다중목표 대화형 추천시스템을 위한 사전 학습된 언어모델들에 대한 성능 평가

(Performance Evaluation of Pre-trained Language Models in Multi-Goal Conversational Recommender Systems)

김태호*, 장형준**, 김상욱***

(Taeho Kim, Hyung-Jun Jang, Sang-Wook Kim)

요 약

본 연구는 대화형 추천 시스템인 다중 목표 대화형 추천 시스템(MG-CRS)에서 사용되는 다양한 사전 학습된 언어 모델들을 고찰하고, 각 언어모델의 성능을 비교하고 분석한다. 특히, 언어 모델의 크기가 다중 목표 대화형 추천 시스템의 성능에 어떤 영향을 미치는지에 대해 살펴본다. BERT, GPT2, 그리고 BART의 세 종류의 언어 모델을 대상으로 하여, 대표적인 다중 목표 대화형 추천 시스템 데이터셋인 DuRecDial 2.0에서 '타입 예측'과 '토픽 예측'의 정확도를 측정하고 비교한다. 실험 결과, 타입 예측에서는 모든 모델이 뛰어난 성능을 보였지만, 토픽 예측에서는 모델 간에 혹은 사이즈에 따라 성능 차이가 관찰되었다. 이러한 결과를 바탕으로 다중 목표 대화형 추천 시스템의 성능 향상을 위한 방향을 제시한다.

■ 중심어 : 추천시스템 ; 대화형추천시스템 ; 다이얼로그 ; 언어모델

Abstract

In this study paper, we examine pre-trained language models used in Multi-Goal Conversational Recommender Systems (MG-CRS), comparing and analyzing their performances of various pre-trained language models. Specifically, we investigates the impact of the sizes of language models on the performance of MG-CRS. The study targets three types of language models - of BERT, GPT2, and BART, and measures and compares their accuracy in two tasks of 'type prediction' and 'topic prediction' on the MG-CRS dataset, DuRecDial 2.0. Experimental results show that all models demonstrated excellent performance in the type prediction task, but there were notable provide significant performance differences in performance depending on among the models or based on their sizes in the topic prediction task. Based on these findings, the study provides directions for improving the performance of MG-CRS.

■ keywords : recommender systems ; conversational recommender systems ; dialogue ; language model

Ⅰ. 서 론

대화형 추천 시스템(Conversational Recommender Systems)은 사용자와 시스템 간의 멀티-턴 대화를 기반으로 (1) 사용자의 선호도를 예측하며 추천 상품을 제시하고, (2) 대화의 맥락에 따라 적절한 응답을 생성하는 추천 시스템을 의미한다. 대화형 추천 시스템은 실시간 대화를 통해 사용자의 선호도를 보다 정확히 파악할 수 있다는 장점이 있어, 학계와 산업계에서 크게 주목받고 있다 [1,2,6-8,12-16].

대화형 추천의 성공 여부는 시스템이 사용자와의 대화 흐름을 얼마나 잘 관리하는가에 크게 의존한다. 최근에는 기존의 대화형 추천 방식과 같이 추천 상품과 응답 생성에만 초점을 맞추는 것이 아니라,

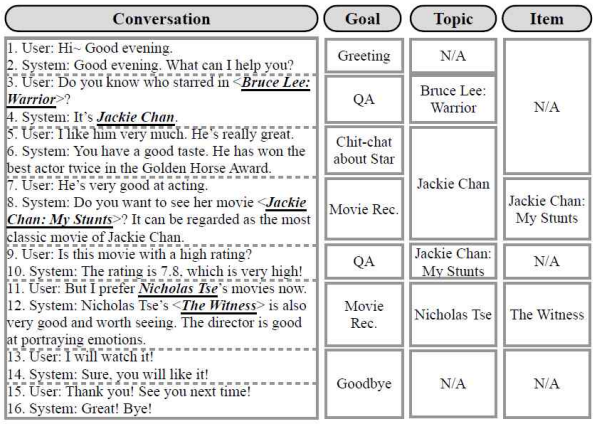
* 학생회원, 한양대학교 컴퓨터소프트웨어 학과 ** 학생회원, 한양대학교 AI응용 학과 *** 정회원, 한양대학교 컴퓨터소프트웨어 학과

이 논문은 2018년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No.2018R1A5A7059549, 그리고 No. 2020-0-01373, 인공지능대학원지원(한양대학교)).

접수일자 : 2023년 06월 26일
수정일자 : 1차 2023년 07월 13일
게재확정일 : 2020년 07월 20일

교신저자 : 김상욱 wook@agape.hanyang.ac.kr

대화 내의 각 서브 대화의 목표를 관리하는 다중 목표 대화형 추천 시스템(Multi-Goal Conversational Recommender Systems)에 대한 연구가 활발히 이루어지고 있다 [2,7,8]. 다중 목표 대화형 추천 시스템은 추천 상품과 응답 발화뿐만



Conversation	Goal	Topic	Item
1. User: Hi~ Good evening. 2. System: Good evening. What can I help you?	Greeting	N/A	N/A
3. User: Do you know who starred in <***Bruce Lee: Warrior***>? 4. System: It's ***Jackie Chan***.	QA	Bruce Lee: Warrior	
5. User: I like him very much. He's really great. 6. System: You have a good taste. He has won the best actor twice in the Golden Horse Award.	Chit-chat about Star	Jackie Chan	
7. User: He's very good at acting. 8. System: Do you want to see her movie <***Jackie Chan: My Stunts***>? It can be regarded as the most classic movie of Jackie Chan.	Movie Rec.		Jackie Chan: My Stunts
9. User: Is this movie with a high rating? 10. System: The rating is 7.8, which is very high!	QA	Jackie Chan: My Stunts	N/A
11. User: But I prefer ***Nicholas Tse***'s movies now. 12. System: Nicholas Tse's <***The Witness***> is also very good and worth seeing. The director is good at portraying emotions.	Movie Rec.	Nicholas Tse	The Witness
13. User: I will watch it! 14. System: Sure, you will like it! 15. User: Thank you! See you next time! 16. System: Great! Bye!	Goodbye	N/A	N/A

그림 1. 다중 목표 대화형 추천 시스템 데이터셋 예시 (UniMIND[2] 논문 제공)

아니라, 각 시스템 응답 발화의 유형(예: 채팅, 질문-답변, 추천 등)을 예측하고, 해당 발화의 주요 토픽 키워드(예: 배우, 감독 등)도 예측하게 된다(그림1 참조).

다중 목표 대화형 추천시스템은 사용자와의 대화를 기반으로 타입 예측, 토픽 예측, 상품 예측, 응답 예측 등 여러 태스크를 수행하게 된다. 이에 따라 여러 자연어 처리 태스크에서 범용적으로 좋은 성능을 보이는 사전학습된 언어모델(예: BERT, GPT, BART)을 사용하면, 더욱 품질 높은 다중 목표 대화형 추천시스템을 구현할 수 있을 것이다. 일례로 UniMIND [2]는 Transformer encoder-decoder 구조 [11]의 BART 모델 [5]을 사용하여 다중 목표 대화형 추천시스템의 각 태스크의 성능을 크게 향상시켰다.

성공적인 다중 목표 대화형 추천시스템을 위해선, 기반 언어 모델을 적절히 선택하고, 이를 각 태스크에 맞게 효과적으로 미세 조정하는 것이 중요할 것이다. 그러나 현재까지의 연구에서는 다중 목표 대화형 추천시스템에서 언어모델들 종류에 따른 성능 비교나, 언어모델의 사이즈에 따른 성능 비교에 대해 심층적인 분석이 충분히 이루어지지 않고 있다. 이에 따라 본 연구에서는 다양한 언어모델들의 크기에 따라 다중 목표 대화형 추천시스템 태스크의 성능이 어떻게 변화하는지를 측정하고 분석하는 것을 목표로 한다.

구체적으로 본 연구는 BERT[3], GPT2[10], 그리고 BART[5]와 같은 세 종류의 사전 학습된 언어모델을 대상으로 하여, 대표적인 다중 목표 대화형 추천시스템 데이터셋인 DuRecDial 2.0[8]에서 '타입 예측' 및 '토픽 예측'의 정확도를 측정하고 비교한다. 또한, 각 모델의 성능평가는 base 모델과 large 모델을 활용하여 실험을 수행함으로써, 모델 사이즈에 따른 성능 변화 역시 조사한다.

실험 결과를 요약하면, (1) '타입 예측'에서는 모든 모델이 뛰어난 성능을 보였고, 모델들 간에 혹은 사이즈에 따라 큰 정확도 차이를 보이지 않았다. 하지만 (2) '토픽 예측'에서는 '타입 예측'에 비해 낮은 정확도를 보였으며, 모델들 간에 혹은 사이즈에 따라 정확도 차이를 보였다. 구체적으로, 타입 예측에서는 모든 모델들이 높은 성능을 보였으나, BERT-large 모델이 가장 좋은 정확도를 보였습니다. 반면, 토픽 예측에서는 BART-large 가 좋은 성능을 보였습니다.

본 논문의 구성은 다음과 같다. 먼저, 2장에서는 MG-CRS 문제 정의와 세 가지 언어모델(BERT, GPT2, BART)를 통해 다중 목표 대화형 추천시스템 문제를 해결하는 방법에 대해 소개한다. 그리고 3장에서는 대표적인 다중 목표 대화형 추천시스템 데이터셋인 DuRecDial2.0에서의 성능 평가 결과를 제시하며, 그 결과의 의미를 분석한다.

Ⅱ. 본 론

본 장에서는 MG-CRS에서 각 언어모델이 어떻게 수행되는지 설명한다. 구체적으로, 2.1절에서는 다중 목표 대화형 추천시스템 문제 정의를 한다. 2.2절에서는 BERT [3]에 대해 소개한다. 다음으로, 2.3절에서는 GPT2 [10]에 대해 소개한다. 마지막으로, 2.4절에서는 BART [5]에 대해 소개한다.

2.1. 다중 목표 대화형 추천시스템 문제정의

다중 목표 대화형 추천시스템(multi-goal conversational recommender sytsem)는 사용자와 시스템 간 대화가 주어졌을 때, 다음 응답 생성에 관련된 정보들을 예측하는 것을 목적으로 한다. 구체적으로 (1) 다음 응답의 타입(e.g., '영화추천', '음악추천', '잡담', '인사', '질의응답')을 예측하는 것, (2) 다음 응답의 토픽 키워드(e.g., '겨울왕국', '마

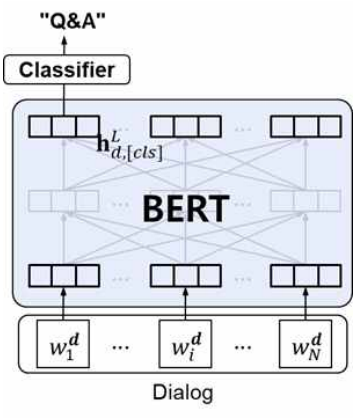


그림 2 BERT 모델 아키텍쳐

블', '제임스 카메론', '송강호')를 예측하는 것, (3) 다음 응답에서 추천할 상품을 예측하는 것, 그리고 마지막으로 (3) 다음 응답 자체를 예측하는 것을 목적으로 한다. 본 논문에서는 다음 응답의 타입과 토픽을 예측하는 것에만 집중하여 설명하고자 한다.

2.2. BERT (Bidirectional Encoder Representations from Transformers) [3]

BERT는 Google에서 2018년 개발한 사전 학습 언어 모델로, Transformer 인코더의 아키텍처를 사용하는 것을 특징으로 한다. BERT는 입력 문장 내 각 단어의 양방향 문맥을 self-attention layer [11]를 통해 참조할 수 있기에, 주어진 단어의 앞뒤에 있는 모든 단어를 참조하여 그 단어의 의미를 이해할 수 있다는 장점이 있다. BERT는 대용량 코퍼스(e.g., Wikipedia, book corpuse)에서 두 가지 사전 학습 태스크를 사용하여 학습되는데, 하나는 Masked Language Model (MLM)이고 다른 하나는 Next Sentence Prediction (NSP)이다. 이를 통해 BERT는 여러 NLP 작업(e.g., QA, natural language inference 등)에서 높은 성능을 보여주었다.

MG-CRS의 타입 예측과 토픽 예측을 수행하기 위해, 우리는 먼저 주어진 사용자와 시스템간의 입력 대화($[w_1^d, w_2^d, \ldots, w_N^d]$)를 BERT를 통해 인코딩한다. 이때, 우리는 이전 연구들과 마찬가지로 BERT의 출력 결과 중 [CLS] 스페셜 토큰의 결과를 입력 대화의 대표벡터 (representation vector)로 간주한

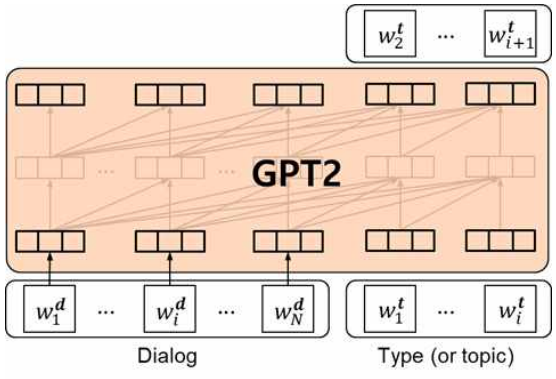


그림 3 GPT 모델 아키텍쳐

다. 그 다음, 입력 대화의 대표벡터를 단일층(single layer) 분류기에 통과하여, 전체 타입(혹은 토픽)의 수 크기만큼의 클래스로 분류한다(그림2 참조). 학습(training) 시에는 정답 타입(혹은 토픽)에 대한 분류 점수가 다른 타입들(혹은 토픽들)에 대한 분류 점수보다 높아지도록 음의 로그 우도(negative log likelihood) 손실함수를 최소화한다. 추론(inference) 시에는 가장 높은 분류 점수를 받은 상위 K개의 타입(혹은 토픽)을 최종 결과로 출력한다.

2.3. GPT-2 (Generative Pre-training Transformer 2)[10]

GPT-2는 OpenAI에서 개발한 모델로, Transformer 디코더 구조 [11] 기반의 단방향 생성 모델이다. GPT-2는 주어진 문맥을 기반으로 다음에 올 단어를 예측하는 것을 기본으로 한다. GPT-2는 BERT와 마찬가지로 대용량 코퍼스를 통해 사전 학습되며, 이 과정에서는 언어를 이해하고 문맥에 맞게 자연스러운 문장을 생성하는 방법을

배운다.

MG-CRS의 타입 예측과 토픽 예측을 수행하기 위해, 우리는 GPT-2의 입력으로 사용자와 시스템 간 대화($[w^d_1, w^d_2, ..., w^d_N]$)를 넣고, 출력으로 목표로 하는 다음 응답의 타입(혹은 토픽)을 텍스트($[w^t_1, w^t_2, ..., w^t_i]$)로서 생성하도록 미세조정(fine-tuning)한다. 추론(inference) 시에는 이전 출력 토큰을 다음 입력 토큰으로 하는 자기회귀(auto regressive) 방법을 활용하여, 타입(혹은 토픽)을 생성한다(그림3 참조).

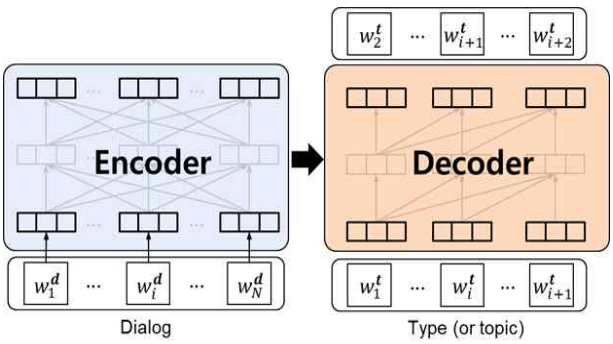


그림 4 BART 모델 아키텍쳐

2.4. BART (Bidirectional and Auto-Regressive Transformers)[5]

BART는 Facebook AI에서 개발한 언어 모델로, Transformer 인코더-디코더 구조[11]를 활용한다. 먼저 BART는 양방향 Transformer 인코더를 사용하여 문장에서 랜덤하게 마스킹 된 토큰을 예측하고, 그 결과를 Transformer 디코더를 사용하여 원래의 문장으로 복원하도록 사전 학습된다. BART는, GPT-2와 달리, 입력 문장의 양방향 문맥을 모두 고려하며 문장을 생성하는 방법을 배울 수 있다는 장점이 있다. BART는 원문 생성, 문서 요약, 번역 등의 여러 NLP 작업에서 높은 성능을 보인다.

MG-CRS의 타입 예측과 토픽 예측을 수행하기 위해, GPT-2와 마찬가지로, 우리는 BART의 입력으로 사용자와 시스템 간 대화($[w^d_1, w^d_2, ..., w^d_N]$)를 넣고, 출력으로 목표로 하는 다음 응답의 타입(혹은 토픽)을 텍스트($[w^t_1, w^t_2, ..., w^t_i]$)로서 생성하도록 미세조정(fine-tuning)한다. 추론 (inference) 시에는 이전 출력 토큰을 다음 입력 토큰으로 하는 자기회귀(auto regressive) 방법을 활용하여, 타입(혹은 토픽)을 생성한다 (그림4 참조).

Ⅲ. 실 험

본 장에서는 MG-CRS 태스크를 푸는 데 각 언어 모델들의 사이즈에 따른 성능을 평가한다.

2.1. 실험세팅

가. 데이터셋

DuRecDial은 목표 지향적이며 지식 기반의 대화 추천 데이터셋으로, 영화, 음악, 식당 등 다양한 분야에 걸친 대화들을 포함하고 있다. 각 시스템의 발화 차례에서의 타입과 토픽(지식 엔티티)가 제공된다. 발화 타입의 종류는 총 21가지이다. 우리는 지식 엔티티를 대화 토픽으로 취급하며, 이 데이터셋은 과거 아이템 상호작용이 포함된 사용자 프로파일과 각 대화와 관련된 지식 베이스를 함께 제공한다.

나. 평가지표

우리는 MG-CRS의 여러 태스크 중 타입 예측과 토픽 예측에 집중하여 모델을 평가한다. 구체적으로 우리는 각 모델을 히트 비율 (hit ratio) 평가지표[1,2,13,14]로 평가하고자 한다. 히트 비율은 모델이 추천한 상위 K개 상품들 중에 실제 정답(즉, 실제 응답 타입 혹은 토픽)이 포함하는지 평가한다.

다. 구현명세

우리는 각 모델을 huggingface PyTorch library*를 통해 구현한다. 구체적으로, huggingface 의 Transformer 라이브러리 중 BERT-base-uncased (110million parameters), BERT-large-uncased (340 million parameters), GPT2 (117 million parameters), GPT2-large (774 million parameters), facebook/bart-base (140 million parameters), facebook/bart-large

* https://huggingface.co/

(400 million parameters) 모델을 사용한다. 학습률(learning rate)은 1e-6부터 1e-4 범위 내에서 조정하였고, 배치 크기(batch size)는 8에서 32 사이에서 조정하였다. 최적화기(optimizer)로는 AdamW를 사용하였다.

2.1. 실험결과

본 실험에서는 다중 목표 대화형 추천 시스템에서 사전 학습된 언어 모델들의 성능을 비교하였다. 이를 위해 BERT, GPT2, 그리고 BART의 base와 large 모델들을 사용하여 '타입 예측' 및 '토픽 예측' 작업을 수행하였다. 우리는 각 시스템의 응답의 타입이 '질의응답', '장소추천', '영화추천', '음악추천' 등 네 가지 경우에 한정하여 실험을 진행하였고 그 결과를 표4에 정리하였다.

먼저, '타입 예측' 결과를 살펴보면, 모든 모델이 질의응답과 장소추천 부분에서 완벽한 성능을 보였다. 영화추천과 음악추천 부분의 성능은 약간 못 미친다. 전반적으로 성능의 차이가 크지 않지만, BERT-large 모델이 가장 높은 성능을 보였다.

반면, '토픽 예측' 결과를 보면, 모든 모델의 성능이 다소 떨어지는 것을 확인할 수 있다. 이는 '토픽 예측' 작업이 '타입 예측' 작업보다 더 복잡하며, 모델이 입력 대화의 세세한 정보를 더욱 잘 이해해야 하기 때문으로 보인다. 모델 별로 살펴보면, BART 모델들이 높은 성능을 보였다.

모델의 사이즈에 따른 성능 차이에 대해 살펴보면, GPT2와 BART의 경우 large 모델이 base 모델보다 전반적으로 더 높은 성능을 보여주었다. 이는 large 모델이 base 모델보다 더 많은 파라미터를 가지고 있어, 복잡한 패턴을 학습하는 능력이 더 높기 때문으로 보인다. 이는 최근 대용량 생성형 언어모델이 모델 사이즈가 커짐에 따라 여러 NLP에서 비약적인 성능을 보였다는 결과와 비슷한 맥락의 결과로 보인다[4,9].

표 1. 각 언어모델의 사이에 따른 타입 및 토픽 정확도 측정 결과

Model	타입 예측					토픽 예측				
	질의 응답	장소 추천	영화 추천	음악 추천	총점	질의 응답	장소 추천	영화 추천	음악 추천	총점
BERT -base	1	1	0.9743	0.9621	0.9771	0.9310	0.5573	0.5250	0.5356	0.5854
BERT -large	1	1	0.9910	0.9743	0.9879	0.9471	0.5617	0.5022	0.5401	0.5804
GPT2 -base	1	1	0.9383	0.9309	0.9521	0.9310	0.6479	0.5242	0.3120	0.5235
GPT2 -large	1	1	0.9699	0.9513	0.9717	0.9563	0.6938	0.5771	0.5068	0.6188
BART -base	1	1	0.9816	0.9567	0.9783	0.9586	0.7070	0.5969	0.5320	0.6372
BART -large	1	1	0.9875	0.9459	0.9771	0.9540	0.7379	0.5954	0.5500	0.6461

종합적으로, 실험 결과는 사전 학습된 언어 모델들이 다중 목표 대화형 추천 시스템에서 높은 성능을 발휘할 수 있음을 보여주었다. 특히 BART 모델들이 전반적으로 우수한 성능을 보였으며, 이는 그들이 비교적 최근에 개발된 모델로, 더욱 향상된 학습 알고리즘과 구조를 가지고 있기 때문으로 해석할 수 있다. 또한, large 모델의 성능이 base 모델보다 전반적으로 높았으며, 이는 더 많은 파라미터가 복잡한 작업을 처리하는 데 도움이 되었음을 보여준다.

Ⅲ. 결 론

본 논문에서는 다중 목표 대화형 추천 시스템에서 사전 학습된 언어 모델들의 성능을 상세 비교 분석

하였다. 우리는 여러 언어모델들이 다중 목표 대화형 추천 시스템에서 높은 성능을 보이며, 특히 사이즈가 커질수록 더욱 좋은 성능을 보인다는 것을 확인하였다. 이를 바탕으로 향후 다중 목표 대화형 추천 시스템 연구에서 더 큰 사이즈의 언어모델을 효과적으로 사용하여 최적화된 성능을 보이기를 기대한다.

REFERENCES

[1] Chen, Qibin, et al., "Towards knowledge-based recommender dialog system," arXiv preprint arXiv:1908.05391, 2019.

[2] Deng, Yeng, et al., "A unified multi-task learning framework for multi-goal conversational recommender systems," *ACM Transactions on Information Systems*, Vol. 41, No. 3, pp. 1-25, Feb. 2023.

[3] Devlin, Jacob, et al., "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv:1810.04805, 2018.

[4] Hu, Edward J., et al., "Lora: Low-rank adaptation of large language models," arXiv:2106.09685, 2021.

[5] Lewis, Mike, et al., "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," arXiv:1910.13461, 2019.

[6] Li, Raymond, et al., "Towards deep conversational recommendations," *Advances in neural information processing systems 31(NeurIPS 2018)*, PP. 9748 - 9758, Red Hook, USA, Dec. 2018.

[7] Liu, Zeming, et al., "Towards conversational recommendation over multi-type dialogs," arXiv:2005.03954, 2020.

[8] Liu, Zeming, et al., "Durecdial 2.0: A bilingual parallel corpus for conversational recommendation," arXiv:2109.08877, 2021.

[9] Liu, P., et al., "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Computing Surveys*, Vol. 55, No. 9, pp. 1-35, Jan. 2023.

[10] Radford, Alec, et al., "Improving language understanding by generative pre-training," 2018.

[11] Vaswani, Ashish, et al., "Attention is all you need," *Advances in neural information processing systems 30(NIPS 2017)*, 2017.

[12] Wang, Xiaolei, et al., "Towards Unified Conversational Recommender Systems via Knowledge-Enhanced Prompt Learning," *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Minin,* PP. 1929 - 1937, New York, USA, Aug. 2022.

[13] Zhang, Jun, et al., "Kers: A knowledge-enhanced framework for recommendation dialog systems with multiple subgoals," *Findings of the Association for Computational Linguistics: EMNLP,* PP. 1092 - 1101, Punta Cana, Dominican Republic, Nov. 2021.

[14] Zhou, Kun, et al., "Improving conversational recommender systems via knowledge graph based semantic fusion," *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining,* pp. 1006-1014, New York, USA, Aug. 2020.

[15] 장준혁, "디지털 소외계층을 위한 지능형 IoT 애플리케이션의 공개 API 기반 대화형 음성 상호작용 기법," *스마트미디어저널*, 제11권, 제10호, 22-30쪽, 2022년 11월

[16] 전민규, 김남규, "텍스트 요약 품질 향상을 위한 의미적 사전학습 방법론," *스마트미디어저널*, 제12권, 제5호, 17-27쪽, 2023년 6월

저 자 소 개

김태호

2013년 한양대학교 컴퓨터소프트웨어학부 (학사)

2019~현재 한양대학교 컴퓨터소프트웨어학과 석박사 통합과정

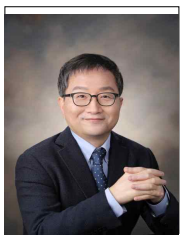
주관심분야 : 추천시스템, 자연어처리

장형준

2013년 아주대학교 교통시스템공학과 (학사)

2022 ~ 현재 한양대학교 AI응용학과 석사과정

주관심분야 : 추천시스템, 자연어처리



김상욱

1989 서울대학교 컴퓨터공학과 졸업 (학사)

1991 한국과학기술원 전산학과 졸업 (석사)

1994 한국과학기술원 전산학과 졸업 (박사)

1991~1991 미국 Stanford University, Computer Science Department 방문연구원

1999~2000 미국 ICM T.J Watson Research Center, Post-Doc.

1995~2003 강원대학교 정보통신공학과 부교수

2003~현재 한양대학교 컴퓨터소프트웨어학부 교수

2009~2010 미국 Carnegie Mellon University, Visiting Scholar

관심분야: 데이터사이언스, 추천 시스템, 소셜 네트워크 분석