토익 문제 풀이 모델 학습을 위한 유의어/반의어 기반 데이터 증강 기법

이정우°†, Aiyanyo Imatitikua Danielle*§, 임희석*†§
고려대학교 컴퓨터학과†, Human-inspired AI 연구소§
{time79779, titi, limhseok}@korea.ac.kr

Synonyms/Antonyms-Based Data Augmentation For Training TOEIC Problems Solving Model

Jeongwoo Lee°†, Aiyanyo Imatitikua Danielle*§, Heuiseok Lim*†§
Department of Computer Science and Engineering, Korea University†
Human-inspired AI Research§

요 약

최근 글을 이해하고 답을 추론하는 연구들이 많이 이루어지고 있으며, 대표적으로 기계 독해 연구가 존재한다. 기계 독해와 관련하여 다양한 데이터셋이 공개되어 있지만, 과거에서부터 현재까지 사람의 영어 능력 평가를 위해 많이 사용되고 있는 토익에 대해서는 공식적으로 공개된 데이터셋도 거의 존재하지 않으며, 이를 위한 연구 또한 활발히 진행되고 있지 않다. 이에 본 연구에서는 현재와 같이 데이터가 부족한 상황에서 기계 독해 모델의 성능을 향상시키기 위한 데이터 증강 기법을 제안하고자 한다. 제안하는 방법은 WordNet을 이용하여 유의어 및 반의어를 기반으로 굉장히 간단하면서도 효율적으로 실제 토익 문제와 유사하게 데이터를 증강하는 것이며, 실험을 통해 해당 방법의 유의미함을 확인하였다. 우리는 본 연구를 통해 토익에 대한 데이터 부족 문제를 해소하고, 사람 수준의 우수한 성능을 얻을 수 있도록 한다.

주제어: 딥러닝, 자연어 처리, 기계 독해, 데이터 증강

1. 서론

현재 딥러닝 기술이 많이 발달되어 여러 분야에서 많은 성과를 거두고 있으며, 사람 수준을 능가하는 모델들 또한 많이 존재한다. 하지만 사람의 능력을 평가하는 실제 시험들에 대해서는 딥러닝 모델로 해결하려는 연구가 많이 진행되고 있지는 않다. 특히 토익은, 과거에서부터 현재까지도 사람의 영어 능력을 평가하기 위한 기준으로써 많이 활용되고 있으나, 이에 대해 딥러닝 모델은 큰 성과를 보이고 있지 않다. 토익에서 Part 1~4는 Listening Comprehension 문제에 해당하고, Part 5~7은 Reading Comprehension 문제에 해당하며, 읽기 능력을 평가하는 Reading Comprehension 문제는 구체적으로 단문 빈칸 채우기(Part 5), 장문 빈칸 채우기(Part 6), 지문의 내용을 이해하고 추론하는 독해 문제(Part 7)로 이루어져 있다. 토익은 이러한 문제 풀이를 통해 사람의 영어 능력을 평가하기 위한 기준으로써 많이 사용되고 있으나, 이에 대해 딥러닝 모델은 사람 수준에서는 매우 쉽게 여겨지는 task에서조차 큰 성과를 보이지 못하고 있다 [1].

우리는 이렇게 딥러닝 모델이 좋은 성과를 내지 못하는 이유를 데이터 문제로 생각한다. 현재 사람의 문제 풀이를 위한 데이터는 많이 존재하나, 기계의 학습을 위해 공개된 데이터는 거의 존재하지 않는다. 더불어 토익 문제 풀이와 관련된 연구가 활발히 이루어지지 않는 이유 또한 관련 데이터가 부족하기 때문인 것으로 사료된다. 즉, 이러한 모델의 훈련 데이터 부족 문제로 인해, 딥러닝 모델이 문장 내에서의 단어 간 관계성 파악에 어려움을 겪게 되어, 사람이 비교적 쉽게 풀 수 있는 문제도 딥러닝 모델은 그에 비해 잘 맞추지 못하는 것으로 파악된다. 이에 우리는 딥러닝 모델의 문장 내 단어 간 관계성 파악 능력을 향상시키고 토익 문제 풀이 모델의 성능을 향상시키기 위해 모델의 훈련 데이터 증강 기법을 제안하였으며, 이를 사용한 결과는 어떠했는지 분석하였다.

우리는 토익의 여러 Part들 중에서 언어 능력의 평가에 유의미하게 사용할 수 있는[1, 2, 3] cloze test[4]인 Part 5 문제에 대해 연구를 수행한다. 또한, 모델의 객관적인 성능 검증을 위해 공식적으로 오픈된 데이터셋이 필수적으로 필요한데, Part 5를 제외한 다른 Part는 현재 공식적인 데이터셋이 존재하지 않는 반면, Part 5는 이를 위한 데이터셋이 Kaggle에 존재하기 때문에 객관적인 성능 검증을 하기에 적합하다[2]. 그러나 이 Kaggle 데이터는 3,625개 밖에 존재하지 않기 때문에, 기계를 학습하기에는 충분한 양이라고 할 수 없다. 이에 본 논문에서는 WordNet을 이용하여 유의어 및 반의어를 기반으로 한 굉장히 간단하면서도 효율적인 데이터 증강 기법을 제안하며, 이를 통해 토익 문제 풀이 task에서 우수한 딥러닝 모델을 생성할 수 있도록 한다.

*교신저자(Corresponding author)
[1]https://github.com/graykode/toeicbert
[2]https://www.kaggle.com/tientd95/toeic-test

2. 관련 연구

빈칸 채우기 task는 크게 주관식 형태의 빈칸 채우기 task와 객관식 형태의 빈칸 채우기 task가 존재한다. 이 중 객관식 형태의 빈칸 채우기 task에서는 적절한 오답 선택지(distractor)를 생성해야 하는데, 이는 실제로 굉장히 큰 비용이 발생하므로 이러한 문제점을 해결하기 위해 자동으로 distractor를 생성하려는 연구들이 이루어지고 있다.

[5] 논문에서는 빈칸 채우기 문제를 만들 때 전문가들이 자신의 경험에 따라 문제를 만들기에 적절한 문장을 선택하고, 문장의 빈칸 부분과 그에 따른 distractor를 생성할 때에도 어느 정도의 휴리스틱한 패턴이 있음에 영감을 받아서 Automatic MCQ Generation System을 제안하였다. CRF를 활용하여 빈칸의 패턴을 추론하며, 대상 단어에 대한 특정한 패턴을 고려하여 distractor를 생성한다. 이후 Google AJAX API를 사용하여 각 distractor candidate와 빈칸 지문을 검색하고, 결과가 없으면 filtering하는 방식을 적용하였다.

[6] 논문에서는 distractor를 생성하는 task에 초점을 맞추어 진행하였다. 이 논문에서는 WordNet과 같은 lexical database를 활용하여 distractor candidates를 생성하는 Candidate Set Generator와, lexical database를 참고하여 생성한 distractor candidates 중에서 가장 적절한 distractor를 선택하는 Distractor Selector를 제안하였다.

[7] 논문에서는 [6] 논문에서 제안했던 Candidate Set Generator, Distractor Selector 구조를 그대로 따르는데, Candidate Set Generator에서 WordNet과 같은 lexical database를 사용하는 것 대신 PLM을 활용해서 distractor를 생성한다는 차이가 있다. 그리고 Distractor Selector에서는 Word Embedding Model인 FastText를 활용해서 정답과 distractor 사이의 유사도를 기준으로 가장 좋은 Distractor를 Rerank하는 방법론을 제안하였다.

본 논문에서는 데이터 증강 연구들에서 자주 사용되는 WordNet을 통해, 유의어와 반의어를 기반으로 데이터 증강의 간편성과 효율성을 극대화하여, 굉장히 간단하면서도 효율적인 데이터 증강 기법을 제안한다.

3. 제안하는 연구

토익의 Part 5 단문 빈칸 채우기 문제는 빈칸이 포함된 문장과 해당 빈칸에 해당하는 정답 선택지, 그리고 오답 선택지로 구성되어 있다. 오답 선택지는 정답 선택지를 고를 때 헷갈릴 만한 것으로 적절하게 생성하는 것이 중요한데, 본 연구에서는 이러한 특징에서 영감을 받아 유의어 및 반의어를 기반으로 하여 데이터를 증강시킬 수 있는 방법을 제안한다. 그림 1은 본 연구에서 제안하는 데이터 증강 방법을 그림으로 표현한 것이다.

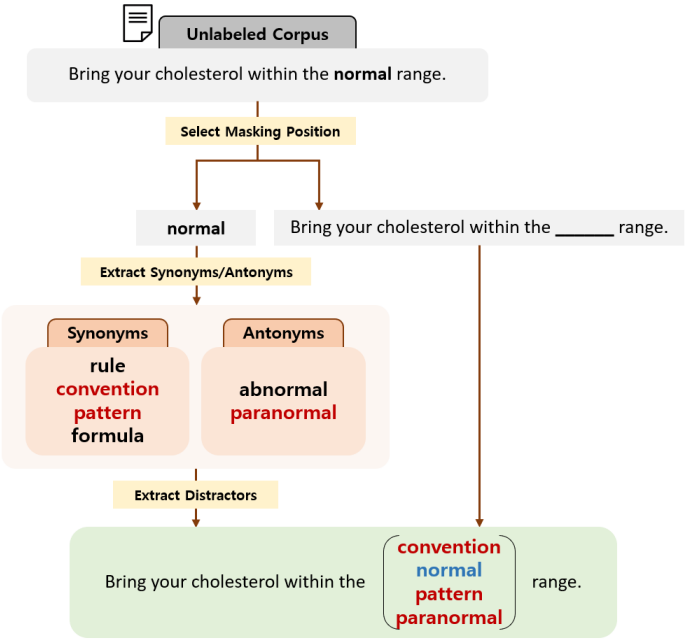

그림 1. 데이터 증강 과정

토익의 Part 5 단문 빈칸 채우기 문제는 빈칸이 있는 하나의 문장이 주어지고, 해당 빈칸에 가장 적합한 단어나 구, 절을 4개의 선택지 중에서 고르는 문제이다. 이와 유사하게 데이터를 증강하기 위해 우리는 AI Hub의 한국어-영어 번역(병렬) 말뭉치에서 영어 문장 데이터를 활용하여, 문장 데이터를 어절 단위로 자르고 그중 하나를 빈칸으로 바꾸어 데이터를 생성하도록 한다. 본 연구에서는 데이터의 품질을 높이기 위해 문장 데이터를 어절 단위로 자르고, 각 어절 양 끝의 구두점 및 특수문자를 제거하여 단어에만 집중할 수 있도록 하였다.

영어 문장 데이터를 어절 단위로 자르고, 각 어절 양 끝의 구두점 및 특수문자를 제거한 후, 각각의 어절을 빈칸으로 하는 토익 데이터를 생성한다. 이때 WordNet을 이용하여 각 빈칸에 들어갈 정답 선택지에 대한 유의어와 반의어를 추출하고, 추출한 유의어/반의어 집합에서 3개를 무작위로 추출하여 오답 선택지를 구성한다. 때로는 유의어와 반의어가 정답이 되는 경우도 있으나, 본 연구에서는 데이터 증강의 간편성과 효율성을 극대화하기 위해 원본 문장에서의 정답이 가장 올바른 답이라고 설정하고, 이러한 데이터가 무수히 많아졌을 때 딥러닝 모델의 문장 내 단어 간 관계성 파악 능력이 향상될 것이라 가정하였으며, 실험을 통해 이를 증명하였다.

4. 실험

본 연구에서는 PLM(Pretrained Language Model)를 활용한 Multiple Choice 관점을 적용하여 모델을 학습한 후 평가를 진행하였다. 사용한 모델은 BERT [8]이며, bert-large-uncased

표 1. Kaggle 토익 데이터와 유의어/반의어 기반 증강 데이터에 대한 실험 결과

Data	Accuracy
Kaggle 토익 데이터	79.06%
유의어/반의어 기반 증강 데이터 + Kaggle 토익 데이터	86.77%

를 사용하여 학습을 진행하였다. batch size는 128, max epoch는 50, Learning rate는 3e-5로 설정하였으며, GPU는 NVIDIA RTX A6000을 사용하였다.

Kaggle 토익 데이터와 본 논문에서 제안한 방법으로 생성한 유의어/반의어 기반 증강 데이터에 대해 학습을 진행하였다. Test는 Kaggle의 토익 데이터 363개로 진행하였으며, 표 1은 실험에 대한 결과이다.

실험에 대한 결과에서 알 수 있듯이, 본 논문에서 제안한 방법으로 생성한 유의어/반의어 기반 증강 데이터가 성능 향상에 유의미한 영향을 준다는 것을 볼 수 있다. 원본 문장에서의 정답이 가장 올바른 답이라고 설정하고, 이러한 데이터를 무수히 많이 증강하여 이를 통해 모델 학습을 진행하는 것이 딥러닝 모델의 문장 내 단어 간 관계성 파악 능력을 향상시키는 것에 도움을 준다는 것을 알 수 있다. 최종적으로 실험 결과를 확인하였을 때, Kaggle 토익 데이터로만 학습을 진행하는 것보다 본 논문에서 제안한 방법으로 데이터를 생성하여 먼저 학습을 진행하고 Kaggle 토익 데이터를 추가적으로 학습하는 것이 훨씬 더 높은 성능을 낸다는 것을 확인할 수 있으며, 데이터 부족 문제 해소에 유의미하게 작용한다는 것을 알 수 있다.

5. 결론

본 연구에서는 사람의 영어 능력 평가를 위해 사용되는 토익에 대해 데이터가 부족한 상황에서 기계 독해 모델의 성능을 향상시킬 수 있는 방안에 대해 연구를 수행하였으며, 유의어/반의어 기반 데이터 증강 기법을 제안하였다. 때로는 유의어와 반의어가 정답이 되는 경우도 있으나, 본 연구에서는 데이터 증강의 간편성과 효율성을 극대화하는 것에 더욱 집중하였으며, 원본 문장에서의 정답이 가장 올바른 답이라고 설정하고 이러한 데이터가 무수히 많아졌을 때 딥러닝 모델의 문장 내 단어 간 관계성 파악 능력이 향상되는 것을 실험을 통해 확인할 수 있었다. 이를 통해 데이터가 부족한 문제가 존재하더라도 기계 독해 모델의 성능을 향상시켜, 우수한 성능을 얻을 수 있다. 향후에는 본 논문에서 제안한 방법을 통해 다른 task에서도 유의미한 성능을 얻을 수 있는지 실험해보고 분석할 예정이다.

감사의 글

본 연구는 과학기술정보통신부 및 정보통신기술기획평가원의 대학ICT연구센터지원사업의 연구결과로 수행되었음(IITP-2022-2018-0-01405). 이 논문은 2021년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(NRF-2021R1A6A1A03045425). 이 논문은 2020년도 정부(과학기술정보통신부)의 재원으로 정보통신기술기획평가원의 지원을 받아 수행된 연구임 (No. 2020-0-00368, 뉴럴-심볼릭 (neural-symbolic) 모델의 지식 학습 및 추론 기술 개발).

참고문헌

[1] S. S. Fotos, "The cloze test as an integrative measure of efl proficiency: A substitute for essays on college entrance examinations?" *Language learning*, Vol. 41, No. 3, pp. 313–336, 1991.

[2] J. Jonz, "Cloze item types and second language comprehension," *Language testing*, Vol. 8, No. 1, pp. 1–22, 1991.

[3] A. Tremblay, "Proficiency assessment standards in second language acquisition research:"clozing" the gap," *Studies in Second Language Acquisition*, Vol. 33, No. 3, pp. 339–372, 2011.

[4] W. L. Taylor, ""cloze procedure": A new tool for measuring readability," *Journalism quarterly*, Vol. 30, No. 4, pp. 415–433, 1953.

[5] T. Goto, T. Kojiri, T. Watanabe, T. Iwata, and T. Yamada, "Automatic generation system of multiple-choice cloze questions and its evaluation," *Knowledge Management & E-Learning*, Vol. 2, No. 3, p. 210, 2010.

[6] S. Ren and K. Q. Zhu, "Knowledge-driven distractor generation for cloze-style multiple choice questions," *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35, No. 5, pp. 4339–4347, 2021.

[7] S.-H. Chiang, S.-C. Wang, and Y.-C. Fan, "Cdgp: Automatic cloze distractor generation based on pre-trained language model," *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 5835–5840, 2022.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.