

혁신성장 청년인재 집중양성 사업

파파고를 품은 관용구 영어 번역기: Tree

2021년 10월 29일

인공지능 자연어처리(NLP) 기반 기업데이터 분석
태조

김예찬, 김태영, 김택현, 심판교, 유현승, 이지수, 진가형

목 차

| | |
|--|--------|
| 1. 프로젝트 개요..... | - 1 - |
| 1.1 프로젝트 기획 배경 및 목표..... | - 1 - |
| 1.2 구성원 및 역할 | - 2 - |
| 1.3 프로젝트 추진 일정 | - 3 - |
| 2. 프로젝트 관련 선행연구 | - 4 - |
| 2.1 관용구 기계번역을 위한 한-영 데이터셋 구축 및 평가방법 | - 4 - |
| 2.1.1 관용구 데이터셋 구축 | - 4 - |
| 2.1.2 기계번역 성능 평가 방안: 블랙리스트 | - 4 - |
| 2.2 한-영 관용구 기계번역을 위한 NMT 학습 방법 | - 6 - |
| 2.2.1 관용구 데이터셋 구성 | - 6 - |
| 2.2.2 실험 및 결과..... | - 7 - |
| 2.3 Classification of Idiomatic Sentences Using AWD-LSTM | - 9 - |
| 2.3.1 개요 | - 9 - |
| 2.3.2 전이 학습을 통한 관용구 분류기 구축 | - 9 - |
| 2.3.3 관용구 문장 데이터셋(TroFi/자체 데이터셋)..... | - 10 - |
| 2.4 선행연구 분석 | - 12 - |
| 2.4.1. 관용구 기계번역 선행 연구의 한계 및 개선사항..... | - 12 - |
| 2.4.2 관용구 분류기 선행 연구의 한계 및 개선사항 | - 12 - |
| 3. 프로젝트 개발 결과 | - 13 - |
| 3.1 프로젝트 Flow..... | - 13 - |
| 3.2 데이터셋 증축 작업..... | - 14 - |
| 3.2.1 의역을 하는 이유..... | - 15 - |
| 3.2.2 작업 내용(작업 방식)..... | - 15 - |

| | |
|----------------------------|--------|
| 3.3 관용구 분류기 | - 15 - |
| 3.3.1. 관용구 분류기 모델 선정..... | - 15 - |
| 3.3.2 한국어 관용구 데이터 증축 | - 19 - |
| 3.4 관용구 번역기 | - 21 - |
| 3.4.1 번역기 성능 향상 대안 | - 21 - |
| 3.4.2 번역기 1..... | - 22 - |
| 3.4.3. 번역기 2..... | - 22 - |
| 3.5 번역 서비스 - Tree | - 23 - |
| 4. 결론 | - 24 - |

1. 프로젝트 개요

1.1 프로젝트 기획 배경 및 목표

최근 기계 번역에 대한 기술적 수요가 늘어나고 있음에도, 기계 번역기의 관용구 번역 성능은 만족스럽지 못한 경우가 많다. 예를 들어, '집안 형편으로 인해 나는 가방끈이 짧다'는 문장을 넣어 기계번역을 시도했을 때, '많이 교육받지 못해 학력이 짧다'는 관용구의 의미 대신 문자 그대로의 '가방끈이 짧다'는 의미의 'have a short strap' 등으로 번역됨으로써 관용구의 번역이 제대로 이루어지지 않고 문장이 직역되는 것을 확인할 수 있다.

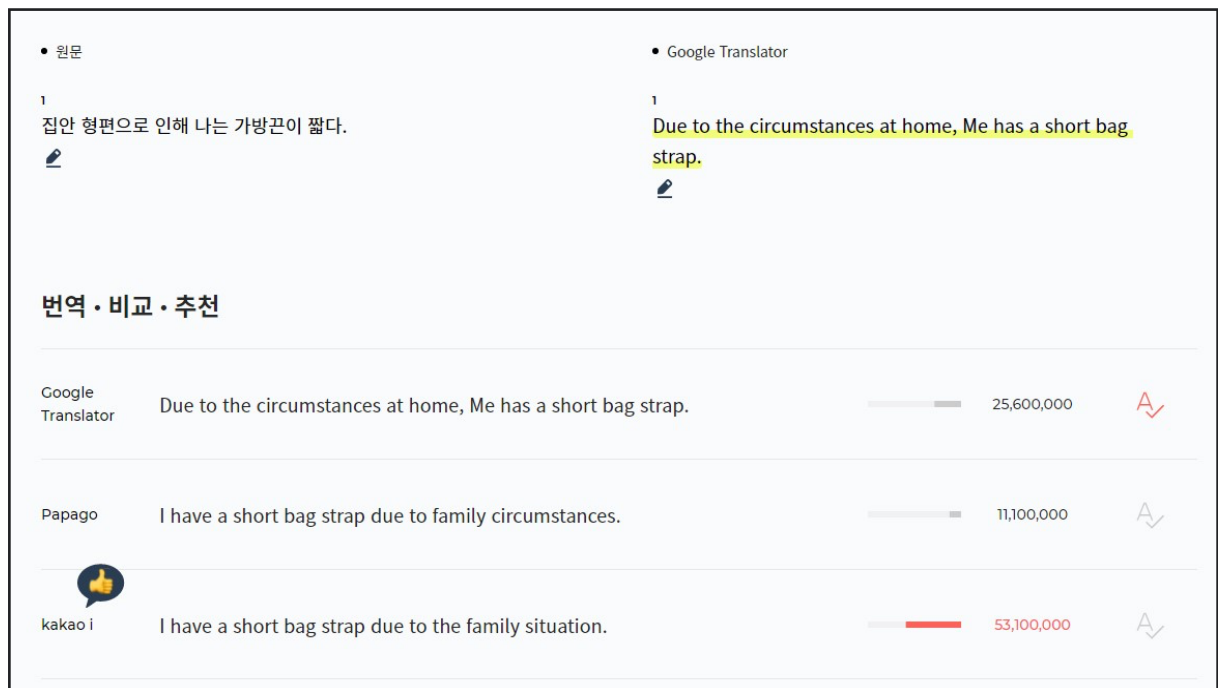


그림 1. 관용구 문장 기계번역 비교(Google Translator, Papago, Kakao i) - 출처 : 지콘스튜디오

이러한 성능적 문제가 있음에도 관용구 번역에 대한 연구나 논문이 부족한 상황이다. 관용구 오번역 문제를 개선할 수 있다면 번역기 실사용자들에게 보다 고품질의 번역 서비스를 제공할 수 있으며, 또한 기계 번역 연구에도 기여할 수 있을 것으로 기대된다. 따라서 태조에서는 관용구를 올바르게 번역해낼 수 있는 번역기 프로젝트를 기획하여 진행하였다.

1.2 구성원 및 역할

| 이름 | 전공 | 역할 | 구현 부분 |
|-----|-----------------|----|---|
| 김태영 | 컴퓨터공학과 | 팀장 | 관용구 분류기 개발, 데이터셋 전처리 및 구축 |
| 김예찬 | 항공정비공학과 | 팀원 | 관용구 분류기 개발, 데이터셋 전처리 및 구축 |
| 김택현 | 영어영문학과 | 팀원 | 관용구 분류기 개발, 데이터셋 전처리 및 구축 |
| 심판교 | 바이오시스템 기계공학과 | 팀원 | 번역 엔진 개발, 데이터셋 전처리 및 구축, 웹 디자인 |
| 유현승 | 경제학과 | 팀원 | 번역 엔진 개발, 데이터셋 전처리 및 구축, 웹 디자인 |
| 이지수 | 컴퓨터공학과 | 팀원 | 번역 엔진 및 API 개발, 웹 사이트 개발, 데이터셋 구축 및 전처리 |
| 진가형 | 영어영문학과 | 팀원 | 번역 엔진 개발, 데이터셋 전처리 및 구축, 웹 디자인 |

1.3 프로젝트 추진 일정

| 구분 | 기간 | 활동 |
|----------------------|-----------------------|---|
| 사전 기획 | 8/13(금) ~ 8/21(토) | 팀 구성 및 주제 브레인스토밍 |
| | 8/24(화) ~ 8/26(목) | 자료 조사 |
| | 8/27(금) ~ 8/31(화) | 논문 데이터 탐색 및 연구 |
| PJT 수행 및 완료 | 9/1(수) ~ 9/14(화) | 프로젝트 수행 1: 논문 구현 및 응용 |
| | 9/15(수) ~ 9/30(목) | 프로젝트 수행 2: 데이터셋 증축 |
| | 10/1(금) ~10/15(금) | 프로젝트 수행 3: 번역기 및 분류기 성능 검증, 프로토타입 제작 |
| | 10/16(토) ~10/29(금) | 프로젝트 수행 4: 데이터셋 증축 및 번역기 업그레이드, 결과보고서 작성 |
| | 11/2(화) | 최종 발표 준비 |
| | 11/3(수) | 최종 발표 |

2. 프로젝트 관련 선행연구

2.1 관용구 기계번역을 위한 한-영 데이터셋 구축 및 평가방법

관용구는 둘 이상의 단어가 결합하여 특정한 뜻을 생성한 어구로 기계번역 시 종종 오역이 발생한다. 이는 관용구가 지닌 함축적인 의미를 정확하게 번역할 수 없는 기계번역의 한계를 드러낸다. 따라서 신경망 기계 번역(Neural Machine Translation)에서 관용구를 효과적으로 학습하고 번역 결과를 평가하려면 관용구에 특화된 번역 쌍 데이터셋과 평가방법이 필요하다. 본 논문에서는 관용구 기계번역을 위한 한-영 번역 쌍 데이터셋을 구축하는 방법과 관용구 번역 결과를 평가하기 위해 블랙리스트를 구축하는 방법을 제안한다.

2.1.1 관용구 데이터셋 구축

한-영 번역 쌍 데이터셋을 구축하기 위해, 논문에서는 우선 표준국어대사전 사이트에서 관용구 목록을 다운로드하였다. 그 결과 총 3,887개의 목록을 확인할 수 있었고, 해당 목록을 AI Hub의 한-영 병렬 말뭉치에서 확인해 본 결과 총 420개의 관용어가 18,808개 문장에서 사용되었음을 알 수 있었다. 이후 데이터셋이 편향되지 않게 하기 위해 동일한 관용구를 가진 문장 개수를 최소 4개, 최대 40개로 제한하여 최종적으로 420개의 관용어를 포함하는 7500개의 한-영 병렬 문장 데이터셋을 구축할 수 있었으며, 논문에서는 이를 KISS(Korean-English Idioms in Sentences Dataset)로 명명하였다.

2.1.2 기계번역 성능 평가 방안: 블랙리스트

기계번역의 성능 평가 방안으로 BLEU 스코어가 흔히 쓰이고 있지만, BLEU 스코어의 경우 정답으로 간주되는 번역 결과와 다른 단어가 번역에 사용되었을 시 실제 번역 결과가 자연스럽더라도 스코어상으로는 낮은 점수를 기록한다는 단점이 있다. 본 논문에서는 이러한 BLEU 스코어의 단점을 보완할 수 있는 방법으로 블랙리스트를 활용한 기계번역 성능 평가를 제안한다.

a. 블랙리스트 구축

관용구는 해당 구를 구성하는 단어의 의미를 그대로 옮겨서는 해석할 수 없기에, 번역기가 관용구의 단어 그대로를 번역한다면 오역에 해당한다.

표 1. 관용구를 포함하는 한-영 번역 쌍 예시

| | |
|---------|--|
| 관용구 | 눈이 높다. |
| 한국어 원문 | 나는 여자 보는 눈이 높아요. |
| 영어 번역 쌍 | I have high standards for women. |
| 관용구 | 마침표를 찍다. |
| 한국어 원문 | 아버지의 명예회복을 위한 김지훈의 기나긴 여정이 마침내 마침표를 찍었다 . |
| 영어 번역 쌍 | Kim Ji-hoon's long journey to restore his father's honor has finally come to an end . |

이와 같은 관용구의 특성에 착안해, 본 논문에서는 단어의 의미 그대로 직역된 문장에서 핵심적인 단어를 블랙 리스트에 추가하여 기계번역기가 관용구를 번역할 시 자주 발견되는 '직역으로 인한 번역 오류'를 탐지할 수 있는 방안을 고안하였다. 기본적으로는 직역으로 인한 오역 단어를 블랙리스트에 추가하지만, 하나의 관용구 당 블랙리스트 단어를 1 개 이상 5 개 이하로 제한하여 블랙리스트의 크기를 조정하였다. 만일 관용구를 기계번역한 영어 문장에 블랙리스트에 해당하는 단어가 1 개 이상 포함된다면 오역으로 평가한다. 본 논문에서는 KISS 에 쓰인 420 개의 관용구에 대한 블랙리스트만을 구축하였다.

표 2. 한-영 관용구 블랙리스트 예시

| 관용구 | 블랙리스트 |
|----------|--------|
| 유명을 달리하다 | famous |
| 눈 높다 | eye |
| 운을 떴다 | lucky |

표 3. 블랙리스트를 이용한 오역 탐지

| | |
|---------|---|
| 관용구 | 눈이 높다 |
| 블랙리스트 | eye |
| 한국어 원문 | 나는 여자 보는 눈이 높아요. |
| 영어 번역 쌍 | I have high standards for women. |

b. 블랙리스트를 활용한 기계번역 성능 평가

이러한 과정을 거쳐 블랙리스트를 구축한 뒤 KISS 속 직역으로 인한 오역을 블랙리스트를 통해 검출, 최종적으로 275개의 3,461개 번역쌍 데이터로 KISS를 정제하였으며, 이 정제 데이터를 국내 주요 기계번역 서비스인 Google Translator, 네이버 Papago와 Kakao i 번역에 돌려 BLEU 스코어와 블랙리스트 평가로 해당 기계번역 서비스의 번역 품질을 평가하였다.

표 4. 기계번역 서비스 번역 품질 평가

| | Google 번역 | Naver Papago | Kakao i |
|------------|-----------|--------------|--------------|
| 블랙리스트 탐지 | 1,179 | 1,093 | 1,049 |
| 블랙리스트 미탐지 | 2,282 | 2,368 | 2,412 |
| 번역 정확도 (%) | 65.93 | 68.41 | 69.69 |
| 평균 BLEU 점수 | 30.04 | 13.47 | 33.83 |

네이버 Papago의 경우 다른 번역 서비스와 비슷한 수치의 번역 정확도를 기록하고 있지만 평균 BLEU 점수는 현저히 낮음을 확인할 수 있다. 반면 블랙리스트 번역 품질 평가 결과의 경우 Papago의 번역이 다른 기계번역과 유사한 정도로 블랙리스트 테스트를 통과하고 있다. 이러한 결과를 바탕으로 블랙리스트 번역 품질 평가 방안이 BLEU 스코어를 보완함을 알 수 있다.

2.2 한-영 관용구 기계번역을 위한 NMT 학습 방법

본 논문은 2.1 선행연구의 연장선으로 한-영 관용구 기계번역에 특화된 데이터셋을 이용하여 신경망 기계번역 모델에 관용구를 효과적으로 학습시키기 위해 특정 토큰을 삽입하여 문장에 포함된 관용구의 위치를 나타내는 방법을 제안한다. 실험 결과, 제안한 방법을 이용하여 학습하였을 때 대부분의 신경망 기계 번역 모델에서 관용구 번역 품질의 향상이 있음을 보였다.

2.2.1 관용구 데이터셋 구성

KISS 7500개 쌍에서 블랙리스트를 이용하여 오역을 제거한 3,377개의 번역 쌍으로 구성된 관용구 번역 쌍 데이터셋을 구축하여 실험에 이용한다. NMT 모델의 학습을 위해 AI Hub에서 제공하는 한-영 번역(병렬) 말뭉치 데이터를 사용하였으며, KISS를 제외하고 중복을 제거한 나머지 AI Hub 데이터를 함께 이용하였다.

표 5. 데이터셋 문장 개수

| | 관용구 포함 문장 쌍 | 관용구 미포함 문장 쌍 |
|-----|-------------|--------------|
| 개발셋 | 500 | 5,000 |
| 학습셋 | 1,877 | 1,577,426 |
| 평가셋 | 1,000 | 10,000 |
| 합계 | 3,377 | 1,592,426 |

표 5와 같이 관용구가 포함된 문장, 관용구가 포함되지 않은 일반 문장으로 각각 나누어 학습 및 평가를 위한 데이터셋으로 분할하고, 이후 표 6처럼 분할한 데이터셋에 실험 조건 별로 다르게 전처리를 적용하여 실험을 수행하였다. 문장 내의 관용구 위치를 나타내기 위해 특수한 토큰 '<idm>'을 관용구 앞에 부착하는 방법을 제안하였다.

1. 관용구가 포함되지 않은 문장으로만 이루어진 학습셋: 1,577,426 문장 쌍
2. 관용구를 포함하는 문장과 관용구가 포함되지 않은 문장으로 이루어진 학습셋: 1,577,426 + 1,877 문장 쌍
3. 2에서 문장 내 관용구의 시작 위치를 <idm> 으로 표기한 학습셋
4. 2에서 문장 내 관용구의 시작과 끝 위치를 각각 <idm>, </idm> 으로 표기한 학습셋

표 6. 관용구 학습 데이터셋 예시

| | |
|---|----------------------------|
| 1 | 관용구 미포함 |
| 2 | 발을 동동 구르는 상황이다. |
| 3 | <idm>발을 동동 구르는 상황이다. |
| 4 | <idm>발을 동동 구르는</idm> 상황이다. |

2.2.2 실험 및 결과

구축한 데이터셋에 형태소 분석과 Byte Pair Encoding 32,000회를 적용하였다. 다양한 NMT 모델로 실험하기 위해 OpenNMT를 이용하였고, NMT 모델은 총 4가지로 OpenNMT의 기본 모델인 2-layer LSTM, 4-layer LSTM, 그리고 4-layer Transformer, 6-layer Transformer 모델이다.

표 7. 실험 결과 BLEU 점수 비교

| NMT 모델 | 2 Layer LSTM | | 4 Layer LSTM | | 4 Layer Transformer | | 6 Layer Transformer | |
|-----------------------------------|---------------|----------------|---------------|----------------|---------------------|----------------|---------------------|----------------|
| 데이터셋 | 관용구 포함 평가셋 | 관용구 미포함 평가셋 | 관용구 포함 평가셋 | 관용구 미포함 평가셋 | 관용구 포함 평가셋 | 관용구 미포함 평가셋 | 관용구 포함 평가셋 | 관용구 미포함 평가셋 |
| 관용구 미포함 학습 | 24.45 | 30.32 | 26.05 | 32.76 | 28.28 | 33.58 | 28.37 | 34.45 |
| 관용구 포함 학습 | 24.37 | 30.23 | 26.45 | 32.63 | 29.09 | 34.09 | 28.73 | 34.22 |
| 관용구 포함 학습 + <idm> 태그 | 24.32 | 30.16 | 26.76 | 36.62 | 29.46 | 34.71 | 28.42 | 34.02 |
| 관용구 포함 학습 + <idm> </idm> 태그 | 24.51 | 30.00 | 25.89 | 32.68 | 27.63 | 33.88 | 27.16 | 33.88 |

실험 결과 문장 내에 토큰을 삽입하여 관용구 위치를 표기하였을 때 관용구가 포함된 문장의 번역 성능이 대부분의 NMT 모델에서 향상되었다. 특히, 관용구 포함 학습 + <idm> 태그 사용에 4 Layer Transformer 모델을 적용했을 경우에 모든 평가셋에서 가장 좋은 성능을 보였다. 그리고 관용구에 <idm> 태그를 시작 위치에만 표시하였을 때에 가장 높은 번역 성능을 보였다.

표 8. 관용구 학습 방법에 따른 번역 결과 비교

| | 관용구 학습 방법 | 번역 결과 |
|----|--------------------------------|--|
| 원문 | 학부모만 발을 동동 구르는 상황이다. | only the parents are anxious . |
| 1 | 관용구 미포함 학습(Baseline) | only parents are rolling their feet . |
| 2 | 관용구 포함 학습 | only parents are rolling their feet . |
| 3 | 관용구 포함 학습 + <idm> 태그 | only parents are struggling . |
| 4 | 관용구 포함 학습 + <idm> </idm> 태그 | only parents are jumping . |

관용구를 효과적으로 학습시키기 위해 표 8과 같이 <idm> 토큰을 삽입하여 관용구의 위치를 표

시켰다. 번역 결과를 비교했을 때 관용구의 시작과 끝 위치를 모두 표기할 경우보다 시작 위치만을 표기할 경우 번역 정확도가 가장 크게 증가함을 알 수 있다.

2.3 Classification of Idiomatic Sentences Using AWD-LSTM

2.3.1 개요

이번 절에서는 본 프로젝트에 적용한 관용구 분류기를 제안한 선행 연구에 대해 소개한다. 관용구란 '두 개 이상의 단어로 이루어져 있으면서 그 단어들의 의미만으로는 전체의 의미를 알 수 없는, 특수한 의미를 나타내는 어구'이다. 관용구를 구성하는 단어들의 의미와 달리, 새로운 의미를 만들어내기 때문에 이를 문장 안에서 자동으로 인식하고 의미를 파악하는 것은 난이도 높은 작업이다. 더불어, 실생활에서 관용구 활용도가 높아 기계 번역 시스템, 대화 시스템, 정보 검색과 같은 자연어 처리 응용 분야에서도 관용구 인식은 해결해야 할 숙제이다. 해당 논문은 주어진 텍스트가 관용구 문장인지, 비관용구 문장인지 분류하는 모델을 제안한다. 또한 일반적인 도메인의 데이터셋으로 사전 학습시킨 모델인 ULMFiT(Universal Language Model Fine-Tuning)을 통해 적은 관용구/비관용구 데이터와 LSTM 모델 중 하나인 AWD-LSTM(ASGD weight-dropped LSTM)을 사용하여 좋은 성능을 이끌어내었다.

2.3.2 전이 학습을 통한 관용구 분류기 구축

전이 학습이란, 학습 데이터가 부족한 분야의 모델 구축을 위해 데이터가 풍부한 분야에서 훈련된 모델을 재사용하는 학습 기법이다. 이 경우, 보통 특정 딥러닝 모델을 사용하는 것보다, 더 적은 양의 학습 데이터를 사용하면서 학습 속도도 빠르고 우수한 성능을 발휘한다. 본 장에서는 대용량 위키텍스트를 가지고 사전학습시킨 ULMFiT 모델을 사용하여 AWD-LSTM 기반 관용구 분류기 모델 개발 과정을 소개한다.

a. ULMFiT

사전 학습은 밑바닥부터 데이터를 학습시키지 않고, 어느 정도 학습이 완료된 가중치 값으로부터 새로운 데이터를 학습시키겠다는 개념에서 착안되었다. 이러한 과정을 통해 효과적으로 레이어를 쌓아서 여러 개의 은닉층도 효율적으로 훈련할 수 있다. 비지도 학습이 가능하기 때문에 레이블이 없는 큰 데이터를 넣어 훈련시킬 수 있다는 것이 큰 장점이다. 또한 모델을 통해 내가 원하는 다운스트림 과제에 적용할 때, 적은 데이터 양으로도 뛰어난 성능을 보여줄 수 있다. ULMFiT는 2018년에 공개된 사전 학습 모델로, 1억 3천만 어절로 이루어진 위키텍스트("Wikitext-103")를 학습 데이터로 사용하였다.

b. AWD-LSTM

사전 학습된 ULMFiT 모델을 통해 우리가 원하는 과제에 적용하는 과정을 '파인튜닝(fine-

tuning)’ 이라 한다. 다시 말해, 파인튜닝이란, 기학습된 모델을 기반으로 아키텍처를 새로운 데이터에 맞게 변형하고 이미 학습된 모델의 가중치로부터 학습을 업데이트하는 방법을 말한다. 본 논문에서는 파인튜닝을 위한 학습 데이터로 TroFi 데이터셋과 자체 구축한 데이터셋을 이용하였고, AWD-LSTM 모델을 사용하여 관용구 분류기를 구축하였다. AWD-LSTM은 기존 LSTM 모델에서 장기 의존성을 강화한 모델로, 모든 은닉층에서 드롭아웃을 적용하고, 확률적 경사 하강법과 가중치 규제 방법을 활용하였다.

c. 관용구 분류기 구축 과정

논문에서 제안하는 전체적인 학습과정은 다음과 같으며 그림 2는 이를 도식화하여 나타낸 것이다.

1. 관용구/비관용구 데이터셋(TroFi Dataset, 자체 데이터셋)을 학습데이터로 사용
2. FastAi 라이브러리를 이용하여 전처리 진행
3. 위키텍스트(‘Wikitext-103’) 데이터로 ULMFiT 사전학습 언어 모델 구축
4. 에서 언급한 학습데이터로 파인튜닝 시도
5. AWD-LSTM을 이용한 관용구 분류기 모델 구축
6. 학습한 모델의 성능 평가. 관용구가 문장에 포함되면 ‘positive(1)’, 포함되지 않으면 ‘negative(0)’을 리턴

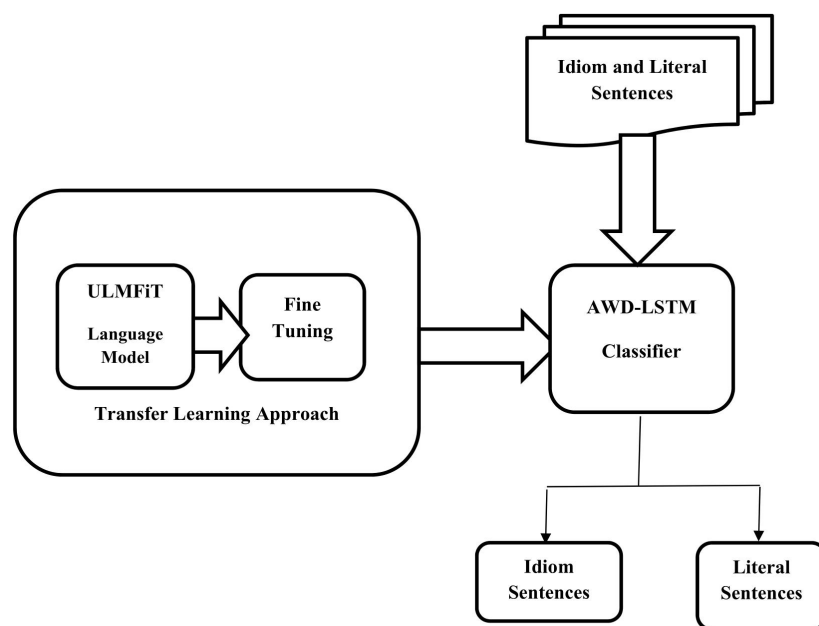


그림 2. 관용구 분류기 아키텍처

2.3.3 관용구 문장 데이터셋(TroFi/자체 데이터셋)

파인튜닝을 위한 관용구 문장 데이터셋으로는 TroFi 데이터셋, 연구실 자체 데이터셋을 이용하

였다. TroFi 데이터셋은 월 스트리트 저널 기사를 크롤링한 데이터로 관용구 문장 2,145개 + 비관용구 문장 1,592개로 이루어져 있다. 자체 데이터셋은 관용구 문장 600개, 비관용구 문장 400개로 구성되어 있다. 표 9는 TroFi 데이터셋의 예시를 보여준다.

표 9. TroFi Dataset Sample

| Id | Text | Label |
|----|---|-------|
| 0 | The debentures will carry a rate that is fixed but can increase based on natural gas prices | neg |
| 1 | Last year the movies were filled with babies | pos |
| 2 | Other magazines may survive five, 10, even 25 or 50 years and then die | pos |
| 3 | It actually demonstrated its ability to destroy target drones in flight | neg |
| 4 | Ever since, Banner has been besieged by hundreds of thrill-seeking callers | pos |

a. 실험 결과

각 데이터의 성능 평가를 위해 F1-score 지표를 사용하였다. F1-score는 정밀도와 재현율의 조화평균으로 각 레이블에 해당하는 데이터 수가 불균형 할 때 주로 사용하는 평가 지표이다. 본 논문에서는 TroFi 데이터셋, 자체 데이터셋에 대해 각각 성능 평가를 진행하였다. 각 데이터셋에 대한 F1-score는 TroFi 데이터셋의 경우, 0.814, 자체 데이터셋의 경우 0.859를 기록하였다.

자동적으로 관용구를 찾아 문맥 안에서의 의미를 파악하는 것은 기계 번역, 정보 검색, 질의응답 시스템과 같은 보다 정확한 자연어처리 어플리케이션을 구축하는 데 도움이 될 것으로 보인다. 따라서 본 논문에서는 이전 학습과 AWD-LSTM을 이용한 관용구/비관용구 문장 분류 모델을 제안하였고 신경망 학습을 통해서도 관용구와 비관용구를 일정 성능 이상 분류할 수 있다는 가능성을 확인하였다.

표 10. 관용구 분류기 평가 결과

| Dataset | Count | Idiom + Literal | Precision | Recall | Accuracy | F1-Score |
|----------|-------|-----------------|-----------|--------|----------|----------|
| Tro-fi | 3737 | 2145 + 1592 | 0.763 | 0.873 | 0.768 | 0.814 |
| In-house | 1000 | 600 + 400 | 0.87 | 0.82 | 0.85 | 0.859 |

2.4 선행연구 분석

2.4.1. 관용구 기계번역 선행 연구의 한계 및 개선사항

기존 연구는 블랙리스트 검출 기법을 제시하였으며, 이를 통해 AI Hub 한-영 병렬 말뭉치에서 기계번역 학습에 사용할 수 있는 관용구 데이터셋을 구축하였고, 관용 표현 주변에 <idm> 토큰을 표시함으로써 관용구 기계번역기의 성능을 높였다.

그러나 기존 연구의 블랙리스트 검출 기법을 사용하기 위해서는 블랙리스트의 구축이 선행되어야 하고, 번역기의 성능을 높이기 위해서는 관용 표현 주변에 <idm> 토큰을 표시해야 하는 작업이 필요하다. 이러한 작업들은 전부 수동으로 이루어진다. 그러므로 논문에서 제시한 방법에는 상당한 시간적 비용이 발생한다는 한계가 있다.

이에 본 프로젝트에서는 블랙리스트 구축을 거치지 않고 관용구를 판별해내어 데이터셋을 구축하는 데 수동 작업을 최소화하려고 한다. 이를 위해 2.3 선행연구를 참고한 관용구 분류기를 이용한다. 관용구 분류기는 관용구 문장과 비관용구 문장을 판별하여 관용구 데이터셋을 빠르게 대량으로 구축하는데 도움이 된다. 또한 기존 연구에서 관용 표현 주변에 표시했던 <idm> 토큰의 위치를 관용구 문장의 맨 앞으로 변경, 해당 문장들로 번역기를 학습하여 논문의 결과와 성능을 비교하려 한다. <idm> 토큰이 문장 맨 앞으로 변경될 시 부착을 자동화할 수 있으므로 새로운 데이터셋 구축시 시간 절감 효과를 기대할 수 있다.

2.4.2 관용구 분류기 선행 연구의 한계 및 개선사항

기존 논문은 사전학습 모델인 ULMFiT와 AWD-LSTM 분류 모델을 이용하여 영어 관용구를 분류하는 방법을 제안했고, Accuracy는 약 76~84%, F1-Score는 약 81~85% 성능을 기록하였다.

영어 관용구 분류기가 사람의 수작업을 대체하기 위해서는 적어도 수작업을 통한 결과물과 최대한 유사해야 한다. 하지만 인공지능 분야에서 100%의 정확도를 내는 건 한계가 있기에, 이러한 관점에서 관용구 분류기를 차용하기에는 신뢰도에 대한 문제가 발생한다. 또한 선행 연구는 영어 관용구 데이터를 이용했기 때문에 한국어 관용구 분류에 대해서는 영어 관용구 분류만큼의 성능을 확보할 수 있을지 실험을 통한 검증이 필요하다.

따라서 본 프로젝트에서는 분류기의 성능 향상과 한국어 대한 적용 가능성을 파악하여 이를 관용구 데이터 증축에 활용하고자 한다. 이를 위해 논문에서 사용한 ULMFiT 모델보다 이후에 공개된 BERT 언어 모델 기반으로 영어 관용구 분류 성능을 비교한 후, 대용량 한국어 데이터로 학습된 사전학습 모델들을 이용하여 한국어 관용구 분류 성능을 확인할 것이다. 이 결과를 기반으로 가장 뛰어난 성능을 보이는 모델을 분류기의 베이스로 활용하려 한다.

3. 프로젝트 개발 결과

이번 장에서는 관용구 번역기 구현을 위한 세부 연구 내용을 기술한다. 관용구 특화 번역기를 만들기 위해서는 관용구 문장에 대해 번역이 잘 되도록 번역기를 어떻게 잘 학습시킬 것인가, 양질의 관용구 병렬 데이터셋을 어떻게 확보할 것인가가 주요 핵심 과제이다. 따라서 본 연구팀은 선행 연구의 <idm> 토큰을 활용하여 번역 성능을 향상시키고, Idiom Classifier를 통해 관용구 데이터를 확보하고자 한다.

3.1 프로젝트 Flow

본 프로젝트의 목표 달성을 위해 관용구 분류 작업, 번역기 구현 작업으로 태스크를 분류하여 진행하였다. 관용구 분류 작업의 경우, SKTBrain에서 발표한 KoGPT2 사전학습 모델을 가지고 한국어 관용구 데이터셋으로 파인튜닝하여 관용구 분류기를 제작한다. 이후 한국어 뉴스 데이터셋인 KCC940을 사용하여 데이터 중 관용구가 포함되어 있는 문장들만 추출한다. 구축한 관용구 문장 데이터셋에는 한글 원문만 존재하기 때문에, 이를 AI Hub 한-영 번역 병렬 말뭉치로 학습시킨 openNMT ver1에 입력값(Source Sentence)으로 넣어 출력값(Target Sentence)을 생성한다. 이렇게 얻어낸 한-영 병렬 데이터를 openNMT ver1에 추가 학습하여 최종 기계 번역 모델을 개발한다. 이후 이 모델을 배포하여 사람들이 실사용할 수 있는 데모 사이트를 제작한다.

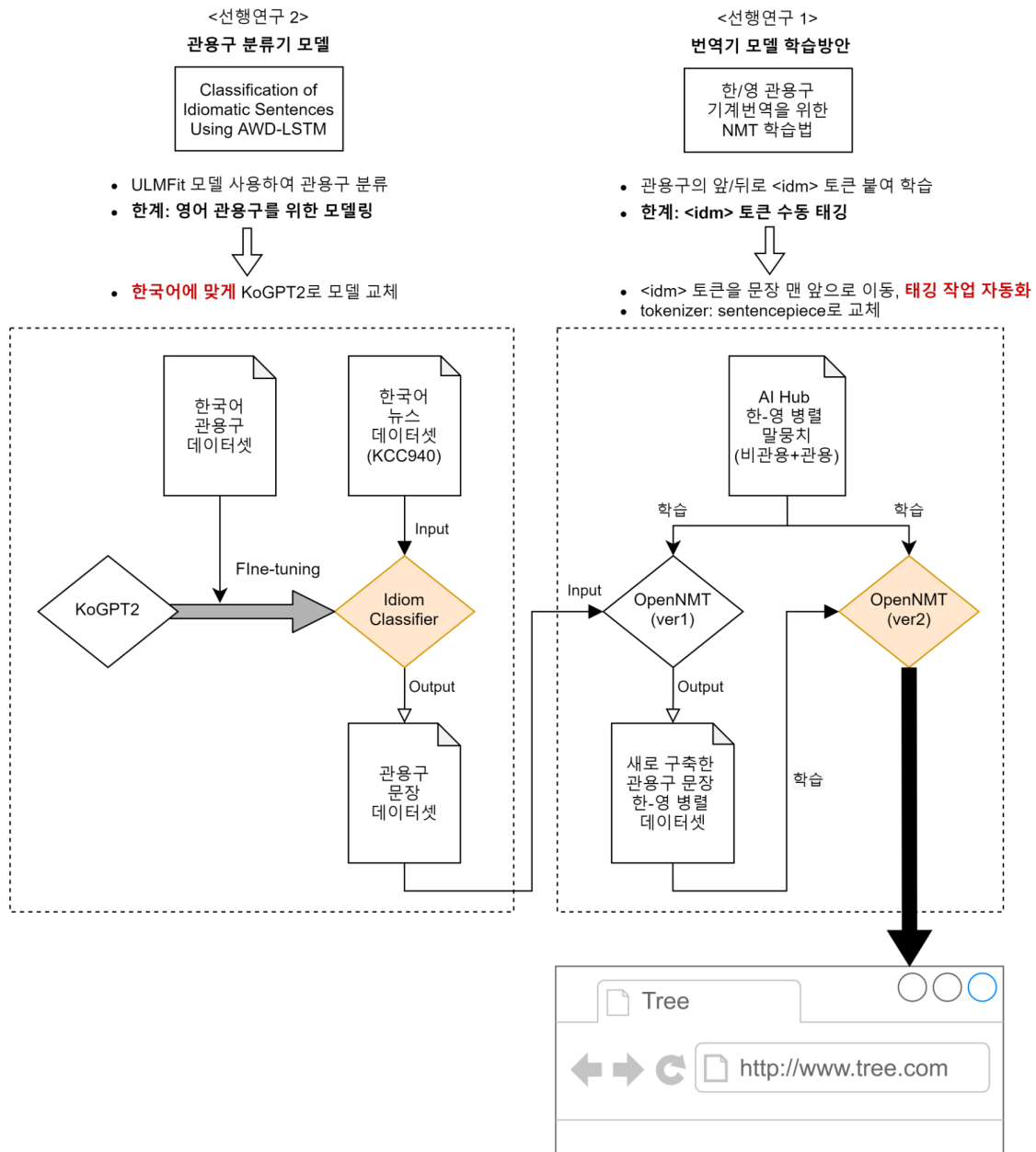


그림 3. 프로젝트 Flow Chart

3.2 데이터셋 증축 작업

논문에서 쓰여진 데이터 3,376개의 관용구 이외의 오역 및 직역으로 쓰여지지 않은 3,979개의 관용구에 대하여 직접 의역을 해주어 양질의 관용구 데이터를 추가로 확보하였다.

3.2.1 의역을 하는 이유

관용구 기계번역을 학습 데이터셋을 구축하는 데는 두 가지 어려움이 있다. 첫째, 관용구를 포함한 문장 자체를 구하는 일이다. 둘째, 관용구 표현에 대한 올바른 영어 번역 문장을 얻어야 한다는 것이다. 본 프로젝트에서는 이 두 가지 어려움을 최소화하기 위해 선행연구에서 KISS 데이터를 정제할 때 제외된 데이터를 살리는 방안을 구상하였다. 직역으로 인해 이용되지 못한 기존의 데이터를 올바르게 번역, 병렬 코퍼스를 추가 구축한다.

3.2.2 작업 내용(작업 방식)

KISS 데이터 7,500개 중 직역 또는 오역 및 직역으로 제외된 데이터는 총 4,124개이다. 이 때 제외된 데이터를 살리되, 영어 문장을 직접 올바르게 번역하는 데는 시간 및 전문성의 한계가 있어, 다음과 같은 방법을 구상하고 실행하였다.

의역을 통한 관용구 병렬 데이터셋 구축 작업 과정은 다음과 같다.

1. 가지고 있는 관용구 리스트 420개에 대해 국립국어대사전을 이용해 정리한 관용구의 의역된 뜻을 참고한다.
2. 원문 내의 관용구를 파파고가 번역할 수 있도록 수작업으로 의역한다.
3. 의역한 한국어 문장을 파파고 API로 번역하여 영어 문장을 얻는다.
4. 번역된 영어 문장과 의역 이전 한국어 문장을 병렬 데이터로 사용한다.

3.3 관용구 분류기

이번 절에서는 관용구 분류기(Idiom Classifier)에 대한 소개와 구축 과정을 서술한다.

3.3.1. 관용구 분류기 모델 선정

본 프로젝트에서 관용구 분류기는 번역기 학습에 필요한 관용구 데이터 구축 과정에 투여되는 수작업을 줄이기 위해 존재하며, 분류기의 입력 데이터로 주어진 말뭉치로부터 관용구 문장과 비관용구 문장을 분류해내는 것이 주된 목표이다. 관용구 문장을 분류하기 위해서는 모델의 성능이 절대적으로 중요하다고 판단했으며, 성능을 높이기 위해 다음의 단계를 거쳐 테스트를 진행했다.

1. 영어 관용구 분류기 성능 비교
2. 한국어 관용구 분류기 모델별 성능 비교
3. 관용구 분류기 모델 선정

위 과정의 Step 별 진행 내용 및 평가 결과는 다음과 같다.

a. 영어 관용구 분류기 성능 비교

우선 영어 관용구 문장에 대해 2장에서 기술한 관용구 분류기 제안 논문의 ULMFiT 모델과 BERT(Bidirectional Encoder Representations from Transformers) 언어 모델의 성능을 비교했다. BERT는 구글에서 개발한 자연어처리(Natural Language Processing, NLP) 언어 모델이며, 최근 여러 자연어 처리 분야에서 기존 딥러닝 모델보다 높은 성능을 내고 있는 사전 학습 모델이다. 또한 BERT는 ULMFiT 이후 제안된 모델이기 때문에 관용구 분류에 더 높은 정확도를 나타낼 것이라 판단하였다. 논문에서 사용한 것과 동일한 TroFi 데이터셋을 이용하여 ULMFiT와 BERT 모델의 성능을 비교한 결과는 아래 표와 같다.

표 11. 영어 관용구 분류기 성능 비교

| Model | Accuracy |
|-----------------------------------|----------|
| ULMFiT + AWD-LSTM | 0.768 |
| BERT (base-multilingual-cased) | 0.844 |

영어 관용구 분류기 성능 비교를 통해 BERT 모델이 관용구 분류 태스크에서 더 좋은 성능을 보이며, 따라서 관용구 분류에 좀 더 적합한 모델임을 확인했다.

b. 한국어 관용구 분류기 모델별 성능 비교

다음으로 한국어 관용구에 대해서도 사전 학습 모델 기반의 분류기가 영어 관용구만큼 높은 정확도를 갖고 분류할 수 있는지 실험했다. 성능 비교 대상 모델은 분류 태스크에 적합하면서 기학습된 사전학습 모델인 KoBERT, BERT-Multilingual, KoGPT2를 선정했다. 성능 비교를 위해 AI Hub 한-영 병렬 말뭉치에서 추출한 관용구, 비관용구 문장을 1 대 1 비율로 구성하였으며, 학습 데이터로 5,401개, 검증 데이터는 400개, 시험 데이터로 1,351개 문장을 사용했다. 시험 데이터를 가지고 각 모델별 성능 평가를 해본 결과는 다음 표와 같다.

표 12. 한국어 관용구 분류기 성능 비교1

| Model | Accuracy | Precision | Recall | F1 Score |
|----------------------|----------|-----------|--------|----------|
| KoBERT | 0.928 | 0.973 | 0.889 | 0.925 |
| BERT Multilingual | 0.965 | 0.965 | 0.97 | 0.97 |
| KoGPT2 | 0.968 | 0.964 | 0.976 | 0.972 |

이후 2차 검증을 위해 AI Hub 한-영 병렬 말뭉치에 포함되지 않은 문장으로 구성된 새로운 테스트셋(관용구 50문장 + 비관용구 50문장)을 통해 다시 한번 성능 평가를 진행하였다. 관용구 문장은 국립국어원 우리말샘에서 제공하는 관용구 용례 문장을 가져왔으며, 비관용구 문장은 뉴스 기사로부터 추출하였다.

KoBERT의 경우, 크게 두 가지 방식으로 모델을 학습하였다. 첫 번째는 사전학습 모델을 Freeze 한 상태로 마지막 FFN(Feed Forward Neural Network)만 학습시키고나서, 작은 학습률로 모델 전체를 학습시키는 방식이며, 두 번째는 처음부터 모델 전체를 학습시키는 방식이다. KoGPT2의 경우 사전학습 모델을 Freeze시켜 진행한 첫 번째 방식만 수행했다. 그 결과, 아래 표와 같이 KoGPT2 모델을 사용했을 때 분류 성능이 가장 높은 것을 확인했다.

표 13. 한국어 관용구 분류기 성능 비교2

| Model | Accuracy | Precision | Recall | F1 Score | Freeze |
|------------------------|----------|-----------|--------|----------|--------|
| KoBERT | 0.92 | 0.92 | 0.92 | 0.92 | |
| KoBERT | 0.94 | 0.94 | 0.94 | 0.94 | ✓ |
| BERT (Multilingual) | 0.95 | 0.93 | 0.94 | 0.94 | |
| KoGPT2 | 0.95 | 0.98 | 0.96 | 0.97 | ✓ |

c. 관용구 분류기 모델 선정

최종적인 한국어 관용구 분류기 모델을 선정하기 위해 세 가지 경우로 나누어 실험을 진행했다.

1. 'KISS+' + 비관용구 7,355개
2. AI Hub 한-영 병렬 말뭉치로부터 추출해낸 관용구 문장 18,808개 + 비관용구 문장 18,808개

3. AI Hub 한-영 병렬 말뭉치로부터 추출해낸 비관용구 문장 30,000개 + oversampling을 통해 얻은 관용구 문장 30,000개

한국어 관용구 분류기의 성능을 최대한 끌어 올리기 위해서는 데이터의 질과 양이 중요하다. 1)은 의역 작업을 거친 'KISS+' 데이터로 수작업을 통해 관용구 문장을 선별했기 때문에 품질이 보장된다. 2)는 2.1 논문에서 구축한 420개의 관용구가 포함된 문장 18,808개를 학습에 이용했다. 1)번 과정에서 제거한 비관용구 문장이 18,808개에 포함된 만큼 상대적으로 품질은 떨어진다고 판단했다. 3)은 oversampling 기법을 사용하여 데이터의 분포를 맞춰준 다음 학습을 진행했다.

1) 학습에 사용한 데이터는 관용구 문장 5,855개와 비관용구 문장 5,855개이며, 검증 데이터는 각 카테고리에서 1,500개씩 랜덤으로 추출했다.

표 14. 1)에 대한 한국어 관용구 분류기 성능 비교

| Model | Accuracy | Precision | Recall | F1 Score | Freeze |
|------------------------|----------|-----------|--------|----------|--------|
| KoBERT | 0.91 | 0.91 | 0.91 | 0.91 | ✓ |
| BERT (Multilingual) | 0.89 | 0.92 | 0.898 | 0.891 | |
| KoGPT2 | 0.95 | 0.95 | 0.98 | 0.97 | ✓ |

2) 학습에 사용한 데이터는 관용구 문장 17,307개와 비관용구 문장 17,307개이며, 정확한 성능 비교를 위하여 검증 데이터를 1)과 동일하게 사용했다.

표 15. 2)에 대한 한국어 관용구 분류기 성능 비교

| Model | Accuracy | Precision | Recall | F1 Score | Freeze |
|------------------------|----------|-----------|--------|----------|--------|
| KoBERT | 0.952 | 0.952 | 0.952 | 0.952 | ✓ |
| BERT (Multilingual) | 0.918 | 0.914 | 0.921 | 0.916 | |
| KoGPT2 | 0.947 | 0.947 | 0.942 | 0.947 | ✓ |

3) 비교적 높은 성능을 보인 KoBERT와 KoGPT를 대상으로 테스트했으며, 상기한 바와 같이 oversampling 기법을 이용하여 학습을 진행했다. 즉, 관용구 문장 수를 비관용구 문장 수와 동일하게 맞춰 준 뒤 학습하였다. 마찬가지로, 검증 데이터는 1), 2)에서 사용한 것과 동일하다.

표 16. 3)에 대한 한국어 관용구 분류기 성능 비교

| Model | Accuracy | Precision | Recall | F1 Score | Freeze |
|--------|--------------|--------------|--------------|--------------|--------|
| KoBERT | 0.953 | 0.953 | 0.953 | 0.953 | ✓ |
| KoGPT2 | 0.954 | 0.955 | 0.955 | 0.953 | |

두 모델의 성능 차이는 미미하였으나 KoGPT2가 조금 더 나은 성능을 보였고, KoGPT2가 KoBERT에 비해 대량의 말뭉치로 사전학습 된 모델이기에 KoGPT2 모델을 최종 한국어 관용구 분류기 모델로 선정하였다.

3.3.2 한국어 관용구 데이터 증축

관용구 번역기 학습에 사용할 관용구 문장을 추가적으로 구축하기 위해 뉴스 기사 데이터셋을 확보하였다. 이는 AI Hub 한-영 병렬 말뭉치 구성의 50%를 뉴스 기사가 차지하고, 일반적으로 뉴스 기사에서 관용구를 적극 활용하기 때문이다. 그에 맞춰서 국민대학교 자연어처리 연구실에서 공개한 뉴스 데이터셋인 KCC(Korea Contemporary Corpus) 원시 말뭉치를 사용하였다.

a. 최종 선택된 모델을 이용한 관용구 문장 분류

각 뉴스 데이터셋에 대해 관용구 분류 작업을 진행했으며 분류 결과는 아래 표와 같다.

표 17. 분류된 각 뉴스 데이터셋의 관용구 및 비관용구 문장 수

| Data | 전체문장 | 관용구 | 비관용구 |
|--------|------------|---------|------------|
| KCC940 | 6,263,454 | 366,236 | 5,897,218 |
| KCC150 | 11,961,347 | 726,668 | 11,234,679 |
| KCCq28 | 1,337,721 | 132,389 | 1,205,332 |

다음으로 관용구로 분류한 문장에서 표본을 추출하여 실제 관용구가 포함된 문장이 맞는지 확인하는 작업을 진행했다.

표 18. 각 뉴스 데이터셋의 관용구 문장 예시

| | |
|--------|---|
| KCC940 | 14일 강병원 더불어민주당 원내대변인은 이날 현안 브리핑을 통해 "여야가 42일 만에 드디어 국회 정상화의 땀을 흘렸다 "며 이같이 말했다. |
| | 각종 여론조사에서 앞서고 있는 오거돈 전 장관은 긴장의 끈을 놓으려 하지 않았다 . |
| | 같은 당 정진석 의원은 "숙의 민주주의를 좋아하는 이 정부가 숙의는커녕 국무회의 심의도 거치지 않고 개헌안을 발표하고 있다"고 꼬집었다 . |
| | 이날 추 대표는 이재명 후보의 개소식에서 이 후보야 말로 문재인 정부와 호흡을 잘 맞출 책임자라고 추켜세우며 당내 일부 '반 이재명' 정서 차단에 나섰다. |
| | 만나는 보수 정치 원로들마다 허를 차고 있습니다 . |
| KCC150 | 여전히 SK에는 김성근 전 감독의 그림자가 짙게 드리워있다 . |
| | 이근호는 다시 갈림길 앞에 섰다 . |
| | 실제 업계에서도 앓는 소리가 나온다. |
| | 그러나 갈 길은 아직 멀다 ¹ . |
| | <u>정신 차리고 전화 끊으세요</u> ² . |
| KCCq28 | 김정은이 핵폭탄 때문에 국제적 제재가 많이 들어오자, 언제 어떤 형태로 IS와 손을 잡을지 모른다"고 주장했다. |
| | <u>부츠가 발에 딱 맞도록 테이프로 감고 뒀다</u> "며 "그래도 이 부츠로 올림픽에 나갈 예정이다" ² . |
| | 박지성은 인사말을 할 때 "매년 이렇게 여러분을 만나는 이 자리가 행복하다"고 <u>말문을 열었다</u> ² . |
| | 그는 " ' 가지 많은 나무 바람 잘날 없다 ¹ '는 속담도 있는데 겨우 38석의 제3당이 오늘로서 3명의 의원 ' 도살장에 끌려가는 소 신세 ¹ ' 같은 느낌을 받았다"고 말했다. |
| | 감독을 필두로해 중심을 딱 잡고 앞으로 나아가는 모습이 보인다"며 "팀이 안 좋을 때는 잡음이 생기기 마련이다 ¹ . |

각주 1 : 표준국어대사전에 관용구로 등록되어 있는 않지만 관용구로 판단할 수 있는 문장

각주 2 : 관용구가 없는 문장

위 표에서 확인할 수 있듯 관용구로 분류된 각 문장을 일부 샘플링하여 확인해본 결과, kCC940, KCC150, kCCq28 순으로 관용구를 잘 추출한 것으로 나타났다. KCC150과 KCCq28의 경우 관용구가 아닌 문장들이 상대적으로 많이 포함되어 있으며, 속담 표현들도 포함되는 것을 확인할 수 있다. 전체 데이터의 양과 관용구 분류기를 통해 추출된 관용구 문장의 개수는 KCC150이 가장 많으나 추출된 관용구 문장의 품질 측면에서 KCC940이 가장 우수하다. 따라서 본 연구에서는 KCC940 데이터에서 추출한 관용구 문장을 번역기 학습에 사용할 데이터로 선정했다. 이 데이터는 번역기를 통해 병렬 한영 병렬 말뭉치로 만들어질 예정이다.

3.4 관용구 번역기

3.4.1 번역기 성능 향상 대안

a. <idm> 토큰 위치 변경

선행 연구에서 관용구 기계번역 성능 향상을 위해 제안했던 <idm> 토큰 부착 방안을 수정하여, 관용 표현 주변이 아닌 관용구 문장 맨 앞에 <idm> 토큰을 부착하여 번역기를 학습시켰다(6 Layer Transformer). 모델을 5만 번 학습시킨 뒤 논문과 성능을 비교해본 결과는 표 17과 같다.

표 19. <idm> 토큰 위치에 따른 BLEU 스코어 비교

| NMT 모델 | 논문 모델 | 수정 모델 |
|-----------------------|---------|-------|
| <idm> 토큰 위치 | 관용 표현 앞 | 문장 앞 |
| 일반 문장 평가셋 | 34.02 | 36.86 |
| 관용구 문장 평가셋 | 28.42 | 30.26 |
| 일반 문장 + 관용구 문장 평가셋 | 36.19 | 36.19 |

BLEU 스코어 비교 결과, <idm> 토큰을 문장 앞으로 옮긴 수정 모델에서 성능이 향상되는 경향을 보였다. 이에 추후 학습하는 번역기는 <idm> 토큰을 관용구 문장의 맨 앞에 표시하기로 하였다.

b. 토큰나이저 변경

선행연구에서는 KoTagger와 Mecab을 토큰나이저로 사용하였으나, KoTagger가 비공개인 관계로 Predict시 Detokenizing이 불가능하다는 문제가 있었다. 이에 토큰나이저를 Sentencepiece로 교체하였다. Sentencepiece를 사용할 시 Detokenizing에 유리하며, <unk> 토큰 발생을 줄여 prediction 결과 개선을 기대할 수 있다.

3.4.2 번역기 1

본 번역기는 선행 연구에서 학습한 데이터 KISS(Korean-English Idioms in Sentences Dataset)를 최대한 활용하여 번역 성능을 높이고자 한다. 따라서 번역기에 사용된 KISS+ 데이터셋(3.2에서 증축한 관용구 데이터셋)을 학습데이터로 추가하였다. 데이터셋의 문장 개수는 표 20과 같다.

표 20. 데이터셋 문장 개수

| 구분 | 일반 문장 | 관용구 문장 | 일반문장 + 관용구 문장 |
|------|-----------|--------|---------------|
| 학습셋 | 1,564,450 | 5,855 | 1,570,305 |
| 평가셋 | 4,500 | 500 | 5,000 |
| 테스트셋 | 9,000 | 1,000 | 10,000 |
| 전체 | 1,577,950 | 7,355 | 1,585,305 |

3.4.3. 번역기 2

분류기(Idiom Classifier)를 통해 KCC940 뉴스 데이터셋에서 366,236개의 관용구 문장을 추출하였다. Google, Papago, Kakao i 번역기가 아닌 자체학습으로 만든 번역기-1(3.4.2 항목)을 통해 366,236개의 관용구 문장을 영어문장으로 번역하여 한-영 병렬 코퍼스를 구축하였다. 추가로 증축한 데이터는 표 21와 같다.

표 21. 번역기 -2 학습 데이터

| | AI-hub | KCC940 |
|--------|-----------|---------|
| 일반문장 | 1,577,950 | X |
| 관용구 문장 | 7,355 | 366,236 |
| 총 데이터 | 1,951,541 | |

번역기 2는 표 21의 학습데이터를 사용하여 20만번 학습을 한 모델이다. 일반문장과 관용구 문장이 섞여 있는 테스트셋과 관용구 문장만 있는 테스트셋을 이용하여 성능평가를 진행했다. 번역기2의 성능 평가 결과는 다음과 같다.

표 22. 번역기 -2 성능 평가

| Epochs | 테스트셋 | BLEU Score |
|---------|----------------|------------|
| 200,000 | 일반 문장 + 관용구 문장 | 48.53 |
| 200,000 | 관용구 문장 | 39.49 |

3.5 번역 서비스 - Tree

태조가 제안하는 관용구 특화 번역기 서비스 Tree는 관용구 번역에서만큼은 파파고 등 대기업의 기계번역기를 품을 수 있는 성능을 보이겠다는 포부에서 명명되었다. Tree에서는 입력된 한국어 문장을 영어로 번역하는 서비스를 제공할 수 있도록 기획하였다. 번역의 핵심이라 할 수 있는 번역 모델은 실사용 성능을 고려, 6 Layer Transformer를 기반으로 고유 데이터셋을 N번 학습시킨 모델을 채택하였다. 이를 웹으로 구현해 서비스화하는 과정에서 Flask, HTML5, CSS3, Javascript의 기술을 사용하였다.

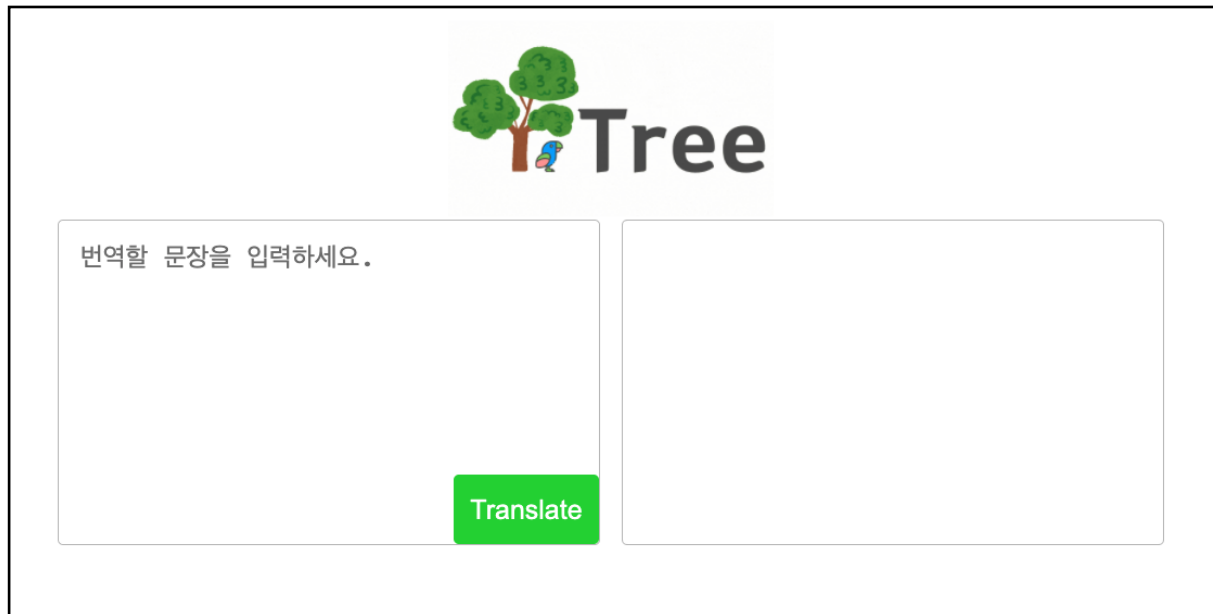


그림 4. 데모 웹 사이트

4. 결론

기계번역 연구에 대한 수요가 날로 늘어나고 있으며, 규칙기반, 통계기반을 거쳐 최근에는 딥러닝 기반의 기계번역 연구가 주로 이루어지고 있다. 그러나 기계번역 분야의 수요 증가 추세에도 불구하고 아직까지 기계번역기의 관용구 번역 성능은 만족스럽지 못한 편이다. 따라서 본 프로젝트는 기계번역에서 관용구 번역 성능을 개선하여 사용자에게 정확한 의미를 전달하는 번역 서비스를 제공하고자 하였다.

이를 위해 본 프로젝트에서는 번역기에 학습시킬 수 있는 양질의 데이터를 확보하는 것을 목적으로 우선 관용구 분류기를 구축하였다. KoGPT2를 이용한 관용구 분류기를 통해 뉴스 데이터셋에서 37만개의 관용구 문장을 추출했다. 그러나 관용구 분류기에서 분류해낼 수 있는 문장은 한국어 뿐이기 때문에 번역기를 학습시키기 위해서는 해당 관용구의 영어 문장을 얻는 과정이 필요하다. 본 프로젝트에서는 오역 이슈가 발생할 수 있는 기존 기계 번역기에 의존하지 않고, 기 확보된 양질의 관용구 병렬 데이터셋으로 자체 번역기를 만들어 관용구의 영어 문장을 얻었다. 이 과정을 통해 관용구에 특화된, 양질의 영문 데이터를 확보했다. 이러한 과정들을 통해 최종적으로 관용구 병렬 말뭉치 37만쌍을 구축하고 번역기 학습에 이용하였다. 그러나 관용구 분류기가 100%의 정확도로 관용구를 분류해 내는 것이 아니기에 분류기를 통해 얻은 관용구 문장의 품질 문제가 상존한다.

번역기의 경우 LSTM 및 Transformer 모델의 설정을 시험해보는 초기 단계를 거쳐 6 Layer Transformer를 기본 모델로 설정하였고, 이후 <idm>토큰을 활용한 학습셋 전처리, 토큰나이저 변경 및 학습 관용구 데이터 증가를 통해 성능을 향상시켰다. 그 결과 선행연구에 비해 관용구 번역 BLEU 스코어는 10점 이상, 관용구와 일반 문장이 섞여 있을 때의 번역 BLEU 스코어는 20점 가까이 향상되었다. 실 활용에서도 기존 번역기와 달리 관용구 문장의 의미를 제대로 번역함을 확인할 수 있었다.

하지만 학습 데이터 문장이 주로 문어체로 이루어져 있기 때문에, 구어체에 대해선 비교적 번역 성능이 떨어짐을 확인하였다. 또한 학습시키지 않은 관용구에 대해서는 오번역 되는 사례가 종종 발견되었다. 전자의 경우, 한-영 병렬 영화 자막 등의 구어체 데이터를 학습에 추가할 시 보완이 가능할 것으로 예상된다. 영화 자막은 전문이 번역을 거쳐 수정되기 때문에 비교적 고품질의 데이터를 확보할 수 있다. 또 구어체 데이터의 학습량을 늘린다면 여러 종결 어미에도 대응할 수 있을 것으로 보인다. 비학습 관용구 오번역 문제의 경우, 국립국어원 표준국어대사전에 수록되어 있는 모든 관용구에 대해 병렬 말뭉치를 구축할 수 있다면 개선 가능할 것으로 예상된다. 이외에도 비지도 학습 기반으로 접근한다면 병렬 말뭉치 구축 없이도 비학습 관용구 오번역에 대한 문제를 해결할 수 있을 것이다.

본 프로젝트에서는 정량적 평가 대신 정성적 비교를 통해 관용구 문장 데이터셋을 선정하였지만 데이터셋의 품질 문제를 보완하기 위해 제안된 방법으로 코퍼스 필터링이 있다. 이는 기계번역의 학습데이터로 쓰이는 병렬 코퍼스의 품질을 높이기 위해 학습데이터의 부적합한 병렬 쌍을 제거하는 방법으로, 데이터 전처리와는 다른 개념이다. 코퍼스 필터링에서는 학습 데이터로 부적합한 문장을 완전 삭제하기 때문에, 코퍼스 필터링을 이용할 경우 고품질의 학습데이터를 구축할 수 있으며 이는 관용구 분류기뿐만 아니라 번역기의 성능향상을 이끌어낼 것이라 예상된다.