질문-단락 간 N-gram 주의 집중을 이용한 단락 재순위화 모델

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

Passage Re-ranking Model using N-gram attention between Question and Passage

Youngjin Jang○, Harksoo Kim
Department of Aritificial Intelligence, Konkuk University

요 약

최근 사전학습 모델의 발달로 기계독해 시스템 성능이 크게 향상되었다. 하지만 기계독해 시스템은 주어진 단락에서 질문에 대한 정답을 찾기 때문에 단락을 직접 검색해야하는 실제 환경에서의 성능 하락은 불가피하다. 즉, 기계독해 시스템이 오픈 도메인 환경에서 높은 성능을 보이기 위해서는 높은 성능의 검색 모델이 필수적이다. 따라서 본 논문에서는 검색 모델의 성능을 보완해 줄 수 있는 오픈 도메인 기계독해를 위한 단락 재순위화 모델을 제안한다. 제안 모델은 합성곱 신경망을 이용하여 질문과 단락을 구절 단위로 표현했으며, N-gram 구절 사이의 상호 주의 집중을 통해 질문과 단락 사이의 관계를 효과적으로 표현했다. KorQuAD를 기반으로한 실험에서 제안모델은 MRR@10 기준 93.0%, Top@1 Precision 기준 89.4%의 높은 성능을 보였다.

주제어: 검색 모델, 재순위화, N-gram 주의 집중, 기계독해

1. 서론

기계독해(Machine Reading Comprehension: MRC)는 주어진 단락에서 질문에 대한 정답을 추출하는 시스템을 의미한다. 최근 사전학습 모델[1]의 등장으로 인해 기계독해 연구도 크게 발전을 이루었다. 하지만 기계독해 시스템은 앞서 언급했던 것처럼 주어진 단락 즉, 정답을 포함하는 단락이 입력된다고 가정한다. 위와 같은 가정은 별도의 검색 과정이 필요한 실제 환경에서 유효하지 않다. 따라서 실제 환경에서는 기계독해 모델에게 질의에 적합한 단락을 제공해줄 수 있는 높은 성능의 검색 모델이 필요하다. BM25[2]와 같은 기존의 검색 모델은 구현이 간단하고 속도가 빠르다는 장점이 있지만, TF-IDF[3] 점수 기반의 어휘 일치 점수로 문서를 검색하기 때문에 의미 정보를 반영하기 어렵다. 이러한 검색 모델의 문제점을 완화하기 위해 검색 모델의 결과 상위 N개를 의미 정보 반영 등이 가능한 알고리즘을 통해 다시 순위를 매기는 연구가 진행되었으며, 이를 재순위화(Re-ranking)라고 한다. 본 논문에서는 오픈 도메인 기계독해를 위한 단락 재순위화 시스템을 제안하며, 검색 모델의 결과 상위 10개를 기반으로 재순위화를 진행한다.

본 논문에서는 질문과 단락 사이의 공통 의미 정보가 질문과 유사할수록 입력 단락이 질문에 적합할 것이라고 가정한다. 위 가정에 대한 예시는 아래의 표 1과 같다.

표 1. 질문과 정답 단락 사이의 어휘 유사도 예시

질문(A)	**세종, 은, 조선, 의,** 몇, **대, 왕, 이**, 야, ?
단락(B)	**세종, 은, 조선, 의,** 4, **대, 왕, 이**, 며, 언어학자, **이**, 다, .
(A)와 (B)의 중복 어휘(C)	세종, 은, 조선의, 대, 왕 이
(A)와 (C)의 어휘 유사 정도	70%

표 1은 형태소 단위로 나누어진 질문과 단락 사이의 중복 어휘 유사 정도를 보여준다. 단락과 질문 사이의 중복된 어휘(C)와 질문을 구성하는 어휘(A)는 70% 정도의 유사 정도를 보이는 것을 확인할 수 있다. 위의 예시를 통해 본 논문에서는 질문(=A)과 공통 의미 정보(=C)를 고정된 길이의 벡터로 표현하여 재순위화 문제를 벡터 간 유사도 계산 문제로 변환하여 재순위화를 수행하고자 한다.

2. 관련 연구

TF-IDF를 이용한 기존 검색 모델은 토큰 매칭 방식에 따라 달라지기 때문에 동음이의어와 같은 의미 정보를 고려하기 어렵다. 이 문제를 해결하기 위해 심층 신경망 기반의 검색 모델이 제안되었다[4-5]. [4]는 MLP(Multi

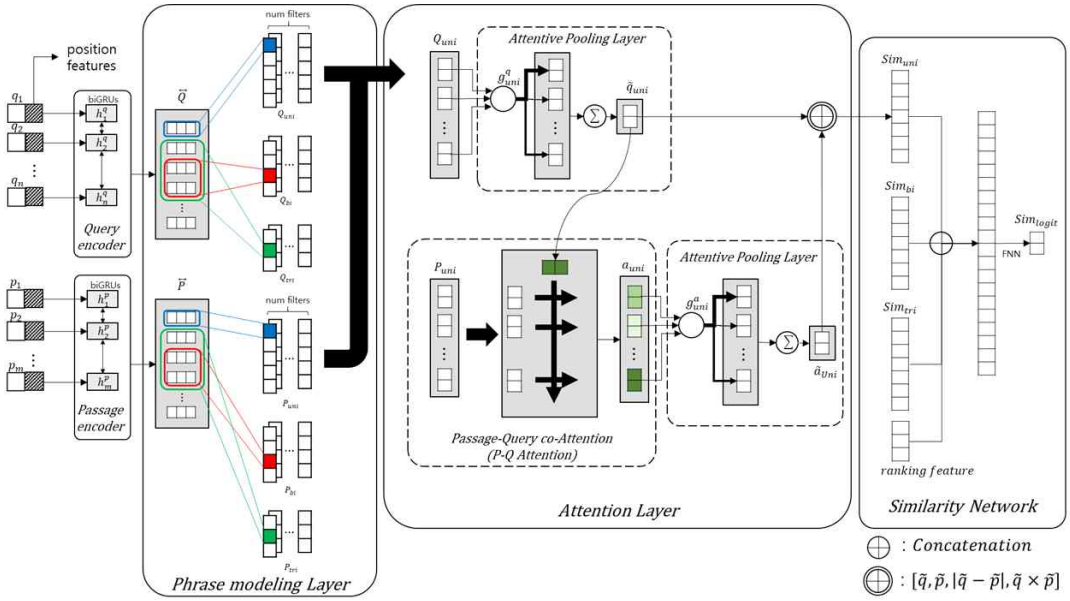

그림 1 제안 모델 전체 구조도

Layer Perceptron)에 의해 생성된 잠재 벡터 공간에서 질의와 문서를 구성하는 모든 토큰 사이의 조합을 통해 관련 점수를 출력하는 모델을 제안했다. [5]는 [4]가 질의와 문서의 길이에 따라 가변적인 벡터 생성 방식이 일관성이 없다고 지적했으며, 이 문제점을 해결하기 위해 길이에 상관없이 비교적 일관적인 합성곱 신경망(Convolutional Neural Network)[6]을 적용한 검색 모델을 제안했다. 위의 검색 모델 연구는 기존의 전통적인 검색 모델 성능을 뛰어넘는 모습을 보였지만, 실제 환경에서는 질의에 대한 문서를 검색하기 위해 모든 문서와 연산을 해야 한다는 단점이 있었다. 이 문제점을 완화하기 위해 심층 신경망을 기반으로 검색 모델의 결과를 재순위화 하는 연구도 진행되었다[7-8]. [7]은 하나의 단락에 대해서만 점수를 예측하는 점이 비효율적이라고 지적했으며, 입력 질의에 대해 검색 모델 결과 N개의 단락을 한 번에 입력 받아 정렬하는 시스템을 제안했으며, [8]은 입력 질문과 단락 사이의 주의 집중과 다중 작업 학습(Multi-task Learning)을 적용한 모델을 제안하여 높은 성능을 보였다.

위와 같이 검색 모델과 관련된 연구는 입력 받은 질의(질문)와 문서(단락) 사이의 상호 정보 계산을 통해 구현되었다. 본 논문에서는 합성곱 신경망을 이용하여 구절 단위의 벡터를 생성하고, 이를 주의 집중 계층에 입력하여 문맥 정보를 효과적으로 반영할 수 있는 재순위화 시스템을 제안한다.

3. 단락 재순위화 모델

위의 그림 1은 본 논문에서 제안하는 시스템의 전체 구조도를 보여준다. 제안 모델은 5개의 세부 모듈(질문 인코더, 단락 인코더, 구절 모델링 계층, 주의 집중 계층, 유사도 측정 네트워크)로 이루어져 있다.

3.1. 질문 인코더와 단락 인코더

질문 인코더와 단락 인코더는 형태소 단위의 문장을 입력 받으며, 각 형태소는 3 가지 유형의 벡터를 연결(concatenation)한 형태로 표현된다. 형태소 벡터 표현에 대한 그림은 아래와 같다.

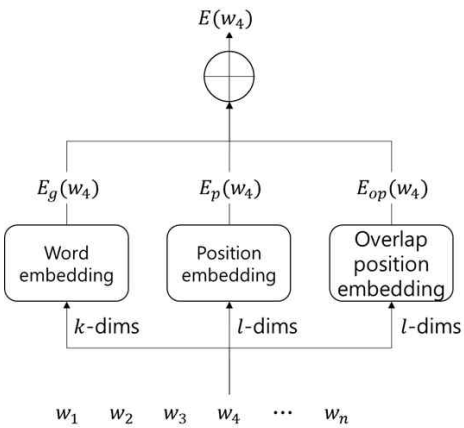

그림 2. 형태소 벡터 표현 방법

그림 2에서 w_i는 질문/단락을 구성하는 i 번째 형태소를 의미한다. $E_g(w_i)$는 w_i에 대한 사전 학습된 GloVe[9] 벡터를 의미한다. $E_p(w_i)$는 w_i에 대한 위치 정보를 의미하고, $E_{op}(w_i)$는 질문과 단락 사이에 서로 겹치는 단어가 몇 번째에 위치하고 있는지에 대한 위치 정보를 의미한다. 위의 과정을 통해 벡터 $E(w_i)$로 표현된 질문과 단락은 GRUs(Gated Recurrent Units)[10]로

구현된 양방향 순환 신경망(Bi-directional Recurrent Neural Network)에 입력되어 각각 $\overrightarrow{Q}, \overrightarrow{P}$로 인코딩된다. 이때, 질문 인코더와 단락 인코더의 가중치는 공유되지 않는다.

3.2. 구절 모델링 계층

본 논문에서 제안하는 구절 모델링 계층은 합성곱 신경망으로 구현되었다. uni/bi/tri-gram convolution 필터를 이용하여 n-gram 자질(feature)을 생성하며, 이때 convolution 계산을 거친 출력 값들은 n-gram 정보를 유지하기 위해 pooling 연산을 거치지 않고 자질 그대로 사용하게 된다. 이에 대한 그림은 아래와 같다.

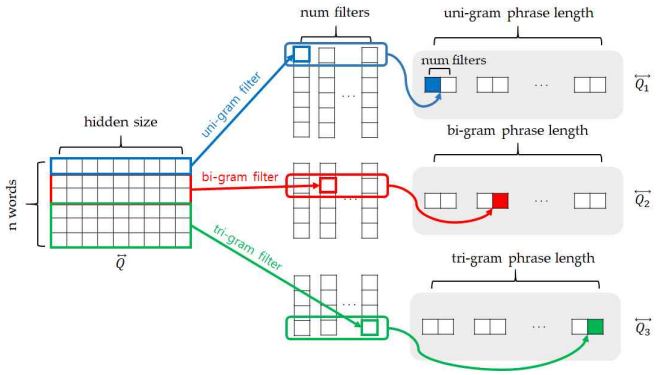

그림 3. 질문 벡터($\overrightarrow{Q}$)에 대한 구절 모델링 예시

그림 3에서 보이는 것과 같이, 인코딩된 $\overrightarrow{Q}$와 $\overrightarrow{P}$는 구절 모델링 계층에 입력되어 구절 단위의 벡터 $\overrightarrow{Q_n}, \overrightarrow{P_n}$로 표현된다.

3.3. 주의 집중 계층

본 논문에서 제안하는 주의 집중 계층은 두 가지의 네트워크(주의 집중 풀링, 상호 주의 집중)로 구성된다. 상호 주의 집중은 질문과 단락 사이의 연관 정도(공통 의미 정보)를 계산하며, [11]에서 제안된 Scaled dot-product attention과 유사하게 계산된다. 이에 대한 수식은 아래의 (1)과 같다.

$$P_{Att}^{n} = softmax(\frac{\overrightarrow{P_n} \cdot \widetilde{Q_n}^{T}}{\sqrt{d_{\widetilde{Q_n}}}})\overrightarrow{P_n} \quad (1)$$

위 수식 (1)에서 $d_{\widetilde{Q_n}}$은 $\widetilde{Q_n}$의 벡터 크기를 의미한다. $\overrightarrow{P_n}$은 구절 모델링 계층을 통해 생성된 단락 n-gram 벡터를 의미하고, $\widetilde{Q_n}$은 주의 집중 풀링[12]을 통해 고정 길이의 벡터로 표현된 질문 n-gram 벡터를 의미한다. 본 논문에서 사용하는 주의 집중 풀링은 P_{Att}^{n}과 $\overrightarrow{Q_n}$을 고정 길이의 벡터($\widetilde{P_n}, \widetilde{Q_n}$)로 표현하기 위해 사용되며, 이에 대한 수식은 아래의 (2)와 같다.

$$\begin{aligned} g_n^q &= softmax(w_n^q(\overrightarrow{Q_n}) + b_n^q) \\ \widetilde{Q_n} &= (\sum g_n^q \times \overrightarrow{Q_n}) \\ g_n^p &= softmax(w_n^p(P_{Att}^n) + b_n^p) \\ \widetilde{P_n} &= (\sum g_n^p \times P_{Att}^n) \end{aligned} \quad (2)$$

위 수식 (2)에서 w_n과 b_n는 각각 가중치 행렬과 바이어스를 의미한다. 주의 집중 풀링을 통해 최종적으로 생성된 고정 길이 벡터 $\widetilde{Q_n}$과 $\widetilde{P_n}$은 두 벡터간 유사도를 측정하는 네트워크에 입력되어, 질문에 대한 입력 단락의 점수를 생성하게 된다.

3.4. 유사도 측정 네트워크

제안 모델은 서론에서 언급했듯, 질문과 단락 사이의 공통 의미 정보가 질문과 유사할수록 질문에 적합한 단락이라고 가정했으며, 나아가 단락 재순위화 문제를 두 벡터 간의 유사도 측정 문제로 풀고자했다. 본 논문에서 사용하는 벡터 간 유사도 측정 네트워크는 [13]에서 제안한 방법으로 아래의 수식 (3)과 같이 계산된다.

$$\begin{aligned} sim(\widetilde{Q_n}, \widetilde{P_n}) &= [\widetilde{Q_n}, \widetilde{P_n}, |\widetilde{Q_n} - \widetilde{P_n}|, \widetilde{Q_n} \times \widetilde{P_n}] \\ sim(\widetilde{Q}, \widetilde{P}) &= w \cdot [sim(\widetilde{Q_1}, \widetilde{P_1}) \oplus sim(\widetilde{Q_2}, \widetilde{P_2}) \\ &\quad \oplus sim(\widetilde{Q_3}, \widetilde{P_3}) \oplus ranking\,feature] + b \end{aligned} \quad (3)$$

위 수식 (3)에서 $[\widetilde{Q}, \widetilde{P}, ...]$는 벡터 간의 연결을 의미하고, w와 b는 가중치 행렬과 바이어스를 의미한다. $\oplus$는 $sim(\widetilde{Q_n}, \widetilde{P_n})$ 벡터 간의 연결을 의미한다. 그리고 최종 유사도 점수 $sim(\widetilde{Q}, \widetilde{P})$ 생성에 사용되는 ranking feature는 입력 단락의 검색 모델 순위를 의미한다. 제안 모델의 학습에 사용된 손실 함수는 cross-entropy가 사용되었다.

4. 실험

4.1. 실험 준비

본 논문에서는 KorQuAD 1.0[14]의 학습 데이터와 검증 데이터를 통해 실험을 진행했다. 제안 모델을 실험하기 위해 KorQuAD에 존재하는 모든 단락을 모아 색인을 진행했으며, 이를 토대로 입력 질문에 대한 BM25의 결과 상위 10개 단락을 추출했다. 하나의 질문에 대해 10개의 질문-단락 쌍을 만들었으며 학습 데이터 603,862 문장 쌍, 평가 데이터 57,723 문장 쌍을 구축하여 실험을 진행했다. 매 epoch 마다 1:5 비율로 Negative Sampling을 진행했으며, 성능 평가 지표로는 MRR@10과 Precision@K를 사용했다.

4.2. 실험 결과 및 분석

표 2. 제안 모델 모듈 적용에 따른 성능 비교

	MRR@10	P@1	P@2	P@3
제안 모델	**93.0**	**89.4**	**95.0**	**96.6**
제안 모델 + w/o ranking feature	91.6	87.5	93.4	95.4
제안 모델 + w/o 구절 모델링 계층	91.8	87.3	94.5	96.1
제안 모델 + w/o 위치 정보 자질	91.2	86.9	93.1	95.0

위의 표 2는 제안 모델의 각 모듈 적용 유무에 따른 성능 변화를 보여준다. 위의 표에서 "+w/o" 기호는 제안 모델에서 제외시킨 모듈이나 자질을 의미한다. 그리고 위치 정보 자질은 중첩 위치 자질을 포함한 실험 결과를 의미한다. 본 논문에서 제안하는 자질이나 모듈이 성능 향상에 도움을 주는 것을 확인할 수 있다. 전체적인 검색 성능을 확인할 수 있는 지표인 MRR@10에 따르면 본 논문에서 사용한 위치 정보 자질이 검색 성능에 영향을 가장 크게 주는 것으로 확인이 되며, Precision@1을 기준으로는 위치 정보 자질이 가장 크게 영향을 주는 것을 알 수 있다. 아래의 표 3은 기존에 제안되었던 모델과의 성능 비교를 보여준다.

표 3. 기존 연구 모델과의 성능 비교

	MRR@10	P@1	P@2	P@3
BM25	91.1	86.7	93.0	95.5
장영진 외 2019[8]	91.2	86.3	93.5	95.5
장영진 외 2020[9]	91.8	88.5	94.0	95.3
제안 모델	**93.0**	**89.4**	**95.0**	**96.6**

표 3에서 BM 25는 학습 데이터 구축에 사용되었던 검색 모델의 성능을 의미한다. [8]은 포인터 네트워크를 이용하여 입력 받은 N개의 문서를 한 번에 재순위화하는 모델을 의미하고, [9]는 다중 작업 학습 기반의 단락 재순위화 모델을 의미한다. 실험 결과에 따르면 제안 모델이 가장 뛰어난 검색 성능을 보이는 것을 확인할 수 있다. BM25와 [8]의 성능 비교를 통해 baseline인 검색 모델의 성능이 충분히 높을 경우엔 재순위화의 성능이 미미할 수도 있다는 것을 알 수 있다. 이를 통해 심층 신경망을 사용하지 않은 검색 모델의 성능이 높은 경우에는 재순위화 모델을 사용하는 것이 오히려 속도 측면에서 비효율적일 수 있다는 점을 알 수 있다.

5. 결론

본 논문에서는 오픈 도메인 기계독해 시스템을 위한 단락 재순위화 모델을 제안했다. 구절 모델링 계층을 적용하여 주변 문맥의 의미정보를 잘 반영할 수 있었으며, 중복 위치 자질을 사용함으로써 어휘 정보도 효과적으로 반영 할 수 있었다. 실험 결과에 대해서 baseline이 되는 BM25의 성능이 상당히 높게 나오는 것을 확인할 수 있는데, KorQuAD라는 제한적인 단락 내에서 검색을 했기 때문에 검색 성능이 높게 나온 것으로 판단된다. 이를 통해 검색 모델 성능이 높은 경우에는, 추가적인 비용을 요구하는 심층 신경망 기반 재순위화 모델 적용에 대해 재고해봐야 한다고 판단된다. 향후 연구로는 검색 모델이 잘 작동하지 못하는 환경에서 제안 모델의 효과를 검증할 수 있는 실험을 진행할 예정이다. 그리고 기계독해 시스템에 적용하여, 정답 단락이 아닌 재순위화 후 최상위로 올라온 단락 얼마나 정답을 포함하고 있는지에 대한 실험을 진행할 예정이다.

감사의 글

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. 2020R1F1A1069737)

참고문헌

[1] J. Devlin, M. W. Chang, K. Lee, K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In NAACL-HLT. 2019.

[2] P. Yang, H. Fang, J. Lin, "Anserini: Reproducible ranking baselines using lucene. ACM Journal of Data and Information Quality", Vol. 10, No. 4, Article 16. 2018.

[3] T. M. Cover, "Elements of information theory", John Wiley & Sons. 1999

[4] J. Guo, Y. Fan, A. Qingyao,W. B. Croft, "A Deep Relevance Matching Model for Ad-hoc Retrieval", In Proceedings of the 25th ACM International on Conference on Information and Knowledge Management (CIKM). ACM, 2016.

[5] Z. Dai, C. Xiong, J. Callan, J. Liu, "Convolutional neural networks for soft-matching ngrams in ad-hoc search", In Proceedings of the eleventh ACM international conference on web search and data mining, pages 126-134, 2018.

[6] N. Kalchbrenner, E. Grefenstette, P. Blunsom, "A convolutional neural network for modelling sentences", Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL), pp. 655-665, 2014.

[7] 장영진, 김학수, 지혜성, 이충희, "'질문-단락'간 주의 집중을 이용한 검색모델 재순위화 방법", 제31회 한글 및 한국어 정보처리 학술대회, pp. 411-414, 2019.10.

[8] 장영진, 권오욱, 김학수, "정보 검색 기반 기계독해 시스템을 위한 단락 재순위화 모델", 2020 한국컴퓨터종합학술대회, pp. 410-412, 2020.07

[9] J. Pennington, R. Socher, C. Manning, "Glove: Global Vectors for Word Representation", Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532-1543, 2014.
[10] J. Chung, C. Gulcehre, K. Cho, Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling", in NIPS 2014 Workshop on Deep Learning, 2014.
[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, L. Polosukhin, "Attention Is All You Need", 31st Conference on Neural Information Processing Systems (NIPS 2017). 2017.
[12] X. Zhou, X. Wan, J. Xiao, "Attention-based lstm network for cross-lingual sentiment classification", Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 247-256, 2016.
[13] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, A. Bordes, "Supervised Learning of Universal Sentence Representations from Natural Language Inference Data", Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp 670-680, 2017.
[14] 임승영, 김명지, 이주열, "KorQuAD: 기계독해를 위한 한국어 질의응답 데이터셋", 한국정보과학회 학술발표논문집, pp. 539-541, 2018