

계층적 정보유형과 문맥정보를 사용한 항공여행대화시스템에서의 예약정보 추출

김세종^o 나승훈 이종혁

포항공과대학교 전자컴퓨터공학부 컴퓨터공학과, 첨단정보기술 연구센터
{sejong^o, nsh1979, jhlee}@postech.ac.kr

Information Extraction for Air Travel Dialogue System Using Hierarchical Information Types and Contextual Features

Se-Jong Kim^o Seung-Hoon Na Jong-Hyeok Lee

Dept. of Computer Science and Engineering, Division of Electrical and Computer Engineering, POSTECH
Advanced Information Technology Research(AITrc)

요 약

대화시스템은 사용자가 자연언어를 사용하여 해당 시스템과 필요한 정보를 주고받는 목적 지향적 에이전트로서 활용되어 왔다. 이러한 대화형 에이전트는 사용자의 입력으로부터 필요한 정보를 정확하게 추출함으로써 이후 처리단계에서의 결과를 향상시킬 수 있다. 본 논문에서는 항공여행관련 대화에서 발생하는 예약정보들 중에서 경유정보, 특히 경유하는 시간 및 날짜에 대한 정보를 효과적으로 추출하는 방법에 대해서 다룬다. 출발·도착정보와 경유정보를 계층적으로 분류하고, 현재 발화되고 있는 문장보다 선행되고 있는 문장들의 예약정보들을 문맥정보로 사용하여 현재 문장에서 추출하고자 하는 정보들을 학습하고 평가하였다. 이를 통해서 얻어진 결과는 출발·도착 및 경유정보를 동시에 고려했을 때보다 효과적인 학습 성능을 보였으며 실제로 시간정보에 대해서는 81.5%, 날짜정보에 대해서는 92.0%의 정확도를 보였다.

1. 서 론

대화란 두 사람 이상이 그들의 언어를 사용하여 서로 마주 대하고 주고받는 말을 가리킨다. 대화시스템은 이러한 대화를 통해 사용자와 시스템 사이에서 정보교환을 할 수 있도록 구현한 자연언어처리기술을 바탕으로 한 사용자 친화적 시스템을 말한다. 특히 대화시스템은 사용자의 요구를 대화를 통해 자동적으로 해결해 주는 목적 지향적 에이전트로서 활용되어 왔으며 기차예약에이전트, 항공예약에이전트, 인터넷쇼핑도우미 등이 그 활용 예라고 하겠다.

이러한 대화형 에이전트는 사용자로부터 입력이 들어왔을 때 형태소 분석 및 패턴매칭 등을 수행하여 해당 입력으로부터 필요한 정보들을 추출하고 이 정보들을 바탕으로 대화유형 등을 판별하는 분석단계를 거친다. 그리고 미리 정의된 대화전략에 따라 답변을 생성하기 위한 정보를 자연언어생성모듈에 전달하여 답변 매칭 및 실제적인 문장 생성을 통해 사용자에게 자연스러운 답변을 제공한다[1]. 대화형 에이전트의 각 처리과정은 그 이후 과정에 직접적인 영향을 주기 때문에 각 과정별로 정확한 처리를 수행해야만 이후 처리단계에서의 결과를 향상시킬 수 있다. 하지만 사용자가 입력한 문장들은 매우 다양해서 형태소 분석 및 패턴매칭만을 사용할 경우 다른 정보들에 비해 상대적으로 심층적인 의미를 가진 정보들을 추출하는데 한계가 있다.

본 논문에서는 이러한 한계점을 감안하여 항공여행관련 대화에서 발생하는 예약정보들 중 경유정보 특히 경유하는 시간 및 날짜에 대한 정보를 효과적으로 추출하

는 방법에 대해서 다룬다. 2장에서는 대화형 에이전트와 관련된 다양한 연구 분야에 대해서 간단히 소개하고 3장에서는 제안하고자 하는 방법을 기존의 일반적인 방법과 비교하여 설명한다. 4장에서는 실제 사용자와 항공예약 전문가 간의 대화내용을 기록한 SRI ATIS Collection[2]을 정제된 코퍼스를 사용하여 실험결과를 제시하고 5장에서는 결론을 내린다.

2. 관련연구

대화형 에이전트는 1966년 키워드 추출에 의한 단순 패턴매칭 기법을 적용한 MIT의 ELIZA를 시작으로 순차 패턴매칭 기법을 적용한 ALICE, 교육시스템으로서 활용될 수 있는 AutoTutor 등 다양한 분야로 개발되어 왔다 [3, 4, 5]. 그리고 이러한 과정들 속에서 에이전트의 대화에 관한 성능 향상을 위해 대화분석 및 대화전략대화문맥 유지, 문장생성에 관한 연구가 진행되어 왔다. 관련 연구 몇 가지를 구체적으로 살펴보면 대용어구 처리는 대화에서 등장하는 지시대명사가 무엇을 가리키는지 알아내는 방법으로 중심화 이론, 결정트리 등을 사용한 연구가 진행되었대[6, 7]. 화행이론은 발화된 문장의 종류를 도메인 독립적, 또는 도메인 의존적인 태그를 사용하여 결정하는 것으로 대화 분석에서 유용하게 사용될 수 있으며 N-gram, HMM(Hidden Markov Model), 자질 선택 기법 등을 사용한 연구가 진행되었대[8, 9].

대화형 에이전트에서 필요한 정보를 추출한다는 것은 특정 도메인에 의해서 이미 정의된 슬롯들을 채워가는

것을 말한다. 앞서도 언급했듯이 정보추출을 위한 방법으로는 형태소 분석 및 패턴매칭을 들 수 있으며 개체명인식(NER, Named Entity Recognition)도 같이 적용할 수 있다. 문장으로부터 추출하고자 하는 정보가 특정 품사일 경우 형태소 분석을 통해 품사정보를 확인할 수 있으며 미리 작성된 패턴과 입력된 문장을 비교하여 원하는 정보 패턴과 일치하는 정보를 찾아 추출할 수 있다. 여기서 패턴이란 정보의 종류를 판단할 수 있는 오토마타나 문장의 형태를 일종의 스크립트로 표현한 격틀 정보 등을 의미한다. 개체명인식 기법의 한 방법으로서 ME(Maximum Entropy)를 사용하여 정보추출을 수행할 경우 후보자질로서 형태소 분석결과 구류음, 분류사전, 도메인 온톨로지를 사용한 연구가 진행되어 왔으며 이밖에도 동의어사전이나 WordNet을 사용하여 정보추출에 도움을 줄 수 있다[1, 10].

3. 제안하는 방법

3.1 정보유형의 계층화

항공여행관련 대화에서 발생하는 예약정보들은 ‘출발장소’, ‘도착장소’, ‘출발날짜’, ‘도착날짜’, ‘출발시간’, ‘도착시간’, ‘항공기회사’, ‘항공기번호’ 등 매우 다양하다. 특히 해외에서는 단일경로가 아닌, 특정 장소를 경유하는 경우가 많아서 경유정보에 대한 처리가 매우 중요하다. 그림 1은 한 번의 경유가 발생했을 경우 해당 정보를 포함한 일반적인 예약정보들을 트리형태로 표현한 것이다.

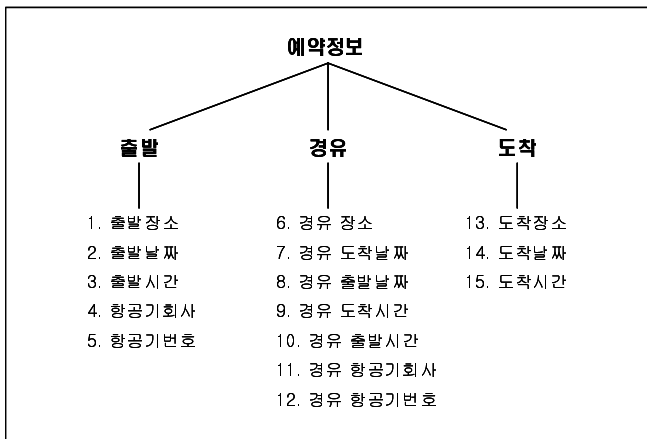


그림 1. 예약정보 트리

위 그림에서 알 수 있듯이 ‘장소’, ‘항공기회사’, ‘항공기번호’는 ‘출발’, ‘경유’, ‘도착’이라는 상위노드에 의해 명확히 구분되어지는데 반해, ‘날짜’, ‘시간’은 상위노드가 ‘경유’인 경우 ‘출발’ 및 ‘도착’이 ‘경유’와 함께 표시되어 명확히 구분되어지지 않는다. 이러한 점은 실제로 예약 정보추출에 있어서 문제를 발생시킬 수 있다. 즉 출발, 도착 및 경유정보를 동시에 고려해야하기 때문에 이를 만족하는 자질을 선택하는데 상대적인 어려움을 겪을 수 있으며 학습할 말뭉치가 한정되어 있을 경우 총 4가지의 예약정보를 한정된 하나의 말뭉치를 통해서 학습해야 한다는 한계점을 가지고 있다. 그림 ‘날짜’와 ‘시간’만을 가

지고 트리를 계층적으로 재구성해보자

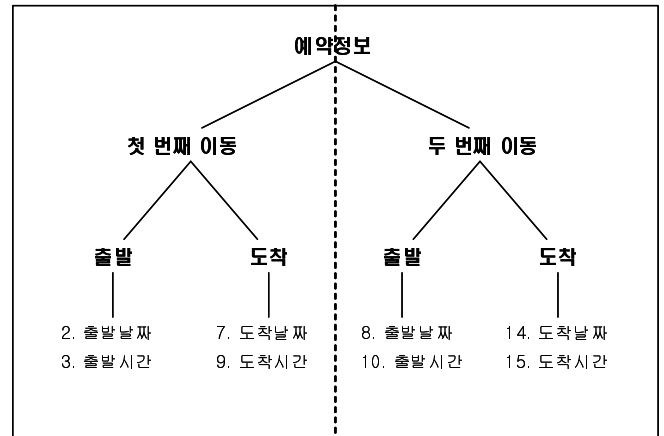


그림 2. 재구성한 날짜 및 시간정보 트리

그림 2는 경유한 순간을 경계로 하여 ‘날짜’와 ‘시간’을 ‘첫 번째 이동’과 ‘두 번째 이동’으로 나누었다. 그리고 하위노드는 ‘출발’과 ‘도착’으로 구분하였다. 이와 같이 정보유형을 계층화하면 ‘출발’과 ‘도착’을 구분할 수 있는 최적의 자질을 상대적으로 쉽게 선택할 수 있으며 ‘첫 번째 이동’과 ‘두 번째 이동’에 대한 새로운 최적의 자질도 쉽게 선택할 수 있다. 또한 학습할 말뭉치가 한정되어 있더라도 ‘출발’과 ‘도착’이라는 2가지의 예약정보를 하나의 말뭉치를 통해서 학습할 수 있으며 ‘첫 번째 이동’과 ‘두 번째 이동’이라는 예약정보에 대해서도 동일한 하나의 말뭉치로 학습할 수 있어서 상대적으로 정확도를 향상시킬 수 있다[11].

3.2 문맥정보의 활용

앞 절에서 소개한 예약정보들 중 출발정보와 도착정보는 다른 예약정보들에 비해 상대적으로 추출하기 쉽다. 항공여행관련 대화말뭉치를 분석해 본 결과 본 정보들은 해당 정보가 포함된 문장 내의 자질만으로도 구분하기 용이했으며 간혹 어떠한 자질도 없이 단독으로 예약 정보가 입력되었을 경우에는 바로 전 문장에 대한 자질을 사용하여 판단할 수 있었다. 예를 들어 “there is one on <United> that leaves <Oakland> at <eleven thirty p.m.>.”이라는 문장이 입력되었다면 ‘eleven thirty p.m.’은 ‘leaves’와 ‘at’이라는 자질 또는 키워드를 통해 출발정보임을 알 수 있다.

그러나 ‘첫 번째 이동’정보와 ‘두 번째 이동’정보를 구분하는 것은 현재 입력된 문장 및 바로 전 문장의 자질만으로는 판단하기 어려운 경우가 많다. 일반적으로 ‘첫 번째 이동’정보와 ‘두 번째 이동’정보를 하나의 문장에 포함하고 있는 경우는 각 장면(scene)의 후반부에 대부분 위치하고 있다. 바로 전 문장에 대해서는 해당정보를 구분하기 위한 자질을 포함하고 있는 경우가 드물었다. 이러한 문제점을 해결하기 위해서는 바로 전 문장뿐만 아니라 그 이전에 선행된 문장들의 정보들을 활용할 필요가 있다. 예약에 관한 대화는 해당 대화에서 예약정보들의 흐름, 즉 ‘특정 예약정보의 등장 유무 및 대화상에서의 등장 위치 또는 순서’에 영향을 받아 진행되기 때

문에 표 1과 같은 가정과 표 2와 같은 문맥정보를 작성하여 활용한다.

표 1. 문맥정보에 적용한 가정

- ※ A: 첫 번째 이동 B: 두 번째 이동
- (1) 현재 문장보다 선행된 문장들의 예약정보들은 A, B를 구분하는데 유용한 기준이 된다.
 - (2) 현재 문장보다 선행된 문장들의 예약정보들은 이미 정확하게 추출되었다.
 - (3) 현재 문장에서 나타나는 예약정보들에 대한 1차 정보(장소, 날짜, 시간, 항공기회사, 항공기번호)는 이미 파악되었다.
 - (4) 현재 문장 이후의 문장은 볼 수 없다.

표 2. 문맥정보

- ※ 만족 1, 비만족 0, 비등장 -1
- ※ A: 첫 번째 이동 B: 두 번째 이동
C: 구분하고자 하는 예약정보의 1차 정보
- * 선행된 ~
- (1) 문장들 중에서 가장 최근(t1)에 등장한 C가 B인 경우
 - (2) 문장들 중에서 두 번째 최근(t2)에 등장한 C가 B인 경우
 - (3)/(4) 문장들 중에서 t1/t2에 등장한 출발C가 B인 경우
 - (5)/(6) 문장들 중에서 t1/t2에 등장한 도착C가 B인 경우
 - (7)/(8) 문장들 중에서 t1/t2에 등장한 출발C와 현재C가 같은 경우
 - (9)/(10) 문장들 중에서 t1/t2에 등장한 도착C와 현재C가 같은 경우
- * 선행된 다른 발화자의 ~
- (11)/(12) 문장들 중에서 t1/t2에 등장한 출발C가 B인 경우
 - (13)/(14) 문장들 중에서 t1/t2에 등장한 도착C가 B인 경우
 - (15)/(16) 문장들 중에서 t1/t2에 등장한 출발C와 현재C가 같은 경우
 - (17)/(18) 문장들 중에서 t1/t2에 등장한 도착C와 현재C가 같은 경우
- * 선행된 ~
- (19)/(20) 문장들 중에서 t1/t2에 등장한 장소가 B인 경우
 - (21)/(22) 문장들 중에서 t1/t2에 등장한 장소와 동일한 장소가 현재문장 내 구분하고자하는 예약정보의 앞쪽에 위치하고 (19)/(20)을 만족할 경우
 - (23)/(24) 문장들 중에서 t1/t2에 등장한 장소와 동일한 장소가 현재문장 내 구분하고자하는 예약정보의 뒤쪽에 위치 위치하고 (19)/(20)을 만족할 경우
 - (25)/(26) 문장들 중에서 t1/t2에 등장한 항공기회사가 B인 경우
 - (27)/(28) 문장들 중에서 t1/t2에 등장한 항공기회사와 동일한 항공기회사가 현재문장 내 구분하고자하는 예약정보의 앞쪽에 위치하고 (25)/(26)을 만족할 경우
 - (29)/(30) 문장들 중에서 t1/t2에 등장한 항공기회사와 동일한 항공기회사가 현재문장 내 구분하고자하는 예약정보의 뒤쪽에 위치하고 (25)/(26)을 만족할 경우
 - (31)/(32) 문장들 중에서 t1/t2에 등장한 항공기번호가 B인 경우
 - (33)/(34) 문장들 중에서 t1/t2에 등장한 항공기번호와 동일한 항공기번호가 현재문장 내 구분하고자하는 예약정보의 앞쪽에 위치하고 (31)/(32)를 만족할 경우
 - (35)/(36) 문장들 중에서 t1/t2에 등장한 항공기번호와 동일한 항공기번호가 현재문장 내 구분하고자하는 예약정보의 뒤쪽에 위치하고 (31)/(32)를 만족할 경우

표 1의 첫 번째와 두 번째 가정은 선행된 문장들의 정보를 예약정보들로 한정해 이미 추출한 예약정보들의 재

사용 및 현재 문장의 예약정보들과 기존의 예약정보들 간의 상관관계를 보이기 위한 것이고 세 번째 가정은 시스템이 기본적으로 갖추어야 할 정보추출 능력을 마지막 가정은 대화의 일반적인 특성을 표현한 것이다

표 2의 문맥정보들은 이러한 가정을 바탕으로 각 정보들의 유형 및 내용들의 상관관계를 여러 가지 면으로 관련지어 작성한 것이다 몇 가지 예를 들어 설명하면 현재 구분하고자 하는 시간정보가 '두 번째 이동'일 경우 선행된 문장들 속에서 '첫 번째 이동'에 대한 정보가 등장했을 것이다. 즉 가장 최근 또는 두 번째 최근에 등장한 시간정보가 '첫 번째 이동'일 경우 그 이후에 등장한 시간정보는 '두 번째 이동'일 가능성이 크다 또한 구분하고자 하는 정보와 위치상으로 가장 근접한 선행된 문장들 속의 항공기회사정보나 항공기번호정보가 두 번째 이동일 경우 해당 정보는 '두 번째 이동'일 가능성이 크다. 그리고 선행된 정보들이 위와 같은 정보유형들을 동시에 모두 만족한다면 구분하고자하는 정보가 '두 번째 이동'일 가능성은 크게 증가할 것이고 이러한 사실은 다양한 문맥정보가 정보유형을 구분하기 위해 활용될 수 있음을 보여준다.

사용자와 에이전트 간의 대화 유형이 다양해짐에 따라 이러한 문맥정보들은 예약정보들을 추출 및 구분하는데 있어서 유용하게 사용될 것이며 전체 문장들을 인식할 필요 없이 필요한 부분만을 인식하는 음성인식 분야에서 여러검사용으로 활용될 수 있을 것이다

4. 실험결과 및 분석

본 논문에서는 실제 사용자와 항공예약전문가 간의 대화내용을 기록한 SRI ATIS Collection을 정제한 코퍼스를 사용한다. 이 코퍼스는 총 66개의 장면(scene)으로 구성되어 있으며 이들 중 경유정보가 포함된 장면만을 선택하여 실험데이터로 사용하였다 선택된 실험데이터는 총 25개의 장면으로 이루어져 있었으나 이를 정제하는 과정에서 40개의 장면으로 재분할하였다 실험데이터의 세부 정보는 표 3과 같다.

표 3. 실험데이터 정보

구분	내용(개수)
시간(t_*)	437
날짜(d_*)	237
장소(l_*)	547
항공기회사(c_*)	233
항공기번호(p_*)	99
문장	1817
대화를 주고받은 횟수	1313

실험은 표 2에서 정의한 내용을 실험데이터에 적용하여 문맥정보들을 추출하고 이를 데이터화하여 일반적인 통계적 학습방법의 단점을 보완한 구조적 위험성 최소화(SRM, Structural Risk Minimize)를 통해 오류를 최소화하는 SVM(Support Vector Machine)기반의 성능평

가를 수행하였다. 특히 SVM은 입력 자질이 많은 경우에도 안정적인 성능을 보이므로 다양한 문맥정보를 고려하는데 있어서 유용하게 활용될 수 있다. 앞에서 선언한 모든 문맥정보들의 값(-1, 0, 1)을 SVM의 벡터공간에 표시하고 최적의 분류경계를 찾아 이에 대한 성능을 평가하였고 모든 문맥정보들을 사용했을 때의 평가결과와 자질 선택 기법을 통해 선택된 문맥정보들만을 가지고 분류한 평가결과를 비교하여 좋은 성능을 보인 방법을 성능평가의 척도로 사용하였다.

예약정보들 중 출발 및 도착정보는 다른 예약정보들보다 쉽게 추출될 수 있으므로 그림 2의 구조를 적용하기 위해 출발 및 도착정보를 이미 추출했다는 가정 하에 수행하였다. 즉 출발시간정보, 도착시간정보, 출발날짜정보, 도착날짜정보에 대하여 각각 '첫 번째 이동'정보와 '두 번째 이동'정보를 구분하는 성능평가를 수행하였다. 각각의 성능평가 이후에는 출발정보와 도착정보를 실험데이터에 추가한 것과 추가하지 않은 것을 시간정보와 날짜정보에 적용하여 출발정보와 도착정보 유무에 대한 비교 성능평가를 수행하였다. 이는 정보유형의 계층화가 어떠한 효과를 발휘하는지 판단할 수 있는 자료가 될 것이다. 또한 실험데이터들의 샘플수가 비교적 적은 편이므로 학습데이터와 평가데이터를 고정적으로 분리하지 않고 총 실험데이터의 10%를 평가데이터로 하고 남은 90%를 학습데이터로 하여 성능평가를 수행한 후 같은 방식으로 평가데이터와 학습데이터를 순차적으로 갱신하여 전체 성능평가를 수행하는 10-fold cross validation 방법을 사용하였다.

4.1 시간정보에서의 성능평가

총 437개의 시간정보들은 277개의 출발시간정보, 160개의 도착시간정보로 이루어져있다. 각각의 성능평가결과는 표 4, 5, 6의 내용과 같으며 각 표에서 사용된 특정항목들의 의미는 다음과 같이 정의된다.

- t_d1: 첫 번째 이동 출발시간으로 판단한 개수
- t_d2: 두 번째 이동 출발시간으로 판단한 개수
- t_a1: 첫 번째 이동 도착시간으로 판단한 개수
- t_a2: 두 번째 이동 도착시간으로 판단한 개수
- **t_d1**: 실제 첫 번째 이동 출발시간 개수
- **t_d2**: 실제 두 번째 이동 출발시간 개수
- **t_a1**: 실제 첫 번째 이동 도착시간 개수
- **t_a2**: 실제 두 번째 이동 도착시간 개수

표 4. 출발시간정보에서의 성능평가결과

	t_d1	t_d2	정확률 (precision)	재현율 (recall)	F-Measure	정확도 (accuracy)
t_d1	193	18	88.9	91.5	90.2	84.8
t_d2	24	42	70.0	63.6	66.7	

표 5. 도착시간정보에서의 성능평가결과

	t_a1	t_a2	정확률	재현율	F-Measure	정확도
t_a1	74	4	74.0	94.9	83.1	81.3
t_a2	26	56	93.3	68.3	78.9	

표 6. 시간정보에서의 비교성능평가결과

		t_d1	t_d2	t_a1	t_a2	정확도
미 사 용	t_d1	179	12	9	11	63.8
	t_d2	18	37	4	7	
	t_a1	45	4	24	5	
	t_a2	27	8	8	39	
사 용	t_d1	192	19	0	0	81.5
	t_d2	25	41	0	0	
	t_a1	0	0	62	16	
	t_a2	0	0	21	61	

특히 출발시간정보에서는 첫 번째 이동 출발시간 개수가 다른 출발시간 개수에 비해 상대적으로 많아 정확도가 가장 높았다. 출발시간정보는 표 2의 문맥정보들 중 자질 선택 기법을 통해서 선택된 3(시간), 4(시간), 21(장소), 25(항공기회사), 31(항공기번호)번 정보에 큰 영향을 받았다.

도착시간정보는 1(시간), 21(장소), 22(장소), 29(항공기회사)번 정보에 큰 영향을 받았으며 실제 평가를 수행할 때도 위의 4가지 정보만을 가지고 평가하였다.

시간정보는 출발도착정보를 문맥정보로 사용했을 경우 명확한 성능향상을 보였으며 출발도착정보 미사용시에는 1(시간), 7(시간), 9(시간), 21(장소), 22(장소), 25(항공기회사), 31(항공기번호)번 정보에 큰 영향을 받은 반면, 출발도착정보 사용시에는 1(시간), 21(장소), 22(장소), 23(장소), 29(항공기회사), 31(항공기번호)번 정보에 큰 영향을 받았다.

즉 각각의 시간정보를 구분하기 위한 문맥정보의 사용은 서로 중복되는 항목도 있었으나 각기 다른 문맥정보를 추가로 사용함으로써 최적의 성능을 보임을 알 수 있었다. 또한 구분하고자 하는 예약정보의 1차 정보와 동일한 유형의 정보만을 가지고 이를 구분한다는 것은 한계가 있으며 다른 유형의 예약정보도 함께 고려해야 해당 예약정보에 대한 정보추출성능을 향상시킬 수 있음을 알 수 있었다.

4.2 날짜정보에서의 성능평가

총 237개의 날짜정보들은 189개의 출발날짜정보, 48개의 도착날짜정보로 이루어져 있으며 각각의 성능평가결과는 표 7, 8, 9의 내용과 같다. 각 표에서 사용된 특정항목들의 의미는 다음과 같이 정의된다.

- d_d1: 첫 번째 이동 출발날짜로 판단한 개수
- d_d2: 두 번째 이동 출발날짜로 판단한 개수
- d_a1: 첫 번째 이동 도착날짜로 판단한 개수
- d_a2: 두 번째 이동 도착날짜로 판단한 개수
- **d_d1**: 실제 첫 번째 이동 출발날짜 개수
- **d_d2**: 실제 두 번째 이동 출발날짜 개수
- **d_a1**: 실제 첫 번째 이동 도착날짜 개수
- **d_a2**: 실제 두 번째 이동 도착날짜 개수

표 7. 출발날짜정보에서의 성능평가결과

	d_d1	d_d2	정확률	재현율	F-Measure	정확도
d_d1	169	4	94.4	97.7	96.0	92.6
d_d2	10	6	60.0	37.5	46.2	

표 8. 도착날짜정보에서의 성능평가결과

	d_a1	d_a2	정확률	재현율	F-Measure	정확도
d_a1	23	3	85.2	88.5	86.8	85.4
d_a2	4	18	85.7	81.8	83.7	

표 9. 날짜정보에서의 비교성능평가결과

		d_d1	d_d2	d_a1	d_a2	정확도
미 사 용	d_d1	166	2	2	3	84.8
	d_d2	8	7	0	1	
	d_a1	12	0	13	1	
	d_a2	3	1	3	15	
사 용	d_d1	168	5	0	0	92.0
	d_d2	7	9	0	0	
	d_a1	0	0	25	1	
	d_a2	0	0	6	16	

출발날짜정보에서는 첫 번째 이동 출발날짜의 개수가 상대적으로 너무 많아 정확도가 가장 높은 반면에 두 번째 이동 출발날짜의 F-Measure결과가 가장 낮았다. 출발날짜정보는 표 2의 문맥정보들 중 1(날짜), 3(날짜), 4(날짜), 10(날짜), 22(장소), 31(항공기번호)번 정보에 큰 영향을 받았으며 실제 평가를 수행할 때도 위의 정보만을 가지고 평가하였다.

도착날짜정보는 1(날짜), 21(장소), 26(항공기회사)번 정보에 큰 영향을 받았으며 출발날짜정보에서와 마찬가지로 위와 같은 최적의 문맥정보만을 사용하여 성능평가를 수행하였다.

날짜정보는 출발도착정보를 문맥정보로 사용했을 경우 상대적인 성능향상을 보였으며 출발도착정보 미사용시 1(날짜), 3(날짜), 9(날짜), 10(날짜), 21(장소), 31(항공기번호)번 정보에, 출발도착정보 사용시 1(날짜), 31(항공기번호)번 정보에 큰 영향을 받았다.

각각의 날짜정보는 앞서 수행한 시간정보의 결과와 마찬가지로 각각의 날짜정보에 최적화된 문맥정보를 사용함으로써 최적의 성능을 보일 수 있었으며 이때 사용된 문맥정보들은 구분하고자 하는 예약정보의 유형 이외의 정보들을 포함하고 있음을 알 수 있었다.

5. 결론

대화형 에이전트를 구현하기 위해서는 자연언어처리 기술에 바탕을 둔 다양한 연구 분야들의 적용이 필요하며 각각의 분야들은 자신의 성능을 최대한 발휘함으로써 성공적인 대화형 에이전트를 구현할 수 있다. 본 논문은 대화형 에이전트의 처리단계들 중 초반부에 위치하는 사용자 입력에 대한 정보추출단계에 대하여 다루었으며 특히 정보유형의 계층화와 계층화된 정보들 중 처리하기 어려운 정보에 대한 문맥정보를 사용한 정보추출의 성능향상을 보여주었다. 또한 본 실험을 통해서 정보의 유형에 따라 다른 문맥정보에 영향을 받는다는 사실을 알아냄으로써 계층화의 유용성을 보다 명확하게 보일 수 있었다. 하지만 한정된 자원으로 인한 대용량 말뭉치에서

의 성능평가를 수행하지 못해 제시한 방법론에 대한 보편적인 효용성을 보일 수 없었다. 향후에 대용량 대화말뭉치를 사용하여 현재의 방법론의 유용성 및 개선점 또는 새로운 방법론에 대한 연구가 필요하며 한국어 적용 및 특정 언어에 특화된 유용한 방법론에 대한 연구가 요구된다. 기존의 개체명인식 방법론을 대화에서의 정보추출, 즉 이미 정의된 정보 슬롯을 채우기 위한 방법으로 대화시스템에 적용하는 연구가 함께 병행되어야 할 것이며 기존의 패턴매칭 등의 기본적인 방법도 다양한 방향으로 연구 및 모색되어야 할 것이다.

감사의 글

본 연구는 첨단정보기술 연구센터를 통한 과학재단 및 2007년도 두뇌한국21사업의 지원을 받았습니다.

참고문헌

- [1] 김수영, 조성배, "순차적 패턴매칭 기법을 이용한 대화형 도우미 에이전트" 한국정보과학회 학술발표논문집, vol. 27, no. 2, pp. 24-26, 2000.
- [2] SRI ATIS Collection, <http://www.ai.sri.com/~communic/amex/amex.html>
- [3] J. Weizenbaun, "ELIZA - a computer program for the study of natural language communication between man and machine," Communication of the ACM, vol. 9, no. 1, pp. 36-45, 1965.
- [4] A.L.I.C.E. Artificial Intelligence Foundation, <http://www.alicebot.org>
- [5] AutoTutor, <http://www.autotutor.org>
- [6] B. J. Grosz, et al., "Centering: a framework for modeling the local coherence of discourse," Computational Linguistics, vol. 21, no. 2, pp. 203-225, 1995.
- [7] W. M. Soon, et al., "A machine learning approach to coreference resolution of noun phrases," Computational Linguistics, vol. 27, no. 4, pp. 521-544, 2001.
- [8] A. Stolcke, et al., "Dialogue act modeling for automatic tagging and recognition of conversational speech," Computational Linguistics, vol. 26, no. 3, pp. 339-373, 2000.
- [9] 김경선, 서정연, "자질 선택 기법을 이용한 한국어 화행 결정," 정보과학회논문지: 소프트웨어 및 응용, vol. 30, no. 3, pp. 278-284, 2003.
- [10] J. H. Eun, et al., "An information extraction approach for spoken language understanding," Proceedings of the 8th International Conference on Spoken Language Processing, 2004.
- [11] A. McCallum, et al., "Improving text classification by shrinkage in a hierarchy of classes," Proceedings of the 15th International Conference on Machine Learning, pp. 359-367, 1998.