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

- 2024-2 언어데이터과학 기말프로젝트 발표-

2021-14058 언어학과 이예은

목차

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 분석**함으로써, 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

중간 발표에서 발견한 문제점:

데이터가 문장 단위로 나뉘어 있지 않아서, 열 단위로 수동문 포함 여부를 감지하면(이진 데이터) 한 데이터에 여러 개의 수동 표현이 포함된 경우 누락이 발생할 수 있음

```
passive_sentences
✓ 0.0s
```

```
['How will our products be shipped to Tokyo?',
 "They'll be sent to Tokyo port by the ship,
 'It was written in very small letters at the
 "It's officially scheduled from July 29 to A
 "Sure, but they'd already known since it was
 'I was told that you hired many employees in
 'Then, what procedures are left before the r
 'Yes, I was told that every employee on my f
 '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?",
```

→ 데이터별로 수동 표현의 개수를 세도록 수정

```python
# 수동문 개수 계산 함수
def count_passive_sentences(text):
    doc = nlp(text)  # 텍스트 처리
    matches = matcher(doc)  # Matcher로 패턴 매칭
    return len(matches)  # 매칭된 패턴 개수 반환
```

2.2. 데이터 가공

- 한국어 피동문을 3개의 패턴으로 나누어 감지
- KoNLPy의 Kkma Class를 사용해 형태소 분석

```
# 피동 접미사 "-이/히/리/기-" 패턴
for tok in tokens:
    if tok[1] == 'VV' and tok[0].endswith(('이', '히', '리', '기')):
        detected_patterns["passive_suffix"].append(tok)

# 보조동사 "-어지-" 패턴
for i in range(len(tokens) - 2):
    if (tokens[i][1] == 'VV' and tokens[i+1][0] == '어' and
        tokens[i+2][0] == '지' and tokens[i+2][1] == 'VXV'):
        detected_patterns["passive_auxiliary"].append((tokens[i], tokens[i+1], tokens[i+2]))

# "-되다", "-받다", "-당하다" 패턴
for tok in tokens:
    if tok[1] == 'VV' and re.search(r'^(되|받|당하)', tok[0]):
        detected_patterns["passive_verb"].append(tok)
```

2.2. 데이터 가공

원문	번역문	kor_passive_count	eng_passive_count
우리 제품은 어떤 방식으로 도쿄에 보내지나요?	How will our products be shipped to Tokyo?	0	1
먼저 배로 도쿄 근처 항구까지 운반하고 그 후 차를 이용합니다.	They'll be sent to Tokyo port by the ship, the...	0	1
교육 안내 메일 하단에 작은 글씨로 적혀있어요.	It was written in very small letters at the bo...	1	1
7월 29일부터 8월 2일까지가 공식적인 기간이에요.	It's officially scheduled from July 29 to Augu...	0	1
네, 메일로 이미 공지를 했으니 다들 알고 있을 거에요.	Sure, but they'd already known since it was no...	0	1
인턴 5명을 포함해서 많은 직원을 뽑았다고 들었어요.	I was told that you hired many employees inclu...	0	1

2.2. 데이터 가공

한국어 데이터 분석에서의 어려움

1) 피동접사와 사동접사 “-이, 히, 리, 기-”의 구분

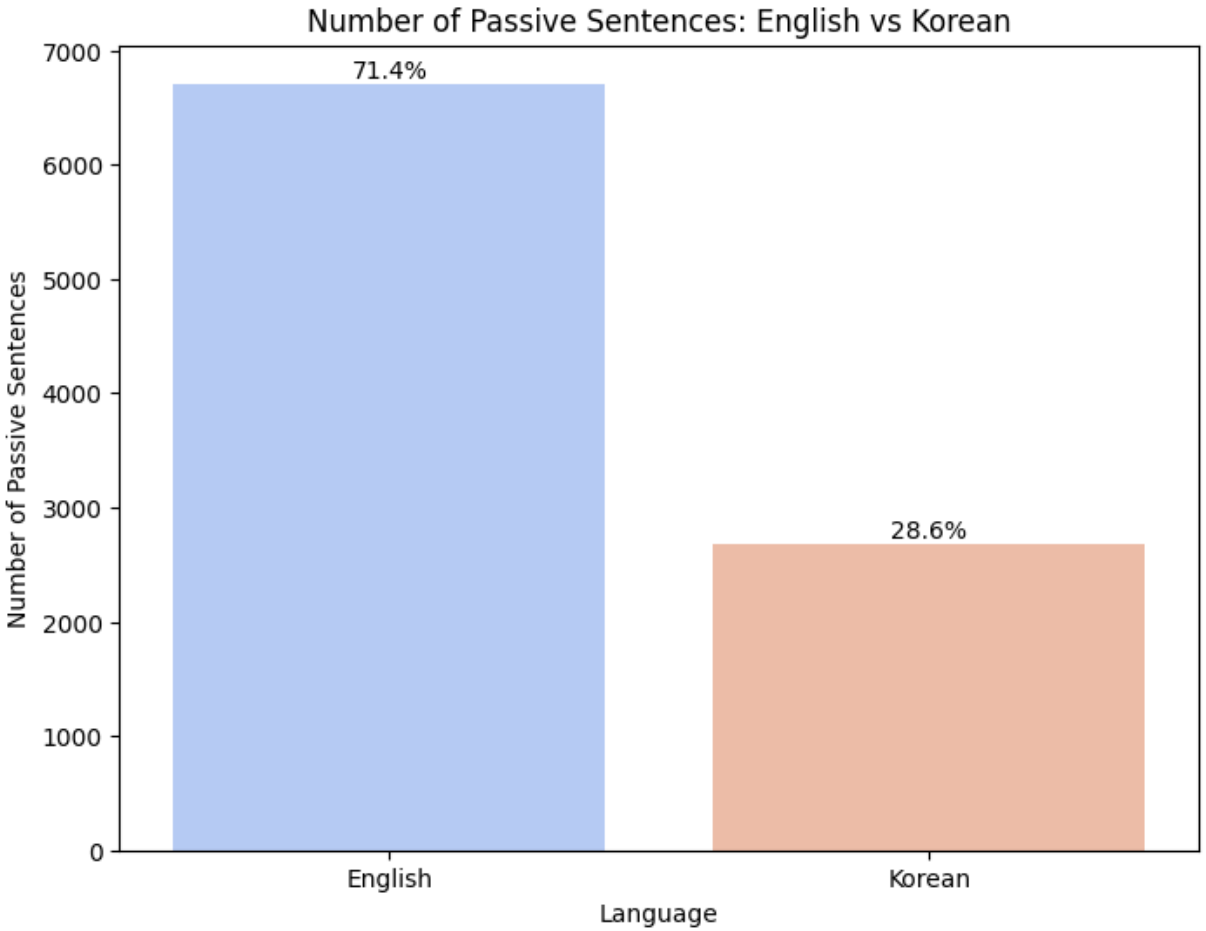
2) 동사 어근 자체에 “-이, 히, 리, 기-”가 포함된 경우

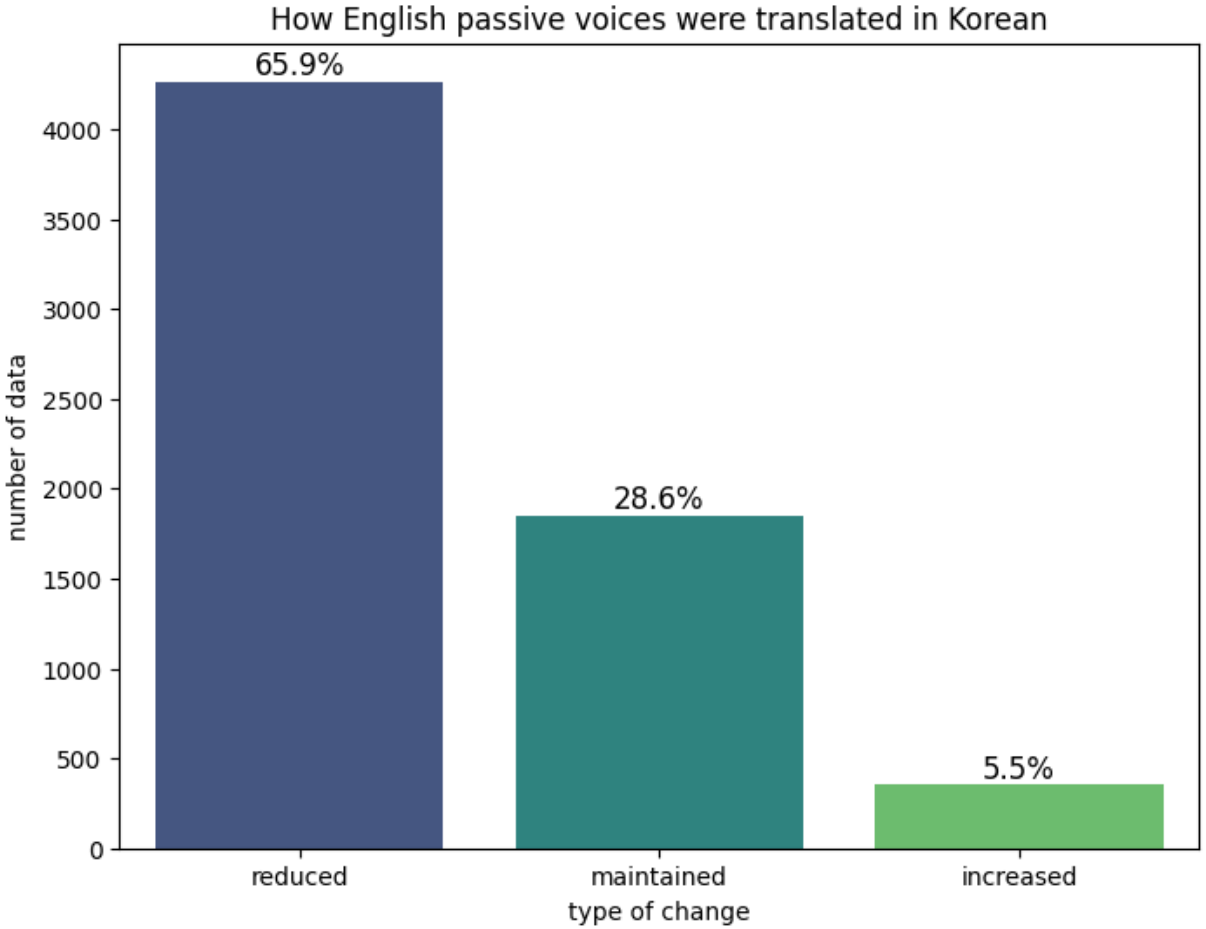
→ 대부분 “드리다” 동사에서 발생하는 문제였으므로 해당 단어만 제외하는 방법

3) “-받다”가 접사가 아닌 동사로 쓰이는 경우

→ “받다”가 동사로 쓰이는 경우 접사와 달리 앞 단어와 공백이 있기 때문에 정규표현식으로 “-받다 " 앞에 공백이 없는 경우만 매칭되도록 함.

3. 연구 결과 및 보완 계획


Number of Passive Sentences: English vs Korean
71.4%
28.6%
Number of Passive Sentences
Language
English
Korean
0
1000
2000
3000
4000
5000
6000
7000


How English passive voices were translated in Korean
65.9%
28.6%
5.5%
number of data
type of change
reduced
maintained
increased
0
500
1000
1500
2000
2500
3000
3500
4000

3. 결론 및 보완 계획

- 한국어의 복잡한 구조로 인해 피동문이 아닌데 피동문으로 감지되는 오류 케이스를 모두 처리하지는 못함. 하지만 이 오류가 전체 연구 결과에 미치는 영향은 미미하다고 판단하고 진행
- 프로젝트를 더 고도화한다면, 영어와 한국어의 수동문/피동문 패턴을 분류해 두었기 때문에 수동문/피동문의 유형별로 번역 양상을 비교해볼 수 있음