●○ 영어 번역 말뭉치 과제

한국어-영어 번역 말뭉치(기술과학)

●○ 개요: 기술과학 한영 말뭉치란?

기술과학 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 기술과학에 특화된 AI 모델의 학습을 위한 목적을 가지고 있으며, 기술과학 분야의 전문 번역기, 기술과학 분야의 용어사전 개발 등에 활용될 수 있다.

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

표1 | 번역기를 활용한 번역의 사례

구분	원문(한국어)	번역문(영어)
1	중계국을 위한 참조 신호를 효율적으로 전송하기 위한 방법이 필요하다.	There is a need for a method of efficiently transmitting reference signals for a relay station.
2	전기자동차와 자율운행자동차의 증가는 지능형 교통 인프라와 빅데이터 분석을 더욱 필요로 할 것이다.	The increase of electric and autonomous vehicles will further require intelligent transportation infrastructure and big data analysis.
3	알고리즘 설계 및 구현의 단순성을 위하여 작업 아이디와 작업 시간대는 모두 숫자로 변환하였다.	For simplicity in algorithm design and implementation, both the job ID and the working time zones were converted to figures.

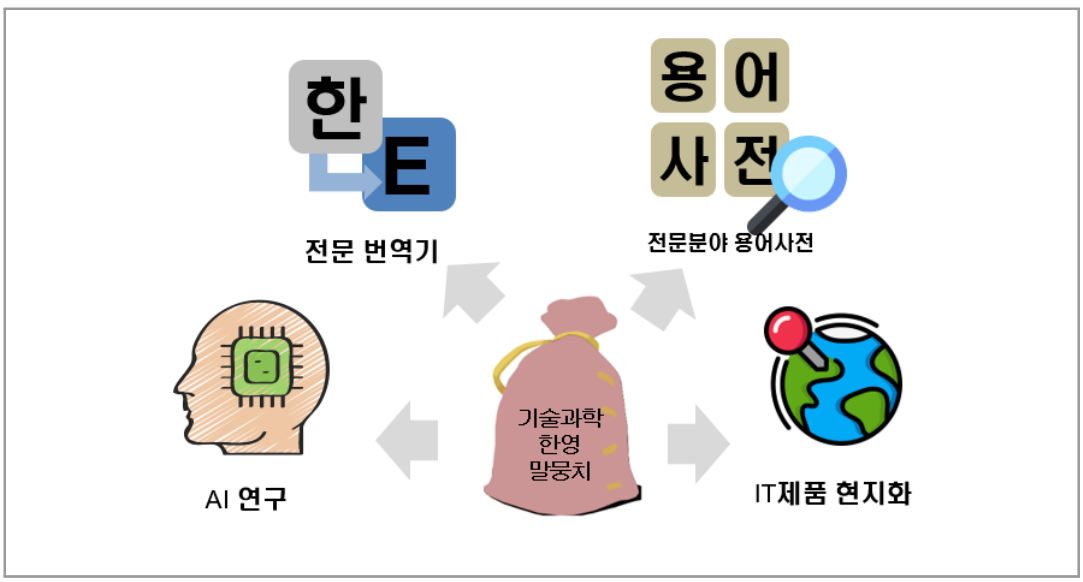

그림1 | 기술과학 한영 말뭉치 활용 사례

데이터셋의 구성

본 데이터셋은 기술과학 분류에 속한 ICT, 전기, 전자, 기계, 의학의 5개 대분야와 그 아래의 18개 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(5개 대분야, 18개 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 | 기술과학 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	ICT, 전기 등 5개 대분야 중 택일	"공학"
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 공학 분야의 말뭉치 데이터 예시이다.


```
{
    "data":
        [
            {
                "ko": "이러한 두 목적 간의 상충관계를 분석하여 균형 있는 최적의 스케줄을 생성하는 것이 요구된다.",
                "en": "It is required to analyze the trade-off between these two objectives to create a balanced optimal schedule.",
```

```
                "domain": "공학",
                "license": "open",
                "style": "문어체"
            },
            {
                "ko": "자원의 용량이 중요한 요소이므로 이를 제약조건으로 하는 스케줄을 생성하는 것은 타당하다.",
                "en": "Since resource capacity is an important factor, it is reasonable to create a schedule with this constraint.",
                "source": "국회도서관",
                "domain": "법률",
                "license": "open",
                "style": "문어체",
                "law_history": ["2015-10-05"]
            },
            {
                "ko": "우선순위 규칙에 기반한 방법들은 최적의 스케줄을 생성하지 못하는 단점을 가지고 있다.",
                "en": "The methods based on priority rules have a disadvantage in that they do not generate an optimal schedule.",
                "source": "한국빅데이터학회",
                "domain": "공학",
                "license": "open",
                "style": "문어체"
            },
            {
                "ko": "각 우선 순위 규칙들에 대한 자세한 정보가 제공되지 못하여 자원 평준화 기능의 사용자에게 혼동을 주고 있다.",
                "en": ""It is confusing to users of the resource leveling function because detailed information on each priority rule is not provided.",
                "source": "한국빅데이터학회",
                "domain": "공학",
                "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.

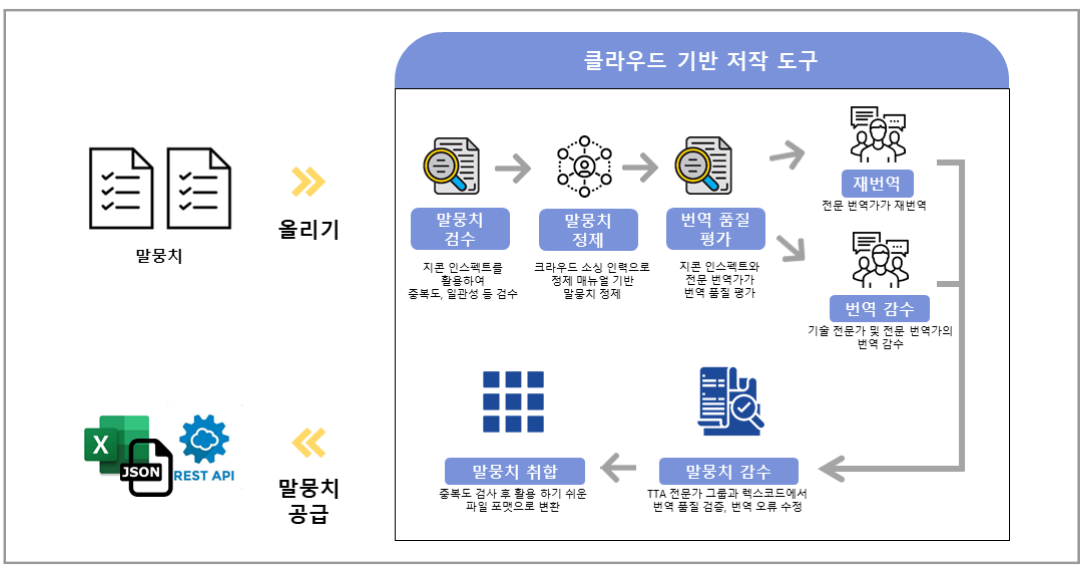

그림5 | 데이터 구축 프로세스

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.

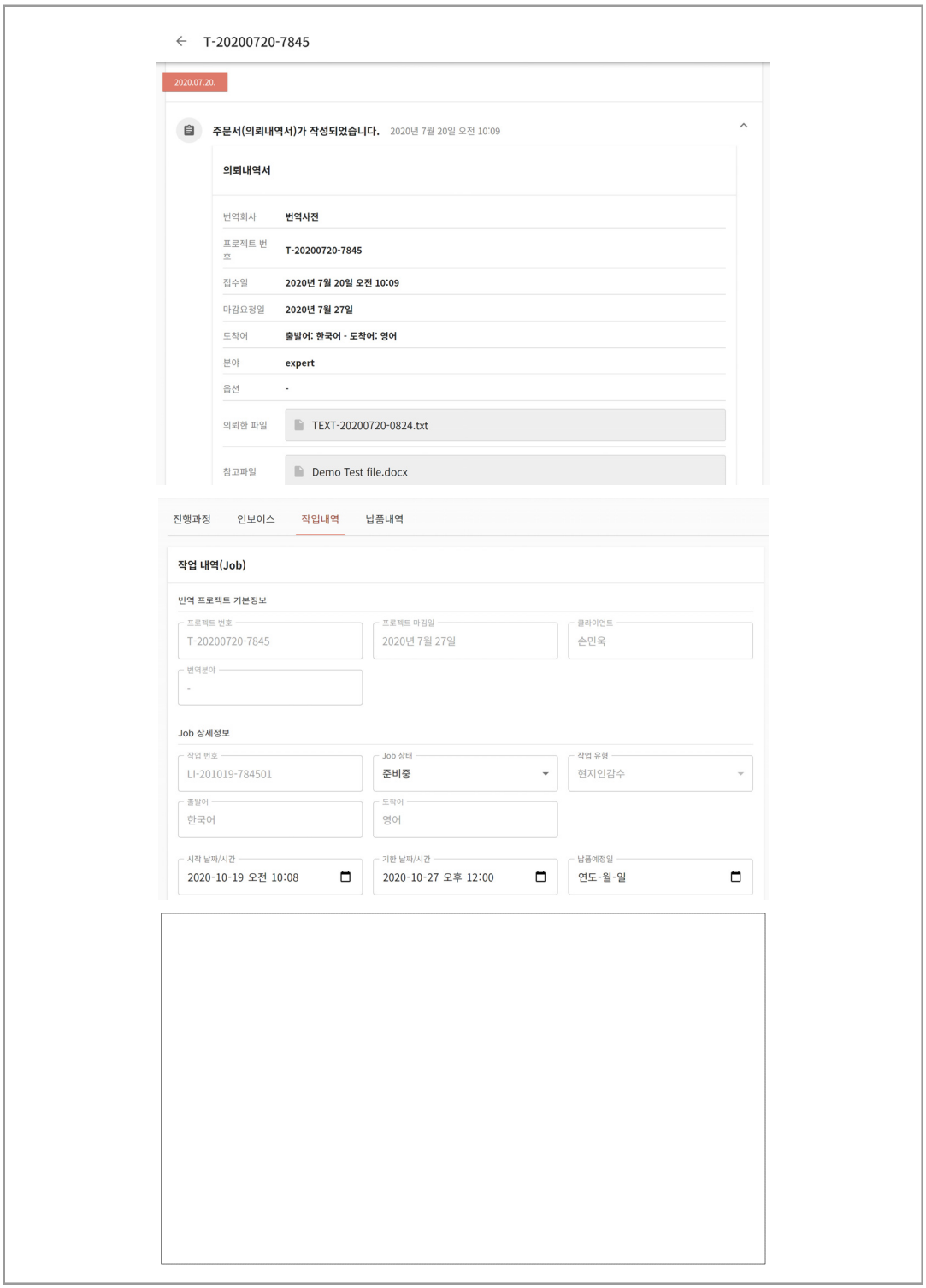

그림6 | 웹 기반 저작도구 화면

●○ 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.

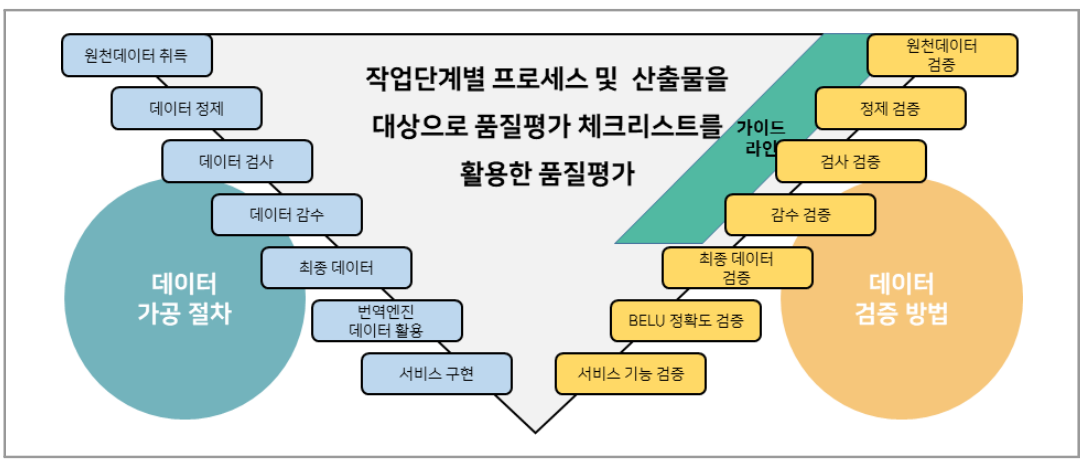

그림7 | 품질 확보를 위한 품질 검수 체계

●○ 데이터 구축 담당자

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