상용화를 위한 신뢰 점수 기반 기계독해 모델

이현구○, 김학수
강원대학교 컴퓨터정보통신공학과
nlphglee@kangwon.ac.kr, nlpdrkim@kangwon.ac.kr

Confidence Score based Machine Reading Comprehension for Commercialization

Hyeon-gu Lee○, Harksoo Kim
Kangwon National University Computer and Communication Engineering

요 약

상용화 서비스를 위한 기계독해 시스템은 출력되는 응답의 정확도가 낮으면 사용자 만족도가 급격히 감소하는 문제가 있다. 응답의 정확도를 높이기 위해서는 모델의 성능을 향상시키거나 신뢰도를 파악하여 확실한 정답만 출력하고 판단하기 모호한 정답은 출력하지 않는 것이 좋다. 또한 현재 주어진 문맥에서 해결할 수 없는 질의의 경우 정답이 없음을 알려줘야 한다. 하지만 모델의 성능을 향상시키기 위해서는 모델이 매우 복잡해져 높은 성능의 하드웨어가 필요하며 추가 데이터가 필요하다. 본 논문에서는 이러한 문제를 해결하기 위해 정답을 찾을 수 있는 질의로만 구성된 말뭉치에서 부정 데이터를 생성하고 신뢰 점수를 계산 할 수 있는 신뢰 노드를 추가하여 정확도를 향상시키는 모델을 제안한다. 실험 결과 응답 재현율은 떨어지지만 신뢰 점수 임계값에 비례하여 정확률이 향상되는 것을 보였다.

주제어: 기계독해, 신뢰 점수, 부정 데이터 생성, 상용화 서비스

1. 서론

기계독해(Machine Reading Comprehension)는 주어진 문맥을 통해 관련된 질문을 해결하는 질의응답 모델이다. 최근 기계독해는 사람을 뛰어넘는 성능을 보이며 상용화 서비스에 계획 및 적용되고 있다. 그러나 상용화 서비스는 정확도가 매우 중요하며 잘못된 결과를 보여줄 시 사용자 만족도는 급격히 감소하게 된다. 특히 주어진 문맥에서 정답을 찾을 수 없거나 확신이 적은 경우 서비스 정확도를 위해 응답을 하지 말아야한다. 최근 이러한 문제를 해결하기 위해 무응답 데이터를 추가하는 방법[1]이 제안되었으나 추가 데이터를 구축하는 비용이 많이 들며 높은 성능을 달성하기 위해 모델이 복잡해지는 문제가 발생한다. 본 논문은 이러한 문제를 완화하기 위해 정답이 있는 데이터만으로 부정 데이터(Negative data)를 생성하고 신뢰 점수(Confidence score)를 이용한 정확도 향상 모델을 제안한다. 제안 모델은 신뢰 점수를 이용하여 확실하지 않은 응답은 출력하지 않기에 응답 재현율은 떨어지지만 상용화 서비스 입장에서 부정확한 답변으로 인한 사용자 만족도 감소를 완화해줄 것이라 기대된다.

2. 관련 연구

기계독해는 SQuAD v1.1[2]이 공개되면서 활발히 연구되기 시작했다. BiDAF[3], R-Net[4]과 같은 초기 모델들은 인코딩 계층, 주의집중 계층, 출력 계층의 기본 구조를 정립하여 높은 성능을 보였다. 주어진 문맥에서 정답을 찾을 수 있는 질문만으로 구성된 SQuAD v1.1이 사람보다 높은 성능을 내자 SQuAD v2.0[1]이 공개되었다. SQuAD v2.0은 주어진 문맥에서 정답을 찾을 수 없는 질의(Unanswerable questions)를 추가하여 응답을 할 수 없는 경우도 판단해야하는 조금 높은 수준의 데이터로 BERT[5]가 공개되면서 좋은 성능을 보였다. 한국어의 경우 SQuAD v1.1과 동일한 형태로 구성된 KorQuAD[6]가 공개되었고 BERT 기반의 발전된 모델을 이용하여 높은 성능을 달성하였다. 그러나 KorQuAD의 경우 상용화에 필수적인 무응답 처리를 할 수 없는 데이터이다. 또한 높은 성능을 내기 위한 BERT의 경우 파라미터의 수가 많아 고성능의 하드웨어가 필요하며 그로 인해 상용화에 어려움이 있다. 본 논문은 무응답이 포함되어있지 않은 데이터 문제와 파라미터 크기를 완화하기 위해 정답이 있는 데이터만을 이용하여 부정 데이터를 생성하고 신뢰 점수 기반의 모델을 통해 성능은 낮지만 가벼운 모델을 개선하여 확실한 정답만 출력해 정확도를 개선하는 방법을 제안한다.

3. 신뢰 점수 기반 기계독해

그림 1은 제안 모델의 흐름을 나타낸다. 전체 구조는 초기 모델 학습, 부정 데이터 생성, 전이 학습(Transfer Learning) 기반 신뢰 점수 모델로 구성된다. 제안 모델에서 기계독해 모델은 GF-Net[7]을 사용한다.

3.1. 초기 모델 학습

초기 모델 학습은 부정 데이터 생성 및 전이 학습을 위한 용도로 크로스 엔트로피(cross entropy)를 통해 기

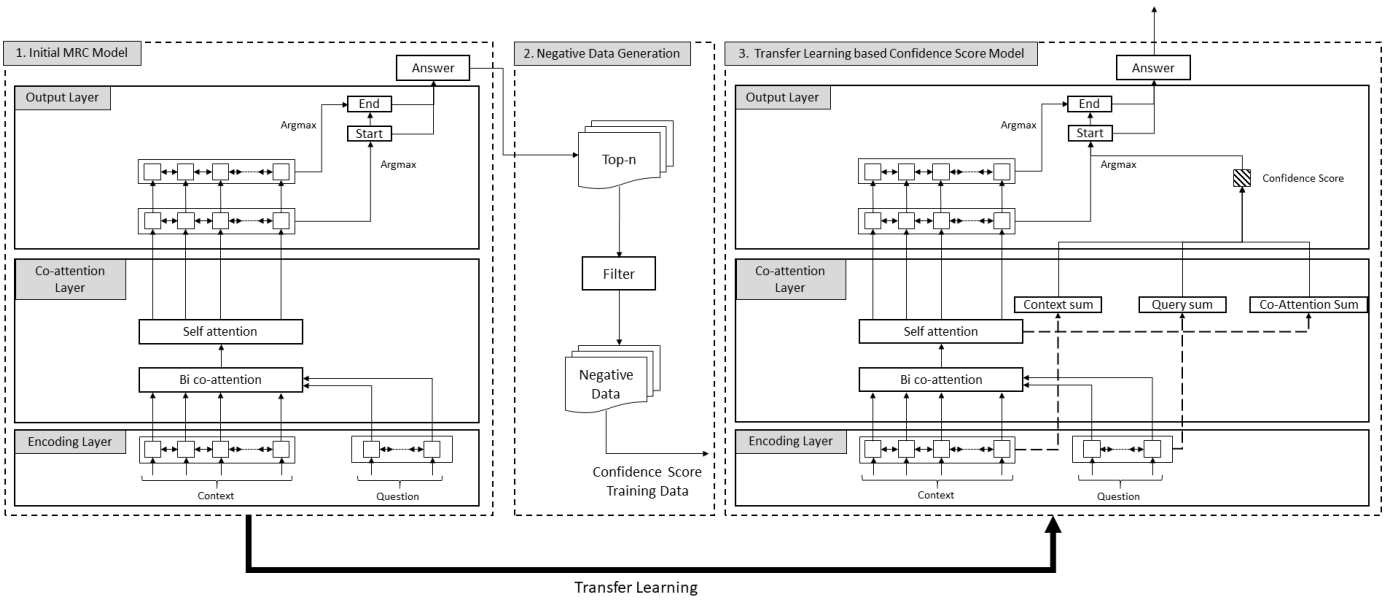

그림 1. 제안 모델의 전체 구조도

계독해 모델을 학습한다. 모델에 사용된 손실 함수(Loss function)는 식 1과 같다.

$$L(\theta) = -\frac{1}{N}\sum_{i}^{N}[\log(p_{y_i^1}^1) + \log(p_{y_i^2}^2)] \quad (1)$$

식 1에서 $p_{y_i^1}^1$와 $p_{y_i^2}^2$는 예측된 확률 분포에서 정답의 시작 위치 y_i^1와 끝 위치 y_i^2의 확률 분포이다.

3.2. 부정 데이터 생성

부정 데이터는 학습된 초기 모델을 통해 생성한다. 그림 2는 부정 데이터 생성의 단계를 나타낸다.

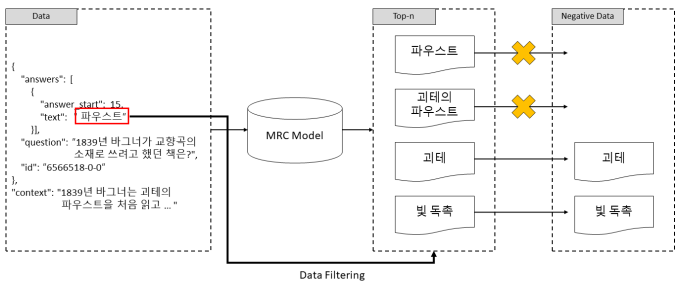

그림 2. 부정 데이터 생성 단계

초기 모델을 통해 학습 데이터에서 top-n 결과를 추출한다. top-n의 결과 중 정답의 어휘가 나타나지 않는 모든 결과를 추출하여 부정 데이터로 사용한다. 예제와 같이 ton-n으로 추출된 [“파우스트”, “괴테의 파우스트”, “괴테”, “빛 독촉”]의 결과 중 실제 정답인 “파우스트”를 제외하고 어휘가 겹치지 않는 [“괴테”, “빛 독촉”]을 부정 데이터로 사용한다. 이는 학습된 모델에서 응답으로 적절하다 할 확률이 높은 데이터로 실제 모델이 해결하기 어려워하는 내용을 부정 데이터로 사용하여 동떨어진 응답(예: 다른 질의의 정답)보다 신뢰 점수를 계산할 때 효과적이기 때문이다.

3.3. 전이 학습 기반 신뢰 점수 모델

신뢰 점수 모델은 초기 학습 모델에 신뢰 점수를 계산하기 위한 노드를 추가한다. 신뢰 점수를 계산하기 위한 노드의 수식은 식 2와 같다.

$$score(C,Q,A) = \sigma(FNN(s(C),s(Q),s(A)))$$
$$s(X) = \sum_{i}^{N} softmax(\frac{XX^T}{\sqrt{d_X}})X \quad (2)$$

식 2에서 C는 문맥, Q는 질의, A는 문맥과 질의간의 상호 집중(co-attention)의 인코딩 벡터이다. $s(X)$는 각 인코딩 벡터를 자가 집중(self-attention)하여 합한 벡터이며 d_X는 벡터 X의 은닉크기(hidden size)이다.

신뢰 점수 노드가 추가된 모델은 전이 학습 기반으로 학습한다. 전이 학습은 기존의 학습된 모델의 파라미터 값을 새로운 모델에 전달하여 빠르게 학습하는 방법으로 본 논문에서는 초기 학습 모델을 전이하여 기본적인 기계독해 기능을 가진 상태에서 신뢰 점수 노드를 미세조정(fine-tuning)하여 학습한다.

신뢰 점수 노드를 학습하기 위한 손실 함수는 식 3과 같다.

$$L(\theta) = -(1-\delta)\log\sigma(score) - \delta\log(1-\sigma(score)) \quad (3)$$

식 3은 sigmoid cross entropy를 나타내며 δ는 부정 데이터의 여부를 나타내며 부정 데이터의 경우 0, 아닐

경우 1의 값을 가진다.

4. 실험 및 평가

4.1. 실험 준비

본 논문에서는 실험을 위해 KorQuAD를 사용한다. KorQuAD는 학습 데이터 60,407개 개발 데이터 5,774개가 공개 되어 있으며 개발 데이터를 평가 데이터로 사용한다.

4.2. 실험 평가

본 논문에서 실험 지표로 완전 일치율(Exact Match), F1-점수(F1-score), 정답 재현율(Recall)을 사용한다. 완전 일치율은 모델이 출력한 응답과 실제 정답이 완전히 일치하면 1 아니면 0으로 계산하고 F1-점수는 실제 정답과 출력 응답간의 부분 일치 성능을 나타낸다. 정답 재현율은 신뢰 점수의 임계값으로 인해 응답하지 못한 비율을 측정하기 위해 사용한다.

표 1과 그림 3은 임계값 별 성능을 나타낸다. x축은 신뢰 점수의 임계값을 나타내고 y축은 성능을 나타낸다.

표 1. 임계값 별 성능 변화

Threshold	Exact Match	F1-score	Recall
초기 모델	0.7246	0.8732	1.0000
0	0.6791	0.8390	1.0000
0.05	0.6962	0.8553	0.9628
0.1	0.7169	0.8728	0.9065
0.15	0.7353	0.8859	0.8585
0.2	0.7472	0.8950	0.8140
0.25	0.7561	0.9006	0.7724
0.3	0.7698	0.9104	0.7335
0.35	0.7801	0.9174	0.7000
0.4	0.7853	0.9201	0.6751
0.45	0.7943	0.9262	0.6434
0.5	0.8030	0.9313	0.6127
0.55	0.8103	0.9353	0.5852
0.6	0.8154	0.9378	0.5546
0.65	0.8219	0.9416	0.5222
0.7	0.8324	0.9458	0.4939
0.75	0.8421	0.9501	0.4617
0.8	0.8482	0.9534	0.4245
0.85	0.8616	0.9579	0.3843
0.9	0.8696	0.9623	0.3386
0.95	0.8920	0.9710	0.2695

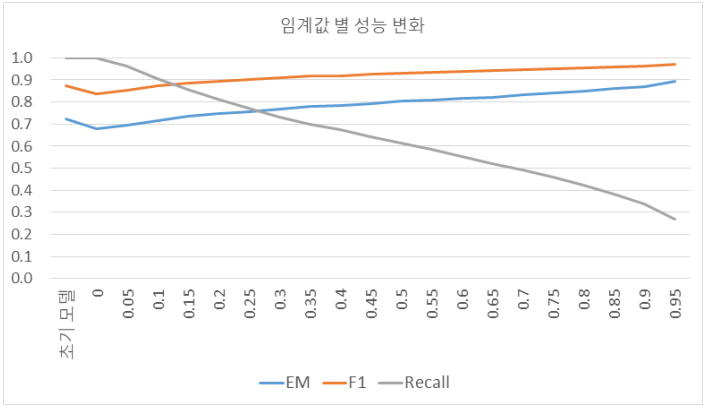

그림 3. 임계값 별 성능 변화

표1과 그림 3에서 “초기 모델”은 3.1장에서 학습된 초기 모델이며 임계값 0의 지표는 전이 학습을 통해 학습된 모델의 성능을 의미한다. 신뢰 점수 모델의 기본 성능이 초기 모델보다 떨어진 이유는 미세 조정을 하면서 기존 정답 위치를 찾아주는 값이 조정되어 악영향을 끼친 것으로 보인다. 실험 결과 임계값을 높일수록 정답 재현율은 감소하지만 완전 일치율과 F1-점수가 오르는 것을 알 수 있다.

5. 결론 및 향후 연구

본 논문에서는 상용화 서비스에 필요한 무응답 처리 및 하드웨어 비용 문제를 완화하기 위한 부정 데이터 생성 및 신뢰 점수 기계독해 모델을 제안하였다. 제안 모델은 무응답 정보가 없는 데이터를 통해 부정 데이터를 생성하고 기존 학습된 모델을 신뢰 점수 모델에 전이 학습하여 신뢰 점수를 측정할 수 있었다. 신뢰 점수를 임계값으로 사용하여 정답 재현율은 떨어지지만 임계값에 비례해 성능이 개선되는 것을 보였다. 향후 연구로 생성된 부정 데이터가 기존 성능에 주는 악영향을 줄이기 위해 새로운 구조의 신뢰 점수 모델을 연구할 예정이다.

감사의 글

본 논문은 한국산업단지공단의 이전기술사업화 과제인 "민원상담 사용자 편의성 증진을 위한 질의 연관문서 자동선별 및 기계문서이해 기반 대화형 응답플랫폼 개발"(주관기관 : 주식회사 포스원) 과제에 기반을 둔 기술논문으로써, 본 논문 작성을 위한 기반을 마련해준 한국산업단지공단에 감사의 말씀을 드립니다.

참고문헌

[1] P. Rajpurkar, R. Jia and P. Liang, Know What You Don't Know: Unanswerable Questions for SQuAD, arXiv preprint arXiv:1806.03822, 2018.

[2] P. Rajpurkar, J. Zhang, K. Lopyrev and P. Liang, SQuAD: 100,000+ Questions for Machine Comprehension of Text, arXiv preprint

arXiv:1606.05250, 2016.
[3] M. Seo, A. Kembhavi, A. Farhadi and H. Hajishirzi, Bidirectional attention flow for machine comprehension, arXiv preprint arXiv:1611.01603, 2016.
[4] W. Wang, N. Yang, F. Wei, B. Chang and M. Zhou, Gated Self-Matching Networks for Reading Comprehension and Question Answering, Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vol 1, pp. 189-198, 2017.
[5] 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.
[6] 임승영, 김명지, 이주열. "KorQuAD: 기계독해를 위한 한국어 질의응답 데이터셋" 한국정보과학회 학술발표논문집, pp. 539-541, 2018.
[7] 이현구, 김학수, "GF-Net 자질 선별을 통한 고성능 기계독해", 2018 한국컴퓨터종합학술회의, pp. 598-600, 2018.