OpenNMT를 활용한 한글 존댓말 변환기의 구현

정준녕[O], 김상영[*], 김성태[*], 이정재[*], 정유철[*]
[O]금오공과대학교 컴퓨터공학과,
[*]금오공과대학교 컴퓨터공학과
e-mail: {jjn00712[O], ksyu0043[*], sungtae1005[*], cssopy[*]}@gmail.com, jyc@kumoh.ac.kr[*]

Implementation of Korean Honorific Converter Using OpenNMT

Jun-Nyeong Jeong[O], Sang-Yeong Kim[*], Seong-Tae Kim[*], Jeong-Jae Lee[*], Yuchul Jung[*]
[O]Dept. of Computer Engineering, Kumoh National Institute of Technology,
[*]Dept. of Computer Engineering, Kumoh National Institute of Technology

● 요 약 ●

최근 발전한 인공신경망 기반 기계 번역은 번역 시 더 자연스러운 번역을 제공한다. 본 논문에서는 기계번역 기법을 이용하여 반말 표현을 존댓말 표현으로 변환하는 기법을 제안한다. 특히, 이를 위해 DCInside의 게시판을 크롤링하고 AI-HUB 데이터와 합쳐 약 20,000개의 자체 데이터 셋을 구축하였으며, 한글 전처리를 위한 4가지 기법 및 OpenNMT 프레임웍의 LSTM 및 Transformer 모듈을 활용하여 실험을 진행하였다. 이를 통해, 반말 표현을 높임 표현으로 변환하는 최적조합을 확인하였으며, 검증시 BLUE점수로 최대 66.53를 획득하였다.

키워드: Machine Translation, OpenNMT, LSTM, Transformer

I. Introduction

딥러닝 기술의 발달로 인공신경망 기반 기계 번역의 성능이 기존 기법들보다 우수해져 현재 대부분 번역 모델에서 인공신경망 기반 기계 번역 기법이 사용되고 있다. 이 연구에서는 단순히 정형화된 방식을 통해 제한적으로 변환하는 것이 아니라 기계 번역을 통해 문장 속에 담긴 표현을 공손한 표현으로 변환하는 방법을 제안한다. 다른 언어로의 번역이 아닌 같은 언어의 다른 표현으로 번역한다는 특수한 상황에서 공손한 표현으로 변환하는 시도로써 LSTM[1], Transformer[2] 두 모델과 띄어쓰기, MeCab[3], BPE[4], SentencePiece[5]을 조합한 전처리 방법에 따른 BLEU 점수[6]를 고찰한다.

II. Preliminaries

1. Related works

네이버[7]와 카카오[8]에서 제공하는 번역기에서 다른 언어를 한국어로 번역할 때 존대 표현으로 바꾸는 기능을 지원하지만 한국어한국어를 존대 표현으로 바꾸어주는 기능은 제공하지 않는다.

2. Data

DCInside 게시판에서 댓글 약 400만개 가량을 크롤링 하여 그중 10,000개를 선별하여 욕설을 제거하고 높임말로 변환해 데이터 쌍을 구축하였다. AI-HUB[9] 데이터의 높임말은 반말로, 반말은 높임말로 변환해 데이터 쌍에 추가하였다. 변환하지 않는 측면에서 도움을 줄 것 같아 변환되지 않는 데이터도 일부 포함하였다.

III. The Proposed Scheme

구축한 데이터를 사용해 OpenNMT[7]를 활용한 LSTM과 Transformer 두 모델을 가지고 실험을 실시한다. Transformer 모델에는 BASE 옵션[2], OpenNMT 옵션[10]을 적용해 실험한다. 데이터의 전처리 단계에서는 띄어쓰기, MeCab 두 토큰화 방식과 Out of Vocabulary를 해결하기 위한 서브워드 토큰화인 BPE와 SentencePiece를 채택하여 실험한다. 형태소 분석기로 조사를 분리할 경우 성능이 향상되는 실험 결과[11]가 있기에 MeCab을 사용할 때 조사를 분리하였다. 위에서 소개한 MeCab 과 서브워드 토큰화를 조합해 띄어쓰기, SentencePiece, MeCab, BPE 를 각각 단독으로 적용한 전처리와 SentencePiece과 MeCab, BPE를 조합한 전처리를 두 모델에 적용해 실험한다. 서브워드 토큰화를 적용한 경우 띄어쓰기

가 제대로 되지 않아 번역된 문장은 KoSpacing[12]을 통해 후처리를 해준다.

IV. Experimental Results

Table 1. 실험 결과 (S.P: SentencePiece, M: MeCab)

	Spacing	MeCab	BPE	S.P	M+BPE	M+S.P
Base	27.01	53.21	26.91	55.58	44.17	60.06
OpenNMT	28.83	52.56	48.70	58.13	54.97	66.53

Table1. 은 Target 데이터에 KoSpacing을 적용한 것과 하지 않은 것 중 더 높은 BLEU 점수를 표시하였다.

LSTM은 띄어쓰기 전처리로 실험한 결과가 좋지 못해 추가적인 실험을 진행하지는 않았다. Transformer의 Base와 OpenNMT 옵션의 실험 결과는 OpenNMT 옵션이 더 높은 BLEU 점수를 보였고 전처리 방법으로 MeCab과 SentencePiece를 함께 적용했을 때 가장 높은 BLEU 점수를 보였다.

Table 2. 일부 모델의 Cross-Validation

실험 옵션	BLEU	GOOD	BAD
OpenNMT (Mecab+SentencePiece)	66.53	0.74	0.26
OpenNMT (Mecab+BPE)	54.97	0.53	0.47
Base (Mecab+SentencePiece)	60.06	0.61	0.39

Cross-Validation은 번역 결과의 무작위 10%의 문장을 선택하여 4명이 GOOD, BAD로 분류한 비율을 표현하였다. 한국어를 한국어로 번역하는 모델에 대해 BLEU 점수를 통한 정확도 측정 방법이 어느 정도 신뢰도 있는 검증 방법이라는 것을 보여준다.

'다 힘들겠지'를 '다 힘드시겠지요'라고 잘 변환하는 결과도 있지만 '나는'을 '저는'으로 변환하는 데이터가 많아 '이가 나는 것'이라는 문장에서 '이가 저는 것'으로 변환하여 출력되는 결과가 있었다. 문맥에 따라 높임 표현으로 변환되지 않아야 하는 부분이 변환되는 결과가 있다는 점에서 개선이 필요하다.

V. Conclusions

본 논문에서 DCInside 데이터와 AI-HUB 데이터를 활용해 높임 표현과 반말 표현으로 말뭉치 데이터를 구축하였다. 구축한 데이터를 다양한 전처리 방법을 조합하여 LSTM, Transformer 두 모델에 대해 실험을 하였다. 실험 결과를 통해 데이터의 품질, 다양성이 더 필요하다는 것을 알 수 있었고 한국어-한국어 번역 모델의 BLEU 측정이 성능 지표가 될 수 있음을 확인하였다.

REFERENCES

[1] Hochreiter, S., and Schmidhuber, J. (1997). "Long short-term memory". Neural Computation, 9(8), 1735-1780.

[2] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... and Polosukhin, I. (2017). "Attention is all you need". In Advances in NIPS (pp. 5998-6008).

[3] Kudo, T., Yamamoto, K., and Matsumoto, Y. (2004). "Applying conditional random fields to Japanese morphological analysis". EMNLP2004 (pp. 230-237).

[4] Sennrich, R., Haddow, B., and Birch, A. (2015). "Neural machine translation of rare words with subword units". arXiv preprint arXiv:1508.07909.

[5] Kudo, T. and Richardson, J. (2018). "Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing". arXiv preprint arXiv:1808.06226.

[6] Papineni, K., Roukos, S., Ward, T., and Zhu, W. J. (2002). "BLEU: a method for automatic evaluation of machine translation". In Proceedings of the 40th annual meeting of ACL (pp. 311-318).

[7] Papago : https://papago.naver.com

[8] Kakao i : https://translate.kakao.com

[9] AI-HUB : https://aihub.or.kr

[10] OpenNMT : https://github.com/jeongwonkwak/OpenNMT-Project

[11] Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. M. (2017). "Opennmt: Open-source toolkit for neural machine translation". arXiv preprint arXiv:1701.02810.

[12] KoSpacing : https://github.com/haven-jeon/KoSpacing