Tree

- 파파고를 품은 관용구 영어 번역기 -

태조 (멀티캠퍼스 C반)

김예찬, 김태영, 김택현
심판교, 유현승, 이지수, 진가형

Table of Contents

Project Overview

TIMELINE

08

2~3주차: 주제 브레인 스토밍

3~4주차: 자료 조사

4~5주차: 논문 데이터 탐색 및 연구

09

1~2주차: 논문 구현 및 응용

3~5주차: 데이터셋 구축

10

1주차: 데이터셋 및 분류기 보완

2주차: 번역기 및 분류기 성능 검증,
Prototype 제작

3주차: 데이터셋 증축 성능 보완

4주차: 최종 결과물 완성
(번역 엔진 웹 탑재 및 디자인 보완)

1. 문제 제기

현재 기계번역기는 관용구 번역을 어떻게 하고 있을까?

미국과 중국의 갈등으로 인해 우리나라까지 불똥이 튄다.

1. 문제 제기

- 현재 기계번역기는 관용구를 실제 뜻과 다르게 '직역'하는 경우가 대다수

2. 프로젝트 목표

관용구의 의미를 '제대로' 번역하는 기계 번역기

3. 프로젝트의 난점 및 해결 방안 탐색

- 관용구를 잘 학습시킬 방법?

  From 한-영 관용구 기계번역을 위한 NMT 학습 방법 – 최민주, 이창기 (강원대, 2020)

관용 표현 앞에 <idm> 표시 후 학습 시키기

누가 봐도 박인비의 우승이 <idm> 불을 보듯 뻔한 상황이었다.

3. 프로젝트의 난점 및 해결 방안 탐색

- 관용구 데이터셋 수집 및 증축이 어려움

From <AWD-LSTM을 이용한 관용구의 분류> (Classification of Idiomatic Sentences Using AWD-LSTM)

– J.Briskilal and C. N. Subalalitha (2021)

관용구 분류기가 있다면 데이터셋을 훨씬 빠르게 구축할 수 있을거야!

Project Details

1. Idiom Classifier

관용구 분류기 개발 Work Flow

STEP 1. 관용구 분류기 성능 비교 (영어)

- 최신 사전학습 모델(BERT)이 관용구 분류 태스크에 더 좋은 성능을 보임

<AWD-LSTM을 이용한 관용구의 분류> 논문 기반 구현

ULMFiT VS **BERT**

(영어 관용구 데이터셋 사용)

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

STEP 2. 관용구 분류기 성능 비교 (한국어)

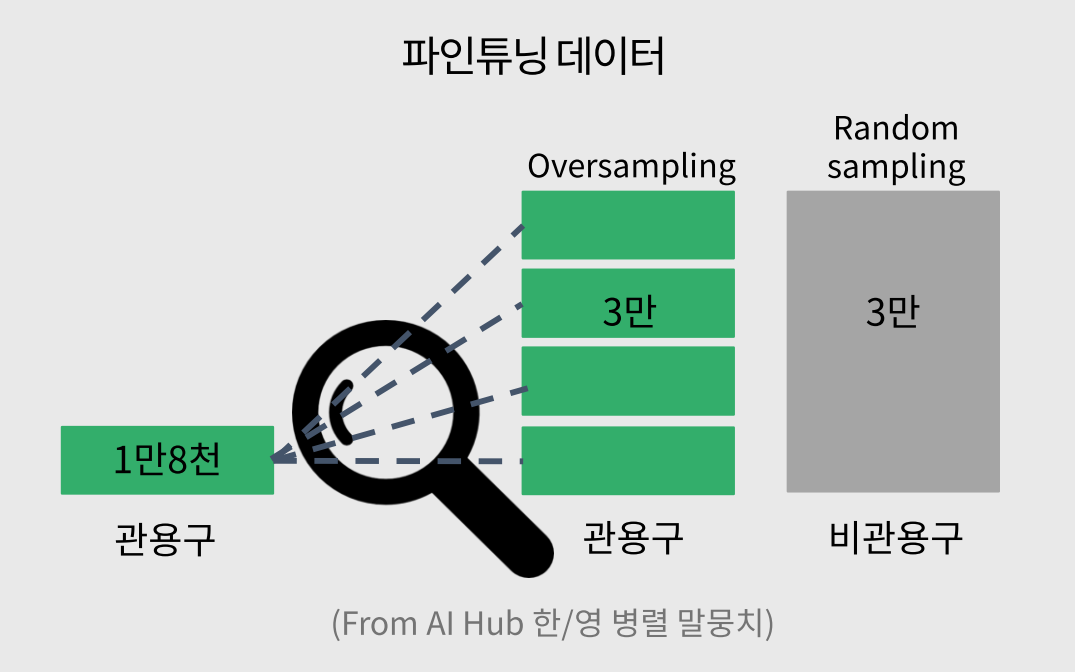
- 영어가 아닌 한국어 관용구 분류에서 90% 이상의 정확도를 보임
- 관용구 분류에 사용 가능하다고 판단!

BERT VS KoGPT2

Model	Accuracy
KoBERT	0.94
BERT Multilingual	0.94
KoGPT2	0.96

STEP 3. 관용구 분류기 모델 선정

- 관용구와 비관용구 비율을 1 : 1로 맞추어 KoBert와 KoGPT2 파인 튜닝
- KoBERT보다 KoGPT2의 사전학습 데이터가 많아 최종 분류기 모델로 선택



KoBERT VS KoGPT2

Model	Accuracy
KoBERT	0.953
KoGPT2	0.954

STEP 4. 한국어 관용구 추출

- 제작한 관용구 분류기를 이용하여 약 37만 개의 관용구 문장 추출

추출 관용구 예시

출처: KCC940 데이터셋(뉴스 기사 문장)

- 각종 여론조사에서 앞서고 있는 오거돈 전 장관은 **긴장의 끈을 놓으려** 하지 않았다.
- 이날 추 대표는 이재명 후보의 개소식에서 이 후보야 말로 문재인 정부와 **호흡을 잘 맞출** 적임자라고 추켜세우며 당내 일부 '반 이재명' 정서 차단에 나섰다.
- 만나는 보수 정치 원로들 마다 **혀를 차고 있습니다**.

2. Idiom Translator

1. 관용구 데이터 증축을 위한 IDEA

- 분류기를 통해 한국어 관용구 약 37만 개를 더 확보했지만 … 동일한 영어 문장이 필요하다!

1. 관용구 데이터 증축을 위한 IDEA

- 관용구 37만 개의 영문 번역을 얻는 방법?

  - 시중 기계번역기로 37만 문장의 영문 번역할 시
    1. 오역 이슈 발생
    2. 금전적 부담(37만 문장 with Papago API …)

  - 그렇다면 자체 미니 번역기를 만들어 영어 문장을 얻어보자!
    1. 일부 데이터를 기계 번역기가 번역할 수 있도록 의역
    2. 의역된 문장을 시중 기계번역기에 넣어 올바른 뜻의 영어 문장 얻기
    3. (의역 전 한글 원문) – (의역의 영번역) 쌍을 1차 병렬 데이터셋으로 구축
    4. 1차 병렬 데이터셋으로 미니 번역기 학습
    5. 미니 번역기에 37만 관용구 넣어 영어 문장 얻기

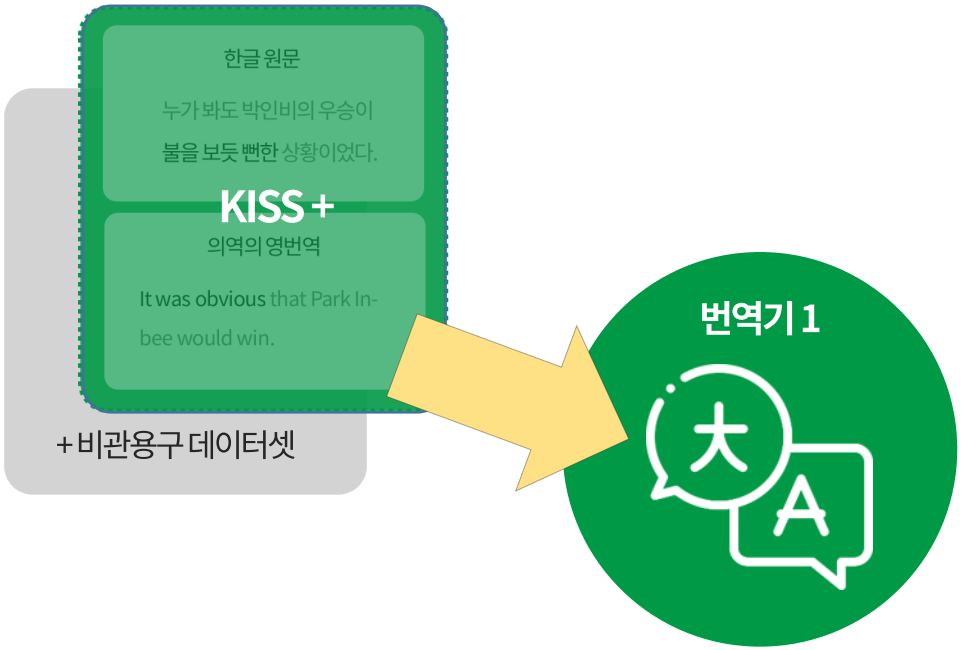
2. 의역 및 관용구 데이터 증축

1. KISS* 데이터셋에서 오역으로 제거된 관용구 데이터 준비
2. 한글 원문 직접 의역 (총 4,124 문장)
3. 의역 문장을 기계번역기에 넣어 1차 관용구 병렬 데이터셋(KISS+)* 구축 (약 7,400 문장)

※ KISS: Korean-English Idioms in Sentences Dataset, 선행 연구 공개 데이터
※ KISS+: KISS의 오역 데이터를 의역 작업으로 수정한 양질의 병렬 데이터셋

2. 의역 및 관용구 데이터 증축

4. KISS+와 비관용구 데이터셋으로 미니 번역기 학습 (번역기 1)



5. 보유한 37만 개의 관용구를 번역기 1에 넣어 영어 문장 얻기

관용구 병렬 데이터 37만 쌍 완성!

3. <idm> 표시 자동화

100% 수작업! 자동화 방법?

- 관용구를 잘 학습시킬 방법?

  From 한-영 관용구 기계번역을 위한 NMT 학습 방법 – 최민주, 이창기 (강원대, 2020)

3. <idm> 표시 자동화

- <idm>의 위치를 문장 맨 앞으로 바꾸어 부착 자동화 + 번역 성능 향상

기존 연구

누가 봐도 박인비의 우승이 <idm> 불을 보듯 뻔한 상황이었다.

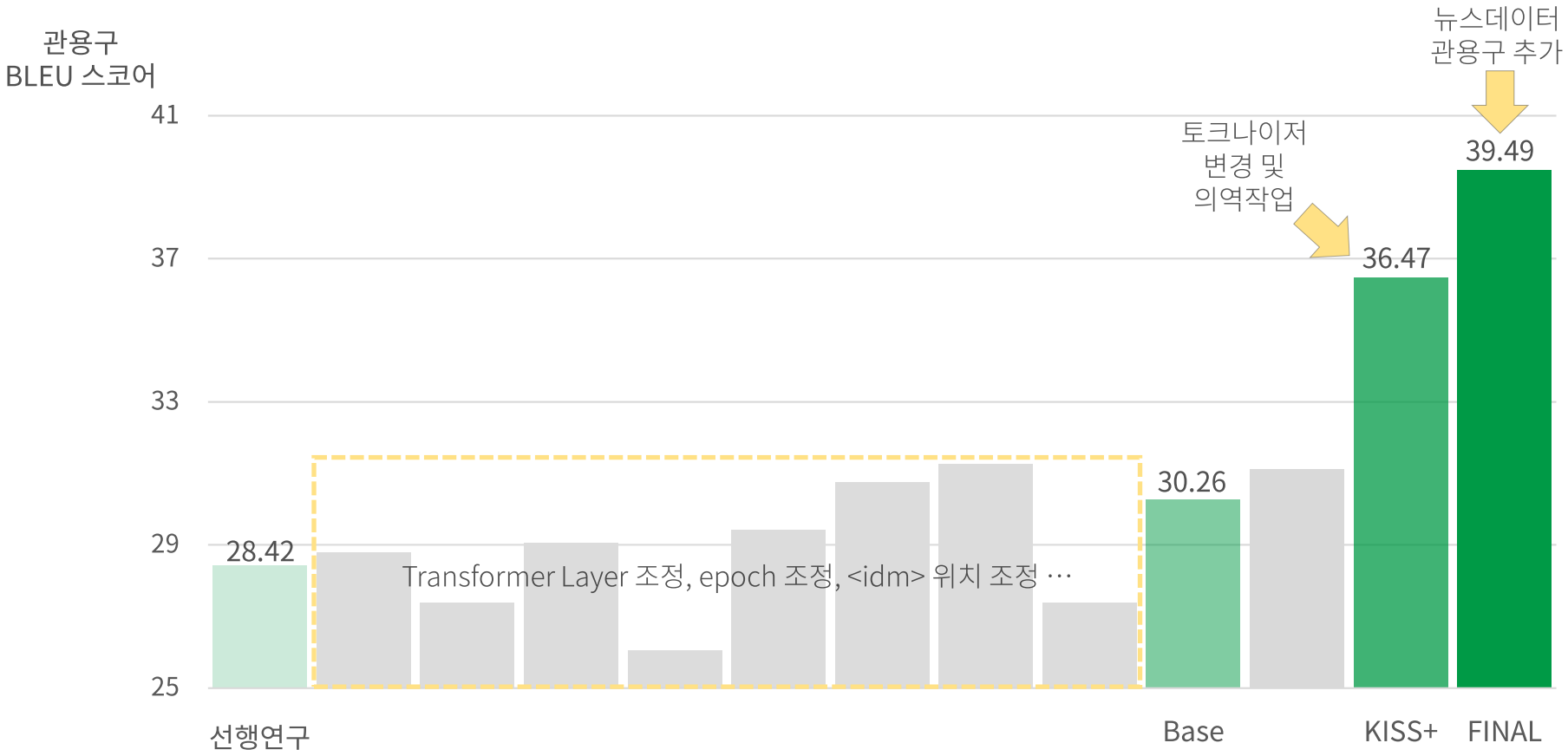
(BLEU 30.26)

변경 후

<idm> 누가 봐도 박인비의 우승이 불을 보듯 뻔한 상황이었다.

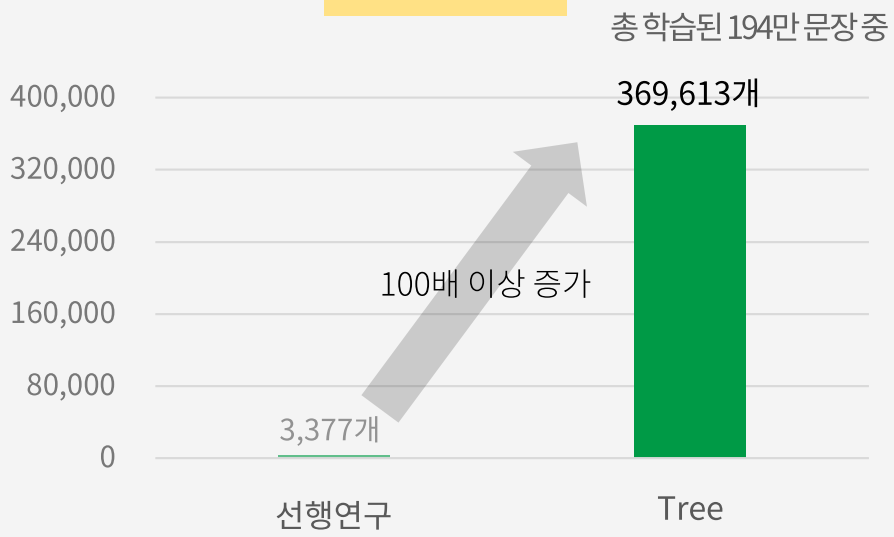
(BLEU 31.11)

4. Tree가 자라온 과정



5. Tree

학습 관용구 수



토크나이저 / epoch

Sentencepiece / 200,000 epoch

(<unk> 토큰 발생 감소 효과)

학습셋 전처리

관용구 문장 맨 앞 <idm>

5. Tree

The conflict between the U.S. and China has ever caused a stir in Korea.

파문을 일으키다,동요시키다

미국과 중국의 갈등으로 인해 우리나라까지 불똥이 튄다.

Google Translator

The conflict between the US and China has sparked a spark in Korea.

'불꽃을 튀기다'의 직역

Papago

The conflict between the US and China sparks even Korea.

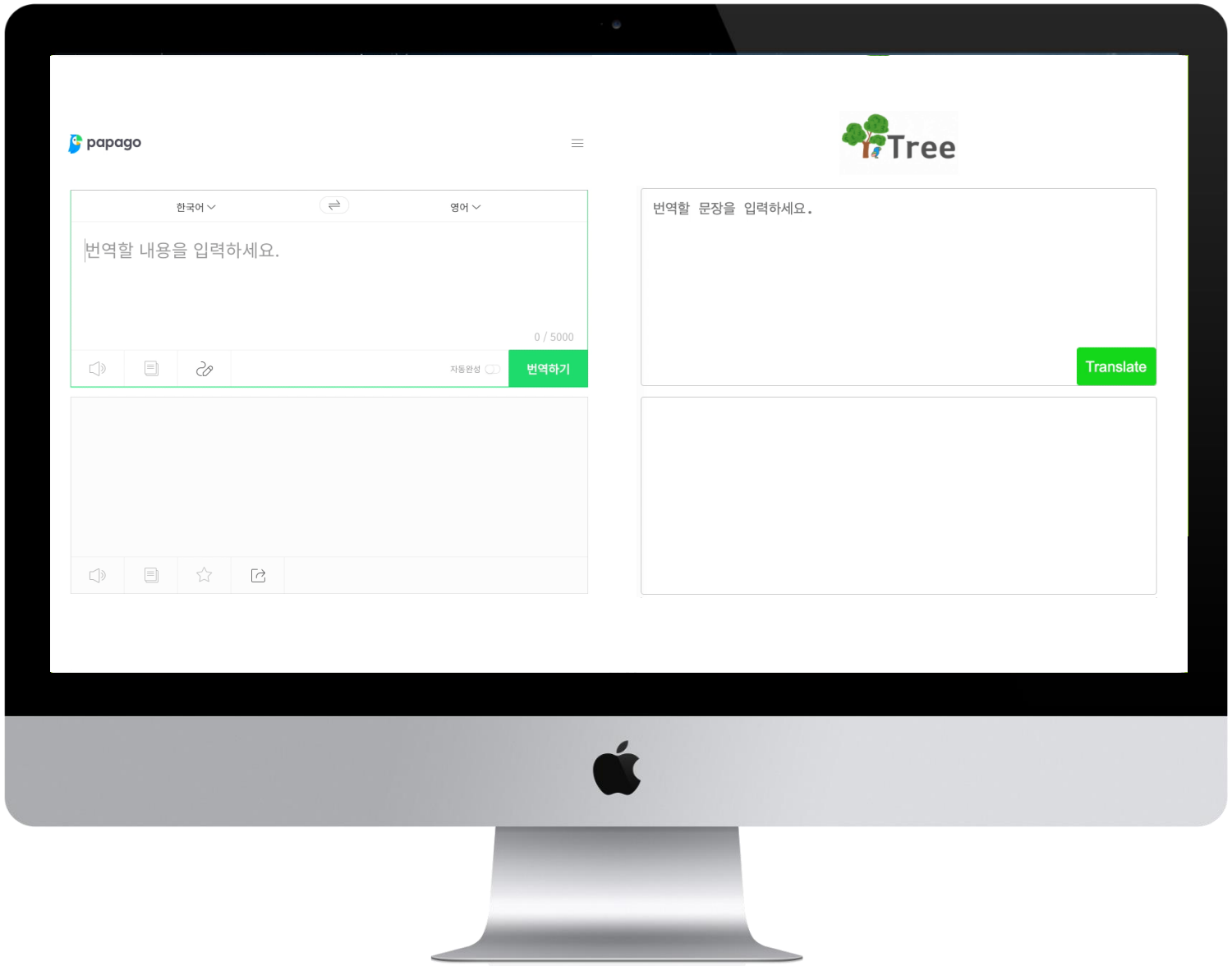
'한국을 유발하다…?'

Kakao i

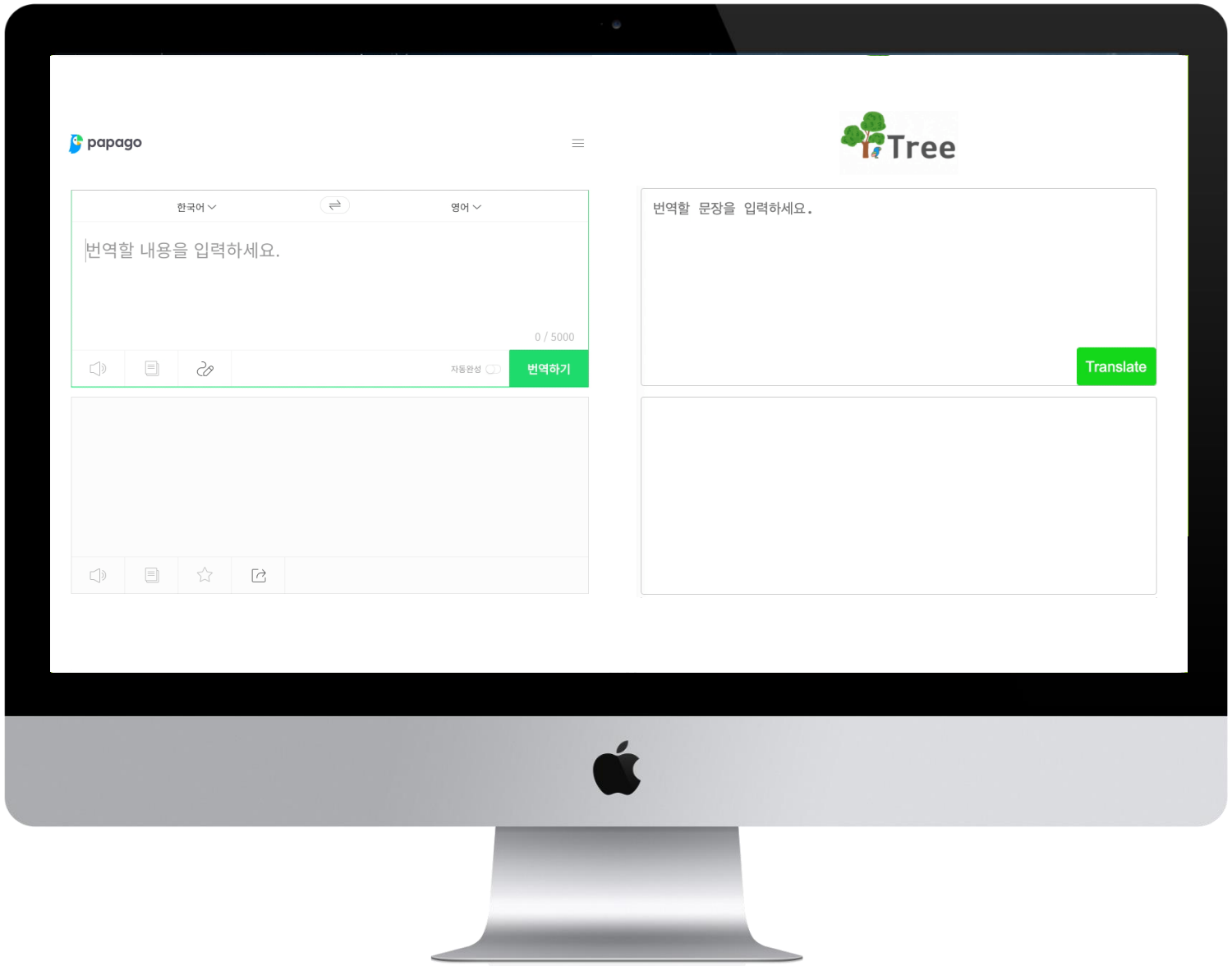
The conflict between the US and China has caused a fire to our country.

'화재를 일으키다…?'

5. Tree


papago
한국어
영어
번역할 내용을 입력하세요.
0 / 5000
자동완성
번역하기
Tree
번역할 문장을 입력하세요.
Translate

5. Tree


papago
한국어
영어
번역할 내용을 입력하세요.
0 / 5000
자동완성
번역하기
Tree
번역할 문장을 입력하세요.
Translate

기술 스택

Model

Web

팀 구성원 및 역할

OpenNMT Developer Team

DEVELOPER

심판교

번역 엔진 개발
관용구 병렬 데이터셋 구축
웹 디자인

DEVELOPER

진가형

번역 엔진 개발
관용구 병렬 데이터셋 구축
웹 디자인

DEVELOPER

유현승

번역 엔진 개발
관용구 병렬 데이터셋 구축
웹 디자인

DEVELOPER

이지수

번역 엔진 개발
관용구 병렬 데이터셋 구축
웹 서비스 구현

Idiom Classifier Team

DEVELOPER

김태영

분류기 연구 및 구현
관용구 병렬 데이터셋 구축

DEVELOPER

김예찬

분류기 연구 및 구현
관용구 병렬 데이터셋 구축

DEVELOPER

김택현

분류기 연구 및 구현
관용구 병렬 데이터셋 구축

THANK YOU!

Q & A

Appendix

Appendix A. 관용구 번역 관련 논문 - 1

관용구 기계번역을 위한 한-영 데이터셋 구축 및 평가방법 (최민주(국민대), 한국정보과학회, 2020.07)

- 논문 내용
  1. 관용구 기계번역을 위한 한-영 번역쌍 데이터셋 구축 (KISS)
  2. 번역 결과를 평가하기 위해 블랙리스트를 구축

- 블랙리스트 : 직역으로 인한 오역을 탐지

표 2 한-영 관용구 블랙리스트 예시

관용구	블랙리스트
꼬집어 말하다	nip pinch twitch
눈 높다	eye
운을 떼다	lucky
유명을 달리하다	famous

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 woman.

Appendix A. 관용구 번역 관련 논문 - 2

한-영 관용구 기계번역을 위한 NMT 학습 방법 (최민주, 이창기(강원대), HLCT, 2020.10)

- 논문 내용 : 신경망 기계번역 모델에 관용구를 효과적으로 학습시키기 위한 방법 제안
  1. 학습 데이터에 <idm> 토큰을 사용하여 문장 내 관용구 위치를 표기하면 NMT 모델 대부분에서 관용구 번역 품질 상승
  2. But 비관용구 문장의 번역 성능이 저하되는 경향 → 비관용구 문장의 번역 품질을 유지하며 관용구 번역을 개선할 수 있는 방법 연구 필요

표 4 실험 결과 BLEU 점수 [8] 비교

	NMT 모델	2 Layer LSTM (OpenNMT Default)		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
1	관용구 미포함 학습 (Baseline)	24.45	**30.32**	26.05	**32.76**	28.28	33.58	28.37	**34.45**
2	관용구 포함 학습	24.37 (-0.08)	30.23 (-0.09)	26.45 (+0.40)	32.63 (-0.13)	29.09 (-0.81)	34.09 (+0.51)	**28.73 (+0.36)**	34.22 (-0.23)
3	관용구 포함 학습 + <idm> 태그	24.32 (-0.13)	30.16 (-0.16)	**26.76 (+0.71)**	32.62 (-0.14)	**29.46 (+1.18)**	**34.71 (+1.13)**	28.42 (+0.05)	34.02 (-0.43)
4	관용구 포함 학습 + <idm> </idm> 태그	**24.51 (+0.06)**	30.00 (-0.32)	25.89 (-0.16)	32.68 (-0.08)	27.63 (-0.65)	33.88 (+0.30)	27.16 (-1.21)	33.88 (-0.57)

Appendix A. 관용구 분류 관련 논문

AWD-LSTM을 이용한 관용구의 분류 (Classification of Idiomatic Sentences Using AWD-LSTM)

(J.Briskilal and C. N. Subalalitha, Expert Clouds and Applications, 2021.07)

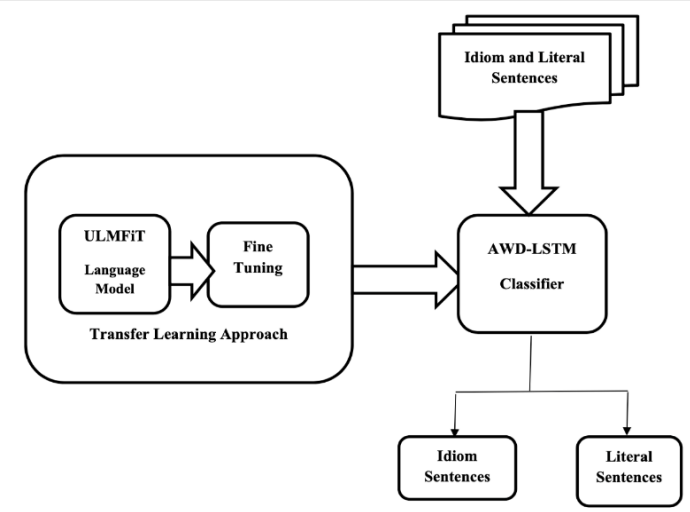
- 논문 내용 : 사전학습 모델을 이용한 관용구 분류

1. Wikitext-103 Dataset으로 사전학습한 ULMFiT 모델을 (관용구+비관용구) 문장 Dataset으로 Fine-Tuning
2. AWD-LSTM Classifier는 관용구 문장과 비관용구 문장을 분류하는 역할

TroFi 데이터셋의 구조

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

관용구 분류기 구조



Appendix B. 관용구 분류기 성능

STEP 1. 영어 관용구 분류기 성능 비교

- 비교를 위해 사용한 데이터셋 : TroFi 데이터셋

Model	Precision	Recall	Accuracy	F1 Score
ULMFiT+AWD-LSTM	0.763	0.873	0.768	0.814
BERT+FFN (base-multilingual-cased)	0.836	0.888	0.844	0.857

Appendix B. 관용구 분류기 성능

STEP 2. 한국어 관용구 분류기 성능 비교

- Fine-tuning 데이터셋 : KISS 데이터셋 (관용구 3,376개 / 비관용구 3,376개)

Model	Precision	Recall	Accuracy	F1 Score
BERT	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	0.91
KoGPT2	0.95	0.98	0.96	0.97

Appendix B. 관용구 분류기 성능

STEP 3. 관용구 분류기 모델 선정

- 최종 Idiom Classifier 모델을 선정하기 위해 3가지 케이스를 테스트
  - 케이스 1 : 관용구 5,855개 + 비관용구 5,855개
  - 케이스 2 : 관용구 17,307개 + 비관용구 17,307개
  - 케이스 3 : 관용구 30,000개 (oversampling) + 비관용구 30,000개
- 각 케이스에 대한 검증 데이터셋은 동일 (관용구 1,500개 + 비관용구 1,500개)
- 테스트 결과

Model	케이스 1	케이스 2	케이스 3
KoBERT	0.91	0.952	0.953
BERT Multilingual	0.89	0.918	X
KoGPT2	0.95	0.947	0.954

Appendix C. 번역기 성능(논문 응용)

1. <idm> 토큰 위치에 따른 번역 성능 실험

- 6 Layer Transformer 모델에 일반 문장, 관용구 문장, 일반 문장 + 관용구 문장의 3가지 테스트셋 각각의 번역 성능을 실험
- 테스트 결과

일반 문장 TEST

학습셋	테스트셋	BLEU
관용표현 앞 <idm>	일반 문장	36.86
문장 맨 앞 <idm>	일반 문장	36.68

관용구 문장 TEST

학습셋	테스트셋	BLEU
관용표현 앞 <idm>	일반 문장	30.26
문장 맨 앞 <idm>	일반 문장	31.11

일반 문장 + 관용구 문장 TEST

학습셋	테스트셋	BLEU
관용표현 앞 <idm>	일반 문장	36.19
문장 맨 앞 <idm>	일반 문장	36.12

Appendix C. 번역기 성능(논문 응용)

2. 의역을 통한 관용구 데이터셋 증축 + 토크나이저 변경 후 번역 성능 실험

- 6 Layer Transformer + Sentencepiece (model type = bpe, 전체 문장 사용)
- 테스트 결과

Epoch	학습셋	테스트셋	BLEU
80,000	문장 맨 앞 <idm>	일반 문장 + 관용구 문장	36.86
80,000	문장 맨 앞 <idm>	관용구 문장	36.68
100,000	문장 맨 앞 <idm>	일반 문장 + 관용구 문장	48.32
100,000	문장 맨 앞 <idm>	관용구 문장	36.48

Appendix C. 번역기 성능(최종)

3. Tree (관용구 36만 문장 증측 후)

- 6 Layer Transformer + Sentencepiece (model type = bpe, 전체 문장 사용)
- 테스트 결과

Epoch	학습셋	테스트셋	BLEU
200,000	문장 맨 앞 <idm>	일반 문장 + 관용구 문장	48.53
200,000	문장 맨 앞 <idm>	관용구 문장	39.49