Annual Conference of KIPS

가사 데이터를 활용한 기존 기계번역 모델의 확장 가능성 연구

권준혁[1], 권민지[1], 김다연[1], 장은수[1], 변종석[1]
전희국[2], 임동혁[1]
[1]광운대학교 정보융합학부, [2](주) 핀다

1. INTRODUCTION

최근 기계번역(Machine Translation)이 크게 발전하면서 전 세계의 언어장벽이 허물어지고 있다. 이러한 기계번역 모델은 학습에 신문 보도와 같은 산문체의 일반적인 논리적 문장을 많이 활용했기 때문에 논리적이고 사실적인 정보를 다루는 텍스트에는 비교적 성능이 높은 것으로 알려져 있다. 하지만 인간의 감정을 다루는 함축적, 예술적 표현들은 기계번역 모델이 처리하는 데 한계가 있다.

따라서 본 연구에서는 기존 기계번역 모델의 한계를 극복하고자 했다. 이를 위해 학습 데이터 세트로 음악 가사 데이터를 활용하였다. 음악 가사는 인간의 감정이 함축적으로 표현된 예술적 표현의 정수이다. 이로인해 기존의 번역모델이 이해하기 힘든 부분들이 존재하기에, 본 연구에 적합한 데이터 세트라고 판단하였다. 이러한 음악 가사 데이터를 활용한 기계번역 모델 연구로 기존 기계번역 모델의 한계점 극복과 확장 가능성을 탐구하고자 한다.

2. METHODOLOGY

데이터 수집 및 전처리

- Web Scraping
  Selenium와 beautifulsoup 모듈을 활용하여 총 600곡의 한국 음악 가사 데이터를 수집하였고, 각 가사 데이터별로 적당한 길이에 맞게 나누어 저장하였다.
- Cloud Translation API
  구글의 Cloud Translation API를 활용하여 한국어 가사에 대응하는 영어 번역 데이터를 얻었다.

	ko	en
0	그리움이 눈처럼	longing like snow
1	쌓인 거리를	the piled up street
2	나혼자서 걷다네	I walked alone
3	미련때문에	because of regret
4	흐르는 세월따라	according to the passing time
...	...	...
23151	너와 나 사이	between you and me
23152	이 묘한 떨림 속에	In this strange trembling
23153	너와 나만의 이 사랑을 느끼죠	I feel this love between you and me
23154	가득 채워진 내 사랑을 놓지 말아 줘	Don't let go of my full love
23155	늘 지금처럼 내 곁에	always by my side like now

[그림1] 음악 가사 데이터 세트

모델 학습

- 전이 학습(Transfer Learning)
  효과적인 학습을 위해 대규모 데이터 세트로 학습된 기계번역 모델을 연구모델의 학습에 이용하였다.
- Pretrained model
  전이 학습을 위한 사전 학습 모델로 자연어 처리 MarianMT 모델을 활용하였다.
- Experiments
  실험을 추적하고, 하이퍼 파라미터 검색 및 최적화를 위해 WandB(Weights & Biases) 플랫폼을 활용하였다.

평가

모델의 성능을 비교하고 최적의 모델을 찾기 위해 다음의 평가 지표를 사용했다.

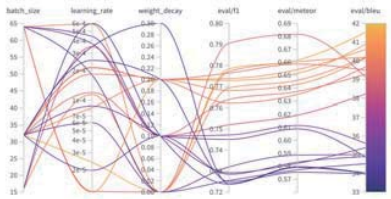
- BLEU
- METEOR
- BERTScore

3. EXPERIMENTAL RESULT

모델 성능 향상을 위해 최적의 하이퍼 파라미터를 찾고자 했다.

batch size, epochs, learning rate, weight decay 등의 파라미터에 범위 또는 값을 지정해서 WandB의 Sweep 기능을 통해 자동으로 파라미터에 따른 학습 결과를 추적했다.

[그림 2] Sweep config



[그림 3] Sweep result

여러 Sweep 중, 좋은 점수를 받았던 하이퍼 파라미터 조합을 대상으로 더 면밀하게 학습을 진행했다. 이를 통해 최적화된 값을 찾아 모델의 성능을 향상시킬 수 있었다.

	F1	METEOR	BLEU
재학습 전	0.6680	0.5613	16.3772
재학습 후	0.8014	0.6914	43.8160

[표 1] 평가 결과

4. CONCLUSION

ko	en
기억 모퉁이에 적혀 있는 네가	You're the one on the corner of your memory
오늘이 한 칸 채 안 남은 그런 시간	There's less than one time today

[표 2] 재학습 전 번역 결과 예제

ko	en
기억 모퉁이에 적혀 있는 네가	You, who was written on the corner of my memory
오늘이 한 칸 채 안 남은 그런 시간	A time like today with less than one space left

[표 3] 재학습 후 번역 결과 예제

- 이번 연구에서는 가사 데이터를 활용하여 번역 모델을 학습시켰다. 기존 번역 모델에 비해 감정표현, 예술적 표현에 관한 번역 성능이 유의미하게 향상된 것을 확인할 수 있었다.
- 더 나아가 학습에 필요한 충분한 데이터 세트를 확보할 수 있다면, 논리적인 텍스트의 번역뿐만 아니라 일상에서 사용하는 함축적, 감정적인 텍스트의 높은 번역 성능을 기대할 수 있다.

5. REFERENCES

- Lukas Biewald. Experiment tracking with weights and biases, 2020. Software available from wandb.com
- Ashish, Vaswani., Noam, Shazeer., Niki, Parmar., Jakob, Uszkoreit., Llion, Jones., Aidan, N., Gomez., Lukasz, Kaiser., Illia, Polosukhin. (2017). Attention Is All You Need.

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학 지원사업의 연구결과로 수행되었음(2017-0-00096)