DBERT: 멀티턴 문맥의 특징을 고려한 대조 학습 기반의 임베딩 모델링+

박상민, 이재윤○, 김재은++
㈜솔트룩스, AI Labs
{sangmin.park, jaeyun.lee, jaieun.kim}@saltlux.com

DBERT: Embedding Model Based on Contrastive Learning Considering the Characteristics of Multi-turn Context

Sangmin Park, Jaeyun Lee, Jaieun Kim++
AI Labs, Saltlux Inc

요 약

최근에는 사람과 기계가 자유롭게 대화를 주고받을 수 있는 자유 주제 대화 시스템(Open-domain Dialogue System)이 다양한 서비스에 활용되고 있다. 자유 주제 대화 시스템이 더욱 다양한 답변을 제공할 수 있도록 사전학습 기반의 생성 언어모델이 활용되고 있지만, 답변 제공의 안정성이 떨어져 검색을 활용한 방법 또한 함께 활용되고 있다. 검색 기반 방법은 사용자의 대화가 들어오면 사전에 구축된 데이터베이스에서 유사한 대화를 검색하고 준비되어있는 답변을 제공하는 기술이다. 하지만 멀티턴으로 이루어진 대화는 일반적인 문서의 문장과 다르게 각 문장에 대한 발화의 주체가 변경되기 때문에 연속된 발화 문장이 문맥적으로 밀접하게 연결되지 않는 경우가 있다. 본 논문에서는 이와 같은 대화의 특징을 고려하여 멀티턴 대화를 효율적으로 임베딩 할 수 있는 DBERT(DialogueBERT) 모델을 제안한다. 기존 공개된 사전학습 언어모델 기반의 문장 임베딩 모델과 비교 평가 실험을 통해 제안하는 방법의 우수성을 입증한다.

주제어: 멀티턴 대화, 유사대화 검색, 임베딩 모델, 학습데이터 구축

1. 서론

자유 주제 대화 시스템은 사람과 기계가 자유롭게 대화를 주고받는 대화 시스템으로 사용자의 흥미를 유발하고 지속적인 대화가 가능하도록 한다. 일반적으로 자유 주제 대화 시스템은 사전학습 기반의 생성 언어모델(Generative Language Model)을 활용하여 자유도가 높은 발화를 생성하고 대화를 진행하지만, 생성되는 발화는 시스템 입장에서 관리하기 어렵고 안정성이 떨어진다는 단점이 있다.

한편 최근에는 방대한 대화 데이터를 사전에 구축해놓고 사용자의 발화와 유사한 경우를 검색함으로써 사용자에게 적절한 답변을 제공하는 검색(Retrieval) 기반의 방법이 사용되고 있다. 예로, 직전 발화와 이전까지 진행된 발화를 검색하여 사용자에게 답변을 제공하는 방법을 챗봇 서비스 '이루다'에서 활용하고 있으며, DPR(Dense Passage Retrieval) 모델을 바탕으로 검색 기반의 방법을 활용하여 자유 주제 대화를 진행하는 페이스북의 '블랜더봇'이 있다[1,2,3].

하지만 대화 시스템에 검색 기반의 방법이 활용됨에도 불구하고 대화 시스템의 대표적인 특징인 멀티턴 대화를 위한 적절한 디자인과 연구가 부족하다.

멀티턴으로 이루어진 대화는 위키백과, 뉴스, 사설 등 일반적인 문서의 문장과 다르게 각 문장에 대한 발화의 주체가 변경되기 때문에 연속된 발화 문장이 문맥적으로 밀접하게 연결되지 않는 경우가 있다. 따라서 기존에 공개된 사전학습 언어모델 기반의 문장 임베딩 모델[4,5,6]을 통해 멀티턴 대화를 임베딩하고 검색하기에는 적절치 않다.

본 논문에서는 이와 같은 멀티턴 대화의 특징을 고려하여 멀티턴 대화를 효율적으로 임베딩 할 수 있는 DBERT 모델을 제안하고 공개된 사전학습 언어모델 기반의 문장 임베딩 모델과 비교 평가 및 분석을 통해 제안하는 방법의 우수성을 입증한다.

2. 제안 방안

표 1은 위키백과, 뉴스, 사설 등과 같이 연속된 일반적인 문장과 대화 도메인에서의 연속된 문장에 대한 예시를 나타낸다. 일반적인 문장의 경우, 연속된 각 문장이 문맥적으로 밀접한 관계로 연결되어 있으며 공개된 SBERT, SIMCSE 등과 같은 사전학습 기반의 임베딩 모델

+ 이 논문은 2022년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No. 2013-2-00109, (엑소브레인-2세부) WiseKB: 빅데이터 이해 기반 자가학습형 지식베이스 및 추론 기술 개발)
++ 교신저자 : 김재은

을 통해 효율적인 임베딩이 가능하다. 하지만 대화에서 연속된 문장은 문맥적으로 밀접하지 않은 관계를 보이며 공개된 임베딩 모델을 통해 벡터를 임베딩하고 표상하기에 적절하지 않다.

본 논문에서는 이와 같은 멀티턴 대화 특징을 잘 반영한 멀티턴 유사 쌍 학습데이터를 자동 구축하는 방안을 제안하고 구축된 데이터를 활용하여 효율적으로 임베딩 할 수 있는 DBERT 모델을 제안한다.

표 1 연속된 일반 문장과 멀티턴 대화 예시

도메인	문장
일반	손흥민은 대한민국의 축구선수이다.
	잉글랜드 프리미어리그 토트넘 홋스퍼에서 활약하고 있으며 대한민국 축구 국가대표이다.
	프리미어리그 공식 베스트 일레븐에 선정되었으며 득점왕을 수상하였다.
대화	이번에 손흥민 경기 봤어?
	응 손흥민 이번에 3골 넣었잖아!
	응 진짜 그때 나 페르시치 난입했을 때 엄청 웃겼어

2.1 멀티턴 대화 데이터 구축방안

본 논문에서는 멀티턴 대화 데이터 구축을 위해 모두의 말뭉치 홈페이지에 구축되어있는 '온라인 대화 말뭉치 2021' [7]를 활용한다. 그림 1은 멀티턴 대화 데이터 구축방안을 나타낸다. '온라인 대화 말뭉치 2021'은 사용자의 스피커 식별자(Speaker Id), 스피커의 발화(Form)를 통해 발화하는 스피커 정보를 알 수 있도록 구성되어있으며 스피커 정보가 바뀌기 전까지를 하나의 발화로 정의하고 총 N개의 발화를 이동 간격(Stride)을 한 칸씩 조정해나가며 멀티턴 대화를 구축한다.

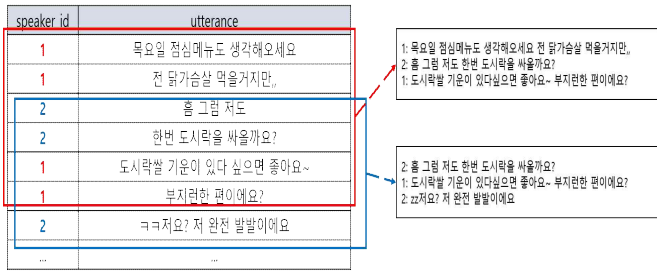
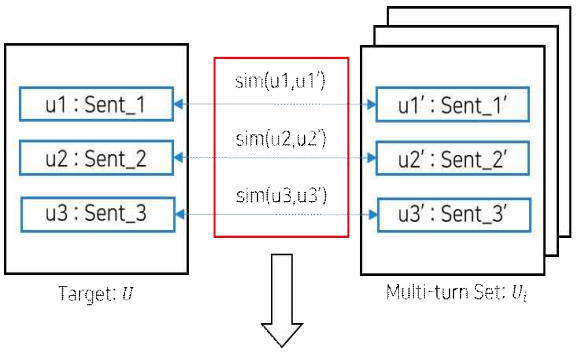


그림 1 멀티턴 대화 데이터 구축방안

2.2 DBERT 모델을 위한 멀티턴 유사 쌍 학습데이터 자동 구축방안

본 논문에서는 멀티턴 특징이 고려된 멀티턴 유사 쌍 학습데이터 구축을 위해 SBERT-IQ 모델[8]을 활용한 학습데이터 자동 구축방안을 제안한다. 본 절에서 타겟(Target)은 유사 쌍을 구하기 위한 멀티턴 대화라고 가정하고 멀티턴 세트(Multi-turn Set)는 타겟을 제외한 나머지 멀티턴 대화로 정의하며, 이에 따른 구축방안은 다음과 같다.

첫 번째, 그림 2에서 나타낸 것과 같이 타겟과 모든 멀티턴 세트 사이 문장 간 코사인 유사도를 SBERT-IQ 모델을 통해 계산한다. 두 번째, 계산된 각 코사인 유사도 점수에 대해 가중 평균 점수를 계산한 후 상위 1번째 점수를 갖는 멀티턴 쌍을 유사 쌍 학습데이터(positive)로 구축하고 그 가중 평균점수를 레이블로 한다. 세 번째, 네거티브(negative) 데이터 세트 구축을 위해 가중 평균 점수가 하위 1000개인 데이터 세트 중 무작위로 1개의 멀티턴 대화 쌍을 뽑고 레이블을 0점으로 부여한다.



$$score(U, U_i) = \alpha * sim(u1, u1') + \beta * sim(u2, u2') + \gamma * sim(u3, u3')$$

그림 2 멀티턴 쌍 유사도 계산 방안

표 2는 제안 방안을 통해 구축한 멀티턴 유사 쌍 데이터 세트의 예시이다.

표 2 멀티턴 유사 쌍 데이터 예시

타입	발화	스코어
타겟	u1 : 아 오늘 날씨 좋다 ㅎㅎ u2 : 날씨도 좋은데 내일 놀러갈래? u3 : 오 그거 좋지 어디로 갈까?	-
긍정	u1 : 날씨 진짜 좋네 ㅎㅎ u2 : 그니까 ㅋㅋ내일 놀래??? u3 : 그래 뭐하고 놀까	0.86
부정	u1 : 배고파요.. u2 : 밥 안먹었어요? u3 : 네 어제부터 다이어트 중이라서요..	0.00

3. 실험 환경 및 학습데이터 통계

본 논문에서 제안한 DBERT 모델은 제안 방안을 통해 자동 구축한 멀티턴 유사 쌍 학습데이터를 활용하여, SBERT 모델과 동일한 구조로 2 에폭, 16배치, AdamW 최적화 기법과 2e-5의 학습률을 사용하여 학습하였다. 표 3은 데이터 구축을 위해 활용한 말뭉치 코퍼스, 멀티턴 대화 그리고 학습, 검증, 테스트 데이터 세트의 통계를 나타낸다.

표 3 실험 데이터 통계

말뭉치(어절)	멀티턴	학습	검증	테스트
3,069,927	1,246,300	169,914	18,880	1,953

4. 실험 평가

표 4는 제안하는 DBERT 모델과 SBERT-IQ 모델에 대한 정량적인 검색성능 평가 결과이다. 평가 방법은 다음과 같다. 첫 번째, 구축된 전체 멀티턴 대화 데이터 세트 중 랜덤하게 추출된 10만 건의 멀티턴 대화 데이터와 테스트 데이터 중 하나의 멀티턴 대화 데이터를 데이터베이스에 저장한다. 두 번째, 테스트 데이터 중 데이터베이스에 저장되지 않은 멀티턴 문장을 쿼리로 하여 코사인 유사도 기반의 전체탐색 알고리즘으로 검색을 수행한다. 세 번째, 검색한 멀티턴 대화 데이터 중 기존 테스트 데이터에서 유사 쌍으로 구축되어있던 멀티턴 대화 데이터가 식별될 경우 정답으로 한다.

이와 같은 평가 방안을 통해 검색성능 평가를 수행한 결과 제안한 모델이 Top-1 검색성능에서 약 15.2%, Top-3 검색성능에서 10.2% 그리고 Top-5 검색성능에서 약 14.0% 이상의 성능 향상을 보였으며 이를 통해 제안한 모델의 우수성을 입증하였다.

표 4 임베딩 모델 별 검색 성능 평가 결과

Top-k	SBERT-IQ(baseline)	DBERT(ours)
1	41.63%	**56.78%**
3	64.00%	**74.24%**
5	72.25%	**86.28%**

5. 결론 및 향후 연구

본 논문에서는 멀티턴으로 이루어진 대화를 효율적으로 잘 임베딩하기 위해 멀티턴 특징을 반영한 멀티턴 유사 쌍 데이터 자동 구축방안을 제안하고 DBERT 모델을 학습하였다. DBERT 모델은 멀티턴 대화의 특징을 고려하여 멀티턴 대화에 대한 임베딩 벡터를 효율적으로 표상할 수 있도록 개선되었으며 멀티턴 대화 검색성능 평가 결과 공개된 임베딩 모델 보다 약 10~15% 이상 향상된 성능을 통해 제안 방안의 우수성을 입증하였다.

제안한 DBERT 모델은 자유 주제 대화 시스템, 목적형 대화 시스템 등 멀티턴 대화를 기반으로 한 다양한 어플리케이션에서 연속된 대화의 문맥을 효율적으로 표상하고 서비스 성능을 높일 수 있을 것으로 예상된다.

향후 연구로는 DBERT 모델 학습을 위해 활용된 사전학습 언어모델을 대화 도메인에 최적화하여 성능을 향상시키는 연구를 수행할 예정이며 하나의 발화에 2개 이상의 의도가 담겨 있는 경우에 중요도가 더 높은 의도에 집중하여 효과적인 임베딩 벡터를 생성할 수 있는 연구를 진행할 예정이다.

참고문헌

[1] Jiho Park. (2021). [online]. Available: https://jiho-ml.com/weekly-nlp-31/, Accessed in 2022.

[2] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, Wen-tau Yih. “Dense Passage Retrieval for Open-Domain Question Answering”. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 6769-6781, 2020.

[3] Jason Weston, Kurt Shuster. (2021). [online]. Available: https://ai.facebook.com/blog/blender-bot-2-an-open-source-chatbot-that-builds-long-term-memory-and-searches-the-internet/, Accessed in 2022.

[4] N. Reimers, I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks", Journal of EMNLP : Natural Language Processing, pp. 3980-3990, 2019.

[5] Tianyu Gao, Xingcheng Yao, Danqi Chen “SimCSE: Simple Contrastive Learning of Sentence Embeddings”, In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp.6894-6910, 2021.

[6] D. Cera, Y. Yanga, S. Konga, N. Huaa, N. Limtiacob, R. Johna, N. Constanta, M. Guajardo-Cespedes, S. Yuanc, C. Tara, Y. Sunga, B. Stropea, R. Kurzweila, "Universal Sentence Encoder," Journal of EMNLP : System Demonstrations, Vol. 1, pp. 169-174, Apr. 2018.

[7] 국립국어원. (2022). [online]. Available: https://corpus.korean.go.kr, Accessed in 2022.

[8] 박상민, 이재윤, 손유리, 김재은, “SBERT-IQ: 키워드 정보량을 고려한 Sentence-BERT 기반의 임베딩 모델”, 한국정보과학회 학술발표논문집, pp.1058-1060, 2022.