딥러닝이 탐구하지 못한 언어와 5가지 태스크

Tech Insider

글·정리 | 이수경 samantha.lee@kakaobrain.com 2016년 3월 알파고와 이세돌 9단이 펼치는 세기의 대결을 두 눈으로 목도한 이후 AI를 제대로 공부해봐야겠다고 결심했습니다. AI의 본진이자 연구소인 카카오브레인으로 걸어들어온 이유입니다. 인공지능 기술과 이로 인해 바뀔 미래 사회를 다루는 글을 통해 사람들과 소통하고 싶습니다.

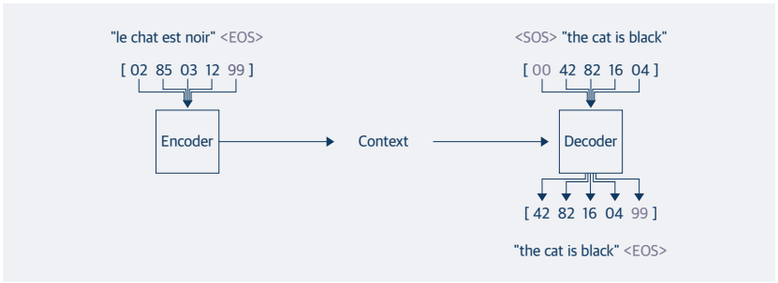
발표·감수 | 박규병 kyubyong.park@kakaobrain.com 카카오브레인에서 언어를 재료로 한 딥러닝 연구를 하고 있습니다. 언어와 교육, 생명정보, 환경 등에 관심이 있습니다. 기술을 중시하지만, 기술이 아닌 사람이 중심이 되어야 한다고 믿습니다. https://github.com/kyubyong에서 그동안의 연구 활동과 사진을 보실 수 있습니다.

그간 딥러닝(deep learning)은 기계 번역(machine translation), 음성 인식(speech recognition), 음성 합성(speech synthesis) 등 여러 응용 분야에서 기술적으로 큰 발전을 이루었다. 그러나 딥러닝으로 충분히 해결하지 못한 언어 관련 과제들은 여전히 많다. 대화 시스템(dialog systems)이나 챗봇(chatbot)이 대표적인 예다. 상업적인 가치가 큰 것으로 알려진 챗봇의 경우 소비자의 만족을 끌어낼 정도의 기술 수준에는 여전히 도달하지 못했다. 그 외 딥러닝이 아직 충분히 해결하지 못한 언어 과제에는 무엇이 있을까?

첫 번째, 비지도 기계 번역(unsupervised machine translation)

NMT(neural machine translation)의 첫 번째 모델인 seq2seq(sequence to sequence) 네트워크[*1]는 지도 학습(supervised learning)[*2] 방식의 기계 번역에 큰 혁신을 가져왔다. 문장 전체의 모든 정보, 즉 단어 의미와 단어 순서, 문장 구조, 단어 간의 의존 관계 등 번역에 필요한 모든 정보(문장 벡터)를 토대로 문맥을 이해하는 자연스러운 번역 결과를 냈기 때문이다.[*3]

[그림 1] 인공신경망 기반의 기계 번역 구조[*4]



[그림 1]에서처럼 'ie chat est noir'를 입력하면 'the cat is black'이라는 번역 문장을 도출하는 원리는 다음과 같다. 먼저 단위정보(단어(word) 또는 토큰(token)) 시퀀스[*5]를 인코더(encoder)에 입력한다. 인코더는 이를 분석해 고정 길이의 벡터 표현(vector representation)을 추정한다. 디코더(decoder)는 이 벡터를 활용해 또 다른 단위정보의 시퀀스를 생성한다.[*6]

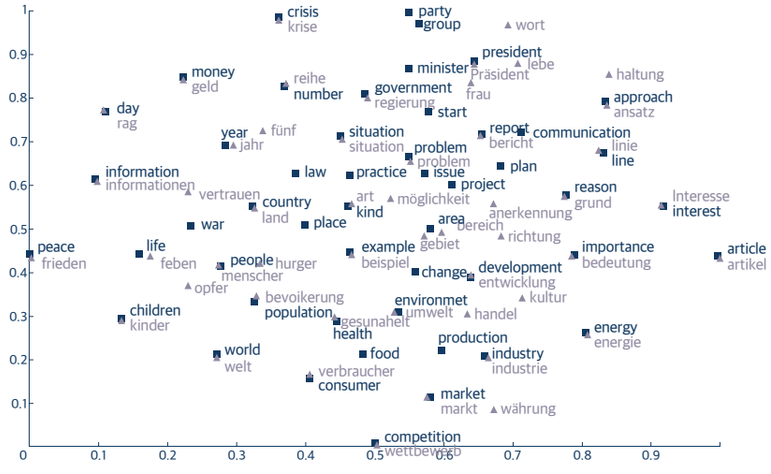
양질의 대규모 병렬 말뭉치(large parallel corpus)가 필요한 지도 학습에서는 데이터 양이 충분치 않으면 학습효과가 거의 없다.[*7] 이처럼 병렬 말뭉치 데이터가 부족한 상황에서는 어떻게 신경망을 훈련해야 할까? 라벨링(labeling)된 데이터와 그렇지 않은 데이터 모두를 훈련에 사용하는 준지도 학습(semi-supervised learning)[*8] 방법이 제안됐으나 큰 성과는 없었다. 최근에는 비지도 학습 교차언어 임베딩(unsupervised cross-lingual embeddings)이 가시적인 성과를 내고 있다.[*9] 훈련 과정은 다음과 같다.

*1 참고 | https://datascience.stackexchange.com/questions/25411/what-exactly-is-theinput-of-decoder-in-autoencoder-setup
*2 설명 | 기존 모델이 가지고 있는 입력과 출력 사이 길이 차이로 생기는 문제를 고정된 상태 벡터를 출력으로 내놓게 함으로써 해결함. 물론 가장 단순한 형태의 seq2seq 모델은 만족할 만한 수준의 성능을 내지 못함. 신경망(neural networks)이 가진 잠재력을 확인한 정도라 해도 무방함(신경망 번역 모델의 진화 과정 https://brunch.co.kr/@kakao-it/155)
*3 설명 | 지도 학습: 정답이 라벨링(labelling)된 데이터로 모델을 학습시키는 방식
*4 참고 | 일상생활 속으로 들어온 기계번역. http://kiss.kstudy.com/public/public2-article.asp?key=50905527
*5 설명 | 시퀀스: 연속된 입력 데이터. 문장(sentences)은 문법과 같은 규칙에 의거한 일련(sequential)의 단어 집합이라는 점에서 시퀀스의 예들 가운데 하나라고 볼 수 있음
*6 참고 | Sequence to Sequence 기술을 이용한 음성 합성 기술개발 동향. http://www.riss.kr/search/detail/DetailView.do?p_mat_type=1a0202e37d52c72d&control_no=2ab1d078bd0d406d4884a65323211ff0
*7 설명 | 머신러닝 성능에 영향을 미치는 요소를 중요도순으로 나열하면 양질의 데이터>좋은 특징(feature)>효율적인 알고리즘이라고 할 수 있음
*8 설명 | 라벨링 데이터를 얻는 일은 사람의 손을 거쳐 일일이 만들어야 하는 만큼 큰 비용을 수반함. 이에 연구자들은 대규모 비(非)라벨링 데이터에 더해 적은 양의 라벨링 데이터를 활용할 경우 학습 정확도가 높아짐을 확인함
*9 참고 | Learning bilingual word embedding

(1) 먼저 언어 A와 언어 B에 대한 단일 언어 말뭉치(monolingual corpora)를 이용해 각각의 단어 임베딩(word embedding)을 생성한다.
(2) 가장 쉬운 단어들만 모아 대역(bilingual) 쌍 샘플 목록을 구축한다.
(3) 대역 말뭉치를 공유 임베딩 공간(shared embedding space)에 매핑(mapping)하는 변환 네트워크(transformation network)를 학습시킨다. 그러면 나머지 다른 단어들도 자동으로 비슷한 위치에 매핑된다.

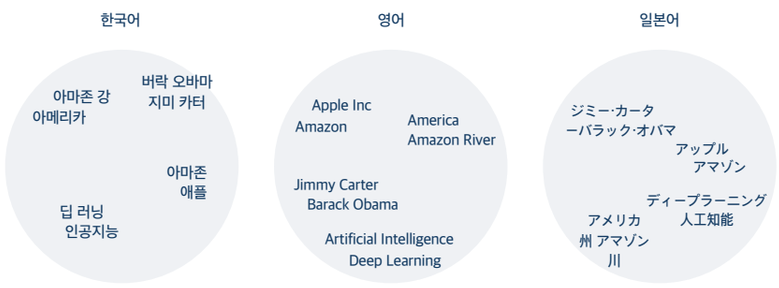
이 방식은 일부 언어 쌍에 대한 지도 학습 기반 교차언어 임베딩보다 성능이 좋은 것으로 알려졌다.*10

[그림 2] 영어와 독일어 말뭉치를 공유 공간에 매핑한 모습*11



하지만 이런 접근 방식에는 근본적인 한계가 있다. 워드 임베딩 그 자체는 비결정론적(non-deterministic)이다. 알고리즘 학습에 활용하는 말뭉치 종류와 정렬 방식, 어떤 알고리즘을 활용하는지에 따라 임베딩 결과는 천차만별로 달라질 수 있다는 의미다. 이로 인해 같은 의미를 나타내는 단어 A(언어 a)와 단어 B(언어 b)가 있으면 각 언어라는 맥락에서 봤을 때 다른 단어 분포(word distribution)를 가질 확률도 배제할 수 없다.

[그림 3] 같은 의미를 뜻하는 단어가 한국어, 영어, 일본어에서 다른 단어 분포를 보인다.*12



아울러 word2vec은 특정 문맥이나 문장 내에서 단어가 갖추는 특정한 의미를 이해하지 못한다. word2vec은 문장 내 한 단어와 같이 출연하는 다른 단어를 관련 단어로 학습하는 것일 뿐 단어의 의미를 학습하지는 않기 때문이다.*13 예를 들어, '민수는 지난주 수술을 받고 다리(의미 1)를 고쳤다'와 '서울 한강에 있는 다리(의미 2)는 총 28개다'라는 문장을 보자. 여기서 학습 모델은 각 문장에 등장하는 '다리'의 의미 차이를 이해하지 못한다. 단지 철자의 다름을 통해 서로 다른 단어를 인식할 뿐이다.

문장별 '다리'의 의미 차이
(의미 1) 사람이나 동물의 몸통 아래 붙어 있는 신체의 부분. 서고 걷고 뛰는 일 따위를 맡아 한다.
(의미 2) 물체의 아래쪽에 붙어서 그 물체를 받치거나 직접 땅에 닿지 않게 하거나 높이 있도록 버티어 놓은 부분.

어휘집(seed lexicon)이 되는 적은 양의 병렬 말뭉치를 초기 학습에 이용한다는 점에서 온전한 비지도학습이라고 보기 힘든 어려운 측면도 있다. 따라서 한 언어에 대한 대규모 단일 언어 말뭉치(monolingual corpora)만을 가지고도 언어 간 번역을 가능케 하는 기술 연구가 필요하다.

단어 임베딩(word embedding)
단어, 문장 등의 의미를 나타내기 위한 분산 표상(distributed representation)으로, 저차원의 벡터 공간으로 표현된다. 대표적인 임베딩 학습 방법으로는 word2vec이 있다.

word2vec*14
같은 맥락에 있는 단어는 가까운 의미를 가진다는 전제에서 출발한다. 한 단어와 함께 출연하는 다른 단어를 관련 단어로 인공신경망에 학습시킨다. 문서 내 연관 단어들은 가까운 곳에 출현할 가능성이 높아 학습을 반복하는 과정에서 주변 단어가 비슷한 두 단어는 가까운 벡터 공간에 놓이게 된다.

두 번째, 맥락 인지 기계 번역(context-aware machine translation)
전형적인 기계 번역은 문장 단위로 번역된다. 그러나 이런 접근 방식은 이상적이지 않을 수도 있다. 예를 들어, [예시 1]처럼 'the bank'가 포함된 a 문장은 '저기 은행(the bank)을 봐라'로 번역될 수 있다. 그러나 b 문장처럼 문맥에 따라서는 '둑(the bank)'으로 해석될 수 있지만 상용 번역기는 '저기 은행을 보세요. 남자들이 낚시하고 있어요'로 번역한다. 문장 단위 번역이 문맥을 고려하지 못하고 있음을 보여주는 단적인 예이다.

[예시 1]
a. Look at the bank over there.
b. Look at the bank over there. **Men are fishing from it.**

문맥을 고려하지 않은 기계 번역이 가진 문제점을 보여주는 또 다른 예시를 보자([예시 2]).

[예시 2]
He took another full piece and chewed it. "It is **a strong full-blooded fish**," he thought. "I was lucky to get **him** instead of dolphin. Dolphin is too sweet. This is hardly sweet at all and all the strength is still in it." - From the Old Band and the Sea

(구글 번역 결과) 그는 또 다른 조각을 가져다가 씹었다. "그것은 **강한 전혈 물고기**입니다." 그는 생각했다. "나는 돌고래 대신에 **그**를 잡아서 운이 좋았다. 돌고래는 너무 달다. 이것은 전혀 달콤하지 않으며 모든 힘이 여전히 남아 있습니다."

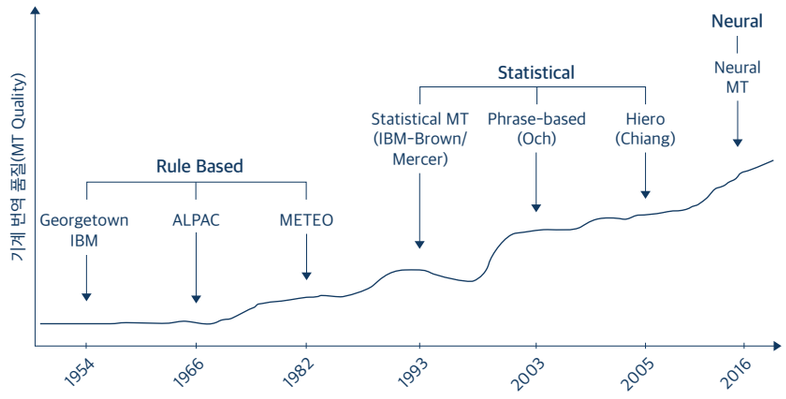
*10 참고 | https://github.com/kweonwooj/papers/issues/100
*11 참고 | http://ruder.io/cross-lingual-embeddings
*12 참고 | http://youngl.org/wp-content/uploads/2017/12/위키와-다국어-검색-이만재-박사.pdf
*13 설명 | 이런 문제를 보완하고자 sense2vec과 같은 모델이 개발되기도 함
*14 참고 | Word2Vec을 이용한 반복적 접근방식의 그래프 기반 단어 중의성 해소, http://www.dbpia.co.kr/Journal/ArticleDetail/NODE06646570#

문맥상 'him'은 '그(사람)'가 아니라 '물고기(fish)'를 지칭한다. 따라서 '나는 돌고래 대신에 그녀석(물고기)을 잡았다'라는 번역이 조금 더 자연스럽다. 그러나 번역기는 문맥을 고려하지 않은 결과를 도출했다. 아울러 톤이 일관적이지 않은 현상(~했다, ~했습니다)도 확인할 수 있다. 그렇기 때문에 자연스러운 결과물을 얻기 위해서는 문법성(grammatical gender)*15, 단수/복수(number or plurality), 경어체(honorifics), 생략(ellipsis)*16과 같은 언어학적 특징(linguistic features)을 고려해야 하는 이유다.

이런 언어학적 특징을 잘 포착하기 위해서는 문맥 파악이 선행되어야 한다. 그렇다면 문맥은 어떻게 정의해볼 수 있을까? 그 방식에 따른 여러 가지 방법론을 생각해볼 수 있다. 한 논문에서는 메모리 네트워크(memory network)를 제안했다. 보통은 번역하려는 문장 바로 앞에 나오는 문장을 함께 학습하면, 실제로 이전보다 번역 성능이 높아짐을 확인할 수 있었다. 하지만 현재 문장을 번역하는 데 필요한 모든 정보가 바로 앞 문장에만 있지 않다는 사실은 직관적으로 알 수 있는 부분이다. 이에 따라 앞서 등장한 모든 정보를 기억하는 외부저장장치(external memory)를 두는 방안도 고려해볼 수 있다. 다만 이 방법은 모델 복잡도가 높아지는 경향이 짙다.

기계 번역의 역사

[그림 4] 기계 번역의 역사 © 스탠퍼드대학교*17



규칙 기반 기계 번역(rule-based machine translation, RBMT)

제2차 세계대전이 끝나고 냉전이 시작된 1950년대, 미국 국방성은 러시아어를 자동으로 번역하는 연구를 시작했다. 이때부터 1970년대에서 1980년대까지는 사람이 실제로 번역하는 프로세스를 본떠서 만든 기계 번역이 고안됐다. 형태소 분석(morphological analysis), 구문 분석(synthetic analysis), 의미 합성(semantic composition)의 과정을 거쳐 중간언어(interlingual)를 만든 다음 역의 과정을 거쳐 중간언어를 타깃 언어의 단어로 맵핑(mapping)하는 식이다. 그러나 안타깝게도 이 방식으로는 완벽에 가까운 기계 번역을 제대로 해낼 수 없음이 드러났다.

통계적 기계 번역(statistical machine translation, SMT)

1988년 IBM은 NLP(natural language processing) 팀을 아무리 크게 키워도 도저히 희망이 안 보인다고 말하기까지 했다. '언어학자 한 명을 해고할 때마다 번역의 정확도가 높아진다'는 프레데릭 제리넥(Frederick Jelinek)'의 말이 우스갯소리처럼 떠돌 정도였다. 그 이유는 언어학자들이 설계한 규칙 기반 시스템보다 확률에 기반한 '바보 같은' 접근 방식의 결과가 더 좋았기 때문이다. 이로써 1990년대를 전후로 SMT 시대가 열렸다.

신경망 기계 번역(neural machine translation, NMT)

인공신경망에 기반한 기계 번역의 기본 구조는 다음과 같다. 먼저 입력 문장을 문장 벡터 형태로 변환, 이를 기반으로 출력 언어의 문장을 생성하는 방식으로 번역을 수행한다. 입력 문장을 문장 벡터로 변환하는 인코더와 출력 문장을 생성해내는 디코더는 인공신경망으로 구성돼 있다. 데이터를 통해 인코더와 디코더의 파라미터(parameter) 값이 조정된다.

*15 설명 | 단어에 남성과 여성, 중성과 같은 성(性)을 부여하는 것
*16 설명 | 동일한 언어 표현 중 생략이 일어나기도 함. 예제 문장은 다음과 같음. All the children had an ice-cream today. Eva chose strawberry (ice cream), Arthur had orange (ice cream) and William (had orange ice cream) too, 《담화분석》 - 현대 영어학 총서9, pp.97~98
*17 참고 | https://nlp.stanford.edu/projects/nmt/Luong-Cho-Manning-NMT-ACL2016-v4.pdf

세 번째, 맞춤법/문법 검사기(grammar checker)

문법이나 맞춤법, 띄어쓰기 등의 오류(error)는 콘텐츠에 대한 신뢰도를 떨어뜨린다. 글을 쓸 때 심적 부담감을 일으키는 요인이다. 이에 사람들은 맞춤법/문법 검사기를 이용해 오류를 최소화하는 데 힘쓴다. 맞춤법/문법 검사기는 다음과 같은 오류를 찾아내 바로잡는다.

(1) 맞춤법/문체(spelling/style) Is this orgnal? → Is this original?
(2) 문법(grammar) I have an apples. → I have an apple.
(3) 연어(collocation)*18 I like powerful coffee. → I like strong coffee.
(4) 어조(tone)/태(voice)*19/격식(formality)

이런 오류를 찾아내기 위해 취했던 전통적인 방법은 다음과 같이 다섯 가지로 볼 수 있다.

(1) 지식 베이스(knowledge base)에 존재하는 지식과 비교 · 대조한다. 매칭되는 결과가 없다면 사용자 입력값에 오류가 있다고 인지한다. 대표적으로 철자 오류는 어휘 사전에 존재하지 않으므로 단순한 형태소 분석을 통해서도 그 오류의 검출 및 교정이 비교적 간단하다고 볼 수 있다.
(2) 대량의 말뭉치에 기반한 통계치를 바탕으로 주어진 문장의 구문을 분석하는 통계(statistics) 기반 분석 방법이 있다. 예를 들어, 사람들이 자주 틀리는 표현을 구문 분석에 활용하는 방식이다.
(3) 몇 번의 추가 편집으로 단어 A를 단어 B로 만들 수 있는지를 측정하는 편집 거리(edit distance, Levenshtein distance)를 이용하기도 한다. 철자 오류와 편집 거리가 작은 단어를 추출하고, 이를 철자 오류가 발생한 단어와 치환하는 식이다.
(4) 주어진 문맥상에서 다음에 나올 적절한 글자, 단어 또는 문장을 예측하는 언어 모델(language model)을 만드는 방법도 있다. 예를 들어, '요금 결재'에서 '결재'는 (1)에서 등록된 올바른 단어이지만 함께 쓰인 '요금'이라는 문맥을 고려해 '결제'가 되도록 바꾼다.*20
(5) 전문가가 만든 규칙(human-generated rules)을 통해 문법적 오류를 수정하는 방법도 있다. 이를 이용하면 언어에 대한 상세한 구조를 정확하게 파악할 수 있다. 규칙이 많을수록 성능을 높일 수 있는 이유다.

지금까지는 대체로 규칙이 더 나은 성능을 보여왔다. 오늘날 가장 많이 사용되는 한국어 맞춤법/문법 검사기 또한 규칙에 기반을 둔 문맥에 의존해 철자 오류를 교정한다. 문제는 전문가의 지식에 기반하여 규칙을 만드는 만큼, 막대한 비용과 시간이 소요된다는 점이다. 아울러 발생 빈도가 높거나 정형화된 오류는 규칙으로 찾아낼 확률이 높으나, 그렇지 않은 유형은 규칙으로도 교정하지 못할 확률이 높다. 딥러닝이 더 나은 성능을 낼 것으로 기대되는 이유다.

맞춤법/문법 검사에 대한 딥러닝의 접근 방식은 크게 두 가지로 나누어볼 수 있다. 하나는 최대우도추정(maximum likelihood estimation, MLE)*21을 이용한 seq2seq이다. 바르게 입력되지 않은 문장 X(source)를 입력받아 바르게 입력된 문장 Y(target)를 반환할 확률을 최대로 하는 모델 파라미터를 학습한다. 다른 하나는 강화 학습(reinforcement learning)이다. 강화 학습은 누적 보상을 최대화해 소프트웨어 에이전트(AI)가 특정 목표에 도달하기 위해 취해야 할 조치를 결정하도록 가르친다. 감독 학습과 달리 강화 학습에서는 라벨링된 입력값과 결과값이라는 데이터 쌍을 필요로 하지 않는다. 인공지능은 자신만의 방식을 찾아 명시적인 지시를 받지 않고도 성능을 개선한다. 여기에서는 성능을 미세조정(fine tuning)할 때 seq2seq를 활용하는 경우가 많다.

그러나 이런 방법론은 규칙에 기반을 둔 전통적인 접근 방식만큼 좋은 성과를 내지 못한다. 먼저 올바른 문장으로만 구성된 데이터 세트를 학습한 딥러닝 알고리즘은 단순히 정답을 내놓을 줄만 안다. 이는 틀린 문장을 입력받아도 오류로 인식하지 못한다는 의미다. '틀린 문장, 교정된

*18 설명 | 특정한 뜻을 나타낼 때 흔히 함께 쓰이는 단어 간의 결합을 뜻함
*19 설명 | 동작을 바라보는 관점 차이로 발생하는 동사의 표현 형식을 일컫는 말
*20 참고 | 어절 N-gram을 이용한 문맥 의존 철자 오류 교정 http://dx.doi.org/10.5626/JOK.2014.41.12.1081
*21 참고 | 주어진 샘플 x에 대해 우도(likelihood)를 가장 크게 해주는 모수를 찾는 방법 https://datascienceschool.net/view-notebook/864a2cc43df44531be32e3fa48769501

문장'으로 구성된 병렬 말뭉치를 만들더라도, 만족할 만한 성능을 보이는 알고리즘을 만드는 데 충분한 규모를 확보할 수 있을지에 대해서도 미지수다. 합성 데이터 생성(synthetic data generation)을 통해 랜덤하게 틀린 문장을 생성하는 방안을 고려해볼 수는 있으나, 괄목할 만한 성과를 내는 전략으로는 충분하지 못할 것으로 보인다.

네 번째, 한정된 자원을 활용한 음성합성(speech synthesis with limited resources)

음성합성(speech synthesis)은 주어진 텍스트를 정확하고 자연스럽게 읽는 음성 데이터를 생성하는 기술을 의미한다. 주어진 텍스트에 대응하는 음성을 출력한다는 점에서 TTS(text to speech)라고 표현하기도 한다.

1990년대 초 규칙 기반 음성합성 기술에 이어, 2000년대에는 대용량 음성 데이터 기반 음편 조합(unit selection) 방식이 주목을 받았다. 전통적인 신호처리 방식 중 대용량 음성 데이터 기반인[*22] 이 방식은 짧은 단위의 음편을 저장한 후 텍스트에 해당하는 음편을 연결해 합성음을 출력한다. 이 방식은 실제 사람의 음성을 녹음한 음편이기에 원음에 가까운 고품질 합성음을 만들 수 있다는 장점이 있는 한편, 음편의 연결 경계가 부자연스럽고 주어진 문장에 항상 똑같이 말하는 문제 등의 단점을 지닌다. 이러한 한계를 극복하고자 통계적 파라미터 방식이 고안됐으며, 이후 은닉 마코브 모델(hidden markov model, HMM)과 심층 신경망(deep neural network, DNN)을 기반으로 한 모델로 발전하였다.[*23] 현재 웨이브넷(WaveNet), 타코트론(Tacotron), 딥보이스(Deep Voice) 등의 LSTM(long short term memory)을 이용한 seq2seq 네트워크가 주로 활용되고 있다.[*24]

오늘날의 음성합성 시스템은 크게 3단계를 거친다. 첫 번째 단계에서는 문장 데이터에서 특징 정보를 분석한다. 두 번째 단계에서는 문장 내 특징 정보로부터 음성 파라미터를 추출한다. 마지막으로 세 번째 단계에서는 음성 파라미터에서 음성을 재구성(vocoder)한다. 문자 단위로 입력된 텍스트로부터 음성을 합칠 수 있도록 하나의 모듈로 이루어진 E2E(end to end) 시스템이라고도 볼 수 있다. 기존 방식과 달리 각 모듈에 대한 전문적인 지식이 필요하지 않아 진입장벽이 낮고, 각 모듈에서 손실(loss)이 누적되는 문제를 해결할 수 있다.[*25]

현재의 음성합성 기술은 인간과 거의 비슷한 음성을 구사할 수 있는 기술 수준까지 도달했다. 특히 지난 2017년 12월 구글(Google)이 발표한 타코트론2(Tacotron2)는 현존하는 최고 음질의 합성음을 만들어내고 있다. 하지만 여기에는 선행조건이 따른다. 원하는 수준의 목소리 품질을 얻기 위해서는 잡음(noise)이 없는 깨끗한 음질을 다량 확보해야 한다는 점이다. 적어도 한 사람의 목소리로 구성된 20시간 이상 분량의 녹음 데이터가 필요하다. 그러나 현실에서는 이런 데이터를 쉽게 얻을 수 없고, 얻을 수 있더라도 스튜디오 녹음 및 편집 비용이 많이 든다. 기존에 보유하고 있는 데이터 세트는 크기 자체가 부족해 배우지 않은 발음과 억양이 생길 수밖에 없다.

따라서 어느 정도 잡음이 있거나 적은 분량의 음성 데이터로도 깨끗한 목소리를 생성할 수 있는 강인한(robust) 인공지능을 만드는 연구가 후행되어야 할 것이다. 현재 1분 정도의 샘플 데이터만으로는 고품질의 음성 합성 엔진 개발은 현실적으로는 불가능에 가깝다. 다만 해당 목소리를 보유한 사람의 목소리인 척 흉내를 내는 모델 정도는 기대해볼 수 있다.

*22 참고 | 한국어 TTS 시스템을 위한 엔드투엔드(end to end) 합성 방식 연구, https://doi.org/10.13064/KSSS.2018.10.1.039
*23 참고 | 지식 기반 프랑스어 발음열 생성 시스템(A knowledge-based pronunciation generation system for French) https://doi.org/10.13064/KSSS.2018.10.1.049
*24 참고 | Sequence to Sequence 기술을 이용한 음성합성 기술개발 동향
*25 참고 | 한국어 TTS 시스템을 위한 엔드투엔드 합성 방식 연구

다섯 번째, 제주어 재활성화(jejueo revitalization)

지구상에는 6000여 개의 언어가 존재한다. 이중 인류의 절반 규모인 50억 명 정도가 20개의 상위권 언어를 구사한다. 그리고 상위권 언어를 제외하면 한 언어를 구사하는 인구가 1만 명도 되지 않는다. 소수가 사용하는 언어의 비중이 절대적으로 높다는 의미다. 소수 언어를 사용할 사람이 점차 없어지면 해당 언어는 지구상에서 사라진다. 학계에서 세계 언어의 50%가 10년 내 소멸할 것으로 전망하는 배경이다. 이는 6000여 개의 언어를 매개로 향유하던 무형자산이 사멸될 수 있음을 시사한다.[*26*27]

제주어 또한 소멸 위기에 처한 언어 중 하나다. 이에 유네스코(UNESCO)는 지난 2010년 제주어를 소멸 위기 언어 5단계 중 사멸 직전인 4단계(아주 심각한 위기에 처한 언어(critically endangered language))로 분류했다. 제주어를 사용하는 인구가 급격하게 감소하고 있기 때문이다.

유네스코에서 소멸 위기에 놓인 언어를 판단하는 9가지 기준[*28]
(1) 세대 간의 언어 전승
(2) 화자의 절대 수
(3) 전체 인구 대비 해당 언어 구사자 비율
(4) 언어 사용 분야의 변화
(5) 새로운 영역, 매체에(해당 언어를 사용하여) 반응하기
(6) 언어 교육과 읽고-쓰기 자료 사용 여부
(7) (해당 언어에 대한) 공식적인 지위 부여/사용을 포함한 정부와 해당 기관의 언어에 대한 태도와 정책
(8) 자신의 언어에 대한 해당 그 지역 사회 구성원의 태도
(9) (해당 언어) 자료의 양과 질

[표 1] 세대 간 언어의 계승 정도

소멸 위기 정도	층위(단계)	언어의 세대 간 계승
안전	-	모든 세대가 해당 언어를 사용하고 있음. 세대 간 계승이 계속됨
취약	1	대부분의 아이가 해당 언어를 사용하고 있음. 그러나 이 언어는 한정된 영역에서 사용됨(보기: 집)
위험	2	아이들이 해당 언어를 집에서 모국어로 배우지 않음. 특히 부모 세대와 그 이전 세대가 사용함
매우 위험	3	조부모들이 해당 언어를 사용함. 부모 세대는 이 언어를 이해할 수 있지만, 그들끼리 또는 아이들과 이 언어를 사용하지 않음
소멸 직전	4	조부모 이상의 세대가 해당 언어의 가장 젊은 사용자임 그들은 단지 부분적으로만 드물게 이 언어를 사용함
소멸	5	해당 언어의 사용자가 더 존재하지 않음

© 유네스코 한국위원회

*26 참고 | 니컬러스 에번스 지음, 김기혁, 호정은 옮김, 《아무도 모르는 사이에 죽다》, 글항아리
*27 참고 | 소멸위기 언어 보존 사례 분석을 통한 제주어 보전 방안, http://www.jst.re.kr/report.do?curPage=4
*28 참고 | http://www.unesco.org/new/en/culture/themes/endangered-languages/faq-on-endangered-languages/
*29 참고 | 시한부 생명 제주어 향후 10년에 달렸다, http://news.chosun.com/site/data/html_dir/2012/07/15/2012071500489.html

대한민국 표준어가 있음에도 제주어의 소멸을 막는 방안을 검토해야 하는 이유는 무엇일까? 제주어가 언어학적 관점에서 연구 가치가 높기 때문이다. 제주어는 훈민정음의 아래아를 사용하는 것은 물론, 표준어에 없는 때가림소(시제) 'ㅇ, ㄴ'을 활용한 제주어만의 문법을 지니고 있다.[*29] 이에 제주도는 제주어를 보전하고자 지난 2007년 '제주어 보전 및 육성' 조례안을 제정, 제주어를 활성화(revitalization)할 다양한 방법을 모색하고 있다. 녹음, 구술 자료집 편찬, 제주어 교육, 연구 및 사전 편찬과 같은 활동이 바로 여기에 해당한다. 그러나 이는 보존(preservation)이나 유지(sustaining)를 위해 필요한 언어 자료화(documentation)에 더 가깝다고 볼 수 있다.

그렇다면 어떻게 하면 사람들이 제주어를 적극적으로 사용하게 만들 수 있을까? 딥러닝을 활용한 표준어-제주어 기계 번역과 제주어 TTS가 가장 현실적인 대안이 될 수 있을 것이다. 물론 이 과정은 쉽지만은 않을 것으로 예상된다. 먼저 '표준어 문장-제주어 문장'이라는 대역 말뭉치 데이터가 충분하지 않다. 제주어로만 작성된 자료가 충분하다고 하더라도, 이를 표준어로 대역하지 못한 자료가 수두룩하다. 그나마 제주특별자치도가 내놓은 《제주어구술자료집(2017)》를 활용해보는 방안을 고려해볼 수 있다. 여기서 추출한 대역 말뭉치 중 글자 수가 5개 이상인 문장 쌍이 대략 6만 3000개가 된다. 상용 번역 서비스(영한/일한/중한 번역)에 활용된 데이터와 비교하면 그 양이 턱없이 모자라다. 추가로 데이터 세트를 확보하고, 이를 학습에 용이한 형태로 정제하는 데 막대한 시간과 비용이 드는 점도 장벽이다. 아울러 큰 틀에서 보면 표준어 어휘를 제주어 어휘로 일대일로 대체하는 작업인 만큼, 딥러닝보다는 형태소 분석 방식이 더 적합하다는 가능성도 완전히 배제할 수 없다.

〈딥러닝이 탐구하지 못한 언어와 5가지 태스크〉 브런치로 연결되는 QR 코드입니다.