

OK-KGD:오픈 도메인 한국어 지식 기반 대화 데이터셋 구축

문선아^o, 김산, 장진예, 정민영, 신사임
한국전자기술연구원, 인공지능연구센터

{seon43, kimsan0622, jinyea.jang, minyoung.jung, sishin}@keti.re.kr

OK-KGD:Open-domain Korean Knowledge Grounded Dialogue Dataset

Seona Moon^o, San Kim, Jinyea Jang, Minyoung Jeung, Saim Shin
Korea Electronics Technology Institute, Artificial Intelligence Research Center

요약

최근 자연어처리 연구 중 오픈 도메인 지식 기반 대화는 많은 관심을 받고 있다. 연구를 위해서는 오픈 도메인 환경을 갖추고 적절한 지식을 사용한 대화 데이터셋이 필요하다. 지금까지 오픈 도메인 환경을 갖춘 한국어 지식 기반 대화 데이터셋은 존재하지 않아 한국어가 아닌 데이터셋을 한국어로 기계번역하여 연구에 사용하였다. 이를 사용할 경우 두 가지 단점이 존재한다. 먼저 사용된 지식이 한국 문화에 익숙하지 않아 한국인이 쉽게 알 수 없는 대화 내용이 담겨있다. 그리고 번역체가 남아있어 대화가 자연스럽지 않다. 그래서 본 논문에서는 자연스러운 대화체와 대화 내용을 담기 위해 새로운 오픈 도메인 한국어 지식 기반 대화 데이터셋을 구축하였다. 오픈 도메인 환경 구축을 위해 위키백과와 나무위키의 지식을 사용하였고 사용자와 시스템의 발화로 이루어진 1,773개의 대화 세트를 구축하였다. 시스템 발화는 크게 지식을 사용한 발화, 사용자 질문에 대한 답을 주지 못한 발화, 그리고 지식이 포함되지 않은 발화 3가지로 구성된다. 이렇게 구축한 데이터셋을 통해 KE-T5와 Long-KE-T5를 사용하여 간단한 실험을 진행하였다.

주제어: 한국어 지식 기반 대화, 오픈 도메인 지식 기반 대화 데이터셋 구축

1. 서론

오픈 도메인 지식 기반 대화는 최근 자연어처리 연구에서 많은 관심을 받고 있다. 연구를 위해 사용되는 대표적인 데이터셋으로 WoW [1], Wizint [2] 등이 있지만 대화와 지식 모두 영어로 이루어져 있다. 이 데이터셋들은 영어로 이루어진 위키백과¹를 오픈 도메인 환경으로 구성하고 대화 데이터셋을 만들었다. 한국어로 이루어진 데이터셋은 간단한 질의응답 형식이거나 오픈 도메인 환경이 구성되지 않은 지식 기반 대화로 오픈 도메인 한국어 지식 기반 대화를 연구하기에 적합하지 않다.

지금까지 오픈 도메인 한국어 지식 기반 대화 연구를 위해서 영어 데이터셋을 기계번역하여 사용하였다. 하지만 이는 두 가지 문제점이 존재한다. 첫 번째로 대화에 사용된 지식은 해당 문화권에 맞춰진 지식으로 한국인이 쉽게 이해하지 못하거나 알지 못하는 내용이 있을 수 있다. 그리고 기계 번역된 데이터셋은 번역체가 남아 있어 자연스러운 대화체로 보이지 않는다. 본 논문에서는 이러한 문제점을 보완하기 위해 한국어로 이루어진 위키백과²와 나무위키³를 기반으로 오픈 도메인 환경을 구성한 뒤 대화를 진행하는 오픈 도메인 한국어 지식 기반 대화 데이터셋을 구축하였다.

데이터셋은 사용자와 시스템의 발화로 이루어져 있고 주어진

The screenshot shows a chat window with the following content:

- 사용자 (User):** 안녕, 영국에 대해 궁금하게 생겼는데 영국의 영어표기에는 어떤 것들이 있어?
- 시스템 (System):** 지식: 영어로는 흔히 United Kingdom, UK, Britain으로 줄여 부른다. 올림픽 등의 스포츠에서는 다소 부정확하지만 관행상 Great Britain이라고 부른다. 영어로는 흔히 United Kingdom, UK, Britain으로 줄여서 불러요. 올림픽 등 스포츠에서는 다소 부정확하지만 관행상 Great Britain이라고 부르고 있어요.
- 사용자 (User):** 영국에는 해가 지지 않는 나라라는 별명이 있었던데, 왜그런지 설명해 줄 수 있어?
- 시스템 (System):** 지식: 한때는 해가 지지 않는 나라라는 별명이 있었. 19세기~20세기에 대영제국의 영토가 절정에 달했을 때 언제나 제국의 어딘가는 낮이었기 때문이다. 19세기~20세기에 대영제국의 영토가 절정에 달했을 때 언제나 제국의 어딘가는 낮이었기 때문이에요.
- 사용자 (User):** 그렇구나, 영국의 근대 역사에 대해서 간단하게 설명해줄래?
- 시스템 (System):** 지식: failed to search knowledge. 제가 찾아드릴 수 없는 정보예요.
- 사용자 (User):** 영국에 대해 잘 알게 된 것 같아. 알려줘서 고마워.
- 시스템 (System):** 지식: no knowledge used. 도움이 됐다니 기뻐요.

그림 1. 오픈 도메인 한국어 지식 기반 대화 데이터셋 예시

주제에 대해 사용자가 궁금한 점을 물어보고 시스템이 답변을 하면서 대화가 진행된다. 이때, 시스템은 크게 3가지 종류의 발화가 가능하다. 먼저 사용자 질문에 대한 지식이 포함된 답변이

¹<https://en.wikipedia.org>

²<https://ko.wikipedia.org>

³<https://namu.wiki>

가능하고 질문에 대한 지식을 검색하지 못해 지식이 포함되지 않은 답변이 가능하고 마지막으로 지식이 필요하지 않아 지식을 포함하지 않은 답변이 가능하다. 지식이 필요하지 않아 지식을 포함하지 않은 답변들은 대부분 대화의 처음과 마지막에 사용자와 간단한 인사를 나누는 발화로 구성되어 있다. 그림 1은 구축된 데이터셋의 샘플 대화를 보여준다. 이렇게 구성된 데이터셋을 학습하기 위해 한국어 거대 모델 KE-T5 [3]와 허깅페이스에 공개된 Long-KE-T5⁴를 사용하여 간단한 실험을 진행하였다. 실험은 크게 두 단계로 시스템 발화에 필요한 지식을 검색하는 지식 검색 모델과 이어질 발화를 생성하는 응답 생성 모델로 이루어진다.

2. 관련 연구

2.1 오픈 도메인 지식 기반 대화 데이터셋

오픈 도메인 지식 기반 대화 연구를 위한 다양한 데이터셋이 존재한다. 먼저 영어 데이터셋으로는 WoW [1], Wizint [2]가 존재한다. 이는 오픈 도메인 환경 구성을 위해 영어로 이루어진 위키백과의 요약(abstract)를 사용하고 사용자와 시스템의 연속적인 발화로 이루어져 있다. 한국어 데이터셋으로는 WoW를 한국어로 번역한 KoWoW [4]가 존재한다. 이는 대화만 번역한 것이 아니라 지식도 모두 번역하였다. 그 외에는 지식 기반 질의응답인 AiHub⁵의 일반 상식 데이터셋이 존재하지만 이는 위키백과의 일부 문서만 사용하여 오픈 도메인 환경을 구축하지 못했고 발화가 이어지는 대화가 아닌 간단한 질의응답만 가능한 데이터셋이다. 그리고 AiHub와 모두의 말뭉치⁶에서 공개한 온라인 구어체 말뭉치 데이터, 일상 대화 말뭉치 등 다양한 형태의 대화 데이터셋이 존재하지만 지식을 기반으로 하는 대화 데이터셋은 존재하지 않는다.

2.2 오픈 도메인 지식 기반 대화 연구

오픈 도메인 지식 기반 대화를 위해서는 크게 두 단계를 거쳐 연구가 진행되고 있다. 먼저 발화에 필요한 지식을 검색하는 지식 검색 모델이 있다. 이는 전통적인 문서 검색 방법으로 TF-IDF [5]와 BM25 [6]가 있고 대표적인 밀집 벡터 기반 문서 검색 방법으로 Transformer [7]의 인코더를 사용하여 대화 내용과 지식 후보 간의 주의 집중 점수(attention score)를 구하여 학습하는 방법이 있다. 지식 후보는 이어질 발화에 필요한 정답 지식과 정답 지식을 제외한 부정 지식이 존재한다. 이 방법은 WoW, BlenderBot1 [8]에서 사용되었다.

그리고 사용자 발화에 이어질 시스템 발화를 생성하는 응답 생성 모델이 있다. 응답 생성 모델은 주로 이전 발화들과 응답

표 1. 오픈 도메인 환경을 위한 지식 통계 정보

	위키백과	나무위키	총합
문서 수	589,882	866,028	1,455,910
문단 수	3,093,213	26,202,822	29,296,035

표 2. 오픈 도메인 한국어 지식 기반 대화 데이터셋 통계 정보

	학습	검증	평가	총합
발화 수	16,031	1,972	1,993	19,996
대화 세트 수	1,418	177	178	1,773
대화 주제 수	103	82	89	103
평균 대화 턴 수	5.628	5.672	5.590	5.629

생성에 필요한 정답 지식을 입력으로 하여 생성한다. 본 논문에서는 언어 모델로 영어와 한국어 생성이 모두 가능한 KE-T5 [3]와 Long-KE-T5를 사용한다.

3. 오픈 도메인 한국어 지식 기반 대화 데이터셋 구축

3.1 오픈 도메인 환경 구축

한국인이 쉽게 접근 가능한 지식으로 오픈 도메인 환경을 구축하기 위해 두 위키 데이터를 사용하였다. 위키백과와 나무위키는 인터넷 사용자가 접근하기 쉽고 담긴 내용을 직접 수정할 수 있다. 문서 단위의 데이터를 전처리한 후 문단 단위로 나누었다. 약 145만개의 문서로 오픈 도메인 환경을 구축하였고 표 1에서 정확한 통계 정보를 볼 수 있다.

3.2 대화 데이터셋 구축

대화 데이터셋 구축을 위한 첫 번째 단계로 대화 주제를 선정하였다. 선정을 위해서 WoW 데이터셋에 사용된 대화 주제를 모두 한국어로 번역한 뒤 수정을 거쳐 선정하였다. 수정은 크게 두 가지 방향으로 진행하였다. 먼저 영어권에서는 익숙하지만 한국에서는 익숙하지 않은 단어를 바꾸어 선정하였다. 예를 들어, 미국의 대표적인 할인점인 ‘월마트’는 한국에서 이미 철수한 상태로 보다 익숙한 ‘대형마트’로 바꾸고 미국의 유명한 가수인 ‘브루노 마스’를 한국의 남녀노소가 아는 가수인 ‘방탄소년단’으로 바꾸는 등의 작업을 진행하였다. 그리고 대화 생성의 자유도를 높이기 위하여 특정 범위에 국한된 대화 주제를 넓은 범위로 바꾸어 선정하였다. 예를 들어, ‘뉴욕, 런던’과 같은 도시 이름을 각각 국가 이름인 ‘미국, 영국’으로 변경하였다. 그리고 ‘채식주의’, ‘채식주의의 역사’와 같이 ‘채식’이라는 큰 범위 안에 속한 주제를 하나로 묶어주었다.

이렇게 선정된 103개의 대화 주제를 가지고 2명의 크라우드 워커마다 각각 사용자와 시스템의 역할을 부여하여 대화 데이터를 수집하였다. 만들어진 대화 데이터셋은 총 1,773개의 대화

⁴<https://huggingface.co/KETI-AIR/long-ke-t5-base>

⁵<https://aihub.or.kr/>

⁶<https://corpus.korean.go.kr/>

세트로 자세한 통계 정보는 표 2를 통해 볼 수 있다. 표 2의 학습, 검증, 평가 데이터는 전체 대화 세트를 임의로 8:1:1 비율로 나누어 구성하였다. 여기서 평균 대화 턴 수는 사용자와 시스템이 대화를 주고 받은 횟수를 측정한 것으로 한 대화 세트 당 약 평균 5.6턴으로 이루어져 있고 학습 데이터에 모든 주제가 포함된 것을 확인할 수 있다.

3.3 데이터셋 특징

오픈 도메인 한국어 지식 기반 대화 데이터셋을 구축하면서 시스템 발화는 크게 3가지로 분류된다. 첫 번째로는 직전 발화에서 사용자가 요구한 지식이 담긴 발화이고 두 번째로는 직전 발화에서 사용자가 지식을 요구했지만 오픈 도메인 환경에서 적절한 지식을 찾지 못해 지식을 담지 못한 발화이다. 이는 그림 1에서 "제가 찾아드릴 수 없는 정보예요."에 해당된다. 이 발화 외에도 "제가 찾을 수 없는 정보예요.", "제가 찾아드릴 수 없는 정보예요.", "제가 답변해드릴 수 없는 정보예요."와 같이 총 4가지 형태로 등장한다. 마지막으로 직전 발화에서 사용자가 지식을 요구하지 않고 시스템도 지식을 사용하지 않은 발화이다. 이는 그림 1에서 "도움이 됐다니 기뻐요."에 해당된다. 이러한 발화는 대부분 사용자와 시스템이 대화의 시작과 마지막에서 인사말을 나누는 과정에서 나타난다. 데이터셋의 시스템의 발화에서 각 분류가 차지하는 비율은 소개한 순서대로 79.75%, 2.39%, 17.86%를 차지한다.

4. 실험 및 분석

4.1 실험

지식 검색 모델은 배치 내 다른 발화의 정답 지식을 부정 지식 후보로 사용하여 학습하였다. 그리고 지식 검색 모델의 성능을 구하기 위해 재현율(Recall) [9]을 사용한다. 이는 학습된 모델의 입력으로 이전 대화 기록 중 최대 n 개의 발화를 입력으로 사용하고 faiss [10]를 사용하여 유사도가 높은 k 개의 문단을 뽑는다. 그리고 유사도가 높은 k 개의 문단이 속하는 문서 중 정답 지식이 포함된 문서가 있는지 측정한다. 이는 변화하는 k 에 따라 $R@k$ 로 표기한다. 표 3에서 지식 검색 모델의 성능을 확인할 수 있는데 본 논문이 소개한 데이터셋은 103개의 대화 주제에 대한 내용만 담고 있어 오픈 도메인 검색 환경에서 검색 성능이 낮은 것을 확인하였다. 대신 데이터셋에 사용된 문서 내에서 검색한 결과인 클로즈드 도메인(Closed-domain) 검색 환경에서는 성능이 올라가는 것을 볼 수 있었다. 그리고 두 환경 모두 KE-T5와 Long-KE-T5의 성능이 비슷한 것을 확인하였다.

응답 생성 모델은 정답 지식과 이전 대화 기록 중 최대 n 개의 발화를 입력으로 사용하여 학습하였다. 그리고 성능 평가를 위해 F1 점수와 ROUGE 점수를 사용한다. 생성된 발화와 실제 응답 발화 간의 점수를 계산했을 때 표 4와 같은 결과를 얻었다.

표 3. 지식 검색 모델 성능(Recall)

검색 환경	모델	R@1	R@20	R@100
오픈 도메인	KE-T5	0.000	0.759	3.739
	Long-KE-T5	0.000	0.759	4.425
클로즈드 도메인	KE-T5	0.506	16.056	43.995
	Long-KE-T5	0.885	15.424	47.914

표 4. 응답 생성 모델 성능

모델	F1	Rouge-1	Rouge-2	Rouge-L
KE-T5	5.390	0.010	0.022	0.082
Long-KE-T5	73.807	26.690	17.277	26.690

KE-T5와 Long-KE-T5의 성능이 비슷했던 지식 검색 모델과 달리 응답 생성 모델은 Long-KE-T5의 성능이 더 뛰어난 것을 확인할 수 있다. 응답 생성모델에서 KE-T5와 Long-KE-T5의 성능차이는 사전학습 모델의 성능 차이에서 비롯된 것으로 분석된다.

4.2 실험 환경

실험을 위한 환경을 다음과 같다. 먼저 실험을 위한 gpu는 nvidia a100 8개를 사용하였다. 지식 검색 모델과 응답 생성 모델에서 이전 대화 기록 중 입력으로 최대 5개의 발화를 사용하였고 지식 검색 모델의 k 는 {1,5,20,100}을 사용하여 성능을 측정하였다. 지식 검색 모델은 KE-T5와 Long-KE-T5의 인코더를 기반으로 하고 Base모델을 사용하였다. 학습률(Learning rate)는 0.0003, 가중치 감쇠(Weight decay)는 0.0001인 AdamW 최적화 방법(Optimizer)를 사용하였으며, 배치(Batch) 크기는 16이고 반복 횟수(Epoch)은 100이다.

응답 생성 모델은 KE-T5와 Long-KE-T5의 조건부 생성 모델을 기반으로 하고 Base모델을 사용하였다. 학습률(Learning rate)는 0.0001, 가중치 감쇠(Weight decay)는 0.0001인 AdamW 최적화 방법(Optimizer)를 사용하였으며, 배치(Batch) 크기는 8이고 반복 횟수(Epoch)은 100이다.

5. 결론

본 논문에서는 새로 구축한 오픈 도메인 한국어 지식 기반 대화 데이터셋을 소개하였다. 오픈 도메인 지식 기반 연구를 위한 데이터셋은 주로 영어로 구성되어 있어 이를 한국어로 연구를 진행하기가 어려움이 있었다. 이를 해결 하기 위해 위키백과와 나무위키를 사용하여 오픈 도메인 환경을 구성하고 103개의 대화 주제를 가지고 평균 5.6턴의 발화와 1,773개의 대화 세트로 이루어진 대화 데이터셋을 구축하였다.

향후 연구로 오픈 도메인 데이터셋을 추가로 구축하고 지식

검색 모델 성능을 올리기 위한 연구를 진행할 예정이다. 데이터셋은 새로운 대화 주제를 가지고 구성할 뿐만 아니라 대화 사이에 지연 시간이 발생하는 멀티 세션 환경을 구성하여 다채로운 오픈 도메인 한국어 지식 기반 대화 데이터셋을 구축할 예정이다. 그리고 지식 검색 모델은 고난도 부정 지식 후보를 추가하여 학습에 활용하거나 검색된 문서의 순위를 재구성하는 방법을 사용하여 성능을 올릴 예정이다. 또한, 세 가지 종류의 발화가 등장하는 시스템 발화에 대하여 이를 먼저 분류하는 분류기를 추가하여 성능을 올릴 예정이다.

감사의 글

이 논문은 2023년도 정부 (과학기술정보통신부)의 재원으로 정보통신기획평가원(No. 2022-0-00320)의 지원을 받아 수행된 연구임

참고문헌

- [1] E. Dinan, S. Roller, K. Shuster, A. Fan, M. Auli, and J. Weston, “Wizard of Wikipedia: Knowledge-Powered Conversational Agents,” *Proceedings of the 7th International Conference on Learning Representations*, 2019.
- [2] M. Komeili, K. Shuster, and J. Weston, “Internet-Augmented Dialogue Generation,” *Proceedings of the 60th Association for Computational Linguistics*, pp. 8460–8478, 2022.
- [3] 신사임, 김산, and 서현태, “KE-T5: 한국어-영어 대용량 텍스트를 활용한 이중언어 사전학습기반 대형 언어모델 구축,” *제33회 한글 및 한국어 정보처리 학술발표 논문집*, pp. 419–422, 2021.
- [4] S. Kim, J. Y. Jang, M. Jung, and S. Shin, “A Model of Cross-Lingual Knowledge-Grounded Response Generation for Open-Domain Dialogue Systems,” *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 352–365, 2021.
- [5] J. Ramos, “Using TF-IDF to determine word relevance in document queries,” 2003.
- [6] S. E. Robertson and H. Zaragoza, “The Probabilistic Relevance Framework: BM25 and Beyond,” *Found. Trends Inf. Retr.*, Vol. 3, No. 4, pp. 333–389, 2009.
- [7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is All you Need,” *Proceedings of the 30th Neural Information Processing Systems*, pp. 5998–6008, 2017.
- [8] S. Roller, E. Dinan, N. Goyal, D. Ju, M. Williamson, Y. Liu, J. Xu, M. Ott, E. M. Smith, Y. Boureau, and J. Weston, “Recipes for Building an Open-Domain Chatbot,” *Proceedings of the 16th European Chapter of the Association for Computational Linguistics*, pp. 300–325, 2021.
- [9] C. Goutte and É. Gaussier, “A probabilistic interpretation of precision, recall and F -score, with implication for evaluation,” *Advances in Information Retrieval, 27th European Conference on IR Research, ECIR 2005, Santiago de Compostela, Spain, March 21-23, 2005, Proceedings*, ser. Lecture Notes in Computer Science, Vol. 3408, pp. 345–359, 2005.
- [10] J. Jeff, D. Matthijs, and J. Hervé, “Billion-scale similarity search with GPUs,” *IEEE Transactions on Big Data*, Vol. 7, No. 3, pp. 535–547, 2019.