한국어 토큰-프리 사전학습 언어모델 KeByT5를 이용한 한국어 생성 기반 대화 상태 추적

이기영○, 신종훈, 임수종, 권오욱
한국전자통신연구원
leeky@etri.re.kr, jhshin82@etri.re.kr, isj@etri.re.kr, ohwoog@etri.re.kr

Korean Generation-based Dialogue State Tracking using Korean Token-Free Pre-trained Language Model KeByT5

Kiyoung Lee○, Jonghun Shin, Soojong Lim, Ohwoog Kwon
Electronics and Telecommunications Research Institute

요 약

대화 시스템에서 대화 상태 추적은 사용자와의 대화를 진행하면서 사용자의 의도를 파악하여 시스템 응답을 결정하는데 있어서 중요한 역할을 수행한다. 특히 목적지향(task-oriented) 대화에서 사용자 목표(goal)를 만족시키기 위해서 대화 상태 추적은 필수적이다. 최근 다양한 자연어처리 다운스트림 태스크들이 사전학습 언어모델을 백본 네트워크로 사용하고 그 위에서 해당 도메인 태스크를 미세조정하는 방식으로 좋은 성능을 내고 있다. 본 논문에서는 한국어 토큰-프리(token-free) 사전학습 언어모델인 KeByT5를 사용하고 종단형(end-to-end) seq2seq 방식으로 미세조정을 수행한 한국어 생성 기반 대화 상태 추적 모델을 소개하고 관련하여 수행한 실험 결과를 설명한다.

주제어: 대화처리, 대화 상태 추적, 토큰-프리 언어모델

1. 서론

대화 시스템은 언어를 매개로 인간과 시스템 간의 상호작용을 가능하게 한다. 일반적으로 대화 시스템은 그 목표하는 바에 따라, 목적 지향 대화 시스템(task-oriented dialogue system)과 챗봇(chatbot)과 같은 재미를 위한 대화 시스템으로 분류된다 [1]. 전통적인 대화 시스템은 파이프라인 방식으로 구성되어 있다. 파이프라인 방식의 대화 시스템은 사용자 발화를 입력받는 음성 또는 텍스트 입력 모듈, 사용자의 의도를 파악하는 언어이해 모듈, 대화 정책(policy)에 따라 시스템의 발화를 결정하는 대화 관리 모듈, 그리고 시스템 발화를 최종 문장의 형태로 생성하는 자연어 생성 모듈로 구성된다. 파이프라인 형태의 대화 시스템은 오류 대처와 도메인 확장에 문제가 있기 때문에 최근에는 심층 신경망에 기반한 대화 시스템이 개발되고 있으며 특히 사전학습 언어모델(pre-trained language model)에 기반한 종단형 방식의 대화 시스템이 좋은 성능을 보이고 있다 [2][3].

대화 상태 추적(Dialogue State Tracking: DST)은 대화 시스템에서 발화 턴을 분석하여 대화 맥락에 포함된 사용자의 요구를 알아내서 사용자의 의도를 파악하는 것을 목적으로 한다. 특히 목적 지향 대화 시스템에서 대화 히스토리와 문맥에 기반하여 사용자 의도를 파악하고 그에 상응하는 최적의 시스템 응답을 생성하기 위해서는 대화 상태 추적이 매우 중요하다.

그림 1은 KLUE 벤치마크 태스크의 DST 태스크[1]에서 WOS (Wizard-of-Seoul) 대화 데이터셋의 일부로서 시스템과 사용자 간의 대화에서 보여지는 대화 상태 정보를 보인다.

그림 1. WOS 데이터셋의 대화 상태 정보 예

그림 1에는 사용자 발화를 분석하여 얻어진 대화 상태 정보가 (belief state)라는 표기로 되어 있다. 그림 1에서 사용자 발화 “서울 중앙에 위치한 호텔을 찾고 있습니다. 외국인 친구도 함께 갈 예정이라서 원활하게 인터넷을 사용할 수 있는 곳이었으면 좋겠어요.”로부터

[1] https://klue-benchmark.com/tasks/73/overview/description

‘숙소-지역’ 슬롯의 값으로 ‘서울 중앙’, ‘숙소-종류’ 슬롯의 값으로 ‘호텔’, ‘숙소-인터넷 가능’ 슬롯의 값으로 ‘yes’가 설정되는데 이와 같이 사용자 발화로부터 가능한 슬롯과 슬롯값을 결정하는 것이 DST의 역할이다.

본 논문은 한국어 토큰-프리(token-free) 사전학습 언어모델인 KeByT5를 백본 네트워크로 사용하여 종단형 seq2seq 방식으로 사용자 발화턴에 대한 대화 상태 정보인 슬롯과 슬롯값을 생성하는 한국어 대화 상태 추적 태스크를 다루며, KLUE 벤치마크 태스크의 WOS-v1.1 데이터셋을 사용하여 제안하는 모델의 학습 및 평가를 수행하였다 [4]. 또한 백본 네트워크인 토큰-프리 한국어 사전학습 언어모델 KeByT5에 대해서도 간략히 설명한다.

2. 관련 연구

최근의 대화 상태 추적 연구는 전통적인 파이프라인 방식에서 심층 신경망을 기반으로 각 모듈이 통합되고 있으며, 더 나아가 사전학습 언어모델 기반의 종단형 방식으로 연구되는 추세이다. [5]는 2개의 인코더를 사용하여 seq2seq 방식으로 대화 상태 추적을 구현한 연구를 소개한다. [6]은 대화 상태 추적을 seq2seq 모델에 적용하였으며 인코더와 디코더를 위해 트랜스포머(Transformers)를 활용하였다. [7]은 생성형 사전학습 언어모델인 GPT-2를 활용한 2단계 생성 기반의 대화 상태 추적 모델을 소개하였다.

또한 멀티 도메인으로의 확장을 고려한 대화 상태 추적 기술에 대한 관심도 많다. 이와 관련하여 자연어 이해 모듈과 대화 상태 추적 모듈을 통합한 자연어 이해 기반 대화 상태 추적을 소개한 연구로는 [8][9] 등이 있다. 미리 정의된 슬롯-슬롯값은 기존의 대화 시스템을 타 도메인으로 확장하기 어렵게 만드는데, [8]은 이러한 문제를 해결하기 위해 BERT 기반의 확장 가능한 대화 상태 추적 모델을 제안하였다. [9]는 슬롯값을 미리 정의해서 후보로 가질 필요 없이 제안하는 세 가지 카피 메카니즘을 사용하여 슬롯 필링(filling)을 수행함으로써 멀티 도메인 환경에서 대화 상태 추적 성능을 개선시키는 방법을 제안하였다.

3. 한국어 토큰-프리 사전학습 언어모델 KeByT5

사전학습 언어모델을 활용하는데 있어, 의학 및 법률 도메인과 같이, 해당 영역에서 주로 사용되는 용어의 차이로 인해 언어 이해 및 전이 학습 능력이 저하되는 경우가 많다 [10]. 이러한 문제는 다양한 다운스트림 태스크에서 미등록어 문제 및 일관되지 않은 생성 문제를 발생시킨다.

토큰-프리 언어모델은 다양한 다운스트림 태스크를 목적으로 할 때 발생하는 도메인 변화 문제에 맞서기 위해 토큰화를 수행하지 않는다, 이를 통해, 토큰-프리 언어모델은 토큰 분리 등 전처리 단계에서 발생하는 정보 손실을 피할 수 있다. 또한, 토큰-프리 언어모델에서는 사전(dictionary) 단위가 바이트(byte) 표현을 모두 포함하고 있어 미등록어 문제가 발생하지 않는다. 이와 관련하여 [11]은 미등록어 문제 완화를 위해 BBPE (Byte-level BPE) 모델을 제안한 바 있다, 이는 다국어를 다뤄야 하는 사전학습 언어모델에서, 동일한 규모의 신경망 파라미터 크기를 갖는 모델에서 어휘 임베딩 파라미터의 비중을 낮춤으로써 더 많은 시퀀스 표현을 학습하는데 도움이 될 수 있다. 또한, 더 많은 파라미터를 입력에 존재하는 오타, 어휘 필터링 등을 우회하기 위해 의도적으로 사람이 알아볼 수 있을 정도로 문자를 변조하는 행위에 강건하기 때문에 더 나은 일반화 성능을 기대할 수 있다.

과거부터 이어져 온 문자 기반(Character-based) 접근방법은 서브워드 단위나 형태소 단위에 비해 요구되는 연산의 양이 높은 반면 성능은 더 낮게 나타나는 경우가 많아 여전히 자리잡지 못하였으나, 사전학습 언어모델에서는 CANINE이나 ByT5, CharFormer와 같은 접근방법이 다양한 문제상황에서 서브워드 모델과도 경쟁할 수 있는 수준으로 보고된 바 있다.

본 논문의 백본 네트워크로 사용한 KeByT5는 한국어 중심의 ByT5 모델을 기본 구조로 학습 후, GBST (Gradient-Based Subword Tokenization) 계층을 인코더에 결합한 뒤, 기반 모델의 가중치를 유지하고, 신규 계층만 초기화하여 추가 학습하는 기법(uptraining)을 통해 학습함으로써 얻어졌다. 이렇게 생성된 토큰-프리 한국어 사전학습 언어모델의 규모, 구성은 아래의 표와 같다.

표 1. 한국어 KeByT5 규모 및 구성

타입	#params	L_{enc}	L_{dec}	D_{ff}	D_{model}
Small	330M	12	4	3584	1472
Base	580M	18	6	3968	1536
Large	1.23B	36	12	3840	1536

4. 한국어 생성 기반 대화 상태 추적 모델

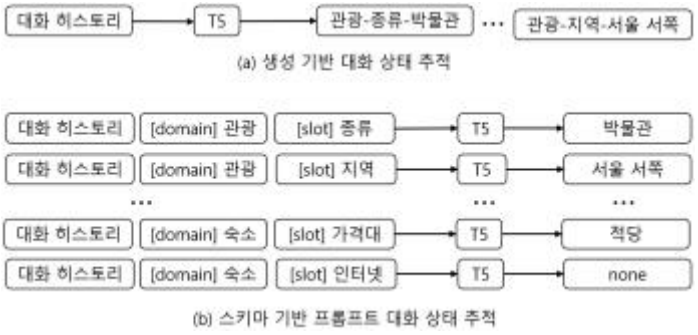

그림 2. 한국어 생성 기반 대화 상태 추적 모델

본 논문에서는 T5 클래스 사전학습 언어모델인 토큰-프리 한국어 사전학습 언어모델을 백본 네트워크로 한 생성 기반의 DST 태스크를 실험하는데 [12]를 참조하였으며, 다음 그림과 같은 2가지 모델을 기본적으로 고려

하였다.

그림 2에서 (a)는 슬롯 및 슬롯값을 직접 생성하는 생성 기반 대화 상태 추적 모델이고 (b)는 스키마 기반의 모델로서 슬롯의 개수만큼 동일한 대화 히스토리가 확장되며, 대화 히스토리에는 도메인 정보와 슬롯 정보가 함께 부착된다.

본 논문의 한국어 생성 기반 DST 모델은 기본적으로 인코더-디코더 구조를 갖는다. 사용자 발화 턴을 U_t라고 하고 이에 대한 시스템 응답을 A_t라고 할 때 대화 문맥 $C_t = \{U_1, A_1, ..., A_{t-1}, U_t\}$로 표현한다. DST 모델 인코더 입력의 토큰 구성은 "[user] U_1 [system] A_1 ... [system] A_{t-1} [user] U_t"이며, 이때 [user], [system]은 각각 사용자 발화와 시스템 응답을 표시하는 세그먼트 토큰을 나타낸다. 인코더 및 디코더의 입출력은 다음과 같다.

$$H_t = Encoder(C_t) \quad (1)$$

$$B_t = Decoder(H_t) \quad (2)$$

마지막으로 모델은 다음을 최대화하도록 학습된다.

$$\sum \log P(B_t | C_t) \quad (3)$$

5. 실험

5.1 데이터셋

본 논문의 토큰-프리 한국어 사전학습 언어모델 KeByT5 기반 한국어 대화 상태 추적 태스크의 성능 평가를 위해 KLUE 벤치마크 DST 태스크의 WOS-v1.1 데이터셋을 사용하였다. WOS-v1.1 대화 데이터셋은 Wizard-of-Oz 방식[13]에 따라, 두 명의 사람이 각각 사용자와 시스템 역할로 수집될 대화의 목표에 따라 대화를 수행함으로써 구축되었다. WOS-v1.1 데이터셋은 {Hotel, Restaurant, Attraction, Taxi, Metro} 와 같이 5개의 도메인에 속하는 대화들로 구성된다. 데이터셋의 특성은 다음과 같다.

표 2. WOS-v1.1 데이터셋

타입	Train	Dev	Test	Total
Dialogues	8000	1000	1000	10000
Single domain dialogues	1806	263	226	2295
Multi domain dialogues	6194	737	774	7705
Total turns	117584	14448	14660	146692
Avg turns per dialogue	14.70	14.45	14.66	14.67
Avg tokens per turns	7.65	7.90	7.84	7.69

실험은 NVIDIA A6000 (48GB) GPU를 사용하여 진행되었다.

5.2 실험 결과

대화 상태 추적 모델의 성능을 평가하기 위해서 본 논문에서는 KLUE 벤치마크에서 제안하는 JGA (Joint Goal Accuracy) 와 slot micro F1 스코어를 사용하였다.

표 3은 한국어 토큰-프리 사전학습 언어모델 KeByT5를 사용한 생성 기반 대화 상태 추적 모델의 성능을 보인다. KeByT5 모델의 경우, base 모델은 JGA, F1 스코어가 각각 77.15%와 96.92%를 보였으며, large 모델의 경우, 각각 78.54%와 97.28%를 보였다. 이와 같은 성능은 베이스라인 모델인 KLUE-RoBERTa-large 모델에 비해 높은 것으로 측정된다.

표 3. 생성 기반 한국어 대화 상태 추적 모델 성능

Model	WOS	
	JGA (%)	F1 (%)
KLUE-RoBERTa-large	50.22	92.23
ETRI-KeByT5-base(ours)	77.15	96.92
ETRI-KeByT5-large(ours)	78.54	97.28

표 4는 동일한 사전학습 언어모델 KeByT5-base 모델을 사용하여 스키마 기반 방식과 seq2seq 생성 방식의 성능 비교를 나타낸다. 이 경우, 동일한 하나의 입력 컨텍스트에 대해 45개의 가능한 모든 슬롯값을 'none'을 포함하여 생성하도록 하는 것보다 단순히 seq2seq 기반으로 슬롯과 슬롯값을 한꺼번에 생성하는 방식의 성능이 좋은 것을 확인할 수 있다. 또한 스키마 방식의 경우 학습 데이터양이 단순 생성 방식에 비해 약 40배 이상 많기 때문에 학습 시간도 상당히 많은 시간이 소요되었다.

표 4. 생성 기반 한국어 대화 상태 추적 모델 비교

Model	WOS	
	JGA (%)	F1 (%)
seq2seq 생성 방식 (본 논문 그림 2 (a))	77.15	96.92
스키마 기반 생성 방식 (본 논문 그림 2 (b))	68.56	84.33

표 5는 동일한 사전학습 언어모델 KeByT5-base 모델을 사용하여 seq2seq 생성 방식으로 실험을 수행하는데, 대화가 진행되면서 생성되는 대화 상태 정보를 입력 컨텍스트에 추가하여 확장하는 경우의 성능을 나타낸다. 참고로, 표 5의 ground truth를 사용한 방식은 추론 과정에서 생성된(generated) 상태(state) 정보가 아니라 정답(reference) 상태 정보를 사용했을 때의 결과를 나타내며, 비교를 위해서 표기하였다. 앞에서의 성능과 비교

했을 때 대화가 진행되면서 생성되는 대화 상태 정보는 오류를 포함하고 있으며, 이러한 오류 있는 상태 정보는 이후의 발화로 계속 전파된다. 따라서 중간 생성 대화 상태 정보를 포함시킬 때의 성능이 포함시키지 않는 때보다 비교하여 성능이 떨어짐을 확인할 수 있다.

표 5. 입력 컨텍스트 확장에 따른 성능 비교

Model	WOS	
	JGA (%)	F1 (%)
seq2seq 생성 방식 (state: ground truth)	82.35	97.66
seq2seq 생성 방식 (state: 생성된 대화 상태 정보)	68.49	95.54

추가적으로 토큰-프리 사전학습 언어모델에서 어휘나 문장의 시퀀스 길이(sequence length)는 토큰 기반 언어모델에 비해 길다. 즉, 한국어 '사랑' 이라는 2음절로 구성된 어휘를 예로 들면 기존 토큰 단위 토크나이저가 해당 어휘를 특별히 분리하여 분석하지 않는 한 길이(length)가 1인 반면 토큰-프리 토크나이저에 의해 분석된 길이는 6으로 정해진다. 따라서 토큰-프리 언어모델을 사용하여 다운스트림 태스크를 미세조정할 때 시퀀스 길이를 증가시켜 주는 것이 필요하다.

6. 결론

한국어 토큰-프리 사전학습 언어모델 KeByT5를 백본 네트워크로 사용하는 한국어 생성 기반 대화 상태 추적 모델에 관한 연구를 수행하였다. 토큰-프리 사전학습 언어모델의 경우 미등록어에 강건한 성능을 보이기 때문에 다운스트림 태스크의 도메인 변화에 좋은 성능을 기대할 수 있다.

대화 상태 추적은 목적 지향 대화에서 사용자 의도를 파악하는 중요한 태스크이다. 본 논문에서는 한국어 토큰-프리 사전학습 언어모델 KeByT5을 사용하여 다양한 실험을 진행하였다. 제안하는 모델은 seq2seq 기반의 복잡하지 않은 구조를 가지고 있으며, KLUE 벤치마크 모델과 비교하여 좋은 성능을 보임을 확인할 수 있었다.

감사의 글

이 논문은 2023년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (RS-2023-00216011, 사람처럼 개념적으로 이해/추론이 가능한 복합인공지능 원천기술 연구)

참고문헌

[1] 권오욱, 홍택규, 황금하, 노윤형, 최승권, 김화연, 김영길, 이윤근, "심층 신경망 기반 대화처리 기술 동향", 전자통신동향분석, 제34권, 제4호, pp.55-64, 2019.

[2] Yang, Yunyi, Yunhao Li, and Xiaojun Quan. "UBAR: Towards fully end-to-end task-oriented dialog system with GPT-2." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 35. No. 16. 2021.

[3] Hosseini-Asl, Ehsan, et al. "A simple language model for task-oriented dialogue." Advances in Neural Information Processing Systems 33, 2020.

[4] S. Park, J. Moon, S. Kim, W. Cho, J. Han, J. Park, C. Song, J. Kim, Y. Song, T. Oh, J. Lee, J. Oh, S. Lyu, Y. Jeong, I. Lee, S. Seo, D. Lee, H. Kim, M. Lee, S. Jang, S. Do, S. Kim, K. Lim, J. Lee, K. Park, J. Shin, S. Kim, L. Park, A. Oh, J. Ha, and K. Cho, "KLUE: Korean Language Understanding Evaluation", arXiv:2105.09680, 2021.

[5] Feng, Yue, Yang Wang, and Hang Li. "A sequence-to-sequence approach to dialogue state tracking." arXiv preprint arXiv:2011.09553 ,2020.

[6] Zhao, Jeffrey, et al. "Effective sequence-to-sequence dialogue state tracking." arXiv preprint arXiv:2108.13990, 2021.

[7] Tian, Xin, et al. "Amendable generation for dialogue state tracking." arXiv preprint arXiv:2110.15659, 2021.

[8] H. Lee, J. Lee, and T. Y. Kim , "SUMBT: Slot-utterance matching for universal and scalable belief tracking," in Proc. Assoc. Comput. Linguist. 2019.

[9] M. Heck et al., "TripPy: A triple copy strategy for value independent neural dialog state tracking," in Proc. Spec. Interest Group Discourse. Dialogue. July. 2020, pp. 35-44.

[10] Kim et al., "A pre-trained BERT for Korean medial natural language processing", Sci Rep 12, 13847 (2022).

[11] Wang, Changhan, Kyunghyun Cho, and Jiatao Gu. "Neural machine translation with byte-level subwords." Proceedings of the AAAI conference on artificial intelligence. Vol. 34. No. 05. 2020.

[12] Lee, Chia-Hsuan, Hao Cheng, and Mari Ostendorf. "Dialogue state tracking with a language model using schema-driven prompting." arXiv preprint arXiv:2109.07506 (2021).

[13] John F Kelley. An iterative design methodology for user-friendly natural language office information applications. ACM Transactions on Information Systems, 2(1):26-41, 1984.