대화 모델링을 위한 음성 언어 이해 기술

류 성 한
포항공과대학교

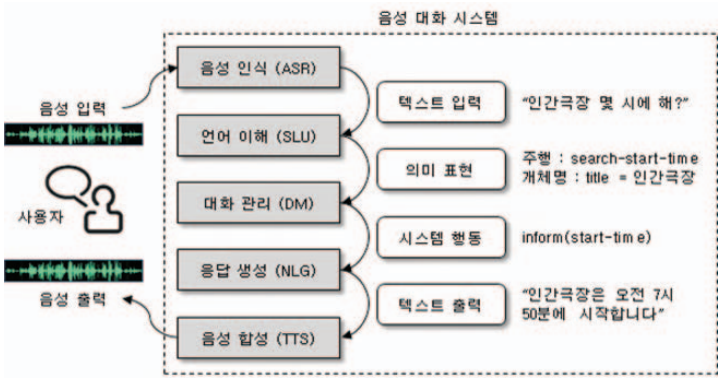
이 근 배
포항공과대학교

Ⅰ. 서 론

대화는 사람과 사람사이의 자연스러운 의사소통 수단이며, 이러한 대화를 사람과 컴퓨터 사이의 상호작용에 적용한 것을 음성 대화 시스템이라고 한다. 음성 대화 시스템은 사용자의 음성 입력을 이해해 사용자가 원하는 서비스를 제공한다. 예를 들어 사용자는 대화를 통해 날씨 정보를 얻고, 주변 맛집을 검색하고, SNS에 글을 작성하고, 기기를 제어하는 등의 일을 할 수 있다. 이러한 음성 대화 시스템은 최근 음성 인식 및 자연어 처리 기술의 비약적인 발전에 따라 학계 및 산업계에서 차세대 인터페이스로 주목받고 있으며, 스마트폰, 스마트 TV 등 다양한 기기에 탑재되어 서비스되고 있다.

음성 대화 시스템은 〈그림 1〉과 같이 크게 다섯 개의 순차적인 과정으로 구성된다.

- 사용자가 발화한 음향 신호를 자연어 문장으로 변환해주는 음성 인식 (Automatic Speech Recognition)



〈그림 1〉 음성 대화 시스템 구조도

〈표 1〉 의미 프레임의 예

예 1	"인간극장 몇 시에 해?"		
	의미 프레임	주행	search-start-time
		개체명	title = 인간극장
예 2	"KBS 뉴스 틀어"		
	의미 프레임	주행	play-program
		개체명	channel = KBS genre = 뉴스
예 3	"그럼 유재석 나오는 거 보자"		
	의미 프레임	주행	play-program
		개체명	cast = 유재석

- 자연어 문장으로부터 사용자의 의도를 이해해 컴퓨터가 처리할 수 있는 형식으로 변환해주는 음성 언어 이해 (Spoken Language Understanding)
- 사용자의 의도에 부합하는 서비스를 제공하기위해 시스템의 행동을 결정하는 대화 관리 (Dialog Management)
- 시스템의 행동을 구체적인 자연어 문장으로 생성해주는 응답 생성 (Natural Language Generation)
- 생성된 자연어 문장을 음성으로 합성해주는 음성 합성 (Text-to-Speech Synthesis)

본 연구는 음성 대화 시스템의 다섯 과정 중 음성 언어 이해 기술의 개요를 소개하고 음성 언어 이해 기술이 풀어야 할 문제점들과 연구 동향을 소개한다.

Ⅱ. 음성 언어 이해 기술 개요

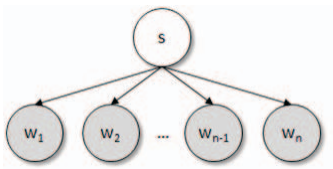
1. 음성 언어 이해의 목표

음성 언어 이해의 목표는 자연어 문장으로부터 사용자의 의도를 이해해 컴퓨터가 처리할 수 있는 의미 표현(semantic representation)으로 변환하는 것이다. 이러한 의미 표현은 대화 관리 모듈이 데이터베이스를 검색하기 위한 질의문을 생성하거나 서비스 에이전트에게 서비스를 요청하기 위한 매개변수로 사용될 수 있다. 대부분의 경우 의미 표현은 〈표 1〉과 같이 속성-값 조합들로 구성된 의미 프레임(semantic frame) 형태로 정의되며, 이 때 의미 프레임의 속성들은 음성 대화 시스템이 제공하는 특정한 영역에 적합하게 정의된 한 개의 주행(dialog act) 및 복수 개의 개체명(named entity)이다. 주행이란 문장에서 사용자가 의도한 행위의 유형을 나타내는 속성이고, 개체명이란 genre, title, channel, start-time, end-time과 같이 사용자가 의도한 행위의 매개변수를 나타내는 속성이다. 예를 들어 "그럼 SBS 드라마 는 언제 하지?" 라는 문장에서 주행의 값은 search-start-time, channel 개체명의 값은 "SBS", genre 개체명의 값은 "드라마"가 된다. 결과적으로 음성 대화 시스템에서의 의미 이해 과정은 한 개의 주행을 식별하는 주행 식별(dialog act identification) 및 복수의 개체명을 인식하는 개체명 인식(named entity recognition)과정으로 구성된다.

데이터 기반 접근법은 비문법적인 자연어 문장도 강인하게 이해 할 수 있으며, 다른 영역에 적용하기 위해서는 방법론을 교체 할 필요 없이 새로운 데이터를 수집하면 되므로 이식성이 뛰어나다는 장점이 있다.

2. 음성 언어 이해 방법

음성 언어 이해 기술은 크게 지식 기반의 접근법(knowledge-based approach)와 데이터 기반 접근법(data-driven approach)으로 나눌 수 있다. 지식 기반 접근법은 전문가에 의해 정의된 문법을 바탕으로 음성 언어 이해 과정을 수행하며 높은 정밀도를 갖는다. 그러나 사람이 사용하는 자연어 문장은 종종 비문법적이기 때문에 지식 기반 접근법은 음성 언어 이해에 실패할 수 있어 낮은 재현율을 갖는다는 한계가 있다. 반면 데이터 기반 접근법은 수많은 문장 및 각 문장에 달린 정답 레이블로 구성된 말뭉치(corpus)를 바탕으로 기계



〈그림 2〉 Naive Bayes classifier의 graphical model

학습(machine learning) 기법을 적용해 음성 언어 이해 과정을 수행한다. 그러므로 데이터 기반 접근법은 비문법적인 자연어 문장도 강인하게 이해 할 수 있으며, 새로운 데이터를 수집해 동일한 방법론을 적용하면 다른영역에 대한 음성언어 이해를 수행할 수 있으므로 이식성이 뛰어나다는 장점이 있다. 일반적으로 학계에서는 데이터 기반 접근법이 주로 연구되고 있으며, 산업계에서는 지식 기반의 접근법과 데이터 기반 접근법을 융합함으로써 높은 성능을 추구한다. 본 연구에서는 데이터 기반 접근법을 바탕으로 주행을 식별하고 및 개체명 인식하는 방법을 소개한다.

3. 주행 식별

문장을 일련의 단어들로 봤을 때, 기계 학습 관점에서 주행 식별은 단어열 $w_{1,n}$이 관측되었을 때 한 개의 주행 s*를 예측하는 sequence prediction 문제이다. 이 때 naive Bayes classifier, Maximum Entropy (MaxEnt), support vector machine (SVM) 등의 기계 학습 기법이 사용된다. 예를 들어 〈그림 2〉과 같은 naive Bayes classifier를 이용한 주행 식별은 주행의 확률 Pr(s)와 주행이 주어졌을 때 단어열의 확률 $Pr(w_{1,n}|s)$의 곱을 최대로 하는 주행 s*를 찾는 과정이며, 이러한 확률 값은 훈련 말뭉치를 통해 학습된다.

$$s^* = \text{argmax}_s \Pr(s) * \Pr(w_{1,n}|s) \quad (1)$$

이러한 주행 식별의 성능은 테스트 말뭉치를 대상으로 주행 식별을 수행한 결과의 정확도를 통해 측정한다.

$$\text{정확도} = \frac{\text{정답을 맞춘 문장의 수}}{\text{모든 문장의 수}} \quad (2)$$

4. 개체명 인식

〈표 2〉 개체명 인식을 위한 BIO 태깅의 예

입력 문장	"김연아 나왔 던 무한 도전 찾 아 줘"
개체명	cast = 김연아 title = 무한 도전
BIO 태깅된 문장	[cast-B 김연아] [O 나왔] [O 던] [title-B 무한] [title-I 도전] [O 찾] [O 아] [O 줘]

단 하나의 주행을 찾아내는 주행 식별과 달리, 개체명 인식은 단어열에서 개체명이 되는 부분을 찾아내는 sequence segmentation과 개체명의 유형을 찾아내는 classification을 수행한다. 일반적으로 sequence segmentation과 classification을 동시에 수행하기 위해 개체명이 레이블링된 문장을 BIO 태깅 기법을 사용해 표현한다. BIO 태깅 기법이란 〈표 2〉와 같이 개체명의 시작을 의미하는 B (Begin), 개체명의 연속을 의미하는 I (Inside), 개체명이 아님을 의미하는 O (Outside) 세 가지 형태의 접미사가 부착된 개체명 레이블을 사용해 문장을 표현하는 것을 말한다.

즉 문장을 일련의 단어들로 봤을 때, 기계 학습 관점에서 개체명 인식은 단어열 $w_{1,n}$가 주어졌을 때 레이블열 $s_{1,n}$*을 예측하는 sequence labeling 문제이다. 이 때 hidden Markov model (HMM), conditional random fields (CRF), Structured SVM 등의 기계 학습 기법이 사용된다. 예를 들어 〈그림 3〉와 같은 HMM을 이용한 개체명 인식은 레이블열의 확률 $Pr(s_{1,n})$과 단어 열의 확률 $Pr(w_{1,n} | s_{1,n})$의 곱을 최대로 하는 최적의 레이블열 $s_{1,n}$*를 찾는 과정이며, 이러한 확률 값은 훈련 말뭉치를 통해 학습된다.

$$s_{1,n}^* = \text{argmax}_{s_{1,n}} \Pr(s_{1,n}) * \Pr(w_{1,n}|s_{1,n}) \quad (3)$$

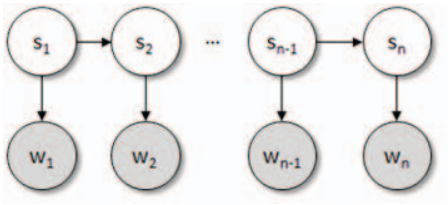
실행 단계에서 최적의 레이블열을 찾는 과정은 동적 프로그래밍 기법을 이용한 Viterbi 알고리즘에 의해 고속으로 수행된다.

이러한 개체명 인식의 성능은 테스트 말뭉치를 대상으로 개체명 인식을 수행한 결과의 정밀도, 재현율, F_1 점수를 통해 측정한다.

$$\text{정밀도} = \frac{\text{올바르게 예측한 개체명의 수}}{\text{예측한 개체명의 수}} \quad (4)$$

$$\text{재현율} = \frac{\text{올바르게 예측한 개체명의 수}}{\text{실제 개체명 수}} \quad (5)$$

$$F_1\text{점수} = \frac{2 * \text{정밀도} * \text{재현율}}{\text{정밀도} + \text{재현율}} \quad (6)$$



〈그림 3〉 HMM의 graphical model

Ⅲ. 음성 언어 이해 기술 연구 동향

Ⅱ 장에서는 음성 언어 이해 기술의 기본 방법론에 대해 다뤘다. 그러나 이러한 언어 이해 기술을 실제 응용에 적용하는 과정에서 성능 및 기능상의 문제가 발생하며, 이러한 문제를 해결하기 위해 학계에서는 다양한 방법론들이 연구되고 있다. 본 장에서는 음성 언어 이해 기술에 풀어야 할 문제점들과 이러한 문제를 해결하기 위한 연구 동향을 소개한다.

1. 자연어의 모호성 (ambiguity)

□ Long-distance dependency 문제

대부분의 음성 언어 이해 기술은 자연어 문장을 전체가 아닌 부분으로 이해한다. 예를 들어 "그럼 SBS 드라마 는 언제 하지?"라는 입력 문장에서 기계 학습 기법을 이용해 "드라마"의 개체명 레이블을 예측하기 위해서는 〈표 3〉과 같이 해당 단어의 일정 거리 이내의 주변 단어들의 조합을 특징(feature)으로 사용한다. 위와 같이 문장을 부분으로 이해하는 방법은 음성 인식 오류를 포함하거나 비문법적인 문장을 이해하는데 강인하다는 장점을 가진다.

〈표 3〉 개체명 인식을 위한 특징 추출의 예

입력 문장	"그럼 SBS 드라마는 언제 하지?"
"드라마"의 특징	word=드라마 word-1=SBS word+1=는 word-2=그럼 word+2=언제 word-bigram=SBS,드라마 +word-bigram=드라마,는 ...

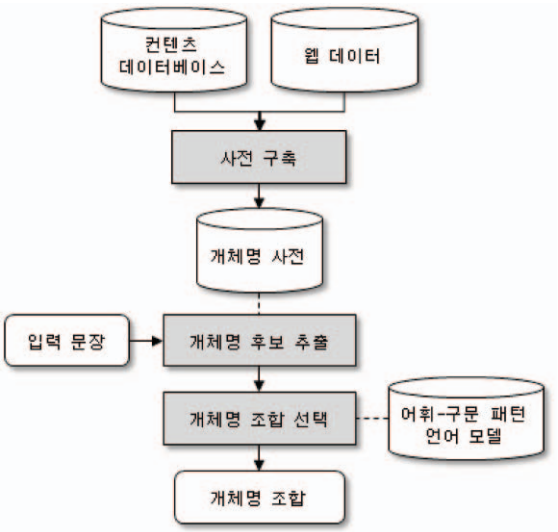
대부분의 음성 언어 이해 기술은 자연어 문장을 전체가 아닌 부분으로 이해한다.

그러나 자연어가 갖고 있는 모호성으로 인해, 길고 복잡한 문장을 이해하기 위해서는 고도의 기술이 요구된다. 예를 들어 비행기 예약에 대한 인터페이스의 한 입력 문장 "...fly from denver to chicago on dec. 10th 1999..."에서 "dec."은 출발하는 달을 나타내는 개체명 depart-month이며, 다른 입력 문장 "...return from denver to chicago on dec. 10th 1999..."에서 "dec."은 돌아오는 달을 나타내는 return-month 이다[1]. 위 문장에서 "dec."의 개체명 레이블을 구별하기 위해서는 해당 단어로부터 여섯 칸 거리에 있는 "fly" 혹은 "return"이라는 단어를 참조해야 하지만, 기존 방법론은 위와 같이 먼 거리에 있는 단어를 참조 할 수 없으므로 개체명 인식에 실패 할 수 있다. 이러한 문제를 long-distance dependency 문제라고 한다.

이러한 문제를 해결하기 위해 문장의 구문(syntax)을 분석해 구문 구조를 활용 할 수 있다. 그러나 현재 구문 분석 기술의 정확도는 충분히 높지 않기 때문에 구문 분석 결과를 사용할 경우 오히려 음성언어 이해 정확도 저하 될 수 있다는 문제가 있었다. 최근 연구에서는 long-distance dependency를 반영하면서 성능 향상에 기여하는 trigger 특징 들을 훈련 말뭉치에서 자동으로 추출하였다[1]. Trigger 특징 (a → b)은 a 와 b가 높은 연관도를 갖는다는 것을 말하며, 예를 들어 입력 문장 "... fly from denver to hicago on dec. 10th 1999 ..."에서는 (return → dec.) 등의 trigger 특징을 추출 될 수 있다. 최종적으로는 훈련 말뭉치에서 추출할 수 있는 trigger 특징 중 실제로 성능 향상에 기여 할 수 있는 특징을 선택해 기계 학습 기법이 이러한 특징을 고려하도록 함으로써 long-distance dependency 문제를 극복한다.

□ World knowledge 없이 이해할 수 없는 문장

개체명 인식은 사람 이름, 위치, 조직과 같이 한 두 단어 수준의 명사 혹은 명사구이며 주변 문맥에 의해



〈그림 4〉 사전 기반 개체명 인식 과정

유추하기 쉬운 개체명에 대해서는 높은 성능을 갖는다. 그러나 단어의 주변 문맥만으로는 개체명을 예측하기 어려운 경우가 있다. 예를 들어 입력 문장 "내셔널 지오그래픽 틀어줘"에서 "내셔널 지오그래픽"의 개체명 유형이 channel-name 인지 아니면 title 인지를 주변 문맥만으로는 판단하는 것은 불가능하다. 그러나 실세계에 "내셔널 지오그래픽"이라는 채널은 존재하지만 프로그램 이름은 존재하지 않는다는 것을 알고 있는 경우 개체명의 유형을 channel-name으로 결정 할 수 있다. 즉 음성 언어 이해를 위해서는 언어적 지식뿐만 아니라 실세계에 대한 지식, 즉 world knowledge가 필요하다.

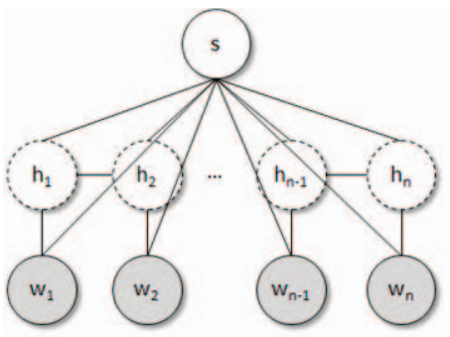
이러한 문제를 해결하기 위해 〈그림 4〉와 같이 각 개체명 유형별로 개체명의 목록을 저장하고 있는 개체명 사전을 구축하고 이러한 개체명 사전을 바탕으로 개체명을 인식하는 방법에 대한 연구가 진행되고 있다. 개체명 사전은 서비스 대상인 컨텐츠의 데이터베이스 혹은 위키피디아 등 웹 데이터를 바탕으로 구축 할 수 있다[2]. 실행 단계에서는 입력 문장으로부터 개체명 사전과 매칭되는 부분을 개체명 후보로 검출한 후, 그러나 실제로 개체명이 아닌 것이 개체명 후보로 검출되는 경우가 발생할 수 있으므로 개체명 후보들을 바탕으로 생성 할 수 있는 개체명 조합 중 가장 자연스러운 문장이

〈표 4〉 개체명 조합에 따른 어휘-구문 패턴의 예

입력 문장	"이 영화 보 면서 녹화 할 래"
조합 1	개체명 조합: ø 패턴: "이 영화 보 면서 녹화 할 래" Logprob.: -13.8, 순위: 2
조합 2	개체명 조합: {genre=영화} 패턴: "이 @genre 보 면서 녹화 할 래" Logprob.: -9.4, 순위: 1
조합 3	개체명 조합: {title=할} 패턴: "이 영화 보 면서 녹화 @title 래" Logprob.: -18.5, 순위: 4
조합 4	개체명 조합: {genre=영화, title=할} 패턴: "이 @genre 보 면서 녹화 @title 래" Logprob.: -14.1, 순위: 3

되는 조합을 선택한다. 예를 들어 입력 문장 "이 영화 보 면서 녹화 할 래"에서 개체명 사전을 바탕으로 title 개체명 후보로 "할", genre 개체명 후보로 "영화"를 검출 할 수 있다. 이러한 개체명 후보들을 바탕으로 〈표 4〉와 같은 선택 가능한 개체명 조합을 생성 할 수 있으며, 각 조합을 바탕으로 문장의 개체명 값을 개체명 유형으로 치환한 어휘-구문 패턴으로 변환한 후 가장 자연스러운 패턴인 "이 @genre 보 면서 녹화 할 래"를 생성한 genre 개체명 "영화"를 선택한다.

또한 사용자가 개체명의 이름 전체를 발화하지 않고 줄여서 발화하는 경우 이를 처리하기 위해 개체명의 약어 사전을 구축해야 한다. 예를 들어 TV 프로그램인 "무한도전"을 "무도"로, "슈퍼스타 K"를 "슈스케"로 발화 할 수 있다. 이러한 약어는 일정한 규칙이 아닌 사회적인 합의에 의해 결정되기 때문에 방대한 웹 데이터를 바탕으로 약어 사전을 구축 하는 방법이 주로 사용



〈그림 5〉 HCRF의 graphical model

된다. 한 연구는 위키피디아의 하이퍼링크, 리다이렉션, 동음이의어 페이지를 바탕으로 약어를 추출했다[2].

2. 기존 기능으로는 처리 할 수 없는 요청

▫ 단일 발화로 복수 여러 사용자 의도를 표현

대부분의 언어 이해 연구는 사용자가 발화하는 각 문장에서 한 개의 주행을 식별하는 것을 목표로 하고 있다. 그러나 종종 사용자는 두 개 이상의 주행을 포함한 문장을 발화하는 경우가 있다. 예를 들어 사용자는 TV 프로그램 영역의 음성 대화 시스템에 대해 "무한도전 예약하고 SBS 틀어줘"와 같이 두 개의 주행 record 및 play-channel 이 검출되는 문장을 발화할 수 있다. 이러한 문제를 해결하기 위해 record#play-channel와 같이 두 개 이상의 단일 주행을 조합한 새로운 다중 주행을 주행 집합에 추가하고, 다중 주행에 해당하는 문장들을 수집해 훈련 말뭉치에 추가함으로써 다중 주행 역시 예측하도록 할 수 있다.

그러나 주행의 개수를 한 문장에서 두 개로 제한하더라도 N개의 단일 주행이 존재하는 음성 대화 시스템의 다중 주행 개수는 N^2개가 된다. 즉 현실적으로 각 다중 주행에 대해 충분항 양의 문장을 수집하기 어려우므로 기존의 naive Bayes classifier, MaxEnt, SVM 등의 기계 학습 기법을 사용하는 경우 훈련 데이터가 충분하지 않아 성능에 악영향을 주는 data sparsity 문제가 발생한다.

현실적으로 각 다중 주행에 대해 충분항 양의 문장을 수집하기 어려우므로 기존의 naive Bayes classifier, MaxEnt, SVM 등의 기계 학습 기법을 사용하는 경우 훈련 데이터가 충분하지 않아 성능에 악영향을 주는 data sparsity 문제가 발생한다.

최근 연구에서는 이러한 문제를 해결하기 위해 기계 학습 기법인 Hidden-state CRF (HCRF)를 사용한다[3]. HCRF는 데이터의 숨겨진 내부 구조를 모델링 하는데 강점이 있다. 예를 들어 주행 record 와 주행 record#play-channel에 해당하는 문장들은 서로 유사한 부분을 갖는다. 기존 기계 학습 기법들이 이러한 양상을 전혀 고려 할 수 없었던 반면, HCRF는 〈그림 5〉과 같이 단어와 주행 사이에 latent variable을 배치함으로써 이러한 양상을 고려해 data sparsity 문제를 완화시켜 다중 주행 식별의 성능을 향상시킬 수 있다.

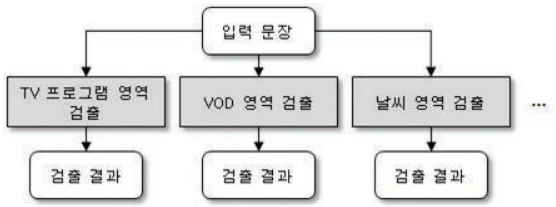
〈표 5〉 단일 발화를 통한 복수 영역에 대한 대화의 예

사용자 : "애니메이션 뭐 있는지 알려줘"
시스템 : "해당하는 TV 프로그램은 다음과 같습니다 – (...) 해당하는 VOD는 다음과 같습니다 – 아이스 에이지 3 (...)"
사용자 : "아이스 에이지 3에 출연한 사람들이 누구지?"
시스템 : "요청하신 TV 프로그램은 존재하지 않습니다. 해당 VOD의 출연진은 다음과 같습니다 – (...)"
시스템 : "그래 이거 보여줘"
시스템 : "해당 VOD를 재생합니다"

▫ 단일 발화로 복수 영역에 대한 서비스를 요청

다중 영역 음성 대화 시스템은 날씨, 음악, 주식, 음식점 등 다양한 영역(domain)에 대한 대화 서비스를 제공한다. 사용자가 발화한 문장으로부터 서비스를 제공하는 여러 개의 영역 중 사용자의 요청에 가장 부합하는 영역을 식별해 언어 이해, 대화 관리, 응답 생성 과정을 수행한다. 문장을 일련의 단어들로 봤을 때, 기계 학습 관점에서 영역 식별(domain identification)은 단어열 $w_{1,n}$이 관측되었을 때 한 개의 영역 s^*를 예측하는 sequence prediction 문제이며 주행 식별과 유사한 기계 학습 기법이 사용 될 수 있다.

그러나 사용자는 한 번의 발화로 두 개 이상의 영역에 대한 서비스를 요청 할 수 있다. 예를 들어 〈표 5〉과 같이 TV 프로그램과 VOD에 대한 서비스를 제공하는 음성 대화 시스템의 경우 사용자의 "애니메이션 뭐 있는지 알려줘"와 같은 문장에 대해 두 영역을 모두 검출 할 수 있어야 한다. 이러한 것을 다중 영역 검출(multi-domain detection)이라고 하며, 기계 학습 관점에서 다중 영역 검출은 단어열 $w_{1,n}$이 관측되었을 때 해당 단어열이 관측될 수 있는 한 개 이상의 영역을 검출하는 multi-label classification 문제이다.



〈그림 6〉 다중 영역 검출 과정의 예

〈표 6〉 개체명 및 미등록어 치환 과정의 예

입력 문장	"my brother is going to party tonight"
개체명 치환	"my brother is going to party @TIME"
미등록어 치환	"my $OOV is going to $OOV @TIME"

이러한 문제를 해결하기 위해 〈그림 6〉과 같이 각 영역별로 문장이 해당 영역에 해당하는지 여부를 판별하는 binary classification을 통해 다중 영역 검출 과정을 수행할 수 있다. 최근 연구에서는 각 영역 및 영역 간의 관계에 대한 정의를 바탕으로 binary classification의 정확도를 향상시키고 이러한 다중 영역 검출을 바탕으로 대화를 수행하는 다중 영역 음성 대화 시스템 구조를 개발하였다[4].

음성 인식 과정에서 오류가 발생하는 경우 이러한 오류가 이후 과정으로 전파(error propagation)되어 음성대화 시스템 전체의 성능에 악영향을 준다.

▫ 시스템이 제공하지 않는 영역에 대한 요청

사용자는 다중 영역 음성 대화 시스템이 어떠한 영역에 대해 서비스를 제공하는지 정확히 알지 못하므로 시스템이 제공하지 않는 영역에 대한 서비스를 요청 할 수 있다. 이러한 문장을 영역 외(out-of-domain) 문장이라고 하며, 〈그림 6〉과 같은 다중 영역 검출 과정에서 아무런 영역이 검출되지 않으면 해당 문장을 영역 외 문장으로 간주해 음성 대화 시스템이 거절 응답을 할 수 있다.

그러나 현실에서 발생할 수 있는 다양한 양상의 영역 외 문장에 대한 충분한 데이터를 수집하는 것은 불가능하므로 데이터 기반 접근법에 기반을 둔 음성 언어 이해는 영역 외 문장 검출에 어려움이 있다. 예를 들어 각 영역별 훈련 말뭉치에 한 번도 나오지 않은 단어, 즉 미등록어(out-of-vocabulary word)가 많이 등장하는 문장은 영역 외 문장일 확률이 높다. 그러나 기계학습 관점에서 훈련 과정에서 학습되지 않은 단어는 특징 공간(feature space)에 존재하지 않아 그 존재를 무시하므로 미등록어가 많이 등장하는 문장을 영역 외 문장으로 검출하는 것이 어렵다.

최근 연구에서는 이러한 문제를 해결하기 위해 기계학습 기법이 미등록어를 고려할 수 있도록 미등록어를 모델링한다[5]. 이 때 TV 프로그램 이름, 사람 이름 등의 고유명사 개체명의 경우 미등록어 일 지라도 영역 외 문장임을 판단하는 기준이 되어서는 안 된다. 그러므로 문장의 개체명 값을 개체명 유형으로 치환한 후 미등록어를 미등록어 태그로 치환한다. 예를 들어 〈표 6〉와 같이 입력 문장 "my brother is going to party tonight"에서 time 개체명인 "tonight"은 "@TIME"으로 변환하고 미등록어 "brother"와 "party"를 미등록어 태그 "$OOV"로 치환한다. 이러한 변환은 훈련 및 실행 과정에서 모두 수행되므로 결과적으로 기계 학습 기법이 미등록어를 고려해 영역 외 문장을 검출 할 수 있게 된다.

3. 오류 전파로 인한 성능 저하

▫ 음성 인식 오류

음성 대화 시스템의 첫 과정은 사용자가 발화한 음향 신호를 자연어 문장으로 변환해주는 음성 인식이다. 즉 음성 인식 과정에서 오류가 발생하는 경우 이러한 오류가 이후 과정으로 전파(error propagation)되어 음성 대화 시스템 전체의 성능에 악영향을 준다. 예를 들어 사용자가 발화한 실제 문장이 "예능 프로 틀어줘" 일 때 "예능 프로 들어줘"로 잘못 음성 인식 되는 경우, 음성 언어 이해 기술은 해당 문장으로부터 사용자가 의도한 play-program 이라는 주행을 알아내는데 어려움을 겪는다.

최근 연구에서는 이러한 문제를 해결하기 위해 음성 언어 이해 과정에서 단어와 발음 정보를 함께 활용한다[6]. 문자소를 음소로 변환하는 Grapheme-to-phoneme (G2P) 변환 기술을 이용해 자연어 문장을 발음열로 변환하고, 음성 언어 이해 과정에서 단어가 다르더라도 발음이 일치한다는 것을 활용한다. 예를 들어 문장 "예능 프로 틀어줘"와 "예능 프로 들어줘"의 발음이 유사하다는 것을 활용한다. 이를 통해 음성 인식 오류가 발생하더라도 강인하게 음성 언어 이해를 수행할 수 있도록 한다.

또한 불확실한 환경에서 행동을 계획하고 결정하는 partially observable Markov decision process (POMDP) 기술을 대화 관리에 적용해 음성 인식 오류로 인해 발생할 수 있는 사용자 의도의 불확실성을 해소 할 수 있다. POMDP 기반의 대화 관리는 사용자의 의도를 확률적인 신뢰 상태(belief state)로 표현하고, 사용자와의 대화를 통해 신뢰 상태를 업데이트해 나가면서 사용자에게 응답을 제공한다. 그러나 이러한 POMDP 기반 대화 관리는 복잡한 문제에 적용 할 경우 훈련에 지나치게 오랜 시간이 걸린다는 한계가 있다. 최근 연구에서는 이러한 문제를 해결하기 위해 신뢰 상태를 간략화된 상태(summary space)로 표현하고, 대화를 통해 추적해야 할 상태를 여러 개로 분할한 composite summary point-based value iteration (CSPBVI) 기법을 통해 POMDP 기반 대화 관리의 훈련 속도를 향상시켜 복잡한 문제에 적용 할 수 있도록 하였다[7].

□ 형태소 분석 오류

어절 단위로 띄어쓰기가 되어있는 한국어 문장을 이해하기 위해서는 형태소(morpheme) 분석을 통해 어절을 의미를 갖는 가장 작은 단위인 형태소 단위로 분리하고 각 형태소를 음성 언어 이해의 기본 단위로 처리할 필요가 있다. 예를 들어 음성 인식 출력 문장이 "예능프로 틀어줘"인 경우, 이를 "예능 프로 틀 어 줘"와 같이 형태소 단위로 분할해야 "예능"에 개체명 레이블 genre-b을 부여 할 수 있다. 그러나 "예능프 로 틀어줘"와 같이 잘못된 형태소 분석 결과를 바탕으로는 그것이 불가능하다. 즉 형태소 분석 과정에서 발생한 오류가 언어 이해 과정으로 전파된다.

〈표 7〉 음절 단위 개체명 인식

입력 문장	"예능프 로 틀 어 줘"
개체명	genre = 예능
BIO 태깅된 문장	[genre-B 예] [genre-I 능] [O 프] [O 로] [O 틀] [O 어] [O 줘]
"능"의 특징	syllable=능 syllable-1=예 syllable+1=능 syllable-bigram=예,능 left-space=false +syllable-bigram=능,프 right-space=false ...

이러한 문제를 해결하기 위해 한국어 음성 언어 이해를 단어가 아닌 음절 단위로 수행하도록 할 수 있다. 기계 학습 관점에서 음절 단위 개체명 인식 과정은 단어열이 아닌 음절열이 주어졌을 때 레이블열을 예측하는 sequence labeling 문제이다. 예를 들어 입력 문장 "예능프 로 틀 어 줘"에 대해 〈표 7〉과 같이 음절 "예"에 genre-B, 음절 "능"에 genre-I를 부여 하게 할 수 있다. 또한 개체명 인식을 위한 특징으로 음절 및 형태소 정보를 모두 활용 할 수 있다. 한국어에서 이와 유사한 방법을 명사 추출에 활용한 바 있다[8].

Ⅳ. 결 론

본 연구는 음성 언어 이해 기술의 개요를 소개했고, 음성 언어 이해 기술이 풀어야 할 문제점을 크게 세 가지 – 자연어의 모호성, 기존 기능으로는 처리 할 수 없는 요청, 오류 전파로 인한 성능 저하 – 로 나누고 각각의 문제를 해결하기 위한 연구 동향을 소개했다. 그러나 모호성은 자연어에서 다양한 양상으로 나타날 수 있으며, 사용자는 언제든지 기존 기능으로는 처리 할 수 없는 새로운 양상의 요청을 할 수 있고, 음성 인식 등의 기술은 아직 완벽히 풀리지 않은 문제이기 때문에 오류 전파의 가능성은 언제나 존재한다. 즉 음성 언어 이해 기술은 풀어야 할 문제가 많으며 응용 분야등을 고려하여 연구하고 적용해야 한다.

감사의 글

본 연구는 미래창조과학부 및 한국산업기술평가관리원의 산업융합원천기술개발사업(정보통신)의 일환으로 수행하였음. [10044508 , 비기호적 기법 기반 인간 모사형 자가학습 지능 원천기술 개발]

참 고 문 헌

[1] Jeong et al., "Practical Use of Non-local Features for Statistical Spoken Language Understanding", Computer Speech and Language, vol. 22, no. 2, pp. 148-170, April, 2008.

[2] Bøhn et al. "Extracting Named Entities and Synonyms from Wikipedia", IEEE International Conference on Advanced Information Networking and Applications, April, 2010.

[3] Xu et al., "Exploiting Shared Information for Multi-intent Natural Language Sentence Classification", Annual Conference of the International Speech Communication Association, August, 2013.

[4] Ryu et al., "A Hierarchical Domain Model-Based Multi-Domain Selection Framework for Multi-Domain Dialog Systems", International Conference on Computational Linguistics. Dec. 2012.

[5] Ryu et al., "Exploiting Out-of-Vocabulary Words for Out-of-Domain Detection in Dialog Systems", International Conference on Big Data and Smart Computing, Jan., 2014.

[6] Wang et al. "Improving Spoken Dialogue Understanding Using Phonetic Mixture Models", International Florida Artificial Intelligence Research Society Conference, May, 2011.

[7] Williams et al. "Scaling POMDPs for Dialog Management with Composite Summary Point-based Value Iteration", AAAI Workshop on Statistical and Empirical Methods in Spoken Dialog Systems, July, 2006.

[8] Lee et al. "A Syllable Based Word Recognition Model for Korean Noun Extraction", Annual Meeting of the Association for Computational Linguistics, July, 2003.

류 성 한

2012년 2월 동국대학교 컴퓨터공학과 학사
2012년 2월~ 현재 포스텍 컴퓨터공학과 통합과정

〈관심분야〉
음성 언어 이해, 자연어 처리

이 근 배

1984년 2월 서울대학교 컴퓨터공학 학사
1986년 2월 서울대학교 컴퓨터공학 석사
1991년 2월 UCLA 전산학 박사
1991년 9월~현재 포스텍 컴퓨터공학과 교수

〈관심분야〉
자연어 처리, 음성 인식, 음성 합성, 기계 번역