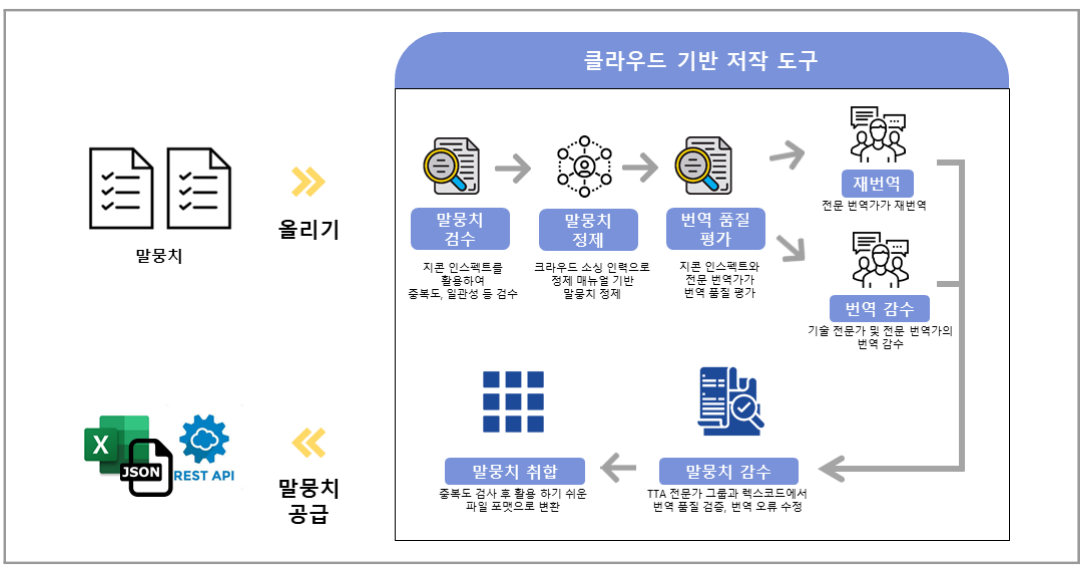

그림5 | 데이터 구축 프로세스

대량의 데이터를 단시간에 효율적으로 구축하기 위해 크라우드 소싱을 도입하였고, 이를 위해 웹 기반의 말뭉치 데이터 저작도구를 자체 개발하여 활용하였다. 데이터 정제 담당, 데이터 가공 담당, 데이터 감수 담당, 데이터 품질 담당은 저작도구를 통해 프로젝트 관리자가 할당한 업무를 언제, 어디서나 처리할 수 있다.

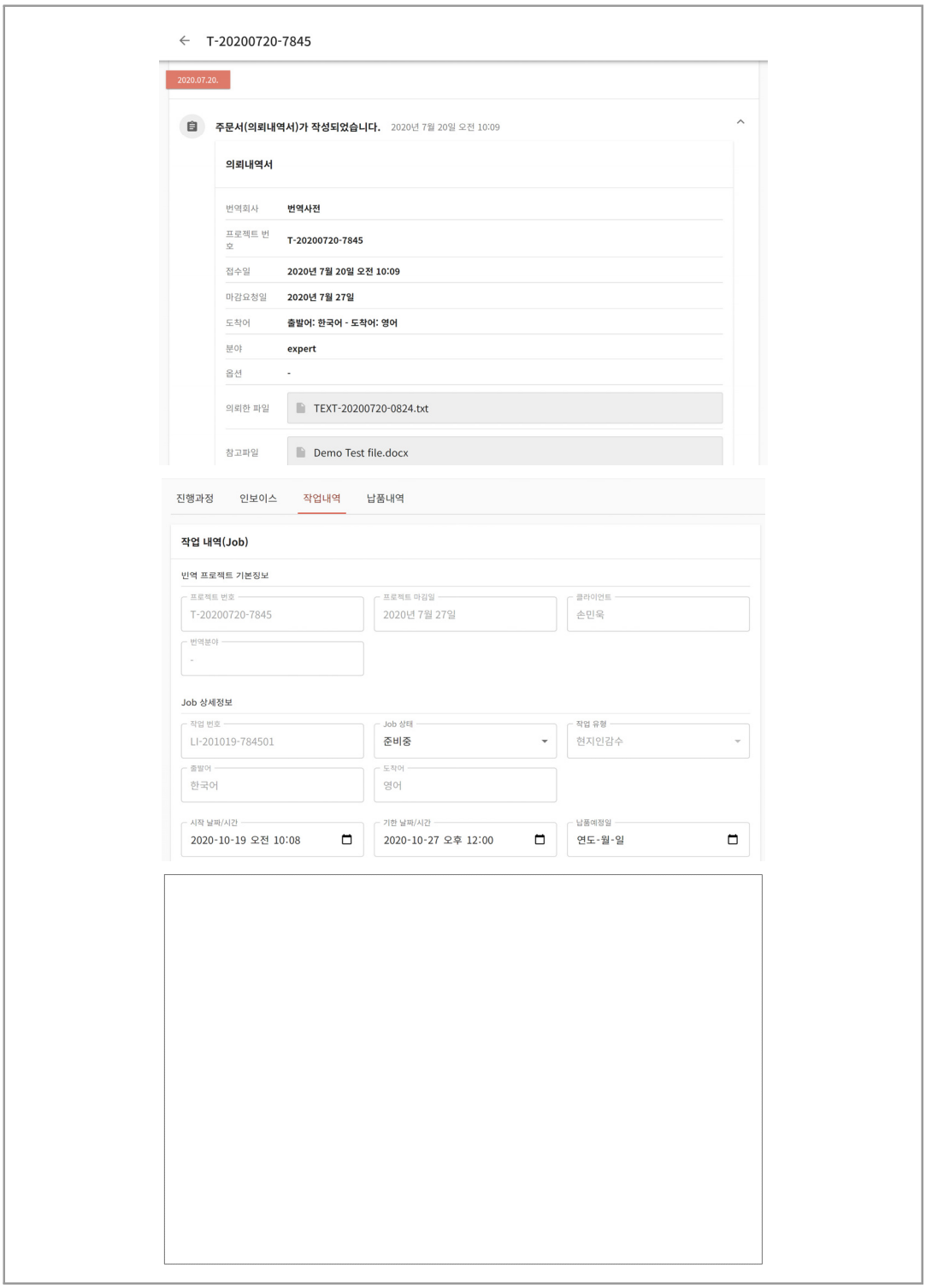

그림6 | 웹 기반 저작도구 화면

●○ 검수와 품질 확보

150만 건의 고품질의 한영 말뭉치를 구축하기 위해 각 데이터 구축 과정의 단계마다 품질기준을 정의하고, 품질 검수 후 기준 통과 시만 다음 단계로 진행이 가능하도록 하였다. 이를 위해 각 단계마다 품질 평가 담당자를 배치하여 품질 관리를 하였고, 품질 기준 미달 데이터의 경우 품질 총괄부서에서 해당 데이터의 사후처리 방법을 결정하였다.

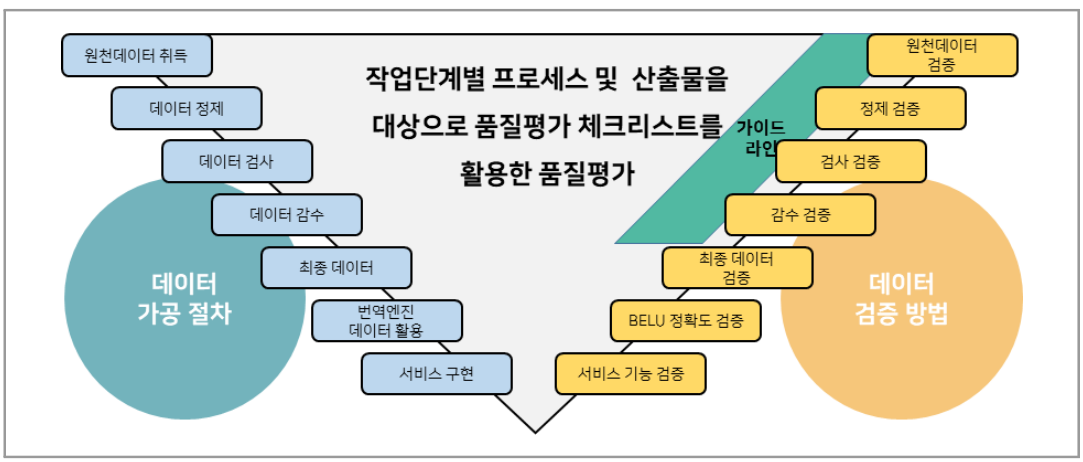

그림7 | 품질 확보를 위한 품질 검수 체계

●○ 데이터 구축 담당자

수행기관(주관) : ㈜트위그팜
(전화: +82-02-1833-5926, 이메일: sunho.baek@twigfarm.net)