대화 시스템의 말뭉치 구축을 위한 Object-Action 반자동 추출기

윤정민○, 황재원, 고영중
동아대학교 컴퓨터공학과
{yjungmin2, sftcap, youngjoong.k}@gmail.com

Semi-Automatic Object-Action Extractor to Build the Utterance Corpus for the Dialogue System

JungMin Yoon○, Jaewon Hwang, Youngjoong Ko
Donga University, Department of Computer Engineering

요 약

본 논문은 대화 시스템에서 사용되는 말뭉치의 구축을 위해 Object와 Action을 반자동으로 추출하는 도구에 대해 기술한다. 제안하는 추출 도구는 형태소 분석과 의존 구문 분석의 결과를 기반으로 적절한 Object와 Action을 추출하는 것에 목표를 두고 있다. 그러나 형태소 분석과 의존 구문 분석의 결과는 여러 가지 오류가 포함될 수 있다. 이러한 오류는 잘못된 Object와 Action의 추출로 이어질 수 있다. 그리고 Object의 추출에 있어 해당 명사의 격이 중요한 정보를 가진다. 하지만 한국어의 특성한 조사의 생략 등으로 인해 격 태깅의 모호성이 발생하게 된다. 따라서 본 논문에서 제안하는 반자동 추출기는 형태소 분석과 의존 구문 분석의 잘못된 결과를 사용자가 손쉽게 수정할 수 있도록 하고 모호성이 발생할 수 있는 Object를 사용자에게 알려주어 올바른 Object와 Action의 추출을 가능하게 한다. 추출기를 이용한 말뭉치의 구축은 1) 형태소 분석 2) 의존 구문 분석 3) Object-Action 추출의 단계로 진행된다. 실험에서 사용된 발화는 관광 회화용 대화 시스템의 숙박, 공항 영역의 500개의 발화이며, 이 중 259개의 발화가 태깅 시 모호성이 발생하는 발화이다. 반자동 추출기를 통해 모호성이 발생한 발화를 태깅한 결과 전체 발화 중 51.8%의 발화를 빠르고 정확하게 태깅할 수 있었다.

주제어: 대화 시스템, 말뭉치 구축, 반자동 추출기

1. 서론

대화 시스템은 인간 사용자와 시스템 에이전트 사이에 정보를 주고받기 위해 자연어 인터페이스를 통해 자연스러운 대화를 수행하는 목적을 가지고 있다. 자연스러운 대화를 위해서는 시스템 에이전트가 사용자의 발화를 분석하여 적절한 응답을 생성 할 수 있어야 한다[1]. 대화 시스템은 일반적으로 자연어 이해, 대화 관리, 자연어 응답생성 등의 모듈로 구성된다[2].

자연어 이해를 위해서는 사용자의 발화의 의도를 파악하는 것이 중요하다. 일반적으로 발화의 의도는 발화내의 주어, 목적어, 동사가 결정하게 된다. 본 논문에서는 이를 Object, Action이라고 정의하고 이를 추출하는 것에 목표를 둔다. Object와 Action은 대화 시스템의 발화에서 사용자가 보다 중요하다고 생각되어지는 명사와 동사가 되게 된다. 발화내의 Object와 Action 예는 그림 1과 같다.

대화 시스템 구축을 위해서는 각 발화의 Object와 Action을 포함하는 말뭉치를 구축할 필요가 있다. 이를 위해서는 사용자가 수작업으로 문장의 형태소 분석 결과와 의존 구문 분석의 결과를 확인하고 수정하는 작업이 필요하며, 이는 많은 시간과 비용이 드는 작업이다. 그러나 만약 손쉽게 Object와 Action을 추출하는 것을 도와주는 도구가 있다면 시간과 비용을 절감할 수 있을 것이다. 하지만 한국어의 특성상 성분의 생략과 동사 어미의 변화[3]가 심하여 언어 분석기의 성능이 떨어지게 되고

이 논문은 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (No. NRF-2013R1A1A2009937)

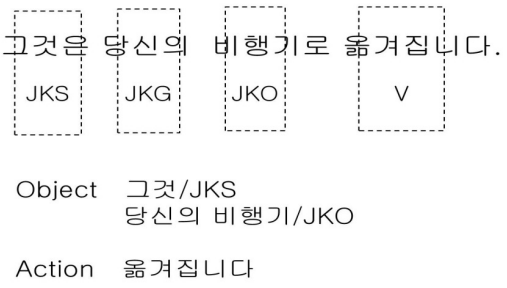

그림 1 Object Action추출 예

올바른 결과를 추출하기가 힘들어 자동 추출에는 한계가 있어 반자동 추출기를 고려하였다. 반자동 추출기에서는 여러 가지 오류를 포함하는 형태소 분석 결과와 의존 구문 분석 결과를 수정할 수 있게 하여, 수정된 결과를 토대로 Object-Action 쌍을 추출할 수 있다. 이렇게 추출된 Object와 Action은 말뭉치 기반 대화 시스템에서 적절한 발화를 찾을 때 사용된다. 이 때, 명사인 Object들이 어떠한 조사와 함께 사용되었는지가 중요한 정보가 된다. 이는 실제 대화 시스템에서 발화의 의도 분석에도 중요하게 이용되게 된다. 따라서 본 추출기에서는 Object에 격조사 태그를 함께 부착하여 Object의 격을 구분할 수 있도록 하였다. 이 경우 모호성이 발생할 수 있는 부분을 사용자에게 제시하여 빠르고 정확하게 태깅할 수 있도록 하였다. 제안하는 추출기의 실행 화면은 그림 2와 같다.

본 논문의 구성은 다음과 같다. 2장에서는 Object-Action 반자동 추출기의 구성에 대해 기술하고 3장에서는 반자동 프로세스, 격조사태그에 대해 기술한다. 4장에서는 본 논문에서 제안하는 추출기의 성능을 평가하고 마지막으로 결론에 대해 기술한다.

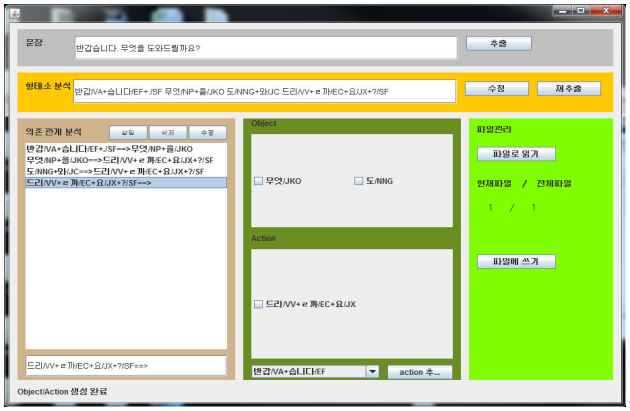

그림 2. Object-Action 반자동 추출기 실행 화면

2. Object-Action 반자동 추출기 구성도

Object-Action 반자동 추출기의 개략적인 구성도는 그림 3과 같다.

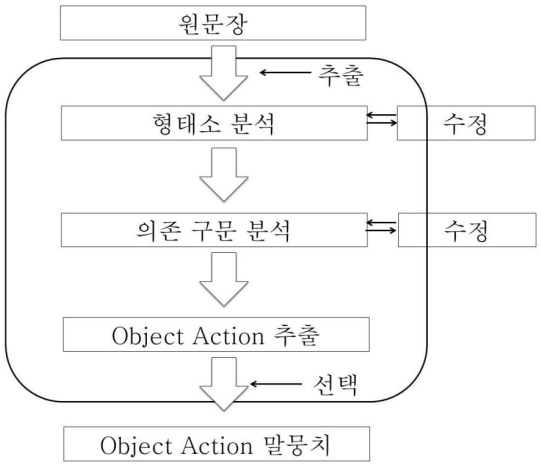

그림 3. Object-Action 반자동 추출기의 시스템 구성도

그림 3의 구조는 논리적 구조를 나타낸 것으로, 크게 3가지의 단계를 따르게 된다. 첫 번째 단계로 문장을 분석하여 형태소 형태로 나타내어 주는 형태소 분석단계, 두 번째 단계로 형태소 분석 결과를 바탕으로 어절 간의 의존 관계를 분석하는 의존 구문 분석단계와 마지막으로 두 단계를 거쳐서 조건에 맞는 Object와 Action을 슬롯으로 나타내는 Object-Action 추출 단계로 구성된다. 앞선 두 단계에서는 형태소 분석기, 의존 구문 분석기가 이용된다.

3. 반자동 프로세스

3.1.프로세스

명사의 추출은 전적으로 형태소 분석기의 성능에 의존한다. 명사의 형태소는 변화가 적기 때문에 형태소 분석기에서 명사로서의 판단은 어렵지 않다. 형태소 분석 결과에서 명사가 연속으로 나열 된 경우 합성명사로 판단하여 본 명사와 합성명사를 모두 슬롯에 나타내어 선택의 폭을 넓힌다. 형태소 분석 결과가 잘못되었을 경우에는 잘못된 형태소의 태그를 수정함으로써 수정된 결과를 슬롯에 나타낼 수 있다. 표 1은 잘못된 형태소 분석결과를 보여준다. 표 2는 형태소 분석 결과에 의존한 잘못된 Object-Action 추출결과를 보여준다.

표 1 잘못된 형태소 분석의 예

문장	성함을 말해주세요.
형태소분석결과	성함/NNG+을/JKO 말/NNG+해/NNG 주세/NNG+요/NNG+./SF

표 2 잘못된 Object-Action 추출의 예

Object	성함/JKO 말/NNG 해/NNG 말해주세요/NNG
Action	x

하지만 동사의 경우 동사 활용이 다양하여 올바른 형태소를 찾기에 어려움이 있다. 따라서 형태소 분석 결과에 이어 각 어절 간의 의존 관계 결과를 이용한다. 의존 관계는 한 구성요소와 다른 구성요소와의 문법적 관계를 밝히는 것을 의미한다. 따라서 어순이 비교적 자유롭고, 문법적 관계에 있는 구성요소들이 불연속적으로 나타날 수 있는 한국어 문장 분석에 적합하다. 의존관계에 따라 지배소는 여러 개의 의존소들을 가질 수 있으나 의존소는 하나의 지배소만을 갖는다[4]. 동사는 지배소이므로 여러 개의 의존소를 가진다. 따라서 의존 구문 분석 결과에서 여러 개의 의존소를 가지는 어절을 Action으로 선택하여 슬롯에 나타낸다.

3.2.격조사 태그

한국어의 경우 격조사에 의해 구문의 의미가 부여되고 하나의 격조사가 서술어의 특징에 따라 다양한 의미를 가지는 특징을 가진다[5]. 본 추출기에서는 말뭉치를 구축할 때 격조사에 대한 태깅을 실시한다. 본 논문에서 사용한 격조사의 예는 표 3과 같다.

표 3 격조사의 예

격조사 JK	주격조사 JKS	-이/-가,-께서,-에서
	목적격조사 JKO	-을/-를
	부사격조사 JKB	-에, -에서,-에게 -에게서, -한테 -(으)로, -와/과...
	관형격조사 JKG	-의
	보격조사 JKC	-이/-가
서술격조사	서술격조사 VCP	-이다

태깅 방식은 각 Object에 붙은 격조사를 분석하여 Object의 명사태그를 제거하고 그 자리에 격조사 태그를 부착하여 표현하였다. 예는 표 4와 같다.

표 4 격조사 태그 부착의 예

문장	철수가 영희를 좋아한다
형태소 분석결과	철수/NNG+가/JKS 영희/NNP+를/JKO 좋아하/VV+다/EF+./SF
Object	철수/JKS, 영희/JKO

3.3. 격조사 태깅 시 모호성 문제

한국어의 특성상 대화문에서 조사 생략은 빈번하게 나타난다. 생략의 정도는 대체로 문법적 관념의 정도에 비례한다. 문법적 관념은 어느 정도의 통사 구조에 의해 예측될 수 있으며, 문법적 관념의 비중이 높은 쪽은 생략이 가능한 것이다[6]. 이처럼 조사의 생략은 격조사의 태깅에서 모호성을 발생시키고 주격조사 자리에 쓰인 보조사 또한 주격과 목적격의 모호성을 유발하게 된다. 따라서 격조사 태깅에 2가지의 모호성이 존재하게 되며 그 예는 다음과 같다.

1) 조사가 생략된 경우
   예: 표 몇 장 필요하세요?

2) 보조사가 사용된 경우
   예: 미용실은 예약해야합니까?

1)의 예에서 '표' 는 주격으로 추출되어야 하나 주격조사가 생략되어 격을 결정할 수 없다. 2)의 예에서는 '미용실' 은 목적격으로 추출되어야 하나 보조사 '은' 의 경우 주격과 목적격으로 모두 사용될 수 있어 모호성이 발생한다. 이 같은 문제로 인해 반자동 추출기의 도움이 없이 태깅을 할 경우 사람이 일일이 이 같은 현상을 파악하여 태깅할 필요가 있어 많은 시간이 소요된다. 하지만 반자동 추출기에서는 이 같이 모호성이 발생할 수 있는 부분을 하이라이트해서 제시하여 태깅하는 사람이 쉽게 모호성이 있는 부분을 발견하고 이를 해결할 수 있도록 하였다.

3.4. 대화 시스템을 위한 Object-Action의 활용

코퍼스 기반 대화 시스템은 사용자의 입력이 들어 왔을 경우, 미리 구축한 인식 가능한 발화 리스트 중에서 사용자가 입력한 발화랑 의미가 가장 유사한 것을 어떻게 찾는가에 의해 대화 시스템의 성능이 결정된다. 이를 위해서는 사용자의 의도를 효과적으로 파악하는 것이 중요한데 Object와 Action이 중요한 실마리가 될 수 있다. 그래서 이를 효과적으로 추출하여 인식 가능한 발화 리스트와의 발화 간 유사도를 계산할 때 중요한 자질로 사용할 수 있다.

사용자가 발화를 입력하게 되면 대화 시스템은 입력된 발화의 Object와 Action을 추출하고 추출된 Object와 Action을 바탕으로 말뭉치 내의 발화들 중 동일한 Object와 Action을 가지는 발화들을 대상으로 유사도를 계산한다. 그리고 코사인 유사도를 기반으로 적절한 발화를 찾게 된다. 입력 발화와 추출된 발화의 예는 그림 4와 같다.

입력 발화 : 예약을 할 수 있습니까?

Object : 예약

Action : 하다

추출 발화 : 예약을 하고 싶습니다.

그림 4 입력 발화와 추출 발화의 예

4. 시스템평가

4.1. 평가 데이터 및 환경

평가에 사용된 데이터와 평가 방법은 다음과 같다.

- 평가 데이터
  : 영어 회화 어플리케이션 'Just Trip'의 숙박, 공항 영역
  : 총 500개의 발화

- 평가 환경
  : Object-Action 반자동 추출기 이용하여 모호성을 가지는 발화의 모호성 해결 및 형태소 분석의 오류 및 구문 분석 결과의 오류를 반자동 추출기를 이용하여 수정하여 오분석 해결

4.2. 반자동 추출기를 이용한 발화의 모호성 해결 및 오분석 결과 수정 결과

표 5 모호성이 존재하는 발화와 오분석 발화의 통계

	발화 수	비율
(1)	94	18.8%
(2)	99	19.8%
(3)	81	16.2%
합계	259	51.8%

(1)은 조사가 생략되어 모호성이 발생하는 발화를 의미하며, (2)는 보조사가 사용되어 모호성이 발생하는 발화이며, (3)은 언어 분석의 결과 오분석되어 잘못된 Object-Action쌍이 추출된 발화이다. 합계의 259개의 발화가 (1), (2), (3)의 발화의 합계랑 차이가 나는 이유는 3가지의 경우에 중복되는 발화 15개를 제거한 결과이다.

이처럼 반자동 추출기를 이용하여 발화의 모호성과 오분석 발화를 수정함으로써 전체 발화의 51.8%에 해당하는 모호성이 존재하는 259개의 발화를 정확히 태깅을 할 수 있어 대화 시스템을 위한 양질의 말뭉치를 구축할 수 있었다.

5.결론

본 논문은 관광회화 대화 시스템의 말뭉치 구축을 위한 Object-Action 반자동 추출기에 대해 기술하였다. Object-Action 반자동 추출기를 활용함으로써 말뭉치 구축 시에 모호성을 빠르게 제거하여 태깅을 위한 시간 절약과 정확률의 향상에 도움을 줄 수 있었다.

현재 이렇게 구축된 관광회화 대화 시스템의 말뭉치는 후처리하여 관광회화 어플리케이션 'Just Trip' 의 발화 말뭉치로 활용되고 있다.

6.참고문헌

[1] 홍금원, 이정훈, 신중휘, 이도길, 임해창. “대화시스템의 로그를 이용한 대화예제의 자동 확충에 관한 연구", 한국HCL학회 학술대회, pp 257-262, 2009.
[2] 김석환, 이창재, 정상근, 이근배. “EPG 정보 검색을 위한 예제 기반 자연어 대화 시스템”, 정보과학회논문지 : 소프트웨어 및 응용 34, pp.123-130, 2007.
[3] 김형준, 임동희, 강승식, 은지현, 장두성. “세종 계획 말뭉치를 이용한 품사 태거의 성능 개선”, 한국컴퓨터종합학술대회 논문집, Vol.34, No.1, 2007.
[4] 홍영국, 이종혁, 이근배. “의존문법에 기반을 둔 한국어 구문 분석기”, 한국정보과학회 학술발표논문집 20, pp 781-784, 1993.
[5] 정석원, 박의규, 나동열, 윤준태. “격 관계와 상호정보를 이용한 한국어 의존 파서”, 한국정보과학회 학술발표논문집, pp 450-455, 2001.
[6] 권재일. “조사의 성격과 그 생략 현상에 대한 한 기술 방법”, 어학연구 25, pp129-139, 1989.