난이도 조절 가능한 어구 단위 빈칸 추론 문항 생성 시스템

강석훈°[1], 이근배[1,2]
포항공과대학교 인공지능대학원[1], 컴퓨터공학과[2]
{sh.kang,gblee}@postech.ac.kr

Difficulty-adjustable Phrase-level Cloze Question Generation System

Seokhoon Kang°[1], Gary Geunbae Lee[1,2]
Pohang University of Science and Technology,
Graduate School of Artificial Intelligence[1], Computer Science and Engineering[2]

요 약

딥러닝을 이용한 언어 모델은 다양한 분야에서 사용되고 있는데, 그 중 교육 분야에선 꾸준히 시험 문항을 자동으로 생성하려는 요구가 존재해 왔다. 그러나 빈칸 추론 문항, 그 중에서도 어구 단위 빈칸 추론 문항은 학습 및 평가 목적으로 널리 쓰이고 있지만, 이를 자동 생성하려는 연구는 상대적으로 드물다. 이에 본 연구에선 masked language modeling (MLM)을 이용한 난이도 조절이 가능한 어구 단위 빈칸 추론 문항 생성 시스템을 제안한다. 본 시스템은 정답 생성 모델의 attention 정보에 따라 지문 내 중요한 어구를 삭제해 오답을 생성하고, 동시에 어구의 삭제 비율을 조절함으로써 더 쉽거나 더 어려운 오답을 만들어낼 수 있다. 평가 결과, 제안한 시스템은 기존 접근법보다 정답과의 유사도가 최고 28.3% 낮았고, 또한 난이도 설정에 따라 쉬운 오답이 어려운 오답에 비해 유사도가 15.1% 낮아, 더 정답과 먼 뜻의 오답을 생성해내었다.

주제어: 문항 자동 생성, 어구 단위 빈칸 추론 문항, 난이도 조절, Masked Language Modeling (MLM)

1. 서론

교육 분야에서는 학습자의 수학 능력을 시험 문제를 통해 평가하므로, 이러한 문제를 높은 품질로 제작하는 것은 중요하다. 하지만 이러한 시험 문항을 만드는 것은 시간과 노력이 많이 들어가는 작업이고, 인터넷의 보급에 따라 전자 학습이 널리 퍼지며 문항 자동 생성의 중요성은 더욱 증가했다. [1, 2] 이전에는 규칙에 기반한 [3], 혹은 미리 정의된 유의어 사전에 기반한 [4] 문항 자동 생성에 대해 연구되었으나, 이러한 자동 생성 시스템은 제작 가능한 문항 유형에, 그리고 생성되는 문항의 품질에 한계가 존재한다. 그러나 최근 딥러닝 기술의 발전에 따라 몇년간 자연어 처리 시스템은 크게 발전했고 [5, 6, 7], 이는 기존 시스템에 비해 더 폭넓은 문항 유형에 대해 더 높은 품질의 문제를 생성하는 시스템을 개발하는 돌파구가 됐다.

대부분의 언어 능력 평가 시험에서 사용되고 있는 시험 문항 유형인 빈칸 추론 문항 유형의 정답과 오답 생성도 그 중 하나이다. 빈칸 추론 문항의 경우, 크게 단어 단위 빈칸 문항과 어구 단위 빈칸 문항으로 나눌 수 있다. 표 1을 보면, 단어 단위 빈칸 문항의 경우 빈칸에 한 단어가 들어가는 반면, 어구 단위 빈칸 문항의 경우 빈칸에 한 어구가 들어가며, 단어가 아닌 어구에 대한 문맥적인 이해를 요구하므로 비교적 더 어려운 유형임을 알 수 있다. 그러나 아직 대부분의 연구는 단어 단위 빈칸 문항의 자동 생성에 대해서만 중점을 두고 있고, 어구 단위 빈칸 문항에 대해서는 사전 연구가 상대적으로 드물다. [8] 문제의 난이도 또한 문제의 품질과 직결되는 중요한 요소이지만, 이를 제어하며 문항을 자동으로 생성하는 것은 더더욱 어렵다.

표 1. 단어 단위와 어구 단위 빈칸 추론 문항 예시

단어 단위	이번에 _____ 에서 옷을 샀는데, 굉장히 마음에 듭니다. 1. 옷가게 2. 은행 3. 학교 4. 서점
어구 단위	이번에 출시된 스마트폰은 배터리 용량이 커, _____ 있습니다. 1. 오랫동안 사용할 수 2. 고장 없이 사용할 수 3. 작업을 빠르게 수행할 수 4. 적은 금액으로 구매할 수

이에 본 연구에선 이러한 문제점을 해결하고자, 난이도 조절이 가능한 어구 단위 빈칸 추론 문항 생성 시스템을 제안한다. 본 연구가 제안하는 시스템은

- 난이도 조절이 가능한 첫 어구 단위 빈칸 추론 문항 생성 시스템이며,
- 정보 삭제를 통해 오답 생성에 최적화된 구조를 가졌고,
- 하나의 모델로 정답과 오답 모두 생성 가능해 효율적이고,
- 자가지도 학습 목표를 가져, 학습에 특별한 데이터가 필요하지 않다.

2. 관련 연구

2.1 빈칸 추론 문항 자동 생성

단어 단위 빈칸 추론 문항에 대한 자동 생성에 대한 사전 연구는 크게 첫번째로 말뭉치에서 지문 선택, 두번째로 지문 내 빈칸의 위치 추정, 마지막으로 빈칸에 들어갈 정답과 오답 생성, 총 세가지로 나눌 수 있다 [9]. [10, 11]의 경우에는 모두 WordNet이나 내부 유의어 사전과 같은 외부 자원을 사용한 연구로, 이러한 연구는 외부 자원에 존재하지 않는 문서에 대해서는 잘 작동하지 않는다는 문제점이 있다. [12]는 이 세가지 모두를 문서의 구문 분석 결과, 단어 발생 횟수나 빈도수, 품사, 문맥적 유사도 등에서 직접 선별한 특징들을 종합해 따로따로 결정했다. 이후 데이터 기반 seq2seq 모델의 등장과 함께, 외부 자원을 활용한 규칙 기반 시스템보다는 딥러닝 기반 자연어 생성에 기반한 연구가 주류를 차지하게 됐다. RevUP [13] 또한 이 세가지 모두를 다룬 연구이지만 데이터 기반 방식을 이용한다는 차이가 존재한다. 해당 시스템은 주제적으로 중요한 문장을 선택해 기계 학습 분류기를 통해 빈칸의 위치를 고른다. 오답 생성에 대해서는 word2vec, 언어 모델 확률, dice coefficient를 이용한다. CLOZER [9]는 masked language model (MLM)의 확률 분포를 이용해 지문 내 빈칸의 위치를 추정하는 방식을 제안했다. [14]는 오답을 생성할 때 빈칸 단어 뿐만 아니라 문맥도 고려하기 위해 BERT를 이용했고, MLM 확률을 통해 생성되는 오답의 난이도 제어가 가능했다.

어구 단위 빈칸 추론 문항의 경우에는 상대적으로 연구의 양이 적다. SWAG [15]은 딥러닝을 이용해 어구 단위 빈칸 추론 문항의 오답을 만든 연구 중 하나이다. SWAG은 상황이 주어진 다음, 이후 무엇이 일어날 지를 4개의 선지중에서 하나의 정답을 고르는 평가용 데이터셋인데, 해당 논문의 저자는 먼저 오답 후보를 생성하기 위해 LSTM 기반 언어 모델을 greedy decoding 방식을 사용해 샘플링했고, 이후 문체 특징을 검출하는 여러 모델들로 adversarial filtering을 수행해 생성된 후보간 문체 특징 차이를 줄였다. 그러나 해당 방식의 경우에는 생성된 오답이 정답에 가까울 수 있는 문제점이 존재하기 때문에, 이를 사람이 직접 검사해 걸러내는 작업이 필요했고, 생성되는 오답의 난이도 또한 조절이 불가능했다.

본 논문에서는 어구 단위 빈칸 추론 문항의 빈칸에 들어갈 정답과 오답 생성에 초점을 두고, 신경망 기반 MLM을 이용해 정답과 오답을 생성하는 데이터 기반 자연어 생성 방식을 이용하되, 난이도 조절이 가능하도록 연구를 진행해보려 한다.

2.2 난이도 제어 문항 자동 생성

난이도 제어가 가능한 문항 자동 생성 시스템은 비교적 최근에 연구가 많이 이루어 지고 있다. [16]은 지문과 난이도가 주어졌을때 난이도에 맞는 질문을 생성하는, SQuAD [17]와 유사한 형식의 독해 평가를 위한 문항의 자동 생성에 대한 연구이다. 해당 논문의 저자가 제안하는 Difficulty-controllable Generation (DQG) framework는 인코더-디코더 구조를 가졌는데, 인코더에서는 지문 토큰과 추가로 정답과 지문간의 유사도를 인코딩했고, 디코더를 각 난이도에 따라 각각 존재하는 변수를 인코더의 최종 은닉 상태와 이은 벡터로 초기화 함으로써 난이도에 맞는 질문을 생성한다. [18]도 동일한 형식의 독해 문항의 생성에 관련된 연구인데, 지문과 목표 난이도만 입력받아 질문과 정답 쌍을 생성한다는것과, item response theory [19]를 이용해 생성된 질문 정답 쌍의 난이도를 예측한다는 것, 또한 문항 생성에 transformer 기반 모델인 BERT, GPT-2를 이용한다는 점에서 다르다. [20]은 질문에 답하는데 여러 단계의 추론이 필요한 multi hop 질문의 생성에 관한 연구인데, 난이도를 모델링하기 위해 질문에서 개체 연결 (entity linking)이 얼마나 확실한지, 그리고 얼마나 특정적인지를 이용했다.

이렇듯 난이도 제어 문항 자동 생성은 현재 활발하게 연구되고 있는 분야이다. 그러나 위와 같은 접근법은 별도의 질문이 존재하지 않는 빈칸 추론 문항의 생성에 적용하기 힘들기 때문에, 본 논문에서는 새로운 방식의 난이도 제어 문항 자동 생성 기법에 대해 제안하려 한다.

3. 제안하는 시스템

어구 단위 빈칸 추론 문항을 자동 생성하는 시스템을 만들기 위한 가장 쉬운 방법은 지문과 이에 대응하는 정답, 오답에 대해 학습해 각각을 생성하는 방법일 것이다. 그러나 불행히도 아직까지 어구 단위 빈칸 문항에 관한 공개된 데이터셋은 존재하지 않으므로, 이와 같은 데이터 없이도 오답을 생성해낼 수 있는 시스템이 필요하다. 이에 본 연구에 자가지도로 학습하는 MLM을 이용해 오답을 생성하고자 한다.

빈칸 추론 유형 문항에서, 오답이 정답이 될 수 없는 주된 이유는 주어진 지문의 일정 부분과 모순되기 때문이다. 이 점에 착안해 본 논문에서는 정답을 출력하도록 학습시킨 신경망 기반 MLM에 지문의 특정 부분을 삭제한 다음 전달함으로써 지문의 일부분과 어긋나는 오답을 생성하고, 또한 정보를 삭제하는 비율을 조절함으로써 생성되는 오답의 난이도 또한 조절하는 어구 단위 빈칸 추론 문항 자동 생성 시스템을 제안한다.

본 시스템은 그림 1과 같이 총 4가지 모듈로 구성된다. 첫번째로 MLM을 이용해 지문의 빈칸에 알맞은 정답을 생성하는 정답 생성 모듈, 다음으로 정답 생성중 계산된 attention 정보를 이용해 지문 내 어구의 중요도 서열을 매기는 어구 랭킹 모듈, 이후 서열 정보를 이용해 지문에서 중요한 정보를 삭제하고, 이를 정답 생성 모듈의 MLM에 입력해 오답 후보들을 생성하는 오답 생성 모듈, 마지막으로 적절하지 않은 오답을 필터링해 최종 오답을 출력하는 오답 필터링 모듈이다.

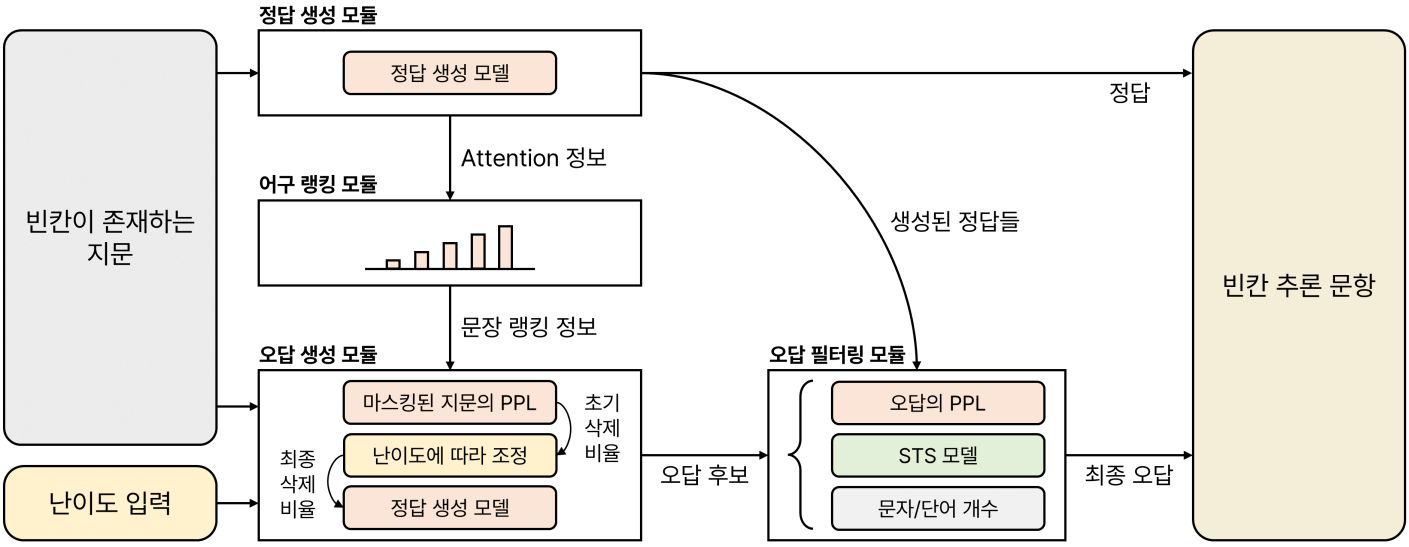

그림 1. 제안한 어구 단위 빈칸 추론 문항 생성 시스템 구조

3.1 정답 생성 모듈

문장에 존재하는 빈칸 안에 문맥적으로 올바른 단어들을 생성하는 태스크가 MLM 학습 목표와 유사한 것에 착안해, transformer 기반 MLM을 이용해 정답 생성 모듈을 구성했다. 다만 사전 학습에서의 마스킹과 빈칸 추론 문항 내에 존재하는 빈칸 사이에는 단어 길이, 구성 성분 등에 차이가 있으므로, 빈칸 추론 문항과 유사한 성격의 마스킹에 대해 추가 학습을 진행한 모델을 사용해 정답 생성을 진행한다. 또한 오답 필터링 모듈에서 정답과 유사한 오답을 걸러내기 위해 텍스트 의미 유사도를 사용하게 되는데, 생성된 하나의 정답 외에도 다른 정답이 존재할 수 있는 가능성에 대응하고자 여러개의 정답을 샘플링한다. 이후 정답을 생성하며 도출된 attention 정보를 **어구 랭킹 모듈**에 넘기고, 샘플링된 정답들은 **오답 필터링 모듈**에 넘긴다.

3.2 어구 랭킹 모듈

지문의 일부를 *임의로* 삭제하게 되면, 같은 비율로 지문의 정보를 삭제해도 중요 어구의 삭제 여부에 따라 생성되는 오답의 난이도가 크게 달라질 수 있다. 따라서 본 논문에서는 정답 생성 모듈의 attention 정보를 통해, 지문내에 존재하는 어구들을 정답을 추론하는데 있어서의 중요도로 서열을 매기는 방법에 대해 제안한다. transformer 기반 모델의 구성 요소 중 self-attention에서는 입력된 각 토큰을 처리하는데에 있어 다른 토큰의 중요도를 계산하게 되는데, 이때 빈칸에 대응하는 토큰을 처리하는데에 있어 중요도로 지문내의 어구들을 정렬하게 된다. 또한 어구들이 여러개의 토큰으로 이루어져 있을 경우엔, 토큰들 중 가장 높은 중요도를 기준으로 정렬한다. 이후 해당 정보를 **오답 생성 모듈**에 넘긴다.

3.3 오답 생성 모듈

일정한 난이도의 오답을 생성하기 위해서는 지문에 존재하는 어구들의 중요도 뿐만 아니라 얼마만큼의 어구를 삭제해 모델에 전달할지, 즉 정보 삭제 비율 또한 중요하다. 그러나 모든 어구에 대해 각각 마스킹을 진행한 후 오답 생성을 진행하는 방식으로 삭제 비율을 정하게 되면 너무 많은 시간과 자원이 소요된다. 이에 본 논문에선 난이도에 따라 적절한 정보 삭제 비율을 생성 과정 없이 추정하고자 한다.

빈칸 추론형 문항에서, 지문 내의 빈칸에 정답이 특별히 어울릴 수록 정답을 추론하기 위한 정보가 많고, 반대로 빈칸에 여러 어구가 어울릴수록 정보가 적을 것이라고 가정할 수 있다. 따라서 어구 랭킹 모듈의 정보를 이용해 지문에 존재하는 여러 어구들을 중요도에 기반해 정렬한 후, 중요도가 낮은 어구부터 차례로 지문에서 삭제하면서 정답에 대한 perplexity (PPL)를 계산한다. 이때 가장 PPL이 크게 증가한, 즉 정답이 생성될 확률이 가장 크게 감소한 삭제 비율부터 지문의 중요 정보를 삭제하고 있다고 볼 수 있으므로, 해당 삭제 비율을 초기 삭제 비율로 정했다. 이후 목표 난이도에 따라 초기 삭제 비율을 그대로 사용하거나 혹은 더 증가시켜 최종 삭제 비율을 정하고, 이를 적용한 지문을 정답 생성 모델에 입력함으로써 오답 후보들을 생성해 **오답 필터링 모듈**에 넘긴다.

3.4 오답 필터링 모듈

오답 생성 모듈은 정보를 삭제함으로써 정답보단 오답을 주로 생성할 수 있는 구조를 가졌다. 하지만 그럼에도 불구하고 정답에 가까운 오답 후보가 생성될 수 있고, 반대로 지문 내용과 크게 벗어나 너무 쉬운 오답 후보 또한 생성될 수 있다. 따라서 본 시스템은 오답 필터링 모듈을 통해, 이러한 적절하지 않은 오답 후보들을 걸러낸 다음 최종 오답을 선정한다.

표 2. 정량 평가 및 정성 평가 결과

모델	난이도 설정	정량 평가 ↓ (평균 유사도)	정성 평가		
			정답 ↓	오답 (어려움) ↑	오답 (쉬움) ↑
베이스라인	-	0.6344	72%	22%	6%
제안한 시스템	오답 (어려움)	0.5357	42%	46%	12%
	오답 (쉬움)	0.4549	36%	34%	30%

오답 필터링 모듈은 정답과 오답간의 의미적 유사도를 통해 정답에 가까운 오답후보와, 정답 생성 모듈의 PPL을 통해 너무 쉬운 오답 후보, 오답과 오답간의 의미적 유사도를 통해 뜻이 중복되는 오답 후보들을 걸러낸다. 추가적으로 생성된 정답과 오답 간의 단어 개수, 글자 개수 차이 또한 필터링 기준이 된다.

4. 실험 및 결과

4.1 실험 방법

제안한 시스템의 정답 생성 모델은 자가지도 학습 목표를 가지는 MLM이므로 학습에 특별한 입력 쌍이 필요하지 않아, 뉴스 데이터셋인 all-the-news-2.0 [1]을 사용했고, 추가적으로 전처리를 진행했다. 또한 본 시스템은 빈칸이 존재하는 지문만을 입력 받기 때문에 추가적으로 빈칸을 만드는 과정을 진행했다. 이때 실제 어구 단위 빈칸 추론 문항과 유사한 빈칸을 만들기 위해, 지문을 여러개의 어구로 나눈 후 2 ∼ 4개의 어구와 3 ∼ 10개의 단어로 구성된 빈칸 후보들을 생성한 후, 단어 길이에 기반한 푸아송 샘플링 ($\lambda = 5$)을 통해 하나의 빈칸을 뽑아 지문을 제작했고, 이 과정을 학습 목표 epoch 만큼 반복해 총 193030 × epoch 개의 예제로 구성된 데이터셋을 제작했다. 어구로 나누기 위해선 flair에서 제공하는 CoNLL-2000의 Chunking 태스크 [21]에 대해 학습된 chunk-english [22] 모델을 이용했다.

평가 데이터로는 중국의 영어 시험으로부터 수집된 데이터셋인 RACE [23]의 전체 validation 셋 지문을 이용했다. 해당 데이터셋의 지문 또한 빈칸이 존재하지 않기에, 마찬가지로 빈칸을 추가하는 작업을 진행했다. 이때 지문의 나머지 내용 만으로 유추할 수 있는 빈칸을 만들기 위해, Huggingface [24]에서 제공하는 bert-large-cased [6]를 이용해 빈칸이 존재할때 원문과 적은 의미 차이를 가지는 빈칸들을 뽑았고, 이후 푸아송 샘플링 ($\lambda = 5$)을 통해 하나의 빈칸을 뽑아 제작한 총 1316개의 지문 중 임의로 500개를 선택해 평가에 사용했다.

정답 생성 모델로는 MLM을 사용했는데, 이때 마스크 토큰 내에 임의 길이의 토큰이 올 수 있어야 하므로 Huggingface에서 제공하는 인코더-디코더 구조를 가진 BART-large [7]를 사용했다. 모델 학습시 하이퍼 패러미터는 batch size 4, learning rate 3e-5로 했고, NVIDIA RTX 4090 1기를 사용했다.

[1]https://components.one/datasets/all-the-news-2-news-articles-dataset

4.2 실험 결과

본 평가의 목적은 정보 삭제을 통한 오답 생성이라는 기법의 평가를 위한 것이므로, 제안된 시스템 중 오답 필터링 모듈은 사용하지 않았다. 베이스라인으로는 SWAG에서 사용된 greedy 디코딩을 사용해, 해당 방법과 비교해 제안된 시스템의 성능을 검증하고자 한다.

4.2.1 정량 평가

정량 평가 기준으로는 의미적 유사도를 이용했으며, 이를 위해 Sentence Transformer [25]의 all-mpnet-base-v2[2]를 이용해 생성된 오답후보들과 정답간의 평균 유사도를 구했다.

표 2로 부터 베이스라인과 제안한 시스템이 도출한 오답 후보들의 평균 유사도를 알 수 있는데, 베이스라인의 경우 정답과 0.6344의 가장 높은 유사도를 가져 오답을 생성하기 보단 정답을 더 많이 생성함을 알 수 있었다. 그에 반해 제안한 시스템의 경우 어려움 설정시 0.5357, 쉬움 설정시 그보다 낮은 0.4549로, 정답보단 오답에 가까운 후보들을 출력하는 것을 볼 수 있었으며, 쉬운 오답을 생성하도록 했을 경우 정답과 더더욱 의미적으로 떨어진 오답 후보들을 생성하는 것을 확인했다.

4.2.2 정성 평가

텍스트 유사도 만으로는 생성되는 오답의 품질을 완전히 평가할 수 없으므로, 평가 데이터 중 임의로 선택된 50개의 지문에 대해 정성적 평가를 진행했다. 사람이 느끼기에 생성된 오답 후보들에 대해 각각 정답에 가까운지를, 또한 오답이라면 어려운지, 쉬운지를 평가해, 각 범주마다 비율을 구했다.

표 2를 보면 베이스라인과 제안한 시스템의 정성 평가 결과를 볼 수 있는데, 베이스라인의 경우 72%의 상당수의 오답 후보들이 실제로는 지문과 문맥이 맞는 올바른 후보임을 보였다. 제안한 시스템의 경우 어려운 오답으로 설정했을때와 쉬운 오답으로 설정했을때 각각 58%와 64%의 오답 후보가 실제로 틀린것으로 확인되었는데, 이를 통해 정량 평가 결과와 마찬가지로 제안된 시스템이 지문과 어울리지 않는 오답 후보를 더 많이 만들어낸다는 것을 알 수 있다. 또한 시스템의 난이도 설정에 따라 실제로 체감되는 난이도에서도 차이가 유의미한 차이가

[2]https://huggingface.co/sentence-transformers/all-mpnet-base-v2

있음을 확인했는데, 어려운 오답을 생성하도록 설정했을때는 46%의 가장 많은 비율의 오답 후보들이 어려운 오답으로 평가되었으며, 쉬운 오답을 생성하도록 했을 경우 2.5배 오른 30%의 오답 후보들이 쉬운 오답으로 평가되었다.

5. 결론

본 논문에서는 난이도 조절이 가능한 첫 어구 단위 빈칸 추론 문항 생성 시스템에 대해 제안한다. 해당 시스템은 지문에서 정보를 삭제해 오답을 생성하는 기법을 사용해 학습에 특별한 데이터가 필요하지 않고, 하나의 모델로 정답과 오답을 모두 생성할 수 있어 효율적이고, 다수의 모델을 이용함으로써 생길 수 있는 문체 차이를 제거했다. 평가 결과, 제안된 시스템은 베이스라인과 비교해 생성되는 오답 후보의 정답과의 평균 유사도가 최고 28.3% 낮아 정답으로부터 더 떨어진 뜻의 오답을 생성했고, 정성 평가에서도 마찬가지로 베이스라인과 비교해 정답이라고 느껴지는 오답 후보 비율이 최대 절반으로 떨어졌다. 또한 본 시스템의 난이도 설정이 쉬움일때, 어려움으로 설정했을 때보다 정답과의 유사도가 15.1% 낮았고, 정성 평가 결과 베이스라인과 대비해 어려운 오답의 생성 빈도는 최대 2배 이상, 쉬운 오답의 생성 빈도는 최대 5배 증가했다.

그러나 본 시스템에 적용된 의미적 유사도를 통한 오답 필터링은 적절치 않은 오답을 걸러내는데에 완전하지 않기 때문에, 향후에는 이를 걸러낼 수 있는 추가적인 방법에 대해 연구할 계획이다. 또한, 제안한 시스템을 다국어로 확장해 저자원 언어에서도 사용할 수 있는 시스템에 대해서도 연구해보려 한다.

감사의 글

이 논문은 2023년 (주)엔에스데블의 UBT Technology AI-R&D 예산 재원으로 수행된 연구임

이 논문은 2023년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2022-0-00223, (세부2) 자폐증 환자의 의사소통 능력 향상을 위한 디지털치료제 개발)

참고문헌

[1] A. Qayyum and O. Zawacki-Richter, *Distance Education in Australia, Europe and the Americas*. Singapore: Springer Singapore, 2018, pp. 121–131. [Online]. Available: https://doi.org/10.1007/978-981-13-0298-5_14

[2] I. R. Goldbach and F. G. Hamza-Lup, "Survey on e-learning implementation in eastern-europe - spotlight on romania," *Proceedings of the International Conference on Mobile, Hybrid, and Online Learning*, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:53524319

[3] T. Alsubait, B. Parsia, and U. Sattler, "Ontology-based multiple choice question generation," *KI - Künstliche Intelligenz*, Vol. 30, No. 2, pp. 183–188, Jun 2016. [Online]. Available: https://doi.org/10.1007/s13218-015-0405-9

[4] G. A. Miller, "WordNet: A lexical database for English," *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*, 1994. [Online]. Available: https://aclanthology.org/H94-1111

[5] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, "A neural probabilistic language model," *Journal of Machine Learning Research*, Vol. 3, pp. 1137–1155, 2003.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Jun. 2019. [Online]. Available: https://aclanthology.org/N19-1423

[7] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7871–7880, Jul. 2020. [Online]. Available: https://aclanthology.org/2020.acl-main.703

[8] G. Kurdi, J. Leo, B. Parsia, U. Sattler, and S. Al-Emari, "A systematic review of automatic question generation for educational purposes," *International Journal of Artificial Intelligence in Education*, Vol. 30, No. 1, pp. 121–204, Mar 2020. [Online]. Available: https://doi.org/10.1007/s40593-019-00186-y

[9] S. Matsumori, K. Okuoka, R. Shibata, M. Inoue, Y. Fukuchi, and M. Imai, "Mask and cloze: Automatic open cloze question generation using a masked language model," *IEEE Access*, Vol. 11, pp. 9835–9850, 2023.

[10] J. Brown, G. Frishkoff, and M. Eskenazi, "Automatic question generation for vocabulary assessment," *Proceedings of Human Language Technology Conference*

and Conference on Empirical Methods in Natural Language Processing, pp. 819–826, Oct. 2005. [Online]. Available: https://aclanthology.org/H05-1103

[11] H. Kunichika, M. Urushima, T. Hirashima, and A. Takeuchi, "A computational method of complexity of questions on contents of english sentences and its evaluation," *International Conference on Computers in Education, 2002. Proceedings.*, pp. 97–101 vol.1, 2002.

[12] M. Agarwal and P. Mannem, "Automatic gap-fill question generation from text books," *Proceedings of the Sixth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 56–64, Jun. 2011. [Online]. Available: https://aclanthology.org/W11-1407

[13] G. Kumar, R. E. Banchs, and L. F. D'Haro, "Revup: Automatic gap-fill question generation from educational texts," *NAACL*, J. R. Tetreault, J. Burstein, and C. Leacock, Eds., pp. 154–161, 2015. [Online]. Available: https://doi.org/10.3115/v1/w15-0618

[14] C. Y. Yeung, J. S. Y. Lee, and B. K.-Y. T'sou, "Difficulty-aware distractor generation for gap-fill items," *Australasian Language Technology Association Workshop*, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:209549356

[15] R. Zellers, Y. Bisk, R. Schwartz, and Y. Choi, "Swag: A large-scale adversarial dataset for grounded commonsense inference," *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

[16] Y. Gao, L. Bing, W. Chen, M. R. Lyu, and I. King, "Difficulty controllable generation of reading comprehension questions," *International Joint Conference on Artificial Intelligence*, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:71147311

[17] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ questions for machine comprehension of text," *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, Nov. 2016. [Online]. Available: https://aclanthology.org/D16-1264

[18] M. Uto, Y. Tomikawa, and A. Suzuki, "Difficulty-controllable neural question generation for reading comprehension using item response theory," *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pp. 119–129, Jul. 2023. [Online]. Available: https://aclanthology.org/2023.bea-1.10

[19] F. M. Lord, *Applications of Item Response Theory To Practical Testing Problems*. Routledge, 1980. [Online]. Available: https://doi.org/10.4324/9780203056615

[20] V. Kumar, Y. Hua, G. Ramakrishnan, G. Qi, L. Gao, and Y.-F. Li, "Difficulty-controllable multi-hop question generation from knowledge graphs," *The Semantic Web–ISWC 2019: 18th International Semantic Web Conference, Auckland, New Zealand, October 26–30, 2019, Proceedings, Part I 18*, pp. 382–398, 2019.

[21] E. F. Tjong Kim Sang and S. Buchholz, "Introduction to the CoNLL-2000 shared task chunking," *Fourth Conference on Computational Natural Language Learning and the Second Learning Language in Logic Workshop*, 2000. [Online]. Available: https://aclanthology.org/W00-0726

[22] A. Akbik, D. Blythe, and R. Vollgraf, "Contextual string embeddings for sequence labeling," *COLING 2018, 27th International Conference on Computational Linguistics*, pp. 1638–1649, 2018.

[23] G. Lai, Q. Xie, H. Liu, Y. Yang, and E. Hovy, "RACE: Large-scale ReAding comprehension dataset from examinations," *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 785–794, Sep. 2017. [Online]. Available: https://aclanthology.org/D17-1082

[24] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush, "Transformers: State-of-the-art natural language processing," *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Oct. 2020. [Online]. Available: https://aclanthology.org/2020.emnlp-demos.6

[25] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 11 2019. [Online]. Available: https://arxiv.org/abs/1908.10084