

(19) 대한민국특허청(KR)
(12) 공개특허공보(A)(11) 공개번호 10-2025-0098328
(43) 공개일자 2025년07월01일

(51) 국제특허분류(Int. Cl.)

G09B 19/06 (2006.01) G06F 40/205 (2020.01)
G06F 40/247 (2020.01) G06F 40/274 (2020.01)
G06Q 50/20 (2012.01) G09B 7/06 (2006.01)

(52) CPC특허분류

G09B 19/06 (2013.01)
G06F 40/205 (2020.01)

(21) 출원번호 10-2023-0189502

(22) 출원일자 2023년12월22일

심사청구일자 2023년12월22일

(71) 출원인

고려대학교 산학협력단

서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)

(72) 발명자

임희석

서울특별시 성북구 안암로 145 자연계캠퍼스 애기
능생활관 311호 NLP&AI 연구실

이정우

서울특별시 동대문구 무학로45길 10, 304호

(74) 대리인

김등용

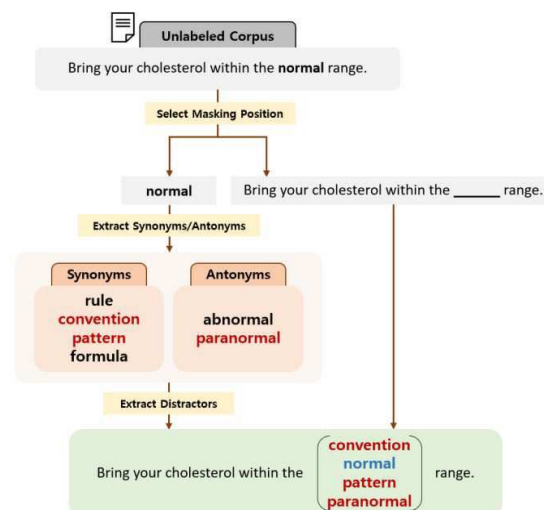
전체 청구항 수 : 총 5 항

(54) 발명의 명칭 토의 문제 풀이 모델 학습을 위한 유의어/반의어 기반 데이터 증강 장치 및 방법

(57) 요약

토의 문제 풀이 모델 학습을 위한 유의어/반의어 기반 데이터 증강 장치 및 방법이 개시된다. 상기 데이터 증강 방법은 적어도 프로세서(Processor)를 포함하는 컴퓨팅 장치에 의해 수행되고, 복수의 영어 문장들을 포함하는 말뭉치가 획득되는 단계, 상기 복수의 영어 문장들 중 하나의 영어 문장을 선택하는 단계, 상기 하나의 영어 문장을 어절 단위로 분할하는 단계, 상기 하나의 영어 문장에서 어느 하나의 어절을 빈칸으로 변환하여, 빈칸을 포함하는 영어 문장과 빈칸과 교체된 정답 선택지를 생성하는 단계, 상기 정답 선택지에 대한 적어도 유의어 및 반의어를 추출하여 오답 선택지 후보를 생성하는 단계, 및 상기 오답 선택지 후보 중에서 3 개를 선택하여, 상기 빈칸을 포함하는 영어 문장, 상기 정답 선택지, 및 3 개의 오답 선택지를 포함하는 증강 데이터를 생성하는 단계를 포함한다.

대표도 - 도2



(52) CPC특허분류

G06F 40/247 (2020.01)

G06F 40/274 (2020.01)

G06Q 50/20 (2013.01)

G09B 7/06 (2013.01)

이 발명을 지원한 국가연구개발사업

과제고유번호	1711193711
과제번호	2018-0-01405-006
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	정보통신방송혁신인재양성
연구과제명	Human-inspired 복합지능 원천기술 연구 및 인력양성
기 여 율	1/2
과제수행기관명	고려대학교산학협력단
연구기간	2023.01.01 ~ 2023.12.31

이 발명을 지원한 국가연구개발사업

과제고유번호	1345365402
과제번호	2021R1A6A1A03045425
부처명	교육부
과제관리(전문)기관명	한국연구재단
연구사업명	이공학학술연구기반구축
연구과제명	Human-inspired AI 연구소
기 여 율	1/2
과제수행기관명	고려대학교
연구기간	2023.03.01 ~ 2024.02.29

명세서

청구범위

청구항 1

적어도 프로세서(Processor)를 포함하는 컴퓨팅 장치에 의해 수행되는 데이터 증강 방법에 있어서,

복수의 영어 문장들을 포함하는 말뭉치가 획득되는 단계;

상기 복수의 영어 문장들 중 하나의 영어 문장을 선택하는 단계;

상기 하나의 영어 문장을 어절 단위로 분할하는 단계;

상기 하나의 영어 문장에서 어느 하나의 어절을 빈칸으로 변환하여, 빈칸을 포함하는 영어 문장과 빈칸과 교체된 정답 선택지를 생성하는 단계;

상기 정답 선택지에 대한 적어도 유의어 및 반의어를 추출하여 오답 선택지 후보를 생성하는 단계; 및

상기 오답 선택지 후보 중에서 3 개를 선택하여, 상기 빈칸을 포함하는 영어 문장, 상기 정답 선택지, 및 3 개의 오답 선택지를 포함하는 증강 데이터를 생성하는 단계를 포함하는 데이터 증강 방법.

청구항 2

제1항에 있어서,

상기 유의어 및 반의어는 어휘 데이터로부터 추출되는,

데이터 증강 방법.

청구항 3

제1항에 있어서,

상기 오답 선택지 후보를 생성하는 단계는,

상기 정답 선택지와 상기 정답 선택지에 대한 유의어 및 반의어를 요청하는 메시지를 어휘 제공 서버로 송신하는 단계; 및

상기 어휘 제공 서버로부터, 상기 정답 선택지에 대한 유의어 및 반의어를 수신하는 단계를 포함하는,

데이터 증강 방법.

청구항 4

제1항에 있어서,

상기 증강 데이터는 TOEIC Part 5의 문제를 풀이하기 위한 토익 문제 풀이 모델을 생성하는 과정에서 학습 데이터로 이용되는,

데이터 증강 방법.

청구항 5

제1항에 있어서,

상기 유의어 및 반의어는 워드넷(WordNet)으로부터 추출되는,

데이터 증강 방법.

발명의 설명

기술 분야

[0001] 본 발명은 딥러닝 학습을 위한 데이터 증강 기법에 관한 것으로, 특히 객관식 형태의 단문 빈칸 채우기 문제 풀이 모델 학습을 위한 데이터 증강 장치 및 방법에 관한 것이다.

배경 기술

[0002] 빈칸 채우기 태스크(task)는 크게 주관식 형태의 빈칸 채우기 태스크와 객관식 형태의 빈칸 채우기 태스크가 존재한다. 이 중 객관식 형태의 빈칸 채우기 태스크에서는 적절한 오답 선택지(distractor)를 생성해야 하는데, 이는 실제로 굉장히 큰 비용이 발생하므로 이러한 문제점을 해결하기 위해 자동으로 오답 선택지를 생성하려는 연구들이 이루어지고 있다.

[0003] Goto 등(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.)은 빈칸 채우기 문제를 만들 때 전문가들이 자신의 경험에 따라 문제를 만들기에 적절한 문장을 선택하고, 문장의 빈칸 부분과 그에 따른 오답 선택지를 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 오답 선택지를 생성한다. 이후 Google AJAX API를 사용하여 각 오답 선택지 후보(distractor candidate)와 빈칸 지문을 검색하고, 결과가 없으면 필터링하는 방식을 적용하였다.

[0004] Ren 등(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.)은 오답 선택지(distractor)를 생성하는 태스크에 초점을 맞추어 진행하였다. 이 논문에서는 WordNet과 같은 어휘 데이터베이스(lexical database)를 활용하여 오답 선택지 후보를 생성하는 후보 집합 생성기(Candidate Set Generator)와, 어휘 데이터베이스를 참고하여 생성한 오답 선택지 후보 중에서 가장 적절한 오답 선택지를 선택하는 오답 선택지 선택기(Distractor Selector)를 제안하였다.

[0005] Chiang 등(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.)은 Ren 등의 논문에서 제안했던 후보 집합 생성기(Candidate Set Generator), 오답 선택지 선택기(Distractor Selector) 구조를 그대로 따르는데, 후보 집합 생성기(Candidate Set Generator)에서 WordNet과 같은 어휘 데이터베이스를 사용하는 것 대신 PLM(Pretrained Language Model)을 활용해서 오답 선택지를 생성한다는 차이가 있다. 그리고, 오답 선택지 선택기(Distractor Selector)에서는 워드 임베딩 모델(Word Embedding Model)인 FastText를 활용해서 정답과 오답 선택지 사이의 유사도를 기준으로 가장 좋은 오답 선택지를 리랭크(Rerank)하는 방법론을 제안하였다.

[0006] 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 듣기 이해력(Listening Comprehension) 문제에 해당하고, Part 5~7은 읽기 이해력(Reading Comprehension) 문제에 해당한다. 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 태스크에서조차 큰 성과를 보이지 못하고 있다.

[0007] 이와 같이, 딥러닝 모델이 좋은 성과를 내지 못하는 이유는 데이터 문제로 판단될 수 있다. 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.

[0008] 이에 본 발명에서는 딥러닝 모델의 문장 내 단어 간 관계성 파악 능력을 향상시키고 토익 문제 풀이 모델의 성능을 향상시키기 위해 모델의 훈련 데이터 증강 기법을 제안한다. 특히, 본 발명은 토익의 여러 파트들 중에서 언어 능력의 평가에 유의미하게 사용할 수 있는 빈칸 채우기 테스트(cloze test)인 Part 5 문제에 대해 것이다 (도 1 참조).

[0009] 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. 그러나, 이 Kaggle 데이터는 3,625개 밖에 존재하지 않기 때문에, 기계를 학습하기에는 충분한 양이라고 할 수 없다. 이에 본 발명에서는 워드넷(WordNet)을 이용하여 유의어 및/또는 반의어를 기반으로 한 효율적인 데이터 증강 기법을 제안하며, 이를 통해 토익 문제 풀이 태스크에서 우수한 딥러닝 모델을 생성할 수 있도록 한다.

선행기술문헌

특허문헌

[0010] (특허문헌 0001) 대한민국 공개특허 제2023-0036623호 (2023.03.15. 공개)
(특허문헌 0002) 대한민국 공개특허 제2022-0068942호 (2022.05.26. 공개)
(특허문헌 0003) 대한민국 공개특허 제2023-0050673호 (2023.04.17. 공개)
(특허문헌 0004) 대한민국 공개특허 제2020-0110400호 (2020.09.23. 공개)

발명의 내용

해결하려는 과제

[0011] 본 발명이 이루고자 하는 기술적인 과제는, 토익 문제 풀이 모델을 위한 유의어/반의어 기반의 데이터 증강 장치 및 방법을 제공하는 것이다.

과제의 해결 수단

[0012] 본 발명의 일 실시예에 따른, 토익 문제 풀이 모델 학습을 위한 유의어/반의어 기반 데이터 증강 방법은 적어도 프로세서(Processor)를 포함하는 컴퓨팅 장치에 의해 수행되고, 복수의 영어 문장들을 포함하는 말뭉치가 획득되는 단계, 상기 복수의 영어 문장들 중 하나의 영어 문장을 선택하는 단계, 상기 하나의 영어 문장을 어절 단위로 분할하는 단계, 상기 하나의 영어 문장에서 어느 하나의 어절을 빈칸으로 변환하여, 빈칸을 포함하는 영어 문장과 빈칸과 교체된 정답 선택지를 생성하는 단계, 상기 정답 선택지에 대한 적어도 유의어 및 반의어를 추출하여 오답 선택지 후보를 생성하는 단계, 및 상기 오답 선택지 후보 중에서 3 개를 선택하여, 상기 빈칸을 포함하는 영어 문장, 상기 정답 선택지, 및 3 개의 오답 선택지를 포함하는 증강 데이터를 생성하는 단계를 포함한다.

발명의 효과

[0013] 본 발명에서 제안한 방법으로 생성한 유의어/반의어 기반 증강 데이터가 성능 향상에 유의미한 영향을 준다. 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 준다. 즉, Kaggle 토익 데이터로만 학습을 진행하는 것보다 본 발명에서 제안한 방법으로 데이터를 생성하여 먼저 학습을 진행하고 Kaggle 토익 데이터를 추가적으로 학습하는 것이 훨씬 더 높은 성능을 내며, 데이터 부족 문제 해소에 유의미하게 작용한다.

도면의 간단한 설명

[0014] 본 발명의 상세한 설명에서 인용되는 도면을 보다 충분히 이해하기 위하여 각 도면의 상세한 설명이 제공된다.
도 1은 토익 Part 5의 단문 빈칸 채우기 문제의 예시를 도시한다.
도 2는 본 발명의 일 실시예에 따른 데이터 증강 방법을 설명하기 위한 흐름도이다.

발명을 실시하기 위한 구체적인 내용

[0015] 본 명세서에 개시되어 있는 본 발명의 개념에 따른 실시예들에 대해서 특정한 구조적 또는 기능적 설명들은 단

지 본 발명의 개념에 따른 실시예들을 설명하기 위한 목적으로 예시된 것으로서, 본 발명의 개념에 따른 실시예들은 다양한 형태들로 실시될 수 있으며 본 명세서에 설명된 실시예들에 한정되지 않는다.

[0016] 본 발명의 개념에 따른 실시예들은 다양한 변경들을 가할 수 있고 여러 가지 형태들을 가질 수 있으므로 실시예들을 도면에 예시하고 본 명세서에서 상세하게 설명하고자 한다. 그러나, 이는 본 발명의 개념에 따른 실시예들을 특정한 개시 형태들에 대해 한정하려는 것이 아니며, 본 발명의 사상 및 기술 범위에 포함되는 모든 변경, 균등물, 또는 대체물을 포함한다.

[0017] 제1 또는 제2 등의 용어는 다양한 구성 요소들을 설명하는데 사용될 수 있지만, 상기 구성 요소들은 상기 용어들에 의해 한정되어서는 안 된다. 상기 용어들은 하나의 구성 요소를 다른 구성 요소로부터 구별하는 목적으로만, 예컨대 본 발명의 개념에 따른 권리 범위로부터 벗어나지 않은 채, 제1 구성 요소는 제2 구성 요소로 명명될 수 있고 유사하게 제2 구성 요소는 제1 구성 요소로도 명명될 수 있다.

[0018] 어떤 구성 요소가 다른 구성 요소에 "연결되어" 있다거나 "접속되어" 있다고 언급된 때에는, 그 다른 구성 요소에 직접적으로 연결되어 있거나 또는 접속되어 있을 수도 있지만, 중간에 다른 구성 요소가 존재할 수도 있다고 이해되어야 할 것이다. 반면에, 어떤 구성 요소가 다른 구성 요소에 "직접 연결되어" 있다거나 "직접 접속되어" 있다고 언급된 때에는 중간에 다른 구성 요소가 존재하지 않는 것으로 이해되어야 할 것이다. 구성 요소들 간의 관계를 설명하는 다른 표현들, 즉 "~사이에"와 "바로 ~사이에" 또는 "~에 이웃하는"과 "~에 직접 이웃하는" 등도 마찬가지로 해석되어야 한다.

[0019] 본 명세서에서 사용한 용어는 단지 특정한 실시예를 설명하기 위해 사용된 것으로서, 본 발명을 한정하려는 의도가 아니다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 명세서에서, "포함하다" 또는 "가지다" 등의 용어는 본 명세서에 기재된 특징, 숫자, 단계, 동작, 구성 요소, 부분품 또는 이들을 조합한 것이 존재함을 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성 요소, 부분품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.

[0020] 다르게 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 모든 용어들은 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 가진다. 일반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥상 가지는 의미와 일치하는 의미를 갖는 것으로 해석되어야 하며, 본 명세서에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않는다.

[0021] 이하, 본 명세서에 첨부된 도면들을 참조하여 본 발명의 실시예들을 상세히 설명한다. 그러나, 특허출원의 범위가 이러한 실시예들에 의해 제한되거나 한정되는 것은 아니다. 각 도면에 제시된 동일한 참조 부호는 동일한 부재를 나타낸다.

[0022] 도 1은 토익의 Part 5 빈칸 채우기 문제의 예시를 도시한다.

[0023] 도 1을 참조하면, 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 헛갈릴만한 것으로 적절하게 생성하는 것이 중요한데, 본 발명에서는 이러한 특징에서 영감을 받아 유의어 및 반의어를 기반으로 하여 데이터를 증강시킬 수 있는 방법을 제안한다.

[0024] 도 2는 본 발명의 일 실시예에 따른 데이터 증강 방법을 설명하기 위한 흐름도이다. 데이터 증강 방법은 적어도 프로세서(Porcessor) 및/또는 메모리(Memory)를 포함하는 컴퓨팅 장치에 의해 수행될 수 있다. 다시 말하면, 데이터 증강 방법을 구성하는 단계들 중 적어도 일부는 컴퓨팅 장치에 포함되는 프로세서의 동작으로 이해될 수 있다. 따라서, 컴퓨팅 장치는 데이터 증강 장치로 명명될 수도 있다. 컴퓨팅 장치는 PC(Personal Computer), 서버(server), 태블릿 PC, 랩탑 컴퓨터 등을 포함할 수 있다.

[0025] 토익의 Part 5 단문 빈칸 채우기 문제는 빈칸이 있는 하나의 문장이 주어지고, 해당 빈칸에 가장 적합한 단어, 구, 또는 절을 4 개의 선택지 중에서 고르는 문제이다.

[0026] 데이터를 증강하기 위해, 임의의 말뭉치가 획득된다. 임의의 말뭉치는 소정의 유무선 통신망을 통하여 말뭉치 제공 서버로부터 수신되거나, USB 메모리 장치와 같은 저장 장치로부터 소정의 입출력 인터페이스를 통해 수신될 수 있다. 실시예에 따라, 임의의 말뭉치는 컴퓨팅 장치에 미리 저장되어 있을 수도 있다. 예시적으로, AI Hub의 한국어-영어 번역(병렬) 말뭉치에서 영어 문장 데이터가 활용될 수 있다.

[0027] 복수의 문장들을 포함하는 말뭉치로부터 어느 하나의 문장이 선택된다. 예시적으로, 선택된 문장은 "Bring your cholesterol within the normal range."일 수 있다. 선택된 문장은 어절 단위로 분할되고, 분할된 어절들 중에서 하나를 빈칸으로 변경할 수 있다. 예시적으로, 선택된 문장의 "normal"이 빈칸으로 변경될 수 있다. 이에 따라, 빈칸을 포함하는 문장(Bring your cholesterol within the ____ range.)과 정답 선택지(normal)가 생성될 수 있다.

[0028] 실시예에 따라, 데이터의 품질을 높이기 위해 문장 데이터를 어절 단위로 자르고, 각 어절 양 끝의 구두점 및 특수 문자를 제거하여 단어에만 집중할 수 있도록 할 수 있다. 이에 따라, (영어) 문장 데이터를 어절 단위로 자르고, 각 어절 양 끝의 구두점 및 특수문자를 제거한 후, 각각의 어절을 빈칸으로 하는 토크 데이터(빈칸을 포함하는 문장과 정답 선택지)를 생성할 수 있다.

[0029] 다음으로, 오답 선택지가 생성된다. 오답 선택지를 생성하기 위해, 정답 선택지에 대한 유의어 및/또는 반의어가 추출될 수 있다. 유의어 및/또는 반의어는 어휘 데이터(어휘 사전, 유의어/반의어 사전 등으로 명명될 수 있음)로부터 추출될 수 있다. 예시적인 어휘 데이터는 워드넷(WordNet)일 수 있다.

[0030] 어휘 데이터는, 유무선 통신망을 통하여, 임의의 시점에 어휘 데이터를 제공하는 서버로부터 수신되거나, USB 메모리 장치와 같은 저장 장치로부터 소정의 입출력 인터페이스를 통해 수신될 수 있다. 실시예에 따라, 어휘 데이터는 컴퓨팅 장치에 미리 저장되어 있을 수도 있다.

[0031] 다른 실시예로, 정답 선택지에 대한 유의어 및/또는 반의어는 유의어 및/또는 반의어를 제공하는 임의의 서버로부터 수신될 수도 있다. 이를 위해, 컴퓨팅 장치는 정답 선택지와 정답 선택지에 대한 유의어 및/또는 반의어를 요청하는 메시지(또는 요청)를 송신하고, 이에 대한 응답으로 정답 선택지에 대한 유의어 및/또는 반의어를 수신할 수 있다.

[0032] 상술한 과정을 통하여 획득된 유의어 및/또는 반의어는 오답 선택지 후보가 된다. 예시적으로, 오답 선택지 후보는 유의어로서 "rule", "convention", "pattern", 및 "formula"를 포함하고, 반의어로서 "abnormal" 및 "paranormal"일 수 있다.

[0033] 오답 선택지 후보 중에서 3 개를 무작위로 추출함으로써, 3 개의 오답 선택지가 마련된다. 이를 통해, 빈칸을 포함하는 (영어) 문장, 정답 선택지, 및 3 개의 오답 선택지를 포함하는 학습 데이터(또는 증강 데이터)가 생성될 수 있다. 예시적으로, 생성된 증강 데이터는 빈칸을 포함하는 문장 "Bring your cholesterol within the ____ range.", 정답 선택지 "normal", 및 오답 선택지 "convention", "pattern", 및 "paranormal"을 포함할 수 있다.

[0034] 또한, 상술한 과정을 반복적으로 수행함으로써, (토크) 문제 풀이 모델을 생성하는 과정에서 활용되는 학습 데이터를 생성할 수 있다.

[0035] 실험 결과

[0036] 본 발명에서는 PLM(Pretrained Language Model)를 활용한 Multiple Choice 관점을 적용하여 모델을 학습한 후 평가를 진행하였다. 사용한 모델은 BERT이며, bert-large-uncased를 사용하여 학습을 진행하였다. batch size는 128, max epoch는 50, Learning rate는 3e-5로 설정하였으며, GPU는 NVIDIA RTX A6000을 사용하였다.

[0037] Kaggle 토크 데이터와 본 발명에서 제안한 방법으로 생성한 유의어/반의어 기반 증강 데이터에 대해 학습을 진행하였다. Test는 Kaggle의 토크 데이터 363개로 진행하였으며, 표 1은 Kaggle 토크 데이터와 유의어/반의어 기반 증강 데이터에 대한 실험 결과를 나타낸다.

[0038] [표 1]

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

[0039]

[0040] 실험에 대한 결과에서 알 수 있듯이, 본 발명에서 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 확인하

였을 때, Kaggle 토익 데이터로만 학습을 진행하는 것보다 본 발명에서 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.

[0041] 이상에서 설명된 장치는 하드웨어 구성 요소, 소프트웨어 구성 요소, 및/또는 하드웨어 구성 요소 및 소프트웨어 구성 요소의 집합으로 구현될 수 있다. 예를 들어, 실시예들에서 설명된 장치 및 구성 요소는, 예를 들어, 프로세서, 컨트롤러, ALU(Arithmetic Logic Unit), 디지털 신호 프로세서(Digital Signal Processor), 마이크로컴퓨터, FPA(Field Programmable array), PLU(Programmable Logic Unit), 마이크로프로세서, 또는 명령(instruction)을 실행하고 응답할 수 있는 다른 어떠한 장치와 같이, 하나 이상의 범용 컴퓨터 또는 특수 목적 컴퓨터를 이용하여 구현될 수 있다. 처리 장치는 운영 체제(Operation System, OS) 및 상기 운영 체제 상에서 수행되는 하나 이상의 소프트웨어 애플리케이션을 수행할 수 있다. 또한, 처리 장치는 소프트웨어의 실행에 응답하여, 데이터를 접근, 저장, 조작, 처리 및 생성할 수도 있다. 이해의 편의를 위하여, 처리 장치는 하나가 사용되는 것으로 설명된 경우도 있지만, 해당 기술 분야에서 통상의 지식을 가진 자는, 처리 장치가 복수 개의 처리 요소(Processing Element) 및/또는 복수 유형의 처리 요소를 포함할 수 있음을 알 수 있다. 예를 들어, 처리 장치는 복수 개의 프로세서 또는 하나의 프로세서 및 하나의 컨트롤러를 포함할 수 있다. 또한, 병렬 프로세서(Parallel Processor)와 같은, 다른 처리 구성(Processing Configuration)도 가능하다.

[0042] 소프트웨어는 컴퓨터 프로그램(Computer Program), 코드(Code), 명령(Instruction), 또는 이들 중 하나 이상의 조합을 포함할 수 있으며, 원하는 대로 동작하도록 처리 장치를 구성하거나 독립적으로 또는 결합적으로(Collectively) 처리 장치를 명령할 수 있다. 소프트웨어 및/또는 데이터는, 처리 장치에 의하여 해석되거나 처리 장치에 명령 또는 데이터를 제공하기 위하여, 어떤 유형의 기계, 구성 요소(Component), 물리적 장치, 가상 장치(Virtual Equipment), 컴퓨터 저장 매체 또는 장치, 또는 전송되는 신호 파(Signal Wave)에 영구적으로, 또는 일시적으로 구체화(Embody)될 수 있다. 소프트웨어는 네트워크로 연결된 컴퓨터 시스템 상에 분산되어서, 분산된 방법으로 저장되거나 실행될 수도 있다. 소프트웨어 및 데이터는 하나 이상의 컴퓨터 판독 가능 기록 매체에 저장될 수 있다.

[0043] 실시예에 따른 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 상기 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 상기 매체에 기록되는 프로그램 명령은 실시예를 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다. 컴퓨터 판독 가능 기록 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(Magnetic Media), CD-ROM, DVD와 같은 광기록 매체(Optical Media), 플롭티컬 디스크(Floptical Disk)와 같은 자기-광 매체(Magneto-optical Media), 롬(ROM), 램(RAM), 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다. 프로그램 명령의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다. 상기된 하드웨어 장치는 실시예의 동작을 수행하기 위해 하나 이상의 소프트웨어 모듈로서 작동하도록 구성될 수 있으며, 그 역도 마찬가지이다.

[0044] 본 발명은 도면에 도시된 실시예를 참고로 설명되었으나 이는 예시적인 것에 불과하며, 본 기술 분야의 통상의 지식을 가진 자라면 이로부터 다양한 변형 및 균등한 타 실시예가 가능하다는 점을 이해할 것이다. 예를 들어, 설명된 기술들이 설명된 방법과 다른 순서로 수행되거나, 및/또는 설명된 시스템, 구조, 장치, 회로 등의 구성 요소들이 설명된 방법과 다른 형태로 결합 또는 조합되거나, 다른 구성 요소 또는 균등물에 의하여 대치되거나 치환되더라도 적절한 결과가 달성될 수 있다. 따라서, 본 발명의 진정한 기술적 보호 범위는 첨부된 등록청구범위의 기술적 사상에 의해 정해져야 할 것이다.

도면

도면1

101. Customer reviews indicate that many modern mobile devices are often unnecessarily ----- .
 (A) complication
 (B) complicates
 (C) complicate
 (D) complicated
102. Jamal Nawzad has received top performance reviews ----- he joined the sales department two years ago.
 (A) despite
 (B) except
 (C) since
 (D) during
103. Gyeon Corporation's continuing education policy states that ----- learning new skills enhances creativity and focus.
 (A) regular
 (B) regularity
 (C) regulate
 (D) regularly
104. Among ----- recognized at the company awards ceremony were senior business analyst Natalie Obi and sales associate Peter Comeau.
 (A) who
 (B) whose
 (C) they
 (D) those
105. All clothing sold in Develyn's Boutique is made from natural materials and contains no ----- dyes.
 (A) immediate
 (B) synthetic
 (C) reasonable
 (D) assumed

도면2

