과학 논문 초록 말뭉치 구축 및 선학습 트랜스포머 기반 초록 자동구조화 방법

김서경 ㅇ 조윤희 허세훈, 정상근*

충남대학교 컴퓨터공학과
kskyung0624@gmail.com, yunhui.cho98@gmail.com, tpgns5247@gmail.com, hugman@cnu.ac.kr

Scientific Paper Abstract Corpus and Automatic Abstract Structure Parsing using Pretrained Transformer

Seokyung Kim^{O} Yunhui Cho Sehun Heo Sangkeun Jung*
Department of Computer Science & Engineering, Chungnam National University

요 약

논문 초록은 논문의 내용을 요약해 제시함으로써 독자들의 연구결과물에 대한 빠른 검색과 이해를 도모한다. 초록의 구성은 대부분 전형적인 경우가 많기 때문에, 초록의 구조를 자동 분석하여 색인해두면 유사구조 초록을 검색하거나 생성하는 등의 연구효율화에 기여할 수 있다. 허세훈 외 (2019)는 초록 자동구조화를 위한 말뭉치 SPA2019 및 기계학습기반의 자동구조화 방법을 제시하였다. 본 연구는, 기존 SPA2019 의 구조화 오류를 바로잡고, SPA2019 에서 추출한 1,346 개의 초록데이터와 2,385 개의 초록데이터를 추가한 SPA2020 말뭉치를 새로이 소개한다. 또한, 다양한 선학습 기반 트랜스포머들을 활용하여 초록 자동구조화를 수행하였으며, 그 결과 BERT-0.86%, RoBERTa-0.86%, ALBERT-0.84%, XLNet-0.86%, DistilBERT-0.85% 등의 자동구조화 성능을 보임을 확인하였다.

주제어: 기계학습, 선학습 기반 트랜스포머, 초록 구조화

1. 서론

과학 논문에서의 초록은 논문의 내용을 요약해서 보여주는 역할을 한다. 초록 내용의 구조나 키워드 등이 색인되어 관리된다면, 독자들이 시스템을 활용하여 유사 구조를 가지는 초록 등을 빠르게 검색하거나 혹은 자동 생성함으로써 연구 검색 혹은 저작 활동의 효율화에 크게 기여할 수 있을 것이다.

[1]에서는 과학논문의 초록 자동구조화와 말뭉치의 중요성에 대해 언급하였고, 약 1,010 개의 초록 및 6,896 개의 문장으로 구성된 말뭉치(이하, SPA2019 로 호칭)를 소개하였다.

본 연구는, SPA2019 의 태그셋을 재정의하고, 기존의 오류에 대해 수정하는 작업을 진행하였으며, SPA2019 에서 추출한 200 개의 초록에 381 개의 초록을 추가 수집하여 3,731 개의 구조단위를 태깅한, 과학 논문 초록 말뭉치 SPA2020 을 소개한다.

또한, Roberta, DistilBERT, RoBERTa, ALBERT, XLNet, DisTilBERT 등의 다양한 선학습 트랜스포머 인코더 기반의 초록 구조 자동화 방법을 활용하여 SPA2020 의 구조화 성능에 대해 보고한다.

본 논문의 구조는 아래와 같다. 2 장에서는 과거에 선행된 본 논문과 유사한 과학논문들에 관한 내용을 서술한다. 3 장에서는 정의한 태그셋과 SPA2020 말뭉치에 대해 소개하고, 데이터에 대한 통계를 보여준다. 4 장에서는 사용한 선학습 기반 초록 구조화 모델들과 각 모델 별 정확도와 F1 Score 를 살펴보고 실험 방법 및 결과를 제공한다. 결론은 5 장에서 제공한다.

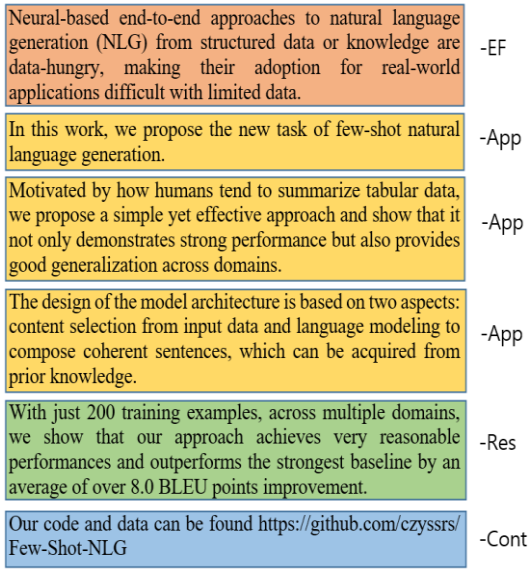

그림 1. 과학 논문 초록 분석 예시

ㅇ 주 저자(Lead Author)
* 교신저자(Correspoding Author)

2. 관련 연구

과거에도 논문 초록을 활용하려는 다양한 연구가 진행되었다. [1]에서는 초록의 자동구조화를 위한 말뭉치를 구축하고 사전 학습된 BERT, RoBERTa 모델을 기본 모델로 하여, 최적학습(Fine-Tunning)하는 방법을 제안하였다. 본 논문에서는 제안된 말뭉치에 데이터 추가와 태그 분류 축소 등의 방법으로 개선하여 새로운 버전으로 제안한다. [2]에서는 텍스트 마이닝 기법을 활용하여 논문 초록에서 나오는 용어들의 출현 빈도를 분석을 통해 기후변화 관련 논문들의 추세와 관심 주제어를 파악하는 연구가 진행되었으며, [3]에서는 제목을 입력으로 사용하여 Attentive neural sequence-to-sequence 를 기반으로 초록을 자동으로 생성하는 것을 보여주었으며, [4]에서는 인용 논문의 효과적인 분류를 위해서 과학 논문의 구조 정보를 인용으로 통합하는 연구가 진행되었으며, [5]에서는 한글 논문 초록에 존재하는 연구방법론을 다양한 알고리즘 (SGD, CNN, LSTM, CNN+Bi-directional GRU, CNN+GRU)을 통해 분류에 있어서 어느 알고리즘이 좋은 성능을 나타내었는지 보여주었다. [6]에서는 절차적 지식 모델링을 활용한 의학 문헌의 초록을 5 개의 의미적인 블록 정보로 분류하였으며 이를 통 하여 목적/해법 문장 추출과 단위 절차 정보 추출 실험을 진행하였다. [7]에서는 생명과학 및 생물의학 주제에 대한 초록 문장 분류에 대한 정보를 담은 MEDLINE 데이터베이스를 통해 자유 검색 엔진 서비스를 제공하였다.

3. SPA2020 말뭉치

3.1 태그셋 정의

[1]의 연구에서는 과학 논문 초록에 공통적으로 나타나는 주요 구조를 분석해 각 문장을 4 개의 대분류, 11 개의 중분류, 1 개의 소분류로 구분하여 총 12 개의 기준으로 분류했다. 우리는 위 연구에서 정확도가 실험결과의 성능이 낮은 점과 (BERT 기준 0.80), 과도하게 세분화된 분류로 인해 사용자가 사용하기가 어렵고 직관력이 떨어진다는 문제점에 집중하였다.

위의 문제를 해결하기 위해, 기존 태그셋에서 중·소분류를 없애고, 대분류만을 활용하였다.

결론적으로 무작위로 샘플링된 과학 논문 초록 581 개에 대해 2 인 데이터 작업자 간의 상호일치율이 약 94%에 달함을 확인하였다.

다음은 정의한 태그셋에 대한 설명이다.

1) EF : 해당 논문에서 제안하는 연구 분야에 대한 사전 설명 Tag 이다. 연구 분야에 대한 배경 설명 및 문제점, 선행 연구 소식·문제점·결과 등이 이에 해당한다.

2) App : 해당 논문에서 제시한 방법론 설명 Tag 이다. 주 방법론 및 방향, 방법론에 대한 구체화된 설명, 방법론의 목표, 방법론 실험 내 수행 및 수행방법 등이 이에 해당한다.

표 1. 데이터 분류 기준에 따른 누적 수 및 비율

분류	태그	정의	누적 수(개)	비율(%)
Establishing Field	EF	분야설명 분야의 문제점 최근 연구 및 소식	1,048	0.28
Approach	App	주 방법론/방향 방법론 구체화 방법론 내 수행 방법론의 목표	1,768	0.47
Result	Res	실험 결과 Competition 결과	783	0.21
Contribution	Cont	코드/결과물 공유 분야에 대한 공헌	132	0.04

3) Res : 해당 논문에서 실행한 실험의 결과에 관한 Tag 이다. 수행한 실험에 대한 수치 결과, 방법론 성능 결과, Competition 결과 등이 이에 해당한다.

4) Cont : 해당 논문의 공헌에 관한 Tag 이다. 코드 및 결과물 공유, 방법론의 분야에 대한 공헌 등이 이에 해당한다.

3.2 말뭉치

본 논문에서 정의한 SPA2020 말뭉치는 SPA2019 의 말뭉치에서 추출한 초록 데이터 1,346 건에 2,385 건을 더한 총 3,731 건의 과학 논문 초록 구조단위들로 이루어져 있다. 우리는 과학 논문 초록에 공통적으로 나타나는 주요 구조를 분석하여 초록의 각 문장을 총 4 개의 대분류로 구분하였다.

381 건의 과학 논문 초록 데이터들은 ACL Anthology (https://www.aclweb.org/anthology/)에서 ACL 학술대회의 2020 년 논문 초록들을 수집한 것이다.

3.3 초록 구조화 예

표 2. SPA2020 태그 기준에 따른 문장의 예시

태그	문장
EF	This paper introduces the Webis Gmane Email Corpus 2019, the largest publicly available and fully preprocessed email corpus to date.
App	We crawled more than 153 million emails from 14,699 mailing lists and segmented them into semantically consistent components using a new neural segmentation model.
Res	With 96% accuracy on 15 classes of email segments, our model achieves state-of-the-art performance while being more efficient to train than previous ones.
Cont	All data, code, and trained models are made freely available alongside the paper.

위의 표 2 는 본 논문에서 제안한 태그셋을 기준으로 과학 논문 [8]의 초록에 데이터 작업자가 태깅을 진행한 예시이다.

4. 문장 분류 성능 평가

4.1 선학습 기반 초록 구조화 모델

BERT[9]는 Transformer 인코더에 기반하여 주어진 문장에 대한 문맥 정보를 양방향으로 이해할 수 있는 모델 중 하나이다. RoBERTa[10]의 경우 BERT 가 위치정보를 적극적으로 고려하지 않는 점을 극복한 모델 중 하나이다. XLNet[11]은 양방향 학습을 진행하여 AR 과 AE 를 취합한 모델이다. ALBERT[12]는 모델의 크기와 비례하는 BERT 의 특징을 개선하여 모델의 크기를 줄인 모델이다. DistilBERT[13]는 BERT 의 큰 크기와 연산속도를 개선한 모델이다.

4.2 실험 및 결과

본 장에서는 논문 초록 각각의 문장을 제안한 태그와 데이터 셋으로 분류하는 실험을 진행한다. 따라서 학습된 모델의 성능을 Accuracy 와 F1 Score 를 통해 측정하였다.

실험에서는 BERT, RoBERTa, ALBERT, XLNet, DistilBERT 를 통해 테스트 셋 30%, 훈련 셋 70%의 비율로 문장 분류 성능 평가를 수행하였다. BERT 는 Base-Cased 모형을 사용하였고, RoBERTa 는 Base 모형을 사용하였다. ALBERT 는 base-v2 모형을 사용하였다. DistilBERT 는 Base-cased 모형을 사용하였다. XLNet 은 base-cased 모형을 사용하였다. 이에 대한 실험 결과는 표 3 과 같다.

표 3 을 보면 EF, App 의 경우, 전반적으로 0.84 이상의 정확도를 보인다. Res 와 Cont 의 경우는 전반적으로 0.75 이상의 정확도를 보인다.

그림 2 에 학습 후, 예측한 값과 실제 값을 비교해서 나타낸 heatmap 을 표시하였다. 가로축은 분류한 태그의 수를 나타내고 세로축은 예측한 태그의 수를 나타낸다.

Res 는 App 과 가장 혼동되어 예측된다. Cont 는 App 과 가장 겹친다. 이 heatmap 을 기반으로 각 태그 간 유사성 혹은 연관성에 대해서 가늠할 수 있다.

태그들 간 서로 혼동하는 이유는 크게 세가지로 예상된다.

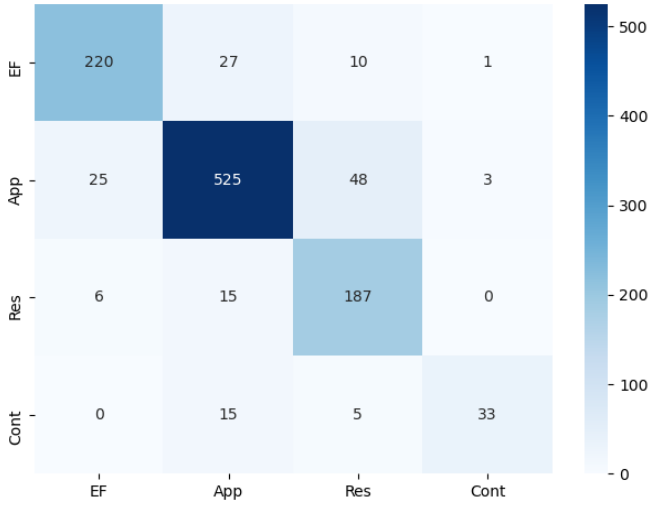

그림 2. 초록 구조 태그 예측 heatmap

1)한 개의 문장이 두 가지 이상의 태그 속성을 가지는 경우
2)문장 자체로만 보았을 때와 초록 전체를 들여다보았을 때의 태그가 다른 경우
3)기계가 배경 지식에 대해 알 수 없는 경우

1)의 경우, “After describing our generative model and an inference procedure for it, we compare gimVI to alternative methods from computational biology or domain adaptation on real datasets and outperform Seurat Anchors, Liger and CORAL to impute held-out genes.” 문장을 예로 들 수 있다. 위 문장은 논문 저자가 논문에서 수행한 내용에 관한 문장이기 때문에 App 태그에 속한다. 그러나, 실험 결과 내용 또한 문장에 들어있기 때문에 Res 의 속성 또한 가지고 있다.

2)의 경우, “This motivates us to propose a new residual unit, which makes training easier and improves generalization.”문장을 예로 들 수 있다. 위 문장은 문장 자체로만 보았을 경우, 방법론에 대한 속성을 설명하는 문장이라 판단해, App 태그로 판단할 수도 있다. 그러나, This 가 이전 문장의 실험에 대한 내용을 의미한다면, 실험 결과에 대한 내용이므로 Res 태그에 해당하게 된다.

3)의 경우, “P-values lower than 0.005 were considered statistically significant.”문장을 예로 들 수 있다. 위 문장은 논문의 주된 방법론에 대한 배경 설명이기 때문에 EF 태그로 분류되어야 한다. 그러나, 기계는 이러한 정보가 배경 정보인지, 방법론에 대한 설명인지를 구분할 수 없다.

표 3. 태그 및 모델별 초록 구조 태깅 성능 표

태그	BERT		RoBERTa		ALBERT		XLNet		DistilBERT	
	Accuracy	F1 Score	Accuracy	F1 Score	Accuracy	F1 Score	Accuracy	F1 Score	Accuracy	F1 Score
EF	0.85	0.85	**0.89**	**0.86**	0.86	0.85	0.88	**0.86**	0.84	0.84
App	0.89	0.88	0.89	0.88	0.87	0.88	**0.90**	**0.89**	0.89	0.87
Res	**0.78**	0.81	0.76	0.81	0.75	0.78	0.75	0.82	0.76	**0.83**
Cont	0.86	0.71	0.78	0.68	0.84	0.69	**0.90**	0.73	0.89	**0.80**
Total	**0.86**	0.81	**0.86**	0.81	0.84	0.80	**0.86**	**0.84**	0.85	**0.84**

5. 결론

본 논문에서는 과학 논문 초록의 문장들을 자동 분류하는 방법을 제안한 [1]의 연구를 확장한 SPA2020 말뭉치를 제안했다. 또한 수집한 데이터셋을 이용하여 BERT, RoBERTa, ALBERT, XLNet, DistilBERT 와 같은 선학습 모델을 통해 학습하고 평가하였다.

향후 연구에서는 본 연구에 문장 간의 관계를 살펴보아 오류를 줄이는 방법을 추가할 예정이다. 추가로, 키워드와 문장 태그를 입력하여 학습된 데이터에서 문장을 추천·제안하는, 논문 초록 작성에 도움을 주는 연구를 진행할 예정이다.

감사의 글

이 논문은 2020 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(No.2020-0-01441)

이 논문은 2019 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (2019-0-00004, 준지도학습형 언어지능 원천기술 및 이에 기반한 외국인 지원용 한국어 튜터링 서비스 개발)

이 논문은 2019 년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(2019R1F1A1060601)

참고문헌

[1] 허세훈, 이병만, 정상근, “초록 구조 분석 및 기계학습을 통한 자동 초록 구조화 방법”, *KSC*, 2019.

[2] 배규용, 박주현, 김정선, & 이영섭, “텍스트 마이닝 기법을 활용한 기후변화관련 식품분야논문초록 분석”, 한국데이터정보과학회지, 24(6), 1429-1437, 2013.

[3] Wang, Q., Zhou, Z., Huang, L., Whitehead, S., Zh ang, B., Ji, H., & Knight, K., “Paper abstract writing through editing mechanism”, arXiv preprint arXiv:1 805.06064, 2018.

[4] Cohan, Arman, Waleed Ammar, Madeleine van Zuylen, and Field Cady, "Structural Scaffolds for Citation Intent Classification in Scientific Publications.", arXiv preprint arXiv:1904.01608, 2019.

[5] 최종윤, 정유철, “과학 기술 연구 동향 파악을 위한 연구방법론의 문장 분류”, 한국 HCI 학회 학술대회, Feb:728-31, 2019.

[6] 송사광, 오흥선, 최윤정, 장혜주, 맹성현, 최성필, & 최윤수, “의료 문헌에서의 절차적 지식 추출을 위 한 단위 절차 추출 연구”, 한국정보과학회 학술발표논문집, 38(1A), 154-157, 2011.

[7] https://www.ncbi.nlm.nih.gov/pubmed/

[8] Janek Bevendorff, Khalid Al Khatib, Martin Potthast, Benno Stein. "Crawling and Preprocessing Mailing Lists At Scale for Dialog Analysis”, *ACL*, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 1151–1158, 2020.

[9] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding.", arXiv preprint arXiv:1810.04805, 2018.

[10] Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov, "Roberta: A robustly optimized bert pretraining approach.", arXiv preprint arXiv:1907.11692, 2019.

[11] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, Quoc V. Le, “XLNet: Generalized Autoregressive Pretraining for Language Understanding”, arXiv preprint arXiv:1906.08237, 2019.

[12] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, Radu Soricut, “ALBERT: A Lite BERT for Self-supervised Learning of Language Representations”, arXiv preprint arXiv:1909.11942, 2019.

[13] Victor Sanh, Lysandre Debut, Julien Chaumond, Thomas Wolf, “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter”, arXiv preprint arXiv:1910.01108, 2020.