의료 분야 기계독해를 위한 다중 스팬 추출 모델

장영진, 김학수
건국대학교 인공지능학과, 건국대학교 컴퓨터공학과
danyon@konkuk.ac.kr, nlpdrkim@konkuk.ac.kr

Long Multispan Prediction Model for Machine Reading Comprehension in Healthcare Domain

Yongjin Jang, Harksoo Kim
Konkuk University Department of Artificial intelligence,
Konkuk University Depaertment of Computer Science and Engineering

요 약

기계독해는 주어진 문서에서 질문에 대한 답변 스팬을 찾아 사용자에게 제공하는 질의응답 작업 중 하나이다. 최근 대규모 언어 말뭉치를 기반으로 한 언어 모델과 대용량의 데이터셋이 공개됨에 따라, 인간을 뛰어넘는 기계독해 모델이 공개되었다. 하지만 대부분의 기계독해 연구는 단순한 형태의 단답형 단일 스팬 추출에 집중했기 때문에 다수의 스팬이나 긴 길이의 스팬을 요구하는 실제 애플리케이션에서 유용하지 않을 수 있다. 실제로 의료 도메인에서 사용자는 바이러스명과 같이 짧고 단순한 정보보다 질병의 증상과 약의 효과, 부작용 등의 자세한 정보를 요구한다. 따라서 우리는 질문에 대한 답변으로 비연속적인 긴 텍스트 스팬을 추출할 수 있는 기계독해 모델을 제안한다. 의료 도메인 질의응답 데이터셋을 대상으로 진행한 실험에서 제안 모델은 모든 평가 지표에서 이전의 기계독해 모델과 비교하여 가장 높은 성능을 보였다.

1. 서 론

기계독해(Machine Reading Comprehensionl; MRC)를 위한 신경망 개발은 빠르게 성장하는 연구 주제가 되었다. 이로 인해 다양한 데이터셋이 공개되었으며[1-3], 대부분 짧은 단일 스팬 답변 추출 작업에 집중했다. 그러나 이러한 단일 스팬 답변은 복잡하고 여러 유형의 정보가 필요할 수 있는 실제 애플리케이션에서 유용하지 않을 수 있다[4]. 또한 기존의 기계독해 데이터셋에서 높은 성능을 보이는 시스템[5-6]이 다중 스팬 답변이나 긴 스팬 답변을 추출해야하는 환경에서 성능이 크게 저하되는 모습을 보였다. 따라서 우리는 사용자의 질문에 대해 충분한 정보를 제공하고 단일 스팬뿐만 아니라 비연속적인 긴 다중 스팬을 제공하는 기계독해 모델을 설계했으며, 다중 스팬을 효과적으로 추출하기 위한 학습 전략을 제안하고자 한다.

2. 관련 연구

최근 기계독해 시스템[7-8]은 사전학습된 언어 모델을 기반으로 질문과 문서 사이의 연관성을 고려하여 답변 스팬을 추출한다. 이러한 기계독해 시스템은 입력 시퀀스에서 정답 스팬의 시작 및 끝 위치를 학습하며, 추론 단계에서는 모든 입력 토큰 위치에 대한 시작 및 끝 위치 점수의 합으로 결정된다. 이 방법은 간단하고 강력하지만, 입력 시퀀스에서 정답 스팬에 해당하는 위치를 학습하기 때문에 구조적으로 단일 답변 스팬 추출로 제한된다. 또한 독립적으로 계산된 시작 및 끝 위치 점수의 합으로 최종 응답 스팬이 결정되기 때문에 두 위치 간의 정보를 공유할 수 없다는 한계가 있다. 다시말해, 대부분의 기존 기계독해 시스템은 입력 문서에 대한 유일한 정답 스팬의 시작과 끝 위치에 대해 학습되기 때문에 다중 답변 스팬 추출에 적합하지 않다[9]. 따라서 본 논문에서는 위 문제를 극복하기 위해 [10]에서 제안한 기계독해 기반 개체명인식 연구에서 영감을 얻은 Span Matrix를 비연속 다중 답변 스팬에 적합하도록 수정한 프레임워크를 제안하고자 한다.

3. 제안 모델

아래의 그림 1은 본 논문에서 제안하는 모델의 전체 구조를 보여준다.

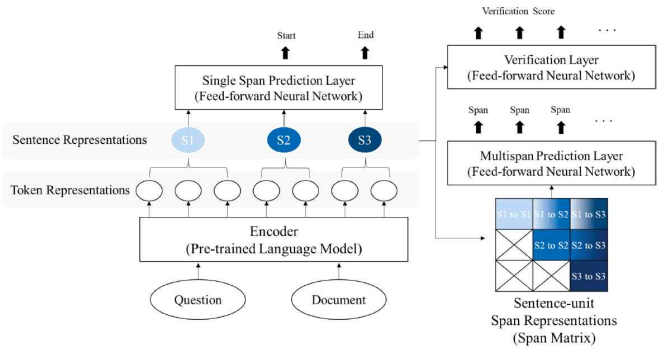

그림 1. 제안 모델 구조도

제안 모델은 크게 세 단계로 구성되어있다. 첫째, 여러 문장으로 구성된 긴 답변 스팬을 추출하기 위해 언어 모델의 토큰 표현에서 문장 벡터을 얻는다. 둘째, 입력 문서에서 가능한 모든 문장 단위 스팬을 나타내는 Span Matrix를 생성한다. 마지막으로 Span Matrix의 모든 요소에 대한 Verification 점수를 계산하고 실험적으로 설정된 Threshold보다 높은 모든 스팬을 답변 스팬으로 결정한다. 구체적으로 문장 단위 벡터는 아래의 그림 2와 같이 언어 모델의 토큰 벡터로부터 얻는다.

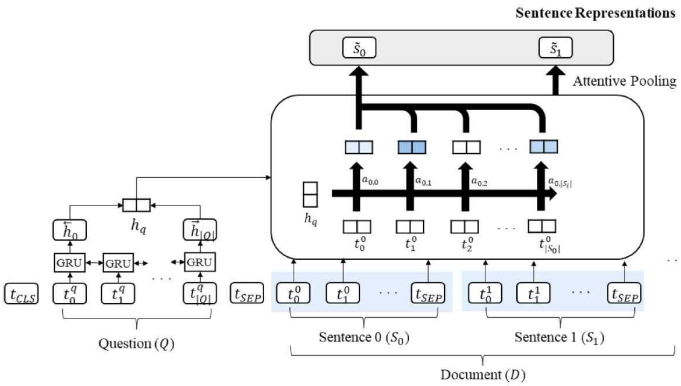

그림 2. 문장 표현 방법

다음으로 우리는 답변 스팬의 시작과 끝 위치를 독립적으로 학습하는 기존의 MRC 프레임워크를 개선하기 위해 아래의 그림 3과 같이 문장 단위로 표현 가능한 모든 스팬을 행렬로 나타내고 Span Matrix에서 하나 이상의 요소를 답변 스팬으로 학습한다.

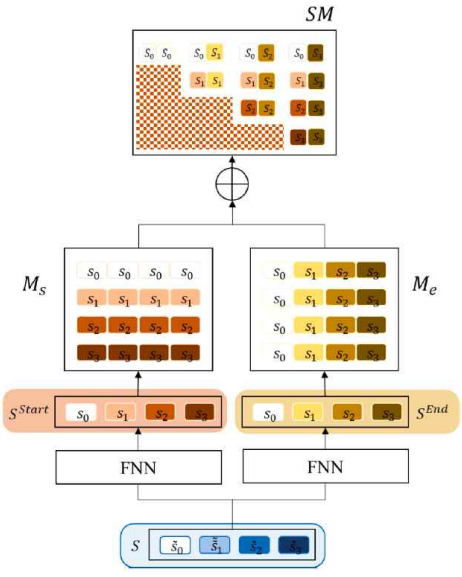

그림 3. Span Matrix 생성 방법

Span Matrix의 각 요소는 입력 문장의 가능한 모든 조합으로 구성된 스팬을 나타낸다. 본 논문에서는 제안모델을 효과적으로 훈련시키기 위해 기존 기계독해 프레임워크에서 사용된 손실 함수와 Span Matrix의 각 행과 열에 대한 CrossEntropy를 함께 사용한다. 추론 단계에서는 Span Matrix의 각 요소의 확률 값이 Threshold보다 높은 스팬을 최종 답변으로 추출한다.

4. 실 험

실험에는 하나 이상의 답변이 부착되어있는 의료 도메인 데이터셋 MASHQA[4]를 사용했다. 약 2만 개의 학습데이터와 약 3천 개의 개발 데이터셋을 포함하고 있으며, 질문에 대한 답변은 평균 2.58개의 스팬이 부착되어있다. 실험 결과에 대한 성능 표는 아래의 표 1과 같다.

표 1. 실험 결과

Model	Precision	Recall	F1 score	EM
RoBERTa + MRC	57.7	19.0	28.6	9.4
MultiCo [4]	58.1	55.9	57.0	22.0
Proposed Model	64.5	61.3	62.9	34.1

표 1에서 RoBERTa+MRC는 RoBERTa를 사용한 기존 기계독해 프레임워크를 의미하고 MultiCo는 시퀀스 라벨링 방법에 기반한 이전 SOTA 모델이다. 표 1에 나와있듯 기존 기계독해 프레임워크는 다중 스팬 추출에 적합하지 않은 것을 알 수 있으며, 제안 모델이 모든 평가 지표에서 가장 높은 성능을 보이는 것을 알 수 있다. 또한 시퀀스 라벨링 기반의 모델인 MultiCo와 비교하여 다중 답변 추론에 제안 모델의 구조가 더 효과적인 것을 알 수 있다.

5. 결 론

본 논문에서는 의료 영역의 기계독해 작업을 위해 길고 비연속적인 여러 스팬을 예측하기 위한 모델을 제안했다. 긴 스팬를 예측하기 위해 제안 모델은 토큰 단위 벡터 문장 단위 벡터로 변환했으며, 문장 단위 Span Matrix를 사용하여 연속되지 않은 답변 스팬을 선택했다. 대표적인 의료 기계독해 데이터셋인 MASHQA에 대한 실험에서 제안 모델은 F1 점수 65.3%로 가장 높은 성능을 달성했으며, 이를 통해 다중 스팬 추출에 제안 모델의 구조가 효과적인 것을 알 수 있다.

감사의 글

이 논문은 2022년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No. 2022-0-00369, (4세부) 전문지식 대상 판단결과의 이유/근거를 설명가능한 전문가 의사결정 지원 인공지능 기술개발)

참고 문헌

[1] Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., & Deng, L. (2016). MS MARCO: A human generated MAchine reading comprehension dataset. *Proceedings of the Neural Information Processing Systems Workshop on Cognitive Computation: Integrating neural and symbolic approaches*.

[2] Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). SQuAD: 100,000+ questions for machine

comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 2383–2392).

[3] Yang, Y., Yih, W. T., & Meek, C. (2015). WIKIQA: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 2013–2018).

[4] Zhu, M., Ahuja, A., Juan, D. C., Wei, W., & Reddy, C. K. (2020). Question answering with long multiple-span answers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings* (pp. 3840–3849).

[5] Seo, M., Kembhavi, A., Farhadi, A., & Hajishirzi, H. (2017). Bidirectional attention flow for machine comprehension. *5th International Conference on Learning Representations*.

[6] Yu, A. W., Dohan, D., Luong, M. T., Zhao, R., Chen, K., Norouzi, M., & Le, Q. V. (2018). QANet: Combining local convolution with global self-attention for reading comprehension. *6th International Conference on Learning Representations*.

[7] Joshi, M., Chen, D., Liu, Y., Weld, D. S., Zettlemoyer, L., & Levy, O. (2020). SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics, 8*, 64–77. https://doi.org/10.1162/tacl_a_ 00300

[8] Zhang, Z., Yang, J., & Zhao, H., (2021), Retrospective Reader for Machine Reading Comprehension. *In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35*, No. 16, pp. 14506-14514.

[9] Segal, E., Efrat, A., Shoham, M., Globerson, A., & Berant, J. (2020). A simple and effective model for answering multi-span questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (pp. 3074–3080).

[10] Li, X., Feng, J., Meng, Y., Han, Q., Wu, F., & Li, J. (2020). A unified MRC framework for named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5849–5859).