# 테크니컬 리포트

2020년 1차
인공지능
학습용
데이터 구축

자연어 영역

## 전문분야 한영 말뭉치

# 개요: “인공지능 학습용 한국어-영어 번역 말뭉치”란?

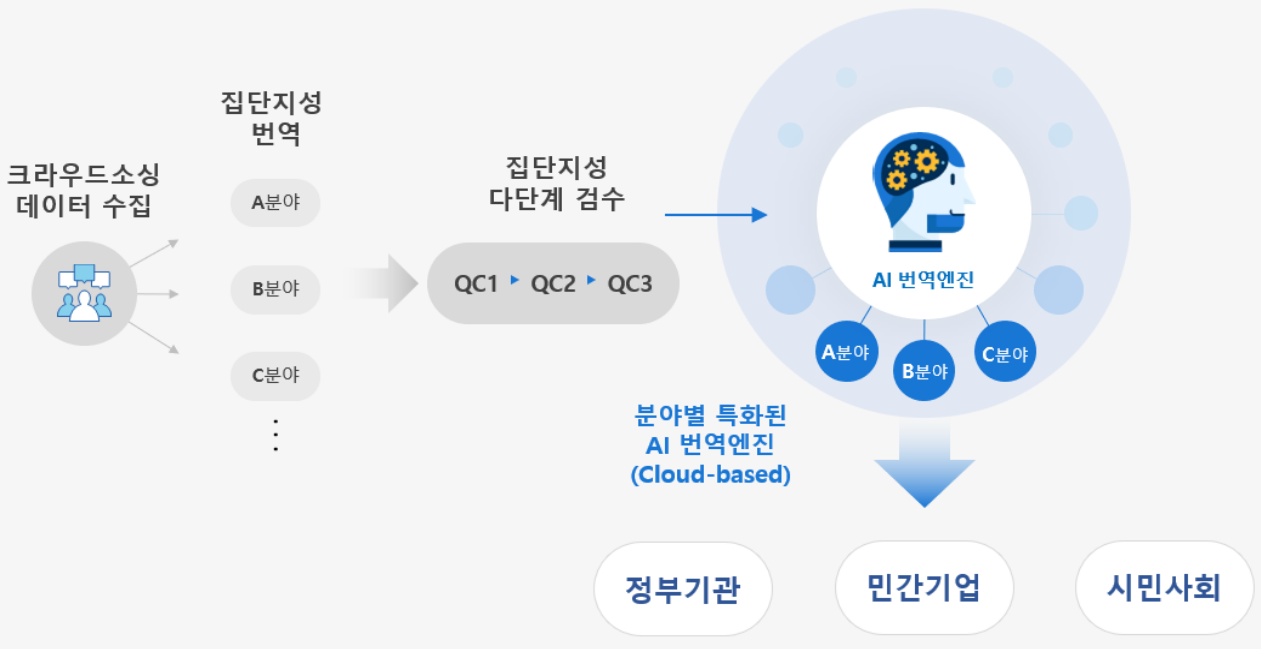
인공지능 학습용 한국어-영어 번역 말뭉치란 AI 번역엔진 구축을 위한 기계학습용 언어쌍을 뜻 한다. 해당 말뭉치는 IT 기술, 가정통신문, 향토문화/음식, 관광, 금융/증시, 의료/보건, 국제스포츠행사, 대법원 판례의 총 8개 전문 분야의 한-영 번역 데이터 구축을 통해 고품질 번역엔진 기계학습의 초석을 다진다.

4차 산업혁명 시대를 맞아 인공지능 번역 엔진 개발을 위한 딥러닝 데이터의 중요성은 해를 거듭할 수록 더해가고 있다. 하지만 영세한 기업들과 연구기관들이 이를 실현하기 위해 필수적인 원천데이터 확보 및 가공·검수 과정에서의 큰 어려움을 겪고 있는 것이 국내의 현실이다.

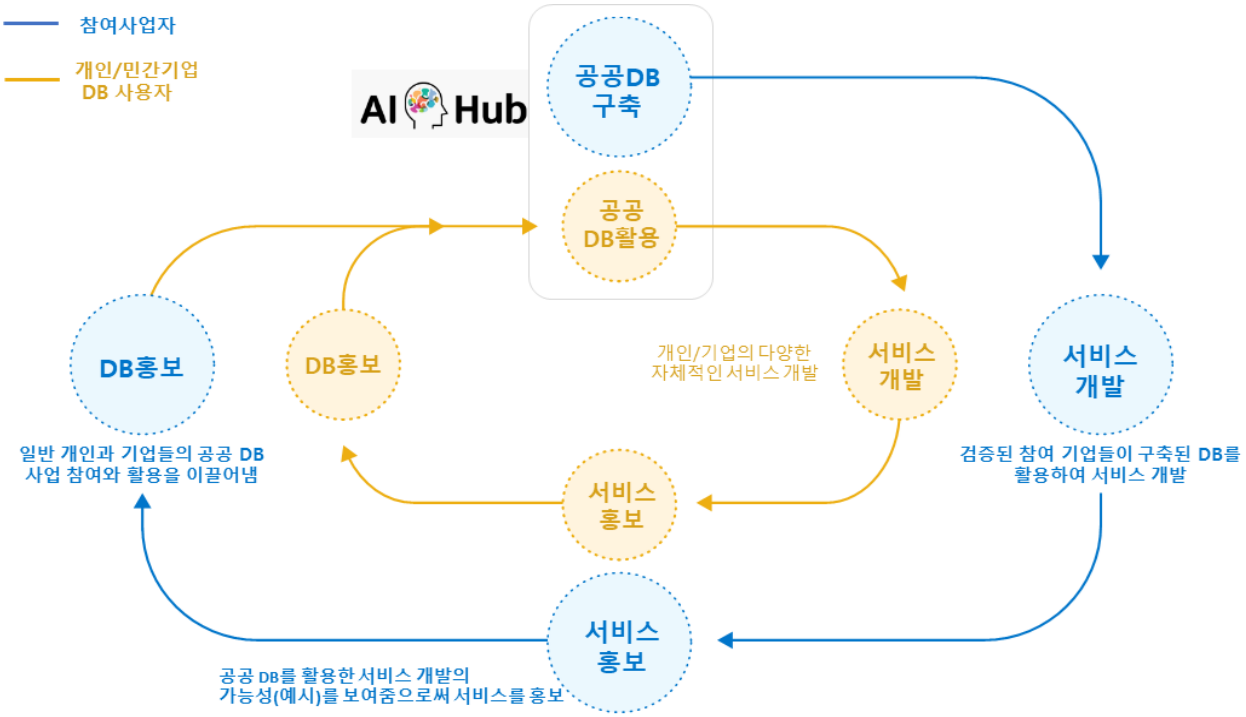
따라서 이번 인공지능 학습용 번역 말뭉치 사업의 목표는 양질의 대규모 학습용 데이터를 구축·공개하여 중소·벤처기업, 스타트업 등 민간 기업들의 인공지능 기술개발을 촉진하고, 이들 기업이 해당 기술을 바탕으로 인공지능 서비스 및 제품을 공급하는 등 인공지능 활용 산업의 활성화를 도모 하는 것이다.

㈜플리토 컨소시엄은 이러한 인공지능 산업의 활성화를 목표로, 한-영 150만 문장(1차 사업 기준)의 분야별 특화 말뭉치 데이터를 구축한다.

인공지능 학습용 번역 말뭉치를 활용한 전문 분야의 특화 서비스 모델의 예시와 이를 통한 개인/민간기업 중심의 AI 데이터 생태계 선순환 조성 프로세스는 아래의 [그림 1]과 [그림 2]를 참고할 수 있다.




**[그림 1] 기존/신규 사업 기반 실제 DB 확장 및 활용 모델**

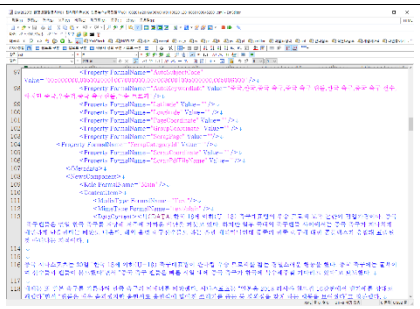
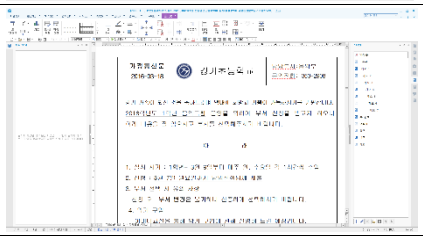
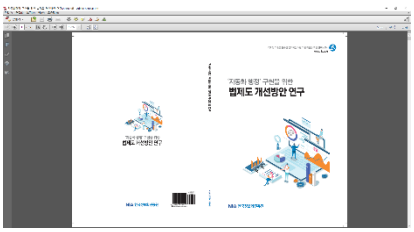



**[그림 2] 개인/민간기업 중심 AI 데이터 생태계 조성 프로세스**

# 데이터의 구성

번역 말뭉치를 구축하기 위한 원천 데이터의 자료 포맷은 XML, HWP, PDF, HTML 방식이며, 이를 각각 디렉토리 별로 분류하여 문장 단위로 DBMS에 저장을 한다. DBMS에선 중복되는 문장, 어절 수에 의한 문장 제외, 그리고 문장의 배분까지 관리를 하게 된다. 이때 일반 EUC-KR이나 KS5601 등의 포맷으로 관리할 경우, 데이터의 소실이 발생할 수 있기 때문에, 데이터들은 UTF-8 포맷으로 저장이 된다. 하단의 [그림 3]에서 실제 DB데이터 구조를 볼 수 있다.

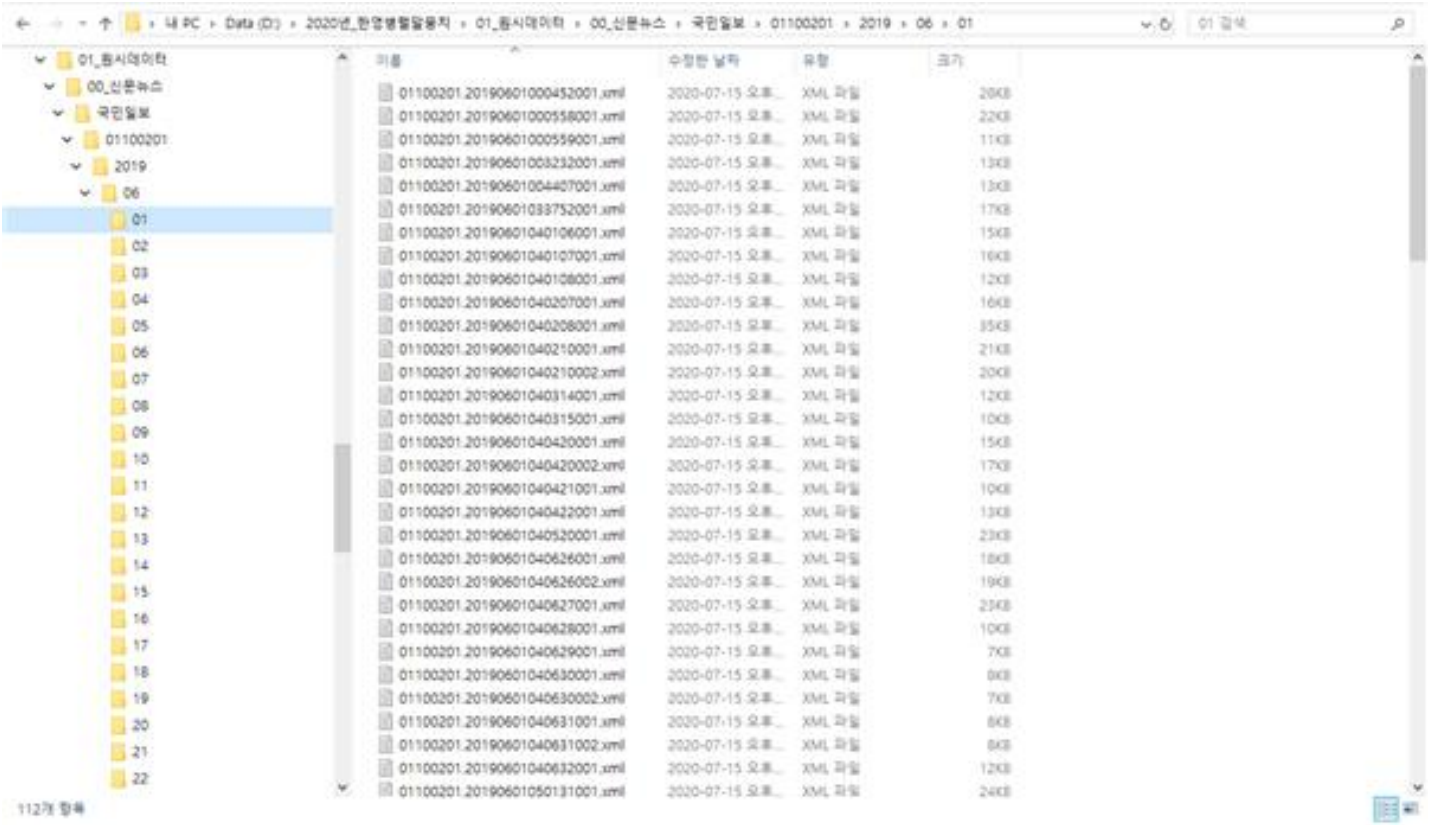
그림과 같은 정제 과정을 거쳐 간단한 분류와 제목 등의 메타정보와 함께 번역팀에 엑셀 형태로 전달이 된다. 이후 번역팀에서는 번역한 결과를 추가하여 최종적으로 엑셀 형태로 말뭉치 데이터를 제공한다.

<table>
<tr><th>포맷</th><th>대상</th><th>예시</th><th>데이터 형식</th></tr>
<tr><td>XML</td><td>신문 자료</td><td>  </td><td rowspan="5">관리용 데이터 구조<br><table><tr><td>id</td><td>int</td></tr><tr><td>파일</td><td>nvarchar(200)</td></tr><tr><td>NewsItemId</td><td>nvarchar(100)</td></tr><tr><td>seq</td><td>int</td></tr><tr><td>URL</td><td>nvarchar(MAX)</td></tr><tr><td>제목</td><td>nvarchar(MAX)</td></tr><tr><td>자동분류</td><td>nvarchar(MAX)</td></tr><tr><td>문장</td><td>nvarchar(500)</td></tr><tr><td>언론사</td><td>nvarchar(150)</td></tr><tr><td>날짜</td><td>nvarchar(150)</td></tr><tr><td>기고자</td><td>nvarchar(50)</td></tr><tr><td>어절수</td><td>int</td></tr><tr><td>used</td><td>nvarchar(10)</td></tr><tr><td>확인</td><td>nvarchar(50)</td></tr><tr><td>구축여부</td><td>nvarchar(50)</td></tr><tr><td>중복</td><td>nvarchar(50)</td></tr></table></td></tr>
<tr><td>HWP</td><td>가정통신문 등</td><td>  </td></tr>
<tr><td>HTML</td><td>웹 사이트<br>수집(openApi)</td><td>  </td></tr>
<tr><td>PDF</td><td>정부기관</td><td>  </td></tr>
</table>

| | | | |
|---|---|---|---|
| | | | 유사도 \| nvarchar(10)<br><br>번역한 이후 데이터는 제목과 URL 제목, url, 분류 항목, 어절수를 포함 형태의 엑셀 파일임 |

**[그림 3] DB테이블 구조**

한 예를 들어 DBMS에 신문 자료의 파일을 저장할 경우, 일반 신문 자료은 텍스트 자료이기 때문에 파일 명 이외의 정보는 대부분 문서 내에 메타데이터로 관리하고 있다. 자세한 관리 구조는 하단의 [그림 4]에서 확인할 수 있다.



**[그림 4] DBMS상 수집 데이터 관리 예시**

## 데이터의 구축 기준과 분포

구축 분야 선정에서는 전문성과 범용성, 이 두 부분의 균형을 적절히 맞추는 것이 매우 중요하다. 너무 깊게 전문 분야에 함몰되면 필요 이상의 고난이도 데이터가 구축되면서 활용할 수 있는 시장이 협소해짐과 동시에 대중적 활용도가 떨어지면서 현실적 수요를 충족시키지 못하는 상황이 발생할 수 있다. 반면, 너무 광범위하거나 평이한 분야는 전문성이 떨어져 증분 학습의 효과를 저해할 수 있다. 기계번역의 수준을 높이기 위해서는 앞으로도 말뭉치를 지속해서 수집할 수 있어야 하므로 전문성을 유지하면서도 일반 수준의 활용도를 만족시키는 말뭉치를 최소한 20~30만 문장 이상 안정적으로 확보할 수 있는 분야를 선정하도록 해야 한다.

아울러 고려할 것은 하나의 전문분야 문장을 번역할 때 그 문장이 전문적인 용어로만 표현되지는 않는다는 것이다. 하나의 문장은 일반적인 언어 규칙을 따르면서 필요한 전문 용어가 곳곳에 포함될 수 있다. 그러므로 기계번역 학습은 해당 분야 내 적정 수준의 전문성을 띤 자료가 선 적용되고, 후에 나머지 전문분야의 자료를 보완하는 형태로 이루어진다.

위의 기준으로 1차 검증 과정을 거친 후보군에서 사업자의 데이터 확보 역량을 고려하여 최종 원천데이터 수집 분야를 선정한다. 상기 내용을 고려한 문장 단위의 세부 특성을 정리해 보면 다음과 같다.

- 공공성과 전문적 활용성을 동시에 염두에 둔 문장
- 전문성이 있지만 지나치게 국소적이지 않아 여러 분야에 적용 가능한 범용적인 문장
- 뉴스와 같이 인공지능 번역 학습의 기본이 되며 활용성이 높은 문장
- 자료 수집이 용이하여 향후에도 안정적으로 확보할 수 있는 자료(자료량, 저작권)

이러한 다양한 사항들을 고려하여 문장의 획득·정제/가공을 위한 기준을 설립하여 이에 따라 데이터 구축을 실행한다. 데이터 구축 과정의 더 구체적인 기준들은 하단의 [그림 5]와 [그림 6]에서 확인할 수 있다.

| 기준 | 내용 |
|---|---|
| 부적합 데이터 제외 | - 분절 오류 문장 삭제 또는 수정<br>- 고유명사 처리 기준 미달 문장 삭제(개인정보, 미등록어 과다 등)<br>- 심각한 비문 삭제 |
| 문장 부호 통일 | - 번역 부적합 기호 처리(이중 꺽쇠, 중점 등)<br>- 문장 종료 처리(분리, 병합, 마침표 등)<br>- 오타, 인코딩 오류 등 교정 |
| 어절수 | - 대략 5~30어절(평균 15어절)의 문장 |
| 중복 문장 | - 2019년도에 구축한 문장을 포함하여 중복된 문장은 제외 |
| 유사도 | - 문장 유사도가 90% 이하인 문장 |
| 메타 데이터 | - ID, 제목, URL 등 메타정보 포함 |

**[그림 5] 문장 정제 기준**

| 기준 | 내용 |
|---|---|
| 고유명사 번역 | - 원문에 이미 번역이 함께 병기된 경우 병기된 내용을 기준으로 작성<br>- 원문에 번역이 없는 경우 검색하고, 바로 찾을 수 없는 경우 #원문고유명사#로 문장에 표기<br>- 언어 고유의 기호도 그대로 유지 |
| 상세 지침 | - 국립국어원 외래어/로마자 표기법 준수<br>- (지명) 로마자로만 표기하지 않고 의미를 함께 작성<br>- (회사명, 모델명) 공식 이름에 따라 작성 |
| 띄어쓰기 | - 단위는 앞 핵심어에 붙여 씀 |
| 기호 | - 여러 종류의 대쉬는 일반적인 하이픈(-)으로 대체<br>- 원문에 강조나 인용 등을 위해 쓰인 따옴표도 번역문에서 따옴표를 사용하여 작성<br>- 원문의 기호를 되도록 영문에서 매칭되게 작성하되, 해당 기호가 영문에서 쓰이지 않는 경우 적절하게 번역 |
| 대문자열의 약어 | 원문의 대문자열은 풀어서 번역하지 않고 그대로 작성<br>- 영문 약어가 국문에서 소문자로 잘못 표기된 경우 대문자로 작성<br>- 원문에 약어와 풀이가 병기되어 있을 때 그 형식을 유지하여 번역 |
| 문장 분절 | - 원문은 한 문장인데, 기계 번역문이 두 문장으로 되어 있는 경우 억지로 한 문장으로 수정하지 않음. |
| 원문에 충실한 번역 | - 원문과 비교 했을 때 동일한 뜻/의도의 문장임을 알 수 있을 정도의 의역을 허용 |
| 맞춤법 오류 | - 오탈자 및 수 불일치된 경우 이를 적절하게 수정<br>- 그 외 단순 맞춤법/문법 오류를 수정하고, 수정하면서 생기는 맞춤법 오류 또한 주의하여 작성 |

**[그림 6] 문장 가공(번역) 기준**

말뭉치는 원시 언어로 된 텍스트 문장과 목적 언어로 된 텍스트 문장쌍으로 이루어진 형태를 말한다. 본 인공지능 한-영 학습용 번역 말뭉치 데이터는 용도가 높은 대략 8개의 분야(분포) 기준 하에 세분화된 말뭉치쌍 구축 목표숫자를 설정한 후 구축에 착수한다. 아래 [그림 7]을 통해 배정된 언어쌍과 분야별 문장 목표수의 규모를 확인할 수 있다.

| 한국어-영어 | |
|---|---|
| 분야 | 1차 구축목표 |
| IT 기술 | 200,000 |
| 가정통신문 | 100,000 |
| 향토문화/음식 | 200,000 |
| 관광 | 200,000 |
| 금융/증시 | 200,000 |
| 의료보건 | 250,000 |
| 국제스포츠행사 | 200,000 |
| 대법원 판례 | 150,000 |
| 구축 목표량 | 1,500,000 |

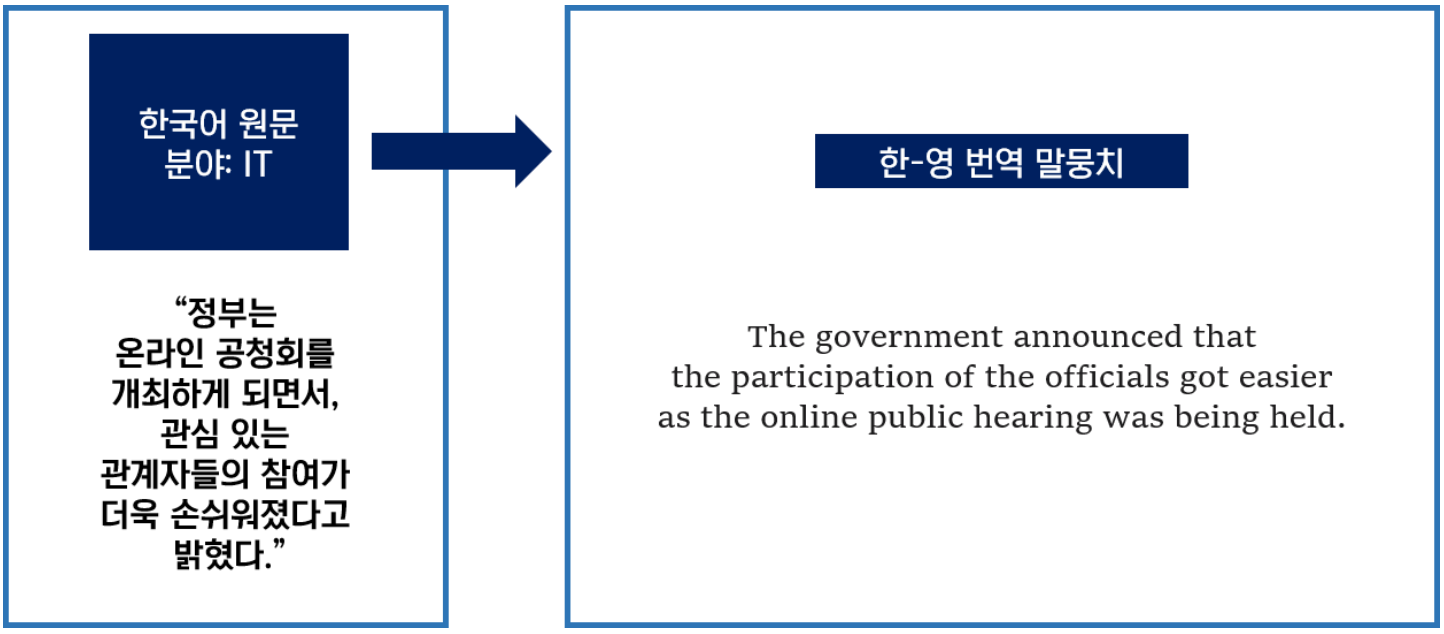
**[그림 7] 한국어-영어 분야별 구축 목표량**

# 데이터 예시

원문 데이터는 획득·정제 과정을 거쳐 간단한 분류와 제목 등의 메타정보와 함께 가공(번역)기관에 엑셀 형태로 전달이 된 후 가공(번역)을 맡은 기관들은 번역한 결과를 추가하여 최종적으로 엑셀 형태로 말뭉치 데이터를 제공한다. 가공 과정이 마무리가 되면 기관들은 [그림 10]과 같은 형태로 납품을 하게 된다. 데이터 가공의 경우 고품질의 학습용 번역 말뭉치를 구축하기 위해선 세부적인 기준하에 진행이 되며 여러 차례의 검증 과정을 거치게 된다. [그림 11]과 같이 잘 정제된 원문 데이터를 한국어-영어 언어쌍에 맞게 번역하여 번역 말뭉치를 구축한다.

| 번역기관 | ORIGIN | TRANSLATION | 분야 | 한국어 어절수 | 영어 단어수 | 길이분류 | 난이도 |
|---|---|---|---|---|---|---|---|
| 플리토 | "가정에서 활용할 수 있는 진로 탐색 자료를 다음과 같이 안내하오니 우리 아이들의 꿈을 기르는 데 활용하시기 바랍니다." | "Career exploration materials that can be used at home are guided as follows, so please use them to raise the dreams of our children." | 가정 통신문 | 17 | 24 | 3 | 중 |
| 플리토 | 정부가 대주주의 사재 출연이나 지분 담보 등을 조건으로 걸지 않았기 때문에 조 회장도 경영권 분쟁에 대한 부담을 잠시나마 내려놓을 수 있을 것으로 보인다." | Chairman Cho is also expected to be able to put down the burden of management rights dispute for a while, as the government has not put a condition on the major shareholders' contribution to private property or equity collateral." | 금융/증시 | 23 | 39 | 4 | 상 |
| 플리토 | "이어 김지찬까지 우전안타를 날렸지만 2루 주자 박민이 홈으로 쇄도하다 원포지션이 투수인 일본 우익수 미야기 히로야의 정확한 원바운드 송구에 아웃되며 아쉽게 이닝이 끝났다.", | Kim Ji-chan then made a right-handed hitter, but the inning ended when the 2nd baseman Park Min rushed toward home and was tagged out on an accurate 1-bound throw by the Japanese pitcher and right fielder Hiroya Miyagi.", | 스포츠 | 22 | 38 | 4 | 상 |

**[그림 10] 인공지능 학습용 번역 말뭉치 데이터 가공 완료 형태**




**[그림 11] 인공지능 학습용 번역 말뭉치의 사례**

# 데이터 구축 과정

우선 실제 번역에 들어갔을 경우, 부적절한 문장을 제외하고 말뭉치로써 조건을 충족시키는 문장만을 구축하기 위해선 번역에 필요한 문장의 2배수를 확보해야 한다. 보다 효율적으로 적합한 문장들을 정제하는 과정은 기본적인 요건들 하에 엑셀, EmEditor, 아래한글등과 같은 도구를 이용하여 이루어진다.

신문이나 저널을 비롯한 웹에서 제공하는 데이터는 생각보다 정제가 되어 있지 않다. 이러한 문장을 정제하는 방법은 최대한 빠르게 많은 문장을 수집한 다음에 기계적 검증을 거쳐 번역에 적합하지 않은 문장은 버리는 것이다. 그런 다음에 맞춤법 등을 빠르게 확인해야 한다.

원천 데이터 형태에 따라 수집 방법이 달라진다. 예를 들어, 신문 자료는 한국언론재단으로부터 구매하며 문서의 구조는 XML 형식으로 되어 있다. 이 경우에는 파이썬 프로그램을 이용하여 XML 문서에서 필요한 문장을 DB에 저장하면 된다. 그러나 웹에서 openAPI 형태로 제공하는 문서는 직접 문서를 수집해야 한다. 만약에 웹 페이지에서 직접 문서를 수집해야 할 때에는 해당 사이트를 수집할 수 있는 프로그램을 만들어서 수집한다. 그 외에도 중복 여부를 검증하고 유사도를 비교하는 등 세밀한 과정을 거쳐 가능한 가장 정확하게 정제된 원문데이터를 번역기관에 제공하는 것을 목표로 한다.

인공지능 학습용 번역 말뭉치의 원문 수집 절차 과정은 아래의 [그림 12]와 같다.

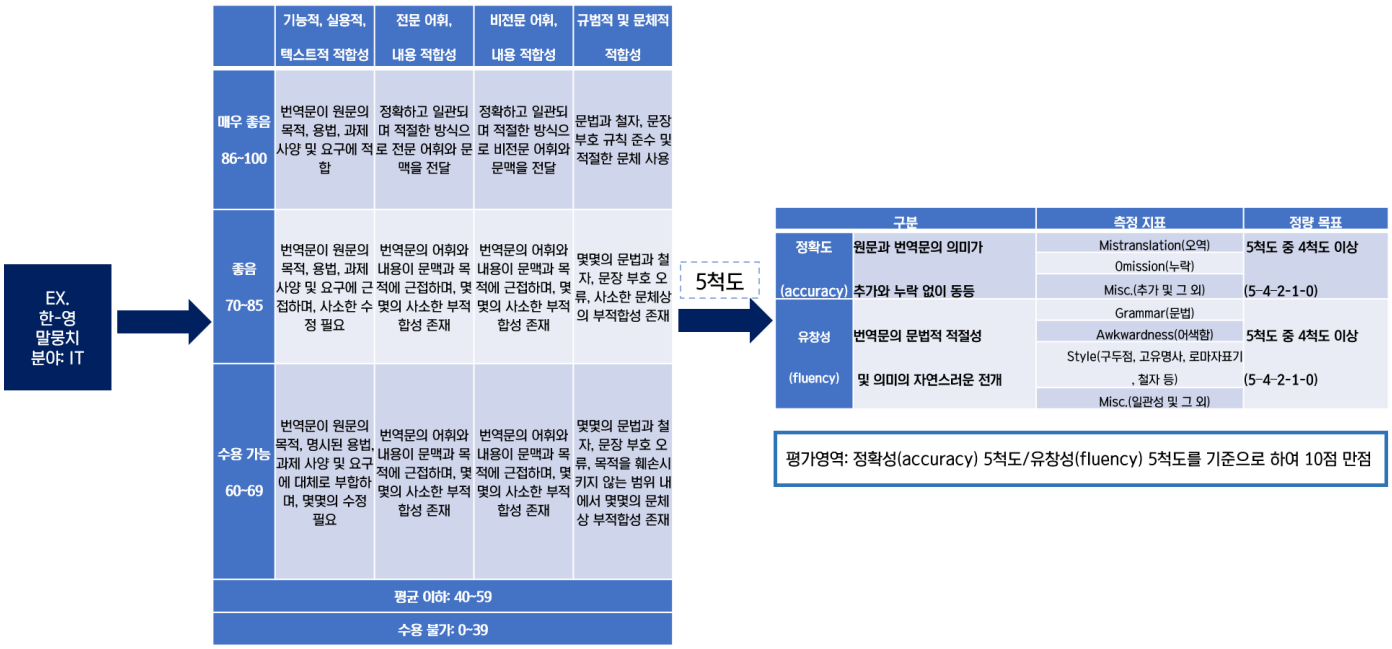

**[그림 12] 획득·정제 절차 공정흐름도**

원문을 공급받은 기관들에선 품질 담당자가 제공받은 원문을 2차로 정제한 후, 번역 담당자들은 준수해야 할 공통 지침 사항 기준 하에 분야 별로 정리된 세트 원문에 대한 가공, 즉 번역하는 과정에 돌입한다.  번역 담당자들은 하루, 혹은 일주일 단위로 할당된 양을 다음 지침에 맞추어 작업한다. NMT를 활용할 경우 분야 혹은 문장 형식에 따라 NMT 엔진을 자유롭게 바꾸어 작업하여 적은 시간과 비용으로 보다 효율적으로 작성한다. [그림 6]와 같은 기준에 따라 번역가들은 좋은 양질의 번역문을 제작하여 [그림 10]과 같은 형태로 5% 샘플링을 하여 검수기관에 번역완료 말뭉치들을 전달한다.

# 검수와 품질 확보

- **TQA 질적 측정 기준**

대량의 데이터를 높은 품질로 생성하기 위하여, 검수 기관들은 전체 가공 납품 물량의 5%의 샘플을 추출하여 문장을 검수한다. 이러한 말뭉치 번역 데이터의 고품질을 유지하기 위한 대표적인 기준으론 'TQA 질적 측정 기준'이 있다. 평가항목은 정확성(오역, 누락, 기타오류) 및 유창성(문법, 어색함, 스타일, 기타오류) 등의 항목으로 나누어 번역 키워드의 사소한 실수와 중대한 실수로 구분하여 객관성 강조하며, 각 5점씩 배점이 되어 10점을 만점으로 한다. 검수 연구원들의 공정하고 타당한 검수결과를 위하여 반복적인 눈높이 맞추기 및 수시로 피드백을 주고받으며 검수작업에 착수한다. 구체적인 과정은 하단의 [그림 13]을 통해 확인할 수 있다.




**[그림 13] TQA 질적 측정 기준 프로세스**

샘플추출 과정에서 구축도구 내 기술상의 오류로 인해 미완성의 문장(번역 또는 검수 미완성, 원문과 번역문 불일치 등)을 수령할 경우 이를 기관들에게 보고하고, 해당 문장에 대한 수정을 진행하는 한편 기술적 오류를 파악하고 개선할 수 있도록 해당 내용을 전달한다. 통계적 결과를 기반으로 추후 기술적 분석으로 품질이 99.5% 이상이 되도록 평가한다. 즉 모집단 오류 비율을 측정하는 p-값(오류의 유의미성)을 p=0.05% 즉 99.95%로 설정한 후 예상치 99.5%는 최소한 지키게 한다. 작년의 경우 p=0.04939% 즉 99.5061%가 되었으며 올해에도 이 정도 수치 이상으로 될 것으로 예상한다.

- **통계 분석(statistical analysis)**

지속적인 품질 관리를 위해서는 객관성과 타당성이 중요한 요소가 된다. 이에 검수기관에서는 언어에 적용 가능한 R-통계를 기반으로 아래와 같은 절차로 품질 관리에 집중한다. 통계 분석은 번역 영역 별 단일 모집단이 임의로 설정한 오류허용 비율에 비추어 오류를 생산하는 집단인지 아닌지를 분석하는 것을 말하며, 과정에서 필요한 추론 통계(inferential statistics)를 위해 이항 분석(binomial analysis) 방법 도입한다. 기본적으로, 표본 크기(sample size)는 클수록 좋지만 전체의 5% 정도와 신뢰구간(confidence level) 95% 허용오차(error tolerance) 5%는 타당하다.

단일모집단 비율검증을 위해서는 기본적으로 이항분포(binominal distribution) 사용한다. 이항검정(binominal test)은 하나의 변수가 두 개의 값을 가질 때, 첫 번째 값을 가지는 확률에 대한 검정 값을 말한다. 관련되어 [그림 15]에서 '이항검정(Binomial test) 통계분석 방법으로 도출한 오류허용 비율에 따른 평가문장 개수 별 ' 오류허용 문장 비율' 표를 확인할 수 있다.

모집단에서 추출하는 표본문장의 개수는 수 천문장 정도만 되어도 충분하며, 본 통계를 위한 표본문장의 개수는 전체 문장의 5%인  75,000 문장 수준이다. 아래 [그림 15]은 몇 개의 문장을 통계 분석 샘플로 삼아야 충분한 샘플이 가능한지의 여부를 대략적으로 알아보기 위하여 모의로 통계 분석을 시뮬레이션을 활용한 수치이다.


**[그림 14] 프로젝트 감수 데이터 관리 시스템 예시**

| 오류허용 비율 | 평가문장 개수 | 실제 오류 허용 개수 | p-값 | 실제 오류 비율 (%) |
|---|---|---|---|---|
| 5% | 100 | 1 | 0.03708 | 1 |
| | 200 | 4 | 0.02645 | 2 |
| | 300 | 8 | 0.03407 | 2.6 |
| | 500 | 16 | 0.03429 | 3.2 |
| | 800 | 29 | 0.03943 | 3.6 |
| | 1000 | 38 | 0.04335 | 3.8 |
| | 1500 | 60 | 0.3935 | 4 |
| | 2000 | 83 | 0.04231 | 4.1 |
| | 3000 | 130 | 0.04885 | 4.3 |
| | 5000 | 225 | 0.0472 | 4.4 |
| | 8000 | 367 | 0.04633 | 4.58 |
| | 10000 | 463 | 0.04574 | 4.6 |
| | 15000 | 705 | 0.04673 | 4.7 |
| | 20000 | 949 | 0.04979 | 4.74 |
| | 25000 | 1193 | 0.04977 | 4.77 |
| | 30000 | 1193 | 0.04818 | 4.79 |
| | 50000 | 2419 | 0.04874 | 4.83 |
| | 80000 | 3898 | 0.04939 | 4.87 |
| 15% | 100 | 8 | 0.02748 | 8 |
| | 200 | 21 | 0.0415 | 10.5 |
| | 300 | 34 | 0.04106 | 11.3 |
| | 500 | 61 | 0.04163 | 12.2 |
| | 800 | 103 | 0.049 | 12.8 |
| | 1000 | 131 | 0.04875 | 13.1 |
| | 1500 | 201 | 0.04306 | 13.4 |
| | 2000 | 273 | 0.04717 | 13.6 |
| | 3000 | 417 | 0.04719 | 13.9 |
| | 5000 | 708 | 0.04929 | 14.1 |
| | 8000 | 1147 | 0.04945 | 14.3 |
| | 10000 | 1440 | 0.04724 | 14.4 |
| | 15000 | 2177 | 0.0482 | 14.5 |
| | 20000 | 2916 | 0.0487 | 14.58 |
| | 25000 | 3656 | 0.04848 | 14.6 |
| | 30000 | 4397 | 0.04839 | 14.65 |
| | 50000 | 7368 | 0.04952 | 14.73 |
| | 80000 | 11833 | 0.04941 | 14.79 |

**[그림 15] 이항검정(Binomial test) 통계분석 방법으로 도출한 오류허용 비율에 따른 평가문장 개수 별 '오류허용 문장 비율'**

오류허용 비율을 5%로 설정할 경우(상황에 따라 비율은 달라질 수 있음)

1) 표본 크기($n$)이 100개일 경우, 오류 문장이 2개(실제 오류 비율:1%)만 되어도 *p*값(p-value) 유의수준 $\alpha$ = .05 초과

2) 표본 크기($n$)이 500개로 늘어나도, 오류 문장이 17개(실제 오류 비율:3.2%)만 되면 *p*값이 유의수준 $\alpha$= .05 초과

3) 그러나 샘플문장이 1500개가 되면, 오류 문장이 61개(실제 오류 비율:4%)가 되고 *p*값이 유의수준 $\alpha$= .05 초과. 이 지점부터는 대략적으로 실제 오류 비율이 5퍼센트에 근접하기 시작

4) 따라서 샘플추출량이 5% 보다 낮거나 높아도 전체 통계 분석에 활용되는 문장의 개수가 수천 개의 수준만 되면 추론통계를 할 수 있는 분량

- **품질검사와 피드백**

말뭉치의 품질 검사를 위해선 구축문장을 검수지침서에 따라 평가영역, 평가항목으로 나누어 품질을 평가하고 그 결과를 컨소시엄 각 업체에 개별적으로, 수시로 전달한다. 이를 전달 받은 컨소시엄 기관은 품질평과 결과를 토대로 내부 번역사에 대한 자체 평가를 진행하며 이를 바탕으로 적절한 피드백을 제공한다. 작업 후반기에 검수물량이 대량으로 몰리는 현상이 발생 방지를 위해 작업량을 효율적으로 배분하며 컨소시엄과의 유기적인 소통을 통해 품질 평가 일정을 사전에 조율한다. 샘플추출 과정에서 구축도구 내 기술상의 오류로 인해 미완성의 문장(번역 또는 검수 미완성, 원문과 번역문 불일치 등)을 수령할 경우 이를 컨소시엄에 보고하며 해당 문장에 대한 수정을 진행하는 한편 기술적 오류를 파악하고 개선할 수 있도록 해당 내용을 전달한다. 이러한 체계적인 과정을 통해 고품질의 말뭉치 품질 검사를 진행한다.

- **학습데이터 적합성 검증**

인공신경망 기계번역(NMT) 기술은 구글, 마이크로소프트 등 국제적인 회사와 네이버, 카카오 등 국내 회사에서 현재까지 어텐션을 이용하는 트랜스포머 딥러닝 기술을 기반으로 성능 좋은 기계번역 결과를 내고 있다. 이 어텐션을 이용한 트랜스포머 딥러닝 기술은 공개되어 누구나 활용이 가능하며, 이를 통해서 구축된 병렬코퍼스를 학습데이터로 활용하는 방안은 이미 검증된 방안이다. 더불어, AUTOML 등 전이학습에도 학습데이터로 활용이 가능하여, 분야별 특화된 번역모델 개발에도 활용 가능하다.

## 데이터 구축 담당자

수행기관(주관) : ㈜플리토 (전화: 010-3630-8864), 이메일: dongwon.seo@flitto.com