Goorm Project#3

Lyrics Machine Translation

< 2조 (2조) >

조현수, 이준엽, 김남준, 김영수, 손동협, 조승연, 정명관

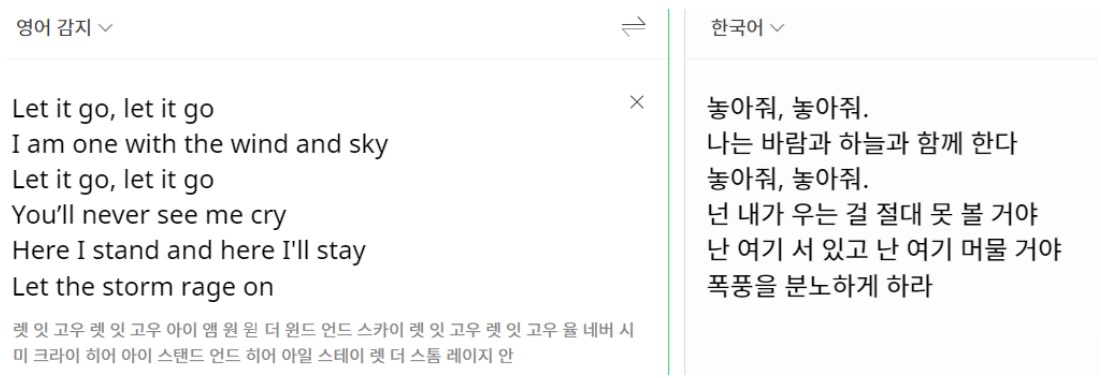
Index

Problem

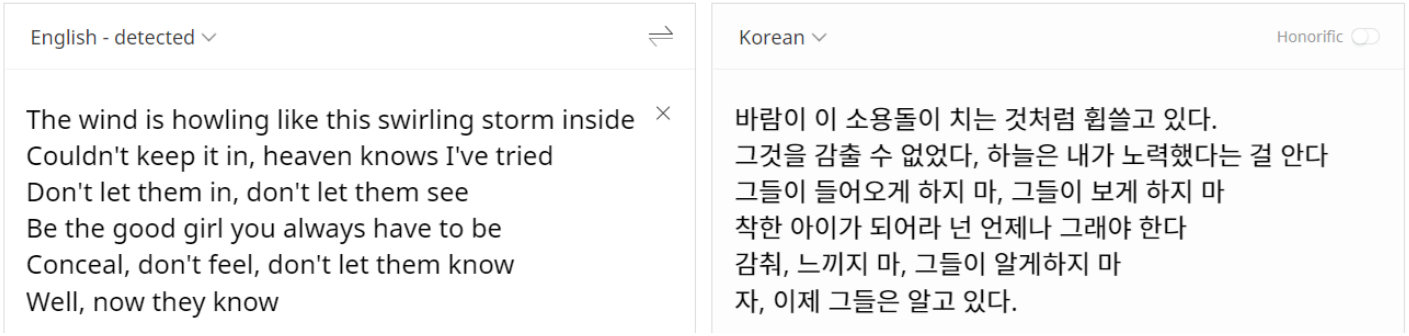
팝송 등 외국 노래들은 유튜브, 멜론 등 음악플레이어에서도 가사가 번역되어 있는 경우가 드물다.

노래 가사는 구어체의 표현 등이 많기 때문에 일반적인 번역기로는 노래 분위기에 맞게 번역이 잘 안된다.

제대로 된 가사 번역본을 찾기 위해서는 블로그 등을 찾아야 하고 유명하지 않은 노래는 찾아도 나오지 않는다.



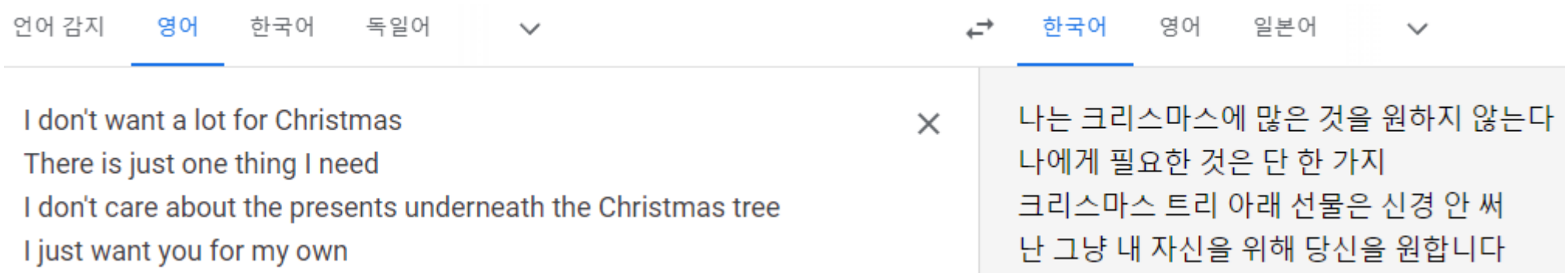
파파고 번역기



구글 번역기

Solution

기존 번역 모델에서 parallel corpus 형태의 노래 가사 데이터로 학습을 시켜 직역이 아닌 노래가사의 느낌이 나도록 번역을 한다.




```
==========================================================================================
Input:
I don't want a lot for Christmas There is just one thing I need I don't care about the presents underneath the Christmas tree I just want you for my own
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
Output:
0: 난 크리스마스에 많은걸 원하지 않아 내가 원하는건 오직 한가지 있지 노을 아래 선물은 신경쓰지 않어 난 널 내 것으로 하고 싶을 뿐이야
==========================================================================================
```

데이터셋 확보 방안

노래 가사 번역 사이트:

https://lyricstranslate.com/ , https://kgasa.com/

- 프로그램을 이용한 번역 X
- 사용자가 번역 후 공유
- 다양한 언어로 약 총 1,265,481개의 parallel corpus 데이터
- 영-한 번역 약 1300여곡
- selenium, requests, beautifulsoup4 라이브러리를 이용한 크롤링

데이터 부족

데이터셋 확보 방안

노래 가사 사이트:

https://www.melon.com/ https://www.lyrics.co.kr/

- monolingual corpus
- selenium, requests, beautifulsoup4 라이브러리를 이용한 크롤링
- back translation 후 기존 데이터에 추가
- 잘못된 번역 제거
- 다른 언어 포함된 가사 제거
- 의미 없는 단어 혹은 문장 제거
- 총 약 24,000개 데이터 수집

monolingual corpus기에 주어진 시간과 GPU 성능에 따라 더 많은 데이터를 수집 가능

구현

- 모델 선정
  - 한-영 가사 번역기: KoBERT-DistilGPT2
  - 영-한 가사 번역기: DistilBERT-KoGPT2
  - m2m100 (Facebook AI)

- 모델 학습
  - Encoder-Decoder구조의 pretrained model 설계
  - AIHub의 parallel data를 이용하여 추가 학습
  - 수집한 데이터를 이용하여 추가 학습

모델 성능 검증

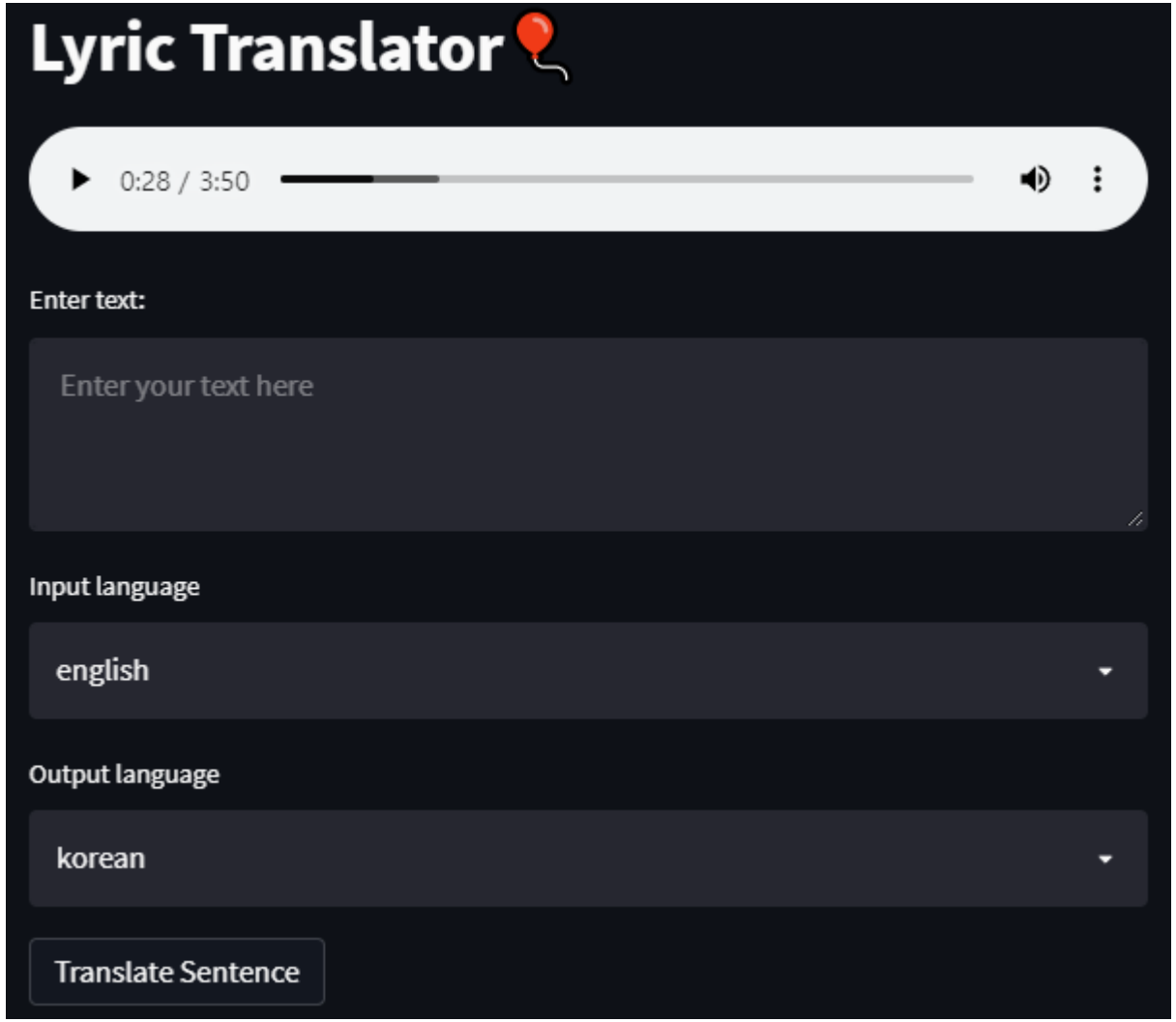
- BLEU score를 이용하여 학습한 모델들과 구글 번역기와의 비교

- 성능 평가는 Test data 2000개를 Train data에서 분리하여 사용

	English -> Korean **(BLEU)**	**Korean -> English** **(BLEU)**
Google Translation	0.1681	0.2459
BERT-KoGPT2	0.0904	-
KoBERT-GPT2	-	0.2767
m2m100	**0.3383**	**0.2772**

어플리케이션 구성

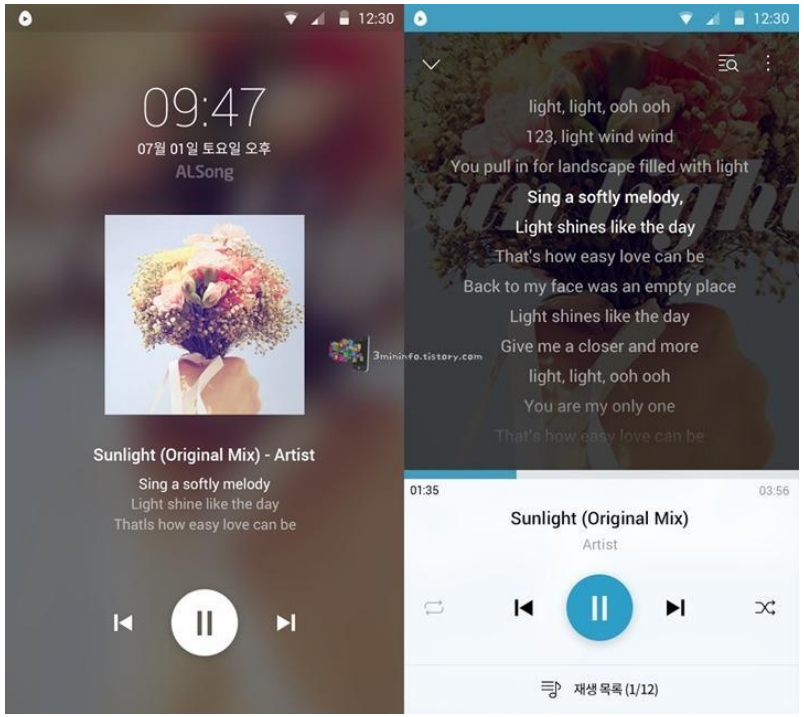
- text box에 한글 혹은 영어의 가사를 입력
- 입력 언어와 출력 언어를 선택
- Translate sentence 버튼 클릭
- 번역된 가사 출력
- 음성 출력

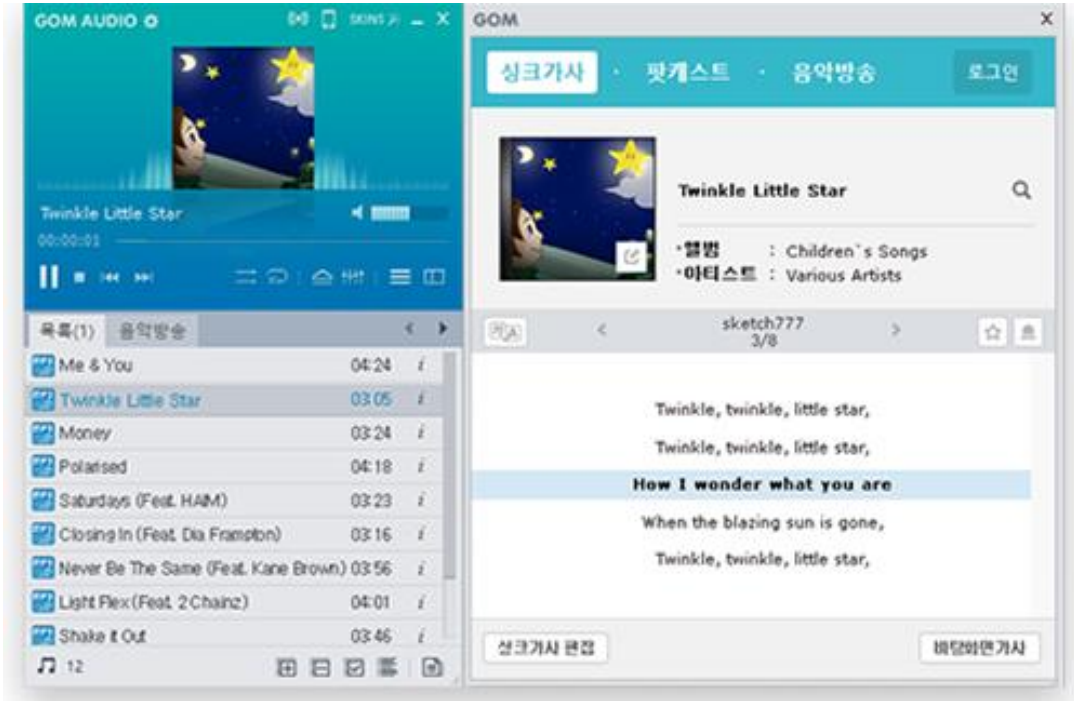


Market Opportunity

기존 번역기와 같이 어플 및 홈페이지 형태로 제작 시 접근성이 좋다.

시장에 존재하지 않는 서비스





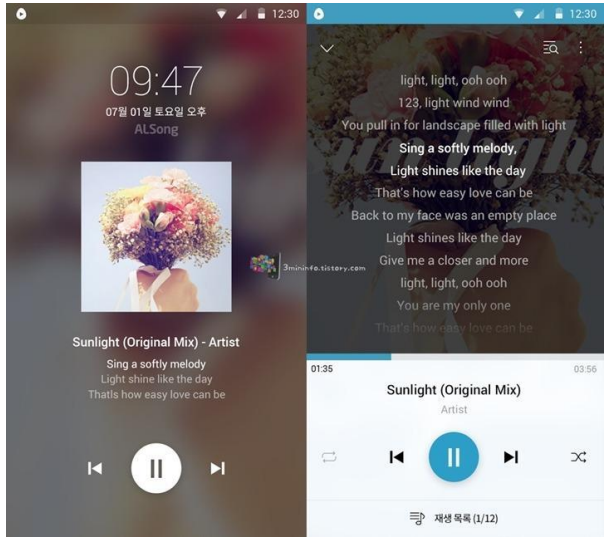
Marketing & Growth Strategy

기존의 가사 어플과 결합하여 번역된 가사 제공

오역 발생 시 사용자가 직접 데이터를 추가해가며 모델이 점점 더 발전

외국에서도 k-pop에 더 쉽게 접근 (실제 사람들이 번역해 놓은 데이터도 대부분이 k-pop)

가사 외에도 Youtube, 영화나 드라마 등 초월번역이 필요한 곳에도 응용 가능



Service Target & Future Work

Service Target

K-pop을 사랑하는 외국인

팝송을 사랑하는 사람

외국어를 잘 하지 못하는 사람

팝송으로 언어 공부하려는 사람

Future Work

데이터셋을 늘려 한-영 외의 다른 언어들까지 가사 번역이 가능하도록 모델 개선

음성인식 기능 추가

스트리밍 어플과 연동하여 실시간 가사 번역

음성합성 기능에 멜로디를 추가하여 가수가 번역된 언어로 노래를 부르는 듯한 기능

가사 데이터 특성 상 완전한 문장으로 번역되지 않을 때가 있음

프로젝트 일정

	4.19	4.20	4.21	4.22	4.23	4.24	4.25	4.26	4.27	4.28	4.29	4.30	5.1	5.2	5.3
아이디어 회의															
데이터 수집 및 전처리															
모델 탐색 및 설계															
모델 성능 향상															
모델 후처리															
Streamlit 구현															
보고서 작성 및 발표															

Live Demo

Streamlit Live Demo:

References

M2M100
Fan, Angela, et al. "Beyond english-centric multilingual machine translation." Journal of Machine Learning Research 22.107 (2021): 1-48.

DistilBERT
Sanh, Victor, et al. "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter." arXiv preprint arXiv:1910.01108 (2019).

BERT
Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

GPT-2
Radford, Alec, et al. "Language models are unsupervised multitask learners." OpenAI blog 1.8 (2019): 9.

Back-Translation
Edunov, Sergey, et al. "Understanding back-translation at scale." arXiv preprint arXiv:1808.09381 (2018).

Q & A

Thank You