강화 학습법을 이용한 효과적인 적응형 대화 전략
(An Effective Adaptive Dialogue Strategy Using Reinforcement Learning)

김 원 일 † 고 영 중 †† 서 정 연 †††
(Wonil Kim) (Youngjoong Ko) (Jungyun Seo)

요 약 인간은 다른 사람과 대화할 때, 시행착오 과정을 거치면서 상대방에 관한 학습이 일어난다. 본 논문에서는 이런 과정의 강화학습법(Reinforcement Learning)을 이용하여 대화시스템에 적응형 능력의 부여 방법을 제안한다. 적응형 대화 전략이란 대화시스템이 사용자의 대화 처리 습성을 학습하여, 사용자 만족도와 효율성을 높이는 것을 말한다. 강화 학습법을 효율적으로 대화처리 시스템에 적용하기 위하여 대화를 주대화와 부대화로 나누어 정의하고 사용하였다. 주대화에서는 전체적인 만족도를, 부대화에서는 완료 여부, 완료시간, 에러 횟수를 이용해서 시스템의 효율성을 측정하였다. 또한 학습 과정에서의 사용자 편의성을 위하여 시스템 사용 역량에 따라 사용자를 두 그룹으로 분류한 후 해당 그룹의 강화 학습 훈련 정책을 적용하였다. 실험에서는 개인별, 그룹별 강화 학습에 따라 제안한 방법의 성능을 평가하였다.

키워드 : 대화 시스템, 적응형 대화 전략, 강화 학습, 주대화와 부대화, Q-학습법

Abstract In this paper, we propose a method to enhance adaptability in a dialogue system using the reinforcement learning that reduces response errors by trials and error-search similar to a human dialogue process . The adaptive dialogue strategy means that the dialogue system improves users' satisfaction and dialogue efficiency by learning users' dialogue styles. To apply the reinforcement learning to the dialogue system, we use a main-dialogue span and sub-dialogue spans as the mathematic application units, and evaluate system usability by using features; success or failure, completion time, and error rate in sub-dialogue and the satisfaction in main-dialogue. In addition, we classify users' groups into beginners and experts to increase users' convenience in training steps. Then, we apply reinforcement learning policies according to users' groups. In the experiments, we evaluated the performance of the proposed method on the individual reinforcement learning policy and group's reinforcement learning policy.

Key words : Dialogue System, Adaptive Dialogue Strategy, Reinforcement Learning, Main-dialogue and Sub-dialogue, Q-learning

· 이 연구(논문)은 산업자원부 지원으로 수행하는 21세기 프론티어 연구개발사업(인간기능 생활지원 지능로봇 기술개발사업)의 일환으로 수행되었습니다.

† 정 회 원 : 삼성전자 영상디스플레이 사업부 연구원
oneill.kim@samsung.com
†† 종신회원 : 동아대학교 컴퓨터공학과 교수
yjko@dau.ac.kr
††† 종신회원 : 서강대학교 컴퓨터학과 교수
seojy@sogang.ac.kr

논문접수 : 2005년 11월 28일
심사완료 : 2007년 12월 14일

정보과학회논문지: 소프트웨어 및 응용 제35권 제1호(2008.1)

1. 서 론

대화 처리 시스템(Dialogue System)의 목표는 사람들간의 자연스러운 대화처럼 사람과 컴퓨터 사이의 의사소통을 수행하는 시스템을 만드는 것이다[1]. 하지만 현재의 대화처리 시스템은 한계적인 대화처리 기술로, 사용자 만족감과 시스템 효율성이 인간의 자연스러움을 충족시키기에는 거리가 멀다. 따라서 사용자 만족도를 높이기 위해서는 대화 처리 시스템은 적용 가능한 기술 범위 안에서 유연하고 적응력 높은 대화 전략을 가질 필요가 있다. 시스템이 각 개인의 습성을 처리 할 수 있는 적응형 능력(adaptability)을 가진 대화 전략을 사용하게 되면 사용자의 질의를 유연하게 처리하고 사용자와도 더 자연스러운 의사소통을 할 수 있기 때문이다.

S1: TV 프로그램 안내 시스템입니다. 무엇을 도와 드릴까요?
U1: 8시 30분에 하는 TV 프로그램을 보여줘.
S2: 8시 30분에 하는 TV 프로그램을 선택하셨습니다.
U2: 어떤 장르의 프로그램을 선택하시겠습니까?

그림 1 대화 처리 시스템의 적응형 능력 예제

그림 1은 TV 프로그램을 안내하는 대화 처리 시스템의 적응형 능력에 관한 간단한 예제이다. "S1"은 시스템이 대화 시작을 위해서 사용자에게 하는 발화이다. 사용자는 "S1" 발화를 듣고 "U1"의 발화로 시스템에게 지시하게 되는데, 이는 사용자에 따라서 다르게 반응할 수 있다. 만약 시스템에 능숙한 사용자이거나, 대화 처리 시스템의 입력부가 제대로 작동되고 있을 경우는 이어 나오는 시스템의 "S2" 발화는 번거로울 것이다. 하지만 시스템 사용에 서툰 초보자이거나, 주변 환경 때문에 입력을 인식하기 어려운 환경이라면 현재 진행 상태 파악을 위해서 "S2" 발화는 필수적이다. 이렇게 사용자 응답을 듣고 시스템이 반응하는 "S2" 발화를 각 상황에 맞게 자연스럽게 처리하려면, 대화 처리 시스템은 적응형 능력을 가지고 발화해야 한다. 즉 적응형 능력이란 시스템에게 스스로 사용자의 현재 상태가 어떤지를 파악하고 그에 맞추어서 스스로를 변화하게 하는 지능을 부여하는 것이다.

본 연구에서는 대화 처리 시스템이 스스로 현재 상태를 파악하고, 시스템과 사용자 경험이 바탕이 되어서 사용자 선호도에 따른 시스템 진화 방법에 관하여 다룰 것이다. 이것은 시스템이 미리 발생할 수 있는 시스템 오류나 사용자의 불만족을 미리 예방하므로 사용자 만족도를 높이고 시스템 효율성을 높여줄 수 있다.

2. 관련 연구

사용자에게 맞춤형으로 대화 인터페이스를 제공해주는 방법을 사용자 모델링이라고 한다[2]. 자연어 처리 시스템을 위한 사용자 모델링은 사용자의 언어적 능력과 관심 범위에 맞추어서 시스템의 사용성을 높일 수 있는 언어 인터페이스를 제공하여 주는 것이다. 하지만 기존에 연구된 자연어 처리의 사용자 모델링은 대부분 파싱(Parsing), 대화 처리 메커니즘, 화용론 처리와 같은 특정 분야에 대하여 중점적으로 연구되고 있으며, 자연어 처리 시스템과 같이 언어 처리 전반을 다루는 시스템에서 어떻게 사용자 모델링을 적용해야 할 지에 관한 연구는 없었다[2]. 초기의 연구로 Wallis와 Shortliffe[3]는 문제의 주제와 선호도에 따라서 설명의 수준을 달리하여서 생성하도록 하는 자연어 시스템을 개발하였다. 이것은 초보자가 아닌 전문가 수준만을 고려한 한 가지 특성을 위한 사용자 모델링이기 때문에 미흡한 수준으로 평가된다[3]. McTear[4]는 대용량의 대화 말뭉치를 이용하여서 자동적으로 사용자 모델을 얻을 수 있는 기계 학습법과 추론을 이용한 방법을 제시하고 있다. 한편 Walker[5]는 음성 대화 처리 시스템의 평가 방법을 제안하였는데, 이것은 대화 시스템의 사용자 모델링을 올바르게 평가하는 방법을 포함하고 있다. Jokinen과 Kanto[6]은 사용자의 역량(Expertise)에 따른 사용자 모델링을 하고 이에 따른 음성 기반 전자 메일 시스템을 사용하게 하는 것을 제안하였다. 이 시스템은 3단계로 사용자의 사용 역량을 분류하고, 사용자에게 미리 정해진 분류 기준에 따라서 사용하게 하였다. Möller[7]는 음성대화처리 시스템의 품질을 결정짓는 요소를 세분화된 계통도를 이용하여서 분류하였으며, 이들에 관한 상관관계를 정리하였다.

Litman과 Pan[8]은 시스템 스스로 대화 전략이 변화할 수 있는 적응형 대화 처리 시스템에 관하여 다루었다. 이 적응형 대화 처리 시스템은 사용자의 발화 음성 인식률이 낮아질 경우 이를 재확인하는 발화에서 적응형 능력이 작동되었다. 만약 시스템이 음성 인식률을 낮은 것으로 판단한다면, 시스템의 대화 전략을 시스템-사용자 혼합 주도형으로 바꾸어서 대화를 진행하게 된다. 한편 국내에서 연구된 적응형 대화 처리 시스템으로는 은지현[9]이 마코프 의사 결정을 이용한 홈 네트워크 기기 제어 영역 대화 관리자를 적용한 시스템을 개발하였으며, 이에 대한 사용자 평가는 아직 수행되지 않았다.

Finnish Interact Project[10]에서는 사용자를 그룹별로 나누고, 각 그룹에 따라서 시스템의 언어 생성 부분에 적응형 능력의 학습을 가능하도록 하였다. Walker [11]는 각 발화를 적용대상으로 대화 처리 시스템에 강화학습 알고리즘을 이용하여서 적응형 능력을 부여시켰다. 하지만 대화의 경우 각각의 발화가 다음 발화에 반드시 영향을 미치는 것이 아니기 때문에 강화 학습의 적용 대상은 앞의 사건이 다음 사건에 영향을 미쳐야 속성에 만족시키지 않을 수 있다. 본 연구에서는 강화 학습에의 적용 사건은 부대와와 주대화로 나누어서 대응하였으며 각 부대화는 각각의 단계로서 일련의 단계를 거치게 된다. 또한 Finnish Interact Project[10]의 제안처럼 사용자의 역량에 따라 다르게 대화 처리 시스템을 학습시켰다. 이에 따라 본 연구의 대화 처리 시스템은 사용자 역량에 따른 주대화와 부대화를 대상으로 강화 학습을 하게 되고, 이 학습 결과를 이용해서 사용자 최적화된 모델링을 얻는 방법을 사용한다.

3. 강화 학습법을 이용한 대화 전략

3.1 강화 학습(Reinforcement Learning)

강화 학습법은 에이전트가, 주어진 환경에 관한 미리 설정된 모델 없이, 보상값(Reward)과 행동(Action)의 상호 작용을 통해서 학습이 일어나는 기계 학습법이다. 강화학습은 다른 기계 학습법과 비교하여서 1) 상호작용을 통한 학습 2) 목적 지향적 학습 3) 외부환경과의 교류를 통한 학습 4) 각 상황에서 보상을 최대화시키는 행동을 할 수 있게 하는 학습의 특성을 가진다.

강화학습은 연속적으로 일어나는 의사 결정 문제를 풀 때, 매 단계마다의 효용성 계산을 반복하면서 수렴하면 최종 목표인 최적화된 해를 구할 수 있다는 벨만 방정식(Bellman Equation)을 이용한 것이다. 여기서 에이전트는 하나의 상태가 다음의 상태에 영향을 미친다는 마코프 속성(Markov Property)을 만족시켜야 한다[12]. 에이전트는 최상의 정책을 얻기 위해 주어진 환경과의 상호작용을 하면서 얻게 되는 시행착오의 경험을 이용한 확률적인 수행과정을 거친다. 이런 특성으로 대화 처리 시스템에 강화학습을 이용하여 적응형 능력을 부여하면, 사용자의 사전 정보 없이도 대화 과정을 거치면서 학습이 가능해진다는 장점을 가지게 된다. 강화학습은 그림 2와 같이 표현할 수 있다[12].

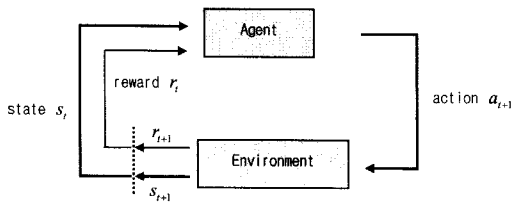

그림 2 에이전트와 외부 환경과의 상호 작용

그림 2는 어떤 시간에서 앞의 행동에서 얻은 보상값(Reward r_t), 현재 시간에서 관측된 상태(State s_t), 다음 시간에서 취하게 되는 행동(Action a_{t+1}), 다음 시간에서 얻게 되는 보상값(r_{t+1}), 다음 시간 단계의 상태(s_{t+1})의 구성되는 요소로 구성되어 있다. 에이전트가 앞의 상태에서 특정 행동을 수행한 결과 현재 상태 s_t로 전이된 상태라고 할 때, 그 행동에 대한 보상값으로 r_t를 받고 있다. 이 r_t는 에이전트의 행동을 결정짓는 요소가 되어서 에이전트의 행동 정책을 강화(Reinforcement)시킨다. 에이전트는 현재의 상태에서 가장 최적의 행동(a_{t+1})을 파악한 후 실행하게 된다. 이 행동은 외부 환경과 반응하여서 다음 상태 s_{t+1}와 보상값 r_{t+1}을 얻게 된다. 다시 이 보상값은 에이전트의 행동 정책을 강화시키게 된다. 이런 에이전트와 외부 환경과의 상호작용은 일련의 최적화된 행동 정책을 얻게 될 때까지 반복하게 되는 것이다. 이와 같은 알고리즘을 통해서 강화 학습은 결국 상태와 행동을 사상시키는 최적의 정책을 구성할 수 있게 된다.

만약 어떤 정책 π에서 행동을 취하였을 때의 기대값은 식 (1)과 같이 표기할 수 있으며, 보상값은 사용자와 시스템간의 최종목표에 있어서 얼마나 유용한지를 가치함수 $Q(s,a)$를 이용하여서 표현할 수 있다.

$$Q^{\pi}(s) = E_{\pi}\{R_t \mid s_t = s, a_t = a\} = E_{\pi}\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \mid s_t = s, a_t = a\} \quad (1)$$

여기서 상태-행동의 평가 함수 $Q(s,a)$값을 예측하고, 이 Q-값에 기반하여서 행동을 결정하는 것을 Q-학습법이라고 한다. 이 $Q(s,a)$는 미리 정해져 있지 않기 때문에, 에이전트는 지금까지의 결과를 바탕으로 점진적으로 값을 찾아간다. 에이전트는 현재 상태에서 가능한 행동 중 $Q(s,a)$값이 가장 큰 값의 행동을 최적의 행동으로 선택하고 수행한다. 그리고 이것을 수행한 후 환경으로부터의 보상값 r을 받고 기존의 $Q(s,a)$를 갱신하며 식 (2)를 바탕으로 진행한다.

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \cdot [r_{t+1} + \gamma \cdot \max_a Q(s_{t+1}, a) - Q(s_t, a_t)] \quad (2)$$

본 연구에서 식 (2)는 ε-탐욕 행동에 기반하여서 진행하도록 하였다. 이것은 에이전트가 학습에서 새로운 행동을 배울 때, 지금까지의 경험을 바탕으로 하여서 최선의 행동을 취할지, 아니면 임의의 행동을 취해서 더 좋은 결과를 기대할지에 관한 딜레마에 빠지는 것을 해결하는 방법이다. 본 연구에서 사용한 ε-탐욕 행동은 $(1-\varepsilon)$의 행동선택은 최적이라고 추정되는 행동을 취하고, ε만큼의 행동은 미래에 더 최적이라고 추정되는 행동(Exploration)을 선택하는 방법이다. ε만큼은 계속적으로 Exploration을 하므로 지역 최적화에 빠지지 않고, 더 빨리 최적의 목표에 도달하게 된다.

3.2 강화 학습법에 의한 최적화 대화 전략

대화처리에서 앞 절에서 다룬 보상값의 평가방법은 중요한 연구대상이다. 사람의 대화 흐름에서는 앞의 발화가 다음 발화를 결정하는 것이 아니기 때문에, 전체적인 대화흐름에서 부분 발화가 대화 목표를 달성하는데 얼마나 유용하였는지 알기 어렵기 때문이다. 기존 연구로 Walker는 PARADISE 방법론을 제안하였고[5], 또한, [11]에서는 그 방법론을 사용하여서, 전체 대화의 효용성을 측정하고 강화학습의 보상값으로 사용하였다. Walker는 여기서 직전에 일어난 발화의 Q값을 반영하는 변수에 1을 놓아서 과거의 모든 보상값을 균등하게 현재에 반영하고 있다.

마코프 속성을 만족시킨다는 것은 각 과정들이 연관되어 진행될 때, 각 과정의 속성들이 다음 과정에 영향을 미치도록 할 경우를 뜻하는 것이다. 하지만 대화의 경우에는 기존 연구의 각 발화 단위가 반드시 다음 발

화에 영향을 미치는 것이 아니므로 마코프의 속성을 충족시키지 않을 수 있다. 이에 따라서 본 연구에서는 대화를 발화 단위가 아닌 각각의 과정으로 이루어졌다고 가정하였으며, 전체 대화 중 하나의 대화 단계를 부대화로 정의하였다. 이에 따라서 하나의 부대화(Sub-Dialogue)는 하나의 상태(Condition)로 가정되어 마코프 속성을 가질 수 있도록 하였다. 또한 Walker는 발화를 통해 일어난 것만 보상값으로 사용하였는데[11], 본 연구에서는 대화 흐름도 반영하기 위해서 보상값을 주대화와 부대화의 두 번에 걸쳐서 평가한다.

이에 따라 본 연구에서는 시스템과 사용자간에 하나의 주제와 이에 연관된 부분 대화를 가지도록 대화 구성을 고안하였으며, Q-학습 알고리즘이 대화 처리 시스템에 맞도록 개선하였다.

$$Q(s,a) \leftarrow Q(s,a) + \alpha \cdot r_D \quad (3)$$

식 (3)은 주대화와 부대화를 반영하기 위해서 각 대화에서 마지막 부대화에 도달하면 저장된 Q값에 대화 평가값(r_D)을 이용하여, 다시 Q값을 갱신하는 식이다. 본 연구에서는 기존 Q-학습 알고리즘 (3-2)과 제안된 식 (3)을 사용한 아래 알고리즘을 만들어 대화 처리 시스템에 적용하였다.

4. 강화 학습을 이용한 효과적인 대화 처리 시스템의 구현

4.1 시스템 구현의 구성

본 연구에서는 3절에서 제시된 개선된 강화 학습법으로 TV프로그램 정보 제공 시스템을 구축하여 실험하였다. TV 프로그램 정보는 데이타 관리의 효율성 및 편의성을 위해서 MY SQL 4.0.2를 이용한 데이타베이스로 저장하였다. 데이타베이스에는 "프로그램 이름", "채널", "프로그램 시작 시간", "프로그램 종료 시간", "프로그램 방영 날짜", "프로그램 장르", "프로그램 캐스팅", "프로그램 등급", "프로그램 제작자", 프로그램 개요" 등의 TV 프로그램 정보를 관계형 데이타베이스 형태로 저장하였다.

사용자는 몇 단계의 과정을 거치면서 원하는 TV 프로그램에 관한 요구 사항들을 입력하고, 시스템은 입력된 정보로, SQL 문장을 만들어서 데이타베이스와의 질의 응답을 하고, 이 얻어온 정보를 사용자에게 맞게 가공하여서 보여준다.

시스템은 사용자와의 대화를 통해서 작업을 수행하도록 구성되어 있는데, 강화학습을 하기 위해서 사용자 발화 입력을 키보드와 마우스를 이용한 정보 입력으로 대체하여 학습하였다. 시스템은 매번 입력된 발화에서 입력 정보와 대화 상태를 추정한 값으로 강화 학습을 하면서 사용자 모델링을 하게 된다. 시스템은 사용자에게 저장된 강화학습 값과 Q-학습 알고리즘에 의하여 발화한다. 사용자와 시스템 간의 공통의 목적을 완수한 후, 즉 대화 종료 후에는 사용자에게 이번 대화의 전체적인 전략을 평가하게 하고, 이 점수에 기반하여 다시 Q-Value를 갱신하도록 하였다.

4.2 개선된 강화학습 알고리즘을 사용한 대화 처리 시스템

본 실험에서 사용한 대화 처리 시스템은 그림 3에서 제안된 알고리즘을 바탕으로 Q-Value의 갱신을 주대화와 부대화의 두 번에 걸쳐서 하도록 그림 4와 같이 구성되었다. 그림 4에서 볼 수 있듯이 시스템은 대화의 흐름을 깨지 않으면서 각 Q-학습법의 상태를 평가할 수 있도록 대화를 부대화의 묶음으로 생각하였고, 이 부대화의 달성 여부를 파악할 수 있는 입력 값들을 이용해서 상태의 보상값으로 사용하도록 구성하였다. 또한

Initialize $Q(s,a)$ arbitrarily
Repeat (for each episode (Dialogue)):
    Initialize s
    Repeat (for each step of episode):
        Choose action a_t for state s_t using policy derived from Q by ε-greedy search
        Tack action a_t, observe reward r_{t+1}, s_{t+1}
        $Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha \cdot [r_{t+1} + \gamma \cdot \max_a Q(s_{t+1},a) - Q(s_t,a_t)]$
        $s_t \leftarrow s_{t+1}$
    until s_t is terminal
$Q(s,a) \leftarrow Q(s,a) + \alpha \cdot r_D$

그림 3 개선된 Q-학습법이 적용된 알고리즘

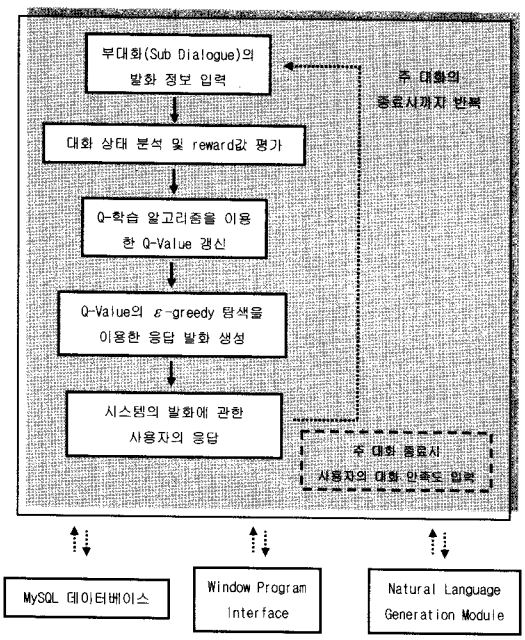

그림 4 사용된 Q-학습 알고리즘 및 시스템 구성도

전체적인 대화에 관한 사용자 만족도와 대화의 흐름에 관한 효율성도 보상값에 반영하기 위해서, 대화가 끝난 후 사용자는 대화의 전체적인 평가를 하고, 이 값으로 Q-값도 갱신한다.

본 대화 시스템에서는 사용자 대화가 1)달성 완료, 2) 달성 소요 시간의 최소화, 3)부 대화의 오류 감소 등을 이루었을 때, 사용자가 시스템에 관하여 긍정적인 만족을 가지는 것으로 가정하였다. 이에 따라서 대화 상태 평가는 달성 여부, 달성시 소요 시간, 부대화에서의 오류 수 등의 스칼라 값으로 변환하여 보상값으로 적용하도록 구성하였다.

4.3 시스템의 실험 과정

본 시스템은 사용자가 그림 5의 과정을 거치면서 원하는 TV 프로그램의 정보를 찾아가게 된다.

위의 검색과정은 각각의 부대화 단계로서 그림 5에 나타난 과정이 끝나면 주대화는 종료된다. 시스템은 각 과정에서 시스템이 필요로 하는 정보를 입력하고 이에

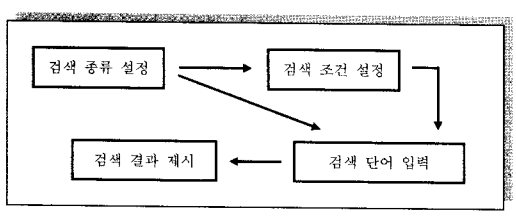

그림 5 시스템의 검색 과정

따라서 시스템의 대화 전략이 들어가 자연어 문장을 만들어 내어서 사용자에게 응답 발화한다. 시스템은 대화 전략은 예제(Example) 제시여부, 응답의 확인(Confirmation) 제시의 여부, 결과의 요약 방법 등이 있다. Q-학습법 알고리즘을 이용하여 각 상태마다 가능한 행동 중 하나를 선택한 후, 이를 활용하여 사용자에게 질의 응답하게 된다.

5. 실 험

5.1 실험 데이타 및 방법

본 연구에서는 실험을 2가지로 나누어서 진행하였다. 첫 번째 실험에서는 각 개인마다의 강화 학습에 따른 훈련을 하고, 사용자 모델링을 알아 보도록 진행하였다. 본 대화처리 시스템을 처음 사용하는 10명에게 약 25회 동안 대화 처리 시스템을 사용하도록 하여서, 강화학습의 학습 진행 여부 추이를 확인하였다. 그리고 강화학습을 통하여 나온 정책을 검증하기 위하여 각 사용자에게 5회에 걸쳐서 대화 시스템을 수행한 결과와, 비교대상으로 삼은 대화전략이 고정된 시스템, 랜덤(Random)하게 작동하는 시스템을 5회에 걸쳐 사용한 후의 결과와 각각 비교하였다.

두 번째 실험에서는 각 사용자 역량에 따른 모델링의 차이 및 강화학습의 타당성을 검증하기 위하여 2개의 그룹으로 나누어서 실험을 진행하였다. 사용자 선별은 시스템의 사용에 따른 역량의 차이로 분리하였는데, 본 시스템을 사용하기 이전에 자연어 처리를 사용하였고, 본 시스템을 30회 이상 사용한 대상을 전문가 그룹으로, 시스템을 전혀 처음 접하는 대상을 초보자 그룹으로 나누었고 개인별이 아니라 각 그룹별로 강화학습을 수행하였다. 또한 두 번째 실험에서도 본 연구에서 제안한 적응형 시스템을 비교하기 위하여, 대화 전략을 고정시킨 시스템, 대화 전략을 랜덤하게 변화시키는 시스템을 두 그룹에게 동일하게 진행하였다. 이 실험은 Finnish Interact Project [10]와 같이 사용자의 역량에 따른 적응형을 적용하기 위해서 실시되었다. 즉, 각 범주에 따라 강화학습을 적용하여 나온 결과를 토대로 사용자 최적화 모델링을 얻는 시스템 방법을 실험하는 데에 그 목표를 두었다.

본 연구 시스템의 목적이 TV 프로그램 정보 안내이므로, 실험 데이타 수집을 위해서 사용자에게 표 1의 과제를 주고 수행하게 하여 강화 학습을 훈련하도록 진행하였다. 사용자의 시스템 평가 만족도를 정량적으로 분석하기 위하여서 -2 ~ 4 사이 점수로 사용자가 평가하도록 하였다.

실험에 참가한 피실험자들과 특성을 정리하면 표 2와 같다.

5.2 사용된 대화 전략

고안된 시스템에서 사용자에게 사용한 대화 전략은

표 1 실험 과제

1. 당신은 지금 밤 7시에 마땅히 할 일이 없어서 TV를 보려고 한다. 7시에 보기에 적당한 프로그램을 찾아보아라.
2. 당신은 지금 KBS1의 뉴스를 녹화하려 한다. 녹화를 위해서 KBS1의 뉴스는 어느 시간대에 하는지 찾아보아라.
3. 당신은 오후 10시대에 드라마를 보려고 한다. 보고 싶은 드라마를 정하기 위해서 오후 10시에 하는 드라마에 누가 출연하는 지를 찾아보아라.

표 2 피실험자 데이타

	대상	데이타의 양	횟수
실험 1	처음 사용자	10명	1인당 30회
실험 2	전문가 그룹: 자연어처리 시스템에 관하여 경험이 있으며 본 시스템을 30회 이상 사용한 사람	각 그룹당 15명	1인당 15회
	초보 사용자 그룹: 시스템을 처음 사용하는 사람		

표 3 시스템에서 사용된 대화 전략

Example 1 (설정)	사용자에게 현재 입력 가능한 것이 무엇인지 제시하면서 진행한다.
Example 2 (해제)	사용자에게 입력 가능한 것을 제시하지 않으면서 진행한다.
Confirmation 1	사용자가 입력한 내용을 다시 보여주지 않는다.
Confirmation 2	매번 사용자가 입력한 내용을 확인하도록 다시 보여준다.
Confirmation 3	입력된 조건으로 검색하기 전, 사용자가 무엇을 제시하였는지 보여주면서 틀린 지, 맞는 지도 묻는다.
Summary 1 (설정)	검색된 결과가 4개 이상일 경우, 검색된 결과물에서 다시 한번 검색을 하게 된다.
Summary 2 (해제)	검색된 결과의 개수에 상관없이 보여준다.

예제(Example) 제시여부, 응답의 확인(Confirmation) 제시의 여부, 결과의 요약 방법 등의 방법으로 표 3과 같은 방법들을 대화에 적용하면서 진행하였었다.

그림 6은 시스템과 사용자간의 대화예제를 보여주는데, 이것은 프로그램 이름으로 "프란체스카"가 들어가 있는 내용을 검색하려는 주대화이다. 세부 부대화에는 검색과정을 거치면서 정보를 입력하도록 하는 요소들로 구성되어 있다. 사용자는 마우스와 키보드를 이용하여서 원하는 TV프로그램 검색을 위한 내용들을 입력하므로 대화 예제에서는 시스템의 응답 발화와 사용자의 행동을 보여준다. "System 1,2"의 발화와 "User 1"의 행동은 시스템이 사용자에게 검색종류를 설정하는 과정이며, "System 3"과 "User 2"는 검색조건을 설정하도록 하고, "System 4,5"와 "User 3"의 발화는 검색단어를 입력하게 하고 있다. 또한 "System 6,7,8"과 "User 4,5" 발화는 사용자에게 검색 결과를 제시하여 주고 있다.

시스템의 응답발화는 표 3에서 제시된 대화 전략들을 사용한다. 예를 들어 System5의 발화는 적응형이 부여되어 Confirmation 3 전략이 들어간 발화이다. 만약 Confirmation 1일 경우라면 아무 발화도 하지 않으며, Confirmation 2의 경우는 "프란체스카를 입력하였습니다."로 발화하는 것으로 세 가지 확인전략 중 사용자에게 맞는 방향을 선택하여 나가면서 발화한다.

5.3 실험 결과 분석

5.3.1 실험 1: 개인별 강화 학습

본 연구의 강화 학습법 알고리즘으로 훈련된 정책을 가진 시스템을 사용자들에게 평가하도록 한 결과, 1.13의 만족도를 보여주었다. 또한 본 연구에서 베이스라인인 랜덤 전략은 사용자의 스타일에 상관없이 대화 전략이 랜덤으로 변하며 고정 전략은 사용자의 스타일에 상관없이 대화 전략이 고정된 것을 의미하였다. 제안한 알고리즘은 대화 평가값이 1.13이고, 랜덤 전략과 고정 전략은 각각 -0.26과 0.78으로 비교되었다.

그림 7은 10명의 사용자에게 본 시스템을 사용토록 하면서 Q-학습을 시켰을 때, 각 사용자가 평가한 대화 횟수에 따른 대화에 관한 평가 값에 대한 평균값을 그

(User5: 입력 버튼을 누른다.)

System10: 원하시는 TV 프로그램 결과를 찾으셨습니까? 대화 결과에 관한 평가를 '점수 트랙바'를 조정하여 입력하여 주십시요.

(User6: 점수 트랙바를 조정하여서 점수를 입력한다.)

그림 6 시스템 대화 예제

표 4 사용자의 평가 값 평균

유저 모델링	사용자 대화 평가 값 평균
Adaptive Modeling	1.13
Random Modeling	-0.26
Fixed Modeling	0.78

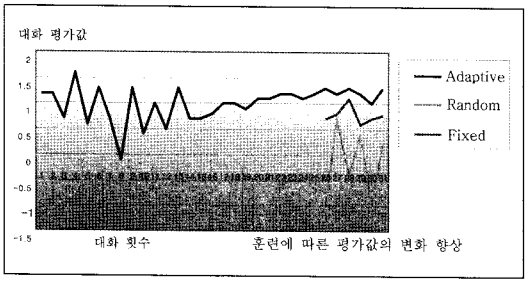

그림 7 개인별 강화학습 훈련 그래프

래프로 나타낸 것이다. 그래프의 오른쪽에 있는 랜덤 전략과 고정 전략의 그래프와 제안된 시스템의 그래프를 비교해 봤을 때, 일정 대화 이상 학습을 수행했을 때 다른 전략보다 더 좋은 성능을 보임을 알 수 있다. 그림 7에서 보는 바와 같이 적응형 대화 시스템의 그래프는 Q-학습이 Exploration/Exploitation을 반복하면서 서서히 최적의 정책을 찾아가므로, 초기에는 발산하지만 서서히 수렴을 하여 나가는 것을 보여 준다.

5.3.2 실험 2: 사용자 역량별 강화학습

표 5와 표 6은 사용자의 역량을 전문가 사용자 그룹과 초보 사용자 그룹으로 나누어서 모델링을 하고, 그 결과를 다시 사용자에게 평가하도록 하였을 때의 결과를 보여 준다.

전문가 그룹을 모델링을 하여 나온 결과를 적용한 것이 표 5이며, 초보 사용자 그룹의 모델링 결과를 적용한 것이 표 6이다. 표 6의 만족도가 표 5보다 전체적으로 떨어지는 것은 초보 사용자 그룹이 아직 시스템을 어떻게 사용해야 할 지 잘 모르고 사용법 또한 사용자마다 달라서, 모델링 값이 모든 사용자를 만족시키기에는 일

표 5 전문가 그룹의 정책 적용 결과

유저 모델링	전문가 그룹	초보 사용자 그룹
Adaptive Modeling	**0.893**	**0.523**
Random Modeling	-0.746	-0.63
Fixed Modeling	0.167	0.445

표 6 초보 사용자 그룹의 정책 적용 결과

유저 모델링	전문가 그룹	초보 사용자 그룹
Adaptive Modeling	**0.324**	**0.579**
Random Modeling	-0.746	-0.63
Fixed Modeling	0.167	0.445

관적이지 못했기 때문으로 분석된다. 또한 두 가지의 그룹으로 나누어서 사용자에게 시스템 만족도를 평가하게 할 경우, 강화 학습 자체가 그룹별로 이루어지기 때문에 개인의 특성을 실험1보다는 잘 반영할 수 없으므로 실험1의 결과보다 낮은 만족도를 보여준다.

하지만 모든 사용자가 개별적으로 강화학습을 한 후에 시스템을 사용해야 한다는 과정이 없어졌기 때문에, 사용자 시스템을 처음 사용할 때 그룹의 대화 전략 정책을 사용한다면 훈련에 걸리는 시간이 절약된다는 점에서 효율적이다.

6. 결론 및 향후 연구 과제

본 연구에서는 시스템과 사용자간의 상호작용을 하는 대화 처리 전략으로 기계 학습법 중 하나인 강화 학습법을 사용하고, 대화에 이 강화 학습법을 적절하게 모델링하는 방법을 제안하였다. 본 연구는 기존의 각 발화를 적용 대상으로 사용한 연구와 달리, 각각의 부대화를 대상으로 삼아서 대화 처리 시스템에 적응형을 부여하고, 사용자 역량에 따른 강화학습을 적용하여 사용자 최적화된 모델링을 얻는 시스템 방법을 제안하였다.

효과적인 강화 학습법의 적용을 위해서 제안된 방법을 정리하면 다음과 같다.

1) Q-학습에 대화처리 전략을 적용시키기 위해서 대화를 주대화와 부대화로 나누었고, 하나의 주대화는 에피소드로, 부대화는 상태로 사상시켜서 강화 학습법에 적용시켰다.
2) 부대화는 주대화에서 각 단계마다 하나의 주제를 완료시키는 대화로 정의하였으며, 이 부대화의 완료여부, 완료시간, 에러 등을 측정한 스칼라 값으로 한 상태의 보상값을 사용하였다.
3) 주대화는 일관성이 있는 시스템과 사용자간의 하나의 대화를 말한다. 주대화가 끝날 때마다 사용자 전체적인 대화전략을 평가하고 시스템에 입력하여서 Q-값을 갱신하게 한다.
4) 다양한 전략이 들어가 있는 시스템 발화를 만들어내기 위해서 ε-탐욕 탐색법(Greedy Search)으로 Q-학습의 값을 찾게 된다.
5) 시스템의 발화는 예제(Example)의 제시 여부, 사용자 발화의 확인법(Confirmation), 프로그램 정보의 요약법(Summary) 등을 이용하여서 사용자에게 다양한 대화 전략을 구사하게 된다.

또한 상기 대화 전략이 적용된 시스템의 강화 학습이 대화의 효율성 및 사용자 만족도가 제대로 학습되었는지를 보기 위하여 각 개인별 실험과, 사용자의 역량별 양태를 파악하기 위한 그룹별 실험을 진행하였다. 개인별 실험에서는 제안된 알고리즘의 성능이 비교된 알고

리즘보다 우수함을 확인하였다. 그룹별 실험에서는 미리 훈련된 강화 학습의 정책을 사용할 경우, 각 개인마다 훈련 과정을 거친 정책보다는 성능은 떨어지지만, 사용자 편의성을 위해 훈련 과정이 단축되므로 시스템의 효율성을 높일 수 있다는 것을 보여주었다.

향후 연구 과제로는 1)음성 언어 처리 시스템과의 연구, 2)대화의 정량적인 평가에 인지과학적인 접근, 3)대화 처리 모델의 수학적 정형화에 관한 연구를 제안한다. 본래 사람의 대화에서 의사소통 도구는 음성인 만큼, 연구 범위를 음성으로 확대하면 다양한 대화의 양태 파악이 가능하다. 따라서 대화 시스템을 음성 언어 처리로 확대하여 연구하면 대화 처리 양상을 보다 잘 반영한 사용자 모델링을 만들 수 있을 것이다. 또한 본 연구는 컴퓨터학에서의 자연어 처리에 초점을 맞춰서 진행되었는데, 적응형 능력을 강건하게 부여하려면 언어학적 요소와 인지과학적 요소를 이용한 대화의 정량적인 평가가 있어야 할 것이다. 마지막으로 본 연구의 목표가 최적화 대화 전략이므로, 이에 대한 기반 연구로 대화의 수학적인 정형화 연구와 최적화된 대화 전략이 존재할 수 있는지에 대한 연구도 필요하다.

참 고 문 헌

[1] T. Dean J.Allen, Y. Aloimomos, Artificial Intelligence Theory and Practice, Addison-Wesley, 1995.

[2] I. Zukerman, D. Litman, "Natural Language Processing and User Modeling: Synergies and Limitations," User Modeling and User-Adapted Interaction Vol.11, pp. 129-158, 2001.

[3] J.W. Wallis, E.H. Shortliffe, "Customized Explanations Using Causal Knowledge," In B.C. Buchanan, E.H. Shortliffe(eds): Rule-based Expert System: The MYCIN Experiments of the standford Heuristic Programming Project, Addison-Wesley Publishing Company, pp. 371-388, 1985.

[4] M. McTear, Spoken Dialogue Technology Toward the Conversational User Interface, Springer-Verlag London, 2004.

[5] M. A. Walker, D. Litman, C. A. Kamm, A. Abella "PARADISE: A Framework for Evaluating Spoken Dialogue Agents," In Proceedings of the 35th Annual Meeting of the Association of Computational Linguistics(ACL 97), pp. 271-280, 1997.

[6] K. Jokinen, K. Kanto, "User Expertise Modeling and Adaptive in a Speech-Based E-mail System," In proceedings of Annual Meeting of the Association of Computational Linguistics 2004(ACL 2004), pp. 87-94, 2004.

[7] S. Möller, "A new Taxonomy of the Quality of Telephone Service Based on Spoken Dialogue System," In proceedings of the 3 th SIGdial Workshop on Discourse and Dialogue, Philadelphia, PA. pp. 142-153, 2002.

[8] D. Litman, S. Pan, "Empirically Evaluating an adaptable spoken dialogue systems," In Proceedings of the 7th International Conference on User Modeling(UM'99), pp. 55-64, 1999.

[9] 은지현, 최준기, 장두성, 김현정, 구명완, "마르코프 의사결정 과정에 기반한 대화 관리 시스템", In Proceedings of the HCI 2007, pp. 475-480, 2007.

[10] K. Jokinen, M. Kaipainen, T. Jauhuainen, G. Wilcock, M. Turunen, J. Akulinen, J. Kussis, K. Lagu, "Adaptive Dialogue System Interaction with interact," In Proceedings of the 3rd SIGdial Workshop on Discourse and Dialogue, Philadelphia, PA, pp. 64-73, 2002.

[11] M. A. Walker, J. Wright, I. Langkilde, "Using natural Language Processing and Discourse features to identify understanding errors in a spoken dialogue system," In Proceedings of the 17th International Conference on Machine Learning, Palo Alto, CA. pp. 1111-1118, 2000.

[12] R. S. Sutton, A. G. Barto, Reinforcement Learning An Introduction, MIT Press, 1998.

김 원 일

2003년 서강대학교 컴퓨터학과 학사. 2003년 독일 Siemens 인턴 근무. 2005년 서강대학교 컴퓨터학과 석사. 2005년~현재 삼성전자 영상디스플레이 사업부 연구원 관심분야는 사용자 인터페이스 설계 및 디자인, 자연어처리, 대화처리, 음성언어처리 등

고 영 중

1996년 서강대학교 수학과 학사. 1996년~1997년 LG-EDS 근무. 2000년 서강대학교 컴퓨터학과 석사. 2003년 서강대학교 컴퓨터학과 박사. 2004년~현재 동아대학교 컴퓨터공학과 조교수. 관심분야는 자연어처리, 텍스트 마이닝, 정보 검색, 대화 시스템, 소프트웨어공학 등

서 정 연

1981년 서강대학교 수학과 학사. 1985년 미국 Univ. of Texas, Austin 전산학과 석사. 1990년 미국 Univ. of Texas, Austin 전산학과 박사. 1990년~1991년 미국 Texas Austin, UniSQL Inc. Senior Researcher. 1991년 한국과학기술원 인공지능 연구 센터 선임연구원. 1991년~1995년 한국과학기술원 전산학과 조교수. 1996년~현재 서강대학교 컴퓨터학과/바이오융합기술 협동과정 정교수. 관심 분야는 한국어 정보 처리, 자연어처리, 대화처리, 지능형 정보 검색 등