

chapter 1

Conversational AI 기반 다국어 자동통역 기술 동향



김상훈 Ⅵ 한국전자통신연구원 책임연구원

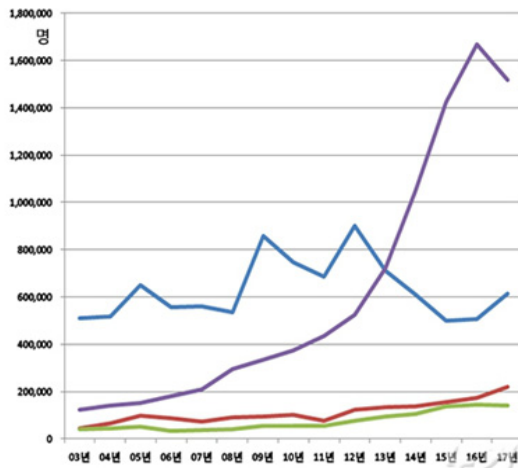
자동통역 기술은 청각지능, 언어지능, 학습지능 등 인간의 지능을 모방하는 복합지능 기술이며, 인간의 능력을 증강하는 초지능 기술로서, 미국의 구글, MS, IBM, 일본 NTT 도코모, 중국 바이두 등 세계 IT 대표기업들 간의 기술개발 경쟁이 치열하다. 국가 간의 경제, 사회, 문화적 융합 가속화로 언어장벽 해소가 국가 경쟁력과 직결됨에 따라 여행/관광, 국제비즈니스, 국제전시/스포츠 행사, UN/국가안보 등 각 산업 분야에서 혁신적인 변화를 가져올 정도로 파급효과가 크다. 이에 세계적으로 자유발화로 끊임 없는 실시간 통역을 목표로 Conversational AI 기반 다국어 자동통역 연구 개발을 추진하고 있는 가운데, 우리도 원천기술 확보를 통한 시장 선점이 요구되고 있다.

I. 서론

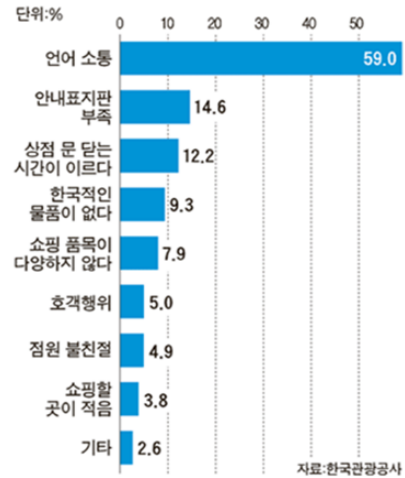
유엔세계관광기구(UNWTO)에 따르면 태국, 싱가포르, 인도네시아, 말레이시아, 중국 등 아시아-태평양 지역 해외관광객이 2019년에 3.5억 명에 달할 것으로 보고된 바 있다. 그리고 전 세계적으로 보급률이 매우 높은 스마트폰은 자동통역, AI 비서 등 음성 기반 인공지능 서비스에 최적화된 모바일 디바이스로, 최근 스웨덴의 에릭슨은 5G 서비스를 시작으로 중국에서 14억 대, 북동아시아나 인도지역은 6.7억 대의 스마트폰이 판매될 것으로 예상한 바 있다. 시장조사기관인 Business Wire Inc.[1]는 자동통역 서비스 시장이

* 본 내용은 김상훈 책임연구원(☎ 042-860-5141, ksh@etri.re.kr)에게 문의하시기 바랍니다.

** 본 내용은 필자의 주관적인 의견이며 IITP의 공식적인 입장이 아님을 밝힙니다.



[3월 외국인 관광객 국내 방문 현황]



[중국인이 한국에서 쇼핑할 때 불편한 점]

〈자료〉 한국관광공사(한국관광 통계자료, 2017)

[그림 1] 여행상황 언어소통 필요성

2020~2025년 동안 CAGR 9.4% 성장할 것으로 예측했으며 여행, 비즈니스 등에서 중국어 수요가 가장 클 것으로 보고 있다. 중국인 12억 명 중 10억 명이 만다린(Mandarin)을 사용하고 있고, 중국의 바이두(Baidu)는 2019년 8월부터 자동통역 API 서비스를 제공하기 시작했다. 한편, 한국관광공사에 따르면 국내 입국하는 해외관광객이 점점 늘어나는 추세이고, 해외 관광객 대상 설문조사 결과 언어소통이 가장 불편하다는 결과가 나온 바 있다([그림 1] 참조).

이처럼 국가 간 인적/물적 교류가 활발해지면서 언어 간 장벽을 허무는 자동통역 기술의 확보는 여행/관광, 국제비즈니스, 국제전시/스포츠 행사, UN/국가안보 등 국가 글로벌 경쟁력과 직결됨을 알 수 있다. Jupiter Research사는 2020년 가장 유망한 12대 미래 기술 중 하나로 실시간 동시통역(real-time translation) 기술을 선정한 바 있고, 2019년 MIT 테크놀로지 리뷰에서는 동시통역 기술을 10대 기술로 선정한 바 있다. 자동통역 기술은 [그림 2]와 같이 다국어 음성인식, 자동번역, 음성합성 등 인간의 복합지능을 모델링하는 고난이도 초지능 기술로서, 미국의 구글, MS, IBM, 일본의 NTT 도코모 등 세계 IT 대표기업들의 기술개발 경쟁이 치열하다. 글로벌화의 가속으로 언어장벽 해소가 각 산업 분야에서 혁신적인 변화를 가져올 정도로 파급효과가 기대됨으로 Conversational



〈자료〉 한국전자통신연구원 자체 작성

[그림 2] 자동통역 요소기술 구성

AI 기반 다국어 자동통역 기술 확보가 무엇보다도 시급하다.

한편, 국내 자동통역 연구는 주로 한국어나 영어 중심의 기술 개발에 주력해 왔지만 삼성 스마트폰, 현대 자동차 등 국내 글로벌 기업들의 제품이 해외 기술과 경쟁을 위해서는 다국어 기술 확보가 매우 중요한 요소가 되었다. 특히, 국내 다문화가족이 늘어남에 따라 119 긴급상황 시 상황판단을 위한 의사소통 체계 구축이 시급하고, K-drama, K-pop 등 한류 콘텐츠에 대한 통역된 자막이 동남아, 남미, 중동 국가 등에 신속히 제공되기 위해서는 다국어 확장이 필요하다. 최근 사례로 보면, UAE 원전 건설현장 회의를 위한 한-영, 한-아랍어 통역기술 적용 가능성이 검토되고 있고, COVID -19 등 비대면 감염병 진단 문의를 위한 다국어 통역이 방역체계에서 필수 기능으로 요구될 것으로 보인다. 이와 같이 경제적으로 중요한 G20 국가의 언어, 사회현안으로 중요한 국내 다문화가족이 사용하는 언어, 국제사회에 기여하기 위한 분쟁지역 언어 등 희소한 언어까지 커버할 수 있는 다국어 자동통역 기술이 필요할 것으로 예상된다[2].

II. 자동통역 연구 동향

1. 국외 동향

최근 딥러닝 기반 인공지능 기술의 발전과 하드웨어적으로는 GPU 고속화로 대용량 데이터 학습이 가능하게 됨에 따라 장기간 기술적, 산업적으로 성장이 정체되어 왔던 음성 인식 및 자동번역 기술이 인간과 비교될 만큼 획기적으로 개선되었다. 이에 이러한 요소기술을 기반으로 하는 자동통역 기술도 실생활에서 요구가 많아지고 적용 분야도 점점 확대됨에 따라 국내외 시장규모도 고성장의 계기를 맞이할 전망이다. 자동통역 기술 수준도 그동안 제한발화나 단문 위주 낭독체 발화를 통역하는 수준에서 대화통역, 강연통역, 회의통역, 전화통역 등 영역이나 발화에 제한이 없는 연속 자유발화형 자동통역 수준으로 인간 중심의 기술로 진화할 것으로 예상되며, 아마존, 구글, 마이크로소프트 등 글로벌 기업을 중심으로 집중 투자가 이루어지고 있다. 자동통역 시장 기술 선점을 위해 일부 업체는 수십 개 다국어를 대상으로 웨어러블, 화상통역 등 초기 서비스 모델을 경쟁적으로 제공하고 있다.

가. 마이크로소프트

통번역 앱인 Microsoft Translator는 70여 개국 자동통번역이 가능하며 텍스트, 음성, 대화, 사진 등 다양한 미디어에 담긴 언어를 번역하는 기능도 제공한다. 특히, 마이크로소프트는 스카이프 화상통화를 통해 영어와 유럽어 간 실시간 통역 서비스를 제공하고 있으며, 스마트워치 통역, 화상통역 등 다양한 서비스 영역에서 구글과 시장 선점 경쟁을 벌이고 있다.

나. 아마존

아마존은 구글, 마이크로소프트와 함께 자동통역 분야 주요 서비스 업체로 AWS에서 Amazon Translate 서비스를 제공하고 있다. 또한, 인공지능 비서인 아마존 알렉사를 통해서도 영어를 다른 48개 언어로 통역해주는 기능을 제공하고 있다. Amazon Translate는 55개 언어 및 변형 언어 간의 통번역 기능을 유료로 제공한다. 기술로는 딥러닝 기법을 사용하였고, 분야에 따른 전문용어, 고유명사 등 번역이 잘 안 되는 용어를 사용자가 별도

로 추가함으로써 성능을 높이고 있다. Amazon Translate에서는 처리한 텍스트의 문자 수를 기준으로 사용한 만큼만 비용을 내는데, 처음 12개월 동안 월별 최대 200만 개의 문자를 무료로 번역할 수 있으며, 100만 자당 15 달러의 사용료를 받고 있다. 번역전문업체인 Intento[3]는 최근 15개 상용 엔진을 대상으로 번역 성능 평가를 시행하였으며, 14개 언어 쌍, 16개 산업 부문 및 8개 콘텐츠 유형에서 Amazon Translate를 2020년 최고의 기계번역 공급자로 선정한 바 있다.

다. 구글

구글은 전 세계적으로 가장 많은 사용자가 구글번역 및 유튜브 동영상 실시간 통역을 이용하고 있다. 43개 언어에 대해 대화를 실시간으로 통역해주고, 103개 언어로 된 텍스트를 번역하는 기능을 가장 먼저 제공했으며, 인터넷 연결 없이 오프라인으로 59개 언어에 대한 번역 기능도 제공한다. 2020년 1월에 CES에서 Live Translation 기능을 프로토타입 수준에서 공개했고, 영어, 유럽어 및 힌디어, 태국어 등 9개 언어에 대해 통역이 가능하며 이 기능은 구글 픽셀버드의 킬러 앱이 될 것으로 보인다.

라. Shenzhen Timekettle Technologies Ltd

중국업체로 40개 언어에 대한 통역 기능을 제공하고, 어댑티브 노이즈 제거를 위해 빔포밍과 뉴럴 네트워크 알고리즘으로 음성입력 품질을 획기적으로 개선하였다. 음성입력감지(VAD) 기술을 적용하여 “Hi Siri”나 “OK Google”과 같이 wake-up을 하지 않고 음성 픽업을 자동으로 수행한다. 인터넷이 안 되는 지역에서도 통역이 가능하다.

마. Waverly Labs

2020년 9월, Waverly Labs는 20개 언어, 42개 방언과 호환되는 통역 서비스를 제공한다고 밝혔는데, 이 서비스는 음성인식 신경망과 결합된 원거리 마이크 어레이를 사용하고 클라우드 기반으로 통역결과를 제공한다. 정확한 번역을 제공하기 위한 개인, 그룹 및 일대일 번역 모드를 제공하고 최대 4명의 참가자와 스마트폰 하나로 통역이 가능하다. 당초 자동통역용 이어셋 개발비를 마련하기 위해 5,000명 이상의 크라우드 펀딩으로 70만 달러 이상을 모금했다.

2. 국내 동향

자동통역을 구성하는 핵심기술인 음성인식, 자동번역, 음성합성 기술은 개발 난이도가 높아 주로 오랜 기간 언어지능, 청각지능 등 인공지능 기술이 축적되어 있는 국가 연구소나 글로벌 기업을 중심으로 연구 개발이 이루어지고 있다. 네이버 파파고는 축적된 음성검색 기술과 번역기술을 통역으로 확대하여 현재 한국어와 영어, 일본어, 중국어 등 총 13개 언어 간 번역이 가능하고, 통역은 스페인어, 불어 등 5개 언어를 지원한다. 인터넷이 안 되는 환경에서도 번역이 가능하다. 삼성에서는 스마트폰 단말에 탑재가 가능한 임베디드 통역기술을 주력으로 개발하고 있다. ETRI는 한국어-영어 자동통역을 시작으로 2012년 일본어, 중국어, 2015년 스페인어, 프랑스어, 독일어, 러시아어, 아랍어 최근에는 베트남어, 태국어, 말레이시아어, 인도네시아어 등 총 13개 언어에 대한 다국어 자동통역 기술을 개발하였다[4]. 2018년에는 평창 동계올림픽 자동통번역 서비스 공식 후원사인 한글과컴퓨터와 함께 지니톡 자동통역 서비스를 올림픽 최초로 공식적으로 제공한 바 있으며(그림 3) 참조), 글로벌 경쟁력을 위해 대상 언어를 늘려 나가고 있다. 또한, 강연, 회의 등 연속 자유발화 자동통역 원천기술 개발에 집중하고 있다.



〈자료〉 한국전자통신연구원 자체 작성

[그림 3] 2018 평창동계올림픽 자동통역 공식 서비스 계획

III. Conversational AI 기반 다국어 자동통역

최근까지 자동통역 기술은 여행이나 일상 등 제한된 영역에서 낭독형 단문을 명료하게 발화해서 통역하는 수준이었으나 국제간 경제, 문화 등 교류가 확대됨에 따라 회의, 강연, 콘텐츠자막 등 자동통역이 필요한 상황이 점차 확대되고 있고, 자동통역 수준도 사용자가 자연스럽게 말하는 자유발화(Conversational Speech)가 가능한 수준으로 요구사항이 높아지고 있다. 이에 이러한 요구수준을 만족하기 위해서는 최대한 자연스러운 자유발화 데이터 구축부터 대화 이해를 위한 복합지능 기술 및 음성인식 오류로 인한 통역한계를 극복해야 하는 좀 더 원천적인 기술 개발이 필요하다. 또한, 국가 위상이 올라감에 따라 국내 다문화가족 119 긴급상황 시 통역 서비스, UN 지원 활동 등 희소한 언어에 대한 다국어 통역 서비스 확대도 필요한 시점이다.

1. 자유발화





자유발화는 즉흥적으로 발성하기 때문에 “계획되지 않은 발화”라고도 하며, 음성적 특징으로는 [그림 4]의 자유발화 예문과 같이 문장이 불완전하거나 말을 더듬기도 하고 문법도 잘 지켜지지 않는다[5].

따라서 이러한 현상을 음성에 담기 위해 최대한 자연스러운 발성을 유도하고자 [그림 5]와 같이 2명 이상이 특정 주제를 가지고 대화하는 데이터를 수집한다. 최근 한국지능정보사회연구원(NIA)에서는 국가주도 인공지능 학습데이터 구축 사업을 통해 대량의 데이터를 구축하고 있으며, 2019년 첫 단계로 약 1,000시간에 해당하는 한국어 자유발화 학습데이터(DB명: KSponSpeech)와 평가데이터를 공개한 바 있다[6].

```
아/ 네 저희가 저희+ 저희 회사와 제휴된 차량 센터를 안내해드릴 겁니다. n/
아/ 네. n/
네 그쪽에서 점검받고 수리하시면은 다른 센터에서 수리하실 때보다 더 저렴하게
비/ 어/ 교체하실 수 있을 것 같거든요. n/
아/ n/
오/ 네 알겠습 오/ 그리고 지금 n/
지금 창기로 엔트하신 것으로 확인되었는데 n/
네네. n/ |
```

〈자료〉 한국전자통신연구원 자체 작성

[그림 4] 자유발화 예문

<p>낭독형 제한발화 음성인식 (As is)</p>	<p>대화형 자유발화 음성인식 (To be)</p>
<ul style="list-style-type: none"> • (낭독형) 발음을 명료하게 또박또박 발성 • (제한발화) 대본있음  <p><대본을 낭독하여 녹음></p> <ul style="list-style-type: none"> • (난이도) 표준규칙에 따라 발성 • (응용사례) 단어나 문장단위 제한발화로 스마트폰 음성검색, 덕데이션, 네비게이션 주소검색, 음성리모콘 등에 제한 범위내 활용  <p><음성검색 및 음성명령></p>	<ul style="list-style-type: none"> • (대화형) 발음이 불명료하고 유창하게 발성 • (자유발화) 대본없음  <p><대본없이 대화상황을 녹음></p> <ul style="list-style-type: none"> • (난이도) 간투사, 생략, 더듬거림, 잘못발성, 사투리 포함 • (응용사례) 사람간 또는 기계와 자유대화로 동시통역, 인공지능 비서, 대화로봇, AI 콜센터, 대화형 전문가시스템 등 활용 급증 예상  <p><대화형 AI 및 동시통역></p>

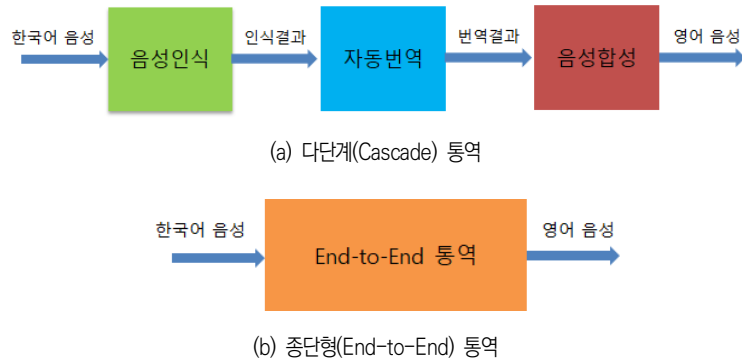
<자료> 한국전자통신연구원 자체 작성

[그림 5] 낭독체 음성과 자유발화 음성의 차이점 비교

2. 종단형(End-to-End) 통역

기존 자동통역 시스템은 [그림 6 (a)]와 같이, 음성인식 모듈과 자동번역 모듈을 각각 학습한 후, 시스템에 음성을 입력하여 나온 음성인식 결과를 다시 자동번역 모듈로 입력해서 최종 통역결과를 얻는 구조이다. 이를 다단계(Cascade) 통역이라고 하며, 이런 경우 음성인식 모듈의 오류가 자동번역 모듈로 전파되어 전체적인 통역시스템의 성능을 현저히 떨어뜨리게 된다. [그림 6 (b)]와 같이 최근 제안된 종단형 통역의 경우, 인식과 번역 학습 모델을 하나로 통합함으로써 음성인식 오류전파를 막을 수 있고, 통역 속도도 개선되는 효과가 있다[7].

일반적으로 종단형 통역은 어텐션(Attention) 기반의 인코더-디코더 구조나 트랜스포



〈자료〉 한국전자통신연구원 자체 작성

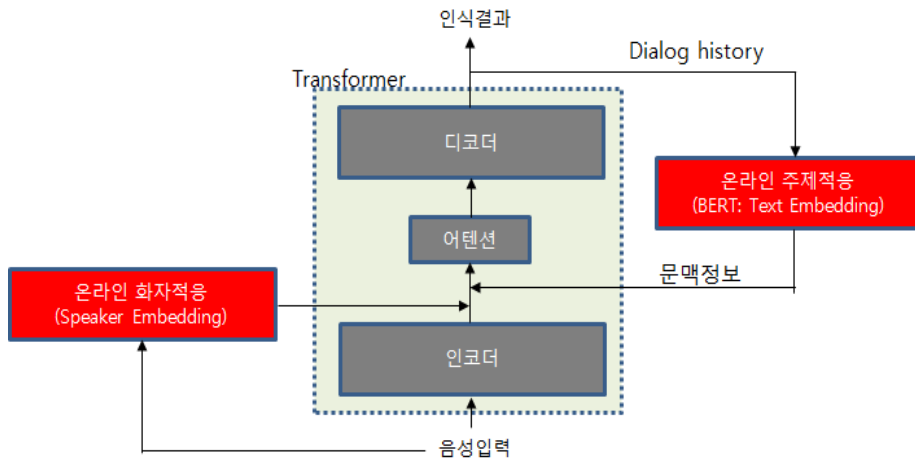
[그림 6] 다단계 및 종단형 자동통역 개념

머(Transformer) 구조가 주로 사용되는데, 실험결과 트랜스포머 구조가 좋은 성능을 보이고 있다. 단점으로는 종단형 통역에 사용되는 학습데이터는 “소스언어 음성, 소스언어 텍스트, 타깃언어 텍스트”로 구성되기 때문에 학습데이터 확보가 용이하지 않고, 또 종단형 인공지능 학습을 위한 데이터가 대략 1만 시간 이상되어야 하므로 이렇게 방대한 데이터를 구축하는 것이 시간이나 비용 측면에서 거의 불가능해 이를 해결하기 위한 연구가 진행 중이다. 최근 타코트론(Tacotron)이라는 인공지능망을 이용한 음성합성 기술이 제안되어 합성음의 명료도나 자연성을 대폭 개선하였고[8], 화자 임베딩(Speaker Embedding) 정보를 이용하면 다양한 음색을 가진 음성생성이 가능하여 종단형 음성인식이나 통역에 학습데이터를 증강하는 목적으로 연구가 활발히 진행되고 있다[9].

종단형 통역이라 하더라도 음성인식 관점에서 성능을 개선하는 연구는 매우 중요하다. 특히, 자유발화인 경우, 화자에 종속적인 측면(예; 발음 불명료, 음색, 말투 등)이 기존 낭독체보다 강하기 때문에 화자적응 기술을 적용하여 좀 더 학습모델에 가까워지도록 해야 한다. 화자적응 방법으로는 [그림 7]과 같은 구조로 i-vector를 특징벡터에 추가하거나 최근에는 화자 임베딩 정보인 x-vector를 이용하는 방법이 연구되고 있다.

자동통역의 경우, 상대방과 대화를 주고받기 때문에 대화의 주제를 트래킹하는 것은 말의 뜻을 제대로 이해하고 의미적으로 정확한 통역에 매우 중요하다.

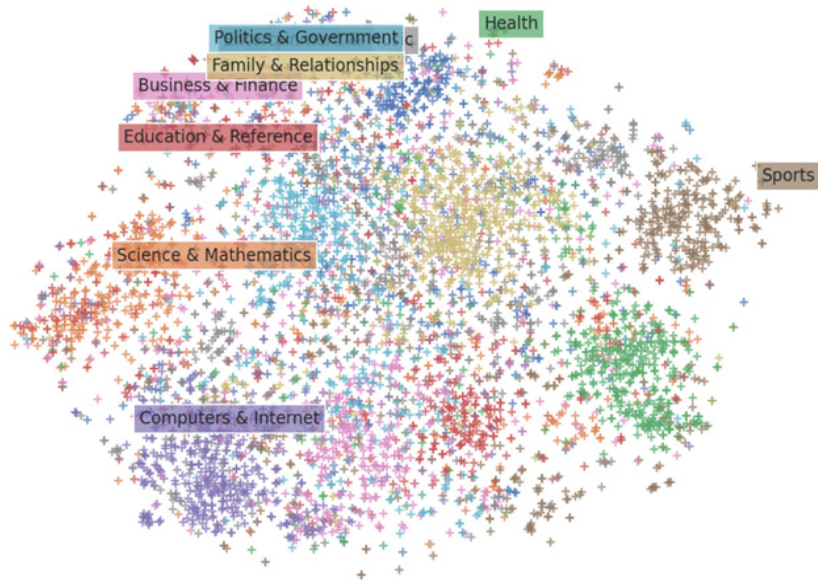
최근 자연어처리 분야에서 BERT(Bidirectional Encoder Representations from Transformer) 임베딩 방식이 개발되어 주제 분류, 유사문장 추출, 질의응답, 문서요약



〈자료〉 한국전자통신연구원 자체 작성

[그림 7] 온라인 주제적응 및 화자적응 구조

등 상당한 성과를 내고 있고, 자동통역에서도 도메인 적응을 위한 유사한 문장을 추출한다든지 [그림 8]과 같이 대화의 주제를 파악한다든지 등의 다양한 적응을 시도하고 있다 [10]. 온라인 주제적응은 이전 발화의 인식결과로부터 BERT를 통해 주제에 해당하는 임



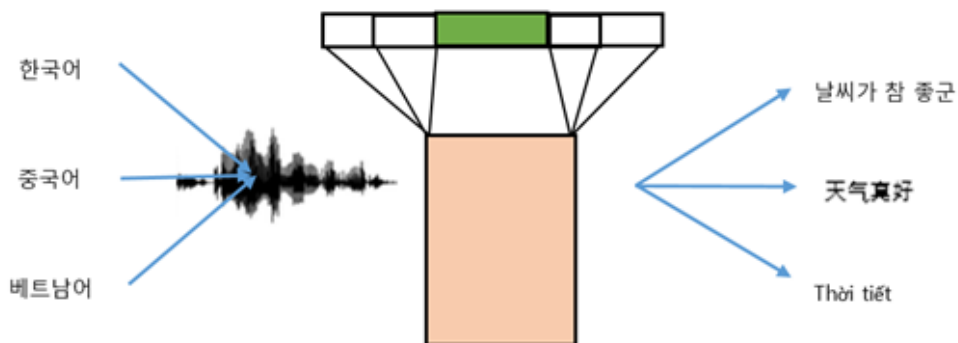
〈자료〉 한국전자통신연구원 자체 작성

[그림 8] 주제별로 구분되는 문장 임베딩 결과

베딩 벡터를 추출하여 특징 벡터에 추가함으로써 음성인식이나 자동통역의 성능 개선이 가능하다. 이와 같이 종단형 통역은 세계적으로 연구 초기 단계로 아직 학습데이터 제약이라는 큰 장애가 있지만 기존 다단계 통역 방식의 한계를 극복할 수 있는 기술로 기대가 된다.

3. 다국어 확장

다국어 확장이 가능해지려면 무엇보다도 다국어 데이터 확보가 중요한데, 통상 종단형 구조의 인공지능을 학습하기 위해서는 수천~수만 시간 이상의 데이터가 필요하다. 한, 중, 영, 일본어의 경우, 공개 데이터를 사용하거나 상용으로 구입해서 사용하거나 국내에서 데이터를 직접 구축해서 확보할 수 있지만, 유럽어나 동남아어, 힌디어, 몽골어, 우즈베키스탄어, 페르시아어 등 다소 희소한 언어의 경우 국외에서도 확보가 매우 어렵다. 따라서 [그림 9]와 같이 이미 구축된 다수 언어를 통합해서 멀티태스크 학습을 하거나, 타코트론 합성기나 신경망 번역기로 종단형 통역용 학습데이터를 증강하는 인위적 생성 방법이 대안으로 연구되고 있다[11].



〈자료〉 한국전자통신연구원 자체 작성

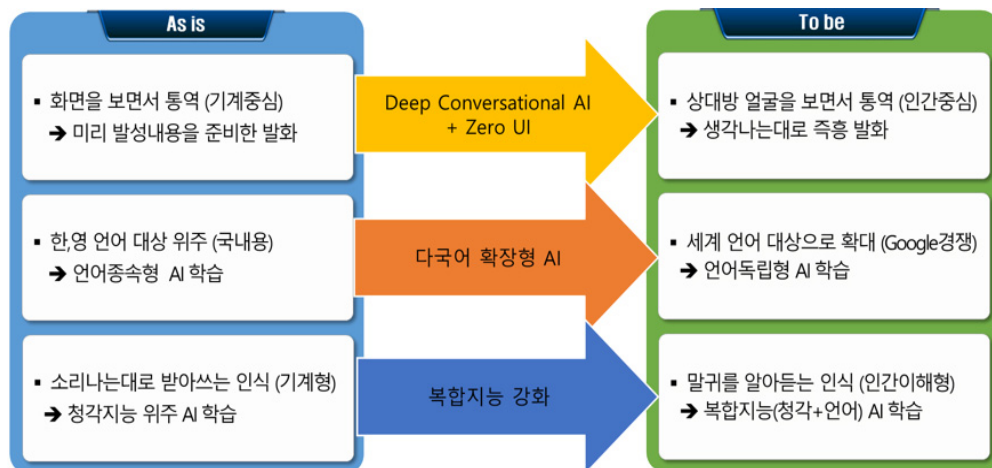
[그림 9] 다국어 데이터를 이용한 멀티태스크 학습

딥러닝 네트워크 구조도 pre-training 모델과 추가 전이학습을 어떻게 해야 하는지에 따라 새로운 구조가 제안되고 있으며 일부 개선 효과를 보이고 있다. 번역 데이터의 경우, 인위적으로 생성할 경우 종단형 통역기 성능에 제한적일 수 있는데, 이에 번역기만 성능 개선을 위한 mono-lingual 텍스트 데이터를 활용한 방법도 연구가 진행되고 있다.

IV. 결론

현재 자동통역 기술은 자유발화 대화형 다국어 통역으로 진화하고 있으며, 학습데이터 부족 문제, 인식오류 전파 문제, 대화 의도 파악 등 Conversational 음성에 나타나는 문제를 해결하는데 주력하고 있고 일부 성과도 보이고 있다. 이러한 문제 외에 끊임 없는 대화형 통역이 되기 위해서는 스마트폰 화면을 터치 후 말해야 하는 인터페이스를 개선해서 사용성을 대폭 높여야 하며, 해외지명, 이름 등 외래어로 된 고유명사 인식이 잘 안 되는 code-switching 문제 등 실용적인 측면에서도 개선해야 할 부분이 많다.

자동통역 기술을 둘러싸고 구글, MS, IBM 등 세계 IT 대표기업들의 기술개발 경쟁이 치열하게 전개되고 있는 가운데, ETRI에서는 자유발화로 끊임 없는 실시간 통역 및 다국어 확장이 용이한 원천기술 개발을 목표로 연구 개발을 추진하고 있다(그림 10) 참조). 이를 통한 원천기술 확보와 기술선점을 토대로 세계 시장을 주도할 수 있기를 기대해 본다.



〈자료〉 한국전자통신연구원 자체 작성

[그림 10] Conversational AI 기반 자동통역 서비스 시나리오

[참고문헌]

- [1] Business Wire Inc., "Global Speech to Speech Translation Market(2020 to 2025) - Growth, Trends and Forecast," Research and Markets, 2020.
- [2] 김승희, 윤승, 조훈영, 최승권, 김상훈, "다국어 자동통역 기술동향 및 응용", 한국전자통신연구원,

- 전자통신동향분석, 26권 5호(통권 131), 2011, pp.1-12.
- [3] Intento Inc., "The State of the Machine Translation," Intento Report, 2020.
 - [4] 김승희, 박준, 김상훈, "자동통역기술, 서비스 및 기업 동향", 한국전자통신연구원, 전자통신동향분석, 29권 4호(통권 148), 2014, pp.39-48.
 - [5] 권오욱, 최승권, 노윤희, 김영길, 박전규, 이윤근, "자유발화형 음성대화처리 기술동향", 한국전자통신연구원, 전자통신동향분석, 30권 4호(통권 154), 2015, pp.26-35.
 - [6] Jeong-Uk Bang, Seung-Hi Kim, Mu-Yeol Choi, Min-Kyu Lee, Yeo-Jeong Kim, "KSponSpeech: Korean Spontaneous Speech Corpus for Automatic Speech Recognition," Applied Sciences, 10(19), 2020.
 - [7] Ye Jia, Ron J. Weiss, Fadi Biadsy, Wolfgang Macherey, Melvin Johnson, Zhifeng Chen, Yonghui Wu, "Direct speech-to-speech translation with a sequence-to-sequence model," Proc. Interspeech 2019, pp.1123-1127,
 - [8] 최연주, 정영문, 김영관, 서영주, 김희린, "한국어 text-to-speech(TTS) 시스템을 위한 엔드투엔드 합성 방식 연구", 말소리와 음성과학, 제10권, 제1호, 2018, pp.39-48.
 - [9] Ye Jia, Yu Zhang, Ron J. Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, Yonghui Wu, "Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis," NeurIPS, 2018.
 - [10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Proceedings of NAACL-HLT, 2019, pp.4171-4186.
 - [11] Andrew Rosenberg, Yu Zhang, Bhuvana Ramabhadran, Ye Jia, Pedro Moreno, Yonghui Wu, Zelin Wu, "Speech Recognition with Augmented Synthesized Speech," ASRU, 2020.