음성언어기술 기반 대화형 질의응답 시스템 개발 동향


김 종 진
네이버 주식회사


김 준 석
네이버 주식회사


김 정 희
네이버 주식회사


이 현 아
네이버 주식회사


서 희 철
네이버 주식회사

Ⅰ. 서 론

HCI(Human-Computer Interface)에 음성입출력 기술을 이용한 사용자 인터페이스와 언어처리 기술을 이용한 사용자 의도 이해를 통합하여 보다 고차원적인 HCI 편리성을 사용자에게 제공하고자 하는 연구 및 시스템 개발은 관련 연구/개발 분야의 오랜 숙원 사업이었다. 그러나, 과거에는 데이터의 부족과 연구개발방법론의 미성숙, 컴퓨팅 파워의 부족 등 다양한 원인에 의해 이러한 HCI 기술이 부분적으로 개발/도입되거나, 개발/도입된 기술도 완성도 측면에서 사용자로부터 큰 호응을 받지 못하였으며, 이로 인한 연구개발로의 생산적인 피드백 또한 많지 않았다. 그러나 인터넷 정보검색 기술이 단어 중심의 키워드 검색에서 문장 형태의 점점 더 복잡한 사용자의 질의어를 처리할 수 있게 되고, 스마트폰과 같은 다기능 휴대/단말 장치의 사용 인구가 증가하면서, 자연스럽게 사용자는 좀더 입출력은 편리하면서도, 검색 질의는 복잡한 의도를 가진 정보검색을 요구하게 되었고, 수요공급의 법칙에 따라 기업체에서는 사용자 요구사항을 만족시킬 수 있는 기술개발과 서비스의 공급에 촛점을 맞추어 연구개발을 진행하고 있다.

본 고에서는 이러한 노력의 일환으로 네이버에서 개발되어 서비스되고 있는 “네이버 링크(LINK)” 서비스를 소개하고, 링크 서비스에 사용된 핵심 요소 기술인 자연어 이해 시스템, 대화 모델 시스템, 음성인식 시스템, 음성합성 시스템, 자동 번역 시스템에 대해여 상술하고자 한다.

Ⅱ. 네이버 링크(LINK) 서비스

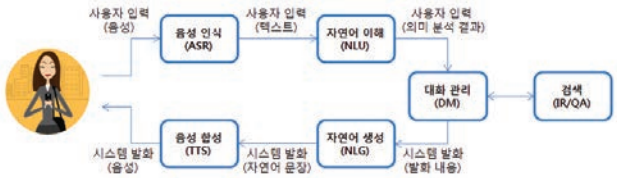
본 절에서는 2011년 11월 네이버에서 개발되어, 〈그림 1〉 과 같이 앱 형태로 서비스되고 있는 “네이버 링크(LINK)” 서비스에 대하여 기술한다.

네이버 링크 서비스는 1) 사용자 검색 질의 로그를 기반으로 대량의 학습 데이터를 생성하고, 2) 대량의 학습 데이터를 활용하기 위해서 데이터 기반 접근 방법을 사용하고, 3) 자연스러운 대화를 위한 지식 정보를 온톨로지 형태로 관리하고, 4) 중의적인 입력 처리를 위해서 하나 이상의 가능한 후보들을 고려하고, 5) 음성인식 및 합성 입출력 인터페이스 방식으로 동작한다.

네이버에는 매일 다양한 사람들이 다양한 형태로 검색한다. 사용자 검색 패턴에는 특정 대상에 대한 질의들이 연속으로 들어오는 경우가 있다. 예를 들어, “서울 날씨”, “내일 서울 날씨”, “내일 강원도 날씨”와 같은 질의들이 차례대로 들어오는 경우이다. 이런 경우, 사용자는 ‘날씨’라는 주제에 대해서 여러 가지 표현으로 자신의 검색 목적을 표현한다. 이처럼 하나의 주제에 대한 연속적인 질의들은 하나의 대화 흐름으로 볼 수 있으며, 이를 기반으로 대화 시스템을 위한 학습 데이터를 자동으로 생성한다. 더불어 서비스 오픈 후에 들어오는 사용자 로그도 학습 데이터로 함께 사용한다.

네이버 자연어 이해 시스템은 1) 다양한 유형의 질의 처리, 2) 자연어 문장의 중의성 해결, 3) 검색 질의의 대상 도메인 고려, 4) 기존 사용자 검색 질의를 활용하고, 5) 계층적인 형태로 의미를 표현한다.

〈그림 1〉 네이버 링크 서비스 앱



〈그림 2〉 네이버 링크 서비스 요소기술

네이버 링크 서비스는 〈그림 2〉와 같은 요소기술로 구성된다.

사용자는 질의하고자 하는 내용을 구어체 형태로 발화하고, 시스템은 이를 음성인식 시스템을 통해 사용자 발화를 발화문 형태의 텍스트로 변환하고, 이를 자연어 이해 시스템을 통해 사용자 입력에 대한 의미분석을 수행한다. 상세 의미 분석된 질의어는 대화관리 시스템에 의해 검색 서버와 연동되어 사용자 질의에 대한 정보를 검색하고, 정보를 사용자에게 텍스트와 이미지를 화면에 어떻게 보여줄 것인지, 어떤 내용을 음성출력으로 사용자에게 들려줄 것인지 결정한다. 특히, 정보를 음성출력하는 경우에는 검색된 정보를 자연어 생성 모듈을 거쳐 문장 형태로 변환하고, 이를 음성합성 모듈을 이용하여 음성출력한다.

Ⅲ. 네이버 링크 서비스의 요소기술

1. 자연어 이해 시스템

네이버 자연어 이해 시스템은 사용자 질의를 분석하는 시스템으로, 정보 검색 시스템, 질의 응답 시스템, 대화 시스템 등에서 사용할 수 있는 형태로 되어 있다. 이를 위해서 네이버 자연어 이해 시스템은 1) 다양한 유형의 질의 처리, 2) 자연어 문장의 중의성 해결, 3) 검색 질의의 대상 도메인 고려, 4) 기존 사용자 검색 질의를 활용하고, 5) 계층적인 형태로 의미를 표현한다.

1.1 다양한 유형의 질의 처리

네이버 자연어 이해 시스템은 네이버 검색 시스템에 입력되는 다양한 형태의 질의를 분석한다. 검색 질의 유형으로는 특정 객체의 속성을 묻는 질의(“변호인의 감독은 누구인가요?”), ‘예/아니오’를 묻는 질의(“박쥐는 포유류인가요?”), 어떤 조건을 만족하는 객체 리스트를 찾는 질의 (“2013년에 개봉한 로맨스 영화는 무엇인가요?”) 등의 유형이 있다. 네이버 자연어 이해 시스템은 입력 질의 유형을 파악하고, 각 질의에서 사용자가 묻고자 하는 의도까지 분석한다.

1.2 중의성 해결

언어처리 분야의 모든 응용이 그러하듯, 자연어 이해에서도 중의성 해결이 큰 과제 중 하나이다. 예를 들어, “요리 영화”라는 질의는 두 가지 의미가 가능하다. “요리”라는 사람이 출연한 영화를 묻는 의미와, “요리”를 주제로 하는 영화를 묻는 의미이다. 전자는 “요리”라는 객체의 ‘출연 영화’ 속성값을 묻는 유형이고, 후자는 ‘주제’가 ‘요리’인 영화 리스트를 묻는 유형이다.

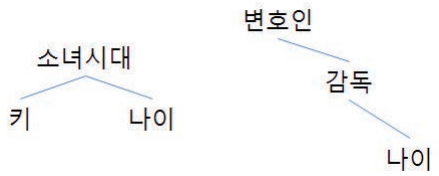
네이버 자연어 이해 시스템은 주변 문맥을 고려해서 중의성 해결을 시도한다. 주변 문맥만으로는 중의성 해결이 안 되는 경우, 가능한 모든 분석 결과를 생성한다.

중의적인 문장에 대해서는 가능한 모든 의미 분석결과를 제공함으로써, 자연어 이해 시스템을 사용하는 정보 검색 시스템, 질의 응답 시스템, 대화 시스템에서 가능한 의미 정보를 모두 사용할 수 있도록 한다.

대화 시스템에서는 하나 이상의 가능한 의미를 기반으로 중의성 해결을 위한 대화를 진행할 수도 있다.

1.3 도메인 기반 자연어 이해

네이버 자연어 이해 시스템은 자연어 이해를 위해서 검색 질의의 도메인을 함께 고려한다. 사용자 질의는 검색 도메인에 따라서 구분할 수 있으며, 자연어 이해 시스템은 질의의 검색 도메인을 파악함으로써 질의 이해의 정확도를 향상시킬 수 있다. 이는 도메인별로 질의를 구성하는 객체, 속성 등의 정보가 다르며, 도메인에 의해서 중의성이 해결될 수 있기 때문이다. 예를 들어, “변호인 관객 수는 몇 명인가요?”와 같은 질의는 ‘영화’ 도메인 질의임을 파악함으로써, “변호인”이 영화 제목임을 명확하게 결정할 수 있다.



(좌: “소녀시대 키 나이”, 우: “변호인 감독 나이”)

〈그림 3〉 계층적 의미표

1.4 사용자 질의 기반 자연어 이해

입력 자체가 아무 제약이나 규약이 없는 자연어 문장이므로 같은 질의 의도이지만 실제 질의의 형태는 무척 다양할 수 있다. 자연어 이해 시스템은 최대한 다양한 형태의 질의를 이해할 수 있는 이해력을 갖춰야 한다. 이를 위해서는 여러 가지 방법론이 존재할 수 있는데, 본 네이버 자연어 이해 시스템은 대량의 검색 질의로부터 유사한 질의들을 추출하여 이를 학습 데이터로 사용하고, 이로부터 유용한 패턴과 문법을 추출하여 자연어 이해를 시도한다.

1.5 계층적인 형태의 의미 표현

자연어 이해의 결과물은 입력 문장의 구조적/의미적 분석 결과이며, 이는 의미 표현 형태로 기술된다. 네이버 자연어 이해 시스템은 의미 표현을 위해서 계층적인 구조를 이용한다. 예를 들어, “소녀시대 키 나이”와 “변호인 감독 나이”는 〈그림 3〉과 같은 형태로 각각 표현된다. 두 가지 모두 객체 + 속성1 + 속성2 의 형태를 취하고 있지만, 의미 표현은 다르다. 이유는 “소녀시대 키 나이”에서 속성2인 “나이”는 “소녀시대”의 속성을 나타내지만, “변호인 감독 나이”에서 속성 2인 “나이”는 “변호인”이 아닌 “변호인 감독”의 속성을 나타내기 때문이다.

2. 대화 모델 시스템

네이버 대화 시스템은 사용자와 대화하는 시스템으로, 대화를 통해서 사용자의 목적을 달성한다. 1960년대 엘리자(ELIZA)부터 최근에는 Apple Siri, Google Now, 삼성전자 S-voice, LG 전자 Q-voice, 네이버 Link 등 다양한 대화형 시스템이 서비스되고 있다.

2.1 대화 모델 기술 연구개발 동향

대화 시스템은 주로 음성 입력을 받아서 음성 출력을 생성한다. 이를 위해서 음성 인식과 음성 합성을 사용한다. 음성 인식으로 생성된 자연어 문장은 자연어 이해 시스템을 통해서 의미 표현으로 변경되며, 의미 표현된 정보를 대화 관리에서 처리하고 시스템 발화 내용을 생성한다. 자연어 생성 시스템은 발화 내용을 기반으로 자연어 문장을 생성한다.

대화 시스템을 특징짓는 가장 중요한 요소는 "대화 관리(Dialogue Management: DM)" 모듈이다. 대화 관리 모듈은 사용자와의 대화 내용을 관리하고, 대화 내용을 기반으로 시스템의 다음 발화 내용을 생성한다.

대화 시스템을 특징짓는 가장 중요한 요소는 "대화관리(Dialogue Management: DM)"모듈이다. 대화관리 모듈은 사용자와의 대화 내용을 관리하고, 대화 내용을 기반으로 시스템의 다음 발화 내용을 생성한다.

대화 관리를 위한 접근 방법으로는 지식 기반 접근 방법과 데이터 기반 접근 방법이 있다. 지식 기반 접근 방법은 대화 개발자가 직접 기술한 지식을 이용하는 방법으로, Voice XML[1]과 CMU에서 개발한 RavenClaw[2] 시스템이 해당한다. VoiceXML은 대화 규칙을 XML 문법에 따라서 기술함으로써 대화 시스템을 만들 수 있도록 한다. CMU의 RavenClaw는 대화 단계를 의미에 따라 구분하고, 각 단계 간의 관계를 계층적으로 표현함으로써 수동 지식 구축을 쉽게 한다.

데이터 기반 접근 방법은 대량의 대화 데이터에서 통계 정보를 추출해서 활용하는 방법이다. 데이터 기반 접근 방법에서는 대화 데이터와 기계 학습 기법을 함께 사용한다. 다양한 기계 학습 기법 중에서 예제 기반 방법[3]과 강화 학습 방법[4]이 많이 적용되고 있다. 예제 기반 방법은 소규모 대화 데이터로도 시스템을 구축할 수 있다는 장점이 있다. 강화 학습 방법은 대화의 최종 목적 달성에 최적화된 대화를 진행한다는 장점이 있지만, 대규모 학습 데이터가 필요하다는 단점이 있다. 이를 극복하기 위해서 가상의 대화 시스템을 만들어서 학습 데이터를 생성하는 사용자 시뮬레이션(user simulation)에 관한 방법도 연구되고 있다.

2.2 네이버 대화 모델 시스템

네이버에서는 〈그림 4〉와 같이 2개의 대화모델을 적용한 서비스를 하고 있다. 〈그림 4(a)〉는 2012년 11월에 서비스를 시작한 네이버 "링크(Link)" 서비스로, 모바일에 장착된 개인 비서 역할을 수행한다. 〈그림 4(b)〉는 2014년 1월에 서비스 시작한 "대화형 검색"이다. 대화형 검색 서비스는 선물추천 관련된 질의에 대해서 대화형으로 검색 가이드를 제공한다.

네이버 대화 관리 시스템은 학습 데이터에서 사용자의 대화 흐름을 잘 파악하고, 대량의 데이터를 적극적으로 활용하기 위해서 데이터 기반 접근 방법을 취하고 있다. 이를 위해서 지금까지의 대화 기록 (dialogue history: dh)을 기반으로 시스

〈그림 4〉 네이버 대화 시스템
(a) 네이버 링크, (b) 대화형 검색

템의 다음 행동(action: a)을 결정하기 위해서 다음 수식을 사용한다.

$$\underset{a}{argmin}\, P(a|dh)$$

대화 기록에는 사용자와 주고받은 내용뿐만 아니라, 음성 인식 결과, 사용자 입력 오류 등의 모든 대화 정보를 함께 이용한다. 시스템의 다음 행동은 다음 발화 내용이 되며, 이 내용에는 사용자의 요청에 대한 답변, 사용자 요청을 이해하지 못한 경우의 문의 사항 등이 있다.

대화 참여자 간에는 공유하는 정보가 있다. 예를 들어, "할아버지"의 성별은 남성이고, 연세가 있는 분이라는 정보를 서로 공유하고 있다. 공유하는 정보가 많을수록 대화 흐름이 원만하고, 빠른 진행이 가능하다. 대화 시스템에서도 이와 같은 정보가 필요하다. 네이버 대화 시스템에서는 2가지 온톨로지 형태로 해당 정보를 관리한다. 하나는 공용 온톨로지(common ontology)이고, 다른 하나는 도메인 온톨로지(domain ontology)이다. 공용 온톨로지는 "할아버지"와 같이 다양한 도메인에서 사용할 수 있는 정보를 가지고 있다. 반면에 도메인 온톨로지는 "경복궁"와 같이 장소/길찾기 등의 특정 도메인의 정보를 가지고 있다.

자연어 문장에는 중의적인 표현이 있을 뿐만 아니라, 음성 인식 품질로 말미암아서 음성인식에서 하나 이상의 후보를 생성할 수가 있다. 하나 이상의 후보들을 대화 관리에서 적절한 형태로 처리할 필요가 있다. 네이버 대화 관리는 이를 처리하는 방법으로 가능한 모든 후보군을 관리하고, 사용자와의 대화를 통해서 해결한다. 이를 위해서 명시적 확인(explicit confirmation)과 암묵적 확인(implicit confirmation)을 함께 사용하고 있다[5]. 네이버 대화 관리에서는 기본적으로 명시적 확인은 지금까지 주고받은 대화 내용과 검색 데이터베이스에 해당 정보가 없는 경우에 사용하고, 암묵적 확인은 하나 이상의 후보가 대화 내용 혹은 검색 데이터베이스에 있으면 사용한다.

3. 음성인식 시스템

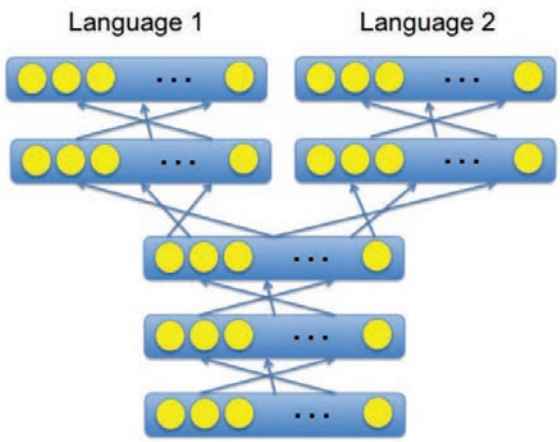
음성인식은 모바일 기기의 사용이 보편화되면서 사용자 입력 수단의 하나로 크게 다시 각광을 받고 있다. 2000년대 중반까지만 해도 인식률의 한계로 인한 서비스 품질의 사용자 만족도가 높지 않아 침체기를 맞이하였다. 그러나, 최근 스마트폰과 같은 모바일 기기 사용자가 늘어나고, 정보검색 및 모바일 사용빈도가 높아지면서, 사용자의 자연스러운 인터페이스에 대한 요구사항이 증대되고, 그 일환으로 음성인식 기술이 다시 크게 부각되고 있다.

또한 이러한 요구에 의해 음성인식 기술을 활용한 서비스가 늘어나고, 서비스를 통해 자연스럽게 사용자 패턴이 수집되고, 시스템 개발에 재활용되면서 점점 더 성능이 향상되는 선순환 구조를 이루고 있다.

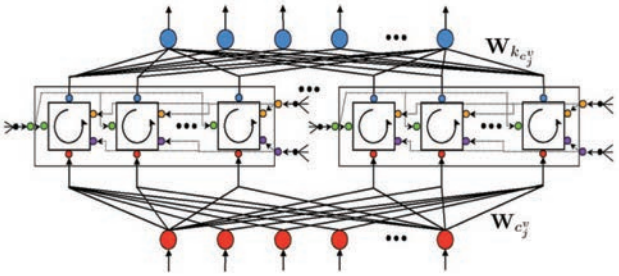
3.1 음성인식 연구 개발 동향

음성인식은 최근 deep neural networks를 이용한 deep learning 기술의 발전으로 인식 성능면에서 많은 향상을 가져왔다. 최근 해외 대부분의 선도 업체들은 neural networks를 이용한 음향 모델을 구축하여 사용하고 있으며[6-7], 차츰 언어 모델도 neural networks로 대체되는 추세이다[8].

Deep learning을 적용한 음향 모델 생성을 위해서는 대규모의 데이터 베이스가 필수적이며, 음성인식 성능에서 서버형 서비스로 대규모 음성 데이터 베이스를 구



〈그림 5〉 DNN을 이용한 다국어 음향모델 훈련



〈그림 6〉 Long Short-Term Memory RNN 구조

축한 업체와 그렇지 못한 업체간의 성능 격차를 벌리는 요인이 되고 있다.

Neural networks를 이용한 음향 모델의 또 다른 장점은 〈그림 5〉와 같은 형태로 음향 모델을 일정 부분 공유하여, 서로 다른 언어의 음성 데이터 베이스를 가지고 다른 언어의 음향 모델을 구축하는데 사용할 수 있다는 점이다[9]. 이는 다국어 음성인식을 개발할 때 큰 장점으로 작용한다.

네이버와 같은 국내 서비스에 기반을 둔 회사는 한국어, 구글과 같은 경우는 영어 음성 데이터 베이스를 많이 구축할 수 있으나, 다른 언어에 대해서는 상대적으로 작은 규모의 데이터 베이스를 구축할 수 밖에 없다. 하지만 deep neural networks을 사용할 경우 다른 언어의 데이터 베이스를 사용하게 되어 작은 규모의 음성 데이터를 보충할 수 있게 된다.

언어 모델의 경우에도 최근 추세는 feed forward neural networks보다는 recurrent neural networks를 쓰는 추세인데, 〈그림 6〉과 같은 long short-term memory recurrent neural networks는 과거 데이터를 현재의 출력을 위해 사용하는 구조로서, 이를 사용할 경우 지금껏 사용하고 있는 3-gram, 4-gram을 넘어 이론적으로 무한대의 이전 단어들을 참고하게 된다[8,10]. 최근 ICASSP에 발표된 [8]에 따르면 RNN, FFNN, 기존 Back-off 순으로 언어모델 성능이 우수하다고 보고되고 있다.

RNN방식은 최근 음향모델 성능에서도 기존 DNN을 넘어서는 성능을 보이고 있어 RNN을 음성인식에 접목시키는 추가 연구가 많이 이루어지고 있다[11].

Neural networks 기반의 성능 향상 기법은 필수적으로 대규모 데이터 베이스를 필요로 한다. 특히 neural networks를 사용하여 음향 모델의 성능을 향상시키기 위해서는 대규모 데이터 베이스의 보유가 필수적이고, 서비스를 통해 이를 지속적으로 축적하여야 성능향상이 가능하다. 이 때문에 대부분의 서비스 업체들이 데이터 확보에 치열한 노력을 기울이고 있으며, Open API 또한 그 전략의 일환으로 볼 수 있다. 회사의 직접적인 서비스와 관계없는 오로지 데이터 수집만을 위한 서비스를 하기도 한다.

3.2 네이버 음성인식 시스템

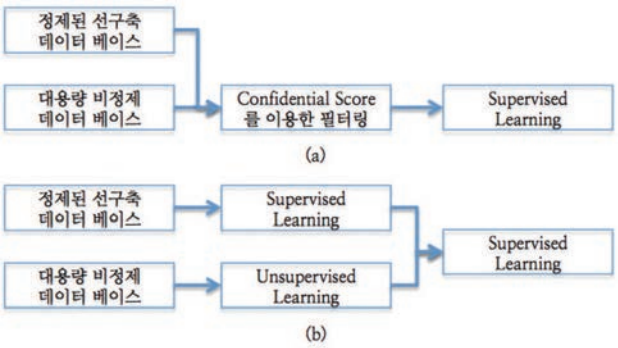
네이버는 2010년 11월 네이버 앱에 음성 검색 기능을 탑재한 이후 꾸준히 새로운 서비스를 출시하며, 키워드 음성인식 형태인 네이버 음성검색과 지식쇼핑, 문장형 음성인식 형태인 링크와 한일통역 그리고 글로벌 회화앱이 출시되었으며, 2013년 7월부터 deep learning 기술을 적용한 음성검색 서비스를 제공하고 있다.

키워드 음성인식 형태인 네이버 음성검색과 지식쇼핑, 문장형 음성인식 형태인 링크와 한일통역 그리고 글로벌 회화앱이 출시되었으며, 2013년 7월부터 deep learning 기술을 적용한 음성검색 서비스를 제공하고 있다.

하지만, 아직 여전히 네이버 및 여러 업체들의 음성인식 서비스들이 사용자들이 만족할 만큼의 성능을 보이지 못하고 있는 것이 사실이며, 이를 개선하기 위한 노력이 꾸준히 진행되고 있다. 이러한 노력들은 알고리즘 개선과 업체들이 보유하고 있는 대규모의 데이터 베이스를 이용하는 두 방향으로 이루어진다.

네이버에서도 음성인식 엔진의 성능 향상을 위해 서비스를 통해 수집되고 있는 음성 데이터 및 네이버 검색 서비스를 통해 들어오는 검색어 데이터를 활용하여 한국어 음성인식 서비스의 경우 높은 정확도를 보이고 있다.

하지만, 사용자 음성 데이터 베이스의 경우 정답이라



〈그림 7〉 대규모 음성 데이터 베이스의 활용

고 확신할 수 없으므로 supervised learning기법으로는 사용할 수 없다. 따라서, 유입된 데이터에 대한 가공이 필요하다. 유입된 데이터를 일일이 사람이 들어보고 정답을 기록하는 방법이 가장 정확하겠지만, 비용이 많이 들고 개인정보 공개와 같은 문제가 있을 수 있다. 따라서 대부분의 업체에서는 〈그림 7〉과 같은 일종의 semi-supervised learning 기법을 이용하여 이를 해결하고 있다.

4. 음성 합성 기술

음성합성 기술은 전통적으로 자동응답 시스템과 네비게이션 시스템, 뉴스읽기와 같은 단문 또는 단방향 형태의 서비스 형태로 주로 사용되었으나, 최근에는 모바일 환경에서 음성입출력 인터페이스를 가진 대화형 정보검색 기술의 사용이 늘어나면서 구어체 및 대화체 중심의 양방향 커뮤니케이션 형태로 그 활용범위가 확대되고 있다. 또한 뉴스 읽기 서비스, 알람앱 서비스 등과 연계되어, 실시간 요약정보 및 안내 정보를 읽어주는 형태로 그 활용 범위가 증대되고 있다.

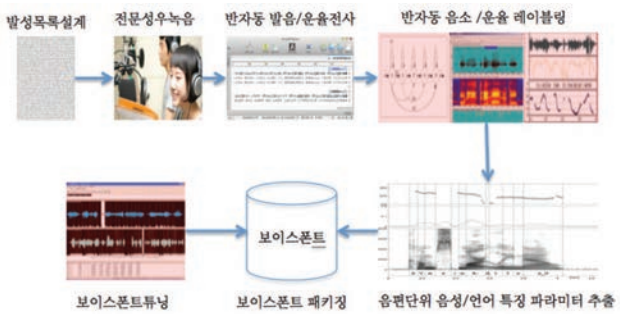
4.1 음성합성 기술 연구 개발 동향

음성합성 기술은 크게 보이스폰트(VoiceFont)를 제작하는 기술과 구축된 보이스폰트를 이용하여 임의의 입력 문장에 대한 합성음을 생성하는 실시간 음성합성 엔진 기술로 나눠서 고찰할 수 있다.

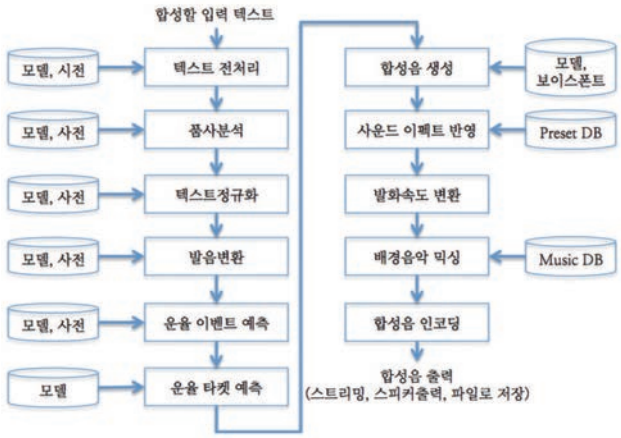
보이스폰트 제작 기술은 〈그림 8〉과 같이 음성 합성 엔진에서 사용할 보이스폰트를 제작하는 기술로 발성목록 설계, 전문성우 녹음, 반자동 발음/운율전사, 반자동 음소/운율 레이블링, 음편단위 음성/언어 특징 파라미터 추출, 보이스폰트 패키징, 보이스 튜닝 기술로 구성된다.

발성목록 설계 기술은 종전에는 음성합성 기술이 주로 단문, 뉴스읽기와 같은 서비스에 활용되고 있어, 설계의 촛점이 주로 평서문과 다양한 음운 환경을 고려한 음소 분포 커버리지를 최대화하는 관점으로 설계되었으나[12], 최근에는 대화형 인터페이스를 지원하기 위한 대화체, 감정표현 음성합성의 필요성이 대두되면서 기존 음운환경 커버리지 뿐만 아니라, 다양한 감정표현, 다양한 대화체 운율 표현이 최적으로 포함된 발성목록을 설계하는 것이 주된 연구방향이며[13], 설계시 고려할 요소가 늘어남에 따라 자연스럽게 발성목록의 양이 증가되는데 이는 개발 비용과 밀접한 관계가 있어, 다양한 음성언어정보 특징을 최대한 포함하면서 최소의 문장수를 설계하는 것이 기술적 과제이다.

전문성우 녹음이란 발성목록을 녹음할 성우를 선정하고, 선정된 성우가 설계된 발성목록을 녹음하는 과정으로 전문성우 녹음 데이터를 보이스 소스라고 부른다. 전문성우 녹음 과정은 알고리듬 측면보다는 장기간 녹음에 따른 품질 유지 등과 같은 엔지니어링 이슈가 많으며 보이스 소스가 무엇이냐에 따라 합성엔진의 출력 음색이 결정된다. 이러한 보이스 소스는 서비스의 유형, 서비스가 겨냥하는 사용자 층에 따라 그 선호도가 매우 다르며, 엔진 개발 측면에서는 하나의 보이스 소스로 모든 서비스와 사용자를 만족시킬 수 없으므로, 서비스와 서



〈그림 8〉 음성합성시스템의 보이스폰트 개발 과정



〈그림 9〉 실시간 음성합성 기능의 구성요소

비스 사용자의 선호도에 최대한 부합되는 보이스 소스를 선별하고, 이를 이용하여 보이스폰트를 빠르게 개발하는 것이 기술적 과제이며, 이는 보이스폰트 제작의 전과정을 최대한 자동화 해야 하며, 자동화의 품질이 매우 높아야 함을 의미하여, 이것을 실현하는 것이 기술적 과제이다. 이를 위해 녹음된 데이터에 대한 정교한 발음별환 및 발음전사, 다양한 운율정보 태깅, 품사 및 구문정보와 같은 언어정보 태깅이 필요하며, 각각의 정보 태깅 과정의 오류가 적을수록 일반적으로 보이스폰트를 빠르게 개발할 수 있으므로, 각각의 기술요소의 성능을 최대화 하는 것이 기술적 과제이다.

또한 서비스가 다국어를 지원해야 하는 경우에는 한국어 뿐 아니라, 외국어를 지원해야 하므로, 각각의 기술요소를 외국어에 대해서도 빠르게 개발하고 성능개선하는 것이 매우 중요하다.

실시간 음성합성엔진 기술이란 〈그림 9〉와 같이 합성할 입력 텍스트가 주어지면, 이에 대해 언어분석을 수행하고, 운율정보를 생성하고, 보이스폰트를 이용해 합성음을 생성하고 출력하는 실시간 기능에 포함된 전 기술을 통칭하며, 실시간 엔진 개발에서 개발된 텍스트 전처리 기술, 텍스트 정규화 기술, 발음변환 기술, 품사분석 기술 등 많은 언어처리, 운율 처리 모듈을 보이스폰트 제작 과정에서도 그대로 사용될 수 있다.

음성합성 엔진 개발에 사용되는 기술요소는 크게 언어종속 언어처리 기술, 언어종속 운율처리 기술, 언어독립 음성합성 기술로 나눠볼 수 있으며, 언어처리 모듈은 다시 텍스트 처리 기술과 전통적인 품사분석 및 구문분석과 같은 언어분석로 나눠볼 수 있다.

텍스트 처리 기술은 매우 언어종속적이고 문화종속적이며, 사용자의 체감과 매우 밀접하게 관련되어 있고, 예외처리가 많은 기술요소이다. 예를 들어 "2NE1"이란 그룹 이름을 "이엔이원" 이라고 읽으면 안되고 "투애니원" 으로 읽어야 한다는 점은 원어민이 아니면 도출하기 어려운 지식으로, 이러한 처리가 매우 잘되어야 한다. 기술적으로는 이러한 지식을 규칙기반으로 정의하기도 어려움이 많고, 무조건 사전 기반으로 처리하기에도 문제가 많으며, data-driven 방식의 통계기반 확률모델[14]만으로도 처리하기 어려우며, 위 3가지 방법을 잘 혼합한 hybrid 접근법을 사용해야 한다.

그 외 언어처리 및 운율처리 기술요소는 주로 Data-Driven 방식으로 개발되며, 모델 훈련을 위한 데이터의 량, 데이터의 특징 파라미터, 감독 학습을 위한 데이터의 태깅, 통계적 활률 모델의 선택 등 전통적으로 기계학습 알고리듬을 이용한 문제해결에서 대두되는 모든 고려사항을 다 포함하고 있다. 특히, 품사분석, 구문분석, 운율 이벤트 예측과 같은 경우에는 주로 태깅 데이터를 기반으로한 통계적 확률모델에 기반한 감독학습을 이용하는 것이 주 연구 방향이며, 태깅 데이터의 개발에 있어 원어민 또는 전문가의 지식이 매우 큰 역할을 한다는 점이 특징적이다. 최근에는 태깅 데이터의 문제점을 해결하기 위해 소량의 태깅된 데이터와 대량의 태깅되지 않는 데이터를 활용하는 semi-supervised 훈련 기법을 도입하여 태깅 데이터 부족 문제를 해결하려는 시도가 있다[15].

합성음 생성 방법과 관련해서는 전통적인 최적 합성단위 탐색 및 음편의 접합을 활용하는 방법과 보코더 기술과 음성인식에서 사용되는 음향모델링 기법을 차용하여 parametric 확률 모델을 이용한 음성합성 기술[16]의 연구가 아주 활발하게 이루어지고 있다. 전통적인 파형 접합 방식은 개발 비용은 크지만 고품질의 합성음을 얻을 수 있어 서버기반, 네트워크 기반의 음성합성 서비스 분야에서 주로 활용되고 있고, parametric 확

률 모델을 사용하는 방법은 저비용, 고속, 저용량 보이스폰트가 요구되는 서비스 분야에서 주로 활용되고 있으며, 대표적으로 스마트폰에 내장되어 통신 음영지역에서도 음성안내를 할 수 있도록 하는 서비스들이 주로 사용한다.

4.2 네이버 음성합성 시스템

네이버 음성합성 시스템[17]은 한/영/일 3개언어를 지원하는 다국어 음성합성 엔진이며, 파형접합 방식의 대용량 서버형과 parametric 확률 모델을 이용한 소용량 단말 내장형 기술로 구성되어 있다.

텍스트 처리 기술은 모두 원어민 개발자가 참여하고 개발하고 있어, 해당 언어 국가의 문화적 사회적 표현의 패턴을 최대로 반영하여 사용자의 체감 성능을 최대화 하고 있으며, rule-based 방법과 data-driven 방식을 혼용하고 있다. 이는 해당 언어 개발 초기에는 해당 언어에 대한 태깅된 데이터가 없어 data-driven 방식으로 모델 훈련이 불가능하기 때문에 먼저 rule-based 방식과 원어민 개발자가 참여하여 반자동으로 태깅된 데이터를 대용량으로 구축하고 이를 활용하여 data-driven 통계적 확률 모델을 훈련시키는 방식을 채택하고 있다.

네이버에서는 2011년부터 기계번역기 관련 연구를 시작하였고, 한일번역기를 시작으로 하여 매년 번역 대상 언어들을 확장해 나가고 있다.

운율 모델링의 경우에는 자동 운율 태깅 및 레이블링 기술을 중심으로 고품질의 대용량 태깅된 데이터 작성 기술을 연구개발하고 있으며, 이를 활용한 통계적 확률 모델기반 감독학습으로 모델을 개발하여 실시간 합성엔진에서 사용한다.

개발된 다국어 합성 시스템은 네이버 글로벌 회화앱, 네이버 링크 서비스, 네이버 사전 예문읽기 서비스 등에 사용되고 있다.

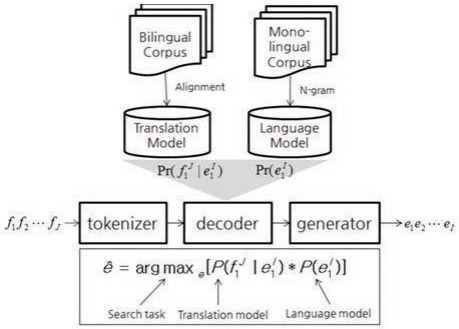
5. 기계 번역 기술

기계번역 기술은 통계적 확률 모델을 이용하여 언어간 자동 번역 결과를 생성해주는 기술로 전통적으로는 번역 결과 그 자체를 활용하는 서비스와 외국어 검색을 위한 번역 결과 활용 서비스로 나눠볼 수 있다. 특히 최근 SNS 사용자의 증가에 따른 언어가 다른 사용자들 간의 커뮤니케이션 지원을 위한 단문 메시지 번역, 또는 음성인터페이스 기술과 결합하여 실시간 자동통역 형태의 서비스로 발전하고 있다.

5.1 기계 번역 기술 연구개발 동향

기계번역의 많은 방법론 중에서 오늘날 가장 많이 사용하는 방식은 통계기반 기계번역(SMT; Statistical Machine Translation) 방식이다. SMT는 〈그림 10〉과 같이 대용량 병렬코퍼스(bilingual corpus)로 부터 자동 학습된 번역모델(TM)과 번역하고자 하는 target 언어로 학습된 언어모델LM의 확률 값을 이용하여 번역한다. 입력문장이 들어오면 우선 tokenizer에서 문장을 쪼갠 후, 번역 후보들을 찾는다. decoder에서는 수 많은 번역 후보들의 network으로부터 최적의 path를 찾아주는 search task를 수행한다. 마지막으로 generator에서 자연스러운 target 언어를 생성해주면 번역이 완성된다.

오늘날의 SMT의 모태가 된 연구는 1991년 IBM의 Watson연구소에서 발표한 논문이다[18]. 기존의 word 단위의 번역이 아니라, phrase (list of words)단위로 번역을 하는 PB SMT(Phrase Based SMT)가 2003년에 나왔고, 오늘날까지 가장 많이 사용되는 번역방법이 되었다[19]. phrase내의 변수들을 통해, 좀 더 다양



〈그림 10〉 SMT 기본구조

한 문장의 구조를 표현할 수 있게 된 것은 David Chiang의 HPB SMT(Hierarchical Phrase Based SMT) 논문부터이다. 한국어와 영어와 같이 어순이 서로 다른 언어들간의 번역의 품질향상에 큰 기여를 한 모델이다[20,21].

HPB SMT방식의 번역기는 속도가 느리다는 치명적인 단점을 가지고 있다. 따라서, pre-reordering 기법을 이용하여 번역하고자 하는 source 언어를 먼저 target 언어와 유사하게 만든 이후에, PB SMT방식으로 번역을 하려는 많은 시도들이 있었다[22,23].

최근의 SMT 분야에서는 DNN(deep neural network)를 이용한 많은 연구들이 진행되고 있다. 구글의 경우 word2vec 모듈을 활용하여 만든 정보를 SMT에 접목하려는 시도를 진행하고 있다[24]. 또한, [25]에서는 DNN을 번역모델(TM) 학습에 필요한 병렬코퍼스의 word alignment에 활용하기도 했다.

5.2 네이버 기계번역 시스템

네이버에서는 2011년부터 기계번역기 관련 연구를 시작하였고, 한일번역기를 시작으로 하여 매년 번역 대상 언어들을 확장해 나가고 있다.

SMT 기반 번역기의 개발과정은, 먼저 번역기 학습에 사용할 훈련 코퍼스를 구축하는 것이다. 단일언어 코퍼스는 인터넷 크롤링 과정을 통해서 구축하며, 병렬코퍼스는 인터넷 크롤링을 통한 수집 및 가공, 번역 업체에 의뢰하여 수작업을 통한 제작 등 다양한 방법으로 구축된다. 네이버의 경우 지식iN 서비스의 어학, 외국어 부분에 사용자들이 올린 번역 요청하기와 그에 대한 답변 데이터를 가공해서 병렬코퍼스를 확장하는데 이용하였다. 번역을 위한 코퍼스 구축이 완료되면, 번역모델(TM)과 언어모델(LM)을 학습하고, 여러 가지 SMT 실험을 진행하게 된다.

다양한 번역 방식들 중에서, PB SMT방식이 적합한지, HPB SMT방식이 적합한지, 혹은 pre-reordering 후에 PB SMT방식이 좋은지는 번역하려는 언어쌍에 따라서 달라지고, 결국은 많은 실험을 통해 가장 좋은 품질을 가지는 번역방식이 선택된다. SMT를 위한 여러 가지 parameter 결정을 위해서는 MERT[27]를 사용했고, 평가 지표는 BLEU[26]을 사용했다. 실험을 통해서 번역 방식과 parameter가 결정된 이후에는 경쟁사 번역기와 비교평가를 진행했다. 이때는 사람이 직접 번역 결과를 평가하는 정성평가도 추가적으로 수행했고, 평가 결과를 분석해서, 오류의 원인을 찾고, 해결방법을 추가해서, 번역 품질을 높이는 작업을 수행하였다.

번역기 서비스화 이후에는 지속적으로 사용자로부터의 피드백을 받아서, 번역 오류를 수정해 나가는 작업을 수행한다. 실제로 번역품질은 바로 유지보수 과정에서 큰 향상이 있음을 서비스 이후에 알게 되었다. 학습용 병렬데이터에 오류가 있는 경우에는 데이터를 수정하고, 규칙으로 잘못된 번역결과가 나타나게 하지 않도록 해야하는 경우도 있고, post-processing등으로 통해서 번역결과를 수정해 주기도 하고, 때로는 번역모델(TM)을 필터링 하기도 하면서 품질을 개선한다. 또한, 번역로그를 통해서 미등록어를 검출해서, 대역어를 수동으로 구축해서, 학습데이터에 추가해주는 작업도 번역 서비스를 이용하는 사람들에게 큰 만족도를 가져다준다. 또한, 언어는 유기체와 같이 계속해서 변화하고, 새로운 용어들도 끊임없이 생겨난다. 따라서, 번역기에 지속적으로 학습용 병렬코퍼스를 추가해주는 일도 번역의 품질을 높이기 위한 중요한 일이다.

Ⅳ. 향후 연구 및 결론

본 논문에서는 음성언어기술 기반 대화형 질의 응답 시스템의 한 예로 네이버 “링크(LINK)” 서비스 시스템을 중심으로 그 요소기술에 대한 관련 연구 최신 동향 및 네이버에서 수행하고 있는 기술개발 내용에 대하여 기술하였다.

대화형 질의 응답 시스템은 텍스트/음성의 멀티 모달 인터페이스를 기반으로 사용자와 시스템이 자연어 형태로 질문을 주고 받으면서 사용자가 원하는 정보나 답을 최적의 단계로 제공하는 것이 목적인 시스템이다. 질의에 구조적/의미적 중의성이 있을 경우 여러 개의 분석 결과도 가능하나, 최대한 중의성을 해소하는 것이 도전

해결해야 할 중요 과제 중 하나이다. 대화 특정 규약이나 제약이 없는 환경에서 입력되는 다양한 자연어 질문에 대해 좀더 강인한 이해력을 갖춘 시스템을 구축하는 것과, 이를 위해 여러 단계의 언어처리 기술과 개체명 인식, 정보 추출, 패턴 추출, 관계 추출 등의 기반 기술의 확보가 매우 중요하다.

또한 음성을 통한 질의어 뿐만 아니라, 스마트폰 기기 등에 내장된 다양한 센서정보, GPS 정보, 로그 정보 등 이질적인 다채널 데이터를 혼합한 사용자 의도 이해 모델의 훈련 및 개발이 매우 중요하며, 스마트폰과 같은 모바일 기기의 경우 개인화된 특성이 강하므로 개인화된 모델의 학습, Long-term 적응기법의 개발 및 실시간 적응 등 매우 많은 기술적으로 해결해야 할 과제가 남아있다.

또한, 단편적인 따라하기식 기술 개발이 아닌 진정한 선도적 기술개발을 위해서는, 사용자, 더 크게는 사람에 대해 더 크고 넓게 이해하기 위한, 인문적/철학적/심리학적/인지적 관점과 같은 비공학적 관점에 대한 관심과 노력이 그 어느 때보다도 중요하게 요구된다.

참 고 문 헌

[1] Voice Extensible Markup Language (VoiceXML) Version 2.0

[2] Raux, Antoine & Maxine Eskenazi, "A Multi-Layer Architecture for Semi-Synchronous Event-Driven Dialogue Management", IEEE Automatic Speech Recognition and Understanding Workshop, 2007

[3] Cheongjae Lee, et al., "Example-based dialog modeling for practical multi-domain dialog system", Speech Communication 51, 2009

[4] Esther Levin, et al., "Using Markov Decision Process for Learning Dialogue Strategies", International Conference Acoustics, Speech and Signal Processing, 1998

[5] Daniel Jurafsky and James H. Martin, "Speech and Language processing", Prentice Hall, 2008

[6] George E. Dahl, Dong Yu, Li Deng and Alex Acero, "Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition," IEEE Trans. Audio, Speech, and Language Processing, vol. 20, no. 1, Jan. 2012.

[7] Geoffrey Hinton, Li Deng, Dong Yu, George Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara Sainath and Brain Kingsbury, "Deep Neural Networks for Acoustic Modeling in Speech Recognition," IEEE Signal Processing Magazine, Nov. 2012.

[8] M. Sundermeyer, I. Oparin, J.-L. Gauvain, B. Freiberg, R. Schluter and H. Ney, "Comparison of Feedforward and Recurrent Neural Network Language Model," Proc. of the ICASSP 2013, pp. 8430-8434.

[9] G. Heigold. V. Vanhoucke, A. Senior, P. Nguyen, M. Ranzato, M. Devin and J. Dean, "Multilingual Acoustic Models Using Distributed Deep Neural Networks," Proc. of the ICASSP 2013, pp. 8619-8623.

[10] Tomas Mikolov, Martin Karafiat, Lukas Burget, Jan Cernocky and Sanjeev Khudanpur, "Recurrent Neural Network based Language Model," Proc. of Interspeech 2010, pp. 1045-1048

[11] Alex Graves, Abel-rahman Mohamed and Geofrrey Hinton, "Speech Recognition with Deep Recurrent Neural Networks," Proc. of ICASSP 2013, pp. 6645-6649.

[12] Isogai, M., H. Mizuno and K. Mano, 2005. Recording script design for corpus-based TTS system based on coverage of various phonetic elements. IEEE International Conference on Acoustics, Speech and Signal Processing, Mar. 18-23, Philadelphia, PA., USA., pp: 301-304. DOI: 10.1109/ICASSP.2005.1415110

[13] J. F. Pitrelli, R. Bakis, E. M. Eide, R. Fernandez, W.

Hamza, and M. A. Picheny, "The IBM Expressive Text-to-Speech Synthesis System for American English," IEEE Transactions on Audio, Speech and Language Processing, vol. 14, iss. 4, pp. 1099-1108, 2006.

[14] Tim Schlippe, Chenfei Zhu, Daniel Lemcke, Tanja Schultz: Statistical machine translation based text normalization with crowdsourcing. ICASSP 2013: 8406-8410

[15] Ziping Zhao, Xirong Ma, Weidong Pei: Semi Supervised Learning for Prediction of Prosodic Phrase Boundaries in Chinese TTS Using Conditional Random Fields. ISNN (2) 2011: 477-485

[16] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, T. Kitamura, Speech parameter generation algorithms for HMM-based speech synthesis, Proc. of ICASSP, pp. 1315-1318, June 2000.

[17] 김종진, 김상진, 김선희, 김형준, 홍진표, 와타나베 리카, NAVER 다국어 음성합성 시스템, 음성통신 및 신호처리 학술대회, 2013

[18] Brown, Peter F., et al. "The mathematics of statistical machine translation: Parameter estimation." Computational linguistics 19.2 (1993): 263-311.

[19] Koehn, Philipp, Franz Josef Och, and Daniel Marcu. "Statistical phrase-based translation." Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1. Association for Computational Linguistics, 2003.

[20] Chiang, David. "A hierarchical phrase-based model for statistical machine translation." Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2005.

[21] Chiang, David. "Hierarchical phrase-based translation." computational linguistics 33.2 (2007): 201-228.

[22] Xu, Peng, et al. "Using a dependency parser to improve SMT for subject-object-verb languages." Proceedings of human language technologies: The 2009 annual conference of the North American chapter of the association for computational linguistics. Association for Computational Linguistics, 2009.

[23] Goto, Isao, Masao Utiyama, and Eiichiro Sumita. "Post-ordering by parsing for Japanese-English statistical machine translation." Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2. Association for Computational Linguistics, 2012.

[24] Mikolov, Tomas, et al. "Efficient Estimation of Word Representations in Vector Space." arXiv preprint arXiv: 1301.3781 (2013).

[25] Yang, Nan, et al. "Word Alignment Modeling with Context Dependent Deep Neural Network." 51st Annual Meeting of the Association for Computational Linguistics. 2013.

[26] Papineni, Kishore, et al. "BLEU: a method for automatic evaluation of machine translation." Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, 2002.

[27] Och, Franz Josef. "Minimum error rate training in statistical machine translation." Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1. Association for Computational Linguistics, 2003.

김 종 진

1995년 2월 원광대학교 컴퓨터공학과 학사
1997년 2월 원광대학교 컴퓨터공학과 석사
2000년 2월 원광대학교 컴퓨터공학과 박사 수료
2000년 7월~2012년 3월 한국전자통신연구원 책임연구원
2012년 3월~현재 (주)네이버 음성합성연구Lab 부장

〈관심분야〉
음성합성, 자연언어처리, 기계학습, 신호처리

김 준 석

1999년 2월 경북대학교 컴퓨터공학과 학사
2001년 2월 포항공과대학교 컴퓨터공학과 석사
2001년 1월~2007년 3월 LG전자기술원 선임연구원
2007년 4월~현재 (주)이버 SMT연구Lab 부장

〈관심 분야〉
기계번역, 음성인식, 검색모델링, 자연언어처리

김 정 희

1996년 2월 서울대학교 전기공학부 학사
1999년 2월 서울대학교 전기공학부 석사
1999년 3월~2012년 1월 LG전자 LG전자기술원
2012년 2월~현재 (주)네이버 딥러닝연구Lab/음성인식개발Lab 부장

〈관심분야〉
음성인식, 자연언어처리, 신경망이론, 기계학습

이 현 아

1994년 2월 경북대학교 전자계산학과 학사
1996년 2월 포항공과대학교 전자계산학과 석사
1995년 12월~1999년 6월 한국전자통신연구원 연구원
2000년 1월~2001년 4월 L&H Korea 연구원
2001년 5월~2004년 9월 보이스텍 연구원
2004년 10월~2006년 1월 코아보이스 연구원
2006년 2월~현재 (주)네이버 자연어처리연구실 부장

관심분야〉
NLU, 개체명 인식, 정보 추출, 관계 추출

서 희 철

1998년 2월 고려대학교 전산학과 학사
2000년 2월 고려대학교 전산학과 석사
2005년 2월 고려대학교 전산학과 박사
2005년 2월~2008년 2월 한국전자통신연구원 선임연구원
2008년 4월~현재 (주)네이버 대화연구팀 부장

〈관심분야〉
자연어처리, 대화시스템