Artficial Intelligence Laboratory

교차성능평가와 적대적 학습을 통한 기계독해 데이터셋의 편향성 검증

의료대용량데이터마이닝 최종발표

정보융합공학과 AI전공
정주경

부산대학교
PUSAN NATIONAL UNIVERSITY

Table of Contents

1. MRC(Machine Reading Comprehension)
2. Subject - Bias analysis of machine reading comprehension dataset
3. Background
    1. Cross-evaluation
    2. Adversarial training
4. Dataset
5. Model
    1. Background(BERT)
    2. RoBERTa
6. Research method
    1. Cross-evaluation
    2. Adversarial training
7. Results
8. Conclusion

Machine Reading Comprehension (MRC)

MRC

- machine(기계)+ reading comprehension(읽고 이해하다)
- 기계독해는 문서의 내용을 기계가 이해하고 이를 기반으로 여러 태스크에서 활용할 수 있는 기술
- 질의응답(QA: Question and Answering)에서 올바른 답변 위해 기계의 강건한 이해 능력이 뒷받침되어야 함

한계

- 기계독해 데이터셋은 많이 구축되어 있지만, 모델이 학습에 사용되지 않은 도메인 데이터가 입력될 경우, 성능이 하락할 수 있음
- 한 가지 도메인의 데이터셋만 학습한 모델은 다른 도메인에서의 성능이 급감하며 낮은 일반화 성능을 갖춘 모델이 될 확률이 높음
- 한 분야에 편향되지 않게 데이터셋을 생성해야 하며 많은 기업에서 크라우드 소싱(Crowd-Sourcing)을 통해 여러 도메인 기계독해 데이터셋을 생성하고 있음
- 하지만 이 역시 편향될 가능성이 있음

Subject - Bias analysis of machine reading comprehension dataset

교차 성능평가

- 여러 도메인 데이터셋을 이용해 각각 학습된 모델을 생성, 학습하지 않은 도메인외 데이터셋도 모델 평가에 활용해 교차 평가 진행
- 이를 통해 학습되지 않은 도메인에 대응하는 모델의 강건성을 평가, 다양한 도메인의 기계독해 데이터셋의 필요성을 확인

적대적 학습

- 데이터셋에 노이즈(Noise)를 주입해 모델의 학습에 사용하는 적대적 학습과 모델의 평가에 사용하는 적대적 평가를 적용
- 모델이 문맥을 온전히 이해하지 않고 반복되는 특정 단서(Superficial cues)를 학습하여 태스크를 수행한다는 연구가 있음
- 문맥에 정답 문장과 유사한 문장을 추가하여 모델을 학습하는 방식으로 적대적 학습을 진행
- 그 후 적대적 평가를 통해 일반 모델과 적대적 학습을 진행한 모델 간 결과를 비교하며 추후 데이터셋 구축에 대한 방향성을 제시

Background - Cross-Evaluation

Cross-Evaluation

표 2 학습말뭉치 별 평가말뭉치 일반화 성능(EM / F1)

Table 2 Test set generalization performance for each training set (EM/F1)

Test Training	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%
ExoBrain v2018	80.13% / 90.75%	73.02% / 87.93%	67.61% / 84.78%

- **ETRI에서 기계독해 데이터셋 교차 평가를 통해 수행한 한국어 기계독해 일반화 성능 평가 연구[1]에서 KorQuad v1.0, NIA v2017, 엑소브레인 v2018 세 가지 기계독해 데이터셋 이용**
- **각 데이터셋마다 모델 학습 후 모든 도메인의 데이터셋으로 평가하여 한국어 기계독해 일반화 성능 평가 진행**

[1] 기계독해 말뭉치의 교차 평가, 블라인드 평가 및 오픈도메인 질의응답 환경 평가를 통한 한국어 기계독해의 일반화 성능평가

Background - Adversarial Training

Superficial cues in MRC

- 이전 연구[2]에 따르면 기계독해 모델이 언어를 이해하지 않고 단순히 Superficial cues(질문과 유사한 어휘와 문장 구조)만을 이용하더라도 높은 성능을 보일 수 있음을 보임
- 실험에서 질문과 유사한 어휘와 구조를 가지도록 하는 적대적 예제를 단락에 추가하였을 때, 큰 성능 하락을 보임

Add adversarial examples in context

Article: Super Bowl 50
Paragraph: "*Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.*"
Question: "*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*"
Original Prediction: John Elway
Prediction under adversary: Jeff Dean

[2] Adversarial Examples for Evaluating Reading Comprehension Systems

Dataset

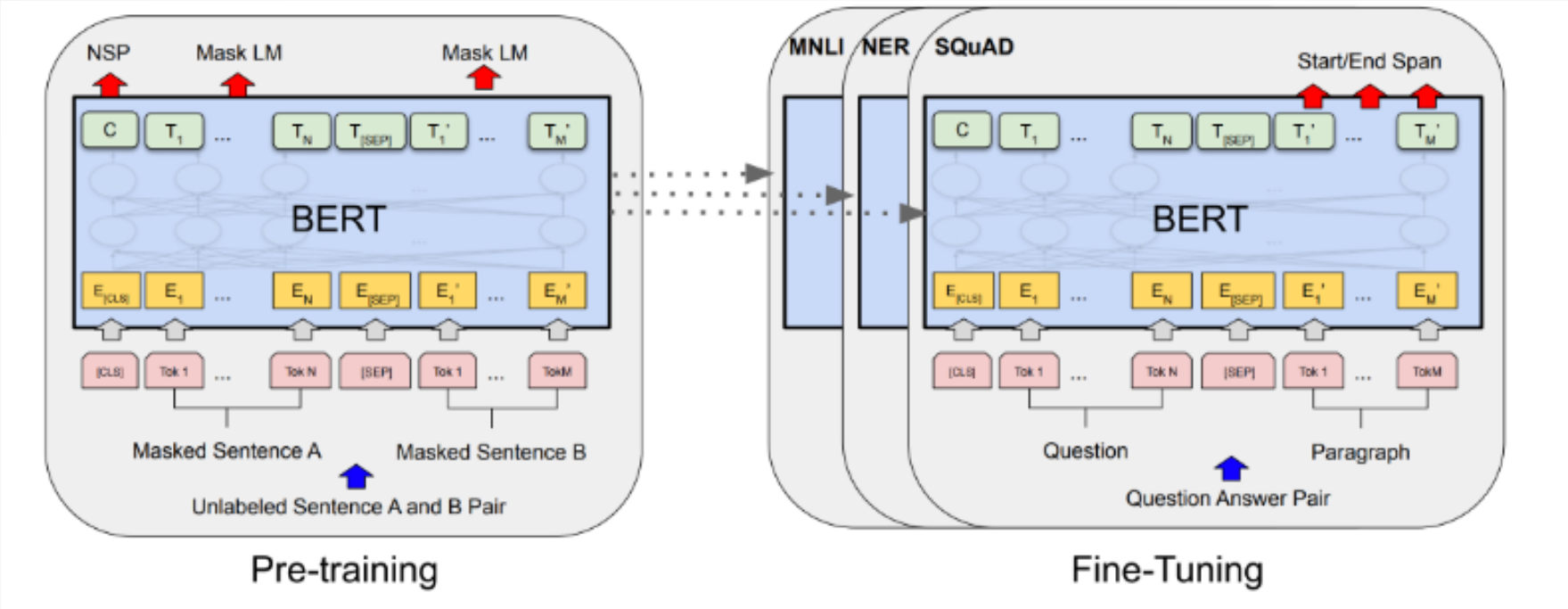
Dataset configuration

- 모
- 행
- 뉴
- 도
- 기

셋


```
"context": "'아빠가 일하는 곳에 와 보니 아빠를 더욱 이해하게 됐어요" - 행정자치부, 「좋은 직장 만들기」 일환 신바람 패밀리 데이 개최 -\n□ 행정자치부(장관 홍윤식)는 22일 오후 직원 가족 등 45명을 정부서울청사로 초청해 다양한 업무공간을 직접 체험하며 서로를 이해할 수 있는 소통.공감 체험 프로그램인 「신바람 패밀리 데이」를 개최했다.\n□ 이날 홍윤식 행정자치부장관은 인사말을 통해 "이곳은 여러분의 자녀 또는 부모님이 국가와 국민의 행복을 위해 늘 고민하고 애쓰는 일터"라며, "오늘 하루 다양하고 즐거운 체험을 통해 꿈과 희망을 갖고 건전하고 밝게 생활해 줄 것"을 당부했다. 홍 장관은 이어 "오늘 행사를 시작으로 매 분기 가족을 초청해 가족 간의 사랑과 직장의 소중함을 느끼도록 하겠다."라고 밝혔다.\n□ 이번 행사에 참석한 직원과 가족들은 장관 집무실을 비롯한 국무회의장, 대한민국 국새(國璽), 스마트워크센터, 정부행정역사관 등 평소 체험하기 힘든 곳을 둘러보며 즐거운 시간을 보냈다.\n□ 이 날 행사에 참석한 ***(남, 12) 어린이는 "아빠가 근무하는 곳이 무척 궁금했는데, 사무실도 둘러보고 무슨 일을 하시는지 알게 되었고, 아빠를 더욱 잘 이해하게 되었다."라며 "나도 아빠와 같은 흠륭한 공무원이 되고 싶다."라는 포부를 밝혔다.\n□ ***(여, 13) 어린이는 "장관님 집무실 의자에 직접 앉아 훗날에 장관이 되는 꿈을 그려 봤다."라고 말했다.\n□ 이번 행사에 관해 행정자치부는 "일회성이 아닌, 직원 유연근무와 자녀 체험학습과 연계해 추진한 것"이라며, "행정자치부는 ①일과 가정이 양립하는 직장문화 조성 ②소통으로 서로 신뢰하는 직장 분위기 조성 ③창의적인 근무 분위기 조성 ④자기계발 및 봉사 나눔 활동의 생활화 등 4개 분야 17개 시책의 「좋은 직장만들기」 프로젝트를 지속적으로 추진하겠다."라고 강조했다.",
        "qas": [
          {
            "qa_type": 1,
            "question_id": "5363243",
            "question": "신바람 패밀리 데이를 개최한 기관은 어디니",
            "is_impossible": false,
            "answers": {
              "text": "행정자치부",
              "answer_start": 74,
              "clue_start": null,
              "clue_text": null,
              "options": null
            }
```

Model-Background(BERT)



BERT

- self-supervised learning 학습을 이용

1. 많은 양의 unlabeled corpus 이용하여 언어 자체에 대해 배워가는 pre-train 단계
2. 특정 도메인의 태스크(down-stream task)를 학습하는 fine-tuning단계로 구성

- 이 두 단계를 차례로 거쳐 학습

학습 모델-RoBERTa

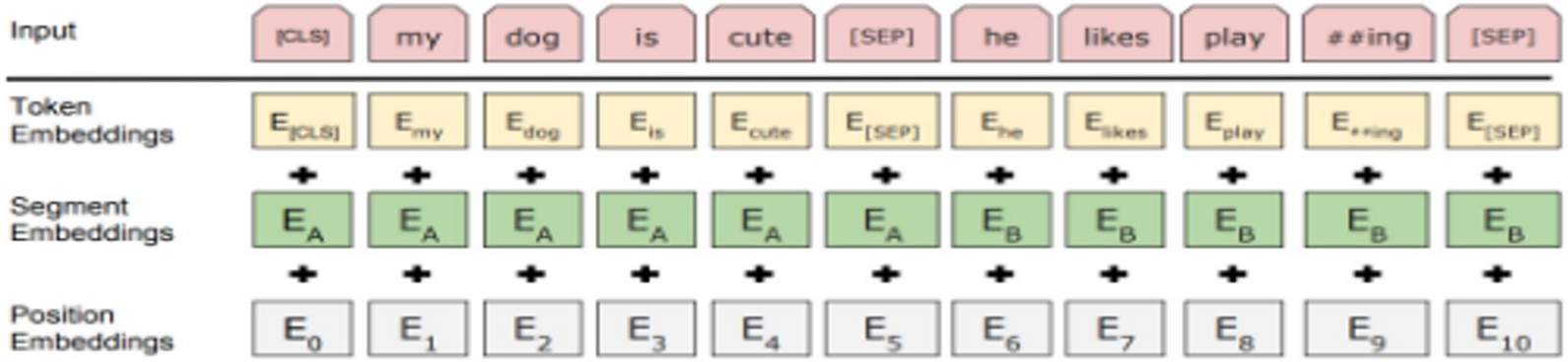


Figure 2: BERT input representation. The input embeddings is the sum of the token embeddings, the segmentation embeddings and the position embeddings.

1. NSP(Next Sentence Prediction) 제거

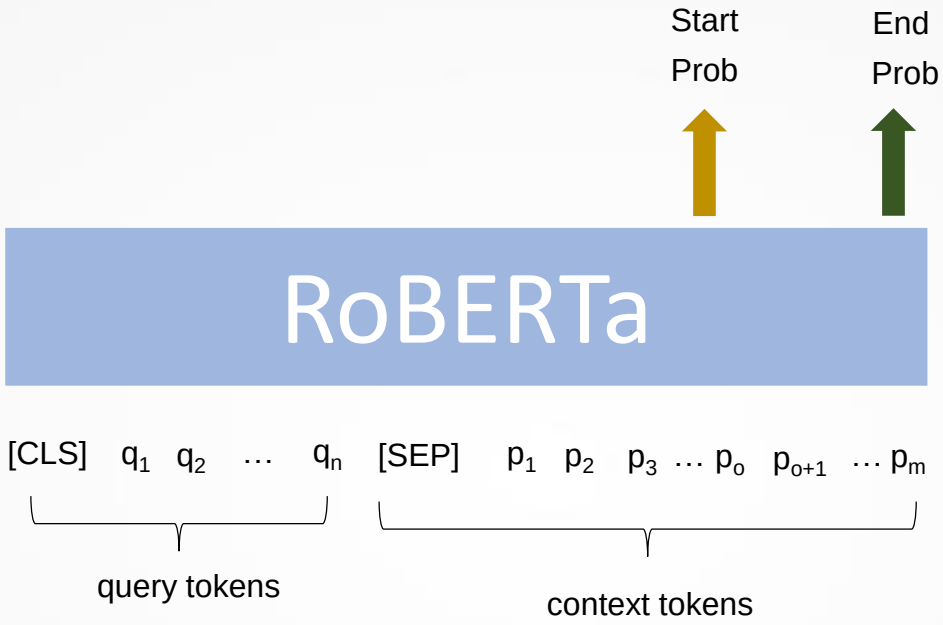
- NSP가 있는 경우와 없는 경우의 여러가지 세팅으로 실험했을 때, 제거된 문장이 더 나은 결과

2. Dynamic Masking

- BERT에서는 Static Masking을 이용, 학습 도중 동일한 Mask를 갖는 입력 문장을 반복해서 보게 됨
- 매 epoch마다 다른 Masking을 사용하는 Dynamic Masking

Research method- cross-evaluation

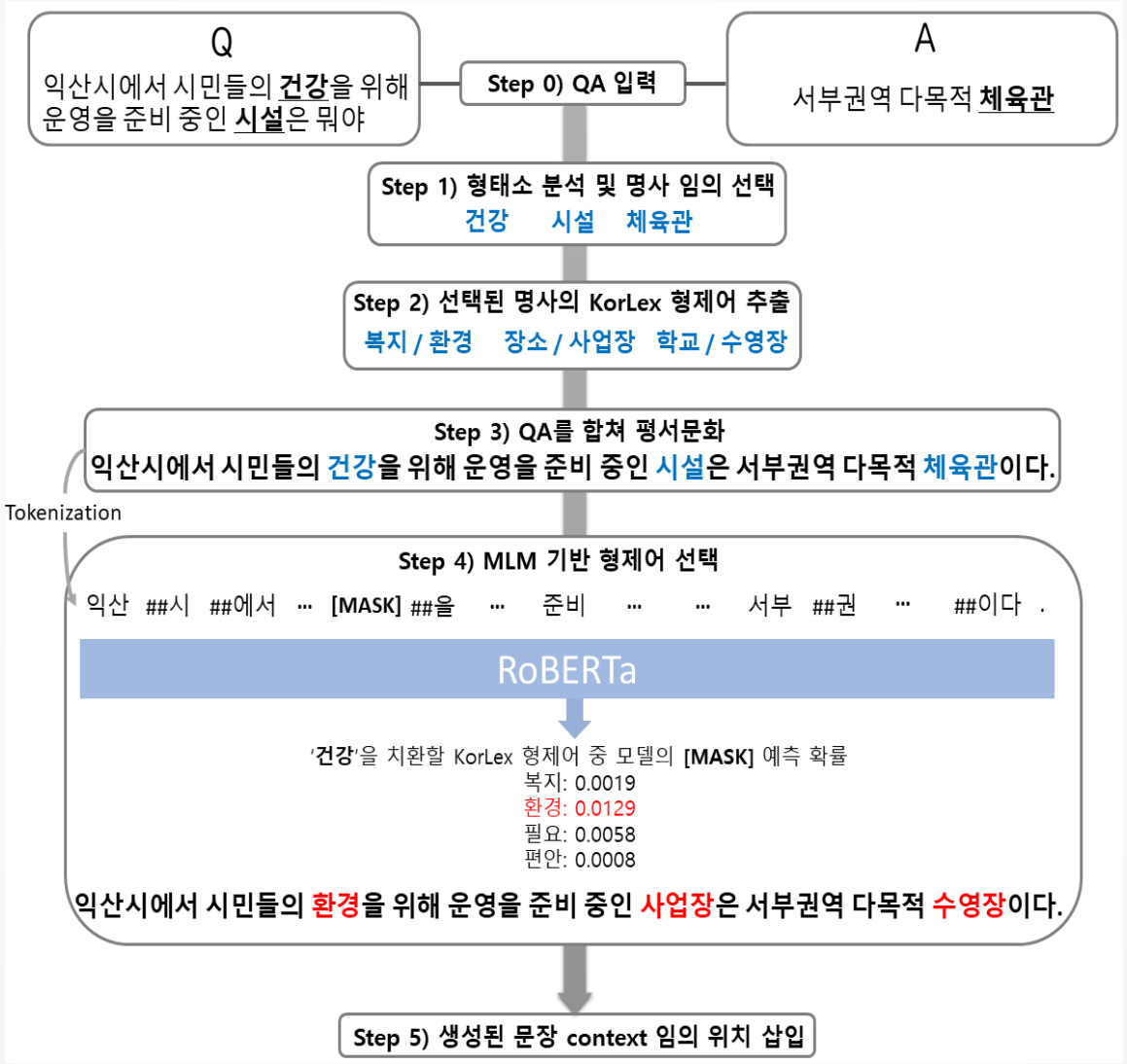
Cross-evaluation



- 각 데이터셋(행정, 뉴스, 도서, 기계독해)으로 RoBERTa를 fine-tuning한 모델 생성
- 모든 데이터셋의 validation dataset을 이용하여 학습된 기계독해 모델의 성능을 평가

Research method-Sentence generation for adversarial training

적대적 학습을 위한 가짜(Fake) 문장 생성



Results

- Sequence length 512, batch-size 24, document stride 128, learning rate 0.0005, warmup ratio 0.1, max_answer_length 20, epoch 4
- 모델은 HuggingFace에서 배포되고 있는 RoBERTa-base이용

Cross evaluation

Train \ Test	행정		뉴스		도서		기계독해		평균	
	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
행정	**89.71**	**96.21**	55.91	72.34	76.67	86.28	49.51	78.10	**70.20**	**83.23**
뉴스	**80.80**	**92.17**	69.29	82.41	74.83	87.01	48.10	72.24	68.25	83.45
도서	74.38	87.12	50.10	66.07	**77.33**	**87.43**	33.72	57.05	52.13	74.41
기계독해	56.99	81.48	34.83	60.84	53.00	73.47	**77.41**	**91.84**	55.55	76.90

- 각각 학습한 모델은 학습 도메인과 동일할 때 높은 성능
- 도서 데이터셋 낮은 성능 => 다른 데이터셋과 도메인 차이 때문에 일반화가 잘 되지 않는 것으로 예상
- 기계독해 데이터셋 낮은 성능 => 초창기에 구축되어 제작 규칙이 확립되지 않아 데이터셋의 도메인 편향이 큰 것으로 예상
- 가장 일반화가 잘 된 데이터셋은 행정 문서

: 일반화 성능을 유지할 수 있도록 다양한 데이터셋 구축 필요

Results

Administrative evaluation

Train \ Test	Origin validation dataset		Adversarial validation dataset	
	EM	F1	EM	F1
행정	90.55	97.15	46.67(-43.88)	80.25(-16.9)
행정(+Adv)	62.90	83.70	73.43(+10.53)	90.67(+6.97)
뉴스	92.83	96.78	47.03(-45.8)	78.46(-18.32)
뉴스(+Adv)	69.23	83.34	74.38(+5.15)	87.25(+3.91)
도서	62.28	89.24	40.24(-22.04)	69.44(-19.8)
도서(+Adv)	68.99	83.50	95.18(+26.19)	97.96(+14.46)
기계독해	62.28	89.24	47.38(-14.9)	82.76(-6.48)
기계독해(+Adv)	58.53	87.72	84.58(+26.05)	96.32(+8.6)

- 각 도메인당 데이터셋 80,000개 제한
- 일반 모델에 적대적 평가를 진행한 결과, 일반 평가보다 성능이 매우 하락
- 적대적 학습을 적용한 모델의 경우 어느 정도 성능을 유지하면서 일반 모델보다 적대적 평가에서 높은 성능을 보임

: 일반 데이터셋은 모델이 잘 맞출 수 밖에 없는 구조로 편향적이게 제작되어, 정답을 포함한 문장과 유사한 구조를 가진 문장을 만났을 때 강건하게 대응할 수 없음을 확인

Conclusion

- 교차 평가 결과에서 일부 데이터셋의 평균 성능이 특히 더 떨어지는 결과를 보임
- 해당 데이터셋(기계독해, 도서)는 일반화가 잘 되지 않음

⇒ 뉴스와 기계독해 데이터셋은 같은 도메인의 문서(뉴스 데이터)를 이용했음에도 일반화 성능 차이가 큼

- 적대적 학습 시 기존 데이터셋과 변형된 데이터셋으로 모델을 각각 학습
- 일반 모델에 적대적 평가를 진행한 경우, 적대적 평가에서 모든 도메인 성능 하락

⇒ 모델은 정답 예측 시 문맥을 이해하기보다 구조적인 특정 단서를 활용함을 알 수 있음

⇒ 데이터셋이 특정 단서를 반복 학습할 수 있게 편향적으로 구축되었음을 확인

: 추후 데이터셋 구축 시 데이터셋에 내재하는 편향을 피하기 위해 다양한 도메인의 활용과 지문에 맞춰진 질문 제작이 아닌 통사 구조 및 어휘의 변형이 필요함