마르코프 의사결정 과정에 기반한 대화 관리자 설계

최 준 기, 은 지 현, 장 두 성, 김 현 정, 구 명 완
KT 미래기술연구소 HCI 연구 담당

Design of Markov Decision Process Based Dialogue Manager

Joon Ki Choi, Jihyun Eun, Du-Seong Chang, Hyun Jeong Kim, Myong-Wan Koo
Human Centric Interface Research Department
Advanced Technology Laboratory, KT
E-mail : {jkchoi, jh06, dschang, hyunj, mwkoo}@kt.co.kr

Abstract

The role of dialogue manager is to select proper actions based on observed environment and inferred user intention. This paper presents stochastic model for dialogue manager based on Markov decision process. To build a mixed initiative dialogue manager, we used accumulated user utterance, previous act of dialogue manager, and domain dependent knowledge as the input to the MDP. We also used dialogue corpus to train the automatically optimized policy of MDP with reinforcement learning algorithm. The states which have unique and intuitive actions were removed from the design of MDP by using the domain knowledge. The design of dialogue manager included the usage of natural language understanding and response generator to build short message based remote control of home networked appliances.

I. 서론

자연언어 대화는 가장 자연스럽고 효과적인 인간과 컴퓨터의 인터페이스로 사용될 수 있다. 최근에는 컴퓨터 이외에 차세대 통신, 주문자형 방송, 홈 네트워크, 지능형 로봇 등 많은 정보를 포함하고 복잡한 인터페이스를 요구하는 장비와 서비스의 증가로 인하여 자연언어 대화 인터페이스의 수요가 늘어나고 있으며 이와 관련된 많은 연구가 진행되고 있다.

일반적으로 자연언어 대화 인터페이스는 크게 세 종류의 모듈로 구성된다. 먼저 음성, 텍스트, 터치스크린 입력 등 멀티모달 기법으로 입력된 자연언어 문장을 이해하는 언어 이해 모듈과 자연언어로 된 응답을 생성하는 응답 생성 모듈, 마지막으로 언어 이해 모듈의 결과를 입력으로 받아서 시스템의 동작과 생성될 응답의 내용을 결정하는 대화 관리자 모듈이 있다.

성공적인 대화 관리자는 사용자와 시스템 간의 자연스러운 혼합 주도형 대화를 유도하며 사용자가 의도한 목적을 정확하게 수행할 수 있도록 한다. 기존의 방법으로는 대화 흐름을 유한 상태 네트워크(finite state network)로 표현하는 방법이 널리 사용되었다[1]. 이 방법은 문제 해결 영역(problem solving domain)의 대화를 처리하기 위해 정확하고 빠르게 설계 될 수 있으나 자연스러운 사용자의 발화가 불가능하며 항상 고정된 시나리오를 따라야 하는 단점이 있었다. 따라서 문제 해결 영역 대화 시스템의 정확한 동작 수행과 사용자의 자연스러운 발화를 가능하게 하기 위하여 마르코프 의사 결정과정 (MDP, Markov Decision Process)에 기반한 대화 관리자가 제안되었다[3]. 이 방법은 사용자의 발화와 그에 대응하는 시스템 동작으로 구성된 대화 코퍼스로부터 훈련된 정책을 선택하여 대화 시스템의 동작과 응답을 결정한다.

본 논문에서는 MDP를 이용한 홈 네트워크의 원격 제어를 위한 대화 관리자를 설계안을 제시하며, 향후 연구 방향을 제시한다.

II. 마르코프 의사결정 기반 대화 관리자

2.1 마르코프 의사결정 과정

MDP는 실세계의 환경을 모델링 하는 유한 개수의 상태(state) 들의 집합과 각 상태간의 천이 확률(transition probability), 각 상태에 따라서 MDP가 미리 학습된 정책에 따라 취하는 행동(action)과 MDP가 수행한 행동에 대하여 환경이 부여하는 보상함수(reward function)으로 구성된다. 이를 식으로 표현하면 MDP는 (S, A, T, R)으로 정의되며 각 기호의 정의는 다음과 같다.

- S: 유한 개수의 상태 s_i들의 집합
- A: 유한 개수의 행동 a_i들의 집합
- T: $T(s_{i+1}, a, s_i) = p(s_{i+1}|s, a)$
- R: $S \times A \mapsto R$,

MDP의 행동을 결정하는 정책(policy)은 강화 학습(reinforcement learning) 방법에 의하여 학습된다[4]. 강화 학습을 통해서 MDP는 대화 코퍼스로부터 각 상태의 최적 행동을 결정함으로써 사용자가 원하는 목적에 도달할 수 있는 정책을 학습한다. 강화 학습은 누적 보상(cumulative reward)을 최대화 하도록 진행된다. 강화 학습은 교사 학습(supervised learning)과 비교사 학습(unsupervised learning)의 중간적인 특성을 띄고 있으며 매 단계의 행동에 대한 보상으로서 평가를 받는다. 이 때 행동의 단계가 정해지지 않은 경우 누적 보상함수는 다음과 같이 표현된다.

$$R_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}$$

위의 식에서 γ는 할인 상수(discount factor)로서 미래에 받게 될 보상이 현재 상태의 가치나 상태-행동의 가치에 반영되는 정도를 조절한다. 즉 γ의 값이 1에 가까울수록 t 시각 이유에 받게 될 보상을 할인하기 않고 반영하게 된다.

MDP는 환경의 특정한 상태 s에 대해서 모든 완벽한 정보를 요구한다. 즉, 환경이 변하였을 때 새로운 상태로의 사상이 명확해야 한다. 따라서 일반적인 대화 관리자의 경우 모든 환경을 MDP의 상태로 표현하기 위해서는 매우 많은 상태들이 요구된다. 많은 상태는 강화 학습을 어렵게 하며 훈련 데이터베이스의 수집이 대규모로 이루어져야 하는 단점이 있다[5]. 또한 환경의 입력, 즉 대화 인터페이스의 경우 사용자의 입력을 시스템이 받아들인 뒤에 이 입력을 대화 관리자가 완전하게 믿을 수 없는 문제가 발생한다. 예를 들어 텍스트 기반의 자연언어를 사용하는 경우 자연언어 이해 모듈의 오류가 생길 수 있으며 음성을 사용자 인터페이스로 사용하는 경우에는 음성인식의 오류가 발생할 수 있다. 이렇게 MDP의 관측을 부분적으로 믿을 수 없는 문제를 해결하기 위하여 부분 관측 MDP (POMDP; Partially Observable MDP)를 사용하는 방법이 제안되어 대화 관리 시스템, 특히 음성 언어 대화 관리 시스템에서 활용되고 있다[2]. 그러나 POMDP 기반의 시스템의 경우 학습 알고리즘이 매우 복잡하다는 단점이 있다.

본 논문에서 설계하는 대화 관리자는 사용자가 입력할 수 있는 환경을 전부 상태로 모델링하지 않고 그룹화하여 사용하는 방법을 제안하였으며, 사용자 입력의 신뢰도를 MDP의 상태 정의 변수로 사용하기 위하여 양자화 하는 기법을 사용하였다.

2.2 기존의 MDP 기반 대화 관리자

기존의 MDP를 기반하는 대화 관리자들은 실버 로봇의 대화 관리자, 비행기 여행 예약을 위한 대화 관리자들이 연구되었다[2][3]. 또한 상태 수가 많아지는 문제를 해결하기 위해서 NJFun 시스템에서는 특정 상태에서 단 한 개의 동작만이 가능하다면 해당 상태는 모델링에서 제외하였고, 시스템의 행동이 여러 가지 가능하여 행동 선택을 해야 하는 상태만을 MDP에서 다루는 방법을 사용하였다[5]. 본 논문에서도 홈 네트워크 제어 영역 코퍼스를 분석하여 시스템의 행동이 규칙으로 제한될 수 있는 상태는 MDP 모델링에서 제거하고 모든 시스템의 행동에 다양성이 존재하는 상태만을 남겨두어 상태의 수를 적게 유지하면서 자연스러운 혼합 주도 대화를 유도하였다.

2.3 홈 네트워크 제어를 위한 대화 관리자

본 논문에서 제안하고자 하는 홈 네트워크 원격제어를 위한 대화 관리자의 구조도는 그림 1과 같다. 대화 관리자의 입력으로는 언어 이해 결과를 사용하고 이후에 영역 지식 규칙을 사용하여 입력된 자연 언어 이해 결과를 확인한다[6]. 영역 지식 규칙은 대화 관리자가 처리할 수 없는 문형의 문장을 입력에서 제거하거나 목적 시스템의 영역 밖에 있는 문장들을 입력에서 제거하고, 사용자가 정의한 영역 규칙들을 적용하는 역할을 한다.

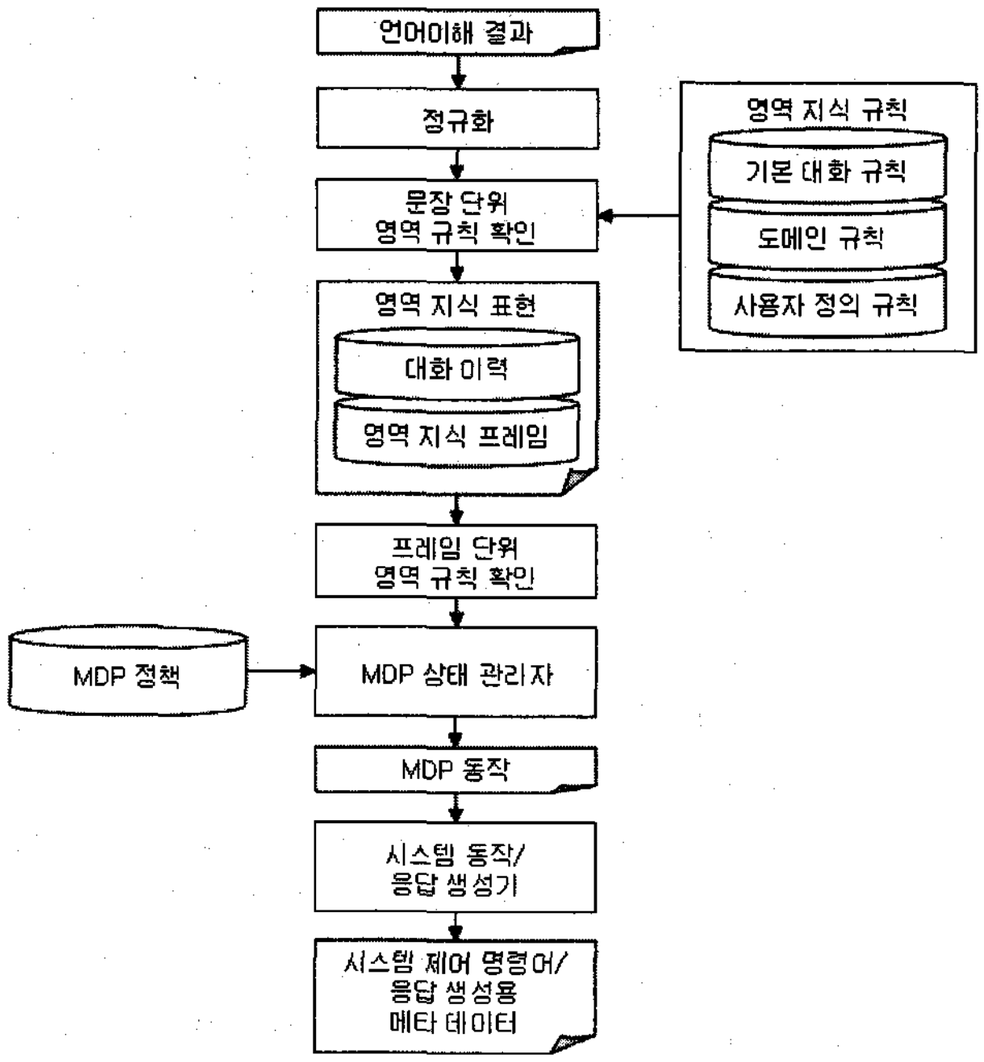


그림 1 홈 네트워크 원격 제어를 위한 대화 관리자의 구조도

영역 규칙에 적합한 문장은 MDP의 상태로 사상되고 학습된 MDP 정책을 이용하여 각 상태에서 적당한 행동을 결정한 뒤 최종적으로 응답이나 시스템 제어문 생성을 위한 화행과 기타 영역 정보를 생성한다. 다음 장에서는 홈 네트워크 제어를 위한 MDP의 상태, 행동, 그리고 정책의 설계에 대하여 설명한다.

III. 마르코프 결정과정의 설계

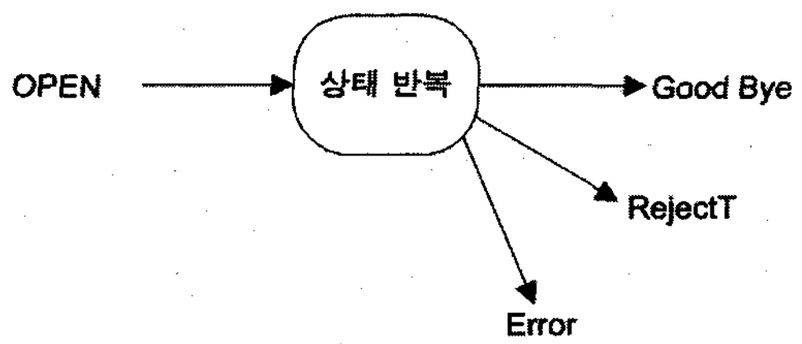
3.1 마르코프 결정과정의 행동

홈 네트워크 제어나 로봇의 제어, 비행기 여행 계획 예약 등 문제 해결 영역의 대화들은 그 목적이 분명하며 시스템이 명확한 행동을 하기 위해서 사용자로부터 얻어야 하는 정보를 미리 예상할 수 있다. 시스템이 원하는 정보를 사용자로부터 자연스럽게 얻어내기 위하여 시스템의 행동은 다음의 표와 같이 대화의 주도 전환에 관련된 행동으로 구성한다.

표 1 MDP의 행동 정의

Action	종류	설명 및 발화 예제
Open		대화시작 “안녕하세요, 홈 네트워크 시스템입니다”
Close		대화종료 "감사합니다, 좋은 하루 되세요"
Specify	D	사용자의 이전 발화보다 자세한 설명을 요구
		"몇 시에 어떤 세탁코스로 예약할까요?"
	R	"예약 세탁에 관한 정보를 더 말씀해 주세요"
Confirm	D	사용자 발화에 대한 확인 과정 "7월 20일 오후 11시 취침모드로 에어컨 가동을 예약하시겠습니까?
	R	"에어컨 가동을 예약하시겠습니까?"
Relax	D	명령어의 속성조건 완화요구 "제습모드는 세탁기 명령어에 적합하지 않습니다. 세탁코스는 울코스, 표준코스, 담요코스가 있습니다. 다시 명령해 주세요"
	R	"세탁기 동작으로 적합하지 않습니다. 다시 명령해 주세요"
Execute		목적 시스템에 대한 직접적인 기기 제어 명령을 수행, 발화 없음
Result		Execute가 정상 수행되었을 때 결과 안내 "세탁기가 정상적으로 예약되었습니다"
Fail		Execute가 비정상적으로 수행되었을 경우의 결과 안내. "홈 네트워크 제어에 실패하였습니다, 오류의 원인은 ***입니다."
Reject	1	이전 발화 1개 만을 오류로 인정. "한 번에 하나의 가전에 대한 명령을 주세요, 세탁기를 예약할까요"
	T	Session 시작 후 이루어진 모든 발화에 대한 내용을 오류로 인정 "명령을 수행할 수 없습니다. 처음부터 명령을 주세요"
Cancel		사용자의 취소 발언에 대한 응답. "명령이 취소되었습니다." "취침모드가 취소되었습니다"

표 1에서 각 행동의 종류 중 D는 사용자에게 가능한 입력을 자세히 설명하여 시스템 주도 대화를 유도하며, 반대로 R은 사용자 주도 대화를 유도한다. MDP의 정책 그래프(policy graph)는 그림 2와 같이 표현 가능하다. 그림 2에서 볼 수 있듯이 MDP의 행동은 일반적인 대화 시스템에서 사용될 수 있는 행동들로 구성되어있으며 영역 의존 정보들은 그림 1에서 표현되어 있는 프레임 단위의 영역 지식과 관련 규칙으로 다루도록 설계하였다. 이와 같은 구조는 다른 영역으로의 대화 관리자의 확장을 손쉽게 할 수 있다.



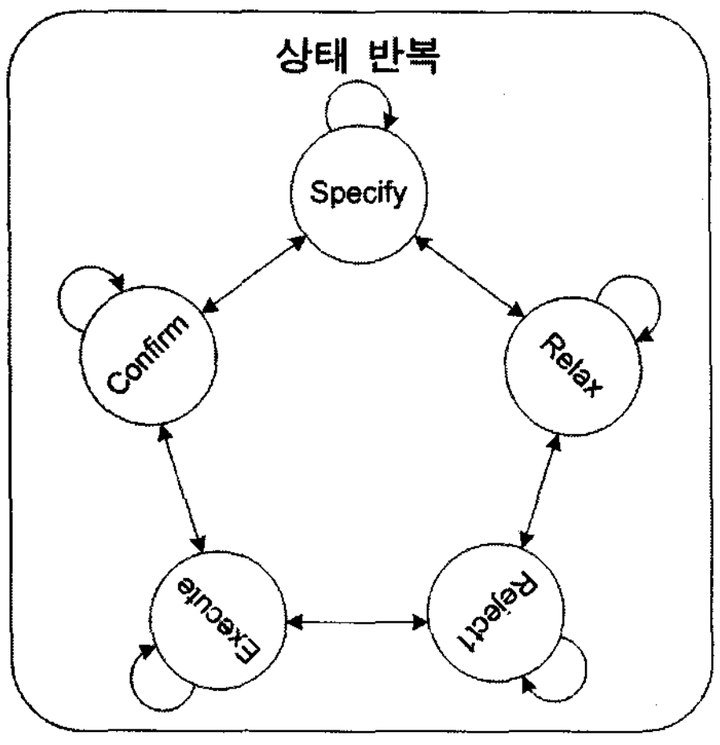


그림 2 MDP의 정책 그래프

MDP의 상태는 혼합 주도 대화 진행에 영향을 미칠 수 있는 요인들을 분류하여 설계하였으며 하나의 환경 변수를 하나의 코드에 할당하여 MDP 상태 이름을 모델링하고자 하는 환경 코드의 조합으로 표현하였다. 상태 이름의 각 비트에 대한 상태 정의 변수의 설명은 아래의 표 2와 같다.

표 2 MDP 상태 이름의 환경 변수 설명

비트	MDP 상태 정의 변수	설명
0	가전제품종류	제어하고자 하는 가전기기의 종류
1	사용자명령의 종류	제어 명령의 종류
2	사용자명령의 속성 채워짐	제어에 사용될 수 있는 속성들이 사용자 제어 명령에 포함된 정도. 가전기기의 위치, 예약 시간, 온도, 모드 등과 같은 속성의 제어문 포함 여부
3	실패반복횟수	사용자가 원하는 하나의 동작을 수행하는 과정 중에 실패한 횟수
4	대화 관리자의 이전 행동	대화 관리자가 이전에 최종적으로 수했던 행동의 종류
5	신뢰도	사용자의 제어 입력에 대한 신뢰도. 자연언어 이해 모듈의 신뢰도 사용 가능

표 2에서 신뢰도 상태 정의 변수는 언어 이해 모듈의 점수를 양자화하여 사용한다. 즉, 언어 이해 모듈의 신뢰도를 미리 학습 자료를 이용하여 구하고, 적절한 신뢰 구간을 학습하여 사용한다.

상태간 천이 확률은 WOZ (Wizard of Oz) 방식으로 수집된 대화 코퍼스에서 구해진다[6]. MDP의 보상함수는 대화의 성공적으로 진행이 되었음을 확인하기 위하여 사용자가 원하는 목적을 이루었을 경우는 보상으로 1을 부여하고 이루지 못한 경우에는 -1을 부여하였다. 그리고 사용자 만족도를 평가하기 위해서 여러 정책들을 사용하여 응답을 생성하였으며 사용자들로 하여금 응답에 대한 만족도를 5단계로 나누어서 평가하게 하였다. 사용자 만족도와 목적 달성 점수를 강화학습의 보상 함수로 사용하였다. 현재 구축 중인 대화 코퍼스의 일부를 사용하여 학습한 결과 사용자가 수동으로 작성한 정책과 유사하게 학습되는 것을 알 수 있었다. 그리고 학습 코퍼스의 확장을 통해서 대화 관리자의 성능이 향상됨이 보고된 바 있다[3].

Ⅳ. 토의 및 향후 연구

본 논문에서는 홈 네트워크 제어 영역을 위한 MDP에 기반한 대화 관리자를 설계하였다. 현재 MDP 훈련을 위한 대화 코퍼스를 수집 중이며, 일부 데이터를 활용하여 초기 버전을 작성하였다. 본 논문에서는 모든 환경 변수를 상태로 표시하는 대신에 환경 변수의 분류를 통하여 많은 상태를 훈련시켜야 하는 문제를 해결하였으며, 영역 규칙의 적용에 따라서 하나의 행동만이 가능한 상태는 모델링에서 제거하였다. 영역 규칙의 적용은 사용자 오류에 대한 예외 행동의 구현을 빠르게 할 수 있다. 또한 양자화된 신뢰도를 상태 정의 변수에 추가시킴으로서 멀티모달 입력의 부정확성에 대비하였다.

향후 연구는 먼저 훈련된 대화 관리자의 성능 평가를 위하여 PARADISE 성능평가[7]와 같은 방법을 사용하여 사용자 만족도 측면과 시스템 동작 수행 정확성에 관련하여 평가를 진행할 예정이며, 또한 사람이 작성한 정책과 직접 비교하는 평가 역시 진행할 예정이다. 이러한 평가 결과를 바탕으로 MDP 상태 정의 변수와 보상함수의 정교화 작업이 수행될 것이다. 또한 작성된 대화 관리자를 사용하여 새롭게 사용자 모델 시뮬레이션을 수행하고, 이를 바탕으로 다시 MDP를 재훈련하는 작업으로 대화 관리자의 성능을 향상시키는 대화 관리자의 안정화 작업을 진행하고자 한다..

안정화 작업 이후에는 양자화된 신뢰도 구간 대신 연속적인 신뢰도를 환경 변수로 사용하고 최적의 신뢰도 경계를 찾을 수 있는 POMDP 기반의 추론 알고리즘으로 확장하고자 하며, 강화된 MDP (augmented MDP)[8]나 구조화된 POMDP (structured MDP)[2]와 같이 학습속도가 개선된 POMDP 기반의 추론 알고리즘으로의 확장 작업을 하여 멀티 모달 인식기와 강하게 결합된 대화 관리자를 개발하고자 한다.

참고문헌

[1] P.A. Heeman, et al, "Beyond Structured Dialogues: Factoring Out Grounding", Proc. ICSLP, 1998

[2] J. Pineau, "Tractable Planning under Uncertainty: Exploiting Structure", Ph.D. Thesis, CMU, 2004

[3] E. Levin, et al, "Using Markov decision processes for learning dialogue strategies", IEEE Trans. on Speech and Audio Processing, vol. 8, pp 11-23, 1998

[4] R.S. Sutton and A. Barto, "Reinforcement Learning: An Introduction", MIT Press, 1998

[5] S. Singh, et al., "Optimizing Dialogue Management with Reinforcement Learning: Experiments with the NJFun System", Journal of Artificial Intelligence Research", Vol. 16, pp 105-133, 2002

[6] 김현정, 은지현, 장두성, 최준기, 구명완, "홈네트워크 제어를 위한 대화관리시스템 설계", 대한음성과학회 추계학술대회 논문집, 2006

[7] M. A. Walker, et al., "PARADISE: A Framework for Evaluating Spoken Dialogue Agents", In Proc. ACL/EACL pp 271-180, San Francisco, 1997

[8] R. Becker, et al., "Solving Transition-Independent Decentralized Markov Decision Process", Journal of Artificial Intelligence Research, Vol. 22, pp 423-455, 2004