ELECTRA 기반 순차적 문장 분류 모델

최기현†, 김학수††, 양성영†††, 정재홍†††, 임태구†††, 김종훈†††, 박찬규†††
강원대학교†, 건국대학교††, 삼성물산†††
pluto32@kangwon.ac.kr, nlpdrkim@konkuk.ac.kr
mike.yang@samsung.com, jackslo@samsung.com, taegu.lim@samsung.com,
jh0010.kim@samsung.com, helpme@samsung.com

Sequential Sentence Classification Model based on ELECTRA

Gi-Hyeon Choi, Hark-Soo Kim,
Seong-Yeong Yang, Jae-Hong Jeong, Tae-Gu Lim, Jong-Hoon Kim, Chan-Kyu Park
Kangwon National University, Konkuk University, Samsung C&T

요 약

순차적 문장 분류는 여러 문장들을 입력으로 받아 각 문장들에 대하여 사전 정의된 라벨을 할당하는 작업을 말한다. 일반적인 문장 분류와 대조적으로 기준 문장과 주변 문장 사이의 문맥 정보가 분류에 큰 영향을 준다. 따라서 입력 문장들 사이의 문맥 정보를 반영하는 과정이 필수적이다. 최근, 사전 학습 기반 언어 모델의 등장 이후 여러 자연 언어 처리 작업에서 큰 성능 향상이 있었다. 앞서 언급하였던 순차적 문장 분류 작업의 특성상 문맥 정보를 반영한 언어 표현을 생성하는 사전 학습 기반 언어 모델은 해당 작업에 매우 적합하다는 가설을 바탕으로 ELECTRA 기반 순차적 분류 모델을 제안하였다. PUBMED-RCT 데이터 셋을 사용하여 실험한 결과 제안 모델이 93.3%p로 가장 높은 성능을 보였다.

주제어: 순차적 문장 분류, 문맥 정보, 사전 학습 기반 언어 모델, ELECTRA

1. 서론

매년 약 250만 건 이상의 과학 분야 논문들이 나오고 있으며[1], 이로 인하여 사용자가 여러 논문들로부터 원하는 정보를 효율적으로 검색하는 과정이 점점 더 어려워지고 있다[2]. 이와 같은 이유로 사용자가 원하는 정보를 효율적으로 찾을 수 있도록 도와줄 수 있는 자동화 도구의 필요성이 나날이 커지고 있다. 사용자는 특정 분야에 대한 관련 논문을 검색할 때, 해당 논문의 요약문을 통해 본인이 찾고자 하는 정보와의 관련 여부를 확인한다. 이러한 과정은 요약문을 구성하고 있는 각 문장들이 수사학적 역할(소개, 방법론, 결론 등)에 맞추어 적절히 구조화 된다면 보다 신속하게 이뤄질 수 있다[2]. 그러나 기존의 과학 분야 요약문들 중 상당 부분이 비정형적이며 이로 인해 사용자의 정보 검색 과정에 방해 요소로 작용한다. 따라서 본 논문에서는 원하는 정보를 신속하게 검색할 수 있도록 하기 위하여 순차적 문장 분류(Sequential Sentence Classification) 작업을 통해 요약문을 구성하고 있는 각 문장들에 수사학적 역할을 자동으로 할당하는 딥러닝(Deep Learning) 기반 접근 방식을 제안하고자 한다.

순차적 문장 분류는 일련의 문장들을 사전 정의된 라벨에 맞추어 분류하는 작업을 의미한다[3]. 대표적인 예시로 과학 분야 학술 논문의 요약문을 구성하고 있는 각 문장들을 대응하는 수사학적 역할에 따라 분류하는 작업이 있다[4]. 문장 수준의 데이터를 입력 받아 해당 문장에 대한 분류를 수행하는 일반적인 문장 분류(Sentence Classification)와 달리 순차적 문장 분류는 문서 혹은 단락(이하 문서) 수준의 데이터를 입력 받는다. 그리고 입력 문서는 순차적인 문장 구조를 갖고 있으며, 문서 내 문장에 대한 정확한 분류를 위해 주변 문장들과의 문맥 정보를 필요로 한다[2]. 본 논문에서는 문장들 사이의 문맥 정보를 반영하여 문서 내에 존재하는 각 문장들에 대한 순차적 문장 분류 작업을 수행하기 위해 사전 학습 기반의 언어 모델을 사용한다.

최근, 문맥 정보를 반영한 언어 표현을 위한 여러 사전 학습 기반 언어 모델이 연구되어 많은 자연 언어 처리 작업의 성능이 향상되었다[3]. 특히, 이와 같은 사전 학습 기반 언어 모델들 중 가장 대표적인 모델인 BERT(Bidirectional Encoder Representation from Transformers)[5]는 여러 트랜스포머 계층(Transformer Layer)[6]으로 구성되어 있으며, 대용량의 일반 말뭉치를 사용하여 마스킹 된 단어 예측(Masked Language Modeling) 작업과 다음 문장 예측(Next Sentence Prediction) 작업을 통해 사전 학습 된다[5]. BERT는 공개 당시, 여러 자연 언어 처리 작업에서 가장 높은 성능을 보였으며 공개 이후, BERT의 문제점을 개선한 여러 모델이 등장하였다[7-8]. ELECTRA[8]는 BERT의 사전 학습 과정에서 적용하였던 입력 토큰들 중 일부 토큰을 [MASK] 토큰으로 변경하여 입력으로 사용하고 변경한 토큰의 원래 입력을 복원하는 마스킹 된 단어 예측 작업을 다음과 같이 변경하였다. 우선, 입력 토큰들 중 일부를 [MASK] 토큰으로 변경하고 변경한 토큰에 대하여 생성기(Generator)에서 의미적으로 적절한 토큰을 생성한다. [MASK] 토큰 대신 임의의 토큰으로 바뀐 전체 입력 토큰들을 판별기(Discriminator)의 입력으로 사용하며, 판별기에서는 입력 토큰이 기존 입력과 동일한지 또는 생성기에 의해 변경되었는지 여부를 예측한다. 이를 통해 기

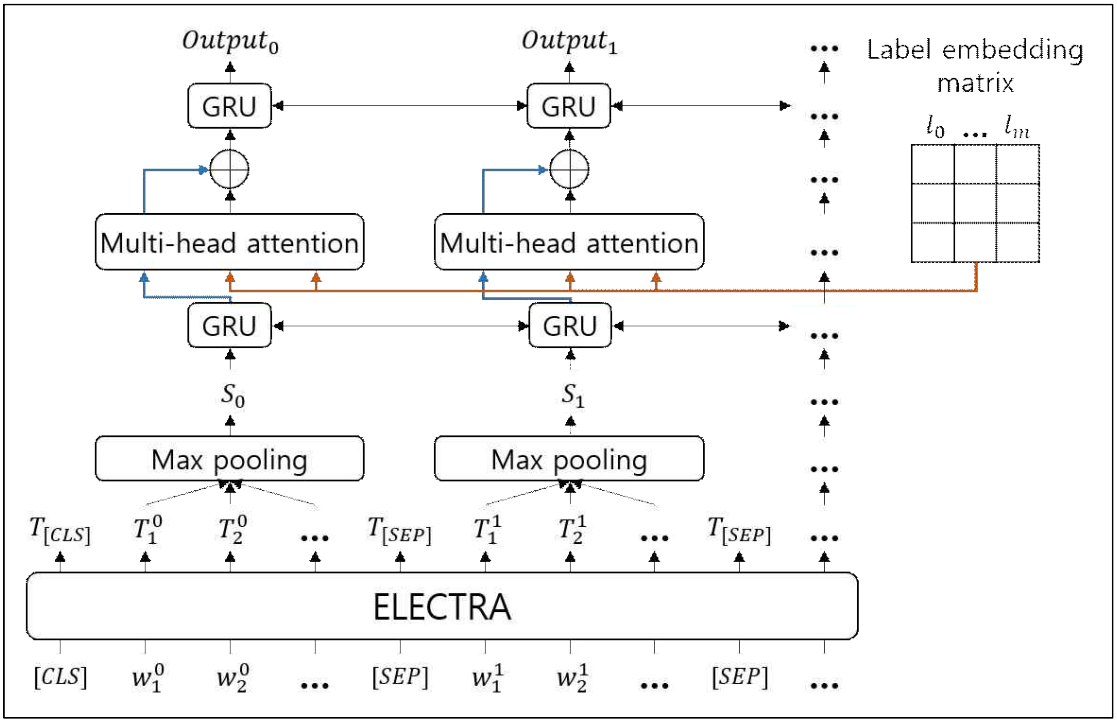

그림 1 제안 모델 전체 구조도

존 BERT 보다 효율적으로 학습할 수 있었으며, 여러 자연 언어 처리 작업에서도 향상된 성능을 보였다. 이와 같은 이유로 본 논문에서는 ELECTRA를 기반으로 문장들 사이의 문맥 정보가 반영된 문장 벡터를 생성하고, 생성된 문장 벡터를 활용하여 순차적 분류를 수행하는 딥러닝 모델을 제안한다.

2. 관련 연구

순차적 문장 분류 작업에 대한 기존 연구는 의학 분야 논문 요약문의 수사적 구조 분석에 집중되어 있었다[2]. 주로 Naive Bayes[9], Support Vector Machine[10], Hidden Markov Model[11]을 기반으로 하는 연구가 진행되었다. 하지만 이러한 연구들은 전문가에 의해 구축된 자질에 많은 의존도를 보이며 자질 구축에 많은 시간과 자원을 소비하는 단점을 갖고 있다. 딥러닝의 등장 이후, 딥러닝을 기반으로 하는 여러 문장 분류 연구들이 진행되었다[12-13]. 하지만 단일 문장에 대한 입력과 그에 대한 분류를 수행하는 방식으로 문장들 사이의 문맥 정보를 고려하지 못하는 단점을 갖고 있었고, 이는 주변 문장이 기준 문장에 대한 분류를 수행하는데 영향을 주는 순차적 분류 작업에서 성능 저하의 원인이 되었다. 최근, 순차적 문장 분류 작업에 보다 직접적으로 접근한 딥러닝 기반 연구들도 진행되었다[14-16]. [14]는 현재 발화에 대한 발화 의도를 분류할 때, 이전 발화에 대한 정보를 반영하여 분류 성능을 높인 모델을 제안하였다. [15]는 Bidirectional LSTM을 사용하여 각 문장을 인코딩한 후 CRF 계층을 사용하여 서로 다른 문장들 사이의 문맥 정보를 반영할 수 있는 모델을 제안하였다. [3]은 BERT를 사용하여 문서 내 문장들의 문맥 정보가 반영된 문장 벡터를 생성하여 분류를 수행하는 모델을 제안하였다.

3. 제안 모델

제안 모델의 전체 구조도는 그림 1과 같다. 문서 내에 존재하는 각 문장들은 ELECTRA를 통해 인코딩 된다. 각 문장을 구성하고 있는 단어들에 대응하는 ELECTRA 출력들을 대상으로 max pooling 연산을 수행하여 문장들 사이의 문맥 정보가 반영된 문장 벡터를 생성한다. 생성된 문장 벡터들은 첫 번째 Bidirectional GRU 계층의 입력으로 사용된다. 다음으로 멀티 헤드 어텐션 메커니즘(Multi-head Attention Mechanism)[6]을 사용하여 각 문장과 분류할 라벨 사이의 관계 정보를 벡터화한 자질 벡터를 생성한다. 그리고 문장 벡터와 해당 문장 벡터에 대응하는 자질 벡터를 연결(Concatenate)한 값을 두 번째 Bidirectional GRU 계층의 입력으로 사용하여 생성된 출력 벡터를 바탕으로 분류를 수행한다.

3.1 문장 벡터 생성

그림 2는 입력 문서 내 첫 번째 문장에 대응하는 벡터를 생성하는 과정에 대한 예시이다. 12개의 트랜스포머 계층으로 구성된 ELECTRA 모델의 마지막 계층의 계산 과정을 다음과 같이 수정한다. 첫 번째 문장에 대한 문장에 대응하는 벡터를 생성하기 위해서 아래 예시와 같이 첫 번째 문장과 관련된 토큰에는 1, 나머지 토큰들에 대해서는 0의 값을 갖는 attention mask를 적용하여 계산한다. 이를 통해 첫 번째 문장에 관련된 토큰 별 출력을 생성할 수 있다. 다음으로 마지막 계층에서 첫 번째 문

장을 구성하는 토큰에 대응하는 출력만 선별하여 추출한다. 그리고 max pooling 연산을 적용하여 해당 문장에 대응하는 문장 벡터를 생성한다. 입력 문서를 구성하고 있는 다른 문장들에 대해서도 동일한 과정을 통해 문장 벡터를 생성한다. 입력 문서를 한번에 ELECTRA 모델의 입력으로 사용하여 문장들 사이의 문맥 정보를 반영하면서 동시에 각 문장들에 대응하는 문장 벡터를 생성하기 위해 위와 같은 방식을 적용한다.

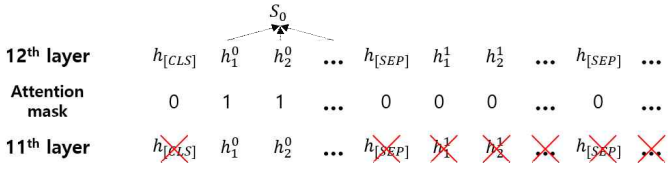

그림 2 문서 내 첫 번째 문장에 대한 벡터 생성 과정

3.2 자질 벡터 생성

인코딩 된 각 문장 벡터들은 문서 내 문장들에 대한 순서 정보를 반영해주기 위해서 첫 번째 Bidirectional GRU 계층의 입력으로 사용된다. 다음으로 멀티 헤드 어텐션 메커니즘을 통해 문장과 분류할 라벨 사이의 관계 정보를 벡터화하여 표현한다. 특정 문장 벡터에 대응하는 자질 벡터는 문장 벡터와 라벨 벡터 사이의 관계 정보를 바탕으로 계산된 어텐션 점수를 바탕으로 전체 라벨 벡터들을 가중합(Weighted-sum)하여 생성되며 이에 대한 수식은 다음과 같다.

$$
\begin{aligned}
&Q = H \\
&K = V = L \\
&Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d}})V \\
&C = MultiHead(Q, K, V) \\
&\quad = Concat(head_0, ..., head_k)W \\
&where\, head_j = Attention(QW^Q, KW^K, VW^V)
\end{aligned}
\tag{1}
$$

수식 (1)에서 H는 첫 번째 Bidirectional GRU 계층의 출력들($S_0', S_1', ..., S_n'$)을 의미하며 이 때, n은 사전 설정한 문서 내 문장의 최대 개수이다. L은 분류할 각 라벨에 대응하는 벡터들($l_0, l_1, ... l_m$)을 나타내며 이 때, m은 분류할 라벨의 개수이다. d는 문장 벡터와 라벨 벡터의 크기, k는 어텐션 헤드의 개수를 의미한다. W, W^Q, W^K, W^V는 무작위로 초기화 된 가중치 행렬을 의미하며 학습 과정 중에 미세 조정(Fine-tuning)된다. 마지막으로 문장 벡터와 자질 벡터를 연결한 값을 두 번째 Bidirectional GRU 계층의 입력으로 사용하며 이에 대한 출력 벡터를 분류기(Classifier)의 입력으로 사용하여 분류를 수행한다.

4. 실험

4.1 데이터 셋

본 논문에서는 제안 모델의 정량적 성능 평가를 위해 PUBMED-RCT 데이터 셋[4]을 사용하여 실험하였다. PUBMED-RCT 데이터 셋은 과학 분야 논문들로부터 추출한 요약문으로 구성되어 있으며, 요약문 내부의 각 문장에 대하여 5개의 수사학적 역할(Background, Conclusions, Methods, Objective, Results)을 부여한 데이터 셋이다. 표 1은 PUBMED-RCT 데이터 셋에 대한 통계를 나타내며 문서 기준으로 표기한 통계이고 괄호 안에 표기된 숫자는 문서 내부 문장들의 개수를 의미한다.

표 1 PUBMED-RCT 데이터 셋 통계 (단위:천)

Train	Valid	Test
15(180)	2.5(30)	2.5(30)

4.2 실험 결과

본 논문에서 수행한 모든 실험에서는 구글에서 공개한 ELECTRA(Base) 모델을 바탕으로 제안 모델 내부의 언어 모델을 초기화 한 후 학습을 진행하였다. 표 2는 제안 모델의 성능 기여를 확인하기 위해 기본 ELECTRA 모델과의 성능 비교를 수행한 결과이며 PUBMED-RCT 데이터 셋을 사용하였다. 기본 ELECTRA 모델은 기존 연구들 중 최고 성능을 보인 [3]의 방식을 바탕으로 입력 토큰들 중 각 [SEP] 토큰에 대응하는 ELECTRA 모델의 출력값을 문장 벡터로 사용하였고 문장 벡터에 단일 완전 연결 신경망(Fully-connected Layer)만을 적용하여 분류하였다. 검증 데이터 셋을 사용하여 가장 높은 성능을 보인 모델을 바탕으로 평가 셋을 적용하여 측정한 micro F1 점수를 표기하였다.

표 2 기본 ELECTRA 모델과 비교 실험

모델	micro F1 점수
기본 ELECTRA 모델	92.27
제안 모델	93.3

실험 결과 제안 모델이 기본 ELECTRA 모델보다 더 높은 성능을 보임을 알 수 있었고, 제안 모델이 순차적 문장 분류에 보다 효과적임을 보여주고 있다. 표 3은 PUBMED-RCT 데이터 셋을 사용하여 기존 연구들과 비교 실험을 수행한 결과를 보여준다. 위에서 수행한 실험과 동일하게 검증 데이터 셋을 사용하여 가장 높은 성능을 보인 모델을 바탕으로 평가 셋을 적용하여 측정한 micro F1 점수를 표기하였다. 실험 결과 제안 모델이 93.3%p로 가장 높은 성능을 보이는 것을 확인 할 수 있었다. 특히, [3]은 과학 분야 관련 일반 말뭉치를 사용하여 사전 학습을 수행한 SCIBERT[16]의 가중치로 BERT를 초기화한 모델을 바탕으로 학습한 모델이다. 하지만 제안 모델이 사용한 ELECTRA 모델은 일반 말뭉치를 바탕으로 사전 학습한 언어 모델임에도 불구하고 더 높은 성능을 보이고 있는 것을 통해 제안 모델에서 적용한 방법론이 순차적 문장 분류 성능 향상에 기여하였음을 알 수 있었다.

표 3 기존 연구들과 비교 실험

모델	micro F1 점수
Dernoncourt et al., 2016 [15]	90.0
Jin et al., 2018 [2]	92.6
Cohan et al. 2019 [3]	92.9
제안 모델	93.3

5. 결론

본 논문에서는 ELECTRA 기반 순차적 문장 분류 모델을 통해 과학 분야 관련 요약문을 구성하는 각 문장들의 수사학적 역할을 자동으로 할당해주는 순차적 문장 분류 작업에 대한 연구를 진행하였다. 제안 모델의 정량적인 성능 평가를 위해 PUBMED-RCT 데이터 셋을 사용한 실험에서 기존 연구들보다 높은 성능을 보여줌으로써 제안 모델이 순차적 문장 분류 작업에 적합하다는 것을 보였다. 하지만 제안 모델은 사전 학습 기반 언어 모델을 사용하고 있기 때문에 입력 길이에 제한이 있으며, 문서 수준 입력을 사용하는 순차적 문장 분류 작업의 특성상 이러한 입력 길이 제한 문제는 매우 치명적이다. 따라서 향후 연구로 입력 길이 제한 문제를 해결할 수 있는 연구를 진행할 예정이다.

감사의 글

본 연구는 삼성물산 산학연구용역 과제의 지원을 받아 수행되었음.

참고문헌

[1] M. Ware and M. Mabe, "The STM report: An overview of scientific and scholarly journal publishing," 2015.

[2] D. Jin and P. Szolovits, "Hierarchical Neural Networks for Sequential Sentence Classification in Medical Scientific Abstracts," In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp.3100-3109, 2018

[3] A. Cohan, I. Beltagy, D. King, B. Dalvi, and D. S. Weld, "Pretrained Language Models for Sequential Sentence Classification," In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 36844-3690, 2019

[4] F. Dernoncourt and J. Y. Lee, "PubMed 200k RCT: a Dataset for Sequential Sentence Classification in Medical Abstracts," In: Proceedings of the Eighth International Joint Conference on Natural Language Processing, Vol. 2, pp.308-313, 2017

[5] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," https://arxiv.org/abs/1810.04805, 2018

[6] A. Vaswani, N. Shazeer, N. Parmar J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," In Advances in neural information processing systems, pp. 5998-6008, 2017.

[7] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," arXiv preprint arXiv:1909.11942, 2019.

[8] K. Clark, M. Luong, Q. V. Le, and C. D. Manning, "ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators," https://arxiv.org/abs/2003.10555, 2020.

[9] P. Ruch, C. Boyer, C. Chichester, I. Tbahriti, A. Geissbühler, P. Fabry, J. Gobeill, V. Pillet, D. Rebholz-Schuhmann, C. Lovis, and A. Veuthey, "Using argumentation to extract key sentences from biomedical abstracts," International Journal of Medical Informatics, Vol.76, No.2, pp.195-200, 2007.

[10] Y. Liu, F. Wu, M. Liu, and B. Liu, "Abstract sentence classification for scientific papers based on transductive svm," Computer and Information Science, Vol.6, No.4, pp.125, 2013.

[11] J. Lin, D. Karakos, D. D. Fushman, and S. Khudanpur, "Generative content models for structural analysis of medical abstracts," In Proceedings of the hlt-naacl bionlp workshop on linking natural language and biology, pp.65-72, 2006.

[12] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," https://arxiv.org/abs/1607.01759, 2016

[13] A. Conneau, H. Schwenk, L. Barrault, and Y. Lecun, "Very Deep Convolutional Networks for Text Classification," https://arxiv.org/abs/1606.01781, 2016.

[14] J. Y. Lee and F. Dernoncourt, "Sequential Short-Text Classification with Recurrent and Convolutional Neural Networks," https://arxiv.org/abs/1603.03827, 2016.

[15] F. Dernoncourt, J. Y. Lee, and P. Szolovits, "Neural Networks for Joint Sentence Classification in Medical Paper Abstracts," https://arxiv.org/abs/1612.05251, 2016.

[16] I. Beltagy, K. Lo, and A. Cohan, "SciBERT: A Pretrained Language Model for Scientific Text," https://arxiv.org/abs/1903.10676, 2019.