

# AI Hub 한국어-영어 번역 말뭉치를 통한 영어 수동문의 한국어 번역 양상 분석

- 2024-2 언어데이터과학 기말프로젝트 중간 발표-

2021-14058 언어학과 이예은

# 목차

## 1. 연구 배경

## 2. 데이터 구축 및 활용 계획

### 2.1. 데이터 선정 및 수집

### 2.2. 데이터 가공

## 3. 추후 프로젝트 진행 계획

# 1. 연구 배경

영어 수동문은 한국어와의 구조적 차이 때문에 번역 과정에서 생략, 능동문으로의 변환, 어색한 표현들이 자주 발생하곤 한다. 이 프로젝트에서는 AI Hub의 한국어-영어 번역 말뭉치를 활용해 영어 수동문이 한국어에서 피동문과 능동문으로 각각 번역되는 양상을 비교 분석함으로써, 이러한 번역 문제를 체계적으로 분석하고 규칙성을 도출하고자 한다.

## 2.1. 데이터 선정 및 수집

- 프로젝트 목적을 달성하기 위해서는 실제 번역 양상을 담고 있고, 한국어와 영어 병렬 데이터가 문장 단위로 구성되어 있는 코퍼스를 사용해야 함
- AI Hub에서 공개한 한국어-영어 번역 말뭉치(뉴스, 법률, 정부 웹사이트, 한국문화 텍스트, 구어체, 대화체) 중 대화체 데이터를 선택

대분류	소분류	상황	Set Nr.	발화자	원문	번역문
비즈니스	회의	의견 교환하기	1	A-1	이번 신제품 출시에 대한 시장의 반응은 어떤가요?	How is the market's reaction to the newly released product?
비즈니스	회의	의견 교환하기	1	B-1	판매량이 지난번 제품보다 빠르게 늘고 있습니다.	The sales increase is faster than the previous product.
비즈니스	회의	의견 교환하기	1	A-2	그렇다면 공장에 연락해서 주문량을 더 늘려야겠네요.	Then, we'll have to call the manufacturer and increase the volume of orders.
비즈니스	회의	의견 교환하기	1	B-2	네, 제가 연락해서 주문량을 2배로 늘리겠습니다.	Sure, I'll make a call and double the volume of orders.
비즈니스	회의	의견 교환하기	2	A-1	지난 회의 마지막에 논의했던 안건을 다시 볼까요?	Shall we take a look at the issues we discussed by the end of the last meeting?
비즈니스	회의	의견 교환하기	2	B-1	그보다는 이번 주 새로운 주제가 더 급한 것 같습니다.	I believe that this week's new issues are more urgent.
비즈니스	회의	의견 교환하기	2	A-2	그럼 새로운 안건으로 회의를 시작하도록 하죠.	Then, let's begin our meeting with the new issues.
비즈니스	회의	의견 교환하기	2	B-2	네, 자료는 여러분의 앞에 미리 준비되어 있습니다.	Sure, the related materials are ready in front of you.
비즈니스	회의	의견 교환하기	3	A-1	이번 주 금요일까지 2천개를 더 주문하라는 건가요?	Do you mean we need to order additional 2,000 items by this Friday?
비즈니스	회의	의견 교환하기	3	B-1	네, 시간이 조금 촉박하기는 하지만 가능해 보이는데요.	Yes, time is running short, but we can manage it.
비즈니스	회의	의견 교환하기	3	A-2	주문은 가능하지만, 수령은 2달 정도 걸릴 것 같네요.	The order is acceptable, but it would take about 2 months until the delivery.
비즈니스	회의	의견 교환하기	3	B-2	이런, 저희는 물건을 최대한 빠르게 받고 싶어요.	Gee, we'd like to receive the items as soon as possible.

## 2.2. 데이터 가공

```
import pandas as pd
!pip install openpyxl

# 파일 읽기
file_path = './data/한국어-영어 번역(병렬) 말뭉치/2_conversation_200226.xlsx'
df = pd.read_excel(file_path)

# 데이터 정보 확인
print(df.info()) # 데이터 구조 및 열 정보 확인
print(df.head()) # 첫 몇 행 출력

# 데이터 처리
df = df.dropna() # 결측치 제거
df = df[['원문', '번역문']] # 필요한 열만 선택
```

원문	번역문
이번 신제품 출시에 대한 시장의 반응은 어떤가요?	How is the market's reaction to the newly rele...
판매량이 지난번 제품보다 빠르게 늘고 있습니다.	The sales increase is faster than the previous...
그렇다면 공장에 연락해서 주문량을 더 늘려야겠네요.	Then, we'll have to call the manufacturer and ...
네, 제가 연락해서 주문량을 2배로 늘리겠습니다.	Sure, I'll make a call and double the volume o...
지난 회의 마지막에 논의했던 안건을 다시 볼까요?	Shall we take a look at the issues we discusse...
...	...
가격표 배치를 잘못해서 혼동을 드렸나 봐요, 죄송해요.	It seems that we didn't place the price tags c...
백화점 포인트로 계산하고 싶은데, 가능한가요?	Can I pay using the department store points?
네, 물론이죠, 전화번호 입력해주시면 됩니다.	Yes, of course, you just need to enter your ph...
입력했어요, 전액 백화점 포인트로 결제하고 싶어요.	I entered it, I want to pay it with all the de...

## 2.2. 데이터 가공

```
● from spacy.matcher import Matcher
```

```
matcher = Matcher(nlp.vocab)
```

```
▼ passive_rules = [
```

```
    [{'DEP': 'nsubjpass'}, {'DEP': 'aux', 'OP': '*'}, {'DEP': 'auxpass'}, {'TAG': 'VBN'}],
```

```
    [{'DEP': 'nsubjpass'}, {'DEP': 'aux', 'OP': '*'}, {'DEP': 'auxpass'}, {'TAG': 'VBZ'}],
```

```
    [{'DEP': 'nsubjpass'}, {'DEP': 'aux', 'OP': '*'}, {'DEP': 'auxpass'}, {'TAG': 'RB'}, {'TAG': 'VBN'}],
```

```
]
```

```
matcher.add('Passive', passive_rules)
```

```
passive_sentences = []
```

```
✓ 0.0s
```

nsubjpass: nominal subject (passive)

aux: auxiliary

VBN: verb, past participle

VBZ: verb, 3rd person singular present

RB: adverb

패턴	대상 문장	특징
패턴 1	<i>The book was written by the author.</i>	기본적인 수동태 형태 (과거형)
패턴 2	<i>The book is written by the author.</i>	현재형 수동태
패턴 3	<i>The book was carefully written by the author.</i>	부사가 포함된 수동태

[python - Detect passive or active sentence from text - Stack Overflow](#)

## 2.2. 데이터 가공

passive\_sentences

✓ 0.0s

```
['How will our products be shipped to Tokyo?',  
 "They'll be sent to Tokyo port by the ship, then transported by vehicles.",  
 'It was written in very small letters at the bottom of the notification email.',  
 "It's officially scheduled from July 29 to August 2.",  
 "Sure, but they'd already known since it was notified by email earlier.",  
 'I was told that you hired many employees including five interns.',  
 'Then, what procedures are left before the release, other than the commercial?',  
 'Yes, I was told that every employee on my floor already submitted.',  
 'It was raised way more than we planned, so we were all very proud.',  
 "That's a great idea. I hope it'll be approved in the general meeting.",  
 'I like the idea, but will it automatically save the data when the computers are turned off?',  
 'I just got an e-mail from the Busan bureau saying that the schedule will be changed due to heavy rain.',  
 "Oh, that's not good. Do you think our KTX train schedules can be changed?",
```

### 3. 추후 프로젝트 진행 계획

- **한국어 텍스트 분석:** koNLPy 라이브러리를 활용하여 피동 접사 및 수동 동사를 감지하여 수동문을 식별
- **데이터 분석:** 구축한 데이터프레임을 pandas로 분석하여 영어 수동문의 한국어 번역 양상을 비교
- **통계 및 시각화:** 피동문과 능동문의 비율을 계산하고, 시각화 기법으로 결과 도출
- **번역 평가:** 문헌에서 제시하는 이상적인 번역과 실제 번역 양상을 비교하여 차이점 분석