BART를 이용한 한국어 자연어처리: 개체명 인식, 감성분석, 의미역 결정

홍승연[O1], 나승훈[2], 신종훈[3], 김영길[4]

[12]전북대학교, [34]한국전자통신연구원
hongsy034@hanmail.net, nash@jbnu.ac.kr, jhshin82@etri.re.kr, kimyk@etri,re.kr

BART for Korean Natural Language Processing: Named Entity Recognition, Sentiment Analysis, Semantic role labelling

Seung-Yean Hong[O1] , Seung-Hoon Na[2], Jong-Hoon Shin[3], Young-Kil Kim[4]

[12]Jeonbuk National University, [34]ETRI

요 약

최근 자연어처리는 대용량 코퍼스를 이용하여 언어 모델을 사전 학습하고 fine-tuning을 적용함으로 다양한 태스크에서 최고 성능을 갱신하고 있다. BERT기반의 언어 모델들은 양방향의 Transformer만 모델링 되어 있지만 BART는 양방향의 Transformer와 Auto-Regressive Transformer가 결합되어 사전학습을 진행하는 모델로 본 논문에서는 540MB의 코퍼스를 이용해 한국어 BART 모델을 학습 시키고 여러 한국어 자연어처리 태스크에 적용하여 성능 향상 있음을 보였다.

주제어: BART, 언어 모델, Transformer

1. 서론

최근 자연어처리 분야에서는 BERT, XLNet, ALBERT 등 Transformer기반의 대용량 사전 학습 모델에 대한 연구가 활발히 이루어지고 있다[1-4]. BERT, ALBERT는 Auto-Encoding 모델로 양방향의 정보를 이용할 수 있는 장점이 있지만 토큰 예측이 독립적으로 이루어져 dependency를 학습할 수 없고 MASK 토큰이 실제 태스크에는 나타나지 않는 문제가 존재한다. 이러한 단점을 개선하기 위해 Auto-Encoding과 Auto-regressive를 결합한 BART를 [5]에서 제안하였다.

본 논문에서는 적은 양의 코퍼스를 이용해 BART를 사전 학습하여 여러 태스크에서 실험을 진행하였다.

2. 관련 연구

최근 많은 자연어처리 태스크에서 사전 학습 모델을 사용하여 fine-tuning을 진행한 모델이 최고 성능을 보이고 있다. 사전 학습 모델은 사전 학습시 여러 objective가 사용되고 있고 크게 AE(Auto-Encoding)와 AR(AutoRegressive)로 나눌 수 있다. Auto-Encoding은 입력 토큰으로부터 입력 토큰을 그대로 복원하는 학습 방법이다. 사전 학습시에는 주로 Mask 토큰을 두어 Denoising하는 방법을 사용한다. Autoregressive는 이전 토큰들을 통해 다음 토큰을 예측하는 학습 방법이다. 일반적인 언어 모델에서 사용되고 있고 단방향성을 가진다.

BERT[1]는 대표적인 AE방법을 통해 학습을 진행하는 모델이다. BERT는 Transformer 기반 사전 학습 언어 모델로 문장 내에 단어들을 랜덤하게 마스킹하고 마스킹된 단어를 예측하는 문제와 연속하는 두 문장의 순서가 적절한지를 예측하는 문제(Next Sentence Prediction) 2가지 태스크를 이용해 학습을 진행한다. 이를 최적하기 위해 나온 모델은 RoBERTa이다. RoBERTa[2]는 BERT모델에서 NSP를 제거하였고 학습시 마스크의 위치를 고정하지 않고 동적으로 할당하는 Dynamic Msking을 통해 개선된 성능을 보였다. ALBERT[3]도 기존의 BERT가 가지고 있는 NSP 적정성에 의문을 가지고 연속되는 두 문장의 순서가 맞는지를 예측하는 SOP(Sentence Order Prediction)를 제안하였고 학습 효율을 높이기 위해 히든 차원과 단어 차원이 동일한 차원을 가지던 구조를 따로 분리하였고 각 layer의 파라미터를 공유하는 방식을 통해 모델의 파라미터 수를 줄이고 학습 시간을 단축시켰다.

XLNet[4]은 AR이 가지는 단방향성 문제와 AE가 가지는 의존 관계를 모델링할 수 없는 문제를 해결하기 위해 Permutation을 통해 AR모델이 양방향 특성을 가지게 하는 Permutaion 언어 모델을 제안하였다.

BART[5]는 AR과 AE가 결합된 모델로 기존의 Transformer기반 Seq2Seq 구조를 따른다. 원본 텍스트를 임의로 변환시킨 후 원본 텍스트를 복구하도록 학습이 진행된다. 간단한 원본 텍스트에 변형을 통해 기존 사전 학습 모델보다 좋은 성능을 이끌어냈다. 특히 기존에 denoising autoencoder들은 한정적인 변형을 가하였지만 BART는 다양한 변형 방법을 적용할 수 있다는 장점이 있다.

한국어 자연어처리에서도 [6-8]에서 여러 사전 학습 모델이 적용되었고 주로 AE기반 모델에서의 학습이 이루어졌다.

3. BART를 이용한 한국어 자연어처리

본 논문에서는 한국어 BART모델을 학습시키기 위해 540MB의 위키코퍼스를 사용하였다. BART의 구조는 Seq2Seq Transformer 구조 [9]를 따르고 있고 구조는

아래 그림 1과 같다. Seq2Seq Transformer는 인코더 블록과 디코더 블록으로 이루어져 있다. 인코더 블록은 멀티헤드 셀프 어텐션과 Position-wise Feed-forward Network 두 개의 sub-layer로 구성 되어 있고 residual connection을 이용하여 결과를 얻는 구조이다. 디코더 블록은 인코더 구조와 비슷한 구조로 되어 있고 차이점은 인코더 결과를 멀티 헤드 어텐션을 적용하는 sub-layer가 중간에 추가된 점이다. 디코더는 순차적으로 결과를 만들어내는 구조이기 때문에 어텐션시 현재 위치보다 뒤에 나온 토큰들은 마스킹을 진행한다. [9]에서 제안한 모델과 다른점은 활성화 함수를 ReLUs대신 GeLUs[10]를 사용하고 있다. GeLUs는 다른 activation에 비해 좋은 성능을 보여 기존의 ReLUs를 대체하였다. BART 학습을 위한 하이퍼 파라미터는 인코더 블록의 개수 12개, 헤드의 개수 12개, 디코더 블록의 개수 12개, 헤드의 개수 12개, 히든 사이즈 768, 드랍아웃 0.1로 설정하여 학습을 진행하였다.

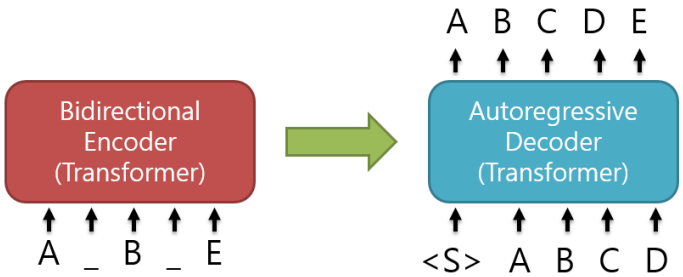

그림 1. BART 모델 구조

3.1 문서 변형(Corrupting Documents)

BART에서 가장 중요한 부분이 입력 텍스트에 변형을 가하는 것이다. [1]의 논문에서는 아래 그림과 같이 다양한 변형을 시도하였다.

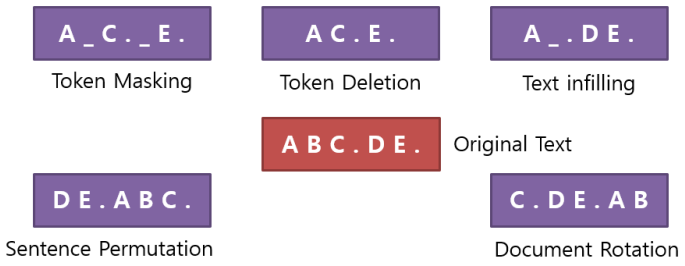

그림 2. 다양한 변형 방법

*Token Masking*은 임의의 토큰을 MASK 토큰으로 변형하는 방법.
*Token Deletion*은 임의의 토큰을 제거하는 방법.
*Text Infilling*은 임의의 Span을 MASK 토큰으로 변형하는 방법.
*Sentence Permutation*은 문서의 문장들을 임의의 순서로 섞는 방법.
*Document Rotation*은 문서에서 임의의 토큰을 선택한 후 시작 토큰으로 설정한 후 재배열 하는 방법.

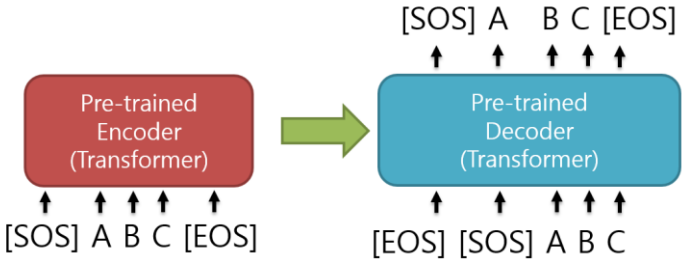

그림 4. Fine-tuning 모델 구조

본 연구에서는 기존 BART 연구에서 가장 좋은 성능을 보인 Text Infilling방법과 Sentence Permutation방법을 사용하여 변형을 하였다. Sentence Permutation을 먼저 적용하고 Text infilling을 수행하였다. *"A B C . D E ."*에서 Sentence Permutation을 수행해 *"D E . A B C ."*를 얻고 Text infilling을 수행하여 *"D E . _ C ."*를 얻는 방식으로 변형을 진행한다.

3.2 하이브리드 토크나이저와 사전 학습

한국어 모델에서의 입력은 일반적으로 형태소 단위의 토큰을 사용하고 있는데 형태소 단위의 토큰은 형태소 분석기의 오류의 전파와 미등록어 문제가 존재한다. 이를 해결하기 위해 [7]에서 하이브리드 토크나이저를 제안하였고 본 논문에서도 이를 채용하였다. 하이브리드 토크나이저는 기본적으로 형태소 단위의 토큰으로 나누고 미등록어인 경우에는 자소 단위 BPE 토크나이저를 사용하여 토큰나이징을 수행한다. 입력은 미등록어가 아닌 경우는 형태소-태그 단위로 구성되고 미등록어인 경우는 자소 단위로 되어있다. 아래 그림 3은 인코더의 입력과 디코더의 입력을 보여준다. 사전 학습을 위한 인코더의 입력은 *"[SOS] 문장 [EOS] [SOS] 문장 [EOS]"* 형태로 이뤄지고 디코더의 입력은 *"[EOS] [SOS] 문장 [EOS] [SOS]"* 형태로 사전 학습을 진행하였다.

원문(문서)
언론/NNG 은/JKS .... 보도/NNG 했/XSV 다/EC ./SF ... ./SF
하이브리드 토크나이징
_언 론 은/JKS .... 보도/NNG 했/XSV 다/EC ./SF ... ./SF
Encoder 모델에서의 입력
[SOS] _언 론 은/JKS .... 보도/NNG 했/XSV 다/EC ./SF [EOS] [SOS] ... [EOS]
Decoder 모델에서의 입력
[EOS] [SOS] _언 론 은/JKS .... 보도/NNG 했/XSV 다/EC ./SF [EOS] [SOS] ...

그림 3. 입력 형태

사전 학습은 변형된 입력 텍스트를 입력으로 디코더의 출력 값을 얻어 원본 문서와의 Cross-entropy를 통해 사전 학습이 진행된다.

4. 한국어 자연어처리 실험

본 논문에서는 감성분석, 의미역 결정, 개체명 인식에 fine-tuning을 적용하여 사전 학습 모델의 성능을 확인하였다. 각 태스크 별 모델은 그림 3과 같이 변형되지 않은 문장을 입력으로 사전학습 인코더와 사전

학습 디코더를 거친 후 디코더의 출력값을 사용하여 fine-tuning을 진행하였다. 실험은 540MB에서 사전 학습한 [6]의 논문의 모델과 BART모델을 사용한 결과를 비교하였다. 실험이 540MB의 저용량에서 이루어졌기 때문에 대용량 실험들과는 따로 비교하지 않았다.

4.1 감성분석

감성분석은 문장으로부터 긍정, 부정 등의 감정을 분석하는 태스크로 주로 긍정/부정으로 이진 분류한다.

한국어 감성분석에 사용한 데이터 셋은 네이버 영화리뷰 감성분석 데이터[11]를 사용하였으며 학습 셋은 142500문장, 개발 셋은 7500문장, 평가 셋은 5만 문장으로 구성되어 있다. 개발 셋은 따로 존재하지 않기 때문에 학습 셋에서 7500문장만 따로 분리하였다. "[SOS] 문장 [EOS]"를 BART모델의 입력으로 하여 Decoder의 출력에서 마지막 벡터를 얻어 긍정, 부정 여부를 판단하였다. LSTM모델은 [6]에서 LSMT만을 사용한 모델의 결과이다.

표 1. 감성분석 성능

모델	ACC
LSTM[6]	79.79%
BERT(형태소 태그)[6]	86.57%
BART Model	**87.03%**

실험 결과 540MB에서 사전 학습한 기존의 모델보다 좋은 성능을 보였다.

4.2 의미역 결정

실험을 위해 Korean Propbank[12]의 Newswire 말뭉치만을 사용하여 학습 데이터를 추출하였고 학습셋은 19602문장, 개발셋은 1152문장, 평가셋 2305문장으로 구성되어 있다. "[SOS] 문장 [EOS]"를 BART모델의 입력으로 하여 Decoder의 출력을 Bi-LSTM CRF에 적용하여 의미역을 결정하였다. 아래 실험은 서술어 인식 및 분류, 논항 인식 및 분류 중 논항 인식 및 분류의 결과를 나타낸다. 평가 지표는 F1-score를 사용하였다. Stacked LSTM-CRF 모델은 [12]에서 제안한 모델이다.

표 2. 의미역 결정 성능

모델	F1
Stacked LSTM-CRF[12]	78.57%
BERT(형태소 태그)[6]	**84.46%**
BART Model	79.14%

실험 결과 540MB에서 사전 학습한 기존의 모델보다 떨어지는 성능을 보였다. 성능 하락의 원인으로 하이퍼 파라미터 튜닝과 형태소 단위 표상 추출 방법의 문제로 보았다.

4.3 개체명 인식

실험을 위해 ETRI의 엑소브레인 언어 분석 말뭉치[13]를 사용하였고 학습셋은 4250문장, 개발셋은 250문장, 평가셋은 500문장으로 구성되어 있다. "[SOS] 문장 [EOS]"를 BART모델의 입력으로 하여 Decoder의 출력을 Bi-LSTM CRF에 적용하여 개체명을 결정하였다. 평가 지표는 F1-score를 사용하였다. LSTM-CRF 모델은 [13]에서 제안한 모델이다.

표 3. 개체명 인식 성능

모델	F1
LSTM-CRF[13]	86.53%
BERT(형태소 태그)[6]	91.58%
BART Model	**91.59%**

실험 결과 540MB에서 사전 학습한 기존의 모델보다 좋은 성능을 보였다.

5. 결론

본 연구에서는 BART 모델을 한국어에 적용하여 여러 태스크에 적용하여 기존의 모델보다 성능 향상을 이뤄짐을 보였다. 향후 생성 모델에 적용한 실험을 진행할 예정이다.

감사의 글

이 논문은 2020년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (R7119-16-1001, 지식증강형 실시간 동시통역 원천기술 개발)

참고문헌

[1] J. Devlin, M. Chang, K. Lee, K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", arXiv:1810.04805, 2018.

[2] Y. Liu, M. Ott, N. Goyal, J.i Du, Mandar S. Joshi, D. Chen, O. Levy, M. Lewis, Luke S. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach", ArXiv, abs/1907.11692, 2019.

[3] Lan, Zhenzhong, et al. "Albert: A lite bert for self-supervised learning of language representations." *arXiv preprint arXiv:1909.11942*, 2019.

[4] Yang, Zhilin, et al. "Xlnet: Generalized autoregressive pretraining for language understanding." *Advances in neural information processing systems*. 2019.

[5] Lewis, Mike, et al. "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension." *arXiv preprint arXiv:1910.13461*, 2019.

[6] 박광현, 나승훈, 신종훈, 김영길, "BERT를 이용한 한국어 자연어처리: 개체명 인식, 감성분석, 의존 파싱, 의미역 결정", 한국정보과학회 학술발표논문집, 407-409, 2019.

[7] 민진우, 나승훈, 신종훈, 김영길, "RoBERTa 를 이용한 한국어 자연어처리: 개체명 인식, 감성분석, 의존파싱". 한국정보과학회 학술발표논문집, 407-409, 2019.

[8] 이영훈, 나승훈, 최윤수, 이혜우, 장두성, “ALBERT를 이용한 한국어 자연어처리: 감성분석, 개체명 인식, 기계독해”, 한국정보과학회 학술발표논문집, 332-334, 2020.
[9] Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems*. 2017.
[10] Hendrycks, Dan, and Kevin Gimpel. "Gaussian error linear units (gelus)." *arXiv preprint arXiv:1606.08415*, 2016.
[11] https://github.com/e9t/nsmc
[12] 배장성, 이창기, “Stacked Bidirectional LSTM-CRFs를 이용한 한국어 의미역 결정”, KIISE 2017.
[13] 나승훈, 민진우, 문자 기반 LSTM CRF를 이용한 개체명 인식, KCC, 2016.