KITE: 한국어 고의 오탈자를 활용한 텍스트 데이터 증강 방법론

Text Data Augmentation Methods using Korean Intentional Typographical Errors

유승욱	최주환	서희재	진교훈	김영빈
Seunguk Yu	Juhwan Choi	Heejae Suh	Kyohoon Jin	Youngbin Kim
중앙대학교 소프트웨어학부 School of Computer Science and Engineering, Chung-Ang University seungukyu@gmail.com	중앙대학교 전자전기공학부 School of Electrical and Electronics Engineering, Chung-Ang University gold5230@cau.ac.kr	중앙대학교 컴퓨터공학부 School of Computer Science and Engineering, Chung-Ang University linkyouhj@gmail.com	중앙대학교 첨단영상대학원 Graduate School of Advanced Imaging Science, Multimedia & Film, Chung-Ang University fhzh123@cau.ac.kr	중앙대학교 첨단영상대학원 Graduate School of Advanced Imaging Science, Multimedia & Film, Chung-Ang University ybkim85@cau.ac.kr

요약문

최근 자연어 데이터를 더욱 효율적으로 활용하기 위해 다양한 텍스트 증강 기법이 제안되고 있다. 본 연구에서는 온라인 상의 댓글 등에서 나타나는 오탈자에 초점을 맞추어 한국어 고의 오탈자를 활용한 텍스트 데이터 증강 기법을 제안한다. 다른 언어와는 달리 한 글자 내에서 다양한 오탈자가 발생할 수 있는 한국어의 특징을 활용해 고의 오탈자를 활용한 텍스트 증강 규칙을 규정하였다. 본 논문에서 제안하는 기법을 통해 증강한 문장은 포함된 오탈자가 늘어남에 따라 기존 문장과의 코사인 유사도 차이가 커짐을 확인하였다. 제안한 방식으로 실험 결과 NSMC, Korean Hate Speech 데이터셋에서 평균 1.1%의 성능 향상을 이룰 수 있었으며, 기존에 제안된 방법과 비교했을 때 더욱 효율적임을 알 수 있었다.

주제어

딥러닝, 자연어처리, 데이터 증강

1. 서 론

1.1 글꼴

최근 텍스트 데이터를 처리하는데 있어 가장 핵심적으로 사용되는 방법은 딥러닝이다. 자연어 처리에 활용되는 모델은 주어진 데이터의 특징을 학습하여 텍스트 분류와 같은 다양한 작업에 활용된다. 그렇기에 딥러닝 기반 모델의 성능을 결정하는 가장 중요한 요소는 데이터의 양과 질이라고 할 수 있다[1]. 그러나 정해진 맞춤법에 따라 정제된 데이터를 통해서만 학습된 모델은 학습이 완료된 이후 오탈자 등의 노이즈가 포함된 실제 데이터에 대해서는 기대한 성능을 온전히 발휘하지 못할 우려가 있다.

텍스트 데이터 중 특히 댓글을 활용한 데이터는 정해진 맞춤법에서 어긋나거나 오타를 포함하는 경우가 빈번하다. 이같은 온라인의 글은 비격식적인 상황에서 작성되었기에 오탈자가 포함되었더라도 보통은 이를 수정하지 않고 방치하는 경우가 대다수이다. 본 연구에서 온라인 상의 텍스트로 구축된 데이터셋을 살펴본 결과 다수의 경우에 오탈자가 포함되어 있음을 확인하였다. 본 연구에서는 이러한 현상에 주목하여 한국어 문장에 고의 오탈자를 포함시키는 방법을 규정하고 이에 기반한 텍스트 데이터 증강 기법을 제안한다.

데이터 증강이란 보유한 데이터의 양을 늘려 모델 학습에 도움을 주는 기법을 말한다. 이는 일반화 성능을 높임으로써 더욱 강건한 모델을 만드는데 목적을 둔다. Easy Data Augmentation (EDA)[2]은 특정 단어의 교체, 삽입, 순서 변경, 삭제를 통한 간단한 방법으로 텍스트 증강을 시도하였다. 한국어 텍스트를 다룬 연구[3]에서도 마찬가지로 EDA 와 유사한 규칙 기반의 데이터 증강 기법을 사용했으며, 역번역을 통해 문장을 생성하고 입력 오타를 활용하는 방식을 추가하기도 하였다. 그러나 문장에서 특정 단어를 임의로 교체하는 등의 방법은 실제 데이터에서 찾아보기 힘든 경우로, 그 특성상 새로 생성된 문장의 의미가 기존 문장과 달라질 가능성을 가진다.

반면 한국어의 특성을 통해 생긴 오탈자는 기존 문장 속 단어 의미를 크게 해치지 않는다. 영어 단어에서는 각 글자가 단어 표현에 있어 일정 비중을 차지하기에 'dog'라는 단어에서 'fog', 'dot'과 같이 한 글자만

오탈자가 발생하여도 본래 의미를 보존하지 못한다. 반면 한국어 단어에서는 각 글자가 초성, 중성, 종성으로 구성되기에 '강아지'라는 단어에서 '겅아지', '강아즈'와 같이 자소 단위로 오탈자가 발생하여도 본래 의미를 일부분 추측할 수 있다.

본 논문에서는 한국어 문장에 고의 오탈자를 포함시켜 텍스트 데이터를 증강하는 방법인 Text Data Augmentation using Korean Intentional Typographical Errors (KITE)를 제안한다. 상단의 예시와 같이 기존 문장 속 단어의 의미를 최대한 해치지 않는 선에서, 실제 데이터에 있음직한 텍스트를 만들어내는 것을 목표로 한다.

글자의 초성에서 오탈자가 발생한다면 이는 영어 단어에서의 예시와 같이 단순한 오탈자가 아니라 의도한 단어 의미를 훼손시킬 가능성이 높기에, 오탈자는 중성 혹은 종성의 위치에만 적용시키고 초성에서는 고려하지 않았다. 그 예로 '강아지'라는 단어에서 '장아지', '강아기'와 같이 초성에 오탈자가 발생한 경우는 읽는 이가 원래 의미를 추측하기 어렵기에 즉시 수정하는 경우가 많다. 이렇듯 영어와 같은 다른 언어와는 달리 한 글자가 초성, 중성, 종성으로 구성되는 한국어로 오탈자를 만들 수 있는 경우는 수없이 다양하기에 고의 오탈자가 적용되는 방법을 상세히 규정하였다.

제안하는 방법을 통해 기존 데이터셋에서 한국어 고의 오탈자가 포함된 데이터로 증강하고, 이를 모델 학습에 활용하여 성능 향상이 가능한지 확인하였다. 앞서 코사인 유사도를 통해 언어 모델이 오탈자로 증강한 문장을 기존 문장과는 얼마나 다르게 받아들이는지 확인해 보았으며, 기존의 제안되었던 텍스트 증강 방법들과 KITE 와의 성능을 비교하기 위한 모델 학습 및 검증을 진행하였다.

2. 고의 오탈자를 활용한 문장 생성 기법

본 연구에서 데이터 증강 기법의 수단으로서 활용한 한국어 고의 오탈자 생성 규칙을 설명한다. 오탈자를 포함한 단어 생성은 2.1 에서 다루며, 2.2 에서는 해당 단어들로 구성된 문장 생성 과정을 설명하였다.

2.1 오탈자 단어 생성 과정

오탈자 단어 생성을 위해 한 단어 내에서 자소 자체를 바꾸는 경우 및 자소 순서를 바꾸는 경우로 구분하여 표 1 에 예시와 함께 정리하였다. 두 경우 모두 두글자 이상의 단어에서만 오탈자를 생성하였는데, 보통 관형사나 감탄사로 사용되는 한글자의 단어에서 오타가 나는 경우는 극히 드물었기 때문이다.

표 1. 한국어 고의 오탈자를 활용한 단어 생성

	사례	예시
자소 자체를 바꾸는 경우	일반자음 혹은 일반모음의 변형	설마 → 섷마, 좋았다 → 좋았드
	이중자음 혹은 이중모음의 탈락	많이 → 맣이, 희생 → 히생
	쌍자음의 탈락	그랬다 → 그랫다
	위의 사례들이 혼용된 경우	그랬다 → 그랳다
자소 순서를 바꾸는 경우	첫번째 글자는 중성까지, 두번째 글자는 종성까지 구성된 경우	지금 → 직므, 대중 → 댖우, 조용히 → 종요히
	첫번째 글자는 종성까지, 두번째 글자는 중성까지 구성된 경우	같이 → 가티 둘이서 → 두리서, 좋네요 → 조헤요

자소 자체를 바꾸는 경우는 키보드 상의 위치에서, 기존 자소와 한 칸 이내에 위치한 임의의 자소로 교체하는 방법이다. 일반자음 혹은 일반모음의 변형 사례 중에서는 '설마'라는 단어에서 '섷마'라는 단어로 오탈자를 만든 경우이며, 자음과 모음 총 26 개의 자소에 대해 적용되도록 하였다. 이중자음 혹은 이중모음의 탈락 사례로는 '많이', '희생'에서 두 자소 중 하나가 탈락되어 각각 '맣이', '히생'와 같이 일반자음으로만 남는 형태로 적용되도록 하였다. 특히 '그랬다'과 같은 쌍자음의 경우는 '그랫다'와 같이 일반자음으로 탈락될 수 있으며, '그랳다'와 같이 다른 일반자음으로 변형이 일어날 수도 있다.

자소 순서를 바꾸는 경우는 한 단어 안에서 특정 글자의 초성, 중성, 종성 속 특정 자소를 다른 글자의 초성, 중성, 종성 속 특정 자소와 교체하는 방법이다. 입력된 단어의 첫번째 글자와 두번째 글자가 각각 어떻게 구성되었는지 파악하여 정해진 규칙에 맞게 서로의 자소를 바꿔주었다. 그 예로 '지금'이라는 단어에서 '직므'이라는 단어로 오탈자를 만들거나, '둘이서'라는 단어에서 '두리서'라는 단어로 오탈자를 만든 경우가 이에 속한다.

2.2 오탈자 문장 생성 과정

한국어 고의 오탈자를 활용한 텍스트 데이터 증강을 위해 기존 문장을 몇배의 비율로 증강할 것인지 뜻하는 '증강 비율', 문장 안에서 오탈자의 비율을 얼마로 설정할 것인지 뜻하는 '오탈자 비율'을 사용자가 임의로 설정할 수 있다. 이같이 각 비율을 설정하였을 때

입력된 문장에 대해서 그림 1 의 과정으로 데이터 증강이 이루어진다.
아래 그림 1 의 예시는 오탈자 비율을 80%로 설정하였기에 문장 내 대부분의 단어에서 오탈자가 생성되었다. 자소 순서를 바꾸는 경우가 적용되어 '조금'이라는 단어에서 '족므'라는 단어로 바뀌고, 자소 자체를 바꾸는 경우가 적용되어 '있던데'라는 단어에서 '있던대'라는 단어로 바뀐 결과를 확인할 수 있다.

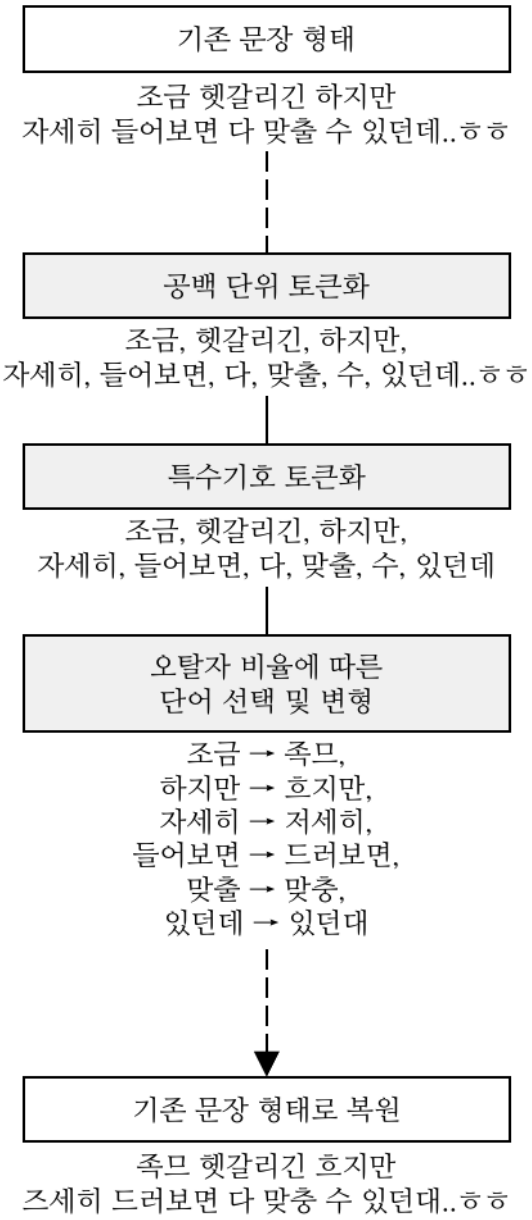

그림 1 한국어 고의 오탈자를 활용한 문장 생성

3. 실험 설계

앞서 제안한 한국어 고의 오탈자를 활용한 텍스트 데이터 증강 기법의 성능을 검증하고자 설계한 실험 과정과 그 결과를 제시한다. 먼저 고의 오탈자로 증강한 문장이 기존의 문장과 얼마나 다른지 비교하고자 임베딩 상에서의 코사인 유사도 차이를 확인해보았으며, 이후 사전 학습된 언어 모델을 텍스트 분류 작업에 Fine-Tuning 하는 과정에 KITE 로 증강한 데이터를 사용함으로써 증강 전후의 성능 차이를 비교하였다.

3.1 실험 데이터셋

텍스트 분류 데이터셋으로는 Naver Sentiment Movie Corpus (NSMC)[4]와 Korean Hate Speech[5] 데이터셋을 사용하였다. NSMC 데이터셋은 네이버 영화 리뷰에 기반한 데이터셋으로 영화를 관람한 사람들의 반응을 긍, 부정으로 이진 분류한 150,000 개의 학습용 데이터로 구성되어 있다. Korean Hate Speech 데이터셋은 인터넷 댓글의 공격성을 3 단계로 분류한 데이터셋으로 약 8,000 개의 학습용 데이터로 구성된다.

3.2 오탈자 비율에 따른 텍스트 증강

본 논문에서 제안하는 KITE 를 활용하여 텍스트를 증강하기 위해서는 문장 안에서 오탈자의 비율을 얼마로 둘 것인지 뜻하는 '오탈자 비율'을 설정해야 한다. 이때 각 비율에 따라 증강한 문장이 기존 문장과 얼마나 다른지 비교하고자 KR-SBERT[6]를 이용하여 임베딩 상에서의 문장 사이 코사인 유사도를 측정하였다.

표 2. KITE 적용 후 코사인 유사도 차이 예시

기존 문장	송중기 시대극은 믿고본다. 첫회 신선하고 좋았다.
증강 문장 예시 1	송중기 시댜극은 미도본다. 처쇠 신선허고 좋았드. (0.735, -0.265)
증강 문장 예시 2	솧중기 세대극은 믿고본다. 처쇠 신선하규 좋았다. (0.580, -0.42)

위의 예시는 오탈자 비율이 80%일 때 증강한 문장이다. 따라서 두 문장 모두 오탈자의 발생 개수는 동일하지만, 오탈자가 발생한 위치에 따라 기존 문장과의 코사인 유사도가 각각 0.735, 0.580 으로 차이나는 모습을 보인다. 표 2 에서 데이터셋 전체로 비교 대상을 늘려 증강 전후의 코사인 유사도를 비교할 때는 Korean Hate Speech 의 테스트 데이터셋을 사용하였고, 증강 비율을 8 배로 두어 많은 양의 증강이 이루어진 상황을 만들었다. 이후 기존 문장과 증강한 문장들 사이 코사인 유사도를 구하여 전체 문장에 대해 평균값을 구하였다.

표 3. KITE 에서 오탈자 비율에 따른 코사인 유사도 차이

오탈자 비율	증강 전후 문장 간 코사인 유사도
20%	0.940, −0.06
40%	0.866, −0.134
60%	0.801, −0.199
80%	0.725, −0.275

그 결과 오탈자 비율이 증가함에 따라 증강한 문장과 기존 문장과의 코사인 유사도가 차이가 커지는 현상을 확인했다. 문장에 많은 오탈자를 추가할수록 기존 문장과 임베딩 상 거리가 멀어지는 것을 의미하며, 이후 활용할 언어 모델이 증강 전후의 문장을 서로 다르게 받아들임을 의미한다. 이는 한국어 자소 단위로 오탈자를 만들면 사람 시각에서는 본래의 의미를 일정 부분 유추할 수 있지만, 언어 모델의 시각에서는 다른 의미를 가진 문장이 된다는 사실을 시사한다.

3.3 증강 방법에 따른 성능 비교

본 논문에서 제안하는 KITE 의 성능을 검증하고자 기존에 제안된 텍스트 증강 방법들과 비교하는 실험을 진행하였다. 사용된 기존 텍스트 증강 방법들 중 Back-Translation (BT)[7]은 가진 텍스트를 외국어로 번역한 뒤 다시 원래의 언어로 재번역함으로써 텍스트를 증강한다. EDA 는 특정 단어의 교체, 삽입, 순서 변경, 삭제를 활용하는 간단한 방식의 텍스트 증강 방법이다.

위의 방법으로 증강한 데이터셋을 한국어로 사전 학습된 KoBERT[8]의 학습 데이터로 이용하여 텍스트 분류를 위한 Fine-Tuning 을 진행하였다. 모델을 학습시키기 위한 Optimizer 로는 AdamW 를 사용하였다. Batch 의 크기는 16 으로, Learning Rate 는 3e-6 으로 지정하고 Cosine Annealing 을 사용하여 20 Epoch 동안 학습을 진행하되 손실값이 4 Epoch 동안 개선되지 않는 경우 Early Stopping 을 주었다. 실험 결과는 표 3 으로 그 중 Original 은 아무런 증강을 거치지 않은 데이터로 실험한 성능을 뜻한다.

표 4. 증강한 데이터셋을 통한 텍스트 분류 성능 비교

	NSMC	Korean Hate Speech	Average
Original	88%	52.8%	70.4%
BT	89.3% (+1.3%p)	53.4% (+0.6%p)	71.35% (+0.95%p)
EDA	88.3% (+0.3%p)	50.6% (−2.2%p)	69.45% (−0.95%p)
KITE	88.7% (+0.7%p)	54.3% (+1.5%p)	71.5% (+1.1%)

사용한 모든 방법에서 증강 비율은 항상 2 배로 두었으며, KITE 에서 오탈자 비율은 NSMC 데이터셋과 Korean Hate Speech 데이터셋에서 모두 60%으로 설정하였다. 실험 결과 각 데이터 증강 방법들을 사용하였을 때 EDA 를 제외하고는 모두 Original 보다 성능 향상을 보였으며, 특히 본 논문에서 제안하는 KITE 가 각 데이터셋에서 기존과 +0.7%p, +1.5%p 으로 평균 +1.1%p 의 차이를 보였다.

특정 단어의 교체 및 삽입 등 간단한 방식을 사용하는 EDA 의 경우 Korean Hate Speech 데이터셋에서 비교적 큰 성능 하락폭을 보였다. 텍스트의 재번역을 통해 증강하는 BT 의 경우 평균 성능의 차이는 KITE 보다 약간 낮으나, 번역에 사용하는 모델에 따라서 증강 소요 시간 및 성능이 일정하지 않으며 다른 증강 방법과 달리 GPU 연산이 요구된다는 단점이 있다. 그에 반해 KITE 를 통한 증강은 별도의 딥러닝 모델을 사용하지 않으며 증강 시간도 각 데이터당 5 분 이내로 신속하게 수행되었다.

4. 결론

본 논문에서는 사람의 실수로 인해 한국어 문장에서 나타나는 오탈자 종류를 상세히 규정하고 이를 활용한 텍스트 데이터 증강 기법인 KITE 를 제안하였다. 이는 별도의 딥러닝 모델을 사용하지 않는 간단한 규칙 기반의 데이터 증강 기법이면서도, 기존 연구에서 주로 사용되었던 특정 단어 교체 및 삽입 등의 방식으로 문장의 본래 의미를 훼손시키지 않는 텍스트 증강 방식이다. 이렇듯 한국어 문장에서의 오탈자가 각 글자의 자소 단위로 발생한다는 점에 기반하여 KITE 를 통한 텍스트 증강 방법을 규정하였다.

본 연구는 실제 데이터에 있을법한 형태로 텍스트 증강을 이루어낸 시도로, 그 과정에서 한국어의 특징을 반영함으로써 증강 과정의 설득력을 확보하였다. 온라인 상에서 일정 양의 한국어 텍스트 데이터를 확보하기 어려울 때 KITE 를 활용하여 오탈자가 포함된 다양한 형태의 한국어 텍스트를 얻을 수 있기를 기대한다.

사사의 글

이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원(NRF-2022R1C1C1008534)을 받아 수행된 연구임.

참고 문헌

1. Luca, A. R., Ursulenau, T. F., Gheorghe, L., Grigorovici, R., Iancu, S., Hlusneac, M. and

Grigorovici, A. Impact of quality, type and volume of data used by deep learning models in the analysis of medical images. Informatics in Medicine Unlocked. 29. Elsevier. 100911. 2022.

2. Wei, J. and Zou, K. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. 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. 6383-6389. 2019.
3. 조진욱, 정민수, 이정훈, 정윤경. 한국어 텍스트 데이터를 위한 변형적 데이터 증강 방법론. 한국정보과학회 2020 한국소프트웨어종합학술대회 논문집. 한국정보과학회. pp. 592-594. 2020.
4. NSMC, Accessed: Nov. 2022. [Online]. Available: https://github.com/e9t/nsmc
5. Moon, J., Cho, W. I., and Lee, J. BEEP! Korean Corpus of Online News Comments for Toxic Speech Detection. In Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media. pp. 25-31. 2020.
6. KR-SBERT, Accessed: Nov. 2022. [Online]. Available: https://github.com/snunlp/KR-SBERT
7. Rico, S., Barry, H. and Alexandra, B. Improving Neural Machine Translation Models with Monolingual Data. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 86-96. 2016.
8. KoBERT, Accessed: Nov. 2022. [Online]. Available:https://github.com/SKTBrain/KoBERT

