멀티태스크 학습을 이용한 대화 상태 추적 시스템

신창욱 [O†], 장두성[‡], 차정원[†]

창원대학교[†], KT[‡]

papower1@changwon.ac.kr, dschang@kt.com, jcha@changwon.ac.kr

Dialogue State Tracking System using Multitask Learning

Chang-Uk Shin[O†], Du-Seong Chang[‡], Jeong-Won Cha[†]

Changwon National University[†], KT[‡]

요 약

우리는 목적 지향 대화에 대한 대화 상태 추적 시스템을 구현한다. 수행된 대화를 입력으로 받아 사용자의 목표, 개체명 인식, 그리고 시스템 화행 총 3가지 태스크로 멀티태스크 학습을 수행해 모델을 작성한다. RNN과 end-to-end memory network, 그리고 pointer generation 메커니즘을 활용하여 입력된 대화 기록으로부터 멀티태스크 추론을 수행한다. 제안 멀티태스크 구조는 사용자의 목표 추론 정확도 95.45%, 개체명 인식 F1 85.61%, 시스템 화행 추론 정확도 85.45%를 달성하였다. 이는 단일 태스크로 추론을 수행하였을 때의 성능 대비 목표 추론 2.72%p, 시스템 화행 추론 2.27%p 상승하였으나 개체명 인식은 0.63%p 하락한 성능이다. 수행된 실험과 분석을 통해, 대화 상태 추론에서 멀티태스크 추론의 유용함을 보인다.

1. 서 론

사용자와 대화로 의사소통을 수행하여 사용자의 목적을 이해하고, 그 목표를 수행해주는 시스템을 우리는 목적 지향 대화 시스템이라 일컫는다. 위의 대화 시스템을 구현하기 위해서는 사용자의 발화를 이해하여 목적을 알아낼 수 있어야 하고, 그렇게 밝혀낸 목적을 수행할 수 있어야 한다.

대화 상태 추적 시스템은 대화 시스템의 하위 모듈로써, 현재까지 진행된 대화를 정리하여 간결하게 표현할 수 있는 '대화 상태'를 작성하는 데 그 목적이 있다. 대화 상태 추적 시스템에 의해 적절한 대화 상태가 작성된다면, 대화 시스템은 그것에 기반하여 시스템 발화를 작성할 수 있게 된다.

대용량의 데이터셋에 기반하여, 대화 상태 추론을 인공신경망으로 수행하고자 하는 사례가 지속되고 있다. [1]에서는 멀티 도메인에서 도메인 간 전이 상황을 해결하고자 시도하였다. TRADE라 명명한 구조를 제안하고 멀티 도메인 목적 지향 대화 데이터셋인 MultiWOZ[2] 데이터셋에 대해 높은 성능을 달성하였다. 인공신경망 기반 모델을 학습하기 위해서는 대용량의 데이터셋이 필수적인데, 대화 상태 추적을 위해 사용될 수 있는 데이터셋의 경우에도 MultiWOZ[2]나 MultiDoGO[3]와 같은 데이터셋이 공개되어 지속적으로 연구되고 있다.

본 연구에서는 입력된 사용자의 발화와 대화 기록에 대한 대화 상태를 작성하는 대화 상태 추적 시스템의 구현을 다룬다. 본 연구에서는 목적 지향 대화에서의 사용자의 목표, 개체명, 그리고 다음 시스템 발화의 화행의 추론을 대화 상태의 속성으로 정의한다. 우리는 위 제시한 문제를 해결하기 위해 대화 기록이 입력되었을 때 사용자의 목표와 개체명, 시스템 화행을 동시에 추론할 수 있는 구조를 제안한다.

대화 상태의 3가지 속성은 서로 강한 연관 관계에 있다고 판단된다. 따라서, 본 논문에서는 3개의 추론이 하나의 구조에서 수행되도록 멀티태스크 학습을 수행한다. 이는 멀티태스크 학습 방법을 통해 각 태스크로부터 발생하는 유용한 정보가 추론기의 공유 계층 등을 통해 서로 공유되어, 각 태스크의 성능을 개선하는 효과를 낸다는 이전 연구 결과[4]에 근거한다.

2. 제안 방법

본 연구에서는 가정 내 비치된 스마트 TV, 스마트 스피커 도메인의 대화 시나리오를 타겟 시나리오로 설정한다. 정해진 시나리오의 데이터셋이 존재하지 않는 상황을 설정하였다. 따라서, 먼저 학습을 위한 데이터셋을 다음과 같은 방법으로 생성하였다.

첫째로, 모델링을 수행할 도메인과 도메인 별 사용자의 목표, 개체명 카테고리, 화행 카테고리를 설정하였다. 그리고 첫 번째 턴 사용자 발화를 직접 작성하였다[표 1].

이어서, 그렇게 작성된 각 발화에 대하여, 사용자 목표, 개체명, 시스템 화행을 부착하였다. 시스템 발화는 각 시스템 화행에 부합하도록 규칙으로 작성하였다.

이어지는 턴은 다시 위 과정을 반복하여 부착하였다. 사용자의 목표가 수행되고 대화를 종료할 수 있다고

판단될 때까지 위 과정을 반복한다.

표 1 첫번째 턴 사용자 발화의 예

발화	개체명
아이유의 노래를 틀어 줘	Singer:아이유
마동석 영화 재생	Actor:마동석
스케줄 등산 추가해	SchTitle:등산
월요일 스케줄 확인	WeekDay:월요일

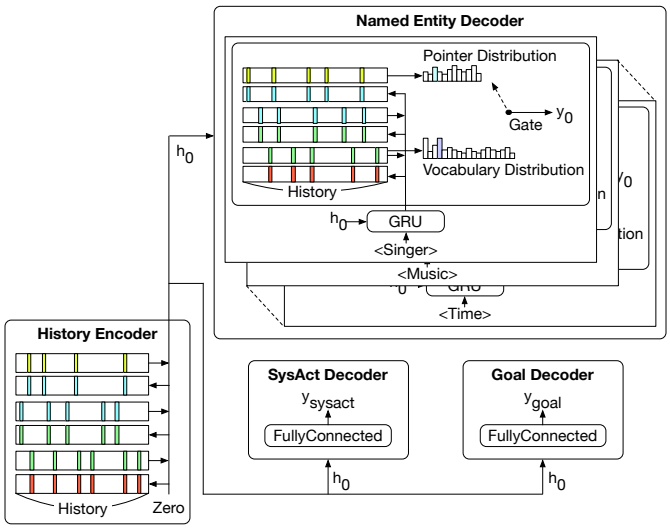

그림 1 제안 인공신경망 구조도

본 연구에서 모델링에 사용하는 인공신경망 구조는 [4]에서 제안된 Mem2Seq의 구조를 변형한 것이다. 본 연구에서 추론하고자 하는 목표는 3가지이고, 개체명의 경우 다시 복수의 카테고리를 가진다. 따라서, 그러한 추론을 수행하기 위해 하나의 인코더에 여러 개의 디코더를 연결하여 멀티태스크 구조를 설정하였다. 아래에 입력된 대화 기록에 대해 추론을 수행하는 과정을 묘사한다.

최초, 입력으로 주어진 대화 기록은 End-to-End Memory Network[5]에 입력되어 하나의 분산 표현으로 인코딩된다. 인코딩이 완료된 분산 표현에 대하여, 총 3개의 디코더가 각각 사용자 목표, 개체명, 시스템 화행을 추론한다.

개체명은 24개의 카테고리 별로 각각 추론되도록 한다. 즉 24개의 Memory Network 디코더가 추론을 수행한다. 음절 단위로 추론되도록 하고, 매 음절 추론마다 pointer generator[6]의 결과와 vocabulary generator의 결과 중 적절한 것을 모델이 선택하도록 한다. pointer generation을 수행하기 위해서는 모든 토큰이 발화 기록에 나타나야 하지만, 그렇지 않은 경우가 있다. 따라서, 만약 추론하고자 하는 음절이 발화 기록에 나타나지 않는 경우, 입력 발화 기록의 첫 번째 토큰이자 특수 토큰인 <NONE> 토큰을 포인팅하도록 하였다. 사용자 목표와 시스템 화행은 입력 발화 기록에 대한 분류 문제로 해결한다. 따라서, 인코딩의 은닉 표현을 단일 계층의 FFNN(Feed-Forward Neural Network)에 입력하여 25개의 사용자 목표와 36개의 화행으로 분류하도록 하였다.

4. 실험

위 기술한 데이터셋 생성 기법으로 작성된 전체 데이터셋의 통계량은 [표 2]와 같다.

표 2 생성된 데이터셋의 통계량

구분	수량	단위
대화의 수	1,938	대화
발화의 수	4,422	발화
대화당 평균 발화의 수	2.282	발화/대화
발화당 부착된 개체명의 수	1.858	개체명/발화

실험에는 전체 대화를 8:1:1로 분할하여 각각 학습, 개발, 평가 데이터셋으로 사용하였다.

Memory Network의 인코더의 hop의 수는 3으로 설정하였다. 모든 은닉 차원과 음절/카테고리 임베딩의 크기는 256으로 설정하였다. Optimizer는 Adam[7], 학습율은 0.001, Dropout은 0.2로 설정하였다.

표 3 태스크별 성능

성능 / 태스크	목표 추론 (Acc)	개체명 인식 (F1)	화행 추론 (Acc)
목표	92.73%	-	-
개체명	-	**86.24%**	-
화행	-	-	83.18%
목표+개체명	92.73%	77.13%	-
목표+화행	94.77%	-	83.41%
개체명+화행	-	71.93%	78.64%
모든 태스크	**95.45% (+2.72%p)**	85.61% (-0.63%p)	**85.45% (+2.27%p)**

표 4 목표+개체명 태스크의 개체명 인식 오류의 예

개체명 인식 오류 발화 기록	오류 개체명
U : "노래 틀어 줘" S : "어떤 노래를 틀어드릴까요?" U : "아이유 노래"	Music : "너와 나"
U : "마동석이 출연하는 컨텐츠"	Actor : "마동석 마동석"
U : "티비 채널 MBC로 변경해"	Channel : "KBS"

[표 3]에 본 논문에서 수행한 실험의 성능을 실었다.

마지막 행의 괄호 내 수치는 단일 태스크 성능 대비 모든 태스크 성능의 상승/하락 수치이다. 실험은 총 3개의 태스크 중 가능한 태스크의 조합을 모두 수행하였다. 목표 추론과 화행 추론 성능은 3개 태스크를 멀티태스크로 학습하였을 때 가장 높은 성능을 달성하였다. 3개의 태스크 중 목표 추론과 화행 추론이 같은 경향을 보이고 개체명 인식만 다른 성능 추이를 보이는 것은 개체명 인식 디코더의 구조가 다른 것, 그리고 개체명 인식 추론만 생성 구조를 채택한 것에서 기인한 것이라 추측된다.

[표 4]에는 '목표+개체명' 태스크에서 평가 데이터셋 내 개체명 인식 오류 중 일부를 정리하였다. '목표+개체명' 태스크와 '개체명+화행' 태스크는 개체명 단일 태스크 대비 개체명 인식 성능이 하락하였기에, 그 원인을 파악하고자 진행한 것이다.

첫번째 샘플은 사용자가 가수 '아이유'의 노래 제목 중 어떠한 것도 발화하지 않았는데 음악의 제목으로써 '너와 나'를 생성하였다. 이는 학습 데이터셋 내에서 단어 '아이유'가 발화된 대화에서는 '너와 나' 또한 높은 확률로 함께 발화되었기 때문에 불필요한 편향이 학습된 것이라 분석된다. 데이터셋을 확장하거나 정제하여 이러한 편향이 학습되지 않도록 개선하여 해결할 수 있다.

두번째 샘플은 배우의 이름인 단어 '마동석'을 2회 연달아 생성하여 오류가 된 샘플이다. 학습 데이터셋에 둘 이상의 배우의 이름이 함께 나타나는 샘플이 일부 포함되어 있다. 그러한 샘플로부터 배우의 이름은 2회 연달아 생성한다는 잘못된 편향이 학습된 것이라 분석된다.

세번째 샘플은 오류로써 같은 카테고리의 다른 개체명 어휘를 생성한 사례이다. 세번째 샘플에서 실제 사용자가 발화한 채널은 MBC인데 시스템이 KBS를 생성한 것은 학습 데이터셋 내 특정 채널의 분포와 무관하지 않아 보인다. 학습 데이터셋 내에서 채널 개체명으로써 KBS의 출현한 횟수는 38회, MBC의 출현 횟수는 13회였다.

5. 결론

본 논문에서는 멀티태스크 학습을 통해 대화 상태 추론을 수행하는 시스템을 작성 및 제안하였다. 제안 구조는 현재까지 수행된 대화 기록을 입력으로 하여, 발화 내 개체명 인식, 마지막 발화 기준 사용자의 목표 추론, 그리고 다음 시스템 발화의 화행을 추론한다. 제안 구조로 수행한 실험에서 단일 태스크 실험 대비 사용자 목표 정확도 2.72%p, 시스템 화행 정확도 2.27%p 향상된 성능을 달성할 수 있었다. 수행된 실험을 통해, 목표 추론과 화행 추론은 비교적 비슷한 성능 추이를 보임을 알 수 있었고, 이러한 양상은 두 태스크의 추론 모듈의 구조의 유사성에 기인하는 것이라 추측된다.

본 연구에서는 대화 데이터셋이 주어지지 않은 상황에서 실제 대화를 모조 생성하고, 그렇게 생성된 데이터셋으로 모델링을 수행하였다. 실제 환경의 대화를 생성하는 작업은 상당한 노력과 시간이 소요되어 대용량의 다중 도메인의 모델을 학습하는 데에는 제한이 있다.

참고 문헌

[1] Wu, Chien-Sheng, Andrea Madotto, Ehsan Hosseini-Asl, Caiming Xiong, Richard Socher, and Pascale Fung. "Transferable multi-domain state generator for task-oriented dialogue systems." arXiv preprint arXiv:1905.08743. 2019.

[2] Budzianowski, Paweł, Tsung-Hsien Wen, Bo-Hsiang Tseng, Inigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. "Multiwoz-a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling." arXiv preprint arXiv:1810.00278. 2018.

[3] Peskov, Denis, Nancy Clarke, Jason Krone, Brigi Fodor, Yi Zhang, Adel Youssef, and Mona Diab. "Multi-Domain Goal-Oriented Dialogues (MultiDoGO): Strategies toward Curating and Annotating Large Scale Dialogue Data." In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 4518-4528. 2019.

[4] Madotto, Andrea, Chien-Sheng Wu, and Pascale Fung. "Mem2seq: Effectively incorporating knowledge bases into end-to-end task-oriented dialog systems." arXiv preprint arXiv:1804.08217. 2018.

[5] Sukhbaatar, Sainbayar, Jason Weston, and Rob Fergus. "End-to-end memory networks." In Advances in neural information processing systems, pp. 2440-2448. 2015.

[6] See, Abigail, Peter J. Liu, and Christopher D. Manning. "Get to the point: Summarization with pointer-generator networks." arXiv preprint arXiv:1704.04368. 2017.

[7] Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." arXiv preprint arXiv:1412.6980. 2014.