지식 기반 다중 대화 시스템을 위한 주의 집중 지식 선택 모델

이도행○, 장영진○, 황금하◆, 권오욱◆, 김학수○
건국대학교 인공지능학과○, 한국전자통신연구원◆
dsdhlee@konkuk.ac.kr, danyon@konkuk.ac.kr, hgh@etri.re.kr, ohwoog@etri.re.kr, nlpdrkim@konkuk.ac.kr,

Attentive Knowledge Selection Model for Knowledge-Grounded Multi-turn Dialogue System

Dohaeng Lee○, Youngjin Jang○, Jin-Xia Huang◆, Oh-Woog Kwon◆, Harksoo Kim○
Konkuk University Department of Artificial Intelligence○
Electronics and Telecommunications Research Institute◆

요 약

지식 기반 다중 대화 시스템은 지식 정보를 포함한 응답을 생성하는 대화 시스템이다. 이 시스템은 응답 생성에 필요한 지식 정보를 찾아내는 지식 선택 작업과 찾아낸 지식 정보를 바탕으로 문맥을 고려한 응답을 생성하는 응답 생성 작업으로 구성된다. 본 논문에서는 지식 선택 작업을 기계독해 프레임워크에 적용하여 해결하는 방법을 제안한다. 지식 선택 작업은 여러 개의 발화로 이루어진 대화 기록을 바탕으로 지식 문서 내에 존재하는 지식을 찾아내는 작업이다. 본 논문에서는 대화 기록 모델링 계층을 활용해 마지막 발화와 관련 있는 대화 기록을 찾아내고, 주의 집중 풀링 계층을 활용해 긴 길이의 지식을 효과적으로 추출하는 방법을 제안한다. 실험 결과, 목적지향 지식 문서 기반 대화 데이터 셋인 Doc2dial 데이터의 지식 선택 작업에서 F1 점수 기준 76.52%, EM 점수 기준 66.21%의 성능을 기록해 비교 모델 보다 높은 성능을 기록하는 것을 확인할 수 있었다.

주제어: 지식 기반 다중 대화 시스템, 지식 선택, 기계독해

1. 서론

지식 기반 다중 대화 시스템(Knowledge-grounded multi-turn dialogue system)은 대화 문맥과 주어진 지식 문서를 바탕으로 대화에 이어지는 응답을 생성하는 시스템이다. 이 시스템은 응답 생성에 필요한 지식 정보를 찾아내는 지식 선택(Knowledge selection) 작업과 찾아낸 지식 정보를 바탕으로 문맥을 고려한 응답을 생성하는 응답 생성(Response generation) 작업으로 구성된다[1]. 아래의 그림 1은 지식 기반 다중 대화 시스템의 구조도이다. 그림 1과 같이 지식 기반 다중 대화 시스템은 대화 문맥에 맞는 지식 선택 작업이 선행되어야 하며, 지식 선택 작업이 제대로 이루어지지 않을 경우, 응답에 필요한 적절한 지식을 반영하지 못하는 문제가 생길 수 있다. 이처럼, 지식 선택 작업은 지식 기반 다중 대화 시스템의 효용성을 결정하는 매우 중요한 작업이라고 할 수 있다. 때문에 지식 기반 다중 대화 시스템 분야의 최근 연구는 응답 생성 연구 뿐만 아니라 정확한 지식을 반영하기 위해 필요한 지식 선택 작업에 대한 많은 연구들도 함께 진행되고 있다[1-6].

지식 선택 작업의 이전 연구는 긴 길이의 대화 기록을 첫 발화부터 마지막 발화까지 모두 사용하거나, 연구자가 정한 규칙(Rule)에 따라 일부 발화만을 사용하였다[2]. 긴 길이의 대화 기록을 그대로 사용하는 경우 응답 생성에 필요한 지식을 찾는데 노이즈(Noise)로 작용하는 대화 문맥이 포함될 수 있는 문제점이 있으며, 대화 기록의 일부만을 사용하는 경우 정답 지식을 찾는데 도움이 되는 대화 문맥을 포함하지 못하는 문제점이 있다.

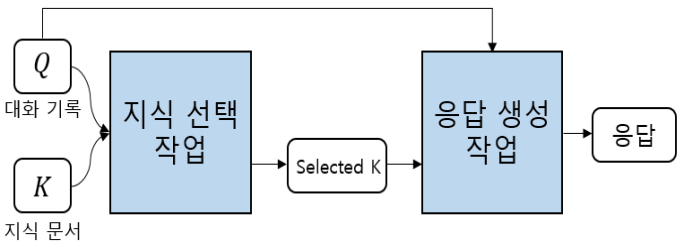

그림 1. 지식 기반 다중 대화 시스템 구조도

따라서 본 논문에서는, 정확한 지식 선택을 위해 현재 질문에 해당하는 마지막 발화와 관련 있는 대화 기록을 찾아내고, 정답 지식을 찾는데 필요한 대화 문맥 정보를 반영하여 높은 성능을 보여줄 수 있는 모델을 제안한다.

2. 관련 연구

최근, 지식 기반 다중 대화 시스템에 대한 연구자들의 많은 관심과 함께, 지식 기반 다중 대화 시스템 구축을 목표로 한 데이터와 새로운 모델들이 등장하고 있다. 지식 기반 다중 대화 데이터 Wizard of Wikipedia 데이터를 공개한 [1]은 지식 문서 기반 대화 시스템을 구축하기 위해 지식 선택 단계와 응답 생성 단계를 각각 진행한 two-stage 구조의 모델과 응답 생성 과정에서 지식

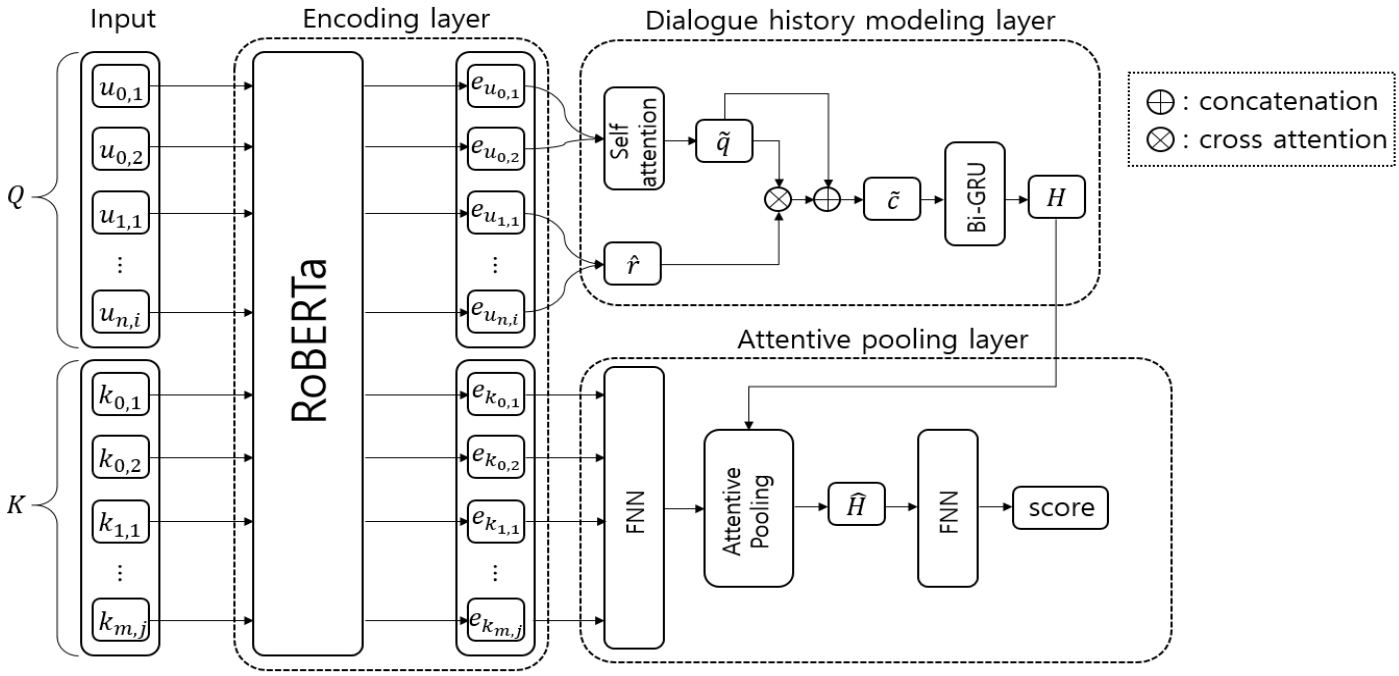

그림 2. 제안 모델 구조도

선택 단계를 함께 수행하는 end-to-end 방식의 모델을 제안했다. 목적 지향 지식 기반 다중 대화 데이터 Doc2dial 데이터 셋을 공개한 [2]는 지식 선택 과업을 주어진 대화 문맥과 문서를 활용해 질문에 대한 답변을 추출하는 스팬 선택(Span selection) 문제로 간주하고, 기계독해 프레임워크[7]를 활용해 해결하고자 하였다.

[1]에서 보고된 바에 따르면, 지식 선택 모델이 응답에 필요한 정확한 지식을 찾아내는 것이 응답 생성에서의 성능을 결정하는 가장 중요한 작업임을 확인할 수 있다. 따라서, 본 논문에서는 지식 문서 기반 대화 시스템을 [1]에서 제안한 two-stage 구조로 간주하고, [2]와 같이 기계독해 프레임워크를 지식 선택 작업에 적용한 모델을 제안하고자 한다.

3. 제안 모델

위의 그림 2는 제안 모델의 전체 구조도이다. 모델의 입력은 $\{Q,K\}$로 표현할 수 있으며, Q와 K는 각각 대화 기록과 지식 문서를 의미한다. 대화 기록 Q는 $Q=\{u_{0,1},u_{0,2},\dots,u_{n,i}\}$로 표현한다. 여기서 n은 대화 기록을 구성하는 발화 수이고, i는 n번째 발화를 구성하는 토큰 수이다. 지식 문서 K는 $K=\{k_{0,1},k_{0,2},\dots,k_{m,j}\}$ 로 표현한다. 여기서 m은 지식 문서를 구성하는 문장의 수이고, j는 m번째 문장을 구성하는 토큰 수이다. 제안 모델은 인코딩 계층(Encoding layer), 대화 기록 모델링 계층(Dialogue history modeling layer), 주의 집중 풀링 계층(Attentive pooling layer)으로 이루어진다. 인코딩 계층은 사전 학습된 언어 모델을 사용해 입력받은 $\{Q,K\}$를 인코딩한다. 대화 기록 모델링 계층은 대화 기록에서 마지막 발화와 나머지 발화 사이의 관계를 계산해 정답 지식을 찾는데 필요한 대화 기록을 압축한다. 주의 집중 풀링 계층에서는 대화 기록 모델링 계층을 통해 만들어진 대화 기록 벡터와 지식 문서 사이의 주의 집중 풀링 연산을 수행하고, 로그 음의 우도(Negative log-likelihood) 손실 함수를 통해 문장 단위의 지식 문서 벡터와 정답 지식 문장과의 손실을 계산한다.

3.1 인코딩 계층

인코딩 계층의 인코더는 사전 학습된 언어 모델 RoBERTa[8]를 사용하였다. 인코딩 계층은 다음과 같이 $[CLS]Q[SEP]K[SEP]$ 형태로 입력받는다. 이때, 입력되는 대화 기록의 순서는 마지막 발화를 제일 앞에 배치하고, 그 이후 발화들을 순차적으로 배치하였다. 인코딩된 대화 기록과 지식 문서는 각각 $\hat{Q}=\{e_{u_0,1},e_{u_0,2},\dots,e_{u_m,i}\}$, $\hat{K}=\{e_{k_0,1},e_{k_0,2},\dots,e_{k_m,j}\}$로 표현한다.

3.2 대화 기록 모델링 계층

대화 기록 모델링 계층은 인코딩된 대화 기록 벡터 $\hat{Q}$에서 정답 지식을 찾는데 중요한 정보를 가지고 있는 마지막 발화와 관련된 대화 기록에 더 높은 가중치를 부여하기 위해 설계되었다.

본 논문에서는, 마지막 발화를 $\hat{q}=\{e_{u_0,1},e_{u_0,2},\dots,e_{u_0,q_i}\}$로 표현한다. 여기서 q_i는 마지막 발화의 토큰 수이다. 마지막 발화를 제외한 나머지 대화 기록 벡터는 $\hat{r}=\{e_{u_1,1},e_{u_1,2},\dots,e_{u_n,r_i}\}$로 표현한다. 여기서 r_i는 대화 기록 벡터의 토큰 수이다.

먼저 마지막 발화 $\hat{q}$를 구성하는 단어들 $\{e_{u_0,1},e_{u_0,2},\dots,e_{u_0,q_i}\}$ 은 마지막 발화 정보를 강하게 반영하기 위해 수식 (1)과 같이 Multi-Head Attention[9]을 이용한 자가 주의 집중 연산(Self attention)을 수행한다.

$$\tilde{q} = MHAttention(\hat{q}, \hat{q}, \hat{q}) \quad (1)$$

식 (1)을 통해 연산된 마지막 발화는 $\tilde{q} \in \mathbb{R}^{q_i \times d}$로 표현된다. 여기서 d는 사전 정의된 hidden_size의 차원 수를 의미한다. 이후, 마지막 발화 벡터 $\hat{q}$와 관련된 대화 기록을 반영하기 위해 $\tilde{q}$와 나머지 대화 기록 벡터 $\hat{r}$사이의 Multi-Head Attention 연산을 수행하며, 그 식은 (2)와 같다.

$$\tilde{r} = MHAttention(\hat{r}, \tilde{q}, \tilde{q}) \quad (2)$$

식 (2)를 통해 연산된 대화 기록은 $\tilde{r} \in \mathbb{R}^{r_i \times d}$로 표현된다. 이후, 식 (1)과 식 (2)에 의해 계산된 두 벡터를 다음과 같이 연결해 준다.

$$\tilde{c} = [\tilde{q}; \tilde{r}] \quad (3)$$

식 (3)에서 (;)은 concatenation 연산을 의미한다. 연결된 대화 기록 벡터 $\tilde{c}$는 $\tilde{c} \in \mathbb{R}^{(q_i + r_i) \times d}$로 표현되며, $(q_i + r_i)$는 $\tilde{c}$를 구성하는 토큰의 수이다. 이후, $\tilde{c}$는 양방향 GRU[10] 계층 연산을 수행하며, 그 식은 (4)와 같다.

$$\begin{aligned} (h_{\tilde{c},1}, ..., h_{\tilde{c},|\tilde{c}|}) &= BiGRU(\tilde{c}) \\ h_{\tilde{c},i} &= [h^f_{\tilde{c},i}; h^b_{\tilde{c},i}] \\ H &= [h^f_{\tilde{c},|\tilde{c}|}; h^b_{\tilde{c},1}] \end{aligned} \quad (4)$$

식 (4)를 통해 연산된 H벡터는 $H \in \mathbb{R}^d$로 표현되며, 주의 집중 풀링 연산을 위한 질의(Query)로 사용된다.

3.3 주의 집중 풀링 계층

주의 집중 풀링 계층은 질의 벡터 H와 지식 문서 각 문장의 토큰 벡터 $\hat{K}_m = \{e_{k_m,1}, e_{k_m,2}, \ldots, e_{k_m,j}\} \in \mathbb{R}^{j \times d}$과의 주의 집중 가중합 연산을 수행하며, 그 식은 (5)와 같다.

$$\begin{aligned} \epsilon^m &= \hat{K}_m \cdot H^T \\ a^m_k &= \frac{\exp(\epsilon^m_k)}{\sum^j \exp(\epsilon^m_j)} \\ \hat{H}_m &= \sum^j a^m_k e_{k_m,k} \end{aligned} \quad (5)$$

식 (5)를 통해 연산된 $\hat{H}$ 벡터는 $\hat{H} \in \mathbb{R}^{m \times d}$로 표현된다. 이후, $\hat{H}$ 벡터는 전방 전달 신경망(Feed-forward neural network)을 통과해 예측 지식 문장의 시작과 끝일 확률을 출력한다. 그 식은 (6)과 같다.

$$\begin{aligned} P_s &= Softmax(W_s \hat{H} + b_s) \\ P_e &= Softmax(W_e \hat{H} + b_e) \end{aligned} \quad (6)$$

식 (6)을 통해 계산된 P_s, P_e에 대한 각 확률 분포는 정답 지식 문장과 각각 로그 음의 우도(Negative log-likelihood) 손실 값을 계산하고, 평가 단계에서는 P_s, P_e의 각 확률 값이 가장 큰 문장 범위를 최종 출력으로 사용한다.

표 1. 실험 결과

Model	F1	EM
Baseline		
+ Last2	74.75	60.24
+ All	55.27	41.18
+ Last2-r	74.30	60.09
+ All-r	74.93	61.73
제안 모델	76.52	66.21

4. 실험

4.1 실험 준비

본 논문에서 사용한 데이터 셋은 목적 지향 지식 기반 다중 대화 데이터 셋인 Doc2dial 데이터이다. Doc2dial 데이터 셋은 총 4개의 도메인으로 이루어져 있으며 총 4,135개의 대화, 487개의 지식 문서로 구성되어 있다. 총 4,135개의 대화는 학습 데이터에 3,474개, 평가 데이터에 661개로 나누어져 있다. 본 논문에서의 실험은 학습 데이터로 학습 후 평가 데이터로 모델의 성능 평가를 진행하였다. 모델의 성능 평가는 F1 점수와 EM 점수를 사용하였다.

실험에서 사용한 비교 모델은 다음과 같다. 본 논문에서는, 공정한 비교를 위해 [2]에서 보고된 BERT-base를 사전 학습 언어 모델로 사용한 대화 기록 입력 형식에 따른 모델들에 대해서, 사전 학습 언어 모델만 RoBERT-large로 바꿔 재학습을 진행하고, 평가하였다. 표 1의 Last2는 대화 기록 중 마지막 2개 발화를 시간 순으로 배치한 모델을 의미한다. All은 모든 대화 기록을 시간 순으로 배치한 모델을 의미한다. r은 대화 기록을 역순으로 배치한 모델을 의미한다.

4.2 실험 결과

표 1을 보면, Baseline 모델 중 모든 대화 기록을 역순으로 배치한 All-r이 F1 점수 기준 74.93%, EM 점수 기준 61.73%로 가장 높은 성능을 기록한 것을 확인할 수 있다. All의 성능을 보면, 나머지 3개의 Baseline 모델과 비교했을 때 F1 점수 기준 약 20% 가까이 낮은 것을 확인할 수 있다. 이는 대화 기록의 마지막 발화가 사전 학습 언어 모델의 제한된 입력 길이로 인해 포함되지 못해 생기는 현상이다. 즉, 본 논문에서 가정했던 대로 마지막 발화가 정답 지식을 찾는데 가장 중요한 정보를 담고 있다는 것을 보여준다.

제안 모델은 F1 점수 기준 76.52%, EM 점수 기준 66.21%로 비교 모델 중 가장 높은 성능을 보이는 All-r보다 F1 점수에서 1.59%, EM 점수에서 4.48% 더 높은 성능을 기록하였다.

5. 결론

본 논문은 지식 기반 다중 대화 시스템의 지식 선택 작업을 해결하기 위해 기계독해 프레임워크를 적용한 모델을 제안하였다. 제안 모델은 대화 기록 모델링 계층을 통해 정답 지식을 찾는데 필요한 대화 기록은 강조하고, 노이즈로 작용할 수 있는 대화 기록은 약하게 반영하는 방법을 효과적으로 적용하였다. 또한, 긴 문장 단위의 정답 지식 문장을 효율적으로 추출하기 위해 주의 집중 풀링 계층을 사용하였다. 실험 결과를 통해 제안 모델의 방법이 가장 높은 성능을 보이는 것을 확인할 수 있었다.

향후 연구로, 다양한 대화 기록 모델링 논문을 참고해 기존 제안 모델보다 대화 기록 정보를 효율적으로 추출하는 방법을 연구할 예정이다.

감사의 글

이 논문은 2019년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (2019-0-00004, 준지도학습형 언어지능 원천기술 및 이에 기반한 외국인 지원용 한국어 튜터링 서비스 개발)

참고문헌

[1] Dinan, E., Roller, S., Shuster, K., Fan, A., Auli, M., & Weston, J. "Wizard of wikipedia: knowledge-powered conversational agents." arXiv preprint arXiv:1811.01241, 2018.

[2] Feng, S., Fadnis, K., Liao, Q. V., & Lastras, L. A. "Doc2dial: a framework for dialogue composition grounded in documents." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 34. No. 09. 2020.

[3] Lian, R., Xie, M., Wang, F., Peng, J., & Wu, H. "Learning to select knowledge for response generation in dialog systems." arXiv preprint arXiv:1902.04911. 2019.

[4] Kim, B., Ahn, J., & Kim, G. "Sequential latent knowledge selection for knowledge-grounded dialogue." arXiv preprint arXiv:2002.07510, 2020.

[5] Zheng, C., Cao, Y., Jiang, D., & Huang, M. "Difference-aware knowledge selection for knowledge-grounded conversation generation." arXiv preprint arXiv:2009.09378. 2020.

[6] Kim, B., Lee, D., Lee, Y., Kim, H., Kim, S., Huang, J., & Kwon, O., "Document-grounded goal-oriented dialogue systems on pre-trained language model with diverse input representation." In Proceedings of the 1st Workshop on Document-grounded Dialogue and Conversational Question Answering. Association for Computational Linguistics. 2021.

[7] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. "Bert: pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805. 2018.

[8] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. "Roberta: A robustly optimized bert pretraining approach." arXiv preprint arXiv:1907.11692. 2019.

[9] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. "Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008). 2017.

[10] Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. "Learning phrase representations using RNN encoder-decoder for statistical machine translation." arXiv preprint arXiv:1406.1078, 2014.