[적은 데이터로] 기계번역 하기

AI 17기 김재윤

Section Project 3

과제: [적은 데이터로] 기계번역하기

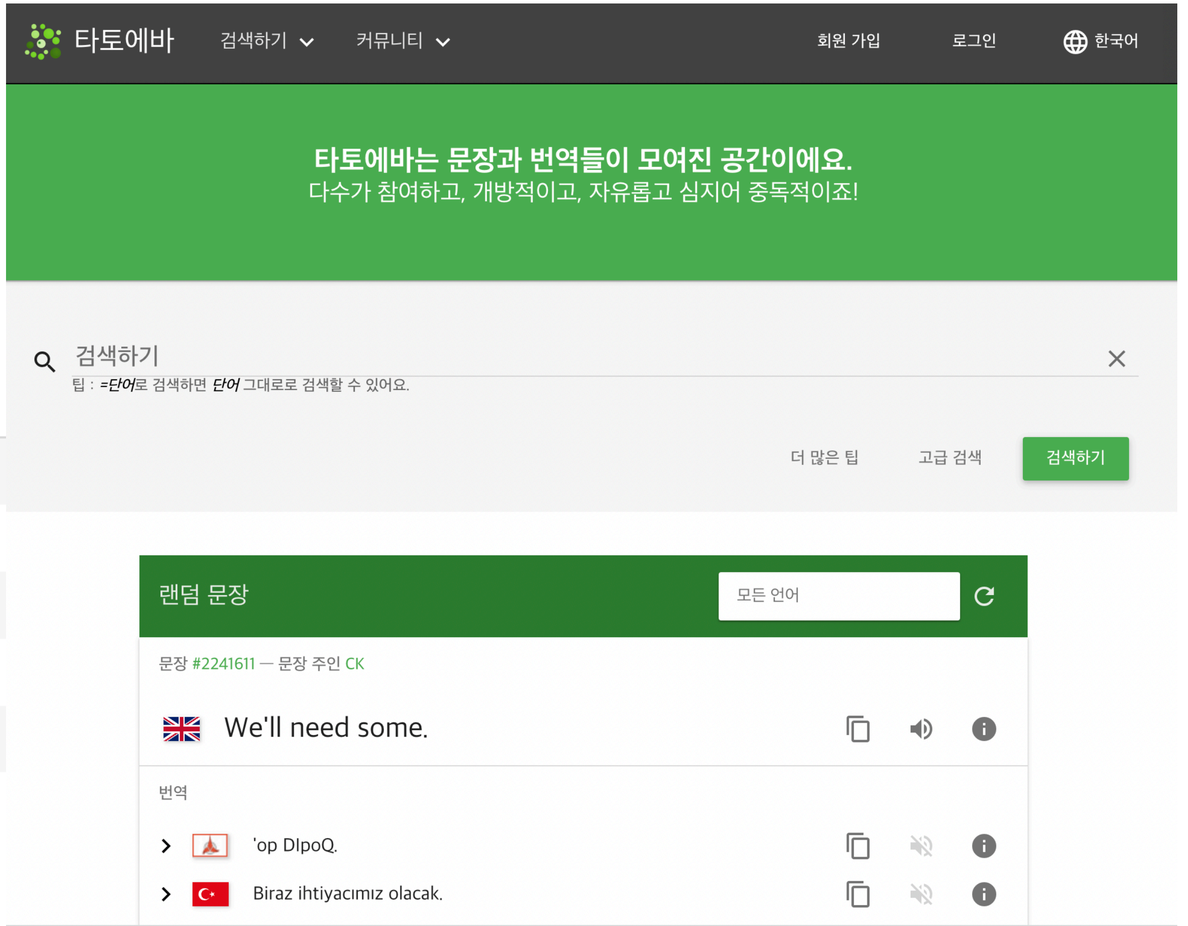
- NLP에 대한 관심때문에, 기계번역이라는 태스크를 선택.
- 이미 많은 연구가 진행되어있고, 구현된 모델도 많은 상황.
- **'제한된 시간 내에, 얼마나 적은 데이터로, 얼마나 정확한 번역이 가능할까?'**
  - cf) 섹션 프로젝트2, 영화 평가 Sentiment Analysis에서, 최적 정확도 86%의 태스크를, 500개 단어만 추출하여 사용해도 80%의 정확도 보장

- **가설: 적은 데이터를 사용해서, 정확도가 크게 떨어지지 않을 것이다!**

사용 자료

- 데이터 : 타토에바 프로젝트 - 독-영 변환자료

	source	target
78505	Tom and I are giving up.	Tom und ich geben auf.
211264	I saw him playing cards with all of them.	Ich sah ihn mit ihnen allen Karten spielen.
64503	His mother is American.	Seine Mutter ist Amerikanerin.
48427	I'm driving you home.	Ich fahre Sie nach Hause.
31543	I heard explosions.	Ich habe Explosionen gehört.
145441	Swimming is a form of exercise.	Schwimmen ist eine Form der körperlichen Übung.



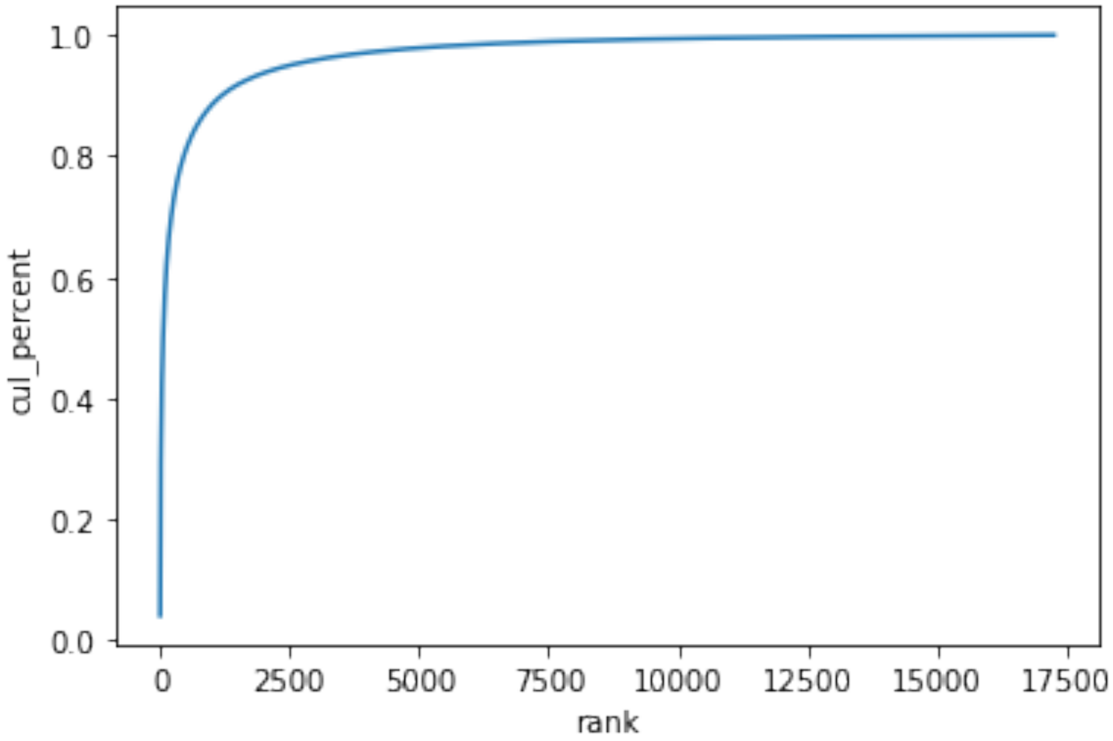
- 코드 : Renu Khandelwal 의 코드(https://arshren.medium.com/) (서-영 번역 예시)

독일어 - 영어 특징

- 인도-유럽 어족의 게르만 어파 서게르만어군의 유이한 메이저 언어.
- 영어의 복잡다단한 역사 때문에 어휘적 유사성이 대단히 높지는 않음
- 문법적 유사성: 어순에 대한 유사성이 높지는 않음. 다만 두 언어 모두 전치사를 사용하며, 주로 Head-last 언어의 구조를 띄고 있음.
- 굴절어라는 특성, 시제 구분의 유사성 면에서는 확고한 이득이 있음.
- 사용 문자는 라틴 알파벳으로 동일하나, 독일어에는 움라우트 문자 3종과 에쓰체트, 총 4개의 추가문자가 있음. 하지만 라틴 알파벳 표기법이 존재하므로 전처리에 용이.

[적은 데이터로]

- 총17260단어 -> 총 260434문장
- 코퍼스 내 단어 누적합
- 1000단어 -> 88%
- 2000단어 -> 93%
- 5000단어 -> 98%
- 10000단어 -> 99%
-



[적은 데이터로]

- 총17260단어 -> 총 260434문장 -> **10만문장 추출 (C)**
- 코퍼스 내 단어 누적합
- 1000단어 -> **13만 문장 (A)**
- 2000단어 ->
- 5000단어 -> 23만 문장 -> **8만문장 추출 (B)**
- 10000단어

- 가설 재정의: A 데이터가 C데이터와 성능차가 크지 않을 것이다!

모델 해설

- word는 특별한 인코딩(tf-idf나 w2v등) 없이 인덱싱.
- LSTM기반 인코더/디코더 사용.
- 디코더는 다음 단어에 올 확률이 가장 높은 단어를 찾아 배치하는 식으로 계속해서 서치/배치를 진행하여 문장을 생성.

결과

A set: 1000words, 130k sent - loss: 0.9495 - acc: 0.6664

B set: 5000words, 80k sent - ??

C set : Full words, 100k sent - loss: 1.2561 - acc: 0.6408

수치상으로 보았을 때는 가설대로 A set이 더 좋은 성능을 보임. 하지만 번역이 정확히 어떤 식으로 진행되는지는 예시 문장을 보아야 알 수 있음.

A set (1000words, 130k sent)

Input Source sentence: tom went there in person

Actual Target Translation: **tom ging persönlich hin**

Predicted Target Translation: **tom ging persönlich hin** gehen haben haben haben h

Input Source sentence: i read a book as i walked

Actual Target Translation: ich las ein buch als ich spazieren ging

Predicted Target Translation: **tom ist nicht sehr groß** aus haben haben haben h

C set (Full words, 100k sent)

Input Source sentence: tom is likely to get married again

Actual Target Translation: **tom wird wahrscheinlich wieder heiraten**

Predicted Target Translation: **tom wird wahrscheinlich wieder heiraten** werden zu

Input Source sentence: whats your favorite kind of sushi

Actual Target Translation: was ist dein lieblingssushi

Predicted Target Translation: **wir müssen tom aufhalten bevor er sich nicht g**

한계점과 개선사항

- 한계점
  - 비교군이 “적합한 성능을 가진, Fully Trained Model”이 아닌 불완전한 모델끼리의 비교.
  - B set의 경우 시간부족으로 트레이닝이 완성되지 못함.
  - 하이퍼 파라미터 조절시간 부족
- 개선사항
  - 더욱 많은 시간을 들여 다양한 테스트를 해볼 수 있다면 좋을 것.