병렬 코퍼스 필터링과 한국어에 최적화된 서브 워드 분절 기법을 이용한 기계번역

박찬준○, 김경민, 임희석

bcj1210@naver.com, totoro4007@korea.ac.kr, limhseok@korea.ac.kr

고려대학교 컴퓨터학과

Parallel Corpus Filtering and Korean-Optimized Subword Tokenization for Machine Translation

Chanjun Park○ , Gyeongmin kim, Heuiseok Lim

Korea University Dept.Computer Science

요 약

딥러닝을 이용한 Neural Machine Translation(NMT)의 등장으로 기계번역 분야에서 기존의 규칙 기반,통계기반 방식을 압도하는 좋은 성능을 보이고 있다. 본 논문은 기계번역 모델도 중요하지만 무엇보다 중요한 것은 고품질의 학습데이터를 구성하는 일과 전처리라고 판단하여 이에 관련된 다양한 실험을 진행하였다. 인공신경망 기계번역 시스템의 학습데이터 즉 병렬 코퍼스를 구축할 때 양질의 데이터를 확보하는 것이 무엇보다 중요하다. 그러나 양질의 데이터를 구하는 일은 저작권 확보의 문제, 병렬 말뭉치 구축의 어려움, 노이즈 등을 이유로 쉽지 않은 상황이다. 본 논문은 고품질의 학습데이터를 구축하기 위하여 병렬 코퍼스 필터링 기법을 제시한다. 병렬 코퍼스 필터링이란 정제와 다르게 학습 데이터에 부합하지 않다고 판단되며 소스, 타겟 쌍을 함께 삭제 시켜 버린다. 또한 기계번역에서 무엇보다 중요한 단계는 바로 Subword Tokenization 단계이다. 본 논문은 다양한 실험을 통하여 한-영 기계번역에서 가장 높은 성능을 보이는 Subword Tokenization 방법론을 제시한다. 오픈 된 한-영 병렬 말뭉치로 실험을 진행한 결과 병렬 코퍼스 필터링을 진행한 데이터로 만든 모델이 더 좋은 BLEU 점수를 보였으며 본 논문에서 제안하는 형태소 분석 단위 분리를 진행 후 Unigram이 반영된 SentencePiece 모델로 Subword Tokenization를 진행 하였을 시 가장 좋은 성능을 보였다.

1. 서 론

과거 기계번역 연구는 규칙기반 및 통계기반 방식을 이용했으나 최근에는 신경망 기반 방식으로 많은 기술적인 성과를 이루어냈다. [1,2,3]. GPU의 등장으로 인한 컴퓨팅 파워의 개선, Tensorflow ,Pytorch등의 오픈소스 프레임워크의 등장으로 인한 개발환경의 개선, 빅데이터의 등장, 좋은 알고리즘의 개발 등의 요인으로 현재 딥러닝을 이용한 다양한 분야에서 엄청난 성과를 보이고 있으며 기계번역도 마찬가지이다.

딥러닝에서 좋은 알고리즘을 이용한 모델의 아키텍쳐를 구성하거나 Hyperparameter Tuning을 이용한 모델 최적화를 하는 행위도 중요하지만 무엇보다 중요한 것은 고품질의 학습데이터를 구성하는 일이다. 즉 인공신경망 기계번역 시스템의 학습데이터 즉 병렬 코퍼스를 구축할 때 양질의 데이터를 확보하는 것이 무엇보다 중요하다. 그러나 양질의 데이터를 구하는 일은 저작권 확보의 문제, 병렬 말뭉치 구축의 어려움 등을 이유로 쉽지 않은 상황이다. 본 논문은 고품질의 학습데이터를 구축하기 위하여 병렬 코퍼스 필터링 기법을 제시하며 이를 통해 데이터 양이 더 적어짐에도 더 좋은 성능을 냄을 보인다. 즉 학습데이터의 양이 중요한 것이 아닌 질이 중요함을 본 논문을 통하여 증명한다. 또한 기계번역에서 가장 중요한 단계는 바로 전처리 단계에서 문장을 어떻게 Subword Tokenization을 진행할 것이냐 이다. 대게 많은

논문들이 단순히 Byte Pair Encoding(BPE) 알고리즘[4]을 이용한 Subword Tokenization 방법을 사용한다. 본 논문은 다양한 실험을 통하여 한국어-영어 기계번역에 최적화된 Subword Tokenization 방법론을 제시한다.

2. 병렬 코퍼스 필터링

Parallel Corpus Filtering이란 양질의 Parallel Corpus를 구축하기 위한 작업이며 좋은 품질의 Corpus만을 선별하는 작업을 의미한다. WMT(World Machine Translation)에서 매년 Shared Task를[5] 열고있으며 Zipporah등 오픈소스도 존재한다.[6]

웹데이터에서 데이터를 크롤링하여 학습데이터로 사용할 경우 많은 노이즈 때문에 학습데이터에 적합하지 않은 데이터들이 많은데 이것을 일일이 사람이 검증하는 것은 쉽지 않은 일이다. 따라서 병렬 코퍼스 필터링 기법을 이용하여 자동으로 학습데이터에 적합하지 않은 데이터를 선별하여 제거하고자 한다.

최근 많은 기계번역 논문들을 보면 데이터 증강 기법으로 Back Translation과 Copied Translation등을 이용하면서 합성 코퍼스를 만들어 성능을 올리려는 사례가 많다. 발상을 달리하여 기계번역에 하위분야 중 하나인 병렬 코퍼스 필터링 기법을 이용하여 데이터를 증강시키는 것이 아닌 데이터를 오히려 줄임으로 성능이 더 높아짐을 보이려 한다. 즉 양이 많은 코퍼스보다 양이 적어도 양질의 코퍼스가 더 좋은 모델을 만드는데 기여함을 증명하고자 한다.

2.1. 제안하는 병렬 코퍼스 필터링 프로세스

Filtering 방법으로 다양한 방식이 있으며 우리는 [7]의 프로세스를 적용하였다.

총 6단계를 거쳐 Parallel Corpus Filtering을 거치게 되며 각 단계의 내용은 아래와 같다.

(1)Unique parallel sentence Filter: 소스와 타겟 쌍이 중복되는 경우 제거한다.

(2)Equal source-target Filter – 소스와 타겟 문장이 동일할 경우 제거한다.

(3)Multiple sources - one target and multiple targets - one source Filters – 동일한 소스 문장이 다양한 타겟 문장과 연결되어 있는 경우나 다양한 소스 문장이 동일한 타겟 문장과 연결되어 있는 경우, 이를 제거한다.

(4)Non-alphabetical filters – 한 문장에 해당 언어의 표기 문자가 아닌 문자가 50% 이상 포함되어 있거나, 한 문장이 연결된 문장에 비해 표기 문자가 아닌 문자를 특히 많이(1:3 이상) 포함하고 있을 경우 이를 제거한다.

(5)Repeating token Filter – 같은 단어가 반복되는 경우 이를 제거한다.

(6) Correct language filter – 각 문장 속 언어를 추정하는 언어 식별 소프트웨어(language identification software)를 사용하여 알맞는 언어로 판단되지 않으면 제거한다

3. 한국어에 최적화된 Subword Tokenization

Subword Tokenization이란 기계번역의 입력문장을 일정한 단위로 쪼개 주는 역할을 의미하며 Out of Vocabulary를 해결하기 위하여 기계번역 전처리 단계에서 반드시 거치는 단계이다. 많은 기계번역 논문들을 보면 Tokenize 방법으로 단순히 BPE만을 적용하여 기계번역을 만드는 모습을 볼 수 있다. 본 논문은 다양한 Tokenization 실험을 통하여 단순히 BPE를 적용하는 것보다 더 좋은 성능을 내는 Tokenization 방법론을 제시한다.

한국어는 교착어로 조사에 따라 다양한 표현이 가능하다. 이에 착안하여 형태소 분석 후 조사를 때어내고 이에 SentencePiece Unigram[8]을 적용하였을 시 성능이 높아짐을 보인다.

4. 실험 및 실험결과

4.1 데이터

데이터 같은 경우 한-영 기계번역을 대상으로 실험을 진행할 것이기에 오픈된 한-영 병렬 코퍼스를 사용한다. 영

어 자막 한-영 병렬 코퍼스인 OpenSubtitles2018[1]을 사용한다.

4.2 모델

모델은 모든 실험에 동일하게 Transformer 모델[3]을 사용하며 Hyperparameters는 아래와 같다. [3]의 모델에서 제안하는 모델의 Hyper-Parameter를 그대로 사용하였다.

<표1> 모델 Hyper-Parameter

Hyper-parameter	Setting
Source Vocabulary	32,004
Target Vocabulary	32,002
Batch Size	4,096
Word Vector Size	512
Attention Head	8
Transformer FF	2,048
Dropout	0.1
Optimizer	Adam
Decay Method	Noam

4.3 병렬 코퍼스 필터링 실험

먼저 병렬 코퍼스 필터링을 진행했을 시 성능 향상 여부를 확인하였다. 데이터양은 아래와 같다. 필터링을 통하여 총 442,050 문장이 필터링 되었다.

<표2> 병렬 코퍼스 필터링 실험에 사용한 데이터

	필터링X	필터링 O
학습데이터	1381190	939140
Validation	5000	5000
Test	5000	5000

실험결과는 아래와 같다.

<표3> 병렬 코퍼스 필터링 실험 결과

	BLEU	BLEU1	BLEU2	BLEU3	BLEU4
Filtering	**7.08**	**28.1**	**12.9**	**7.4**	**4.6**
FilteringX	4.18	23.1	9.1	4.5	2.3

필터링을 진행했을 시 BLEU 점수가 약 3점가량 상승하는 것을 알 수 있었으며 이를 통하여 기계번역 전처리 단계에서 병렬 코퍼스 필터링 작업은 필수적인 작업임을 알 수 있었다.

4.4 한국어에 최적화된 Tokenization실험

실험은 크게 3가지 기존에 나와있는 방법론과 2가지 제안하는 방법론을 이용하여 실험을 진행하였다.

1. BPE를 이용한 Tokenization
2. SentencePiece Unigram 정보를 이용한 Tokenization
3. Mecab (형태소분석)을 이용한 Tokenization
4. 조사 분리 후 Sentence Piece Unigram을 이용한 Tokenization
5. 복합명사 분해 + 조사 분리 + SentencePiece Unigram을 이용한 Tokenization

1[4]은 대부분의 기계번역에서 적용하는 Tokenization 방법이며 2[8]는 구글에서 2018년 제안한 Tokenization방법으로 1보다 우수한 것으로 알려져 있다. 3[2]은 한국어 형태소 분석 단위 Tokenization 방법론이다.

4는 본 논문이 제시하는 방법론으로 한국어의 특성에 맞게 조사를 분리하기 위하여 형태소 분석 후 Sentence Piece Unigram 정보를 이용한 Tokenize를 이용하는 방법이다. 즉 2개의 Tokenization 방법론을 파이프라인으로 연결한 2단계 Tokenization 방법론이다.

5는 4에 복합명사 분해를 추가한 방법이다. 복합명사 분해기 같은 경우 [9]의 복합명사 분해기를 이용한다.

지난 실험에서 병렬 코퍼스 필터링을 적용한 것이 성능이 더 좋은 것을 확인했음으로 이번 실험에서는 기본적으로 필터링이 진행된 데이터를 가지고 실험을 진행한다.

실험결과는 아래와 같다.

[1] http://opus.nlpl.eu/OpenSubtitles-v2018.php

[2] https://bitbucket.org/eunjeon/mecab-ko/src/master/

<표4> Subword Tokenization 실험 결과

	BLEU	BLEU1	BLEU2	BLEU3	BLEU4
BASE	7.08	28.1	12.9	7.4	4.6
BPE	9.67	31.8	15.5	9.0	5.6
SP+Unigram	10.49	31.5	15.4	9.0	5.8
MECAB	11.57	34.7	17.9	10.6	6.5
MECAB+SP+UNIGRAM	**12.22**	**35.8**	**18.9**	**11.7**	**7.7**
+ 복합명사분해	11.79	33.8	17.4	10.5	6.7

BASE는 병렬 코퍼스만을 진행하고 Tokenization를 진행하지 않은 모델이다.

먼저 기존 방법론인 BPE와 SentencePiece Unigram정보, 형태소 단위 토큰화에서는 형태소 단위 토큰화가 가장 좋은 성능을 보여주었다. 즉 한국어 관련 기계번역을 만들 때는 BPE나 Sentencepiece를 사용하는 것보다 형태소 단위분리가 더 좋은 성능을 보임을 알 수 있었다.

실험결과 제안하는 방식인 형태소 분석기를 통한 조사 분리 후 Unigram 정보를 이용한 Sentencepiece Unigram 정보를 이용한 Tokenization을 사용하니 기존 방법론인 3가지 방법론보다 BLEU 점수가 향상됨을 확인하였다.

복합명사를 추가로 분해할 경우 기존 3가지 방법론 보다 더 높은 성능을 보였으나 형태소 분석기를 통한 조사 분리 후 Unigram 정보를 이용한 Sentencepiece Unigram정보를 이용한 Tokenizer를 사용한 결과보다는 높지 않은 결과를 보였다. 즉 너무 많이 분리하는 것은 오히려 성능의 악영향을 미칠 수 있음을 시사한다.

5. 결론

NMT의 모델자체를 연구하는 분야도 존재하나 본 연구와 같이 고품질의 학습데이터를 구축하기 위한 연구 및 전처리에 대한 연구도 중요하다고 생각된다. 향후 한국어 병렬 코퍼스에 알맞은 다양한 Parallel Corpus Filtering기법을 연구해볼 예정이며 한국어 뿐만 아니라 영어, 유럽권 언어에 대해서 전처리 기법을 연구해볼 예정이다.

6. 감사의 글

본 연구는 과학기술정보통신부 및 정보통신기술진흥센터의 대학ICT연구센터지원사업 (IITP-2018-0-01405), 2017년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No.NRF-2017M3C4A7068189).

7. 참고문헌

[1]Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural Machine Translation By Jointly Learning To Align and Translate. In ICLR, pages 1–15

[2]Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In Proc of EMNLP.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Advances in neural information processing systems, pages 5998–6008, 2017.

[4] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In Proc. of ACL

[5] Rikters, Matīss, Impact of Corpora Quality on Neural Machine Translation(2018), In Proceedings of the 8th Conference Human Language Technologies - The Baltic Perspective (Baltic HLT 2018) [3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Advances in neural information processing systems, pages 5998–6008, 2017.

[6] H. Xu and P. Koehn, Zipporah: a Fast and Scalable Data Cleaning System for Noisy Web-Crawled Parallel Corpora, Emnlp (2017), 2935–2940. http://www.aclweb.org/anthology/D17-1319%0Ahttp://aclweb.org/anthology/D17-1318.

[7] Rikters, Matīss, Impact of Corpora Quality on Neural Machine Translation(2018), In Proceedings of the 8th Conference Human Language Technologies - The Baltic Perspective (Baltic HLT 2018) [3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Advances in neural information processing systems, pages 5998–6008, 2017.

[8] Taku Kudo, John Richardson, SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing, EMNLP2018

[9] 박찬준, 류법모 (2018), "2 단계 한국어 복합명사 분해기", 제 30 회 한글 및 한국어 정보처리 학술대회 논문집, 2018.10, pp. 495-497