

다국어 채팅 서비스 앵무새톡 Project Report

Table of Contents

[Project Overview](#)

[Project 세부 & 적용 방법론](#)

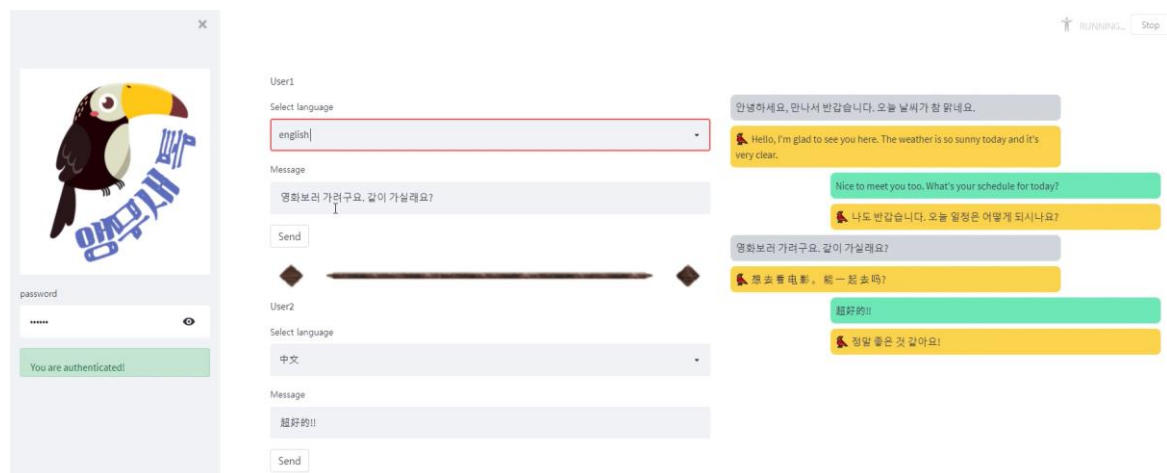
[Model Evaluation](#)

Project Overview

- 소개 : 자신의 언어를 설정하면, 언어에 맞게 상대방의 메시지를 실시간으로 번역하여 메시지 원문 하단에 번역된 결과를 보여주는 다국어 채팅 서비스 데모 제작.

(한국어-영어-중국어 3개 언어에 대한 양방향 번역 지원)

- 웹 UI



- 학습에 이용한 데이터셋

- AI Hub 제공 번역 데이터셋

- 한국어-영어 번역 말뭉치 (기술과학) : <https://aihub.or.kr/aidata/30719>
- 한국어-영어 번역 말뭉치(사회과학): <https://aihub.or.kr/aidata/30720>
- 한국어-영어 번역 말뭉치(말뭉치): <https://aihub.or.kr/aidata/87>
- 한국어-중국어 번역 말뭉치(기술과학): <https://aihub.or.kr/aidata/30722>
- 한국어-중국어 번역 말뭉치(사회과학): <https://aihub.or.kr/aidata/30721>

- Huggingface 공개 데이터셋

- https://huggingface.co/datasets/news_commentary
- https://huggingface.co/datasets/wmt20_mlqe_task2
- https://huggingface.co/datasets/wmt20_mlqe_task1,
- <https://huggingface.co/datasets/alt>,
- https://huggingface.co/datasets/ted_iwlst2013,
- https://huggingface.co/datasets/un_pc

Project 세부 & 적용 방법론

• Multi-way MNMT 방식의 다국어 번역 모델 구조

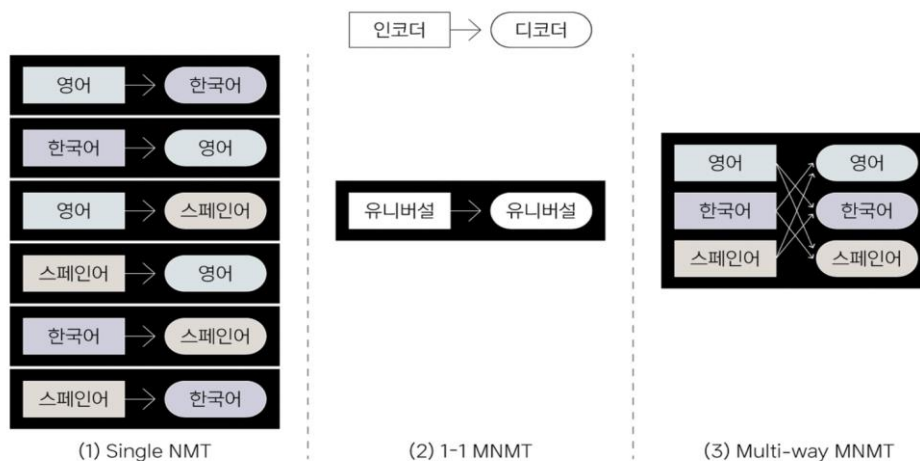


그림) 카카오엔터프라이즈 테크엔

- Multi-way MNMT 방식의 모델 구조는 언어별 인코더와 디코더를 두고 여러 방향에서 이를 공유하는 구조.

- Single NMT 방식은 포괄하는 언어가 증가할수록 학습할 모델의 수가 기하급수적으로 증가하고, 1-1 MNMT 방식은 하나의 거대한 모델이 모든 언어를 포괄해야 하기 때문에 모델의 크기가 충분하지 않을 시, 성능상의 병목과 Entanglement 현상이 발생하는 문제가 존재.

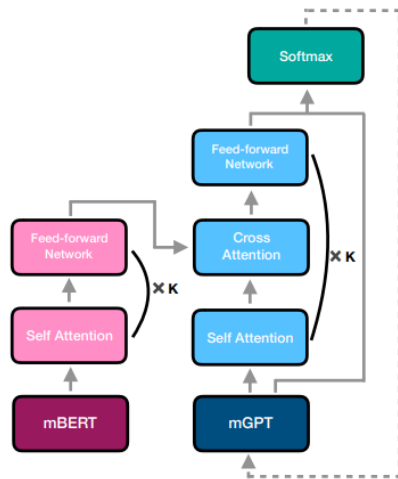
* Entanglement : Output 문장을 영어로 출력하려는 현상. Target 언어가 영어가 아닌데에도 영어 문장을 생성하려는 문제로 주로 1-1 MNMT 모델 구조에서 발생한다.

- 1-1 MNMT를 잘 수행하는 모델이 등장하고 있지만 여전히 기업환경에서는 현실적으로 해결하기 어려운 문제들이 산재.

- Reference : [Revisiting modularized multilingual NMT to meet industrial demands](#)

• 번역 모델 Architecture

- Reference : [Multilingual Translation via Grafting Pre-trained Language Models](#)
- 위 논문에서 제시된 Graformer 모델을 참조해 구현.



• Graformer Architecture

- mBERT와 mGPT를 두고 이들의 Output을 연결하는 Grafting Module을 추가하여 번역을 수행하는 구조.
- GPT2 모델은 트랜스포머의 디코더 구조를 사용하기 때문에 인코더의 출력 결과를 디코더에 전달하는 Cross Attention 파트가 존재하지 않고, 이 역할을 Graft module이 수행하는 구조이다.

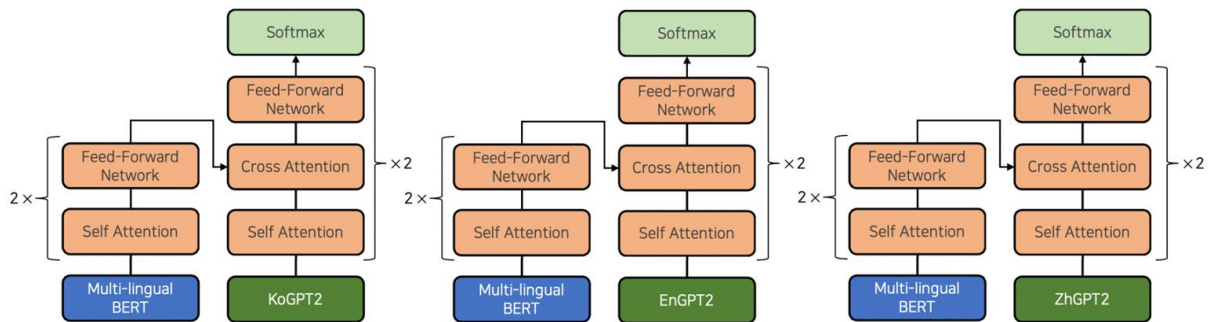
• Graformer 모델 선정 이유

1. 기존의 잘 학습된 BERT와 GPT 모델 활용 가능
2. Grafting Module을 이용한 구현의 용이성
3. 다국어 번역 태스크 SOTA 모델 중 하나인 mBART를 능가하는 좋은 번역 성능을 보이고 있다.
4. 주어진 Resource 내에서 시도할 수 있는 가장 좋은 대안 중 하나.

• Graformer 문제점 개선

- 원 논문에서의 Graformer 모델은 1-1 MNMT 구조. ⇒ 앞서 이야기한 1-1 MNMT 구조의 문제점을 인지하고 주어진 상황에 맞게 Multi-way MNMT 구조로 변경하여 사용.
- 각 언어별 인코더와 디코더의 학습 과정이 공유되기 때문에 순차적인 학습이 필요. ⇒ 프로젝트에 주어진 기간을 고려하여 병렬적으로 학습이 가능하도록 인코더만 Multi-lingual BERT 모델을 이용.

• 최종 번역 모델 구조도



• 번역 모델 경량화 방법론

- Weight sharing , Weight pruning, Quantization 등 다양한 모델 최적화 방법론 중 실질적으로 모델 파라미터 수와 Inference 시간을 줄일 수 있는 Knowledge Distillation 방법론 선정.
- 관련 논문 조사 후 아래의 두가지 방법론 선정.

- 1) [TinyBERT](#)
- 2) [Weight Distillation](#)

• 경량화 방법론 선정 이유

- TinyBERT
 - 원 논문에서 보여주고 있는 경량화 성능이 굉장히 좋았고, Knowledge Distillation 방법론에서 중요한 요소 중 하나인 Distillation Objective가 Transformer 모델 구조에 적합하다고 판단.
- Weight Distillation
 - 기존의 연구들이 Teacher 모델의 어떤 정보를 Student 모델에 전달해 줄 것인지에 대한 연구였다면, 해당 논문은 Teacher 모델의 잘 학습된 파라미터를 Student 모델에 효과적으로 전이하는 방법에 대해 다루고 있다.
 - 논문에서는 이전 연구들이 단순히 Teacher 네트워크의 일부 Layer를 그대로 가져와 사용하였지만, 좀 더 효과적으로 Student 네트워크의 초기 파라미터를 구성하는 것이 중요하다고 주장하고 있으며 이에 관련된 내용이 TinyBERT 논문에서도 언급된 바가 있어 함께 적용해줄 방법론으로 선정하였다.

Model Evaluation

- 번역 성능 (SacreBLEU score 기준)

	Teacher Model	Student Model
한국어	32.90	27.54
영어	30.03	21.89
중국어	53.88	48.07

- 경량화 결과 (Pytorch Profiler 이용)

	Teacher		Student	
	CPU time total	CUDA time total	CPU time total	CUDA time total
한국어	24.831s	1.429s	15.167s	0.957s
영어	39.022s	2.876s	27.127s	2.387s
중국어	32.186s	1.887s	18.997s	1.133s

약 40% 감소

- 경량화는 기존 인코더 레이어 12개, 디코더 레이어 12개로 구성되어 있던 Teacher 모델에서, 각각 4개의 인코더 레이어와 디코더 레이어를 갖는 Student 모델로 학습을 진행.
- 결과적으로 모델 파라미터 수(330M → 210M)와 추론 시간 모두 약 40%가량 감소하였다.

- 번역 문장 예시

- AI Hub 데이터 (Teacher / Student 모델 수행 결과)

입력 문장:

외부 전자 장치(301)의 ISP 종류는 전자 장치(101)의 ISP 종류와 서로 다를 수 있다.

He has also served as Volkswagen's general design manager since 2002 and has won the German Federal Design Grand Prize four times.

생성 문장:

The ISP type of the external electronic device 301 may be different from that of the ISP type of the electronic device 101.

또, 2002년부터 폭스바겐의 디자인 총괄 책임자로 근무하고 독일 연방디자인 대상 4차례도 수상하였다.

정답 문장:

The ISP type of the external electronic device 301 may be different from that of the electronic device 101.

또 2002년부터 폭스바겐의 디자인 총괄책임자로 근무, 독일연방 디자인 대상을 4번이나 수상했다.


- Ted talks 평가 데이터 (Teacher / Student 모델 수행 결과)

입력 문장:

这也不奇怪，我们的地球上人口暴涨，其中有半数以上居住在城市。


(Laughter) What were we just talking about?

생성 문장:

 그것도 이상하지 않은데, 우리 지구상에는 인구가 폭증하고 있고 그 중 절반 이상이 도시에 거주하고 있다.

 (웃음), 뭐 그냥 얘기하고 있을까요?

정답 문장:

 놀랄 것도 없지만, 현재 전체 인구의 절반 이상이 도시에 거주하고 있고,

 (웃음) 무슨 얘기를 하고 있었죠?