음성 언어 처리 기술, 어디까지 왔나

이경님
엔씨소프트 에이아이센터 스피치랩

1. 들어가기

사람이 의사소통하려고 사용하는 일반적이고 효과적인 수단은 언어(말과 글)이고, 음성은 인간의 가장 자연스러운 의사소통 방식이다. 말의 내용과 발화자의 감정을 인식하고 의미를 이해하여, 상황에 따라 자연스러운 대화를 주고받기 위해 필요한 음성 언어 처리 기술은 인간의 자연어 발화를 컴퓨터가 자동으로 이해하고 처리하는 알고리즘을 연구하는 분야로, 대화형 개인 비서 에이전트, 인공 지능(AI) 스피커, 자동 통번역, 음성 대화 질의응답(QA) 시스템 등 다양한 응용 서비스 사례를 들 수 있다.

이미 모바일 메신저 서비스를 통해 많은 양의 소통이 이루어지고 있고, 최근에는 채팅봇이 콜센터 안내, 쇼핑 도우미 및 고객 상담 등의 다양한 서비스를 하면서 대화형 상거래가 가속화하고 있다. 현재는 문자 기반으로 그 서비스가 제공되고 있으며 음성 인식 성능이 기대 수준에 이른다면 음성 인터페이스로 그 기능을 확장·연계할 수 있을 것이다. 아이티(IT) 시장 조사 기관 가트너(Gartner)에 따르면 2019년에는 스마트폰과 사용자 간의 상호 작용 중 20%가 가상 개인 비서(Virtual Personal Assistants)를 통해 이루어질 것이라고 한다. 또한 오는 2020년까지 개인용 기기는 70억 대,

웨어러블 기기는 13억 대, 그리고 사물 인터넷(IoT) 기기는 57억 대로 늘어날 것으로 전망하고 있다. 이 중 최소 20억 대의 기기 및 사물 인터넷 장비가 누르지 않고 제어할 수 있는 제로터치 사용자 인터페이스(UI) 기반으로 작동할 것으로 전망하고 있다. 성공적인 생태계가 활성화된다면 앞으로 무한한 확장 가능성을 기대할 수 있을 것이다.

서비스의 선두격인 애플 '시리(Siri)'(2011)를 시작으로 마이크로소프트 '코타나(Cortana)'(2014), 구글 '어시스턴트(Assistant)'(2016) 등 스마트폰 기반 대화형 개인 비서와 아마존 '에코(Echo)'(2014)와 구글 '홈(Home)'(2016), 애플 '홈팟(HomPod)'(2016)을 비롯하여 에스케이티 '누구'(2016), 케이티 '기가지니'(2017), 카카오 '미니'(2017)와 네이버 '웨이브' 및 '프렌즈'(2017) 등 국내 기업들도 최근 음성 인식 기반의 인공 지능 스피커들을 출시하고 있다. 스마트폰을 비롯, 스마트 스피커, 스마트 홈 허브 기능이 있는 셋톱박스, 티브이(TV), 냉장고 등 스마트 가전으로 음성 인식 기술이 급격히 확산되고 있는 점을 고려할 때, 터치와 버튼이 아닌 음성으로 기계를 제어하고 소통하는 새로운 세상이 가져올 미래 삶의 변화를 기대해 볼 만하다.

글로벌 기업들도 로컬 서비스에 국한하지 않고 다국어 서비스 지원 및 확장을 준비하고 있는 상황에서 특히, 한국어 음성 언어 처리 기술 및 응용 서비스 개발 동향과 성능 개선을 위해 진행되고 있는 연구들을 소개하고자 한다.

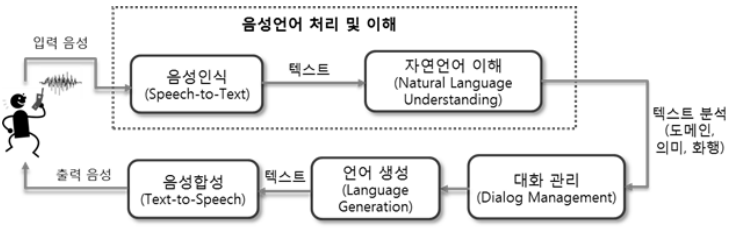
2. 음성 대화 인터페이스 및 응용 서비스

음성 대화 인터페이스는, 사람들 간에 또는 사람과 기계 간에 나타나는 인간의 자연어 음성을 컴퓨터가 듣고 이해하여 주어진 상황에 따라 적절하

게 대응하면서 대화를 나누는 음성 대화 시스템(Spoken Dialog System) 구축을 궁극적인 목표로 한다. 사용자의 음성이 입력되면 음성 인식을 통해 나온 텍스트 결과로부터 자연 언어를 이해하여 텍스트 심층 분석 결과를 구하게 되고, 언어 생성기는 적절한 응답 문장을 생성하고, 음성 합성을 하여 스피커 출력으로 음성을 재생한다. 해당 요소 기술들을 독립적으로 구성할 수도 있지만, 연속성을 갖는 선순환 구조이기 때문에 음성 언어 처리 관점에서 보면 음성 인식과 언어 이해 분야를 밀결합(tightly-coupled)하여, 음성 신호 처리 기술부터 언어 처리 영역까지 포괄하는 '음성 언어 이해'라는 기술 분야로 좀 더 사람의 발성 언어를 잘 반영하여 성능 향상의 한계를 넘어서려는 시도를 하고 있다.

한편 과거의 음성 인식 기술은 아나운서가 책을 읽듯이 발성하는 음성을 대상으로 하는 낭독체 음성 인식 기술이 주로 연구 대상이었으나, 딥러닝 및 잡음 처리 기술의 발전으로 인해 현재는 사람 간의 자연스러운 대화 음성을 대상으로 기술 고도화가 이루어지고 있다.

그림 1 음성 대화 인터페이스 시스템 구성



음성 대화 서비스를 대표하는 인공 지능 스피커와 음성 대화 로봇이 제공하는 업무 기능은 크게 1) 기기 제어 기능, 2) 웹 정보 검색 및 질의응답(QA), 3) 채팅 등으로 구분할 수 있다.

표 1 인공 지능 대화형 서비스 기능

	기기 제어 및 서비스 연동 기능	정보 검색 및 질의응답	일반 채팅
기능	음성 명령 등을 통해 특정 기기 기능에 액세스하고 대화형으로 상호 작용	연계된 포털 사이트에서 사용자가 원하는 정보 검색	지능형 가상 비서와 일상 대화
발성 예	"엄마에게 늦는다고 문자 보내 줘." "달력상에 내일 일정이 있는지 체크해." "음악 소리 줄여 줘."	"유에스(US) 달러 오늘 환율은 얼마야?" "유에스(US) 달러와 호주 달러 간 환율 알려 줘."	"날씨 참 화창하고 좋네." "자기소개해 봐." "나 좀 피곤해."

현재 제공되고 있는 서비스들은 공통적으로 음악, 스마트 홈, 날씨, 일정 관리, 알람, 뉴스 브리핑 등 준비된 도메인에 관한 음성 명령은 잘 처리하고 있다. 특히, 음악과 관련한 명령은 예를 들어 '곡명/가수명' 뒤에 '플레이'나 '틀어 줘'와 같은 명령어가 음악 재생이란 범위에 국한될 수 있도록 음악 스트리밍 서비스와 연결함으로써 도메인별로 음성 인식률과 자연어 처리 기능을 향상시킬 수 있다. 이렇게 도메인(영역)별로 서비스 범위를 확장하는 전략은 성공적인 상용화 비결이기도 하다.

표 2 국내 인공 지능 스피커 서비스 모델 추진 동향(2017년 11월 기준)

구분 \ 기업	에스케이티	케이티	카카오	네이버
디바이스	'누구', '누구 미니'	'기가지니'	'카카오 미니'	'웨이브', '프렌즈'
출시일	2016년 8월	2017년 1월	2017년 11월	2017년 9월/10월
인공 지능(AI) 플랫폼	누구	기가지니	카카오 아이(i)	클로바(Clova)
음성 호출어	아리야, 팅커벨, 레베카, 크리스털	기가지니, 지니야, 친구야, 자기야	헤이 카카오	클로바, 샐리야, 제시카, 짱구야, 피노키오

구분 \ 기업	에스케이티	케이티	카카오	네이버
주요 기능	음악 감상, 일정 관리, 날씨, 주문배달, 아이피 티브이(IP TV), 교통정보(T맵)	홈 아이오티(IoT) 제어, 홈 캠, 음악 감상, 웹 검색, 아이피 티브이(IP TV)	음성 명령으로 카카오톡 이용, 검색, 날씨, 음악 추천, 뉴스 확인	음악 추천, 음성 메모, 일정 관리, 알림/타이머, 뉴스 브리핑, 음성 검색
연동 서비스	티(T)맵, 비 티브이(B tv)	올레 티브이(tv), 지니뮤직	카카오톡, 멜론, (카카오택시, 카카오톡 주문 등 확장 예정)	네이버 뮤직, 티브이(TV) / 셋톱 박스

3. 음성 인식 요소 기술

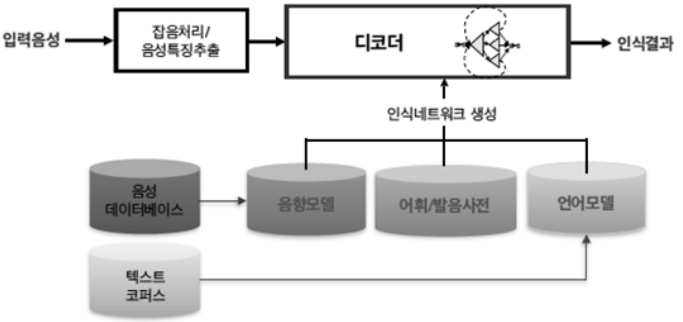
대화형 인공 지능 서비스의 관문인 음성 인식 기술에 대해 소개하고자 한다. 인공 지능 기술과 마찬가지로 음성 인식도 오랜 기간 발전을 거듭해 오면서 성능 개선의 한계로 인해 한때 암흑기를 겪기도 했다. 최근 몇 년 동안 혁신적인 성능을 보이게 된 것은 클라우드 서버 및 고성능 지피유(GPU)와 같은 하드웨어의 눈부신 발전과 빅데이터 기반 대용량 분산 처리 기술 활용이 그 배경이다.

음성 인식을 위한 많은 데이터 및 다양한 지식은 음향학적 관점 및 언어학적 관점의 두 가지 방향에서 볼 수 있다. 음향학적 관점에서는 화자, 배경 잡음, 마이크로폰 등의 다양한 환경을 나타내는 데이터를 활용할 수 있고, 언어학적 관점에서는 어휘, 문법, 문맥 등을 모델링하기 위한 많은 데이터 및 언어 정보를 정확하게 추출하여 지식 정보로 활용할 수 있다.

음성 인식 시스템은 크게 음성/언어 데이터로부터 인식 네트워크 모델을 생성하는 오프라인 학습 단계와 사용자가 발성한 음성을 인식하는 온라인 탐색 단계로 나눠 볼 수 있다. 음성 인식 엔진은 크게 음성과 언어 정보라는

중요한 사전 지식을 사용해 음성 신호로부터 문자 정보를 출력하게 되는데, 이때 개념적으로 음성 신호를 문자 기호로 해석한다는 차원에서 음성 인식 알고리즘을 디코더(decoder)라고 부르기도 한다. 디코딩 단계에서는 학습 단계 결과인 음향 모델(AM; Acoustic Model), 언어 모델(LM; Language Model)과 발음 사전(Pronunciation Lexicon)을 이용하여 입력된 특징 벡터를 모델과 비교, 스코어링을 하여 단어 열을 최종 결정 하게 된다.

그림 2 음성 인식 시스템 기본 구성도

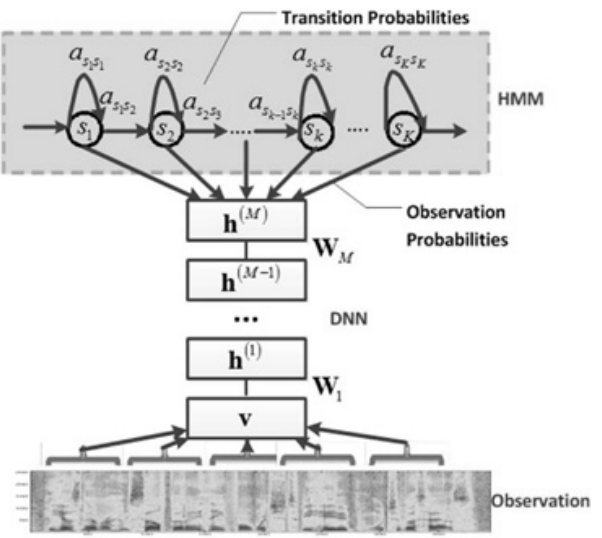


음향 모델링은 해당 언어의 음운 환경별 발음의 음향적 특성을 확률 모델로 대표 패턴을 생성하는 과정이고, 언어 모델링은 어휘 선택, 문장 단위 구문 구조 등 해당 언어의 사용성 문제에 대해 문법 체계를 통계적으로 학습하는 과정이다. 또한 발음 사전 구축을 위해서는 텍스트를 소리 나는 대로 변환하는 음소 변환(G2P; Grapheme-to-Phoneme) 구현 과정이 필요하며, 표준 발음을 대상으로 하는 발음 변환 규칙만으로는 방언이나 사용자의 발화 습관과 어투에 따른 다양한 패턴을 반영하기 어려운 경우가 있어 별도의 사전 구축이 필요하게 된다.

음성 인식 성능은 음성 데이터베이스의 크기와 품질에 비례하여 성능

향상이 이루어진다. 상용 서비스에 적용되는 음향 모델은 대부분 확률 통계 방식인 HMM(Hidden Markov Model) 기반으로 이루어졌으며, 2010년대 들어서면서 딥러닝 기반의 HMM/DNN 방식으로 단어 인식 오류율 기준으로 약 20% 정도의 성능 향상을 이끌었다(<그림 3>). DNN과 HMM을 결합하는 방법은 <그림 3>과 같이 HMM의 각 상태 확률 분포를 모델링하는 데 사용되는 GMM을 DNN으로 대체하는 것으로, 그 외의 모델 구분 단위, 단위별 학습 자료 자동 생성 및 모델 결합을 통한 문장 인식 확장 등은 HMM에서의 방식을 다수 그대로 사용하는 반면 DNN을 추정해야 하는 파라미터가 많아 학습 시간이 많이 소요된다.

그림 3 DNN-HMM 음소 인식기 구조



최근에는 시퀀스-투-시퀀스(sequence-to-sequence) 방식의 RNN (Recurrent Neural Network) 기반으로 속도와 성능 면에서 좋은 결과를

내기 시작했다. 음성 인식에서도 번역어(end-to-end) 학습 방식의 발전으로 일련의 오디오 특징을 입력으로 일련의 글자(character) 또는 단어들을 출력으로 하는 단일 함수를 학습할 수 있게 되었다. <그림 4>에서와 같이 이 함수들은 중간에 음소 단위 또는 발음 사전의 단위 변환을 거치지 않고도 긴 일련의 오디오에 대해 디코딩을 수행할 수 있다는 특징이 있다. <그림 5>에서 오른쪽 CTC (Connectionist Temporal Classification) 모델은 입력 데이터와 레이블 사이의 음성 정렬(alignment) 정보가 없어도 학습이 가능하게 된다. 성공적인 예로 바이두의 'Deep Speech2'의 경우 소음이 많은 환경에서 성능을 높이는 데 목표를 두고 개발되었으며 다양한 말투, 사투리, 시끄러운 환경에서의 음성 인식 정확도를 97%까지 높였다. 바이두는 이를 위해 9,600여 명의 7천 시간 길이 음성 샘플과 15가지 종류의 소음을 더해 10만 시간 가량의 샘플을 확보한 것으로 발표하였고, 중국어에 대해 적용하기 시작하였다(<그림 5>).

최근 마이크로소프트의 발표 논문에 따르면, 전화망 기반의 '스위치보드(SWITCHBOARD)' 말뭉치를 평가 기준으로 사람(전문 속기사)의 전사(transcription) 오류율을 5.8%에서 5.9% 정도로 보고 있는데, 그보다 더 낮은 5.1%의 단어 오류율을 발표하면서 국외 언론에서는 음성 인식 기술이 인간의 능력을 능가하는 새로운 마일스톤을 제시하고 있다고 소개하고 있다.

그림 4 종단 간(end-to-end) 학습 기반 음성 인식 기술 발전 방향

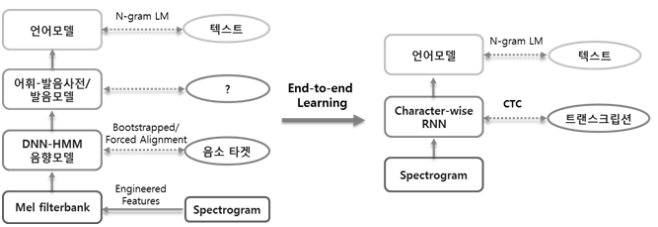
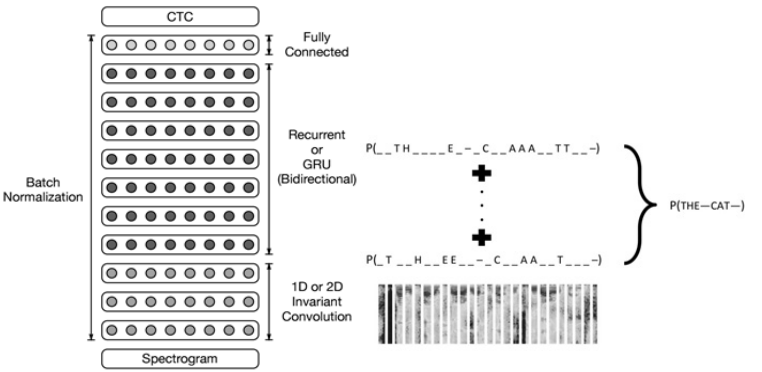


그림 5 RNN 구조와 CTC 모델 설명



마지막으로 음성 인식 결과의 정확도를 높이기 위해서는 문법 구조를 잘 반영한 언어 모델이 필요하다. 대표적인 언어 모델링 방법은 통계적인 방법에 따라 n개의 단어열에 대한 출현 빈도를 확률값으로 나타내는 엔그램(N-gram) 기법이다. 정교한 엔그램을 생성하기 위해서는 다양한 언어 자원(Corpus: 말뭉치)뿐만 아니라 실제 서비스에서 나타나는 언어 양상을 모델링할 필요가 있다. 이를 위해서는 대규모 말뭉치 기반의 언어 모델링 기술이 필수적이다. 음성 인식 서비스의 경우 서비스 어휘의 수는 기하급수적으로 증가하며, 특정 도메인으로만 대상 영역을 한정할 수 없는 특징이 있기 때문에 언어 모델의 대용량화와 지속적인 확장 기능이 필요하다. 또한 미관측 어휘(OOV: Out-of-Vocabulary)에 대한 언어 모델링의 한계를 해결할 수 있는 방법으로서 빅데이터로부터 대규모 텍스트 말뭉치를 얻을 수 있음을 전제로 하고 있다. 대표적인 예로 2006년 공개된 구글 엔그램(Web 1T 5-gram Version 1)의 경우, 1천만급(13K) 어휘를 갖는 1조 단어 규모의 5-gram 웹 기반 통계 정보이다.

여전히 기존 방식의 한계로는 학습 말뭉치에서 관측되지 못한 어휘 열에

대한 확률값의 불안정한 추정 문제와 함께 n값의 제약으로 인해 히스토리를 충분히 반영하기 어렵다는 점을 들 수 있다. 이를 해결하기 위해 도메인 특화를 위해서 다양한 언어 모델 적응(Adaption) 기법으로 모델을 견고하게 확장하는 방법과 언어 거리 간 제한을 극복하기 위해 구문 정보를 활용하거나 신경망 기반으로 언어 모델을 학습하는 등의 다양한 접근 방법들이 있다. 위에 언급된 종단 간(end-to-end) 학습 방식의 경우 RNN 구조를 이용한 글자(단어, 음절) 기반 언어 모델을 적용하는 방법과 word2vec과 같은 단어 임베딩 연구도 활발히 진행되고 있다.

4. 언어 학습을 위한 자유 발화형 한국어 음성 언어 처리

한국어의 경우, 음성 인식 디코딩 과정에서의 탐색 공간 및 계산 효율을 위해 단어가 아닌 형태소 기반의 인식 단위를 사용한다. 문자 기반의 형태소 분석과는 달리 텍스트 원형을 유지하는 음가 기반의 의사 형태소로 분할하고, 그 활용과 결합 확률을 고려하여 어휘 분할을 수행한다. 무제한급 자연어 음성 인식에서는 형태소 분석에 기반 한 어휘 선정의 문제와 함께 인식 대상 어휘의 제한으로 인한 미관측 어휘(OOV) 발생에 따른 인식 오류 발생이 성능에 영향을 끼치게 된다.

또한 방대한 양의 텍스트 데이터를 자동으로 형태소 분할을 하는 경우 텍스트 입력 오류뿐만 아니라 분석 오류 등으로 인해 의미 없는 어휘가 포함되는 경우가 자주 발생한다. 이에 따라 한국어의 경우 형태소 분석 성능, 어휘 선정 및 언어 모델의 확장을 연계하여 고려해야 전반적으로 인식 성능을 개선할 수 있다. 따라서 다양한 어휘, 문법 등을 분석함으로써 무제한급 자연어 음성 인식을 위한 언어 지식을 체계화할 수 있다. 다양한 환경에서 다양한 화자가 발성한 사용자 로그 정보들은 그 자체가 거대한

말뭉치 데이터로서 음성 언어 처리 기술의 성능을 향상시키는 주된 리소스가 된다. 특히 음성 인식의 성능 개선을 위해서는 음성 로그뿐만 아니라 방대한 분량의 텍스트 말뭉치 수집 및 한국어 음성 언어 처리 기술 확보가 필수적이다.

앞서 소개한 도메인 특화 서비스 등과 같이 답변이 가능한 발화를 제외하고 영역이 정해지지 않은 자유 발화에 대해서는 여전히 인식률이 떨어지기 때문에 "잘 알아듣지 못했습니다."라는 답변이나 사용자 발성 가이드를 제시하고 있다. 수집된 음성 데이터의 전사 데이터의 경우 그 비용과 확장의 한계 때문에 관측 가능한 충분한 말뭉치를 최대한 확보하는 것이 중요하다. '21세기 세종계획' 결과물인 세종 말뭉치와 같은 공개 말뭉치를 포함하여 웹 사이트를 통해 획득할 수 있는 웹 문서, 신문 기사, 게시판, 댓글, 누리소통망서비스(SNS) 데이터뿐만 아니라 드라마, 소설, 강연 자료 등 파일 단위의 텍스트 자료들을 그 수집 대상으로 한다.

예상된 시나리오를 넘는 좀 더 자연스러운, 사람의 음성 언어를 인식하고 이해하기 위해서는 대규모의 한국어 구어체 어휘 확보가 절실하다. 주로 모바일 메신저나 오픈 채팅 창에서는 문자로 정보를 전달하고 있지만, 구어체에 매우 가깝다. 다만 좀 더 빠르고 간략하게 하기 위해 축약어나 발음나는 형태로 적기도 하는데 이러한 다양한 대화 현상을 반영할 수 있어야 한다. 물론 사람과 사람 간의 대화와 사람 대 기계의 대화 내용은 서로 다를 수 있지만 인공 지능 페르소나를 통해 인간과의 대화에 가깝다고 가정하고자 한다.

사람들은 다른 사람들이 볼 수 있도록 전례 없이 많은 글을 쓰고 있지만, 개인 정보 이슈를 포함하여 로그 수집 및 활용 방안에는 많은 제약이 따른다. 그럼에도 존재하는 대규모의 데이터를 이용하는 것이 가장 그럴 듯한 언어 모델을 가장 잘 학습할 수 있을 것이다.

대화체 말뭉치로서 매우 유용한 것은 채팅에서 사용되는 문장들로, 채팅

의 형태에는 여러 사람이 함께 이야기하는 형태, 일대일로 이야기하는 형태, 쪽지를 주고받는 형태를 포함하여 공개 채팅, 비밀 채팅 등 다양한 대화 채널이 존재한다. 물론 채팅 입력 시스템을 포함하여 어떤 말을 할지, 어떤 스타일로 질문을 하고 정보를 교환하게 될지 선정하는 것이 매우 어렵다. 목적 기반(Task-oriented)의 시스템에서도 단순 정보 교환 및 일반 오픈 채팅이 이루어지기 때문에 수집 범위를 한정 짓기보다는 대규모의 데이터를 이용하는 방안을 선택하고 있다.

게임 도메인 확장을 고려하여 인벤(inven) 사이트의 게시판 글과 답변뿐만 아니라 엔씨소프트의 피시(PC) 게임 '리니지'와 모바일 게임 '리니지엠(M)' 채팅창에서 주고받은 대화 내용을 모니터링하면서 대화체 말뭉치를 보강하고 있다. 익명성에 기반 한 전체 공개 대화 내용만을 대상으로 하며, 비속어나 금칙어는 입력 시 '***'로 표기되기 때문에 그 원문은 알 수 없다. 전체 오픈 채팅방에서는 각 서버마다 접속한 인원 전원이 볼 수 있으며 각종 질문과 답변, 시세 문의, 아이템 판매, 생활 뉴스 이야기 등 다양한 주제가 복합적으로 일어나고 있다. 불특정 다수와의 대화가 공개적으로 오가며, 자주 만나는 캐릭터들과의 대화, 아이템 획득 시 축하 인사 메시지 등이 다수이다. 게다가 건전한 채팅 문화 정착을 위한 캠페인을 오랜 기간 시행해 왔기 때문에 비공개 채팅창에서 오가는 내용보다는 전달 내용이 명확하고 깨끗한 편이다. 그럼에도 비공식적, 비격식적 문서의 양이 절대적으로 늘어나면서 불완전한 문장 또는 문법에 어긋나는 표현들도 함께 늘어나게 된다. <표 3>에서는 문어체 문장과는 다른 채팅 문장의 특성을 살펴볼 수 있다.

표 3 채팅 문장 특성 분석

채팅 특성	예
신조어, 도메인 특화 단어	혈창, 혈원, 잊섭, 유디, 드상, 축드상, 혈톡, 프리섭
발음 변형 및 축약	드가봐야(들어가 봐야), 걍(그냥), 일루와(이리로 와), 젤루(제일로)
띄어쓰기 오류	단검이라스턴이읍음, 이거88찍으세, 템제대로나오면
겹받침 탈락 및 오타	하겟어, 훔쳣나, 햇죠, 힘듧어, 무었이, 했써
감정 표현 관련 기호	ㅋㅋㅋ, ㅜㅜ, ㅎㅎ, @@, ——;
단음소 음절 및 낱자 표기	ㅇㅇ, ㅅㄱ, ㅍㅍㅍ, ㅅㅇㅅ, 그만 ㅈ ㅏ～～～～, ㅎㅏㅇㅣ
한글-숫자 혼용	4강6셋, 축젤2, 18분에33번가자
외국어 및 한/영 키 오류	dkssudgktpdy 38, can` ban bua`, gai di tang di dao hok

일정 기간 모니터링을 통한 데이터 유효 건 중에서 출현 빈도수를 기반으로 분석해 본 결과, 다음과 같이 동일한 발음으로 인해 철자를 혼용하거나 의도적으로 발음을 예측해서 강조 문구로 사용하는 경우가 많이 발견되었다. 이런 현상은 특히 소통할 때 철자를 확인하는 것보다 속도나 효율을 우선시하기 때문이기도 하다.

- 많이[마니]: 마니(많이) 먹어라, 마니(많이) 아프냐
- 했다[핻따], 했지[핻찌]: 잘햇지(잘했지), 말햇지(말했지), 잘했따(잘했다), 망했따(망했다)
- 좋아[조아]: 조아짐(좋아짐), 조았어(좋았어), 아주 조아(좋아), 조아조아(좋아 좋아)
- 조쿠먼(좋구먼), 아랐어(알았어), 머찌다(멋지다), 마자요(맞아요)

맞춤법 및 철자 오류에 관해서는 언어가 사용되는 한 그 기준과 변화

수용에 대해 끊임없는 고민이 필요하다. 맞춤법 오류는 끊임없이 발생 가능하고, 얼마나 희한한 철자가 입력되는지 보면 경이로울 정도이다. 자연스럽게 사용자들은 고빈도로 노출되는 문장을 정답으로 생각하게 되는데, 마찬가지로 음성 인식 결과가 내놓는 출력도 인식 오류뿐만 아니라 철자 오류 없이 결과를 내 주는 것이 필요하다.

대화체 문장을 반영하기 위해 생각보다 많은 양의 입력 오류를 처리하고 필터링해야 하며, 또한 게임을 주제로 한 대화가 주를 이루기 때문에 도메인 특화 작업에 필요한 신규 어휘 등록 및 분석 정교화 작업이 필요하다. 교과서적인 문장을 기준으로 코드화된 텍스트 분석 및 처리기는 많은 사람들이 실제 사용하는 언어에 적용하기 위해서 별도의 추가 작업이 필요하게 된다.

표 4 텍스트 분석 오류 예

오류 타입	입력 예	형태소 오분석 예제	입력 띄어쓰기 보정 및 사용자 사전 반영
띄어쓰기	저도첨이라 저도첨봐요 예형님	저/MM 도첨/NNG 이/VCP 라/EC 저/NP 도/JX 첨/MAG 봐요/VV+EC 예형/NNP 님/XSN	저/MP 도/JX 첨/NNG 이/VCP 라/EC 저/NP 도/JX 첨/MAG 봐요/VV+EC 예/IC 형/NNG 님/XSN
신조어/ 축약어	출첵 푸귀	출/VV+ETM 첵/NNP 푸/IC 귀/NNG('푸른 귀걸이'의 약어)	출첵/NNG 푸귀/NNG
어체 변형	안녕하세염 안녕하세여 안냐세요	안녕/NNG 하/XSV 세염/NNG 안녕/NNG 하/XSV 세여/EP+EF 안/NNG 냐/VCP+EC 세요/NNG	안녕/NNG 하/XSV 세염/EP+EF 안녕/NNG 하/XSV 세여/EP+EF 안냐/NNG 세요/EP+EF

※ 세종 품사 태그 사용: NNG(일반 명사), NP(대명사), MM(관형사), VV(동사), VCP(긍정 지정사), IC(감탄사), EC(연결 어미), EF(종결 어미), JX(보조사), EP(선어말 어미), XSV(동사 파생 접미사), XSN(명사 파생 접미사) (단, 형태소 분석기마다 결과가 다를 수 있음.)

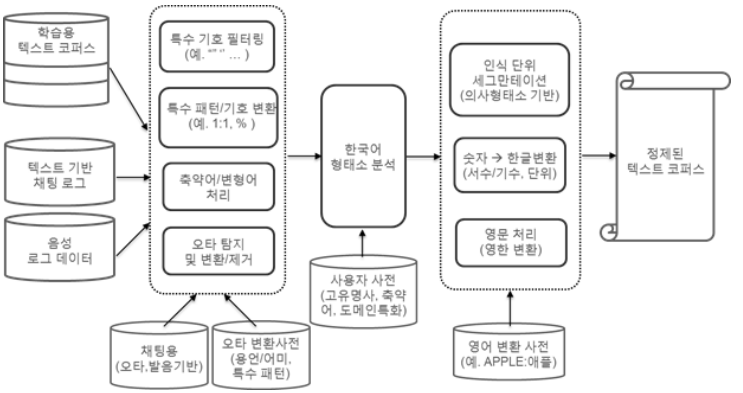
철자법에는 개인적인 감정이 포함되어 있지 않아 사용자의 감정과 상태를 구별하지 못한다. 감성 대화가 가능하게 하려면 양상(modality) 정보를 이해할 수 있어야 한다. 말투의 변화를 반영하려면 '안녕하세요'라는 인사말과 같은 경우 <표 5>와 같이 뒤에 용언의 활용 어미로 '-하세여/-하세염/-하세욤/-하세용/-하세욧/-하셈' 등과 같이 구분할 수 있어야 한다.

표 5 발음 변형 기반의 이형태 표현

표준 입력	변형된 대체 표현
뭐하니	뭐하냐, 뭐하노, 뭐하냥, 뭐하나, 뭐하농, 뭐하누, 뭐하남, 모하니
반가워요	반가워용, 반가워여, 반가워욤, 방가워요, 방가워용, 방가워여, 방가워유, 방가워염
안녕하세요	안녕하세여, 안녕하세용, 안녕하세욤, 안녕하세염, 안녕하세유 안뇽하세요, 안뇽하세여, 안뇽하세용, 안뇽하세욤, 안뇽하세염 안냐세요, 안냐세여, 안냐세용, 안냐세욤, 안냐세염

기존 음성 인식 시스템에서는 '-하세요'를 표준 어휘로 정하고, 발음 사전의 허용 범위 내에서 발음의 다양성을 반영하는 방식으로 수행하였다. 이런 경우 명시적인 표출형 어휘와 암시적인 발음 다양성을 발음 사전에 넣어 처리하는 것 사이에 혼잡도를 어떻게 수용할 것인지 판단이 필요하게 된다. 등록 기준이 정해지면 발생 빈도에 기반을 둔 통계 모델을 채택하는 것이 일반적일 것이다.

그림 6 언어 모델 학습용 말뭉치 정제를 위한 텍스트 처리 프로세스 예



5. 앞으로의 발전 방향

자유 발화형 음성 인식이 어려운 이유는 발성 중 '음', '저', '어'와 같이 예상치 못한 간투사가 수시로 사용되며, 말더듬, 어휘의 도치 현상, 동일 어휘의 반복이나 구간 재발성 등으로 인한 빈번한 비문법적인 발성에 기인하는데, 이와 같은 비정형 언어 처리는 학습을 통한 모델링으로는 여전히 한계가 있다. 기존 어휘 체계를 전문가 지식 기반으로 규칙화하거나 통계적 방식으로 처리하는 것은 확장성 측면과 문제 해결의 범위가 여전히 제한되기 때문에 새로운 방식의 언어 모델이 필연적으로 개발되어야 한다. 이러한 통계적 방식의 단점을 극복하고 비정형 자연어를 효과적으로 인식하기 위해 연산 처리 속도 및 정확도 향상을 위한 기술적 발전과 함께 현재 다양한 딥러닝 기술이 활발하게 연구되고 있다.

한국어의 경우, 인식 단위로 의사 형태소를 사용하기 때문에 후처리 모듈에서 인식 결과를 어절 단위로 재구성하는 과정이 필요하며, 일반적으로 숫자나 영문의 경우 변환해 주는 텍스트 정규화 과정 또한 필요하다. 또한

음성 인식 결과가 완벽하지 않기 때문에 오류 보정을 위한 노이즈 채널 모델과 같은 후처리 방식을 적용하여 그 정확도를 향상시킨다. 외부에서 제공하는 음성 인식 에이피아이(API)를 사용하는 경우에는 음성 인식 엔진이 블랙박스이겠지만, 선순환적으로 언어 모델을 구성하는 데 후처리 보정 기술을 적용한다면 별도의 처리 과정의 부담을 줄이면서 인식 성능의 향상도 가져갈 수 있을 것이다.

현재 제공되고 있는 스마트폰 음성 인식 애플리케이션이나 인공 지능 스피커의 경우, 단말에서는 음성 녹음 및 기본 처리를 수행하고 클라우드 서버로 전송해서 (음성 인식 및 서비스에 필요한) 작업 수행 후 결과를 단말로 재전송하는 서버-클라이언트 방식을 갖는데, 네트워크가 끊긴 상태이거나 제한된 환경의 전용 디바이스에서 수행될 수 있는 서비스 방식도 함께 고려되어야 할 것이다. 최근에는 <그림 1>에서 음성 인식과 언어 이해에 해당하는 두 단계를 하나의 엔진에 담아 음성-의미 해석(Speech-to-Meaning) 과정을 한 번에 처리하여 반응 속도와 정확도를 높이는 기술이 소개되고 있다. 기기 제어와 같은 음성 명령이나 간단한 조회 등 특화된 서비스에 매우 적합한 기술로 보이며, 복합 질문에 대해서도 한 번에 답을 내놓을 수 있을 것으로 기대할 수 있다.

참고 문헌

한국사전연구사 편집부(1994), ≪컴퓨터 정보용어대사전≫, 한국사전연구사 (언어정보처리, 음성정보처리).

권오욱, 최승권, 노윤형, 김영길, 박전규, 이윤근(2015), “자유발화형 음성대화 처리 기술동향”, Electronics and Telecommunications Trends 30-4, 26-35쪽.

박전규(2016), “인공지능 기술 개발 어디까지 왔나? 딥러닝 기반의 음성인식 기술”, ≪컴퓨터월드≫.

김상훈, 조재원(2017), “말하는 대로 통역에서 비서까지, 음성인식 기술”, ≪융합연구리뷰≫ 3-6.

과학기술정책연구원(2017), “지능형 개인비서 시장 동향과 국내 산업 영향 전망”, ≪동향과 이슈≫ 제35호.

김학수(2017), “인공지능 음성언어 비서 시스템의 자연언어처리 기술들”, ≪정보과학회지≫ 35-8.

Dahl, George E., et al(2012), “Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition.” IEEE Transactions on audio, speech, and language processing 20.1, 30-42.

Hinton, Geoffrey, et al(2012), “Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups.” IEEE Signal Processing Magazine 29.6, 82-97.

Amodei, Dario, et al(2016), “Deep speech 2: End-to-end speech recognition in english and mandarin.” International Conference on Machine Learning.

Stolcke, Andreas, and Jasha Droppo(2017), “Comparing human and machine errors in conversational speech transcription.” arXiv preprint arXiv:1708.08615.