기계독해 데이터셋의 교차 평가 및 블라인드 평가를 통한 한국어 기계독해의 일반화 성능 평가

임준호○, 김현기
한국전자통신연구원
{joonho.lim, hkk}@etri.re.kr

Evaluating Korean Machine Reading Comprehension Generalization Performance using Cross and Blind Dataset Assessment

Joon-Ho Lim○, Hyunki Kim
ETRI

요 약

기계독해는 자연어로 표현된 질문과 단락이 주어졌을 때, 해당 단락 내에 표현된 정답을 찾는 태스크이다. 최근 기계독해 태스크도 다른 자연어처리 태스크와 유사하게 BERT, XLNet, RoBERTa와 같이 사전에 학습한 언어모델을 이용하고 질문과 단락이 입력되었을 경우 정답의 경계를 추가 학습(fine-tuning)하는 방법이 우수한 성능을 보이고 있으며, 특히 KorQuAD v1.0 데이터셋에서 학습 및 평가하였을 경우 94% F1 이상의 높은 성능을 보이고 있다. 본 논문에서는 현재 최고 수준의 기계독해 기술이 학습셋과 유사한 평가셋이 아닌 일반적인 질문과 단락 쌍에 대해서 가지는 일반화 능력을 평가하고자 한다. 이를 위하여 첫 번째로 한국어에 대해서 공개된 KorQuAD v1.0 데이터셋과 NIA v2017 데이터셋, 그리고 엑소브레인 과제에서 구축한 엑소브레인 v2018 데이터셋을 이용하여 데이터셋 간의 교차 평가를 수행하였다. 교차 평가결과, 각 데이터셋의 정답의 길이, 질문과 단락 사이의 오버랩 비율과 같은 데이터셋 통계와 일반화 성능이 서로 관련이 있음을 확인하였다. 다음으로 KorBERT 사전 학습 언어모델과 학습 가능한 기계독해 데이터셋 21만 건 전체를 이용하여 학습한 기계독해 모델에 대해 블라인드 평가셋 평가를 수행하였다. 블라인드 평가로 일반분야에서 학습한 기계독해 모델의 법률분야 평가셋에서의 일반화 성능을 평가하고, 정답 단락을 읽고 질문을 생성하지 않고 질문을 먼저 생성한 후 정답 단락을 검색한 평가셋에서의 기계독해 성능을 평가하였다. 블라인드 평가 결과, 사전 학습 언어 모델을 사용하지 않은 기계독해 모델 대비 사전 학습 언어 모델을 사용하는 모델이 큰 폭의 일반화 성능을 보였으나, 정답의 길이가 길고 질문과 단락 사이 어휘 오버랩 비율이 낮은 평가셋에서는 아직 80%이하의 성능을 보임을 확인하였다. 본 논문의 실험 결과 기계독해 태스크는 특성 상 질문과 정답 사이의 어휘 오버랩 및 정답의 길이에 따라 난이도 및 일반화 성능 차이가 발생함을 확인하였고, 일반적인 질문과 단락을 대상으로 하는 기계독해 모델 개발을 위해서는 다양한 유형의 평가셋에서 일반화 평가가 필요함을 확인하였다.

주제어: 한국어 기계독해, 일반화 평가, KorBERT

1. 서론

기계독해는 기계의 자연어 독해 수준을 평가하기 위한 태스크 중의 하나로, 자연어로 표현된 질문과 단락이 주어졌을 때, 해당 단락 내에서 표현된 질문의 정답을 찾는 태스크이다 [1-2]. 영어, 한국어, 중국어 등의 언어에 대해 딥러닝을 이용한 많은 연구가 수행되고 있는 분야이다 [3-5].

기존 기계독해 접근방법은 각 단어의 임베딩 벡터를 기반으로, RNN과 같은 문맥 반영 레이어를 거쳐서, 질문과 단락 사이의 상호 집중(Attention) 매커니즘을 이용하여 정답의 시작과 끝을 인식하는 방법을 많이 사용하였다 [6-7]. 하지만, 최근 BERT 연구 이후에는 대용량 원시 말뭉치로부터 학습한 사전 학습(pre-training) 언어모델을 이용하여, 질문과 단락을 하나의 단위로 입력 후, 해당 단위 내에서 멀티-헤드 자가 집중(Multi-head Self-Attention) 매커니즘을 이용하여 정답의 시작과 끝을 인식하는 방법을 많이 사용하고 있다 [8]. 이와 같이 대용량 말뭉치로부터 학습한 사전 언어모델을 이용할 경우 많은 데이터셋에서 가장 우수한 성능을 보임이 증명되었다 [9-10].

본 논문에서는 현재 최고 성능을 보이는 사전 언어모델 기반 기계독해 기술의 일반화 성능 평가를 목표로 한다.

이를 위하여 첫 번째로, 학습셋과 동일하게 구축된 평가셋에서의 성능 평가와 더불어 다른 평가셋에서의 일반화 성능을 평가한다. 기계독해 일반화 평가를 위하여 한국어에 대해 공개된 KorQuAD v1.0 및 NIA v2017 기계독해 데이터셋과 엑소브레인 과제에서 구축한 엑소브레인 v2018 평가셋을 이용하여 평가셋 간의 교차 평가를 수행하고자 한다 [11].

다음으로, KorBERT 사전 학습 언어모델과 학습 가능한 기계독해 데이터셋 21만 건 전체를 이용하여 학습한 기계독해 모델에 대해 블라인드 평가셋 평가를 수행한다. 블라인드 평가를 위하여 일반분야에서 학습한 기계독해 모델이 법률분야에서의 성능을 평가한다. 그리고 일반적

	학습집합 수량	평가집합 수량	정답 길이 평균 (표준편차)	정답 길이 10 형태소 이하 질문 비율	질문-단락 사이 어휘 오버랩 평균	정답 내 동사구 포함 질문 비율
KorQuAD v1.0	60,406	5,773	5.53 (3.35)	91.88%	75.59%	3.18%
NIA v2017	85,199	15,039	11.59 (13.61)	66.50%	67.77%	8.52%
엑소브레인 v2018	67,470	2,044	12.40 (14.91)	69.67%	56.39%	14.97%

표 1 KorQuAD v1.0, NIA v2017, 엑소브레인 v2018 데이터셋 통계 분석

인 기계독해 데이터셋이 특정 단락을 보고 질문을 생성하는 것과 반대로 질문을 먼저 생성 후, 단락을 검색한 블라인드 평가셋에서의 성능 평가를 통하여 일반화 성능을 평가한다.

본 논문의 구성은 다음과 같다. 2장에서는 기계독해 일반화 평가와 관련된 관련 연구를 소개하고, 3장에서는 본 논문에서 실험한 기계독해 모델에 대해 소개하고, 평가셋 간 교차 평가 결과 및 평가셋 통계분서 결과, 블라인드 평가 결과에 대해 소개한다. 마지막으로 4장에서 결론에 대해 논의한다.

2. 관련 연구

본 연구에 직접적인 동기를 제공한 연구로는 [12] 연구가 있고, 최근 딥러닝 모델 일반화를 위하여 데이터셋 오버피팅(over-fitting)에 주의해야 함을 주장한 연구로 [13]연구가 있다.

(Yogatama et al, 2019)는 딥마인드에서 2019년 발표한 연구로, 일반 언어지능(general linguistic intelligence)을 한 태스크에서 학습한 언어의 어휘, 문법, 문맥, 화용 지식을 다른 새로운 태스크에 빠르게 적용하는 능력으로 정의하였다. 그리고 현재 최고 수준으로 알려진 BERT 및 ELMO 기반 접근 방법의 일반 언어지능 능력을 평가하고자 하였다. 이를 위하여, SQuAD 데이터셋에서 학습한 BERT 기반 기계독해 모델을 이용하여 TriviaQA 및 QuAC 등 다른 기계독해 평가셋에서의 일반화 성능을 평가하였다. 그리고 기계독해를 학습한 모델을 MNLI 태스크에 추가 학습 시, 기존 기계독해 데이터셋에서의 성능 하락(catastrophic forgetting)을 평가하였다. 실험 결과 [12] 연구는 현재 딥러닝 기술의 성능이 특정 평가셋에서 과평가(over-estimated)되어 있으며, 일반 언어지능을 위하여 더욱 정교한 지속 학습(continual learning) 및 메모리 모듈(memory module)의 필요성을 주장하였다.

다른 관련 연구로 [13] 연구는 현재 딥러닝 모델의 성능이 태스크 자체에 대한 학습 외에 데이터셋을 구축한 작업자의 특징까지 학습하여 평가 결과가 과적합(over-fitting)되어 있음을 주장한 논문이다. [13] 연구는 OpenBookQA, CommonsenseQA, MNLI 데이터셋에 대해서 작업자 ID를 자질로 입력하였을 경우 작업자 편향성으로 인하여 각 태스크의 성능이 2~4%이상 개선됨을 확인하였고, 데이터 구축 수량이 많은 작업자는 데이터로부터 자동으로 작업자ID도 인식이 가능함을 확인하였다. 또한, 각 데이터셋의 학습 및 평가 집합을 작업자 별로 구분하여 평가하였을 경우 일반화가 잘 되지 않음을 실험적으로 확인하였다. 실험 결과를 바탕으로 [13] 연구는 데이터셋 구축에 있어서 작업자 편향성을 관리해야 하고, 평가셋 구축 작업자는 학습셋 구축 작업자와 분리되어야 과적합 평가를 방지할 수 있음을 주장하였다.

3. 한국어 기계독해 일반화 성능 평가

본 절에서는 KorBERT 언어모델 기반 기계독해 모델을 이용하여 다중 데이터셋 간의 교차 평가 및 블라인드 평가셋에서의 기계독해 일반화 성능에 대해 소개한다. 세부적으로 3.1절에서 평가에 사용한 KorBERT 기반 기계독해 모델에 대해 설명한다. 3.2절에서 KorQuAD v1.0 데이터셋, NIA 2017 기계독해 데이터셋, 엑소브레인2018 평가셋에 대한 통계 분석 결과를 소개하고, 3.3절에서는 교차 평가 결과를 소개한다. 3.4절에서는 블라인드 평가셋에서의 기계독해 일반화 성능을 소개하고, 3.5절에서는 사전 학습 언어모델의 적용 여부에 따라 블라인드 평가셋에서의 일반화 성능을 비교한다. 마지막으로, 3.6절에서 기계독해 모델에 다른 데이터셋을 추가 학습하는 학습 과정(curriculum)과 사후 망각(catastrophic forgetting)에 대하여 실험 및 논의한다.

3.1 기계독해 모델

기계독해 일반화 평가 실험에 사용한 기계독해 모델은 현재 가장 우수한 성능을 보이고 있는 사전 학습 언어모델 기반 기계독해 방법을 사용한다. 사전 학습 모델로는 형태소 기반 KorBERT 언어모델에 추가 사전 학습을 적용한 모델을 사용하고, 형태소분석기는 엑소브레인 OpenAPI 형태소분석기를 사용하였다 [14-15].

기계독해 태스크 학습 시 파라미터는 sequence length 512, batch 32, document stride 128, learning rate 3e-5, warmup ratio 10%, max_answer_length 30을 모두 동일하게 적용하였다.

3.2 평가셋 별 통계 분석

평가셋 / 학습셋	KorQuAD v1.0	NIA v2017	엑소브레인 v2018
KorQuAD v1.0	**87.11% / 94.67%**	66.32% / 81.83%	59.68% / 76.15%
NIA v2017	75.54% / 88.82%	**77.11% / 90.35%**	**73.23% / 88.40%**
엑소브레인 v2018	80.13% / 90.75%	73.02% / 87.93%	67.61% / 84.78%

표 2 학습 셋 별 평가셋 일반화 성능 (EM / F1)

기계독해 교차 평가에 사용한 데이터셋은 KorQuAD v1.0 데이터셋, NIA 2017 기계독해 데이터셋, 엑소브레인 2018 데이터셋을 사용하였다. 각 데이터셋 별로 학습/평가집합 구성 및 통계 분석 결과는 표1과 같다. 학습/평가집합 구분은 KorQuAD v1.0은 표준 구분을 따르고, NIA v2017 및 엑소브레인 v2018은 랜덤으로 구분하였다. 정답 길이는 정답 단락을 형태소 분석 후, 시작과 끝 정답 경계의 형태소 수로 계산하였고, 질문-단락 사이 어휘 오버랩은 질문 내 모든 형태소를 대상으로 계산하였다. 정답 내 동사구 포함 질문은 정답의 경계 중에 동사, 형용사, 동사 파생 접미사 형태소를 포함한 질문 비율로 계산하였다.

표1을 살펴보면, KorQuAD v1.0 데이터는 정답의 평균 길이가 짧고, 질문-단락 사이 어휘 일치율이 높으며, 정답 내 동사구 포함 비율이 낮음을 알 수 있다. NIA v2017과 엑소브레인 v2018 평가셋은 정답의 길이가 평균 11~12 형태소로 긴 정답을 다수 포함하고 있으며, 엑소브레인 v2018평가셋이 질문-단락 오버랩 비율이 가장 낮으며, 정답 내 동사구 포함 질문 비율이 높은 것을 알 수 있다.

3.2 데이터셋 간 교차 평가 결과

KorQuAD v1.0, NIA v2017, 엑소브레인 v2018 데이터셋에서 학습 후, 각 평가셋에서의 성능 평가는 표2와 같으며, 각각 EM과 F1 성능을 나타낸다.

실험 결과를 살펴보면, KorQuAD v1.0 학습셋에서 학습 시, KorQuAD v1.0 평가셋에서 94.67%로 가장 우수한 성능을 보였으나, 정답의 길이가 길고 질문-단락 사이 어휘 일치도가 낮은 평가셋에서는 각각 81.83% 및 76.15%로 성능이 낮게 측정됨을 알 수 있다.

NIA v2017 학습셋에서 학습 시, NIA v2017 평가셋에서 90.35%로 가장 우수한 성능을 보였으며, KorQuAD v1.0 및 엑소브레인 v2018 평가셋에서도 88.82%와 88.40%로 상대적으로 성능 하락 폭이 작아서, 가장 균형있는 기계독해 성능을 보임을 확인할 수 있다.

마지막으로, 엑소브레인 v2018 학습셋에서 학습한 경우, KorQuAD v1.0에서 90.75%로 가장 높은 성능을 보이고, 엑소브레인 v2018 평가셋에서는 NIA v2017 평가셋보다 낮은 84.78%를 보였다. 이는 엑소브레인 v2018셋이 NIA v2017 셋 대비 정답 길이가 짧은 문제를 다수 포함하여 정답의 길이가 짧은 질문에 대해서는 일반화를 수행하나, 정답의 길이가 긴 질문의 경우 NIA v2017보다 학습데이터 수량이 부족하여 성능 차이가 발생한 것으로 예상된다.

3.3 블라인드 평가셋 평가 결과

본 절에서는 KorQuAD v1.0, NIA v2017, 엑소브레인 v2018 학습셋 21만 데이터를 이용하여 우수한 성능의 기계독해 모델을 구축한 후, 블라인드 평가셋에서의 성능을 측정하고자 한다.

블라인드 평가셋으로는 구축자가 먼저 질문을 만들고 이후 위키백과에서 정답 단락을 검색하여 구축한 엑소브레인 NQ(Natural Questions) 평가셋과 국회도서관에서 구축한 법령분야 기계독해 평가셋을 이용하여 일반분야에서 학습한 기계독해 모델의 타 분야에서의 일반화 정도를 측정하고자 한다.

엑소브레인 NQ 평가셋과 법령 평가셋의 통계 정보 및 예시 질문/단락은 표3과 같으며, 블라인드 평가셋이 기존 KorQuAD v1.0, NIA v2017, 엑소브레인 v2018 평가셋 대비 정답의 길이가 길고, 질문-단락 사이 오버랩 비율이 낮으며, 정답 내 동사구를 포함한 질문의 비율이 높음을 알 수 있다. 각 평가셋의 예시 질문, 단락, 정답은 아래와 같다.

<엑소브레인NQ 평가셋 예>
- 질문: 애니메이션 코코 줄거리 알려줘

	평가집합 수량	정답 길이 평균 (표준편차)	정답 길이 10 형태소 이하 질문 비율	질문-단락 사이 어휘 오버랩 평균	정답 내 동사구 포함 질문 비율
엑소브레인 NQ 평가셋	1,385	24.88 (33.96)	47.59%	40.51%	35.09%
엑소브레인 법령 평가셋	441	20.56 (29.86)	56.69%	54.01%	31.97%

표 3 블라인드 평가셋(엑소브레인NQ, 국회도서관 평가셋) 통계 분석

평가셋 / 학습셋	KorQuAD v1.0	NIA v2017	엑소브레인 v2018	**엑소브레인 NQ 평가셋**	**엑소브레인 법령 평가셋**
KorQuAD v1.0 + NIA v2017 + 엑소브레인 v2018	87.70% / 94.99%[1]	79.71% / 91.78%	76.51% / 90.07%	**56.67% / 79.60%**	**52.38% / 79.77%**

표 4 KorBERT 기반 기계독해 모델의 블라인드 평가셋 평가 결과

평가셋 / 모델	엑소브레인 v2018	엑소브레인 NQ 평가셋	엑소브레인 법령 평가셋
독립 MRC 모델 v1 (사전학습 언어모델 미사용)	81.40% / 89.21%	42.52% / 67.74%	47.39% / 71.60%
KorBERT 기반 MRC (사전학습 언어모델 사용)	76.51% / 90.07%	56.67% / 79.60%	52.38% / 79.77%

표 5 BERT 적용 이전 및 이후의 일반화 성능 비교

- *단락: 《코코》(영어: Coco)는 2017년 제작된 미국의 애니메이션 뮤지컬 영화이다. 픽사 애니메이션 스튜디오가 제작하고 월트 디즈니 픽처스가 배급한다. 리 언크리치 감독의 아이디어를 기반으로 하였으며 애드리안 몰리나가 작가와 공동제작을 맡았다.[1] 영화의 주된 이야기는 12살 소년 미겔이 백 년 전의 미스터리와 관련된 일련의 사건에 휘말리고, 끝내 가족과 놀라운 재회를 하게 된다는 내용이다. ...*
- *정답: 12살 소년 미겔이 백 년 전의 미스터리와 관련된 일련의 사건에 휘말리고, 끝내 가족과 놀라운 재회를 하게 된다*

<엑소브레인 법령 평가셋 예>

- *질문: 국회 인사청문특별위원회가 인사청문과 관련된 자료의 제출을 요구받기 위한 정족수는 어떻게 되나요?*
- *단락: 위원회는 그 의결 또는 재적의원 3분의 1 이상의 요구로 공직후보자의 인사청문과 직접 관련된 자료의 제출을 국가기관·지방자치단체, 기타 기관에 대하여 요구할 수 있...*
- *정답: 의결 또는 재적의원 3분의 1 이상의 요구*

KorQuAD v1.0, NIA v2017, 엑소브레인 v2018 학습셋을 이용하여 KorBERT 기반 기계독해 모델의 성능 평가 결과 및 블라인드 평가셋 평가 결과는 표4와 같다.

KorBERT 기반 기계독해 모델 평가 결과, KorQuAD v1.0에서는 94.99%, NIA v2017에서는 91.78%, 엑소브레인 v2018에서는 90.07%로 학습셋에 포함된 평가셋에서는 우수한 성능을 보임을 확인하였고, 동일 모델을 일반분야가 아닌 법령 분야 평가셋 및 자연스러운 질문 평가셋(질문-단락 사이 어휘 오버랩 비율이 낮은 평가셋)에 적용한 결과 80% 이하로 낮은 일반화 성능을 보임을 확인하였다.

3.5 사전학습 언어모델 적용에 따른 일반화 성능 비교 평가

본 절에서는 사전학습 언어모델 적용 여부에 따른 블라인드 평가셋에서의 일반화 성능 차이를 평가하도록 한다.

BERT 이전의 기계독해 방법은 질문과 단락 사이의 집중(attention) 매커니즘을 이용하여 기계독해 태스크를 위한 별도의 모델을 설계하고, 해당 모델을 이용하여 단락 내 정답의 경계를 찾는 방법을 이용하였다 [6-7]. BERT 이전의 기계독해 방법과 KorBERT를 이용한 기계독해 방법 사이의 일반화에 대한 비교 평가는 표5와 같다.

독립 MRC 모델 v1은 학습셋을 NIA v2017과 엑소브레인 v2018 학습셋을 사용하였고, KorBERT 기반 MRC 모델은 KorQuAD v1.0, NIA v2017, 엑소브레인 v2018 학습셋을 사용하였다.

실험 결과, 독립 MRC 모델 v1은 학습셋과 동일한 평가셋에서는 BERT 기반 MRC 모델 대비 EM은 4.89% 우수하고, F1은 0.8% 정도 성능이 낮게 평가되었다. 하지만, 블라인드 평가셋에서는 약 8~12% 정도의 큰 성능 차이가 발생하였다. 언어모델을 사용하지 않은 기계독해 모델의 경우 학습셋에 대한 오버피팅(over-fitting)에 취약하고, 블라인드 평가셋에서 성능 하락이 큼을 알 수 있고, 사전 학습된 언어모델을 사용할 경우 대용량 원시 말뭉치에서 학습한 지식을 활용하여 블라인드 평가셋에서 성능 하락을 완화하는 정규화(regularization)효과가 있음을 알 수 있다.

3.6 데이터셋 학습 과정(curriculum)과 사후 망각(catastrophic forgetting) 평가

본 절에서는 기계독해 학습셋을 사용하여 1차 학습한 기계독해 모델에 2차 기계독해 학습셋을 사용하여 학습하였을 경우, 기존에 학습한 1차 모델의 성능이 2차 학습 시 발생하는 사후 망각을 평가하고자 한다. 2차 학습 시 학습 파라미터는 1차 학습과 동일하게 적용하고,

1) max_answer_length 파라미터를 25로 설정 시, KorQuAD 성능은 EM 87.80% / F1 95.08% 이다.

학습셋 \ 평가셋	KorQuAD v1.0	NIA v2017	엑소브레인 v2018
1차학습셋: KorQuAD v1.0 + NIA v2017 2차 학습셋: 엑소브레인 v2018	**87.47% / 94.76%** **82.50% / 92.26%** **(-4.97%/-2.50%)**	**77.24% / 90.47%** **75.68% / 89.57%** **(-1.56%/-0.90%)**	69.47% / 85.36% 70.54% / 86.71% (+1.07%/+1.35%)
1차학습셋: KorQuAD v1.0 + 엑소브레인 v2018 2차 학습셋: NIA v2017	**87.14% / 94.72%** **78.74% / 90.71%** **(-8.40%/-4.01%)**	74.11% / 88.31% 77.49% / 90.58% (+3.38%/+2.27%)	69.17% / 85.47% 74.41% / 89.19% (+5.24%/+3.72%)
1차학습셋: NIA v2017 + 엑소브레인 v2018 2차 학습셋: KorQuAD v1.0	80.07% / 91.19% 87.80% / 94.99% (+7.73%/+3.80%)	**77.69% / 90.66%** **70.54% / 85.19%** **(-7.15%/-5.47%)**	**73.67% / 88.69%** **64.82% / 81.34%** **(-8.85%/-7.35%)**

표 6 기계독해 추가 학습 시 사후 망각 평가 (첫 줄은 1차 학습셋으로 학습한 평가 결과, 두 번째 줄은 2차 학습셋으로 추가 학습한 평가결과이며, 마지막 줄은 1차와 2차 사이 성능 차이를 나타냄)

warmup ratio 파라미터는 적용하지 않았으며, max_answer_length는 30으로 평가하였으며, 평가 결과는 표6과 같다.

KorQuAD v1.0과 엑소브레인 v2018로 1차 학습 후, NIA v2017로 2차 학습하였을 경우, NIA v2017과 엑소브레인 v2018에서 각각 2.27%, 3.72% 성능이 개선되나, 정답의 양상이 다른 KorQuAD v1.0에서 -4.01% 성능 하락이 발생하였다.

NIA v2017과 엑소브레인 v2018에서 1차 학습 후, KorQuAD v1.0에서 2차 학습하였을 경우, KorQuAD 평가셋에서는 94.99%로 최고 성능을 보였으나, NIA v2017 및 엑소브레인 v2018 평가셋에서는 기존 1차 학습 모델 대비 각각 -5.47%와 -7.35%로 성능 하락이 큼을 알 수 있다.

실험 결과, 추가 학습데이터의 평균 정답 길이 및 질문-단락 사이 어휘 오버랩 비율과 같은 특징이 평가셋과 유사한 경우 성능이 개선되나, 다른 유형의 평가셋에서는 성능이 하락하고, 길이가 긴 정답을 학습한 모델에 길이가 짧은 정답을 추가 학습한 경우 큰 하락 폭을 보였다.

4. 결론

본 논문에서는 한국어 기계독해 모델의 일반화 성능을 평가하기 위하여, 평가셋 간의 교차 평가와 블라인드 평가를 수행하였다. 평가셋 간의 교차 평가 결과, 각 데이터셋의 정답의 길이, 질문-단락 사이 어휘 오버랩 비율의 통계와 기계독해 일반화 성능 사이에 관련이 있음을 확인하였다. 더불어 블라인드셋 평가 결과, 사전 학습 언어 모델을 사용하지 않은 기계독해 모델 대비 사전 학습 언어 모델을 사용하는 모델이 큰 폭의 일반화 성능을 보였으나, 질문과 단락 사이 어휘 오버랩 비율이 낮은 평가셋에서는 아직 80%이하의 성능을 보임을 확인하였다.

실험 결과 기계독해 태스크는 특성 상 질문과 정답 사이의 어휘 오버랩 및 정답의 길이에 따라 난이도 및 일반화 성능 차이가 발생함을 확인하였고, 일반적인 질문과 단락을 대상으로 하는 기계독해 모델 개발을 위해서는 다양한 유형의 평가셋에서의 일반화 평가가 필요함을 확인하였다.

감사의 글

이 논문은 2019년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2013-2-00131, 휴먼 지식증강 서비스를 위한 지능진화형 Wise QA 플랫폼 기술 개발)

참고문헌

[1] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang, “Squad: 100,000+ questions for machine comprehension of text”. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 2383-2392, 2016..

[2] Pranav Rajpurkar, Robin Jia, and Percy Liang, “Know what you don’t know: Unanswerable questions for squad”, arXiv preprint arXiv:1806.03822, 2018.

[3] https://rajpurkar.github.io/SQuAD-explorer/

[4] https://korquad.github.io/category/1.0_KOR.html

[5] Chih Chieh Shao, Trois Liu, Yuting Lai, Yiying Tseng, and Sam Tsai. “Drcd: a chinese machine reading comprehension dataset”. arXiv preprint arXiv:1806.00920, 2018.

[6] Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. “Bidirectional attention flow for machine comprehension”. In Proceedings of the International Conference on Learning Representations, 2017.

[7] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes, “Reading wikipedia to answer open-domain questions”. arXiv preprint arXiv:1704.00051, 2017.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and

Kristina Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding”, In North American Association for Computational Linguistics (NAACL), 2019.
[9] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. “Xlnet: Generalized autoregressive pretraining for language understanding”, arXiv preprint arXiv:1906.08237, 2019.
[10] Yinhan Liu, et al., “RoBERTa: A Robustly Optimized BERT Pretraining Approach”, arXiv preprint arXiv:1907.11692, 2019.
[11] AIHub, AI데이터 중 일반상식 데이터, http://www.aihub.or.kr/content/142
[12] Dani Yogatama, et al., “Learning and evaluating general linguistic intelligence”, arXiv preprint arXiv:1901.11373, 2019.
[13] Mor Geva, Yoav Goldberg, Jonathan Berant, “Are We Modeling the Task or the Annotator? An Investigation of Annotator Bias in Natural Language Understanding Datasets”, arXiv preprint arXiv:1908.07898, 2019.
[14] http://aiopen.etri.re.kr/service_dataset.php
[15] 이충희, 임준호, 임수종, 김현기, “기분석사전과 기계학습 방법을 결합한 음절 단위 한국어 품사 태깅”, 정보과학회논문지, 43(3), pp. 362-369, 2016.