SOV 언어와 교차 언어 모델 사전 학습을 활용한 한국어-제주어 기계 번역 개선

김정우[1,‡], 배채묵[2,‡], 김미수[3,*]
[1]전남대학교 에너지자원공학과 학부생
[2]전남대학교 인공지능학부 학부생
[3]전남대학교 인공지능융합학과 교수
[‡]공동 1저자, [*]교신저자
jeongwu510@.naver.com, barch0713@gmail.com, misoo.kim@jnu.ac.kr

Improving Korean-Jeju Machine Translation Using SOV Languages and Cross-Lingual Model Pretraining

Jeong-Wu Kim[1], Chea-Muk Bae[2], Misoo-Soo Kim[3]
[1]Dept. of Energy and Resources Engineering, Chonnam University
[2]Dept. of Artificial Intelligence, Chonnam University
[3]Dept. of Artificial Intelligence Convergence, Chonnam University

요 약

제주어 보존과 활성화를 위해, 본 연구에서는 교차 언어 언어 모델을 활용하여 저자원 언어인 제주어의 기계 번역 성능을 개선하고자 한다. 특히, SOV 구조를 가진 언어를 사전 학습에 사용하여 구조적으로 유사한 언어를 통해 한국어→제주어 번역의 성능을 평가했다.

1. 서론

제주어-한국어의 기계 번역은 제주어의 보존과 활성화를 위해 중요하다[1]. 그러나 제주어와 같은 저자원 언어는 데이터 희소성으로 인해 기계 번역 성능이 여전히 제한적인 문제가 있다. 교차 언어 모델 (cross-lingual language model, XLM)은 다중 언어에서의 의미적 및 문법적 공통점을 학습한 것으로, 저자원 언어의 자연어 처리 성능을 크게 향상시킬 수 있기 때문에 제주어와 같이 저자원 언어의 경우에 보다 효과적이다. XLM의 효과를 극대화하기 위해서는 유사한 구조를 가진 언어를 사전 학습시키는 것이 필요하다[2].

본 연구는 제주어와 SOV (주어-보어-목적어)로 어순이 같은 타밀어[3]를 사용해 XLM의 사전 학습을 수행하여 한국어를 제주어로 번역하는 기계 번역의 성능을 높이고자 한다. 이를 통해 한국어와 제주어 간의 상호 이해를 증진시키고, 나아가 제주어의 보존 및 활성화에 기여할 수 있다.

2. 관련 연구

Park et al. (2019)은 제주어-한국어 병렬 데이터셋을 구축했으며, 양방향 번역 성능을 평가하였다[1]. 실험 결과, 제주어에서 한국어로의 번역에서는 BLEU 점수 67.70을 기록한 반면, 한국어에서 제주어로의 번역에서는 BLEU 점수 43.31에 그쳤다. 이는 한국어→제주어 번역 성능이 상대적으로 낮다는 것을 보여주며, 그 원인으로는 번역 모델이 제주어의 고유한 어휘적·문법적 특성을 충분히 반영하지 못한 점이 지적된다. 이러한 한계를 극복하기 위해서는 제주어와 같은 저자원 언어의 특성을 효과적으로 학습할 수 있는 접근이 필요하다.

Zheng et al. (2022)는 한국어→제주어 번역 성능 향상을 목표로 중국어, 일본어, 한국어를 사전 학습시킨 다중언어 모델을 제안하였다[4]. 실험 결과, 세 언어를 모두 학습시킨 모델은 BLEU 점수 64.04를 기록하였으며, 일본어와 한국어만을 학습시킨 모델은 BLEU 점수 70.10을 기록하였다. 일본어는 한국어와 마찬가지로 SOV 어순을 가지는 반면, 중국어는 SVO 어순을 사용한다. 이를 통해 어순이 유사한 언어로 사전 학습된 모델이 더 우수한 성능을 보인다는 사실을 보였다. 본 연구에서는 한국어와 유사한 타밀어로 XLM을 사전 학습하여 번역 성능을 높이고자 한다.

3. 제안 방법

XLM 사전 학습을 위해 SOV 구조를 공유하는 타밀어를 활용한다. 우선 KoBART 모델에 타밀어 데이터를 사전 학습하고, 이후 한국어-제주어 데이터를 학습시켜, 모델이 타밀어와 유사한 문법적 구조를 가진 제주어의 언어적 패턴을 습득하도록 한다. 이를 위해 Translation Language Modeling을 사용하여 두 언어의 문장 표현을 정렬하였다[2].

4. 실험

4.1. 데이터셋

표 1은 실험에 사용한 데이터 셋을 요약한다.

<표 1> 실험 데이터셋

	한국어-제주어	한국어-타밀어
Train Dataset	75,232	9,256
Valid Dataset	25,077	3,085
Test Dataset	25,078	3,086

한국어-제주어 데이터로는 Park et al. (2019)이 구축한 데이터 셋을 사용하였다[1]. 해당 데이터 셋의 문장당 평균 단어 수는 8.3개로 나타났으며, 이 중 단어 수가 8개 이상인 문장들을 선별하여 학습에 사용하였다.

사전 학습을 위한 한국어-타밀어 병렬 데이터 셋이 충분히 존재하지 않기 때문에 역번역을 활용하여 데이터 셋을 구축한다. 이를 위해 영어-타밀어 데이터 셋을 Google 번역 API를 사용하여 영어 부분을 한국어로 번역하여 데이터 셋을 구성하였다.

4.2. 실험 환경

모델은 NVIDIA RTX 3080 GPU 환경에서 학습하였다. 배치 사이즈는 64, 에포크는 10, 학습률은 2e-05로 설정하였다.

4.3. 결과

표 2는 실험 결과를 요약한다. KoBART에 한국어-제주어만 학습시킨 모델(Basline)의 BLEU 점수는 77.0296으로 평가되었다. 타밀어로 사전 학습된 모델(Tamil Pretrained)의 경우, BLEU 점수가 77.0347로 소폭 상승한 것을 확인할 수 있다. 이는 사전 학습에 사용된 타밀어가 한국어와 제주어의 SOV 구조적 유사성을 기반으로 번역 성능에 긍정적인 영향을 미쳤다는 것을 시사한다. 비록 점수 차이는 미미하지만, 이는 저자원 언어인 제주어 번역에서 구조적으로 유사한 언어를 활용한 사전 학습이 성능 향상에 기여할 수 있음을 보여준다.

<표 2> KoBART 모델의 기계 번역 성능 평가

Model	BLEU
Baseline	77.0296
Tamil Pretrained	77.0347

5. 결론

본 연구는 타밀어와 같은 SOV 구조를 가진 언어를 사전 학습에 적용하여, 한국어→제주어 기계 번역 성능을 향상시킬 수 있음을 입증했다. 비록 BLEU 점수의 상승 폭은 작았지만, 이는 사전 학습의 방식이 기계 번역 모델의 성능 개선에 구조적 유사성이 중요한 역할을 한다는 것을 보여준다. 이를 통해 향후 저자원 언어의 번역 성능을 더욱 높일 수 있는 가능성을 제시하며, 나아가 제주어와 같은 소수 언어의 보존과 활성화에 기여할 수 있다.

사사

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 소프트웨어중심대학사업(2021-0-01409)과 과학기술정보통신부 및 정보통신기획평가원의 인공지능융합혁신인재양성사업(IITP-2023-RS-2023-00256629), 대학ICT연구센터사업(IITP-2024-RS-2024-00437718)의 연구 결과로 수행되었음

참고문헌

[1] Park, K., Choe, Y. J., and Ham, J. "Jejueo datasets for machine translation and speech synthesis." arXiv preprint arXiv:1911.12071 (2019).

[2] Conneau, Alexis, and Guillaume Lample. "Cross-lingual language model pretraining." Advances in Neural Information Processing Systems, vol. 32, 2019.

[3] Khanittanan, Wilaiwan (aka Kanittanan). South Asian Languages: Structure, Convergence and Diglossia, New Delhi, Motilal Banarsidass, 1986, pp. 174-178.

[4] Zheng, F., Marrese-Taylor, E., and Matsuo, Y. "Improving Jejueo-Korean Translation With Cross-Lingual Pretraining Using Japanese and Korean." Proceedings of the 9th Workshop on Asian Translation, International Conference on Computational Linguistics, Gyeongju, Republic of Korea, 2022, pp. 44-50.