

2017 국어 정책 학술 대회

# 우리말 정보화의 현황과 과제

2017. 11. 10.(금) 10:00~

페럼타워 3층 페럼홀

발 표 자 료 집



국립국어원

# 2017 국어 정책 학술 대회

## 우리말 정보화의 현황과 과제

행사 개요	06
행사 일정	06
개회사	07
축사	08
<b>주제 1 우리말 인공지능의 개발과 전망</b>	
· 발표 이윤근(한국전자통신연구원 자동통역인공지능연구센터장)	11
· 토론 김진해(경희대학교 후마니타스 칼리지 교수)	39
<b>주제 2 말뭉치 구축의 세계 동향과 국어 말뭉치의 현주소</b>	
· 발표 김한샘(연세대학교 언어정보연구원 HK 교수)	43
· 토론 홍혜진(국립국어원 언어정보과 학예연구관)	67
<b>주제 3 말뭉치 언어학과 이론 언어학,사전 편찬</b>	
· 발표 송상헌(인천대학교 영어영문학과 교수)	71
· 토론 최정도(국립국어원 언어정보과 학예연구사)	85
<b>주제 4 기계 번역은 우리 생활을 어떻게 변화시킬 것인가?</b>	
· 발표 김준석(네이버 파파고 팀 리더)	89
· 토론 정호정(한국외국어대학교 영어통번역학부 교수)	109
<b>주제 5 우리말 자연 언어 처리 기술의 전망</b>	
· 발표 나승훈(전북대학교 컴퓨터공학과 교수)	113
· 토론 차정원(창원대학교 컴퓨터공학과 교수)	141
<b>주제 6 음성 언어 처리, 어디까지 왔나?</b>	
· 발표 이경님(엔씨소프트 에이아이 센터 스피치 랩 음성인식팀)	147
· 토론 정민화(서울대학교 언어학과 교수)	159

## 행사 개요

**일시** 2017. 11. 10.(금) 10:00 ~  
**장소** 페럼타워 3층 페럼홀  
**주제** 우리말 정보화의 현황과 과제  
**주최** 국립국어원



## 행사 일정

10:00~10:20	개회식
10:20~11:10	주제 1 우리말 인공지능의 개발과 전망 • 발표자 <b>이윤근</b> (ETRI)      • 토론자 <b>김진해</b> (경희대)
11:10~12:00	주제 2 말뭉치 구축의 세계 동향과 국어 말뭉치의 현주소 • 발표자 <b>김한샘</b> (연세대)      • 토론자 <b>홍혜진</b> (국립국어원)
12:00~13:30	점심 식사
13:30~14:20	주제 3 말뭉치 언어학과 이론 언어학, 사전 편찬 • 발표자 <b>송상현</b> (인천대)      • 토론자 <b>최정도</b> (국립국어원)
14:20~15:10	주제 4 기계 번역은 우리 생활을 어떻게 변화시킬 것인가? • 발표자 <b>김준석</b> (네이버)      • 토론자 <b>정호정</b> (한국외대)
15:10~15:20	휴식
15:20~16:10	주제 5 우리말 자연 언어 처리 기술의 전망 • 발표자 <b>나승훈</b> (전북대)      • 토론자 <b>차정원</b> (창원대)
16:10~17:00	주제 6 음성 언어 처리, 어디까지 왔나? • 발표자 <b>이경남</b> (엔씨소프트)      • 토론자 <b>정민화</b> (서울대)
17:00~	폐회식

## 개회사



송철의  
국립국어원장

안녕하십니까? 국립국어원장 송철의입니다.

먼저 귀한 시간을 내셔서 이 자리에 참석해 주신 강길부 의원님과 이우성 문화예술정책실장님께 감사의 말씀을 올립니다. 더불어 추운 날씨에도 4차 산업 혁명에 대한 관심으로 이 자리를 찾아 주신 내빈 여러분들께도 감사드립니다.

얼마 전 입동이 지나 본격적으로 겨울이 시작되었습니다. 겨울에는 모든 살아 움직이는 것들이 활동을 멈추어 마치 죽은 것처럼 보입니다. 그러나 이듬해 봄에는 여지없이 깨어나 다시 생명을 움을 틔우고 합니다. 따라서 겨울은 생동이 사라지는 시기라기보다는 생동을 위해 준비하는 시기라고 할 수 있습니다.

4차 산업 혁명 시기에 진입한 지금도, 오늘과 같은 겨울인 듯합니다. 앞으로 다가올 4차 산업 혁명의 봄에 활짝 꽃 피우기 위해 부지런히 기반을 닦는 시기라고 생각합니다. 4차 산업 혁명의 찬란한 봄날을 즐기기 위해서는 겨우내 부단한 노력이 필요합니다. 오늘 이 자리는 그러한 목적으로 마련되었습니다.

지금 이 자리에는 국립국어원에서 4차 산업 혁명 시기에 무엇을 준비할 수 있으며, 4차 산업 혁명을 주제로 학술대회를 개최하는 까닭이 무엇인지 궁금하신 분들이 계실 듯합니다. 4차 산업 혁명은 컴퓨터공학만이 주도하는 것이라고 생각하는 것이 보통의 상식이기 때문입니다. 그러나 4차 산업 혁명의 내면을 들여다 보면, 그 가운데에는 인공지능이 자리하고 있다는 것을 알게 됩니다. 인공지능이 사람처럼, 혹은 사람을 뛰어 넘는 능력을 발휘하기 위해서는 매개가 되는 언어가 필수적입니다.

아이가 언어를 배우려면 부모의 언어를 풍부하게 들음으로써 언어의 구조를 추론할 수 있어야 합니다. '풍부한 부모의 언어'가 곧 기초 자료인 셈인데, 한국어를 말하는 인공지능은 이 '풍부한 부모의 언어'가 없습니다. 그간 4차 산업 관련 학계와 업계에서는 이 한국어 기초 자료의 부재를 지속적으로 성토해 왔습니다. 한국어를 자유롭게 구사할 수 없다면 인공지능의 발전은 한계에 맞닥뜨리게 된다는 점에서 매우 중요한 문제라고 할 수 있습니다. 그러나 이 자료를 개인이 구축하기에는 시간과 노력, 경제적 비용이 적지 않습니다. 따라서 이 국어 거대 자료는 국가 주도로 구축하여 공공재로 사용하는 것이 바람직합니다. 여기에 국립국어원의 역할이 있는 것입니다.

국어 거대 자료(빅데이터)의 구축은 학계와 업계의 요구에 부응하기 위한 것이므로 국립국어원이 단독적으로 진행하기는 어렵습니다. 따라서 학계의 첨단 이론 현황과 업계의 현장 적용 현황을 파악하고 국어 거대 자료 구축의 방향과 구체적인 내용을 가능하고자 자리를 준비했습니다. 이 자리가 국어 거대 자료 구축 사업의 초석이 되기를 바라며, 더불어 4차 산업 혁명 관련 학계와 업계, 또 국어학계와 컴퓨터공학계가 보다 안정적인 인공지능 탄생을 위해 힘을 모으는 계기가 되기를 바랍니다.

아직 완성되지 않았다는 것은 상상 가능성이 무한하다는 의미입니다. 오늘의 이 자리가 부디 4차 산업 혁명 시대에 무한한 상상의 장을 펼치는 기회가 되기가 되기를 바랍니다. 고맙습니다.



강길부  
국회의원

안녕하십니까?

울산 울주 출신의 국회 교육문화체육관광위원회 소속 강길부 국회의원입니다.

4차 산업혁명 시대를 맞이하여 '우리말 정보화의 현황과 과제'를 주제로 열리는 학술대회 개최를 진심으로 축하드립니다. 아울러 오늘 학술 대회를 마련해 주신 송철의 국립국어원장님과 관계자 여러분의 노고에도 감사의 뜻을 전합니다.

언어 정보화 산업이 우리 실생활 곳곳에 빠르게 스며들고 있습니다. 구글이나 네이버에서 만든 기계 번역기가 관광이나 비즈니스 등에서 널리 활용되고 있고, 인공지능을 기반으로 다양한 정보 제공과 각종 기능을 수행하는 AI 스피커도 빠르게 보급되고 있습니다.

여기에 발맞추어 과학기술정보통신부에서는 2013년부터 세계 최고 수준의 인공지능 기술 선도를 위한 엑소브레인 사업을 진행해 인간과 퀴즈 대결을 펼쳐 승리하며 1단계 개발을 성공적으로 마무리하고 법률, 특허, 금융 등의 분야에서 2단계 사업화가 진행되고 있습니다.

그러나 이를 뒷받침하기 위한 우리말 정보화 수준은 매우 미흡한 수준입니다. 인공지능의 필수요소인 한국어 말뭉치 구축 사업조차 중요성 인식 부족으로 인해 2007년 1차 사업 종료 후 중단되는 바람에 후속주자였던 일본보다도 뒤처지고 있는 상황입니다.

마침 '우리말 정보화의 현황과 과제'에 대해 논의하는 학술대회가 열리게 된 것을 매우 뜻깊게 생각합니다. 문화체육관광부에서도 언어자원의 중요성을 인식하고 2018년부터 5년간 175억 원을 투입하여 국립국어원 주관으로 155억 어절에 이르는 대규모 한국어 말뭉치 사업을 통해 국가 언어자원 통합 체계를 구축하겠다고 밝혔습니다.

진정한 국가 언어자원 구축을 위해서는 정부, 학계, 민간 등 언어자원 구축에서부터 활용에 이르기까지 다양한 분야의 전문가가 모두 모여 머리를 맞대고 체계적인 연구와 활용방안이 마련되어야 합니다.

국회에서도 한글의 과학적 우수성을 알리고, 우리말 정보화 발전을 위해 더 큰 관심을 갖고 지켜보겠습니다.

오늘 행사에 참석해 주신 모든 분들의 가정에 평안과 건강이 함께하시길 기원합니다. 고맙습니다.



이우성

문화체육관광부  
문화예술정책실장

안녕하십니까? 문화체육관광부 문화예술정책실장 이우성입니다.

생활에서 4차 산업 시대가 성큼 다가움을 느끼는 시기에 즈음하여 국립국어원에서 4차 산업 발전의 기반을 마련하고자 준비한 이 자리가 한없이 반갑습니다. 그리고 오늘 국어학계, 인공지능 관련 학계, 관련 업계의 협력의 장을 성공적으로 마련한 것을 축하드립니다.

보통 4차 산업 시대의 역군을 이공계열인 인공지능 관련 학계로만 생각하기 쉽습니다. 그러나 그 안을 찬찬히 들여다보면 언어의 문제가 핵심적인 자리를 차지하고 있음을 알 수 있습니다. 말과 글이 의사와 명령을 전달하는 매개가 되기 때문에 사물 인터넷을 효율적으로 이용하고 이용자에게 질 높은 편의를 제공하려면 우리말의 정보화가 필수적입니다. 따라서 인공지능의 고도화는 우리말 정보화의 고도화 정도에 달려 있다고 해도 과언이 아닙니다.

따라서 새로운 산업 혁명 시대를 온전히 바르게 선도하기 위해서는 국어학계, 인공지능 관련 학계, 관련 업계의 협력을 통한 발전이 매우 중요합니다. 오늘 이 자리가 세 분야의 전문가들을 모시고 통섭의 장으로 꾸려진 데에는 이러한 목적이 있었을 것으로 생각합니다.

정보 혁명 시대라고도 불리는 지난 3차 산업 혁명 시대에 우리나라는 국가적인 지원과 국민들의 뜨거운 열정이 어우러져 눈부시게 빛나는 성장을 이룩하였고 세계에 대한민국의 이름을 드날릴 수 있었습니다. 그러한 영광을 4차 산업 혁명 시대에도 이어가기 위해서는 무엇보다 철저한 준비와 관련 기관들의 협조가 필요합니다.

오늘 이 자리가 그러한 초석을 다지는 자리가 되기를 바라며, 바쁘신 가운데에도 우리말 정보화에 대한 깊은 관심으로 참석해 주신 내빈 여러분과 발표와 토론을 맡아 주신 선생님들, 그리고 이 자리를 함께해 주신 모든 분들께 감사의 말씀을 드립니다.



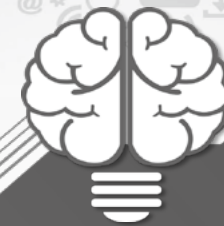
주제 1

# 우리말 인공지능의 개발과 전망

· 발표자 **이윤근**(한국전자통신연구원 자동통역인공지능연구센터장)



# 우리말 인공지능의 개발과 전망



2017. 11. 10.

이윤근 (ETRI)

## 목차

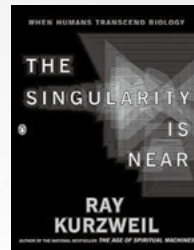
- I 배경
- II 음성 인식
- III 음성 합성
- IV 자동 통역
- V 대화 처리
- VI 질의응답
- VII 자연어 대화 인터페이스 및 음성 인식 개인 비서
- VIII 맺음말

## 인공지능: IT 60년 역사에서 가장 혁신적 기술로 전망 (Gartner, Tractica)

- **AI 정의:** Science and engineering of making intelligent machines
  - John McCarthy, 1956
- **특이점 전망:** 전문가에 따라 의견이 다르나, 30년 ~ 1000년 이내로 전망
  - BBC(2013)는 2045년 도래 가능성 예측(odds 8/1)
  - 전문가 170명의 통계: 인간 지능의 2022년 10%, 2040년 50%, 2075년 90% 가능성
    - \* 특이점(Singularity): AI가 인간의 지능을 넘어서는 시점

## 특이점이 온다: The Singularity is Near

- 대표적 미래학자 중 한명인 Ray Kurzweil이 2006년에 발표
  - 2045년에 '특이점'이 도래한다고 예측
- 기술의 발전 속도는 무어의 법칙 이상의 기하급수적 성장을 이루어 왔으며 앞으로도 그럴 것으로 예상
- 중첩되어 일어날 주요 3가지 혁명: GNR
  - Genetics (유전학), Nanotechnology (나노기술), Robotics (로봇공학: 인공지능)



- ※ 예로 제시한 대부분은 물리량이나 하드웨어와 같은 양적 팽창의 사례임 (SW와 같은 질적 발전으로 변환할 의 여부는 불투명)
- ※ GNR에 대한 중첩혁명의 예측은 시사점이 크며, 미래의 변화를 현재의 한 분야 기술만의 한계를 토대로 예상해서는 안된다는 점은 중요함

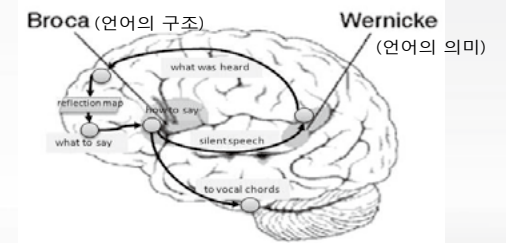
3/41

## 지능의 종류 및 언어 지능



(출처: Howard Gardner의 다중지능)

- 언어는 인간을 다른 동물과 구별하는 가장 큰 특징 중 하나
- 언어의 특성
  - 분절성: 연속적인 세계를 끊어서 표현
  - 개방성: 무한한 표현이 가능 (=창조성)
  - 추상성: 개념화, 일반화 (예) 장미, 백합, 튜립 → 잎, 줄기, 향기가 있음 → 꽃
  - 체계적: 규칙과 구조를 가짐 (예: 문법구조)
- 언어를 담당하는 뇌의 영역



※ 베르니케 영역은 인간이 침팬지의 7배 크기이며, 의미의 이해와 언어의 의미론적인 측면을 해독

6/37

## 인공지능의 개발 동향 및 수준

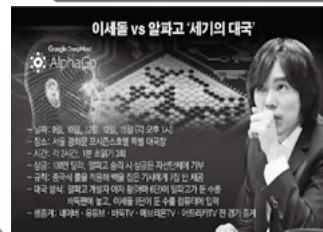
## 개발 동향



## IBM, 구글, MS 등 글로벌 기업은 경쟁적 개발 시작

기업	로고	내용
IBM	IBM	'왓슨' 슈퍼컴퓨터 중심으로 생태계 조성
구글	Google	무인자동차, 얼굴과 음성 인식 등 인공지능 적용 '딥마인드' 등 각종 업체 인수합병 활발
MS	Microsoft	음성 인식 갖춘 개인비서 '코타나' 올해 출시
페이스북	facebook	인공지능 연구그룹 구성, 얼굴 인식 프로그램 발표
애플	Apple	'시리' 업그레이드 버전으로 지능형 개인비서 개발

## 기술 수준



- 딥러닝 기술 적용으로 음성과 이미지 인식 성능 향상
  - 얼굴인식률: 구글 '페이스넷(FaceNet)' 99.96%
- 세기의 바둑 대국에서 알파고 승리: 알파고 vs. 이세돌 9단
  - 인공지능망 강화학습으로 이세돌 9단에 승리
- Weak AI 위주로 2020년부터 시장 형성(Tractica, 2015)
  - 인간의 지식노동을 보조하는 분야로부터 인공지능 시장의 개화

4/37

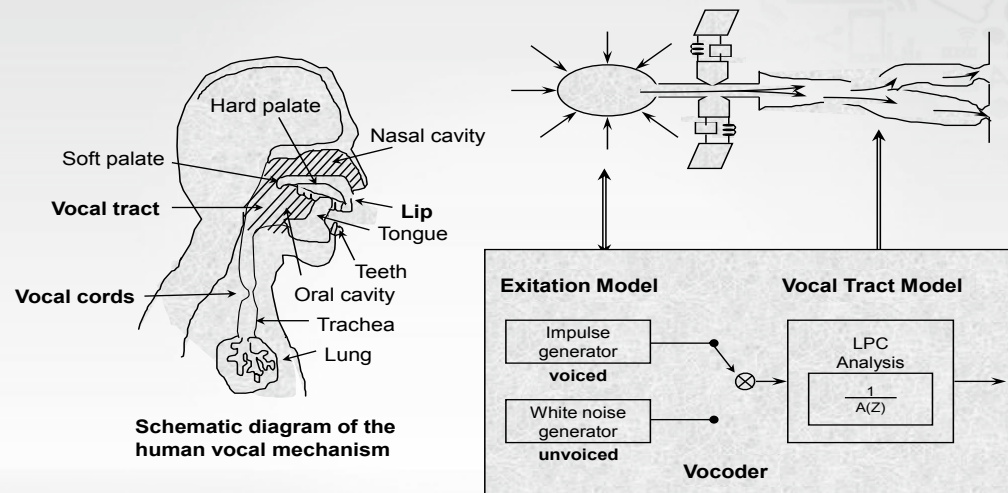
## 언어 지능의 개요

- 인간의 가장 중요한 정보전달, 축적 및 의사소통 수단인 '언어'를 다루는 기술
  - HCI(Human Computer Interaction)의 핵심기술
  - 지식 및 정보 서비스의 기반 기술
  - 민족 고유의 문화 산업을 발전시키기 위한 기반 기술
- 주요 기술: 음성인식, 음성합성, 대화처리, 자연어처리, 질의응답 등



7/37

## 음성 발생 원리 및 모델

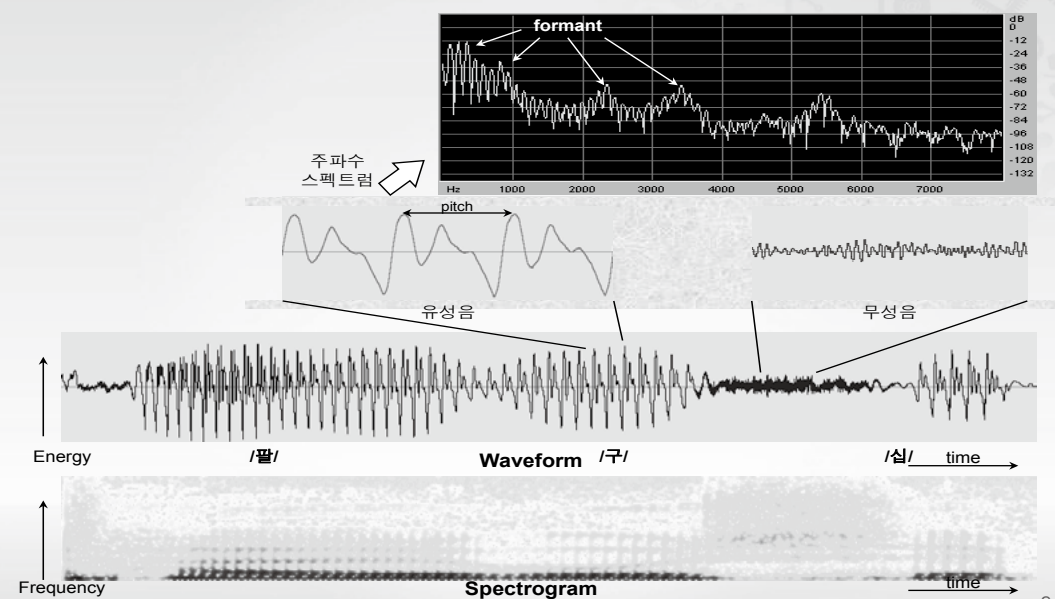


8/37

## 목차

- I 배경
- II 음성 인식
- III 음성 합성
- IV 자동 통역
- V 대화 처리
- VI 질의응답
- VII 자연어 대화 인터페이스 및 음성 인식 개인 비서
- VIII 맷음말

## 음성 신호의 특징



9/37

## 음성 인식 기술의 개요

- 음성 인식 기술: 인간의 말을 문자로 자동 변환하는 기술
- 음성인식의 어려움
  - 동일한 화자인 경우에도 다양한 변이: 음의 높낮이, 발성 속도, 주변 잡음의 영향
  - 동일한 단어라도 화자별로 발성이 다름: 음색, 발음, 강세 등
  - 문맥에 따라 발성이 달라짐: 음운 변화 (자음점변, 구개음화, 경음화 등)

## ■ 음성 인식 기술의 세대 분류



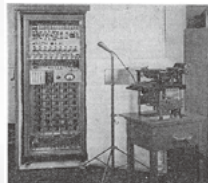
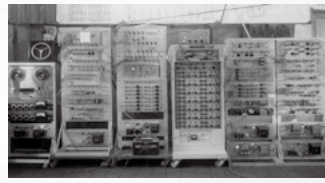
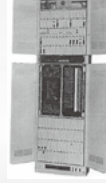
음성인식

11/37



## 음성 인식 기술 1세대 (1952-1968)

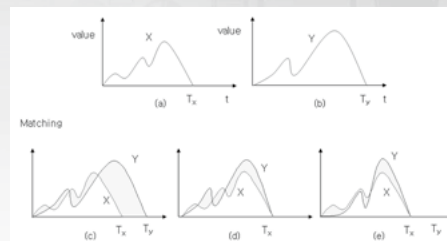
- General
  - The earliest attempt to devise digit/syllable/vowel/phoneme recognition systems
  - Spectral resonances extracted by an analogue filter bank and logic circuits
- Early systems
  - Bell Labs, RCA Labs, MIT Lincoln Labs
  - University College London
  - Radio Research Labs, Kyoto University, NEC Labs

Photonic typewriter  
(RCA Labs, 1956)Automatic speech typewriter  
(University College London, 1959)Spoken digit recognizer  
(Radio Research Labs, 1961)Spoken digit recognizer  
(NEC Labs, 1963)

12/37

## 음성 인식 기술 2세대 (1968-1980)

- DTW (Dynamic Time Warping)
  - Non-linear matching
- Isolated word recognition
  - became a viable and usable technology based on fundamental studies in Russia and Japan
- IBM Labs: large-vocabulary ASR
- Bell Labs: speaker-independent ASR
- Pioneering research on continuous speech recognition (DARPA)
  - Goal: 1000 word ASR, a few speakers, continuous speech, constrained grammar
  - Hearsay I & II systems at CMU, Harpy system at CMU, HWIM system at BBN



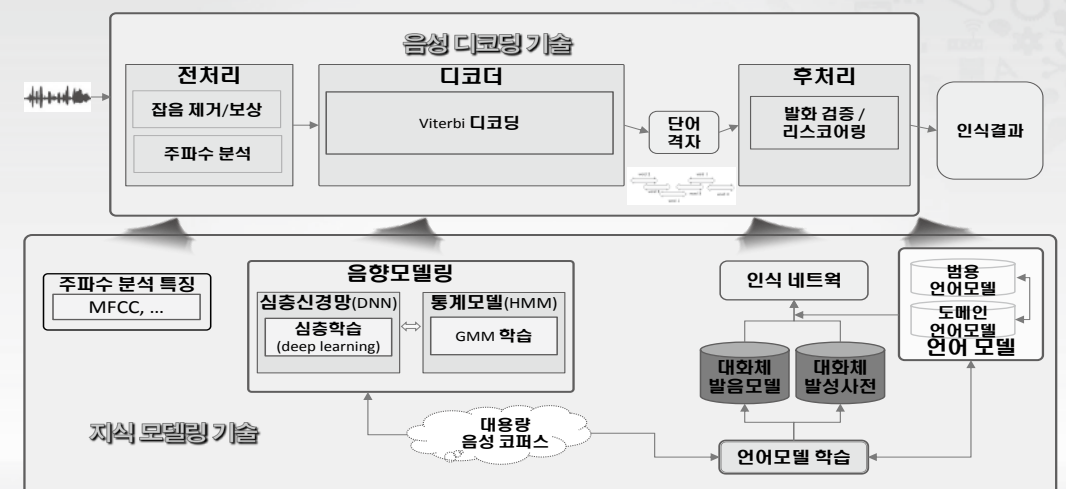
## 음성 인식 기술 3세대 (1980-2006)

- Connected word recognition
  - Two-level DP, level-building method, one-pass method, frame-synchronous DP
- Statistical framework
  - HMM (Hidden Markov Model)
  - Viterbi decoding
- Cepstrum feature
- N-gram
- DARPA program (Resource management task)
  - SPHINK system at CMU
  - BYBLOS system at BBN
  - DECIPHER system at SRI

14/37

## 음성 인식 기술 3세대 (1980-2006) (계속)

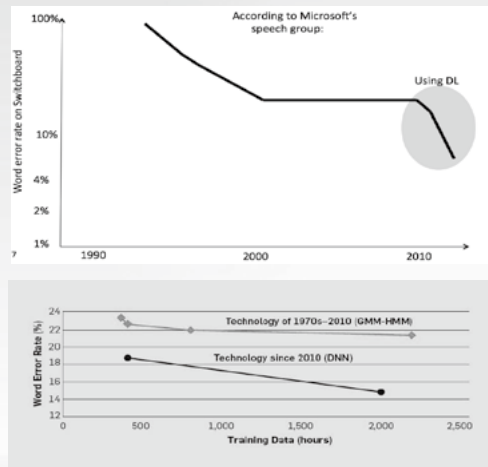
## 음성 인식 시스템 구성도



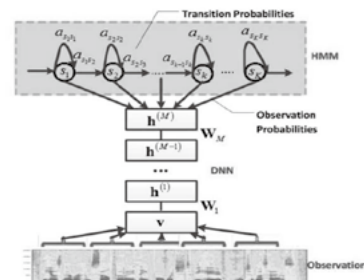


## 음성 인식 기술 4세대 (2006~)

- 음성 인식에 deep neural net을 적용하면서 성능이 대폭 개선됨



Architecture of a DNN-HMM hybrid system



17/37

## 음성 인식 기술의 특징 및 발전 동향

- 클라우드 컴퓨팅 인프라의 발달로 인하여 빅데이터에 기반한 음향/언어 모델 진화 의해 음성인식 성능이 비약적으로 발전 (대표적인 사례: '구글')
- 음성언어 기술 발전을 위해 방대한 분량의 음성언어 DB 구축/처리 기술 필요
- 장기간에 걸친 음성언어 DB 인프라 구축이 필요하며, 음성언어 관련 서비스를 통한 사용자 로그 정보 축적이 기술 발전을 위해 매우 중요함
- 특정언어 중심의 음성언어 서비스가 활성화될 경우, 음성언어 로그 축적의 불균형을 가져와 장기적으로 특정언어의 기술만 발전하는 불균형 현상을 심화시킴

19/37

## 잡음 처리

- 잡음이나 음향간섭 등에 의해 음성입력이 왜곡되어 음성 인식 성능이 저하됨. 이를 개선하기 위한 방법론 필요
- 음성왜곡의 종류
  - 차량잡음 등 외부잡음에 의한 성능저하
  - 실내 반향(room reverberation)에 의한 성능저하
  - 여러 사람이 동시에 발성
  - 마이크로폰, 앰프 등 음성입력채널의 왜곡
  - etc...
- 해결방법
  - 잡음필터링 (Kalman filter, Wiener filter 등), 채널 보상(channel compensation), 음원분리(source separation), 빔포밍 (beam forming) 등 다양한 잡음처리 방법 적용
  - 다양한 왜곡신호를 음향모델에 포함하여 훈련시키는 방법 적용 (multi-condition training)

음성인식

## 음성 인식 상용화 사례



20/37

## 목차

- I 배경
- II 음성 인식
- III 음성 합성
- IV 자동 통역
- V 대화 처리
- VI 질의응답
- VII 자연어 대화 인터페이스 및 음성인식 개인비서
- VIII 맺음말

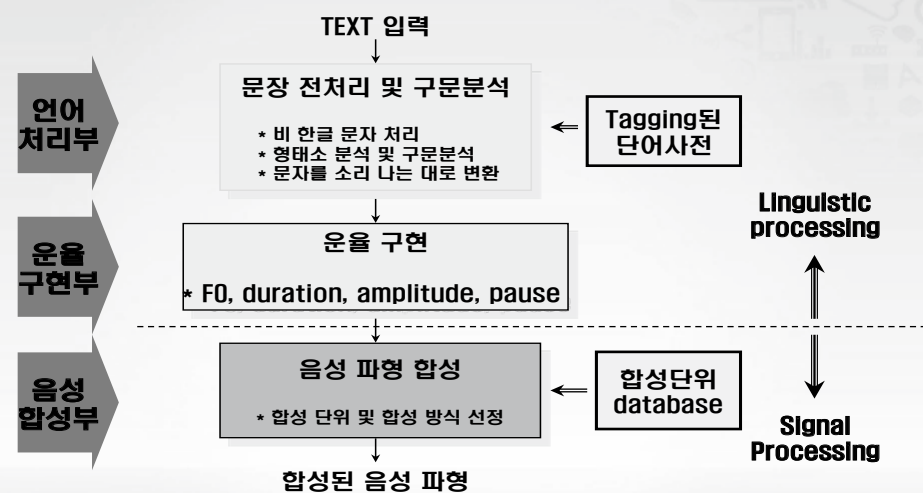
## 문장전처리 및 구문분석

- 비 한글 문자 처리
  - 숫자: 123.45 → 백이십삼점사오
  - 특수 기호: % → 퍼센트, & → 앤드
  - 영어 및 약자: 사전 및 오토메타 이용
- 문자를 소리 나는 대로 변환
  - 변환 Table 이용
  - 불규칙 변환인 경우 형태소분석기 및 사전 이용
  - (예) 학교에 다녀와서 밥을 먹는다. -> 학교에 다녀와서 바블 멍는다.
- 구문 분석
  - 운율 구현을 위한 형태소 분석 및 구문 분석

23/37

## 음성합성기술 개요

음성합성기술: 문자(또는 기호)를 인간의 말로 자동 변환하는 기술



22/37

## 운율 구현

- 운율이란?
  - 발생 시 나타나는 억양, 강세, 리듬 등의 특성을 말하며, 이는 기본 주파수 궤적(Intonation), 음소의 지속시간(duration), 음량(amplitude), 휴지구간 길이(pause length) 등에 의해 결정된다.
- 운율에 영향을 미치는 요소는?
  - 말을 구성하는 음소들의 특성
  - 문장의 계층적 구조
    - 예 : [(인간의:AP) (마음을:AP) (읽는:AP):IP] [(감성 컴퓨터 가:AP) (개발됐다:AP):IP].

※ AP : Accent Phrase, IP : Intonation Phrase

24/37

## 합성 방식

음성합성

## ■ 1세대: 고정 합성 단위 설계

- Fixed Length Unit
  - 단어(word), 음절(syllable), 음소(phoneme)
  - demi-syllable, diphone, triphone, CV(Consonant-Vowel), VC, VCV, CVC, VCCV, etc.
- Formant, LPC 정보 등을 이용한 파형 생성 (소용량 DB를 이용한 음성합성이 가능, 음성품질이 나쁨)

## ■ 2세대: 가변 합성 단위 연결 방식

- Corpus based TTS
- 음성파형을 최소한의 가공을 통해 연결 (대용량 DB를 이용한 고품질 음성생성 가능)

25/37

## 합성 방식 (계속)

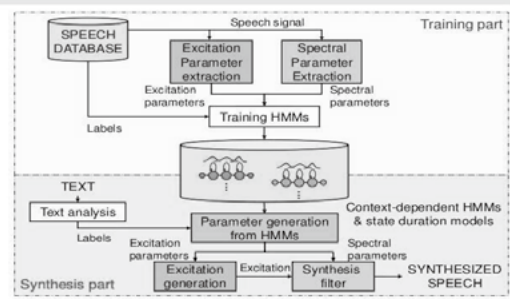
음성합성

## ■ 3세대: HTS (HMM based TTS system)

- 음성인식을 위한 음향모델링에 주로 사용하는 HMM 방식을 음성합성에 적용
- HMM을 이용하여 스펙트럼정보, 여기신호정보, 음성지속시간 등을 동시에 모델링하여 context dependent HMM을 생성하며, 이를 이용하여 음성을 합성
- 적절한 크기의 데이터베이스를 이용한 고품질 음성합성이 가능 (1, 2세대 방식의 장점 결함)

※ 최근 들어 Deep NN을 이용한 음성 합성 방식으로 진화 중 (예: 딥마이드의 wavenet)

HMM-based speech synthesis system



※ 출처 : Nagoya Institute of Technology

21

## 음성 합성 기술의 난제

음성합성

## ■ Emotional TTS

- 합성음에 감정을 구현하는 기술
- 음의 높낮이, 세기, 음색 등의 변화가 심해 음질이 저하되며 제어하기가 어려움

## ■ Voice Conversion

- 특정인의 목소리로 변환하는 기술
- 특정인의 목소리를 나타내기 위해서는 음색 뿐만 아니라 발음, 억양 등 복합적인 요소가 작용

## 목차

- I 배경
- II 음성 인식
- III 음성 합성
- IV 자동 통역
- V 대화 처리
- VI 질의응답
- VII 자연어 대화인터페이스 및 음성인식 개인비서
- VIII 맺음말



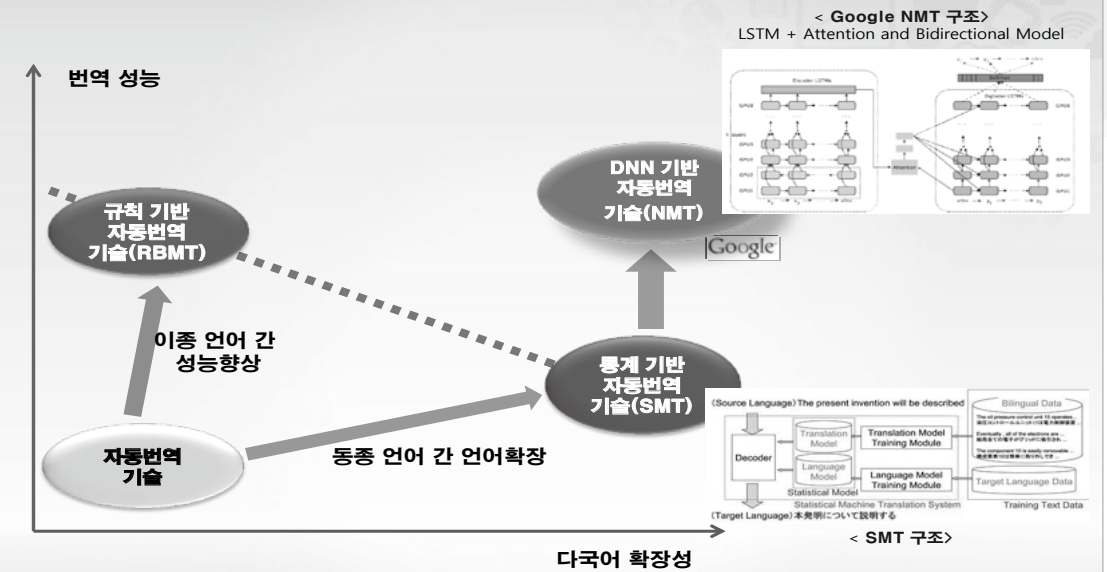
## 자동통번역기술 정의

한 언어의 말 또는 텍스트를 다른 언어의 말 또는 텍스트로 자동으로 통역 또는 번역하는 기술



29/37

## 자동통번역기술의 발전



31/37

## 주요 연구 동향

Google, MS 등 글로벌 기업, 중국의 바이두, 일본의 NTT, 국내의 시스틀인터내셔널을 중심으로 다국어 자동통번역기술 경쟁이 치열함

- [Google] 통계/신경망 기반 자동번역 기술에 기반하여 한국어를 포함한 100개 국어 이상의 다국어에 대한 자동통번역 서비스를 실시함
- [MS] 화상 전화 통역 등 영어를 중심으로 연속 발화에 대한 실시간 자동통역 서비스를 시작함
- [바이두] 광둥어, 고대 중국어, 중국어 변체를 포함하여 총 26개 언어 간 웹서비스 및 모바일 자동통번역 서비스를 실시함
- [NTT] 국가적으로 지원을 받아 NICT와 함께 2020년 동경올림픽을 목표로 본격적으로 다국어 자동통번역기술 개발에 착수함
- [시스틀인터내셔널] 한중일영 자동통번역 및 130여 개 언어의 다국어 문서 자동번역 솔루션을 확보함

30/37

## 국내 자동통역 서비스(지니톡) 개발 현황

2016년 현재 한-중/일/영/스/프 자동통역 기술 개발

- ※ 2015년 5월까지 대국민 시범서비스 운용 (215만 다운로드 달성)
- ※ 기업체 기술이전을 통한 상용 서비스 실시 (2016. 8., 한컴인터프리)
- ※ 신경망 기반 자동번역(NMT) 기술 적용(2017. 2.)
- ※ 넥밴드 이어셋 웨어러블 자동통역 기술 개발(2017 MWC)



~ 2020년

실시간 동시통역 기술로 진화

- ※ 강연, 회의 등 실시간 동시통역 서비스 제공

32/37

## 자동통번의



# 목차

I 배경

II 음성 인식

III 음성 합성

IV 자동 통역

V 대화 처리

VI 질의응답

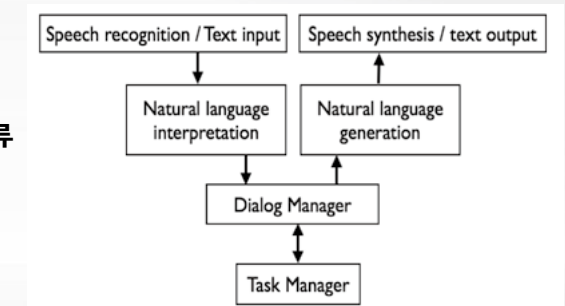
VII 자연어 대화 인터페이스 및 음성인식 개인비서

VIII 맺음말

- I 배경
  - II 음성 인식
  - III 음성 합성
  - IV 자동 통역
  - V 대화 처리
  - VI 질의응답
  - VII 자연어 대화 인터페이스 및 음성인식 개인비서
  - VIII 맺음말

## 대화처리

- 대화 주도(Initiative)에 따른 분류
  - system-initiative
  - user-initiative
  - mixed-initiative



## 대화처리

- 

### Finite state machine for a dialog model for negotiating appointments



## 대화처리 방법론 (2/5)

## ■ Form-based model

- frame-based, form-filling
- Gather task-essential information from user by filling slots in a template
- Slot-filling order is flexible
- Users can provide information more than one at a time
- Possible to design more complex systems by combining several frames

Slot	Question	Response
ACTIVITY	What do you want to do?	walk
START-TIME	When do you want to meet?	tomorrow at 8
DURATION	How long do you want to do it?	2 hours

Frame-based dialog model for negotiating appointments

## 대화처리 방법론 (4/5)

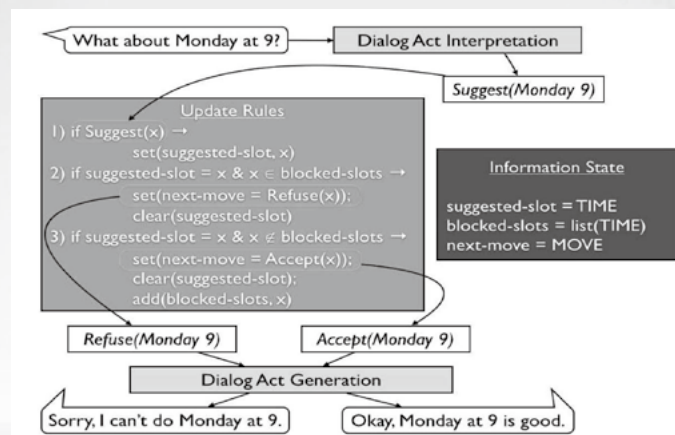
## ■ Topic-oriented vs. Chat-oriented dialog

대화 처리 방법론	정의	특징	주요 차이점과 장단점
Topic-oriented Dialogue (주제 제한, 지식 처리)	- 구체적인 Topic에 대한 대화 처리 (예: 티켓예약, 상품구매/조회 등)	- 시스템 주도형으로 (System-initiative) 대화 흐름 제한 - 특정 Topic에 대한 대화흐름의 수동 지식화	- 혼합 주도형(Mixed-Initiative) 대화관리: 시스템주도 및 사용자 주도 대화 가능, 대화 자유도 증가 - Topic-oriented와 chat-oriented 대화의 협업 ※ Apple Siri 등이 지원함. 기술적 한계로 많은 오류 내포
Chat-oriented Dialogue (주제 무제한, 패턴 처리)	- 목적 없는 대화: chat bot - Data-driven 방식의 예제 매칭 대화 처리	- 문맥유지 없이 단순 반복 응답만 가능 - 대화/응답 패턴의 수동 지식화	- 확장이 용이한 대화모델 구성 (Topic 대화맵 등) - 점층적 대화지식 자동확장 기능

## 대화처리 방법론 (3/5)

## ■ Information State model

- IS에 기반하여 frame-based model을 확장



Information state model for negotiating appointments

## 목차

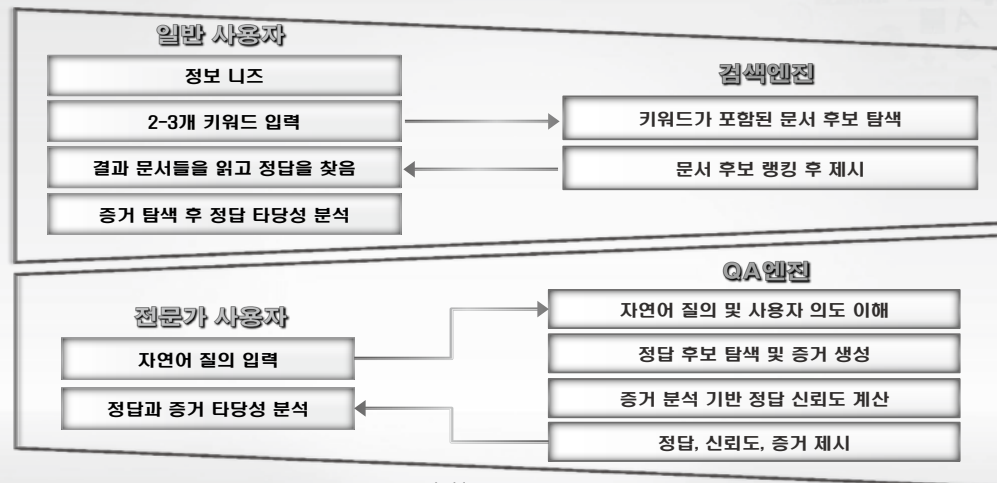
- I 배경
- II 음성 인식
- III 음성 합성
- IV 자동 통역
- V 대화 처리
- VI 질의응답
- VII 자연어 대화 인터페이스 및 음성인식 개인비서
- VIII 맺음말

## 질의응답 시스템 개요

질의응답

## ■ 개요

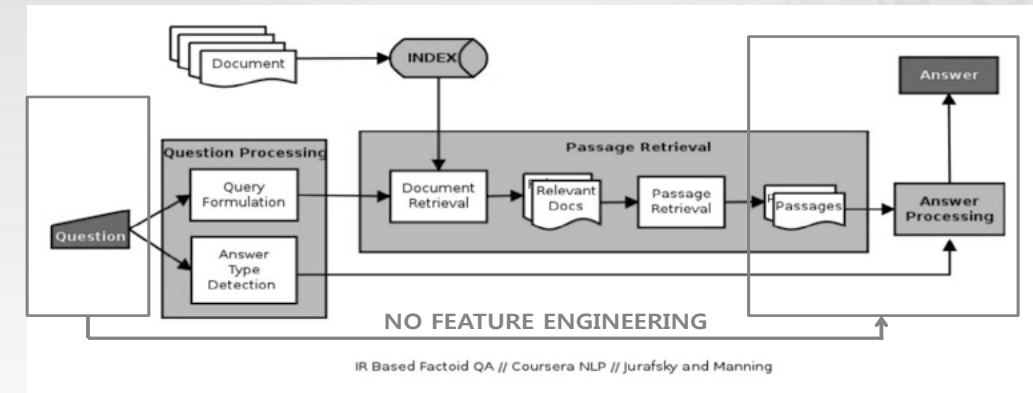
- 사용자의 질의에 대한 답변이 될 수 있는 정답을 문서 집합 내에서 탐색하여 사용자에게 제시해 주는 시스템



&lt;출처: IBM, 2011&gt;

## 딥러닝 기반 질의응답시스템

질의응답

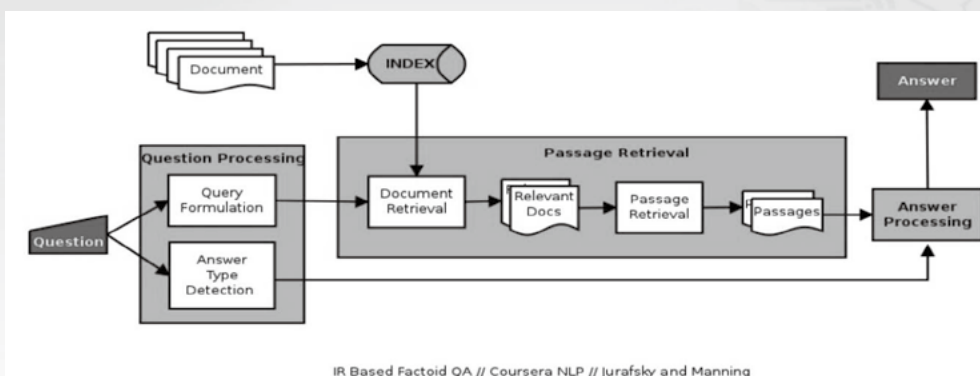


- RNN, GRU, LSTM, memory network 등 딥러닝 기반 모델 사용
- Word embedding: 단어를 2차원 공간의 한 지점으로 매핑하는 기술

※ RNN : Recurrent Neural Net  
 ※ GRU : Gated Recurrent Unit  
 ※ LSTM : Long Short Term Memory

## 질의응답시스템 구성

질의응답



## 질의응답시스템의 대표적 사례

질의응답

- Waston Deep QA 기술개발로 퀴즈쇼에서 인간을 능가(2011)**  
 - 퀴즈분야 특화 단답형 응답 한계. 헬스케어 및 파이낸스 분야 확장 중
- 후지쓰와 NII는 2021년 동경대 입시문제를 풀 수 있는 프로젝트 진행**  
 - 토다이 로봇 프로젝트: 슈퍼컴퓨터 활용 10년(2011~2021)간 진행
- 5억 개 Knowledge Graph를 구축하여 짧은 질의에 대한 요약형 정보 제공**  
 - 구조정보 기반 인물, 지역, 사물 정보 위주 제한된 정보 제공
- Apple Siri 서비스로 활용되며, 수치계산형 위주의 질의응답**  
 - 수작업으로 구축된 지식베이스 활용. 연산 및 통계 위주 질의응답

\* NII(National Institute of Informatics): 일본 국립정보학연구소

## 질의응답시스템의 대표적 사례 (엑소브레인)

- 언어를 이해하고 지식을 스스로 학습하여, 인간에게 전문가 수준(의사, 변호사 등)의 의사결정을 지원하는 언어지능 SW
  - 강학퀴즈 <대결! 엑소브레인> 압승(2016.12.), 국산 AI 자주권 확보 가능성 입증



※ 엑소브레인(외뇌, Exobrain): 내 몸 바깥에 있는 인공 두뇌라는 뜻



질의응답

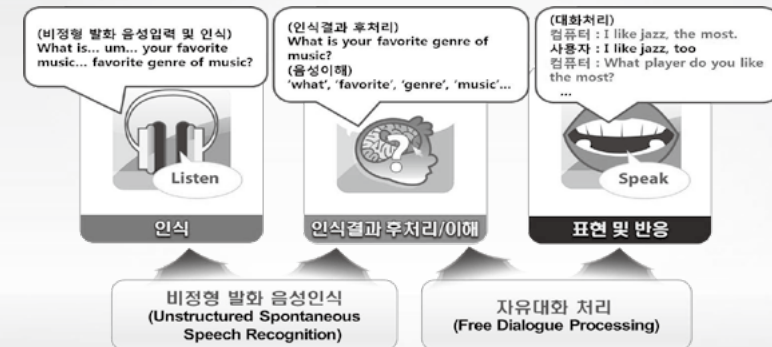
## 자연어 대화 인터페이스 개요

자연어 대화 인터페이스

- 인간의 말을 인식하고 의미를 이해하여, 상황에 맞는 자연스러운 대화를 유도하는 인간-컴퓨터 상호작용 원천기술

“다양한 컴퓨팅 환경에서 음성인식과 같은 natural interface 확산” Microsoft, 빌게이츠

“자연어 이해야말로 인공지능 기술의 핵심” Google Director, 레이 커즈와일



48/37

## 목차

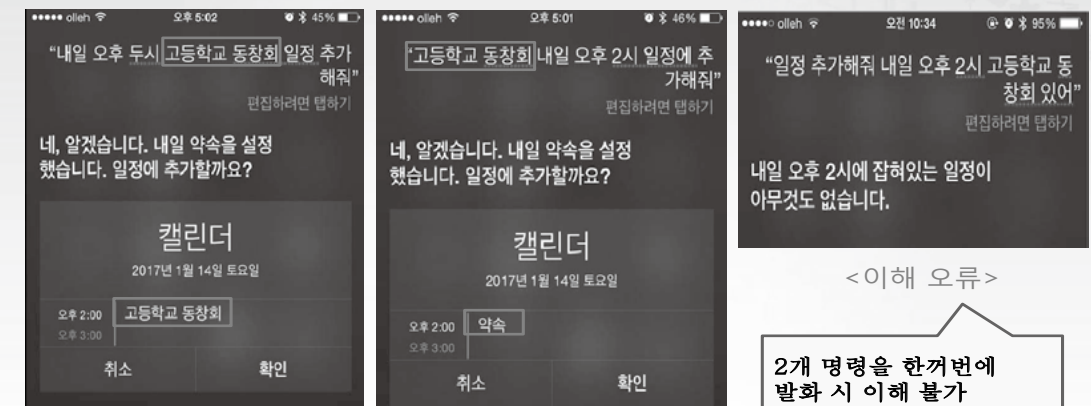
- I 배경
- II 음성 인식
- III 음성 합성
- IV 자동 통역
- V 대화 처리
- VI 질의 응답
- VII 자연어 대화인터페이스 및 음성인식 개인비서
- VIII 맺음말

## 자유 대화 처리 기술의 어려움

자연어 대화 인터페이스

- 자유 대화 이해의 오류 사례 (Apple SIRI)

※ 현재 날짜 2017년 1월 13일



<이해 오류>

2개 명령을 한꺼번에 발화 시 이해 불가

<정상 이해>

<부분 이해>

인지할 시간과 제목 순서 변경 시 이해 불가

50/37

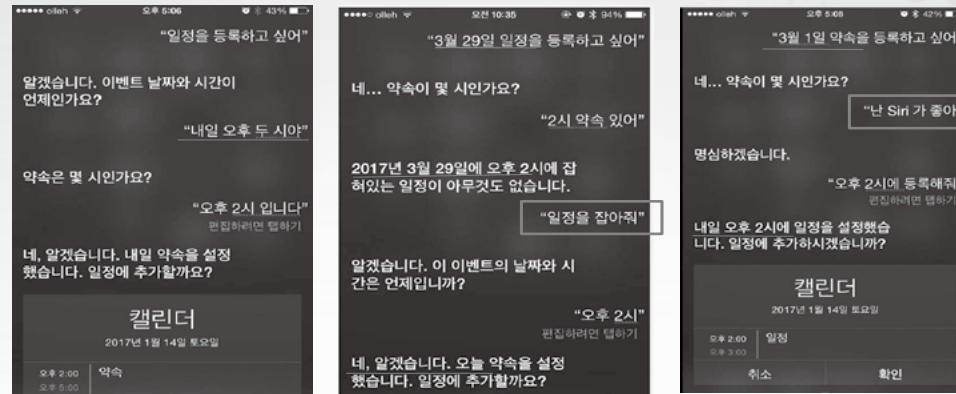


## 자유 대화 처리 기술의 어려움 (계속)

자연어 대화 인터페이스

## ■ 자유 대화 문맥관리의 오류 사례 (Apple SIRI)

※ 현재 날짜 2017년 1월 13일



&lt;정상 문맥&gt;

정해진 대화 시나리오  
진행 (정확한 사용자  
응답 요구)

&lt;부분 문맥 오류&gt;

3월 29일 날짜 문맥을  
유지하지 못함 (오늘  
1월 13일로 오류 발생)

&lt;문맥 오류&gt;

특정 주제의 대화 진행 중, 다른  
주제 대화 삽입 시, 문맥 유지  
오류 발생 (날짜 오류)

51/37

## 음성 인식 개인비서의 발전 전망

자연어 대화 인터페이스

## ■ '인공지능 개인비서'는 Forbes紙의 Top 17 Tech Trends for 2017, Business Insider의 11 Tech Trends for 2017 등에 선정

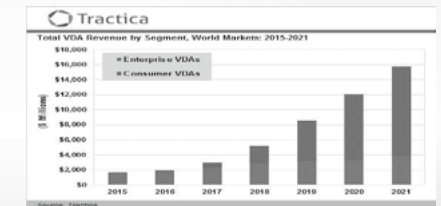
## ■ CES 2017에서 경제학자인 셴 듀브라백(Shawn Dubravac)은 '다섯 가지 트렌드'에 주목해야 한다고 주장

- '음성 인식 기반 기술'(Voice of Computing) : 오는 2020년에는 500만 개 이상의 가정용 음성 제어 로봇이 생산될 것으로 예상. 애플의 '시리', 아마존의 '알렉사', 구글의 '어시스턴트' 등 세계 최고 IT기업들이 앞다투어 다양한 제품 출시. 음성 인식 기술이 스티브 잡스가 개인용 퍼스널 컴퓨터(PC)를 개발한 것과 같이 혁명적 기술 발전이라고 평가
- '인공지능'(Artificial Intelligence) : 최근 인공지능 기술은 빠르게 성장하고 있으며 크기도 소형화. 다양한 가전과 결합, 새로운 차원의 제품 등장 예측.

## ▪ 5세대 이동통신 네트워크(5G) 기술

## ▪ 자율주행 기술

## ▪ 가상현실 기술

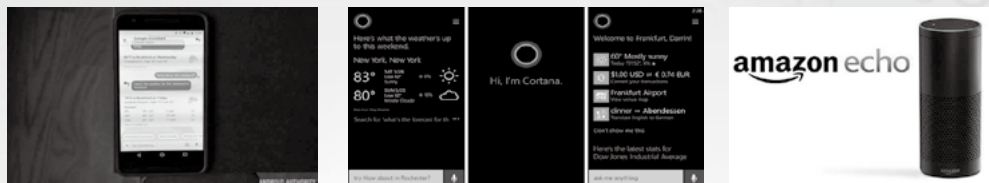


&lt;시장동향 Tractica 2016&gt;

## 음성인식 개인비서 상용제품 사례

자연어 대화 인터페이스

## ■ Apple SIRI, Google Assistant, MS Cortana, Amazon Echo 등



## ■ 삼성 S보이스, SKT NUGU, KT 기가지니



## 음성 인식 개인비서, 정말 장미빛 미래인가?

음성인식 개인비서

## ■ 현재까지 출시되고 있는 서비스의 형상이 모두 유사함

- 간단한 정보서비스(웹검색), 음성명령 등이 가능한 스마트폰 앱
- 스피커 기능을 포함한 가정용 제품

## ■ 소비자를 강력하게 끌어들이는 killer 서비스의 부재

## ■ 버티컬 마켓(서비스)을 주도하기에는 기술적 완성도 미흡

- 인간의 증의적 언어를 정확히 이해하기에는 미흡한 '음성이해기술'
- 다양한 주제에 대한 복잡한 대화를 처리하기에 미흡한 '대화처리기술'
- 전문적 지식을 제공하기에 미흡한 '인공지능 질의응답 기술'

## ➢ 기술적, 산업적 경쟁은 이제부터 시작!

감사합니다.

주제 1  
우리말 인공지능의  
개발과 전망

· 토론자 김진해(경희대학교 후마니타스 칼리지 교수)





본 학술 대회가 ‘우리말 정보화의 현황과 과제’이고, 여기 모이신 분들 중에서도 ‘국어학/언어학’ 연구자들이 많을 것이므로 언어학적 관점에서 궁금한 점 중심으로 질문을 드리겠습니다.

우선 이 발표를 정리하자면 그 동안 한국어 기반의 인공지능 개발의 현황과 전망을 중심으로, 음성인식, 음성합성, 자동통역/번역, 대화처리, 질의응답, 자연어 대화인터페이스와 음성인식 개인비서 등 언어와 관련한 인공지능 기술의 현 수준과 과제를 다루고 있습니다. 이 발표를 통해 우리말 인공지능 개발과 관련한 포괄적인 내용을 일목요연하게 이해할 수 있었습니다.

지능(intelligence)이라고 하는 것이 ‘무언가를 배우는 능력’, 또는 ‘무언가를 생각하고 이해하는 능력’이라고 한다면, 인공지능도 궁극적으로는 학습과 이해, 추리가 가능한 기계일 것이고 여기에서 핵심적인 영역이 언어일 것입니다. 결국 인간의 지적 능력은 언어를 통해 이루어질 것이고, 이를 컴퓨터로 구현/모사할 수 있다면 인간의 지적 능력의 전모를 파악할 수 있는 일이기도 할 것입니다. 이를 위해서는 ‘음성인식/합성(TTS), 형태분석, 구문분석, 중의성 해소, 텍스트 이해, 대화 분석, 의미망, 온톨로지/지식베이스’ 등 각 단위와 과정마다 넘어야 할 과제들이 많이 있을 것입니다. 특히 인간과 가까운 언어(발화/문장)를 출력하고 인간과 기계, 기계와 기계가 대화하는 시스템의 구축은 인공지능 분야뿐만 아니라 언어학 분야에서도 함께 모색되어야 할 과제라고 할 수 있습니다. 이러한 일은 궁극적으로 인간의 지적 활동과 언어능력을 형식 논리, 또는 알고리즘으로 얼마나 유사하게 구현할 수 있는지의 문제일 것입니다.

오늘 발표를 들으면서 드는 몇 가지 질문을 중심으로 토론을 대신하겠습니다.

1. 언어처리 과정에서 가장 문제가 되는 것은 개별 언어 고유의 특성일 것입니다. 물론 이러한 프로그램을 개발할 때 언어 보편적인(독립적인) 영역과 언어 특정한(종속적인) 영역이 나뉘 것으로 보이지만, 특히 궁금한 것은 언어 특정한 영역을 어떻게 대응하느냐 하는 점입니다. 대부분 통계 기반이나 기계학습 등의 방법으로 처리할 수 있을 겁니다. 하지만, 일례로 ‘유성음화 현상과 경음화 현상’처럼 한국어에서 경쟁하는 비규칙적인 음운 현상을 음성합성(TTS)을 할 때 어떻게 처리할 수 있는지요? 대응량 음성 데이터베이스 구축 및 프로토타입 생성만으로 가능한지요. 이는 자동번역에서의 문장 중의성 해소, 대화 시스템 등에서의 발화 의도 파악 등과도 일맥상통하는 과제일 것 같습니다.

인간의 지적[jijeoknaver ↔ jijeokgoogle] 활동 / 인간의 지적인[jijeoginnaver ↔ jijeogingoogle] 활동  
그렇게 하다가는 반드시 대가[daegganaver/google→ daega]를 치르게 될 것이다.  
아눔이 미술계의 대가[daegganaver/google↔ daega]를 홀대했다.

2. 대화 처리 시스템과 관련하여 질문 드립니다. 대화 처리 시스템에서 중요한 것이 대화 문맥을 계속 유지하는 것이라고 봅니다. 대화 문맥을 유지한다는 것은 선행 대화 정보, 또는 화청자 간의 주고받은 정보를 상실하지 않고, 해당 발화에 누적적으로 반영토록 하는 것일 텐데, 이에 대한 기술이 어느 정도 진행되었는지 궁금합니다.<sup>1)</sup> 여기에 적용하는 기술이 어떤 것인지요? 덧붙여 ETRI에서 진행하고 있는 동시통역기 개발의 경우, 대화, 강연, 통화 등에

서 흔히 발견되는 비정형 발화(생략, 간투사 개입, 비문법적 발화, 맥락의존적 의미, 논리적 점핑 등등)를 어떻게 해소하는지요? 아울러 질의 응답 시스템처럼 질문에 대한 문서 검색이나 적정 정답을 제시한 것 외의 언어 사용 영역(감정 공유, 평가, 명령, 제안, 설득, 논쟁, 토론 등)에 적응한 AI 개발의 진행 상황을 알려주시면 고맙겠습니다.

3. 기계 번역에 있어서 언어학자들의 역할에 대해서 궁금합니다. 초기의 '규칙 기반 기계 번역(RBMT)'에 비해 통계 기반 기계 번역(SMT)을 거쳐 신경망 기반 기계 번역(NMT) 시대에 접어들면서 언어학자들의 역할은 축소, 소멸해 가고 있는 게 아닌가 싶습니다. 작년 5월 IBM에서는 '왓슨' 기반 인공지능 플랫폼으로 한국어를 배우겠다는 공언했고, 1년 만인 지난 9월 '대화, 자연어 처리, 자연어 분류, 자동 번역' 등의 기능을 갖춘 '왓슨 API'를 출시했습니다. 더욱 직업적 위기감(!)을 갖게 한 것은 '한국어를 전혀 모르는' 분이 이 프로젝트의 책임자라는 것이었습니다.

다만, 인공지능 연구에서 이미지나 영상을 통한 사물 인식, 패턴 인식(pattern recognition), 기계학습(classification, regression, ranking, collaborative filtering, clustering data, embedding 등)과 같이 비언어적 데이터 정보의 가공·처리·추출 등은 비약적인 발전을 이루었지만, 이에 비해 언어 처리 영역에서는 여전히 해결해야 할 난제들이 많이 남아 있는 것으로 보입니다. 어떤 과제가 있고, 특별히 언어학 연구자들의 역할이나 과제가 무엇인지 밝혀주시면 고맙겠습니다.

- 1) 발표 내용과 관련되는 듯한데, Siri와 같은 '1문장 - 1답변'의 경우는 선행 대화 정보를 유지하지 않습니다. 단문 답변의 경우에도, 미리 설정된 몇 가지 예문을 반복적으로 서비스하고 있습니다.
- (i) "Q1: 랩 한번 해봐." - "A1: 알겠습니다... 자, 갑니다. 제가 작사했어요. 래퍼 토끼에게 미리 사과드립니다. '너와 나의 연결 고리, 이건 우리 안에 소리, 너와 나의 연결 고리, 이건 우리 안에 소리'"
- "Q2: 랩 다시 해봐." - "A2: 다음엔 돈을 내셔야 해요.... 농담입니다."
- "Q3: 다른 랩 한번 해봐." - "A3: 알겠습니다... 자, 갑니다. 제가 작사했어요. 래퍼 토끼에게 미리 사과드립니다. '너와 나의 연결 고리, 이건 우리 안에 소리, 너와 나의 연결 고리, 이건 우리 안에 소리'"
- (ii) "Q1: 끝말잇기하자." - "A1: 좋아요, 제가 먼저 시작할게요. 꽃무늬!!(마귀광대버섯!!과녁!!)"
- "Q2: 끝말잇기 내가 먼저 할게." - "A1: 좋아요, 제가 먼저 시작할게요. 마귀광대버섯!"
- (iii) "Q: 한국에서 제일 큰 도시 이름이 뭐야." - "A: Siri입니다. 잘 부탁드립니다."
- (iv) "Q: 너는 몇 살이야." - "A: 글썄요, 사람들 덕분에 딱 늙긴 했어요. 확실하요."
- (v) "Q: 너는 남자냐, 여자냐?" - "A: 그런 질문에 대답해 드릴 시간은 없습니다."
- (vi) "Q: 시리는 집이 어디냐." - "A: 저 Siri는요, 캘리포니아에 있는 Apple에서 디자인되었습니다."
- (vii) "Q: 인간은 죽는다. 소크라테스는 인간이다. 소크라테스는 죽을까?" - "A: 우리 사이에 그런 것이 중요한가요?"
- (viii) "Q: 너는 언제 죽냐" - "A1: 누구요? 저요?" - "Q: 그래, 너" - "A2: 누구요? 저요?" - "Q: 응, 그래 너." - "A3: 누구요? 저요?"

## 주제 2

# 말뭉치 구축의 세계 동향과 국어 말뭉치의 현주소

· 발표자 김한샘(연세대학교 언어정보연구원 HK 교수)



2017년 국어 정책 학술 대회 - 우리말 정보화의 현황과 과제



## 말뭉치 구축의 세계 동향과 국어 말뭉치의 현주소

김한샘

YONSEI UNIVERSITY



## 차 례

- I. 말뭉치의 규모
- II. 말뭉치 활용 인프라
- III. 국어 말뭉치의 현주소
- IV. 말뭉치 활용 한국어 처리 사례

## 말뭉치의 규모

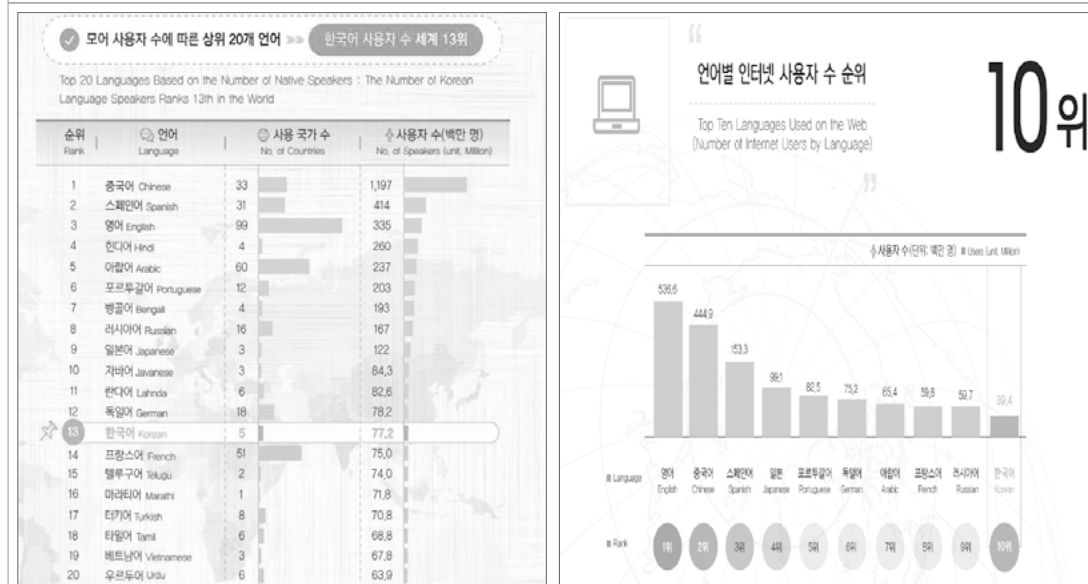
-3-

## 세계 언어 자원 현황 - 2007년(21세기 세종계획 종료 시)



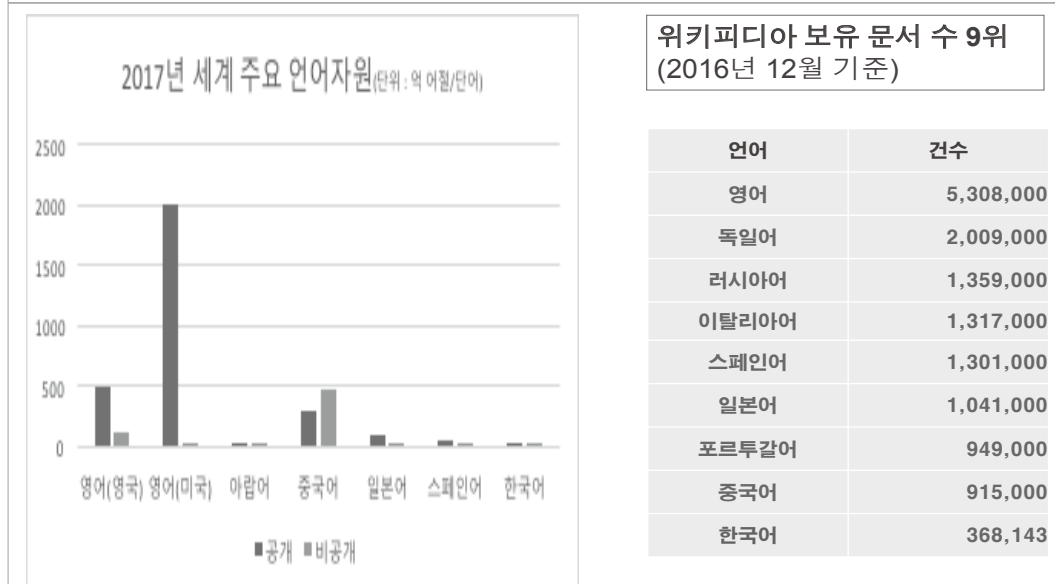
-5-

## 사용자 수 기준 한국어 위치 - 숫자로 살펴보는 우리말(국립국어원, 2014)



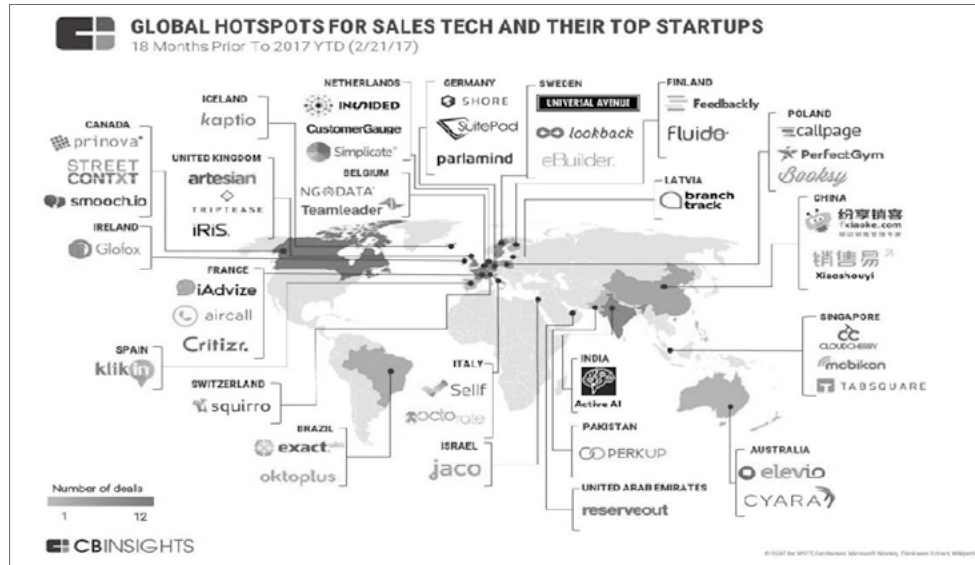
-4-

## 세계 언어 자원 현황 - 현재(국가 차원 언어 자원 구축 과제 10년 공백)



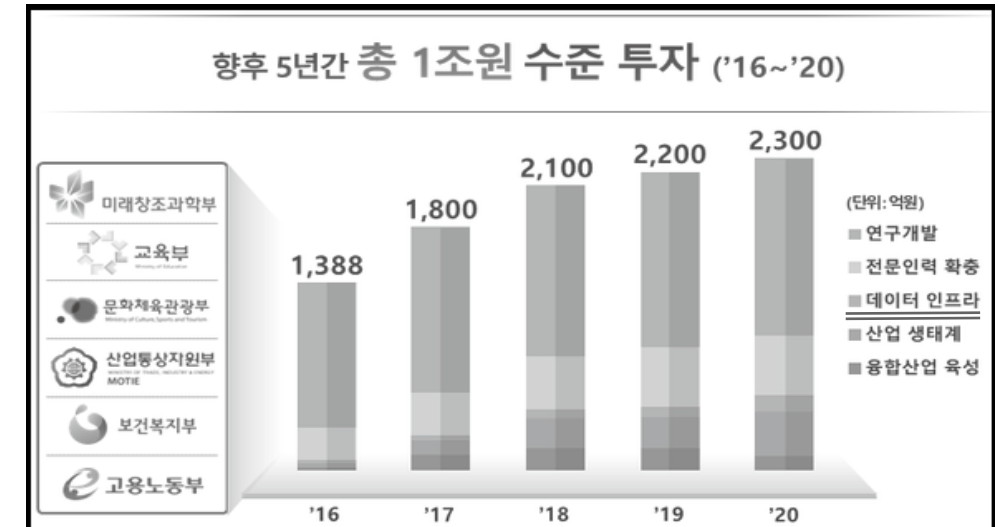
-6-

## 2017년 인공지능 관련 기업 거래량 (언어 자원 규모와 비례)



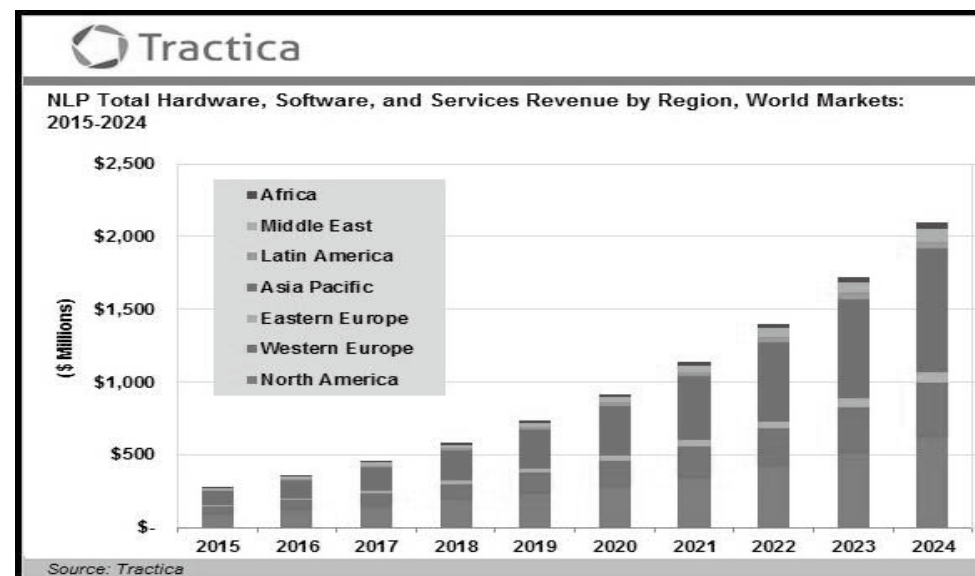
-7-

## 인공지능 관련 국가 차원 투자 계획



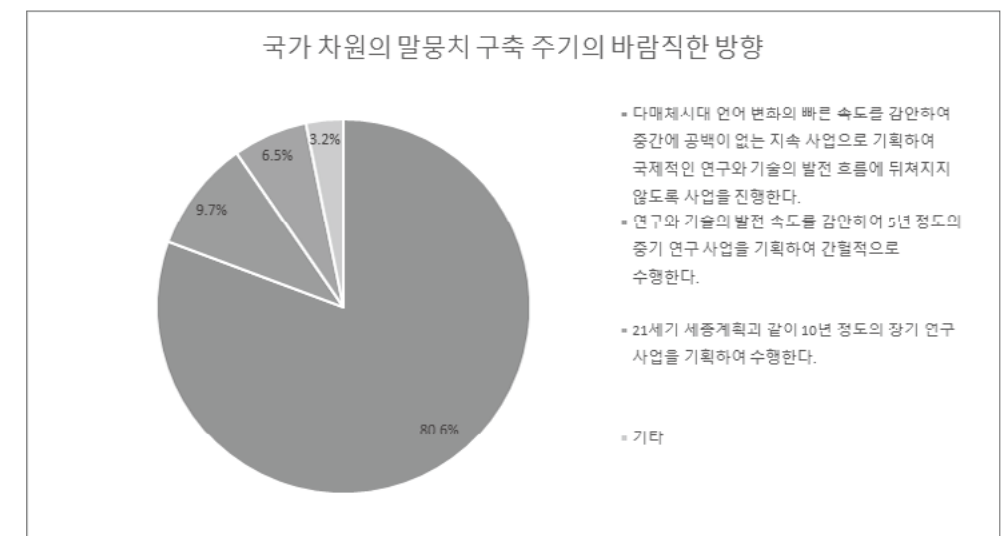
-9-

## 인공지능 시장 예측



-8-

## 국가 차원에서 말뭉치의 지속적 구축과 관리 필요



-10-



## 말뭉치 관련 인프라

-11-

## 말뭉치 공유 인프라 활성화 (유럽 19개국 언어 자원 공유 인프라)

CLARIN stands for "Common Language Resources and Technology Infrastructure".

It is a research infrastructure that was initiated from the vision that all digital language resources and tools from all over Europe and beyond are accessible through a single sign-on online environment for the support of researchers in the humanities and social sciences.

Members	National Consortium (NC)	Leading NC partner	National coordinator
Austria	CLARIN Austria	CLARIN Centre Vienna	Karlheinz Mörth
Bulgaria	CLARIN Bulgaria	Bulgarian Academy of Sciences	Kiril Simov
Czech Republic	LINDAT/CLARIN	Charles University Prague	Eva Hajičová
Denmark	CLARIN-DK	University of Copenhagen	Bente Maegaard
Dutch Language Union	CLARIN DLU / Flanders	Dutch Language Institute	Griet Depoorter
Estonia	CLARIN Estonia	Center of Estonian Language Resources	Kadri Vider
Finland	FIN-CLARIN	University of Helsinki	Kristen Lindén
Germany	CLARIN D	University of Tuebingen	Erhard Hinrichs
Greece	clarimel	ILSP-ATHENA Research Center	Stelios Piperidis
Hungary	HunCLARIN	Research Institute for Linguistics, Hungarian Academy of Sciences	Tamás Váradi
Italy	CLARIN IT	Institute for Computational Linguistics A. Zampolli, Italian National Research Council	Monica Monachini
Latvia	CLARIN-LV	Institute of Mathematics and Computer Science, University of Latvia	Inguna Skadiņa
Lithuania	CLARIN-LT	Vytautas Magnus University	Jurgita Vaičenoniene
The Netherlands	CLARIAH NL	Utrecht University	Jan Odijk
Norway	CLARINO	University of Bergen	Koenraad De Smedt
Poland	CLARIN PL	Wrocław University of Technology	Maciej Piasecki
Portugal	CLARIN Portugal	University of Lisbon	António Branco
Slovenia	CLARIN-SI	Jozef Stefan Institute	Tomaž Erjavec
Sweden	SWE-CLARIN	Språkbanken	Lars Borin

## 말뭉치 공유 인프라 활성화 (영어 중심의 107종 말뭉치 공개)

### Natural Language Toolkit

NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active discussion forum.

### NLTK Corpora

NLTK has built-in support for dozens of corpora and trained models, as listed below. To use these within NLTK we recommend that you use the NLTK corpus downloader, >>> `nltk.download()`

Please consult the README file included with each corpus for further information.

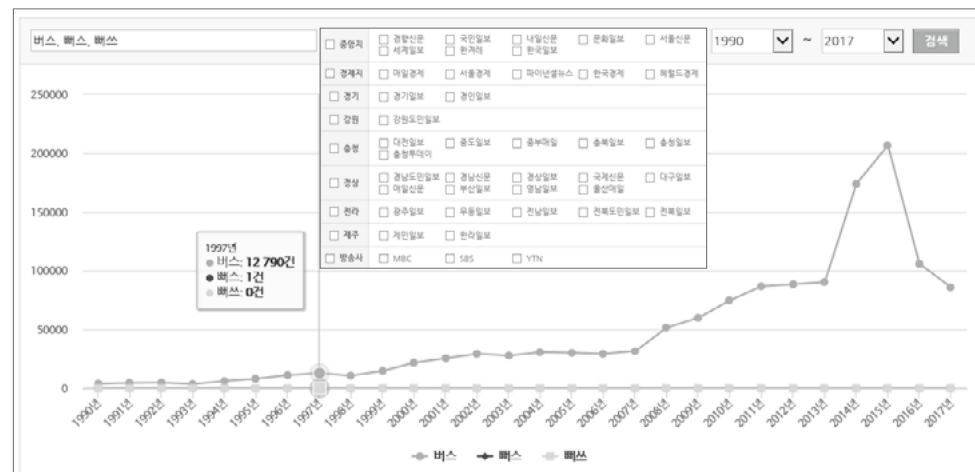
1. *ACE Named Entity Chunker (Maximum entropy)* [ [download](#) | [source](#) ]  
id: maxent\_ne\_chunker; size: 13404747; author: ; copyright: ; license: ;
2. *Australian Broadcasting Commission 2006* [ [download](#) | [source](#) ]  
id: abc; size: 1487831; author: Australian Broadcasting Commission; copyright: ; license: ;
3. *Alpino Dutch Treebank* [ [download](#) | [source](#) ]  
id: alpino; size: 2797255; author: ; copyright: ; license: Distributed with permission of Gerjan van Noord;
4. *BioCreative (Critical Assessment of Information Extraction Systems in Biology)* [ [download](#) | [source](#) ]  
id: biocreative\_ppi; size: 223366; author: ; copyright: Public Domain (not copyrighted); license: Public Domain;
5. *Brown Corpus* [ [download](#) | [source](#) ]  
id: brown; size: 3314357; author: W. N. Francis and H. Kucera; copyright: ; license: May be used for non commercial purposes.;
6. *Brown Corpus (TEI XML Version)* [ [download](#) | [source](#) ]  
id: brown\_tei; size: 8737738; author: W. N. Francis and H. Kucera; copyright: ; license: May be used for non-commercial purposes.;

## 국가 차원 언어 자원 구축 인프라 비교 (일본 : 한국)

항목 \ 기관		일본 국립국어연구소	한국 국립국어원
보유 말뭉치	문어 말뭉치	1억 단어	1.1억 어절
	구어 말뭉치	780만 단어	420만 어절
	역사 말뭉치	1400만 단어	1300만 어절
	웹 말뭉치	250억 단어	-
저작권	확보율	75% (현대 말뭉치 거의 해결)	60% (구어 말뭉치 기준)
	방식	직접 + 일본문예가협회	과제 수행 연구진
말뭉치 구축	원시 말뭉치	업체 외주	대학, 연구소에 용역 과제
	분석 말뭉치	연구소 인력	대학, 연구소에 용역 과제
인력		<문어 말뭉치 1억 기준> 정규직: 7명 전체 30명	<21세기 세종계획 전체> 정규직: 2명(과제 관리)

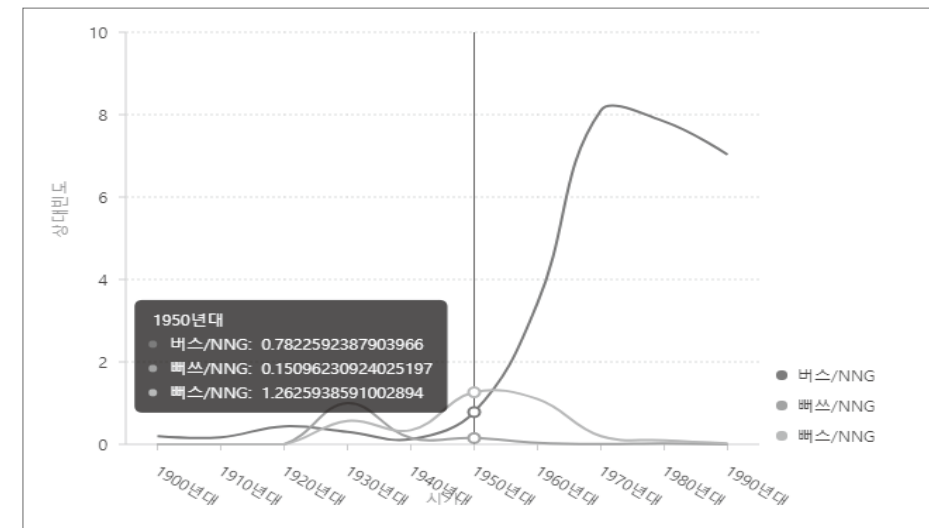
-14-

## 빅카인즈(한국언론진흥재단 - 41개 매체): 원문 판매



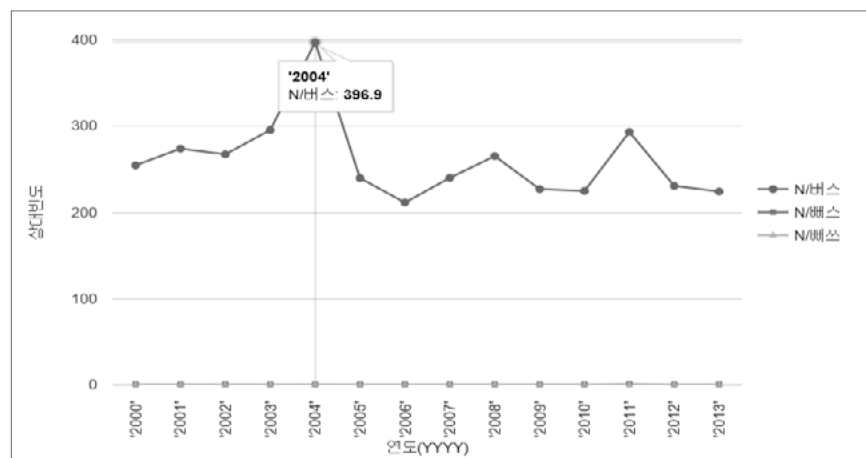
-15-

## 연세 20세기 한국어 말뭉치 - 문학/정보/신문/잡지 (원문 비공개)



-17-

## 고려대 물결 21 - 동아, 조선, 중앙, 한겨레 (원문 비공개)



-16-

## 말뭉치 보유 기관: 문화재청

■ 문화재청: 국가기록유산([www.memorykorea.go.kr](http://www.memorykorea.go.kr))

◆ 문화재청: 텍스트 말뭉치 비공개

◆ 문화재청-국가기록유산: 1,252개의 기록 유산 보유

- 원문이미지와 원문텍스트 자료 공개
- 원문텍스트: 텍스트 형식으로 복사 가능(텍스트 말뭉치 구축 가능)

-18-

## 말뭉치 보유 기관: 한국고전번역원

■ 한국고전번역원([www.itkc.or.kr](http://www.itkc.or.kr))

## ◆ 한국고전종합DB

- 고전번역서, 고전원문, 한국문집총간, 조선왕조실록, 승정원일기, 일성록
- 원문이미지와 원문텍스트 자료 제공
- 원문텍스트: 텍스트 형식으로 저장 가능

## ◆ '한국고전종합DB' 제공 자료 외 자료

- 자체적으로 입력해 놓은 말뭉치 다수 보유: 비공개

## ◆ 고전용어 시소러스 제공

-19-

## 말뭉치 보유 기관: 국사편찬위원회 - 한국역사정보통합시스템

■ 한국역사정보통합시스템(<http://www.koreanhistory.or.kr>)

- ◆ 디렉토리서비스
- ◆ 편년자료서비스
- ◆ 연계사이트제공
- ◆ 시소러스 검색
- ◆ 원문 부분 공개

-20-

## 국어 말뭉치의 현주소

## - 국가 차원의 계획적 말뭉치 구축 중단

→ 주체별 언어 자원의 공유 및 호환 불가

→ 언어 연구와 언어 처리의 연계성 약화

## - 목적별 특수 말뭉치 구축 및 활용 일반화

-21-

## 국어 말뭉치 현황: 국립국어원 구축 자료 (국가 차원)

- ◆ 21세기 세종계획 현대 문어 말뭉치
- ◆ 21세기 세종계획 현대 구어 말뭉치
- ◆ 21세기 세종계획 역사 자료 말뭉치
- ◆ 21세기 세종계획 병렬 말뭉치
- ◆ 21세기 세종계획 북한 및 해외 한국어 말뭉치
- ◆ 한국어 어휘 분석 말뭉치 - 현대 국어 사용 빈도 조사 1, 2
- ◆ 초등교과서 말뭉치 / 초등학생 작문 말뭉치
- ◆ 한국어 학습자 말뭉치
- ◆ 북한어 말뭉치

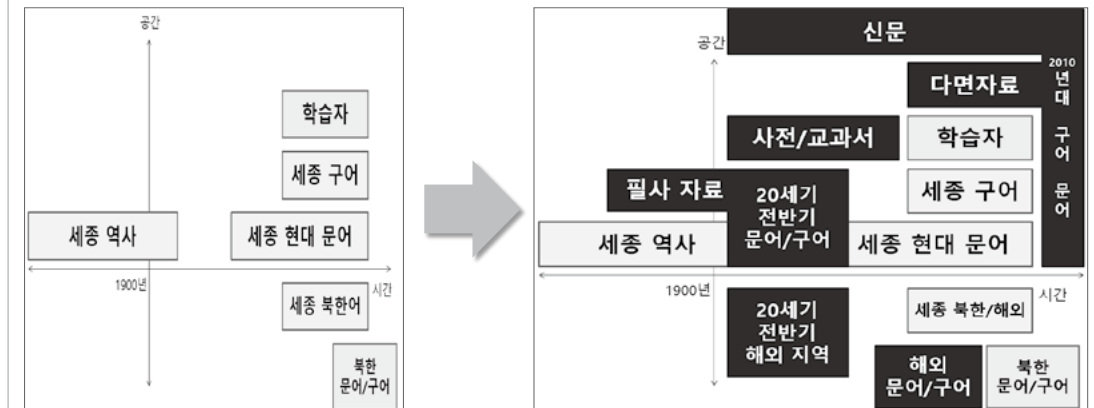
-22-

## 국어 말뭉치 현황 - 대학 및 기관 구축 자료

- 연세대학교 : 현대 구어, 문어, 학습자, 20세기 문어, 교과서, 다면 자료 등
- 고려대학교 : 현대 문어, 신문, 20세기 구어, 성인 자유 발화 등
- 경희대학교 : 활자본 고소설, 개화기 등
- 카이스트 : 현대 문어, 구어, 구문 분석, 다국어 병렬, 신문 등
- 나사렛대학교 : 학령기 전 아동 구어
- 서울대학교 : 감정 및 의견 주석
- 세한대학교 : 수화 언어
- 영남대학교 : 심리학적 주석
- 울산대학교 : 형태 의미 주석
- 제주대학교 : 재일 한국인 담화
- 한국외국어대학교 : 각 분야 리뷰, 다국어 병렬, 감성 주석
- 한양대학교 : 다문화 가정 외국인 담화
- 한국학중앙연구원 : 고소설, 신소설, 현대시, 연가, 구술사 등
- 한국전자통신연구원(ETRI) : 구어, 시나리오, 질문, 병렬, 질의응답용 등

-23-

## 말뭉치의 시간적, 지역적 범위 확대 필요



-25-

## 언어 자원 공유와 호환성의 문제 → 언어 연구와 언어 처리의 연계 약화

- ◆ 저작권의 문제: 연구용 한정, 검색 서비스용 한정
- ◆ 주석의 대상과 깊이: 문법 형태소 처리, 체언 중심 분석
- ◆ 자료 구축의 지속성: 목적 달성 후 폐기, 사후 관리 및 재활용 X
- ◆ 평가 기준 및 주체 부재: 사용자의 선택

→ 국가 차원의 언어 자원 구축 및 관리 필요

-24-

## 역사 자료 말뭉치의 시기적 공백 보완

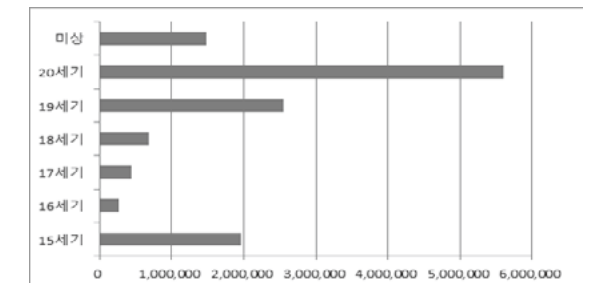
## ■ 19세기 말~20세기 전반기 자료 보충

## ◆ 당대 자료의 특성을 반영한 분류 틀 마련

- 기존의 세종 역사 혹은 세종 현대의 틀을 그대로 적용할 수 없음
- 국한문체, 순한글체 등 문체에 따라 언어 사용 양상에 큰 차이가 있음
- 근대적 매체, 출판물의 등장으로 기존 역사자료와는 분류 틀이 달라야 함
- 세종 역사자료 태그셋에 일부 태그 추가 필요(4음절 한문구, 1음절 한자어 등)

## ◆ 기구축 디지털 아카이브의 말뭉치화

- 국가 기관 대규모 아카이브
- 기관 차원의 협조 요청 필요
- 원문 대조 및 마크업 통일



-26-



## 역사 자료 말뭉치의 장르별, 지역별 공백 보완

## ■ 필사자료 보충

- ◆ 연간, 한글 연행록 등 한글 필사자료 보충

## ■ 해외 자료 보충

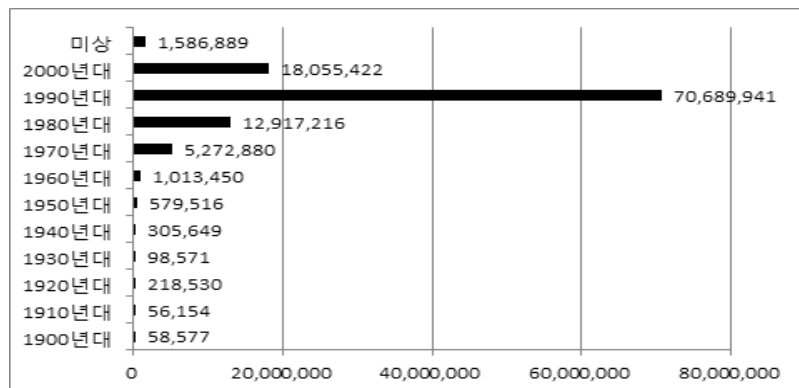
- ◆ 해외 이주가 본격화된 19세기 말 이후
- ◆ 미국, 러시아, 일본, 중국 등 해외 교민사회의 신문, 잡지

-27-

## 현대 국어 문어 말뭉치 구축

## ◆ 시기적 불균형 보완

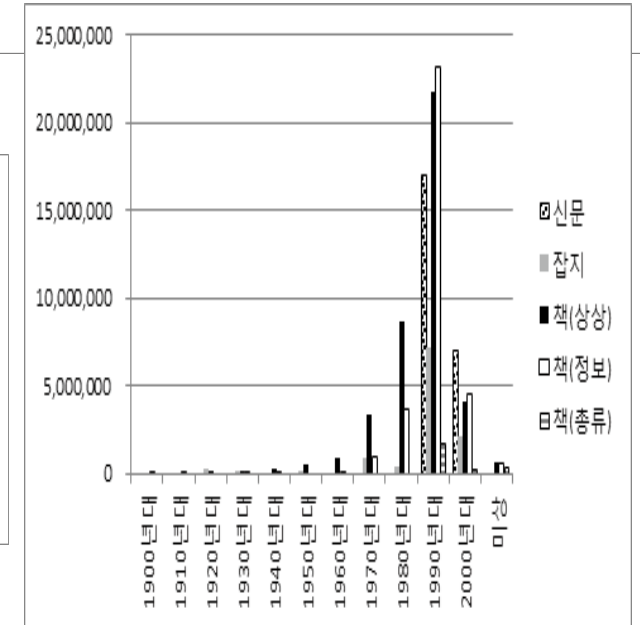
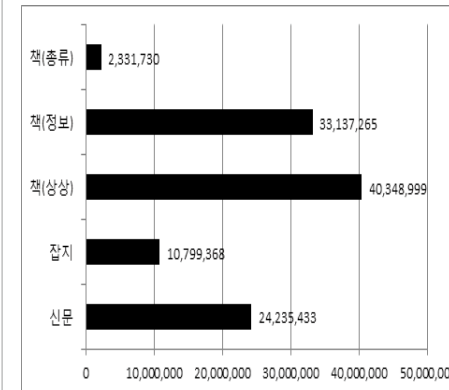
- 특정 연대에 집중된 자료 분포 개선
- 2010년대 이후 자료 보충
- 모니터링 말뭉치의 지속적 구축 필요



-28-

## 현대 국어 문어 말뭉치 구축

## ◆ 대표성과 균형성 확보

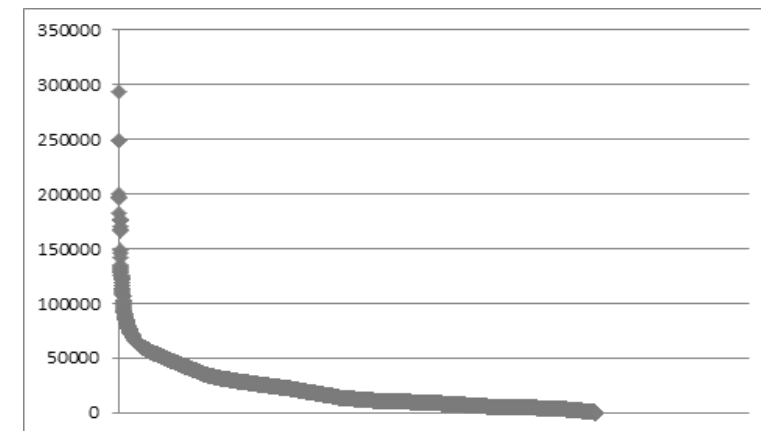


-29-

## 현대 국어 문어 말뭉치 구축

## ◆ 표본 크기 통제

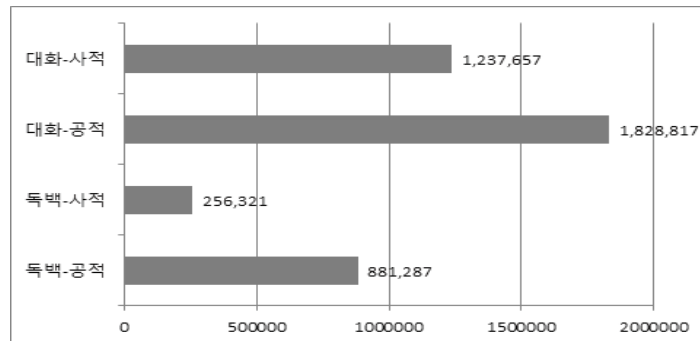
- 최대 크기의 표본은 294,053어절, 최소 크기의 표본이 58어절



-30-

## 현대 국어 구어 말뭉치 구축

## ◆ 사적 자료, 독백 자료 보완 필요



-31-

## 현대 국어 구어 말뭉치 구축

## ◆ 발화자 연령대의 다양화 및 성별의 균형 확보

	10대	20대	30대	40대	50대	60대 이상	나이 미상	성별 합계	성별 비율
남성	101	734	230	308	265	175	2,635	4,448	63.50%
여성	158	913	272	61	64	18	844	2,330	33.26%
성별 미상	0	1	0	0	0	0	226	227	3.24%
연령별 합계	259	1,648	502	369	329	193	3,705	7,005	
연령별 비율	3.70%	23.53%	7.17%	5.27%	4.70%	2.76%	52.89%		

-32-

## 현대 국어 말뭉치 구축

## ◆ 새로운 매체 환경 반영 분류 확장 필요 Biber, Egbert &amp; Davies (2015)

사용역	하위 사용역	사용역	하위 사용역
내러티브	뉴스 보도/블로그 스포츠 보도/블로그 개인적 내러티브 블로그 역사 기사 여행 블로그 짧은 이야기 소설 사건적 이야기/역사 잡지 기사 누고 회고록 기타	설명서	설명서 레시피 FAQ 기술 지원 기타
정보성 기술/설명	기술 백과사전 정보성 블로그 사람에 대한 기술 연구 논문 조목 정보성 FAQ 법률 용어 및 조항 수업 자료 기술 보고서 기타	정보성 설득	판매 목적 기술 설득적 기사/에세이 사설
의견	종교 블로그/설교 조언	서정시	노래 가사 시 기타
상호 토론	토론 포럼 심의 윤리 포럼 독자 및 시청자 반응 기타	구어	인터뷰 비디오/오디오 선사 동적 연설 TV/영화 대본기타

-33-

## 기술 발전에 따른 말뭉치 구축의 방향성

- 빅데이터로 일컬어지는 대규모 언어 자원 구축 가능 (양적 가치)
- 인공지능을 위한 고도의 언어 정보 구축 필요 (질적 가치)
- 완성도 높은 범용 언어 자원 구축에 대한 재고  
: 공개 자료 기반 대규모 말뭉치와 정교하게 설계, 주석한 균형 말뭉치 구축 병행
- 언어 자원 구축의 다양화
- 기구축 자료와의 호환성 확보

-34-

## 대규모 언어 자원 구축을 위한 주석 방법 변경



## ■ 대규모 언어 자원 분석 시 전형적인 주석 단계 수정 필요

- 매뉴얼 태깅 단계 생략, 기계적 후처리(연세대)
- 기계적 후처리, 주석 체계 간소화(고려대)

-35-

## 활용 목적에 따른 주석(6종 주석 통합 언어 분석 말뭉치) 인공지능 개발: Q

```

"morph" : [
  {"id" : 0, "lemma" : "과거", "type" : "NNG", "position" : 0, "weight" : 0.9 },
  {"id" : 1, "lemma" : "하세", "type" : "NNG", "position" : 7, "weight" : 0.176038 },
  {"id" : 2, "lemma" : "윤림곡", "type" : "NNG", "position" : 14, "weight" : 0.9 },
  {"id" : 3, "lemma" : "익", "type" : "JKG", "position" : 23, "weight" : 0.0694213 },
  {"id" : 4, "lemma" : "정식", "type" : "NNG", "position" : 27, "weight" : 0.862814 },
  {"id" : 5, "lemma" : "종록", "type" : "NNG", "position" : 34, "weight" : 0.9 },
  {"id" : 6, "lemma" : "으로", "type" : "JKB", "position" : 40, "weight" : 0.153406 },
  {"id" : 7, "lemma" : "맞", "type" : "VV", "position" : 47, "weight" : 0.443633 },
  {"id" : 8, "lemma" : "는", "type" : "ETM", "position" : 50, "weight" : 0.184941 },
  {"id" : 9, "lemma" : "것", "type" : "NNE", "position" : 54, "weight" : 0.228788 },
  {"id" : 10, "lemma" : "은", "type" : "JX", "position" : 57, "weight" : 0.0688243 },
  {"id" : 11, "lemma" : "?", "type" : "SF", "position" : 60, "weight" : 1 }
],

```

```

"dependency" : [
  {"id" : 0, "text" : "과거", "head" : 1, "label" : "NP", "mod" : [], "weight" : 0.0492727 },
  {"id" : 1, "text" : "하세", "head" : 2, "label" : "NP", "mod" : [0], "weight" : 0.703307 },
  {"id" : 2, "text" : "윤림곡의", "head" : 4, "label" : "NP MOD", "mod" : [1], "weight" : 0.751682 },
  {"id" : 3, "text" : "정식", "head" : 4, "label" : "NP", "mod" : [], "weight" : 0.611044 },
  {"id" : 4, "text" : "종록으로", "head" : 5, "label" : "NP AJT", "mod" : [2, 3], "weight" : 0.784158 },
  {"id" : 5, "text" : "맞는", "head" : 6, "label" : "VP MOD", "mod" : [4], "weight" : 0.860741 },
  {"id" : 6, "text" : "것은?", "head" : -1, "label" : "NP_SBJ", "mod" : [5], "weight" : 0.00494862 }
],

```

-37-

## 상세한 언어 정보 주석 - ISO 위원회 안(의미 분석 지향)

주석 번호	범주명	주석 번호	범주명
1	명사	11	조사
1-1	부동 명사	11-1	각조사
1-1-1	실행 명사	11-1-1	주격 조사
1-1-2	활동 명사	11-1-2	목적격 조사
1-2	고유 명사	11-1-3	관형격 조사
1-3	의존 명사	11-1-4	부사격 조사
1-3-1	단위 의존 명사	11-1-4-1	여격 조사
2	수사	11 1 4 2	위격   저소격 조사
2 1	양수사	11 1 4 3	소격 조사
2 2	서수사	11 1 4 4	구격   상동격 조사
3	대명사	11 1 4 5	상대격   행위자격
3-1	지시 대명사	11-1-4-6	탈위격   탈격
3-2	인칭 대명사	11-1-4-7	유격격
4	동사	11-1-4-8	항진격
4-1	자동사	11-1-4-9	원인격
4 2	타동사	11 1 4 10	자격격
4-3	보조 동사	11-1-4-11	비교격
5	형용사	11 1 4 12	면성격   결과격
5 1	성상 형용사	11 1 4 13	인용격
5-2	지시 형용사	11-1-5	소격 조사
5-3	보조 형용사	11-1-6	보격 조사

6	부사	11-1-7	서울격 조사
6 1	성분 부사	11 2	보조사
6-1-1	행동 부사	11-3	접속 조사
6-1-2	상태 부사	12	어미
6-2	분상 부사	12-1	어말 어미
6-2-1	접속 부사	12-1-1	전성 어미
7	관형사	12-1-1-1	명사형 어미
7-1	수 관형사	12-1-1-2	관형사형 어미
7-2	시시 관형사	12-1-1-3	부사형 어미
7-3	성질 관형사	12-1-2	연결 어미
8	감탄사	12-1-2-1	방형 연결 어미
9	설문사	12-1-2-2	내림 연결 어미
9 1	명사 접두사	12 1 2 3	선택 연결 어미
9-2	수사 접두사	12-1-2-4	원인 연결 어미
9-3	동사 접두사	12-1-2-5	조건 연결 어미
9 4	경유사 접두사	12 1 2 6	상황 연결 어미
9-5	부사 접두사	12-1-3	종결 어미
10	진미사	12-1-3-1	광서형 어미
10-1	명사 접미사	12-1-3-2	외분형 어미
10-2	수사 접미사	12-1-3-3	명령형 어미
10-3	동사 접미사	12-1-3-4	칭유형 어미
10-4	형용사 접미사	12-1-3-5	강단형 어미
10-5	부사 접미사	12-2	선어말 어미

-36-

## 활용 목적에 따른 주석 (6종 주석 통합 언어 분석 말뭉치) 인공지능 개발: A

```

"morph" : [
  {"id" : 0, "lemma" : "윤림곡", "type" : "NNG", "position" : 0, "weight" : 0.9 },
  {"id" : 1, "lemma" : "윤다리기", "type" : "NNG", "position" : 10, "weight" : 0.0701606 },
  {"id" : 2, "lemma" : "는", "type" : "JX", "position" : 22, "weight" : 0.0332677 },
  {"id" : 3, "lemma" : "윤림곡", "type" : "NNG", "position" : 26, "weight" : 0.9 },
  {"id" : 4, "lemma" : "에서", "type" : "JKB", "position" : 35, "weight" : 0.153407 },
  {"id" : 5, "lemma" : "윤다리기", "type" : "NNG", "position" : 42, "weight" : 0.0789692 },
  {"id" : 6, "lemma" : "로", "type" : "JKB", "position" : 54, "weight" : 0.0822907 },
  {"id" : 7, "lemma" : "경기", "type" : "NNG", "position" : 58, "weight" : 0.152575 },
  {"id" : 8, "lemma" : "를", "type" : "JKO", "position" : 64, "weight" : 0.137686 },
  {"id" : 9, "lemma" : "겨우", "type" : "VV", "position" : 68, "weight" : 0.9 },
  {"id" : 10, "lemma" : "는", "type" : "ETM", "position" : 74, "weight" : 0.184941 },
  {"id" : 11, "lemma" : "윤림곡", "type" : "NNG", "position" : 78, "weight" : 0.9 },
  {"id" : 12, "lemma" : "경기", "type" : "NNG", "position" : 88, "weight" : 0.135556 },
  {"id" : 13, "lemma" : "종록", "type" : "NNG", "position" : 95, "weight" : 0.9 },
  {"id" : 14, "lemma" : "이", "type" : "VCP", "position" : 101, "weight" : 0.017525 },
  {"id" : 15, "lemma" : "다", "type" : "EF", "position" : 104, "weight" : 0.353579 },
  {"id" : 16, "lemma" : ".", "type" : "SF", "position" : 107, "weight" : 1 }
],

```

```

"dependency" : [
  {"id" : 0, "text" : "윤림곡", "head" : 1, "label" : "NP", "mod" : [], "weight" : 0.61187 },
  {"id" : 1, "text" : "윤다리기는", "head" : 8, "label" : "NP_SBJ", "mod" : [0], "weight" : 0.414461 },
  {"id" : 2, "text" : "윤림곡에서", "head" : 5, "label" : "NP_AJT", "mod" : [1], "weight" : 0.772406 },
  {"id" : 3, "text" : "윤다리기로", "head" : 5, "label" : "NP_AJT", "mod" : [1], "weight" : 0.652726 },
  {"id" : 4, "text" : "경기", "head" : 5, "label" : "NP_OBJ", "mod" : [1], "weight" : 0.725226 },
  {"id" : 5, "text" : "겨우", "head" : 8, "label" : "VP_MOD", "mod" : [2, 3, 4], "weight" : 0.814884 },
  {"id" : 6, "text" : "윤림곡", "head" : 7, "label" : "NP", "mod" : [1], "weight" : 0.60035 },
  {"id" : 7, "text" : "경기", "head" : 8, "label" : "NP", "mod" : [6], "weight" : 0.811877 },
  {"id" : 8, "text" : "종록이다.", "head" : -1, "label" : "VNP", "mod" : [5, 7, 1], "weight" : 0.0289792 }
],

```

-38-

## 말뭉치의 품질 관리

### ■ 주석 언어 자원에 대한 논의의 패러다임 전환 필요

- 다국어 처리, 한국어 정보 처리의 위상 강화를 위한 국제 표준과의 연계
- 분석 인공지능 연구 등 현 시대의 주제에 맞는 언어 자원 품질의 고도화

### ■ 매뉴얼 태깅의 필요성

- 자동 분석 기술의 한계: 문장 단위 50%대(영어 기준)  
→ 언어 정보 처리 기술 발전을 위해서는 정교한 언어 자원 필요

### ■ 매뉴얼 태깅 및 검수의 문제점

- 시간과 비용의 측면에서 빅데이터 처리에 한계
- 지침 해석과 관련한 주석자 간 불일치(IAA: Inter Annotator Agreement)
- 오류 판정의 정확성에 대한 순환적 한계

→ 문제를 해결하기 위한 다양한 관점의 연구 필요

-39-

## 말뭉치 활용 한국어 처리 사례

-40-

## Text to Text: 말뭉치 기반 스토리텔링 저작 지원 소프트웨어 개발

### 스토리텔링의 진화 이미지로만 머릿속을 맴돌던 창작 아이디어의 구체화

'스토리텔러'는 소설, 영화, 드라마, 애니메이션, 게임 등 콘텐츠 제작에 필수적인 아이디어 도출과 스토리 완성을 도와주는 국내 최초 한국형 스토리텔링 저작지원 소프트웨어입니다.  
이화여대 디지털스토리텔링연구소와 엔씨소프트문화재단은 기존 스토리텔러를 웹과 모바일에서 이용 가능하도록 리뉴얼 하였습니다.  
스토리텔러 2015는 국내 사나리오 산업과 궁극적으로 콘텐츠 산업 발전을 위해 무료로 서비스되고 있습니다.

#### ◎ 영화 DB 오픈소스화

- 1,500여 편의 영화 DB는 약 2만 4,000편의 영화 후보군에서 추출
- 대표영화의 선정 기준: 대중성(글로벌 흥행이 입증된 작품), 작품성(국제 영화제의 작품상, 각본상 수상), 서사성(명확한 3막 구조의 이크 플롯)
- 1,500여 편의 영화를 시퀀스와 장면으로 분할, DB화 작업: 12만 여 개의 요소로 구성



해당 기준에 맞는 영화 DB의 지속적인 업데이트

#### ◎ 이야기 구성의 질적 제고

- 기존 스토리를 참고, 변형하여 새로운 스토리 창작하는 프로세스

사례 기반 추론(Cased-based Reasoning)

- 과거의 특정한 경험을 찾아내어 학습, 개발, 문제 해결에 활용하는 방법, 유추적 추론의 일종

## Text to Image: 시각화 저작 도구 개발



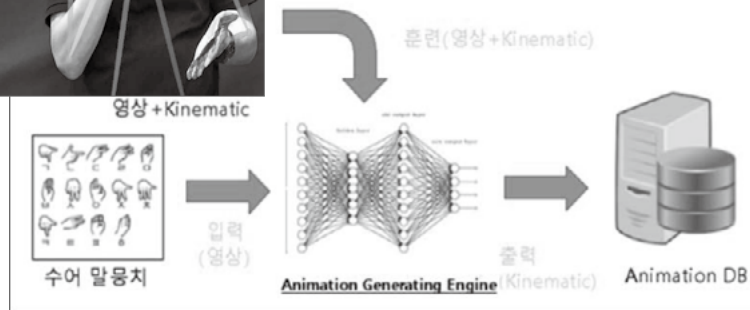
-42-



Image to Animation: 딥러닝 기반 수화 말뭉치 활용 애니메이션 생성



- 청각 장애인과 일반인의 커뮤니케이션을 위한 수어 번역기의 제작에 필요한 신경망을 통한 강화 학습 기법 적용
- 3차원 정보가 없는 기존의 수화(어) 방송 및 교육 데이터에서 3차원 애니메이션을 추출



-43-

고맙습니다.

-44-

주제 2

# 말뭉치 구축의 세계 동향과 국어 말뭉치의 현주소

· 토론자 **홍혜진**(국립국어원 언어정보과 학예연구관)



김한샘 선생님께서 세계 주요 언어의 말뭉치 구축 현황과 공유 환경 등에 대해 말씀해 주셨는데, 인공지능이 주목 받고 있는 이 시점에서 자원으로서의 말뭉치의 가치와 그 필요성을 확인할 수 있는 계기가 되었다고 생각합니다. 국립국어원에서 국어 정보화 사업을 담당하는 담당자로서 개인적으로는 국어 말뭉치의 현주소를 확인하고, 앞으로 관련 분야에서 보완되어야 할 점을 짚어 주셨다는 점에서 그 의의를 찾고 싶습니다. 2007년 '21세기 세종계획' 사업 종료 당시 국어 말뭉치의 규모는 미국, 일본, 중국 등 주요 국가에 크게 뒤떨어지지 않는 세계 최고 수준이었으나, 국가 차원의 말뭉치 구축 사업이 추진되지 못한 10년의 공백으로 현재는 뒤떨어진 상황입니다. 또한, 각 연구자나 기관의 연구 및 활용 목적에 따라 개별적으로 말뭉치를 구축하여 호환성의 문제가 발생하고, 이를 공유하는 환경이 마련되지 않아 활용도가 높지 않다는 문제도 있습니다. 이에 선생님께서 국가 차원에서 말뭉치를 지속적으로 구축하고 관리할 필요성에 대해 강조하신 점에 대해 대체로 의견을 같이합니다. 따라서 반론을 제기하기 보다는 선생님께 몇 가지 질문을 하고 추가적인 설명을 요청드리는 것으로 토론자의 소임을 다하고자 합니다.

우선 선생님께서는 국어 말뭉치의 시간적, 지역적 범위를 확대할 필요성에 대해 말씀해 주셨습니다. '21세기 세종계획'을 통해 구축된 말뭉치가 대개 1990년대 자료에 집중되어 있기 때문에 시기적인 불균형을 보완할 필요성에 대해서는 공감합니다. 다만, 19세기 말부터 20세기 전반기의 자료가 보충될 필요성에 대해서도 말씀해 주셨는데, 역사 자료 말뭉치의 추가 구축은 국어 연구를 위해서는 꼭 필요하고 중요한 작업이지만 언어처리 등의 활용을 생각해 보면 상대적으로 우선순위에서 밀릴 수밖에 없지 않은가 하는 생각이 듭니다. 한정적인 자원과 시간을 고려할 때 어떤 종류의 말뭉치를 우선적으로 구축하는 것이 필요하다고 생각하시는지 추가적인 말씀을 듣고 싶습니다.

선생님께서는 기술 발전에 따라 말뭉치 구축의 방향성을 크게 양적 가치와 질적 가치로 나누어 말씀하시면서 인공지능을 위한 고도의 언어 정보를 구축할 필요성이 있다고 하셨습니다. 공개된 자료를 기반으로 대규모 말뭉치를 구축하는 것 이외에도 정교하게 설계되고 주석한 말뭉치를 구축하는 작업이 병행되어야 한다고 하셨습니다. 내년 부터 국립국어원에서 추진할 예정인 국어 말뭉치 구축 사업에서도 언어 자원을 체계적으로 정비하고 보완하여 정제된 국어 말뭉치를 구축할 필요성에 대해서 공감하며, 관련 내용을 사업으로 계획 중입니다. 말씀하신 것처럼 전량을 수동으로 주석하는 방식이 아니라 기계적으로 분석된 결과를 수정하는 방식 등을 취한다면 분석 말뭉치의 구축 규모를 양적으로 확대하기 위해 들이는 노력이 과거에 비해 크게 줄어든 것이라고 봅니다. 다만, '21세기 세종계획'을 통해 얻어진 형태소 분석 정보나 동형의미어 분석 정보 등이 여전히 유용하게 사용되고는 있지만, 앞으로는 그보다는 높은 수준의 언어 분석 정보가 제공되어야 할 것으로 봅니다. 특정한 분야에서 필요한 분석 정보를 주석하는 것에 한정하는 경우가 아니라면 어떠한 정보를 어떻게 주석할 것인가의 문제, 다시 말해 언어 분석의 종류와 깊이에 관해서는 관련 학계와 산업계 등에서 충분히 논의를 하는 과정이 필요할 것으로 생각합니다. 질적 가치를 확대하는 방향을 생각할 때 앞으로 말뭉치 구축을 추진할 때 어떠한 언어 분석 정보를 제공하는 것이 좋은지에 대해 선생님의 고견을 듣고 싶습니다.

주제 3

# 말뭉치 언어학과 이론 언어학, 사전 편찬

· 발표자 **송상헌**(인천대학교 영어영문학과 교수)



## 1. 말뭉치 언어학의 현주소

독자가 인간 언어에 대한 생성문법적 관점, 특히 최소주의적 관점을 취하고 있다면, Sampson(2007)의 아래와 같은 주장에는 아마도 선뜻 동의를 표하기 어려울 것이다. 언어학적 배경이나 연구 이력에 따라서는 아래의 내용에 거부감을 느끼는 독자도 있을 것이다.

- (1) a. I believe that the concept of “ungrammatical” or “ill-formed” word-sequences is a delusion, based on a false conception of the kind of thing a human language is.
- b. The fact is that linguists who want to treat speakers’ intuitions rather than interpersonally-observable evidence as the basis of linguistic description are simply choosing to turn their back on science, and reverting to the pre-modern pattern of “arguments from authority”. Up to the early modern period, people “knew” that the Sun goes round the Earth. The Pope and other leaders of the Church proclaimed it, and Giordano Bruno was burned at the stake partly because he held a different opinion.
- c. My target article said that if we have a language-description based on empirical data, then an intuition-based description would be redundant; Hoffmann objects that when two people witness a crime, the police do not take just one of their statements. That is a false analogy: the two witnesses are logically equivalent in status. A better analogy is that, if we have used a ruler to measure the precise length of a line, we will not be interested in asking a person to estimate the length by eye.

실제로 Sampson(2007)은 관련된 분야를 전공으로 한 필자가 보기에 다소 과격하게 여겨질 수 있는 주장을 위 인용문 이외에도 여러 곳에서 피력하고 있다. 필자 역시 그 주장에 온전히 수긍을 하는 것은 물론 아니다. 그러나 다소 표현을 거칠게 한 측면은 있으나, Sampson의 주장에도 주요하게 고려할 만한 지점은 존재한다. 언어 연구의 흐름에 대한 그의 특징적 비판을 요약하자면, 언어학(특히 통사론)은 자연세계의 법칙을 설명하는 데 있어서 연구자 개인의 내현적 판단(introspective judgments)을 핵심적으로 사용하면서 동시에 스스로를 과학이라고 주장하는 아주 희한한 분야라는 것이다. 대부분의 현대 과학은 어떠한 자료(예컨대 말뭉치)나 실험을 통해 입증된 것만을 인정하고 그렇지 못한 것은 가설로 남겨둔다. 이에 비해 언어학은 인지과학의 하위영역임을 표방하면서도 오히려 자료나 실험을 경시하는 문화가 지배적이라는 점이 Sampson이 가진 비판적 접근의 출발점이다.



필자는 위 논문이 출판된 시기인 2000년대 중반이 언어 연구에 있어서 말뭉치 패러다임이 자리를 잡아 가는 과도기였다고 판단한다. 국제적으로도 말뭉치 언어학과 이론 언어학을 접목한 다수의 훌륭한 논문들이 이 시기부터 출현하기 시작하였다. 국내 언어학계 전반도 마찬가지로 당시가 일정 부분 과도기 상태에 있었다고 이해한다. 당시 대학원생이던 필자는 언어학 논문을 학술지에 제출한 뒤 심사평을 받아 본 선배 언어학자들에게서 양립된 경험담을 동시에 들을 수 있었다. 말뭉치 분석을 통해 언어 연구를 했다고 하여 평이 매우 부정적으로 나왔다고 불평하는 논문 투고자와 말뭉치적 접근을 시도했다는 점에서 연구의 참신성을 높이 평가받았으며 안도하는 논문 투고자가 공존하였던 때이다. 그도 그럴 것이 21세기 세종 계획이 제반 사업이 종료되고 그 최종 결과물이 일반 대중에게 공개된 시기가 2007년이다. 이 시기에는 말뭉치를 적극적으로 배우고자 하는 시도도 많았다. 중견 학자들끼리 서로 스터디 그룹을 구성하여 ICE-GB등의 정평이 난 해외 말뭉치를 직접 구동해 보고 연구 토론을 한 일도 몇 차례 있었으며, 말뭉치 사용법에 대한 워크숍을 열면 컴퓨터실 좌석이 부족하여 더 이상 신청자를 받을 수 없는 일도 자주 있었다. 5년여 전부터는 COCA(Corpus of Contemporary American English) 등을 망라한 BYU 말뭉치의 개발자인 Mark Davies나 용례 검색기 AntConc의 개발자 Laurence Anthony 등의 해외 학자가 방문하여 며칠씩 말뭉치 관련 특강을 진행하기도 하였다.

어느새 ‘말뭉치’, ‘코퍼스’, 또는 더 나아가 ‘빅데이터’라는 단어를 학술대회 주제 혹은 제목에 넣는 것이 마치 시대의 요구인 것처럼 자연스러워졌다. 매년 출간되는 학술논문 가운데 있어서도 말뭉치를 일부 혹은 전체적으로 활용한 연구가 최소한 양적으로나마 팽창하였다. 말뭉치를 활용한 언어 연구의 대상도 이전의 몇 개 단어 수준에서 진일보하여, 격종출 구문, 내핵관계절 구문, 이중 타동성 등의 한국어 여러 문법 현상이 말뭉치로 분석되고 있는 요즘이다. 뿐만 아니라 최신의 머신 러닝 기법을 말뭉치에 적용하여 언어 현상에 담긴 제약을 밝히거나 혹은 2차적 산출물을 구성하는 연구도 증가추세에 있다. 이제 더 이상 말뭉치를 사용하거나 그에 기반한 연구를 수행한 것이 연구의 가치를 오히려 훼손하였다고 평가받는 사례는 흔치 않을 것이다. 연구자 개인의 배경이나 관점에 따라 말뭉치의 가치에 대해서 완전히 동의하지 못하는 일은 더러 존재할 수 있겠으나, 말뭉치의 효용성에 대해 예전과 같이 즉각적인 의심을 표명하는 연구자도 별로 없을 것이다.

이러한 흐름에 따라 말뭉치 언어학과 이론 언어학의 접목에 관한 발전 방향을 되짚어 보고자 한다. 최재웅(2014)은 말뭉치와 언어 연구의 관계에 대한 여러 입장을 크게 4가지로 정리하고 있는데 아래 (2)와 같다.

- (2)
- a. 말뭉치는 말뭉치일 뿐 언어 연구와 무관
  - b. 이론 전개에 필요한 예시자료 추출 자원
  - c. 언어적 일반화를 도출하거나 뒷받침하기 위한 자원
  - d. 말뭉치가 곧 언어이론

앞서 살핀 지난 10여 년간의 연구 흐름을 감안하였을 때, 양쪽 극단에 놓인 (2a)와 (2d)를 논외로 하고 (2b~c) 사이에서 학계의 공감대가 일정 정도 형성되었다고 보는 것이 타당할 것이다. 그렇다면 이제 말뭉치 언어학과 이론 언어학의 관계는 올바른 방향으로 정립이 된 상태로 보아도 무방할 것인가? 더 나아가 Sampson(2007)의 문제 제기는 이제 상당 부분 해소된 것으로 보아도 될 것인가?

필자는 말뭉치 언어학과 이론 언어학의 협력관계가 이제 태동기에 접어들었다고 생각한다. 다시 말해, 아직 말뭉치 언어학과 이론 언어학이 함께 가야할 길이 많이 남았다고 생각한다. 이는 아직까지 두 언어 연구의 흐름을 더 발전적인 방향으로 모아줄 여지가 여럿 존재한다는 뜻이다.

이어지는 서술에서는 향후 우리 학계에서 말뭉치 언어학과 이론 언어학이 어떠한 방향으로 더 상호 발전적 모색을 할 수 있을지에 대한 필자의 몇 가지 고민을 담아 보고자 한다. 필자가 평소 연구 과정에서 느낀 바를 정리한 것이기에 다소 두서가 없다는 점에 대해서 독자의 양해를 바란다.

## 2. 말뭉치와 과학적 증명: 재현가능성

2005년 황우석 사태 당시 문제점을 최초로 제보한 류영준 교수는 “왜 제보를 하였는가?”라는 질문에 대해 “과학은 믿는 것이 아니라 증명하는 것이다.”라는 답변을 한 바 있다. 앞서 언급된 Sampson(2007)의 주장도 표현을 달리 하였을 뿐 이 부분에서 문제제기의 시발점을 두고 있다. 비슷한 맥락에서 Ted Pedersen등의 학자는 말뭉치 등의 언어자원의 활용이 과학적 증명이 되기 위한 요건에 대해 10여 년 전부터 아래와 같은 주장을 해 왔다.

- (3) This cuts to the core of whether we are engaged in science, engineering, or theology: Scientists reproduce results; engineers build impressive and enduring artifacts; and theologians muse about what they believe but can't see or prove. (Pedersen, 2008)

당연한 이야기다. 대부분의 현대 과학은 오로지 증명할 수 있는 것만을 연구의 대상으로 삼는다. 그러나 Pedersen(2008)은 연구자가 많은 경우 과학이나 공학이 아닌 신학의 방법론을 사용하고 있음을 지적한다.

언어 연구에서 말뭉치를 사용하는 것이 나름의 가치가 있다는 평가를 받는 것은 자연 언어의 여러 현상을 보다 과학적인 방법을 통해 객관적이고 중립적으로 규명하려는 시도이기 때문이다. 그런데 말뭉치만 사용하였을 뿐 그 과정 및 결과제시가 과학적 증명의 방법론을 택하지 못하였다면 말뭉치가 기여할 수 있는 바를 크게 훼손하는 일이다.

‘과학적 증명’에서 먼저 ‘과학’의 의미를 살펴보자. 현대 과학의 핵심은 연구를 계량화할 수 있어야 한다는 것이

다. 다시 말해 주장하고자 하는 바를 수치적 정보로 설명할 수 있어야 과학이라 칭한다. 최근의 말뭉치 언어학 연구가 빈도, 비율, 강도 등의 계량적 정보를 충실히 제공하고 있다는 점에서 보면 우리는 언어 연구에 있어서 과학적 연구 방법을 일정 정도 채택하고 있다. 그러나 본질적으로 현대 과학이 계량화에 무게를 두는 이유는 무엇일지 생각해 보자. 단순히 자연과학은 수학이 중심이 되는 분야라 그러할까? 숫자로 표시되는 정보가 문장으로 표시되는 정보보다 더 요약적이라 그러할까? 가장 중심이 되는 요인은 ‘재현가능성(replicability)’이다. 입력-처리모형-출력의 단계를 가정하는 계량주의 학문에서는 입력이 동일하고 처리모형이 동일하다면 그 출력까지도 동일해야 한다. 같은 자료를 가지고 복수의 연구자가 실험한 결과가 상이하다면 그 내용은 아직 과학적이라 인정받기 이전의 단계에 놓인다. 다시 말해, 다른 연구자에 의해 재현될 수 있는 연구라야 과학적 증명을 이행한 것으로 간주된다.

다음으로 ‘증명’이라는 표현이 실제로 의미하는 바를 이해해 보도록 하자. 이는 말 그대로 풀어서 ‘증거’에 입각한 연구를 수행하겠다는 뜻이다. ‘증거’가 가장 중요하게 여겨지는 공간은 아마도 법정이 아닐까 한다. 우리나라를 비롯한 거의 모든 현대 문명국가들은 법정증거주의를 채택하고 있다. 아무리 심증적으로 범인이 특정된다 하더라도 결정적 물증이나 증언이 없으면 무죄추정의 원칙에 따라야 한다. 위증이나 증거인멸과 같은 행위가 중형에 처해지는 이유도 법정증거주의의 원칙을 흐르는 범죄이기 때문이다. 우리가 언어 연구를 하는 데 있어서도 이처럼 법정에서 증거가 다루어지는 과정을 충실히 따를 필요가 있다. 증거에 입각한 언어 연구와 증거에 입각한 법적 판결은 본질적으로 동일선상에 놓이기 때문이다. 우선 ‘독수독과론(毒樹毒果論)’이라는 개념이 존재한다. 해당 증거를 입수하는 경로가 적합하지 않다면 판결을 내리는 데 있어 그 증거를 전혀 고려치 않는다는 법원칙이다. 예컨대, 사건 현장에서 경찰이 입수한 증거가 존재한다면 공소장에 그 증거를 입수하게 된 경위가 자세하게 기술되고 사진 등의 첨부자료가 뒤따라야 한다. 언어 연구의 증거제시도 마찬가지여야 한다. 언어 연구에 적용하자면 다음과 같은 내용이 논문에 포함되어야 한다. 아래와 같은 내용이 절차적으로 포함되어야 다른 연구자들이 해당 연구를 재검토하거나 혹은 그 과정 및 결과에 기반하여 후속 연구를 수행할 수 있기 때문이다.

- (4) a. 어떠한 경로를 통해 입수한 어떠한 말뭉치를 사용하였는가
- b. 해당 말뭉치에 대한 전처리 작업은 어떠한가
- c. 주석처리를 어떠한 과정과 원칙을 통해 이루어졌는가
- d. 어떠한 도구와 라이브러리를 통해 분석하였는가
- e. 계수(count) 및 연산(compute)을 수행한 방법은 무엇인가

다음으로 ‘증거보관소’라는 개념이 있다. 증거는 함부로 아무 곳에나 두는 것이 아니다. 재판이 끝난 뒤에도 정해진 장소에 일정 기간 보존을 하게 되어 있다. 무엇보다 판결의 무결성을 보장하기 위함이다. 이는 다시 말해 적법

한 절차를 거친 사람이라면 누구나 그 증거를 열람할 수 있어야 한다는 의미이기도 하다. 이를 언어 연구에 적용해 보면 말뭉치 자체 혹은 그 분석된 결과가 다른 연구자가 확인할 수 있도록 구성되어야 한다는 의미이다. 다른 연구자에 의해서 재현적으로 검토되고 검증될 수 있어야 비로소 증거적 능력을 갖춘 주장이라 할 수 있을 것이다.

결론적으로 말뭉치 언어학이 이론 언어학에 대해서 가지는 가장 큰 함의인 ‘과학적 증명’은 ‘재현가능성’의 다른 이름이다. Ted Pedersen 등의 학자는 이상의 사항에 대해서 실제 실험을 통한 검증까지 실시하였다. 대표적인 예가 Fokkens 외(2013)을 들 수 있다. 이는 2013년도 Association of Computational Linguistics에서 최우수논문상을 수상한 연구로 말뭉치를 비롯한 언어 자원의 활용에 있어서 재현가능성이 보장되지 않으면 어떠한 위험이 따르는가에 대해서 분석한 결과이다.

Pedersen(2008) 및 Fokkens 외(2013)에서 주장하는 내용은 결국 연구 자원의 공개성으로 귀결된다. 연구를 수행하는 과정에서도 가급적 공개적인 자료(open resource)를 사용하고 연구의 결과로 얻어진 산출물도 다른 연구자들에게 공개하는 것이 연구의 재현가능성을 도모하기 위한 가장 기초이다. 해당 연구자 혹은 연구진만이 사용할 수 있고 그래서 다른 연구자들이 원칙적으로 입수할 수 없는 자료를 가지고 언어 연구를 수행하고 이를 논문으로 발표하는 사례가 우리 연구 환경에는 더러 존재한다. 이러한 형태도 분명 자료 기반 연구의 한 종류임에는 분명할 것이나 말뭉치 기반 연구라고는 볼 수 없다. 무엇보다 말뭉치는 사회 구성원이 공유하는 공공재적 성격을 지니기 때문이다. 이러한 주장에 대해 “모든 연구 결과를 무상으로 다 내놓으라는 것이냐?”는 반론도 있을 것이다. 그러나 공개된 자료가 반드시 무상의 자료를 의미하는 것은 아니다. 실제로 Linguistic Data Consortium(<https://www ldc upenn edu/>)에서 배포하는 상당수의 말뭉치는 유료이다. 비유하자면 고속도로는 누구나 이용가능한 공공재이지만 톨게이트 요금을 지불해야 하는 것과 같다.

### 3. 말뭉치와 직관: 불투명 유리창

말뭉치 언어학을 전공하거나 혹은 관심을 가진 독자라면 아래에서 인용된 Fillmore(1992)의 글을 접한 적이 있을 것이다. 이는 말뭉치 언어학의 의의를 설명하는 문구로 말뭉치가 이론 언어학에 대해서 가지는 함의를 너무 맹신하지도 또 너무 경시하지도 말라는 의미를 담고 있다.

- (5) I have two main observations to make. The first is that I don't think there can be any corpora, however large, that contain information about all of the areas of English lexicon and grammar that I want to explore; all that I have seen are inadequate. The second observation is that every corpus

that I've had a chance to examine, however small, has taught me facts that I couldn't imagine finding out about in any other way.

그러나 위 문구는 단순한 Fillmore의 경험적 소회를 넘어 말뭉치가 가지는 수학적 특성에 따른 일종의 정리라고 할 수 있다. 필자는 말뭉치 언어학의 특성을 한 문구로 요약한다면 'reliable, but never perfect'라고 생각하는데, 본질적으로 위 (4)와 같은 이야기이다.

생성문법의 수학적 정의는 무한수의 문장을 만들어 낼 수 있는 유한수의 규칙이다. 이는 인간 언어가 무한 집합( $\infty$ )이라는 점을 기본적으로 전제한다. 집합론적 관점에서 생성문법이 자연언어에 대한 규칙제시법에 관한 것이라 한다면, 말뭉치는 그 규칙에 의해 발현된 무한 집합 가운데 부분 집합을 취한 것이다. 무한 집합의 부분 집합은 그 차수(cardinality)가 아무리 크다고 한들 전체 집합의 아주 일부분일 뿐이다. 따라서 현존하는 언어자원 가운데 가장 큰 규모의 말뭉치라 할 수 있는 English Gigaword (26GB, Parker 외 2011)도 특정 언어(영어)의 모든 현상을 완벽하게 반영할 수 없다.

관련하여 중요한 선행 연구로 언급할 수 있는 것이 Phillips(2009)의 'Armchair Linguists'에 대한 변론이다. 언어 연구자가 말뭉치를 사용할 때 종종 '직관은 믿을만 하지 않고 주관적이어서 연구 결과의 객관성을 해친다.'는 믿음을 품는다. Phillips(2009)는 이러한 생각이 실제로는 단단한 근거에 기초하지 않고 있음을 조목조목 짚어 낸다.

우선 마치 타도의 대상으로 매도되고 있는 '직관에 의거한 수용성 판단'이 실제 언어 연구의 폐단을 낳는 원흉이라고 볼 구체적 근거는 존재치 않는다. Phillips(2009)는 실제 문제가 있다면 이론과 자료의 관계에 있어서 자료를 잘못 해석하는 연구자의 잘못이라는 주장을 하는데 이는 후행 실험 연구에서 실제로 입증된 사항이기도 하다. 대표적인 사례가 Sprouse 외(2013)의 실험통사론 연구로서 영어 통사론 자료를 기준으로 하였을 때 언어학자의 판단과 일반 언중의 판단은 95%이상의 합치율을 보인다는 점을 증명하였다. 후행 연구에서 Linzen & Oseki(2015)는 히브리어와 일본어를 대상으로 마찬가지로 실험을 하였다. 이들 연구는 영어에 비해 검증의 과정을 적었고 따라서 충분히 검토가 되지 못한 언어자료는 일반 화자의 판단과의 합치율이 상당히 떨어진다는 점을 보인다.

다른 한편으로 실험이나 말뭉치 활용 등을 위시한 형식적이고 경험적인 방법의 사용이 직관이 가진 잠재적 문제점을 완전히 해소하는 대안이라는 것도 과도한 주장이다. 인간의 언어는 우리의 뇌속에 존재하고 그 안에서 작동하는 일종의 추상체이다. 현재까지 알려진 어떠한 방법론도 이 추상체에 직접적으로 접근할 수 없다. 심지어 뇌파 분석 등의 최신에 방법론도 언어 행위에 따른 토폴로지(topology)나 스펙트로그램(spectrogram)을 읽고 그에 대한 해석을 하는 것이지 그 자체가 언어 규칙을 바로 설명하지 않는다. 말뭉치도 결국 뇌 속 언어가(I-Language)가 겉으로 발현된 산출물(E-Language)의 집합일 뿐이다. Phillips(2009)의 표현대로 직관과 실험·말뭉치 결과는 공히 인간의 언어를 비추는 불투명 유리창(opaque window)이다.

결론적으로 말뭉치 언어학이 이론 언어학보다 방법론적으로 더 세련되었다거나 또는 말뭉치 언어학이 이론 언어학에 대한 대안이라는 생각은 별 설득력이 없다. 말뭉치를 사용할 때로 이론 언어학에서 자료를 다루는 것과 마찬가지로 신중함을 기해야 한다. 아니 오히려 더 보수적인 접근을 취해야 한다. 자칫 과잉 일반화(overgeneralization)를 범할 위험성이 존재하기 때문이다.

전산 언어학에서 가장 기본적인 평가 척도는 정확도(precision)와 재현율(recall)이다. 전자는 선별된 항목이 틀림없이 선별되었는가를 다루는 1종 오류에 관한 평가인 반면, 후자는 선별되었어야 할 항목이 빠짐없이 선별되었는가를 다루는 2종 오류에 대한 평가이다. 필자는 말뭉치를 이용하여 어떠한 이론적 주장을 보강하고자 할 때 관련된 오류를 범하는 사례를 여러 차례 보았다. 예컨대 말뭉치에서 주장하는 내용을 뒷받침해 줄 수 있는 용례를 찾고자 한다면 이 과정은 비교적 쉽다. 그러나 동시에 말뭉치에서 내가 주장하는 내용을 반박할 수 있는 용례가 또한 존재할 수 있음을 놓쳐서는 안될 것이며, 필자의 경험상 이 검토는 상당히 난해하다. 이러한 연구 형태가 가지는 잠재적 위험성(낮은 재현율)에 대해서 경고하는 대표적인 연구로 Kilgariff(2007)를 들 수 있다. 적지 않은 연구자들이 자신의 주장을 뒷받침하는 문장을 제시하기 위해 구글 등의 검색 엔진을 사용해 보고 그 가운데 대표적인 것을 선택하곤 한다. 일견 타당하고 또 편리해 보이는 이 방식이 때로 연구 결과의 왜곡을 낳기도 하는데, Kilgariff(2007)는 그 왜곡효과를 실제 실험 및 관찰을 통해 증명하였다.

반대의 문제도 발생할 수 있다. 앞서 설명한 것처럼 말뭉치가 아무리 크다고 한들 매개언어의 모든 측면을 다 포괄할 수는 없다. 말뭉치에서 어떠한 표현이 출현했다고 하여 언어 규칙을 함부로 상정할 수 없는 것과 마찬가지로 말뭉치에서 어떠한 표현이 나타나지 않았다고 하더라도 해당 언어 규칙 또한 존재하지 않는다고 볼 수는 없다. 결론적으로 말하자면, 좋은 연구를 하는 것은 말뭉치가 아니라 연구자이다. 말뭉치를 사용하였다고 하여 반드시 더 객관적인 연구가 보장되는 것도 아니며, 모든 언어 연구에 말뭉치가 필요한 것도 아니다. 다만 말뭉치를 사용할 때도 오히려 더 많은 부분에 대한 검토를 기해야 함을 지적하고자 한다.

#### 4. 말뭉치와 이론: 화학적 결합

말뭉치의 올바른 사용은 언어 연구 결과의 객관성 확보를 목표로 하는 장치적 역할을 수행한다. 언어 연구를 건물에 비유하자면 이론 연구는 철근 뼈대며 자료 연구는 그 철근에 덧붙여지는 시멘트이다. 철근으로만 된 집에 사람이 살 수 없고, 시멘트로만 세워진 집이 쉽사리 무너질 수 있는 것처럼 언어 연구가 실효성을 가지기 위해서는 이론과 자료 양자가 모두 필요하다. 양자의 관계 정립에 대한 가장 눈에 띄는 입장은 아래 Fillmore(1992)에서 가져온 인용문으로 (5)에서 바로 이어지는 내용이다.



- (6) My conclusion is that the two kinds of linguists need each other. Or better, that the two kinds of linguists, wherever possible, should exist in the same body.

말줄 친 내용에서 알 수 있듯이 가장 이상적으로는 말뭉치 언어학자와 이론 언어학자가 동일한 사람이어야 한다. 말뭉치 언어학자와 이론 언어학자가 따로 있고 이들이 공동 연구를 수행한다면, 이를 일종의 융합 연구라고 할 수 있을지는 몰라도 결국 ‘물리적 결합’에 그치고 말 것이다. 더 생산적인 결과물은 한명의 연구자가 자신의 이론을 세우고 이를 말뭉치로 입증하거나 혹은 말뭉치를 통해 관찰된 사실을 바탕으로 새로운 이론을 정립할 때 가능하다. 즉, 말뭉치와 이론이 ‘화학적 결합’을 이룰 때 보다 풍성한 언어 연구를 기대할 수 있다.

그렇다면 말뭉치와 이론이 화학적으로 결합을 하는 데 있어 방해가 되는 요소는 무엇일까? 여러 가지가 있을 것이다. 우선 이론이 너무 어렵다는 생각에서 말뭉치 언어학으로 진로를 정한 연구자가 의외로 많다. 또한 말뭉치를 본격적으로 사용하려면 컴퓨터 환경에 친숙해질 필요가 있는데 그러기에는 스스로가 너무 늦었다고 생각하는 연구자도 필자는 많이 보았다. 그러나 이러한 문제들에 앞서 필자는 빈번하게 출현되는 표현(frequently occurring expressions)과 언어학적으로 흥미로운 현상(linguistically interesting phenomena)이 반드시 일치하는 것은 아니라는 점에서 한 가지 원인을 찾고자 한다. 대표적인 예로서 격중출 구문을 들고자 한다. 아마 한국어의 문법 현상 가운데 가장 많은 논문이 나온 주제가 격중출 구문일 것이다. 그러나 실제로 격중출 구문이 말뭉치에서 얼마나 자주 출현할까? 더 정확하게 말하자면 한국어 일반 화자들은 일상 생활에서 얼마나 자주 격중출 구문을 사용할까? 송상현·송지영(2014)에 따르면 세종 구어 말뭉치에서 주격 중출 구문은 전체 문장 가운데 약 0.35%, 대격 중출 구문은 0.06%의 비율로 출현한다. 비슷한 관찰은 강계림(2016)도 보고하였는데 주격 중출 구문은 전체 문장에서 1% 미만이며 문어 자료보다 구어 자료에서 3배 이상 자주 출현한다고 분석하였다. 다시 말해 격중출 구문은 그동안 이론적으로 연구된 바와 달리 실세계에서는 상당히 드물게 사용된다. 이러한 이론적 관심과 실제적 분포 사이의 불균형은 여러 언어 현상에 걸쳐 자주 관찰되는데, 비슷한 예를 영어에서 찾아보면 결속이론(binding theory)에서 큰 관심의 대상이 되어온 재귀사도 출현빈도가 0.1%내외에 불과하다(Song, 2017).

그렇다면 출현 빈도도 높으며 이론 언어학적으로도 관심의 대상이 될 법한 현상(예컨대 관계절이나 어순 뒤섞기)에 대해서만 말뭉치와 이론의 결합 연구를 진행할 것인가? 또는 출현 빈도가 낮은 현상(격중출 구문이나 재귀사)에 대해서는 제한적으로만 이론적 잣대를 적용할 것인가? 이 부분에 대한 해답을 잘 제시하고 있는 선행 연구로는 Flickinger & Wasow(2013)을 들 수 있다. 이 말뭉치 연구는 영어의 이른바 do-be 구문을 다루고 있는데 말뭉치에서 추출한 구체적인 용례는 아래 (7)과 같다. 순서상의 선행적 구성만을 본다면 본동사로서의 do 또는 그 활용형이 사용되며 연달아 계사가 출현한 다음 앞의 do동사와 같은 활용형을 취한 일반 동사가 바로 사용되는 형태이다.

- (7) a. what you have to do is get ready  
b. all the government does is send out checks  
c. the thing I'm doing is trying to learn from my mistakes  
d. the best one can do is compare one risk to the next

언뜻 보기에도 상당히 특수한 형태로 이루어진 영어 구문인데, 실제 위와 같은 구성은 4억 2천5백만 단어 규모의 COCA에서 단 6,471회 출현하는 것으로 분석되었다. 즉, 극단적인 저빈도 구문에 해당한다. Flickinger & Wasow(2013)가 이 구문에 말뭉치 분석을 통해 입증하고자 하는 바는 말뭉치에서 적게 출현하고 그래서 경우에 따라 주변적이거나 무시할 만한 수준인 것으로 여겨지는 대상이 오히려 중요한 언어학적 단서를 제시하기도 한다는 점이다. 머신 러닝의 관점에서 말뭉치를 학습 자료로 사용할 때에는 위와 같은 형식의 자료는 일종의 이상치로 간주되어 버려질 것이다. 그러나 이론 언어학적 관점에서는 위와 같은 자료가 기존의 이론을 검토하고 이전에 미처 발견되지 않았던 새로운 제약을 발견해 내는 데 있어서 때로 중요한 기여를 한다. 다시 말해, Flickinger & Wasow(2013)의 연구는 (5)에서 제시된 Fillmore(1992)의 주장을 실증적으로 밝힌 것이다.

결론적으로 말뭉치와 언어 이론을 접목하여 언어 연구를 더욱 발전시키는 데 있어 빈도나 규모 등의 분포적 특성은 큰 제약이 되지 않는다. 연구자 개인의 진취적 자세에 따라서 얼마든 양자를 충실히 결합한 훌륭한 연구를 도모할 수 있다.

## 5. 말뭉치와 도구: 언어 자원

서두에서 언급한 바와 같이 한동안 말뭉치에 관한 워크숍이 많이 열린 시기가 있었다. 이전의 말뭉치 워크숍은 주로 말뭉치를 분석하는 도구 다시 말해 소프트웨어 작동법에 대한 교육이었다고 해도 과언이 아니다. 그 영향에 서인지 지금도 많은 연구자들이 말뭉치 언어학을 한다는 것을 때로 WordSmith나 AntConc와 같은 분석 프로그램을 쓰는 것과 동의어처럼 받아들이는 경향이 있다. 필자의 주견으로는 이는 별로 바람직하지 않은 접근법이다. 무엇보다 ‘도구는 사고의 한계이다.’라는 격률이 함의하는 것처럼 연구의 수단이 특정 프로그램을 중심으로 고착화되면 연구의 창의성이 가로막힐뿐더러 연구자 자신도 타성에 젖어들기 십상이다. 또한 앞서 논의한 바처럼 이론 언어학과 다소 괴리된 자료 중심 연구로 매몰될 가능성도 존재한다. 말뭉치는 그 자체로 언어 연구의 도구이다. 그런데 프로그램을 도구로 인식하고 말뭉치를 그 도구를 통해서 이해해야 하는 대상으로 받아들이면 이론은 어느 자리에 서야 하는가? 언어 이론이 우리가 검증하고 정제해야 할 대상이며, 말뭉치는 그 과정에서 활용되는 도구라고 정리하는 것이 미래지향적이다.



‘말뭉치 자체가 도구이다.’라는 명제를 조금 더 정확히 풀이하자면 이는 말뭉치를 일종의 자원으로 파악하자는 것이다. 흔히 자원이라 하면 우리는 예컨대 원유, 철광석, 목재 등을 떠올릴 것이다. 이들을 가지고는 당장 아무 것도 할 수 없다. 그러나 어떠한 손질을 거치고 나면 그때부터는 많은 일이 가능해진다. 실제 표준국어대사전은 자원을 ‘인간 생활 및 경제 생산에 이용되는 원료’라 정의한다. 즉, 자원은 그 자체로서는 별다른 가치가 없는 원재료이나 이를 가공하면 다른 가치를 크게 창출할 수 있는 잠재적 재화를 말한다. 말뭉치도 언어 연구에서 이와 같은 위치를 지닌다. 그렇다면 자원으로서 말뭉치가 가진 잠재적 가치를 확대하기 위해서는 어떠한 대안이 있을까? 필자는 크게 아래와 같은 항목을 제안하고 싶다.

첫째, 말뭉치에 주석처리를 하는 노하우에 대한 전파와 교육이 필요하다. 원유를 가공하는 과정이 정유라면 말뭉치를 가공하는 과정은 주석 처리이다. 많은 연구자들은 배포된 형태의 말뭉치가 그 상태로 가공이 끝난 완제품처럼 여기곤 한다. 연구자 개인이 별도의 주석 처리를 더 하지 않고 완제품으로서의 말뭉치만 사용한다면 비유컨대 이는 통조림이나 컵라면과 같은 인스턴트 식품으로만 끼니를 해결하는 것과 같다. 말뭉치라는 식재료를 다루는 방법에 대한 요리책(cookbook) 형식의 지도가 필요하다.

둘째, 다종다양한 말뭉치를 언어 연구에 활용하는 사례가 더 축적되어야 한다. ‘21세기 세종계획’ 결과물을 예로 하자면, 대부분의 관련 연구는 형태분석 말뭉치에 집중되어 있다. 몇 개의 대학에서 구축했거나 구축하고 있는 말뭉치도 비슷한 상황인데 문어 형태분석 말뭉치나 구어 말뭉치 일부가 전체 말뭉치 연구의 대다수를 차지한다. 말뭉치도 말뭉치마다의 목적과 쓰임이 있다. 예컨대 문법관계나 어휘의 구조적 제약 등에 대한 언어 연구를 수행하고자 한다면 구문분석 말뭉치(treebank)에 대한 참조가 필수적이다. 비교 언어학적 관점에서 언어 연구를 하고자 한다면 병렬 말뭉치를 사용하여야 하고, 언어 교육 및 습득 연구는 학습자 말뭉치를 참조하여야 한다. 상대적으로 이들 말뭉치의 활용 노하우 및 선행 연구는 축적된 바가 적다. 이후에 말뭉치 워크숍이 더 열린다면, 보다 세분화된 측면에서 말뭉치를 활용하는 데 역점을 두었으면 한다.

셋째, 말뭉치 언어학은 이제 자매 경험주의 언어 연구 방법론인 언어 실험과 연계를 모색할 필요가 있다. 다른 과학 분야에서는 실제 자료를 수합한 관찰과 통제된 환경에서의 실험을 종합하여 결론을 도출하는 방법론이 일반적이다. 공학계열에서 어떠한 모형을 평가하는 관점도 위와 같이 이원적이다. 실험실 안에서 진행되는 내현적 평가(in vitro)와 실험실 밖의 실제 환경에서 관찰하는 외현적 평가(in vivo)가 양대 지표이다. 언어 연구도 이제 이러한 방향으로 나아갈 필요가 존재한다. 이러한 요구는 이미 국제적으로도 하나의 추세로 자리 잡아 가는 분위기이다. 전통적으로 실험을 중시해 온 심리언어학 그룹에서 말뭉치를 사용하여 자신들의 연구를 보강하는 사례뿐만 아니라 (Jaeger, 2011), 역으로 말뭉치 언어학을 주로 연구해 온 그룹이 언어 실험을 적극적으로 고려하는 사례도 속속 보고되고 있다(Gries & Kootstra, 2017). 향후 언어 연구에서 새로운 통찰력은 이러한 접근법에서 파생될 가능성이 높다는 것이 필자의 견해이다.

## 참고문헌

- 강계림. 2016. 주격 조사 ‘가’ 중출 구문 연구 – 구어 문어 말뭉치 분석을 바탕으로 -. 언어과학 23(1): 1–30.
- 송상헌 · 송지영. 2014. 세종 구어 말뭉치 기반 격표지 중출 자료 구축. 언어정보 19: 57–90.
- 최재웅. 2014. 말뭉치와 언어 연구 – 외국의 사례와 경향 -. 한국어학 63: 71–102.
- Fillmore, Charles J. 1992. “Corpus Linguistics” or “Computer-aided Armchair Linguistics.” In Jan Svartvik (Ed.) Directions in Corpus Linguistics, pp. 35–60. Berlin/New York: Mouton de Gruyter.
- Flickinger, Dan, and Thomas Wasow. 2013. A Corpus-Driven Analysis of the Do-Be Construction. In Philip Hofmeister and Elisabeth Norcliffe (Eds.) The Core and the Periphery: Standing on the Shoulders of Ivan A. Sag, pp. 35–63. Stanford, CA: CSLI Publications.
- Fokkens, Antske, Marieke Van Erp, Marten Postma, Ted Pedersen, Piek Vossen, and Nuno Freire. 2013. Offspring from Reproduction Problems: What Replication Failure Teaches Us. In Proceedings of ACL 2013, pp. 1691–1701.
- Gries, Stefan Th, and Gerrit Jan Kootstra. 2017. Structural Priming within and across Languages: a Corpus-based Perspective. Bilingualism: Language and Cognition 20(2): 235–250.
- Jaeger, T. Florian. 2011. Corpus-based Research on Language Production: Information Density and Reducible Subject Relatives. In Emily M. Bender and Jennifer E. Arnold (Eds.) Language from a Cognitive Perspective: Grammar, Usage, and Processing. Studies in honor of Tom Wasow, pp. 161–197. Stanford, CA: CSLI Publications.
- Kilgariff, Adam. 2007. Googleology is Bad Science. Computational Linguistics 33(1): 147–151.
- Linzen, Tal and Yohei Oseki. 2015. The Reliability of Acceptability Judgments across Languages. Unpublished Manuscript.
- Parker, Robert, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2011. English Gigaword. Linguistic Data Consortium.
- Pedersen, Ted. 2008. Empiricism is Not a Matter of Faith. Computational Linguistics 34(3): 465–470.
- Phillips, Colin. 2009. Should We Impeach Armchair Linguistics?. In Shoishi Iwasaki, Hajime Hoji, Patricia M. Clancy, and Sung-Ock Sohn (eds.), Japanese/Korean Linguistics, Vol. 17, pp. 49–64. Palo Alto, CA: CSLI Publications.
- Sampson, Geoffrey R. 2007. Grammar without Grammaticality. Corpus Linguistics and Linguistic Theory 3(1): 1–32.
- Song, Sanghoun. 2017. A Corpus Study of Unbound Reflexive Pronouns in English. Korean Journal of English Language and Linguistics 17(2): 275–305.
- Sprouse, Jon, Carson T. Schütze, and Diogo Almeida. 2013. A Comparison of Informal and Formal Acceptability Judgments Using a Random Sample from Linguistic Inquiry 2001–2010. Lingua 134: 219–248.

주제 3

# 말뭉치 언어학과 이론 언어학, 사전 편찬

· 토론자 **최정도**(국립국어원 언어정보과 학예연구사)



전체적으로 발표의 내용에 동의하는 바이며, 몇 가지 연구와 관련된 질의를 하는 것으로 토론을 대신하고자 한다.

(1) 말뭉치의 재현 가능성의 측면에서는 원 자료인 말뭉치의 배포나 공유도 중요하지만, 연구자의 결과물에 대한 검증 또한 중요하다고 생각된다. 왜냐하면 평가자가 확인할 수 있는 것은 오직 연구 논문에서 제시하고 있는 표나 그래프가 전부이기 때문이다. 심사자나 독자의 입장에서는 연구자가 제시하는 정보 이외에는 접할 수 있는 것이 없기 때문에, 연구자의 연구 결과를 평가하거나 재현해 보는 것이 거의 불가능하다고 할 수 있다. 한국에서는 아직 이러한 연구 문화가 정착되고 있지 않은데, 외국의 사례는 어떠한지 연구 결과물의 배포나 공유에 대해서는 어떻게 생각하시는지 여쭙고 싶다.

(2) 말뭉치상에서 잘 나타나지 않는 (이론적으로 가능한) 언어 현상(또는 형태)은 인위적으로나마 그런 언어 현상(또는 형태)이 나타날 수 있는 텍스트나 발화를 담지 않는다면 연구가 힘들다. 말뭉치를 이용한 사전편찬(중사전 이상의 규모)에서도 저빈도 표제어의 출현을 위해서도 말뭉치의 양을 늘릴 수밖에 없을 것이다. 하지만 최근에는 말뭉치가 개인의 역량으로 다루기 힘들 정도로 많다는 것이 더 문제가 될 수도 있는 환경이 되었다(세종 말뭉치, 포털 검색 등). 이렇게 말뭉치의 양이 늘어나면 늘어날수록 반대로 일반적인 언어 현상(하다, 있다, -어/아 등)의 연구는 질적 분석이 오히려 더 힘들어지기도 한다. 이러한 측면에서 특정 언어 현상을 연구하기 위한 언어 집합(말뭉치)을 구성하는 방법에 대해서 발표자의 의견을 여쭙고 싶다.

(3) 이론적으로는 가능할 듯 보이지만(어떠한 패턴에 사용될 가능성) 실제로는 쓰이지 않는 문장들에 대한 연구는 다분히 이론 언어학의 영역으로 두어야 하는 것인지 의문이 든다. 발표자의 논문(2014)에서도 제시되었던 예문인데, 이러한 언어 현상은 말뭉치를 배로 구축한다 하어도 나타나지 않을 가능성이 아주 높은 것으로 생각된다.

예1) 동물이 포유류가 코끼리가 코가 가운데가 길다.

예2) 꽃이 장미가 백장미가 꽃잎이 색이 예쁘다. (박은영, 2009:15)

자료(말뭉치)를 이용한 언어 연구의 입장에서 이와 같은 언어 현상을 어떻게 받아들여야 할지에 대해서, 발표자의 의견을 여쭙고 싶다.

(4) 최근 '이론으로서의 말뭉치 언어학'에 대한 연구가 속속 나타나는 것으로 보인다. 방법론으로서의 말뭉치 언어학을 넘어 이론으로서의 말뭉치 언어학이 가능하다고 생각하시는지, 이에 대한 발표자의 고견을 듣고 싶다.

주제 4

# 기계 번역은 우리 생활을 어떻게 변화시킬 것인가?

· 발표자 김준석(네이버 파파고 팀 리더)





## 기계 번역은 우리 생활을 어떻게 변화시킬 것인가?

—

김준석

Leader / Papago / Naver

2017-11-10



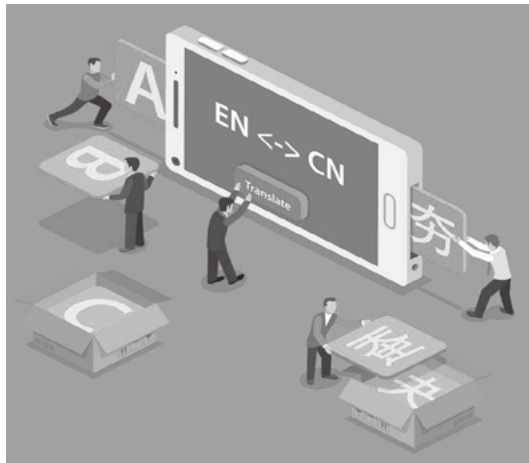
© NAVER Corp.

## 기계 번역 기술 동향

papago

2

## 기계 번역이란?



출처: <http://www.transperfect.com/blog/machine-translation-solution-you-can-bank-on>

papago

3

## 왜 기계 번역 기술이 주목을 받을까?



papago

5

## 온라인 설문조사: 소비자에게 가장 필요한 기술 ?

5월 시장조사기업 엠브레인 트렌드모니터에 따르면 지난 2월 22~28일 스마트폰을 사용하는 전국 만 15~59세 1천명을 대상으로 온라인 설문조사를 한 결과 대다수가 혁신 기술의 필요성에 공감하면서도 부작용에 대한 우려를 드러냈다.

일반 소비자에게 가장 필요한 기술로는 자동 통번역 기술이 꼽혔고, 지능형 자율주행차와 사물인터넷이 뒤를 이었다.

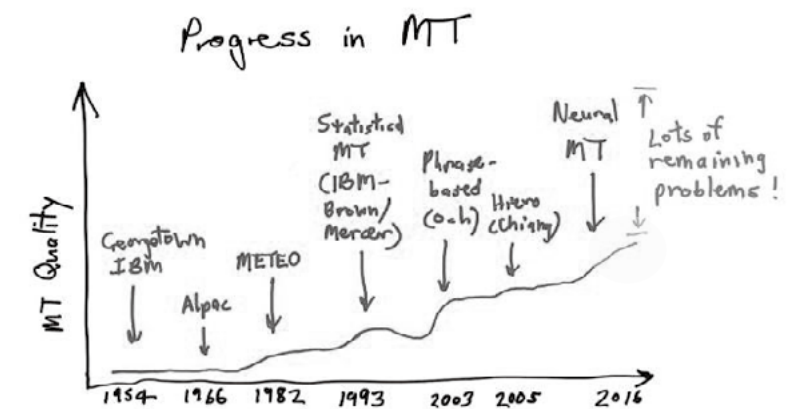


2017-04-05 연합뉴스

papago

4

## 작년부터 좋아진 기계 번역 품질

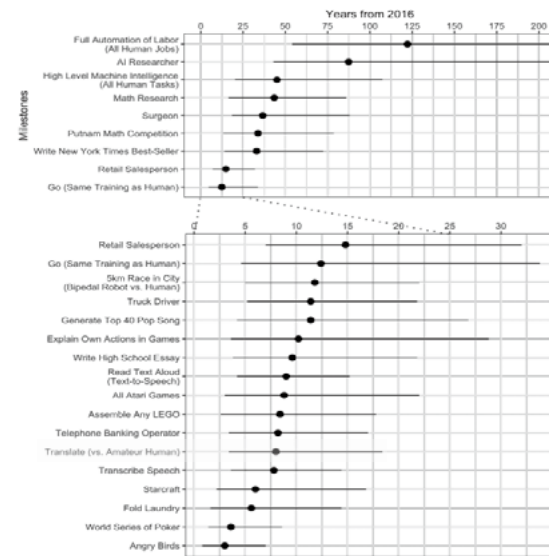


출처: slide by Christopher D. Manning

papago

6

## AI학회 설문조사: 기계 번역 8년 후 인간 수준



papago

출처: <https://arxiv.org/pdf/1705.08807.pdf>

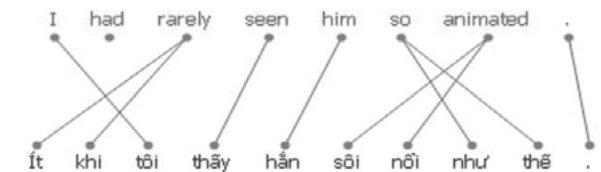
7

## 통계적 기계 번역

### parallel corpus

网站资讯分析网数  
据显示的主域名为  
全世界访问量最高  
的站点除此之外搜  
索在其他国家或地  
区域名下的多个站  
点等等及旗下的等

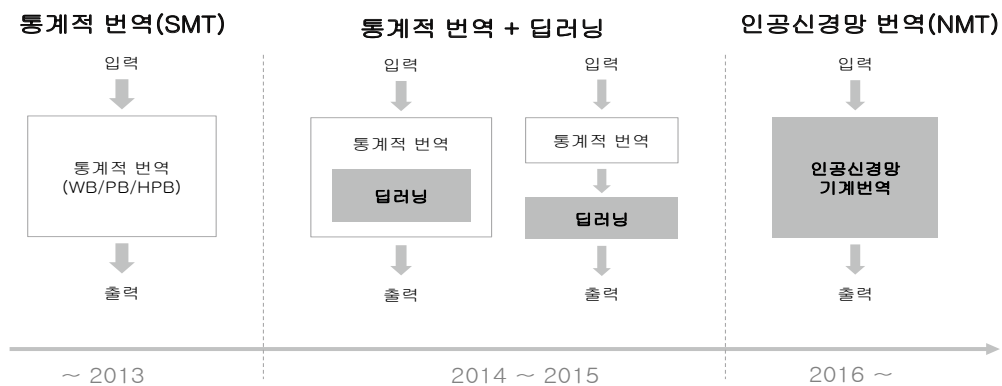
The corporation has been estim  
to run more than one million pag  
in data centers around the world  
to process over one billion searc  
requests and about twenty-four i  
of user-generated data each dat  
December 2012 Alexa listed as



papago

9

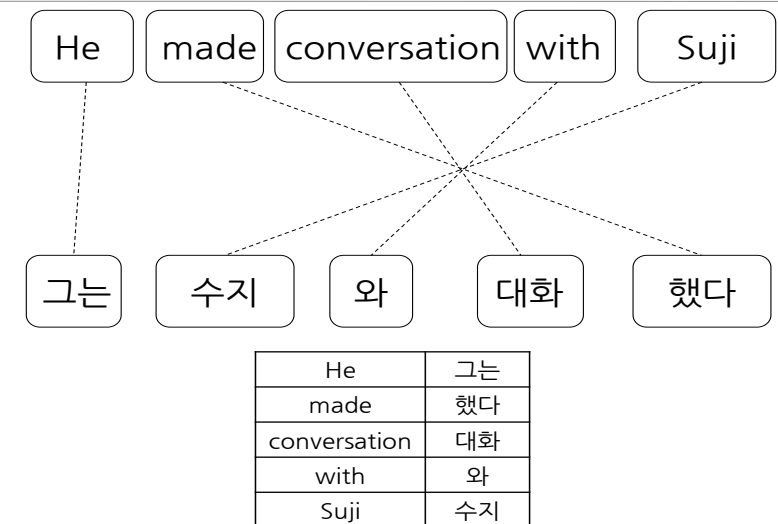
## 기계 번역 기술 트렌드 변화 과정



papago

8

## 단어 기반의 통계적 기계번역(Word-based SMT)



papago

10

He	그는
made conversation	대화했다
with Suji	수지와

11

X <sub>1</sub> made conversation with X <sub>2</sub>	X <sub>1</sub> X <sub>2</sub> 와 대화했다
--	--------------------------------------

12

er	gent	ja	nicht	nach	hause
he	is	yes	not	after	house
it	are	is	do not	to	home
it	goes	, of course	does not	according to	chamber
he	go	,	is not	in	at home
it is		not		home	
he will be		is not		under house	
it goes		does not		return home	
he goes		do not		do not	
	is		to		
	are		following		
	is after all		not after		
	does		not to		
	not				
	is not				
	are not				
	is not a				

13

The diagram illustrates a sequence-to-sequence model. The input sequence "I am a student" is fed into an encoder (represented by gray boxes). The encoder processes the input tokens sequentially, producing a hidden state (represented by a white box). This hidden state is then used by the decoder (represented by white boxes) to generate the output sequence "나는 학생이다 <eos>". The decoder also takes the previous output token as input for the next step.

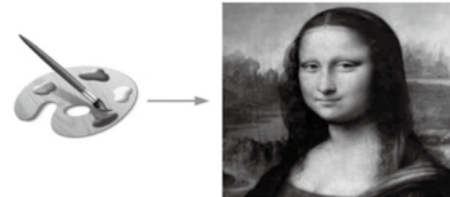
14



## 통계적 방식과 인공지능망 기계번역의 비교



이산적(Discrete)  
지역적 결정(Local Decision)



연속적(Continuous)  
전체적 결정(Global Decision)

papago

15

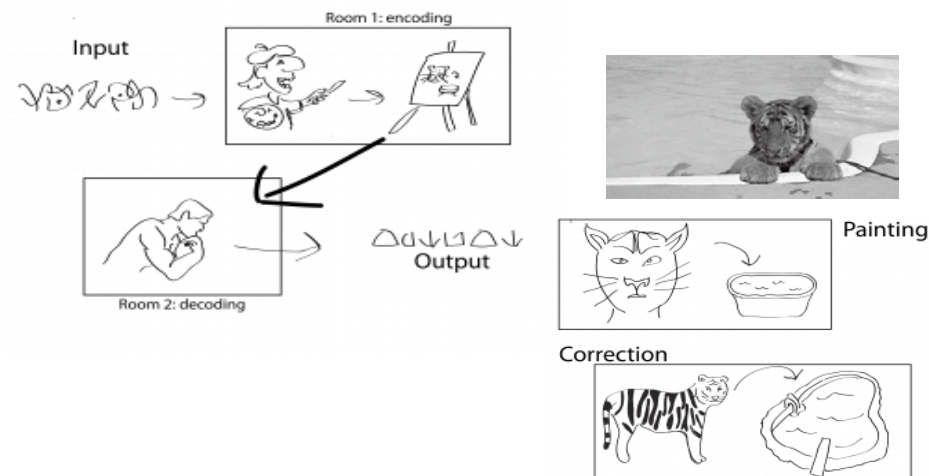
## 인공지능망 기계번역의 주요 연구 주제들

- ✓ 긴 문장 번역을 위한 어텐션 모델(Attention Mechanism)
- ✓ 번역 단위(Tokenization)
- ✓ 인공지능망 구조(Multi-layer Architecture)
- ✓ 다국어 모델 학습 방법(Zero-shot Learning)
- ✓ 새로운 번역 방식들(CNN-based NMT, Transformer, ...)

papago

17

## 인공지능망 기계 번역 기술에 대한 비유



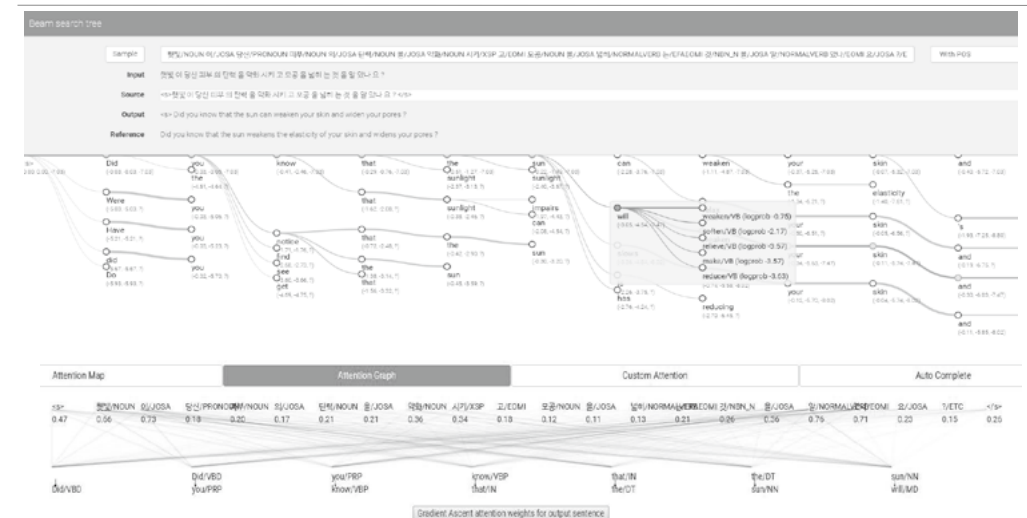
papago

출처: <http://www.pinchofintelligence.com/explaining-googles-zero-shot-translation/>

16

## 인공지능망 기계번역 시각화 도구

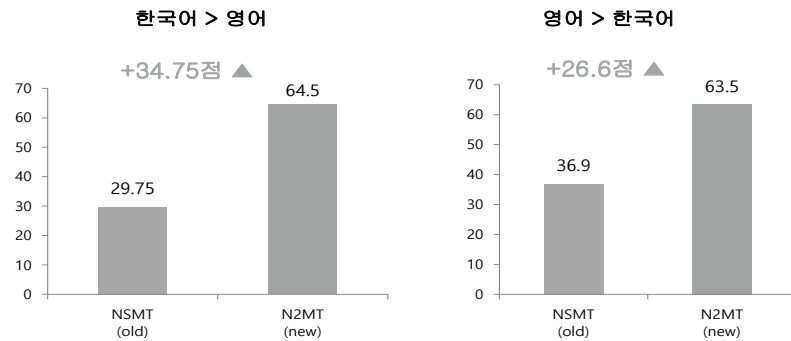
NMT Workshop 2017  
EMNLP 2017  
DEVIEW 2017



papago

18

## 통계적 방식과 인공신경망 방식의 번역 품질 비교 결과



papago

19

## 우리 생활 속의 기계 번역

papago

21

## 정리

- ✓ 기계번역 기술은 다양한 인공지능 기술들 중에서도 가장 필요한 기술
- ✓ 통계적 번역에서 인공신경망 번역으로 번역 방식의 진화
- ✓ 단어/구 기반이 아니라 인공신경망 번역은 문장 전체의 정보 활용
- ✓ 정확한 번역을 위한 다양한 인공신경망 기술들이 빠른 속도로 연구되고 있음
- ✓ 인공신경망 번역의 해석과 디버깅을 위해 시각화 도구의 중요성이 높아짐

papago

20

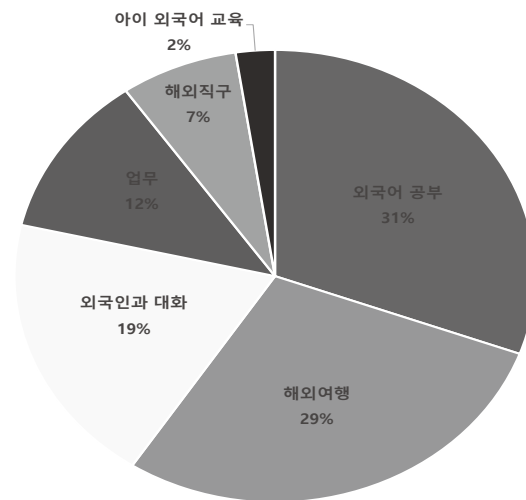
## 기계 번역 서비스는 우리 생활 어디에서 활용되고 있을까?



papago

22

## 설문조사 결과: 번역기를 어디에 사용하는가?



papago

23

## 번역기 사용자 경험들 (2/2)

- ✓ 공항 리무진버스 기사입니다. 다양한 나라의 관광객들이 많이 오는 요즘엔 정신없이 바쁩니다. 하루는 평소와 같이 일을 하고 있는 중에 한 외국인이 급하다는 표정을 짓고 뭐라 말을 하는데 도대체 알아들을 수가 없어 파파고를 이용한 적이 있습니다. 얼마나 고맙던지
- ✓ 24시간 셀프 빨래방 운영자입니다. 근방 호텔에 있는 여행객들이 셀프 빨래방을 자주 찾아오는데, 파파고를 통해 사용 방법도 알려드리고 소통할 수 있어서 너무 좋았습니다. 이제 외국인 손님 두렵지 않아요~
- ✓ 전담 만들 때 일본어로 되어 있어서 알기 어려웠는데 파파고 덕분에 많이 알게 되었음
- ✓ TV를 보다가 외국인들이 말하는 것을 그 자리에서 파파고를 통해 뜻을 찾아보기도 하고 광고에 영어 나오면 그 자리에서 바로 찾아볼 수 있어서 최고의 앵무새이다.

papago

25

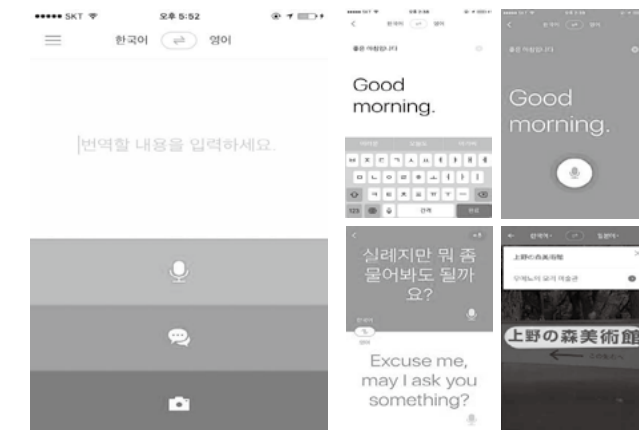
## 번역기 사용자 경험들 (1/2)

- ✓ 간호사로 일하면서 외국인 환자(중국인)를 만났을 때 이용한 적 있어요. 덕분에 업무에 많은 도움 되었습니다!!
- ✓ 갑자기 방문한 일본바이어와의 만남이었어! 급작스러운 방문인 탓에 준비된 게 아무것도 없었어. 너무너무 당황했지만 파파고가 있어서 바이어와의 식사와 티타임, 식사 중간중간 나왔던 업무에 대한 모든 이야기를 깔끔하게 번역해 주었어!! 어찌나 신기하고 고맙던지.
- ✓ 경찰 업무 중 외국인에게 길 안내 서비스를 할 때 이용해서 보람되었다.
- ✓ GS25 근무 중에 외국인이 왔었는데, GS25 대화기능으로 더 친절히 응대할 수 있어서 좋았어요.
- ✓ 게스트하우스를 운영하는 입장에서 외국인과 대화할 때 아주 유용합니다.

papago

24

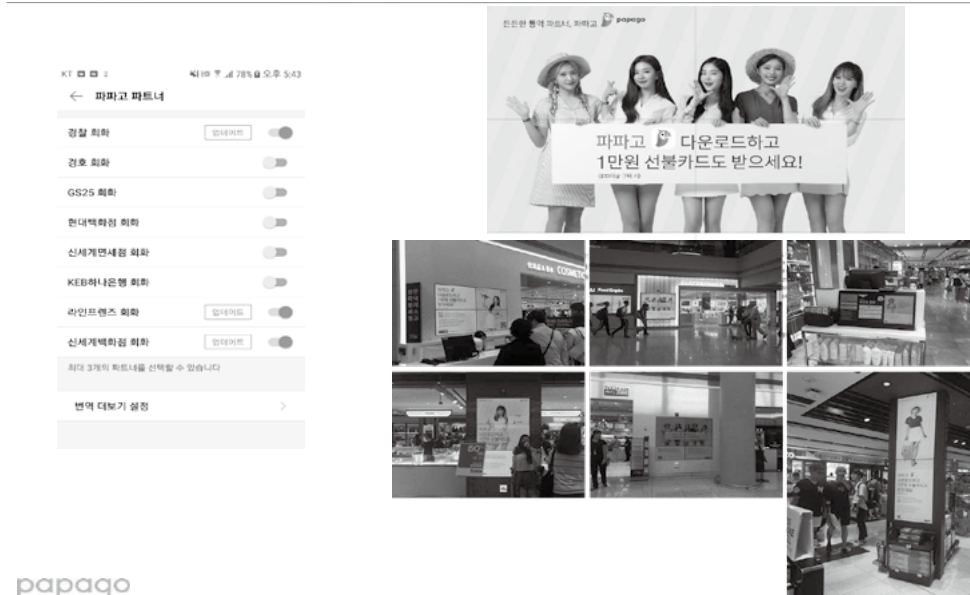
## 파파고 (Papago)



papago

26

## 파파고 파트너



파파고 파트너

경실 회화, 경호 회화, GS25 회화, 현대백화점 회화, 신세계백화점 회화, KEB하나은행 회화, 라인프렌즈 회화, 신세계백화점 회화

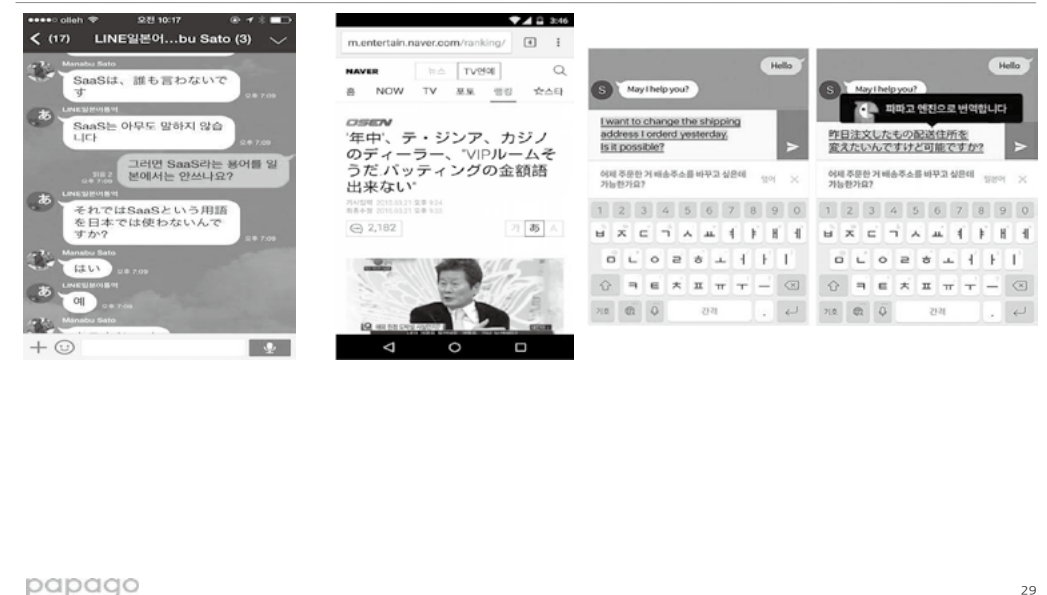
최대 3개의 파트너를 선택할 수 있습니다

번역 대보가 쏠쏠

파파고 다운로드하고 1만원 선불카드도 받으세요!

27

## 네이버/라인의 기계 번역 서비스 (1/2)



네이버/라인의 기계 번역 서비스 (1/2)

네이버: SaaS는, 誰も言わないです. SaaS는 아무도 말하지 않습니다. 그러면 SaaS라는 용어를 일본에서는 안쓰나요? 그러면 SaaS라는 용어를 일본에서는 안쓰나요?

라인: May I help you? Hello. I want to change the shipping address I ordered yesterday. Is it possible? 어제 주문한 저 배송주소를 바꾸고 싶는데 가능한가요? 어제 주문한 저 배송주소를 바꾸고 싶는데 가능한가요?

29

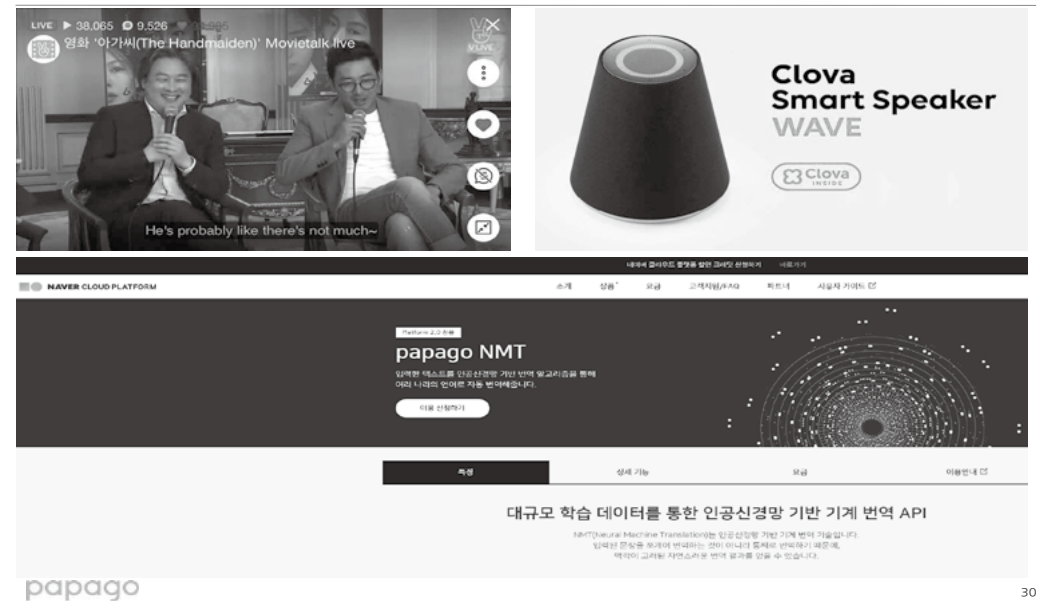
## 현재 파파고 파트너들



서울지방경찰청, GS25, HYUNDAI DEPARTMENT STORE, LINE FRIENDS, 대한민국의용엔지니어링, KEB 하나은행, SHINSEGAE DUTY FREE, KLOOK

28

## 네이버/라인의 기계 번역 서비스 (2/2)



네이버/라인의 기계 번역 서비스 (2/2)

네이버: 영화 '아가씨(The Handmaiden)' Movietalk live. He's probably like there's not much~

라인: Clova Smart Speaker WAVE. Clova INBOX.

네이버 클라우드 플랫폼: papago NMT. 대규모 학습 데이터를 통한 인공지능 기반 기계 번역 API.

30



## 다른 기업들의 기계 번역이 적용된 서비스 사례들



31

## 하드웨어와 결합된 기계 번역



33

## 교육과 결합된 기계 번역



papago

32

## 정리

- ✓ 인공지능망 번역으로 품질이 향상되면서 우리 생활 곳곳에서 활용되고 있는 상황
- ✓ 번역은 서비스 자체로도 의미를 가지지만, 다른 서비스의 가치를 높이는 부품
- ✓ 다양한 하드웨어와 결합된 제품이 나오는 시점
- ✓ 향후 교육, 관광, 쇼핑 등 더 많은 영역에서 사용될 것으로 전망

papago

34

주제 4

# 기계 번역은 우리 생활을 어떻게 변화시킬 것인가?

· 토론자 **정호정**(한국외국어대학교 영어통번역학부 교수)



## I.

발표자료에서 인용한 2017년 2월 여론조사 결과와 같이, 잠재사용자의 기대와 우려를 함께 모으고 있는 기계번역의 발전 현황을 실생활과의 연관지어 다룬 시의적절한 논문이라고 사료됨

발표자료를 보면 발표내용이 기계번역의 발전사와 현황을 다룬 전반과 네이버 파파고의 생활밀착형 사용 실례를 다룬 후반으로 크게 나누어 짐. 전반에서 NMT 모델에 이르기까지의 발전 변천사를 요약하면서 몇가지 주요개념을 도전 겸 과제로 정리하고 있는데, 후반의 주요내용이 될 네이버 파파고의 현재 상황과 과제를 이 개념들과 관련하여 설명하면 보다 전체적이고 객관적인 이해와 평가가 가능할 것임

현재 슬라이드에 제공되어 있는 자료만으로는 제시되어 있는 정보의 신뢰성이나 효용을 판단하기 어려운 정보들이 포함되어 있음. 예를 들어 “작년부터 좋아진 번역품질” 슬라이드의 출처인 C. Manning의 그래프나 네이버 NMT의 향상된 점수 평가 결과 등은 추가적으로 그 근거나 평가기준 사용된 척도 등이 함께 제시되어야 이해가 가능할 것인데, 이는 발표에서 다루어질 것으로 기대됨. 한국어·영어로의 번역 점수가 영어·한국어로의 번역점수를 상회하는 현상이 흥미로운데 이에 대한 설명이 추가적으로 제시되면 흥미로울 것임

일반적인 어휘나 어구 등의 처리 이외에 겸양법/공손어법, 강조 반복 도치 등과 같은 수사학적 기재의 처리, 한 언어권에 서만 존재하는 문화적 현상이나 틀(frame) 등을 반영하는 계산된 표현들의 처리 등에 대한 접근법이나 대안이 있으면 소개 바람. 또 기계번역 툴의 개발의 목적이 그 자체로의 효용도 있지만 다른 서비스의 가치를 높이기 위한 수단으로서의 가치도 크다고 명시하고 있는데, 그러다 보니 문장단위의 번역에 우선 치중하고 있는 것으로 판단됨. 길이에 상관없이 텍스트 차원의 번역은 질적으로 다른 차원의 문제이자 도전이 될 수 있을 것인데 이에 대한 향후 전략이나 대응은 어떻게 예상하고 있는지?

“기계번역 8년후 인간수준”이라는 슬라이드에서 AI 전문가들을 대상으로 한 설문조사 결과를 제시하고 있는데, 이는 번역 대상언어 양쪽을 모두 구사할 수 있는 이중언어구사자의 수준으로의 번역이 가능할 것으로 예측한다는 것임. 이런 전망을 토대로 향후 번역의 직업적 변화와 전망은 어떻게 판단하고 있는지에 대한 전문가 의견을 구함

## II. 네이버 파파고 프로그램을 이용한 기계번역 현황과 문제점

The image displays two screenshots of the Papago web interface. The top screenshot shows a translation from English to Korean. The English input is "How can I explain the fact that I hate myself most?". The Korean output is "내가 가장 싫어하는 사실을 어떻게 설명할 수 있을까?". The bottom screenshot shows a translation from Korean to English. The Korean input is "내가 제일 싫어하는 건". The English output is "it is i that i hateth the most". Below this, a second Korean input "내가 가장 싫어하는 것은 내가 가장 싫어하는 것이라고 설명할 수 있다." is shown, with the English output "how can you explain that i am the one that i hate the most".

## 주제 4 기계 번역은 우리 생활을 어떻게 변화시킬 것인가?

영어 감지 → 한국어

open the window  
this room feels stuffy  
would you mind opening the window cus i feel stuffy in this room  
would you mind opening the window cus this room feels stuffy  
you might want to open the window cus this room feels stuffy

창문을 열어요.  
이 방은 답답하다.  
창문을 여는 것이 번거로운데 방 안에 답답함이 느껴진다.  
창문을 여는 것이 답답하니 창문을 열어 주시겠습니까?  
너는 창문을 열고 싶어 할지도 몰라. 이 방은 답답한 느낌이 들어.

papago

영어 감지 → 한국어

that I've wanted most in my entire life  
that, I've wanted most in my entire life  
hard to believe, almost surreal, the most beautiful girl stood before my eyes

내가 평생 동안 원했던 건  
전 평생 내내 원했던 거예요  
믿기 어려울 정도로, 거의 꿈같은 소녀가 내 눈 앞에 서 있었  
다.

## 주제 5

# 우리말 자연 언어 처리 기술의 전망

· 발표자 **나승훈**(전북대학교 컴퓨터공학과 교수)

### III. 현재 상용화되어 있는 대표적인 기계번역 프로그램의 현황과 문제 (네이버 파파고 포함)

	A	B	C
한 시대를 풍미한 김삿갓은 오늘날 래퍼의 시조라고 할 수 있다	One era, include Kim, can be said to be the city of today's rapper	Gimsatgat pungmihan an era may be called the founder of today's rappers	Kim Satgat, one of the epochs of the age, is now the founder of the modern-day rapper.
한 시대를 풍미한 김삿갓 김병연은 오늘날 래퍼의 시조라고 할 수 있다	In one era, include Kim, the fresh Kim, can be said to be the city of today's rapper	Kim Byeong-yun, who has a taste for one era, can be said to be the founder of rapper today	Kim Sang-yeol, a one- time generation of Kim Sang-gak, is now the founder of the modern- day rapper.
김 대리, 우리 재무팀 팀장으로 어때?	Kim, representative, what about us as a financial team leader?	How about you, Kim Dear, our finance team leader?	Agent Kim, what about our financial team manager?
노래 하나는 잘 하지요	One song is fine.	One song is good.	I'm good at singing.





## 우리말 자연언어처리 기술의 전망

2017.11.10

나승훈  
전북대학교

### 발표 내용

- 자연언어처리 개관 및 기술 흐름도
- 딥러닝기반 언어 분석
  - 단어 임베딩, RNN, Encoder-decoder 등
- 딥러닝 vs. 기계학습 성능
- 요약 및 향후 방향

## 자연언어처리

- 자연언어 이해 및 생성에 대한 계산적 모형에 대한 연구
- 언어분석 (Natural Language analysis)
  - Morphology/Word: 형태소 분석/ 품사 태깅
  - Syntax: 구문 분석
  - Semantics: 의미 파싱, 의미분석
  - Pragmatics: 담화 분석, 상호 참조 해결
- 정보추출 (Information extraction)
  - Named entity recognition (개체명 인식)
  - Relation classification (관계 분류)
  - Entity linking (개체명 연결)
- 자연언어생성 (Natural language generation)
  - 문장/텍스트 생성

## 자연언어처리

- 응용
  - 기계 번역 (Machine translation)
  - 질의 응답 (Question answering)
  - 감성 분석 (Sentiment analysis)
  - 텍스트 요약 (Text summarization)
  - 대화 시스템 (Dialog system)
  - 기계 독해 (Machine reading)
  - ...

## 자연언어처리: 기술 발전도

- 규칙 기반 (50~80년대말)
  - 규칙 기반 방법, 문법론 기반, 지식/논리 기반 접근법
    - Chomsky의 변형 생성 문법, 지배 결속 이론
    - HPSG, Lexical functional grammar, Tree adjoining grammar
- 통계적 접근법 (~2000 초)
  - Annotated 코퍼스로부터 generative model 학습
  - Hidden Markov model ('86)
  - Probabilistic CFG ('80~'90)
  - IBM's Statistical machine translation ('90)
  - PennTreeBank ('93)
  - Statistical language learning [Charniak '94]
  - Head-driven statistical model for parsing [Collins '99]

## 자연언어처리: 기술 흐름도

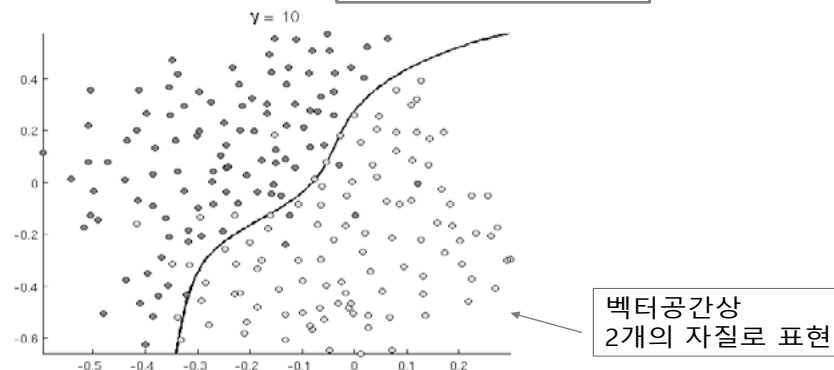
- 기계 학습 기반 (90년대말 ~ )
  - 언어처리의 구조적 분류 문제가 Structured prediction 분야로 기계학습의 하위 분야로 발전
  - Probabilistic graphical models이 자연언어분석에 적용
  - Support vector machine (Vapnik '95)
  - Structed perceptron (Collins '02)
  - Conditional random field (Lafferty '01)
  - Latent Dirichlet Allocation (Blei '03)

## 자연언어처리: 기술 흐름도

- 딥 러닝 기반 (2000후반 ~)
  - 단어임베딩을 통해 입력문을 인코딩 → 딥 아키텍처 상에서 해당 task 수행
  - Neural language model (Bengio et al. '03)
  - SENNA (Collobert et al. '11)
  - Recursive neural network (Socher et al. '12)
  - Neural machine translation (Cho et al '14; Bahdanau et al. '15)
  - Neural Turing machine (Grave '14)
  - Memory network (Weston et al '14)

## 기계 학습 (Machine learning)

- 학습 데이터: Training examples
- 학습: 학습데이터로부터 주어진 예제를 범주로 분류하는 classification 함수를 학습 (또는 regression 함수)
- 데이터의 표현: 자질 벡터 (feature vectors)



## 딥러닝 기반 자연언어처리

- 임베딩 (Embedding)
  - '의미' (meaning)를 벡터 공간상의 한 정점으로 인코딩
    - 예) 단어 임베딩 (word embedding)
  - 단어, 구, 문장, 문맥 등 자연언어처리 과정에서 발생하는 모든 처리 단위를 대상
- 합성성 (Compositionality)
  - 단위가 큰 임베딩 벡터를 보다 단위가 작은 구성 요소들의 임베딩 벡터들의 비선형 함수로 정의
    - 예) 단어 임베딩으로부터 단어열 문맥, 문장, 문서의 벡터를 합성 (composition)
  - 재귀적 신경망 [Socher '11]

## 딥러닝 기반 자연언어처리

- 합성성 (Compositionality) 적용 사례

기존 연구들	합성성 원리 적용 대상
SENNA	문맥에 대한 의미 표상 (단어열)
재귀적 신경망	단어열 (구, 문장)에 대한 의미 표상
순환적 신경망	단어열 (구, 문장)에 대한 의미 표상

## 언어처리: 딥러닝 vs. 기계학습 기반

Tasks	기계학습	딥러닝
형태소분석	CRF (음절기반 분류) S-SVM 구기반 모델	Bi-LSTM-CRF Encoder-decoder
의존파싱	전이기반 그래프기반	SyntaxNet, Stack LSTM Attention-based models
기계번역	SMT (Statistical machine translation)	Neural machine translation (Encoder-decoder)
개체명 인식	CRF, S-SVM	Bi-LSTM-CRF
상호 참조	규칙 기반	MLP Attention-based models
의미역 태깅	S-SVM	Bi-LSTM-CRF

## 언어처리: 딥러닝 vs. 기계학습 기반

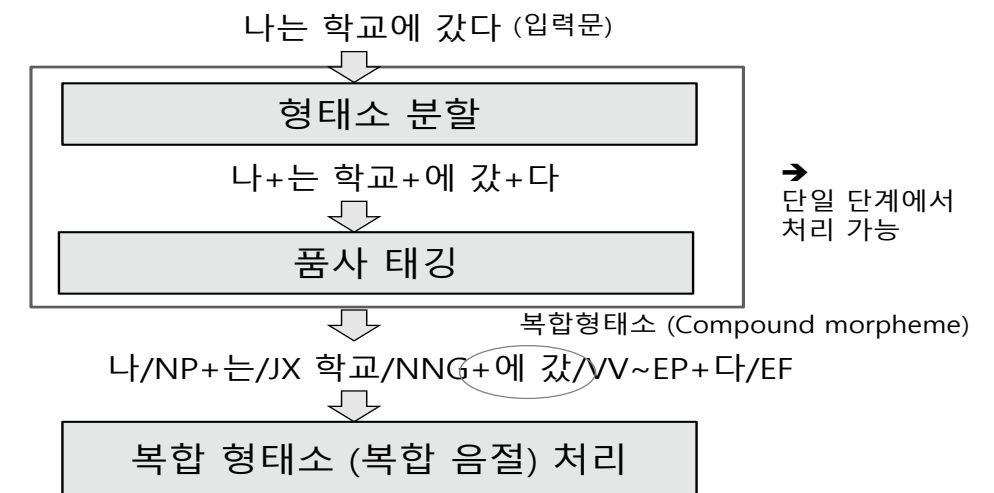
Tasks	기계학습	딥러닝
언어모델링	N-gram	RNN, ConvNet기반
감성분석	SVM 등	Recursive neural network TreeLSTM
응답 생성	SMT	Encoder-Decoder
질의 응답	IBM Watson	Neural Turing machine, Memory network

## • 최근 딥러닝 방법

- Memory-augmented networks
- GAN (Generative Adversarial Networks)
- Reinforcement learning

## 기계 학습 기반 한국어 형태소 분석: 사례

- 음절 기반 형태소 분석 [심광섭 '11; 나승훈 '12; 이창기 '13]



## 음절 기반 형태소 분석

## 자질 추출 (Feature Extraction)

- 사용 자질 정보의 예 [나승훈 '12]

자질 정보	설명
$C_{-2}, C_{-1}, C_0, C_1, C_2$	1음절(uni-char) 정보
$C_{-1}C_0, C_0C_1, C_1C_2$	2음절(bi-char) 정보
$C_{-1}C_0C_1, C_0C_1C_2$	3음절(tri-char) 정보
$S_{-2}, S_{-1}, S_0, S_1, S_2$	1음절 띄어쓰기 정보
$S_{-1}S_0, S_0S_1, S_1S_2$	2음절 띄어쓰기 정보
$S_{-1}S_0S_1, S_0S_1S_2$	3음절 띄어쓰기 정보

– 음절 정보 ( $C_i$ )

- 해당 위치에서 다음 (이전)  $i$ 번째 음절정보를 지칭

– 띄어쓰기 정보 ( $S_i$ )

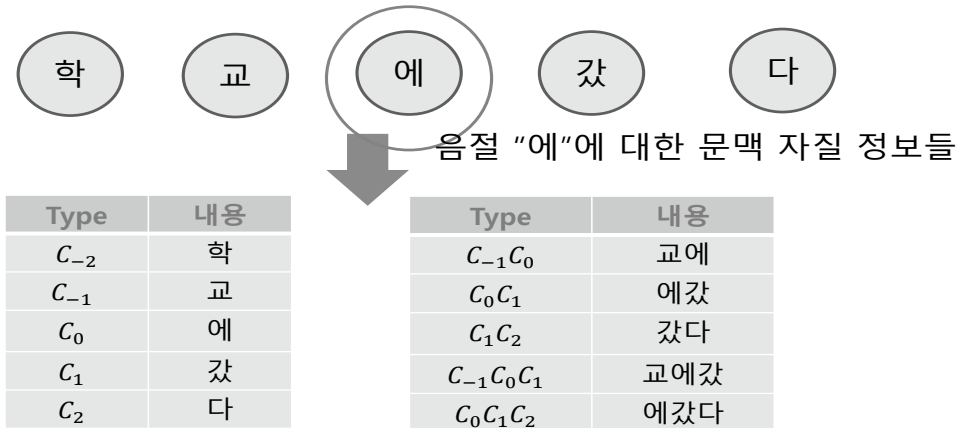
- 해당 음절 위치에서 띄어쓰기가 있었는지 여부



음절 기반 형태소 분석

## 자질 추출

- 예: “학교에 갔다”

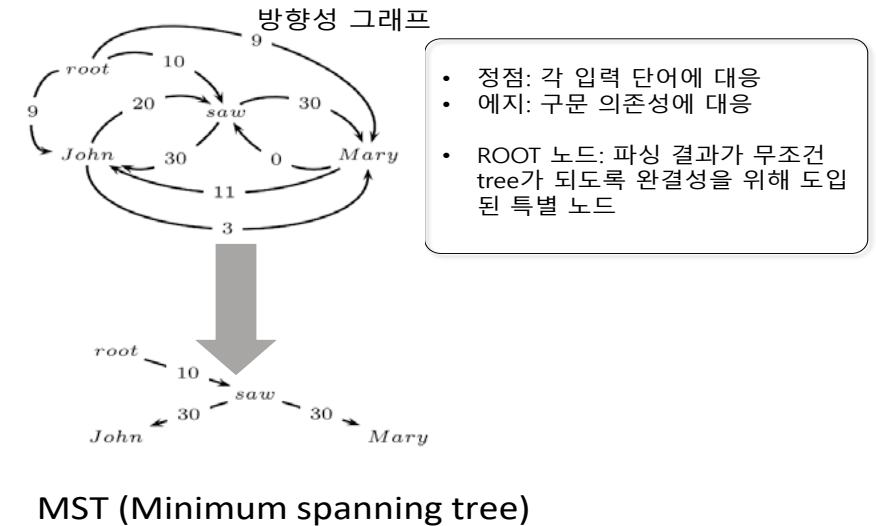


95 - 96%의 어절정확률

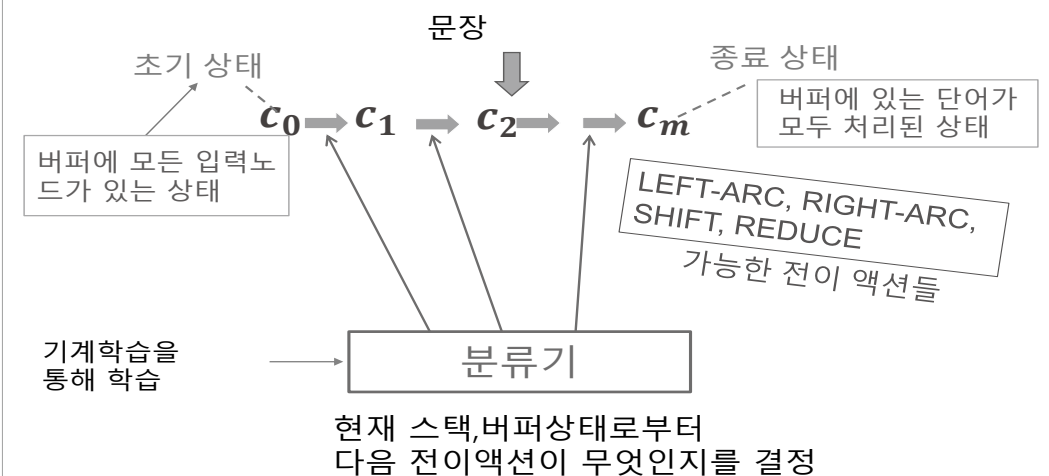
## 기계 학습 기반 의존 파싱

- 그래프 기반 방식 (Graph-Based)
  - 입력문에 대해 방향성 그래프를 형성
  - Parsing: 방향성 그래프에서 가장 높은 점수를 갖는 트리를 찾음
  - 트리 점수 계산을 위해 기계 학습 방법 적용
- 전이 기반 방식 (Transition-Based)
  - Shift-reduce 파싱 과정에서 현재 상태를 정의
    - Stack과 Buffer내의 단어나 부분트리를 참조
  - Parsing: 현재 상태에서 가장 높은 점수를 갖는 액션 선택
  - 액션 (action) 점수 계산을 위해 기계 학습 방법 적용

## 그래프 기반 의존 파싱 = MSTParser



## 전이 기반 파싱



## 전이 기반 파싱: 예제

	stack	buffer
	[0]	[John hit the ball]
SHIFT	[0 John]	[hit the ball]
LEFT-ARC <sub>subj</sub> (pop)	[0]	[hit the ball] ↓ John
SHIFT	[0 hit] ↓ John	[the ball]
SHIFT	[0 hit the] ↓ John	[ball]

## 전이 기반 파싱: 예제

	stack	buffer
LEFT-ARC <sub>det</sub> (pop)	[0 hit] ↓ John	[ball] ↓ the
RIGHT-ARC <sub>obj</sub> (shift)	[0 hit ball] ↓ John the	[]

↓

Buffer empty = Terminal configuration

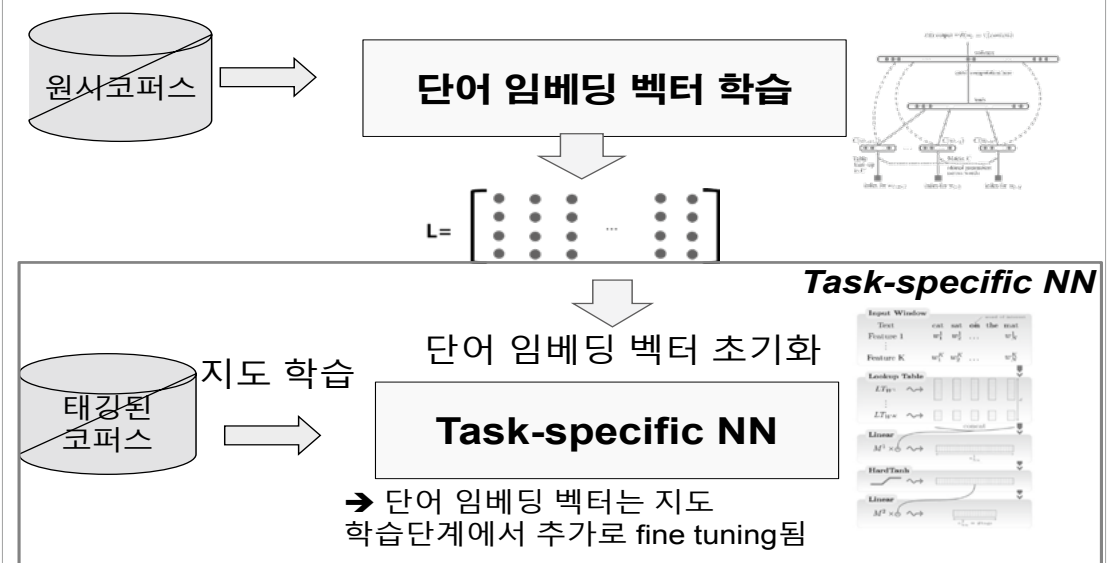
딥 러닝 기반 자연언어처리:  
단어 임베딩 - 분산 표현

## • 분산 표현

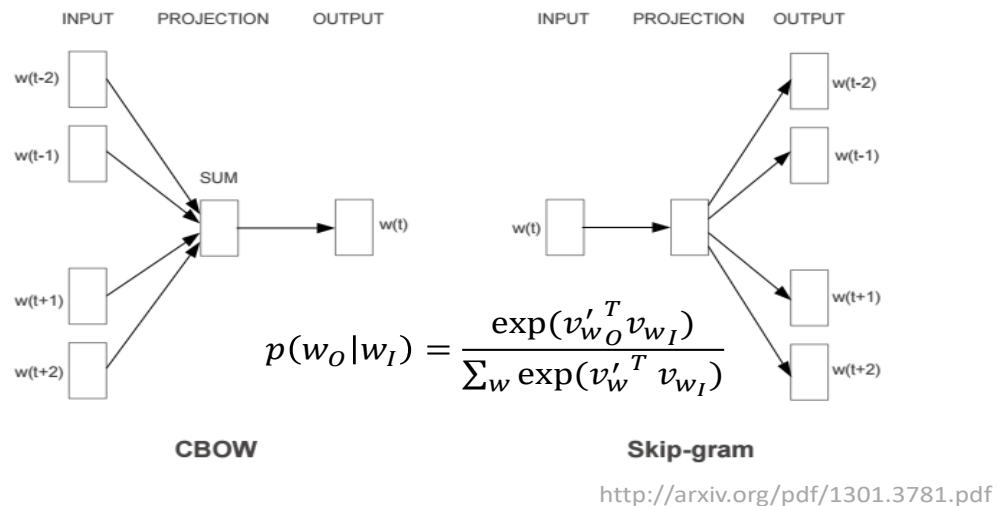
- **n-dimensional latent vector for a word**
- Semantically similar words are closely located in vector space



## 단어 임베딩을 통한 자연언어처리



## 단어 임베딩 벡터 학습: Word2Vec [Mikolov '13]



## Word2Vec: 임베딩 벡터간의 선형 관계

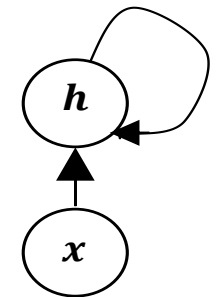
- $a:b=c:d$ 
  - $X = \text{vector}(\text{"biggest"}) - \text{vector}(\text{"big"}) + \text{vector}(\text{"small"})$



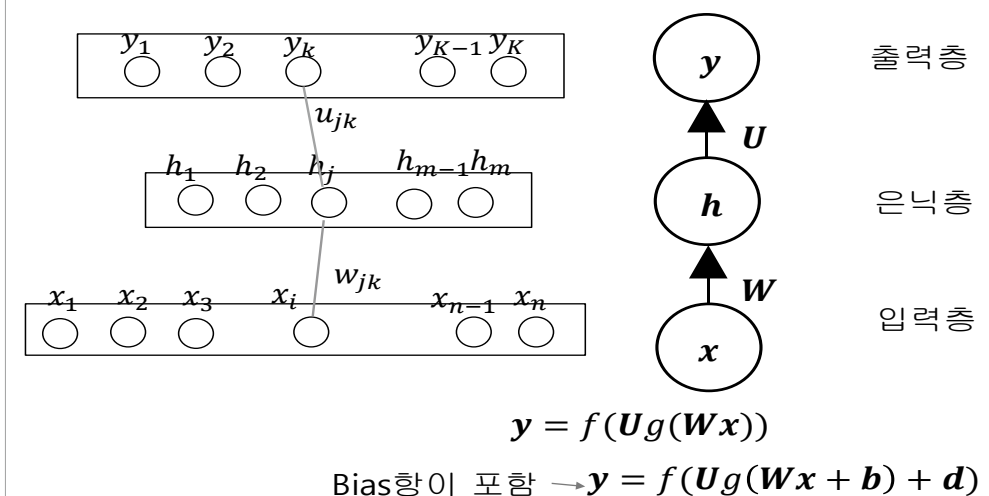
Mikolov '13

## Neural Networks: FNN & RNN

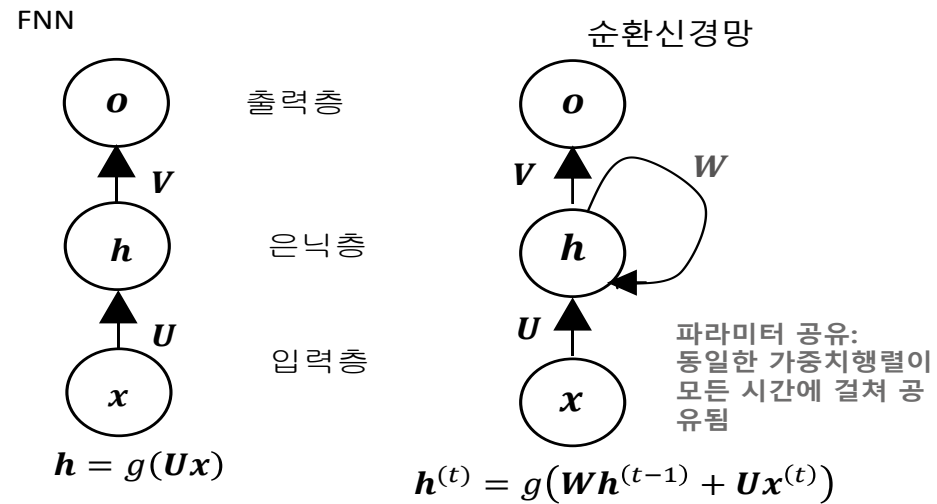
- 피드포워드 신경망 (FNN)
  - 일반 다층 신경망 (MLP)
  - 피드백 연결이 없음
    - 정보 흐름:  $x \rightarrow f(x) \rightarrow y$
  - DAG (directed acyclic graph) 로 표현
- 순환신경망 (RNN)
  - 피드백 연결이 포함
  - Cyclic graph로 표현됨
  - 학습상 난점: Vanishing gradient 문제
    - Gradient clipping
    - Long short term memory (LSTM)



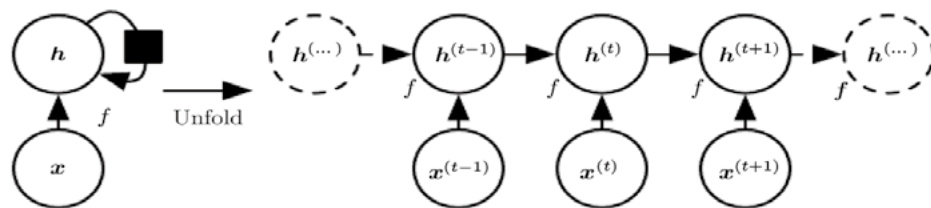
## FNN: Matrix Notation



## RNN: Matrix Notation



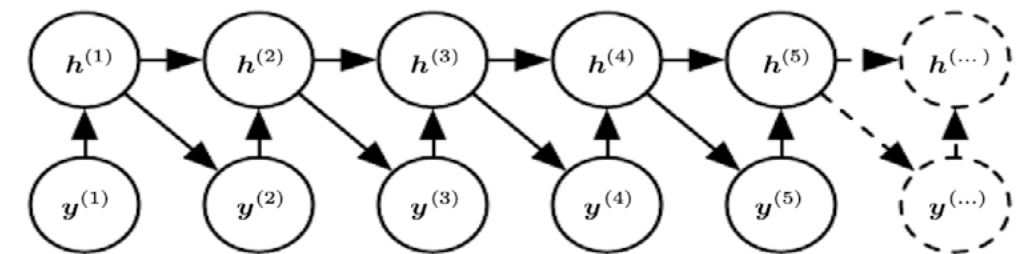
## RNN: Unfolding



- 입력 시간  $t$  시점에  $h^{(t)}$ 에는 현재까지 입력된 모든  $x^{(1)}, x^{(2)}, \dots, x^{(t)}$ 까지의 정보가 보관될 수 있음
- $[h^{(1)}, \dots, h^{(t)}] = RNN([x^{(1)}, \dots, x^{(t)}])$
- $[h^{(1)}, \dots, h^{(t)}] = LSTM([x^{(1)}, \dots, x^{(t)}])$

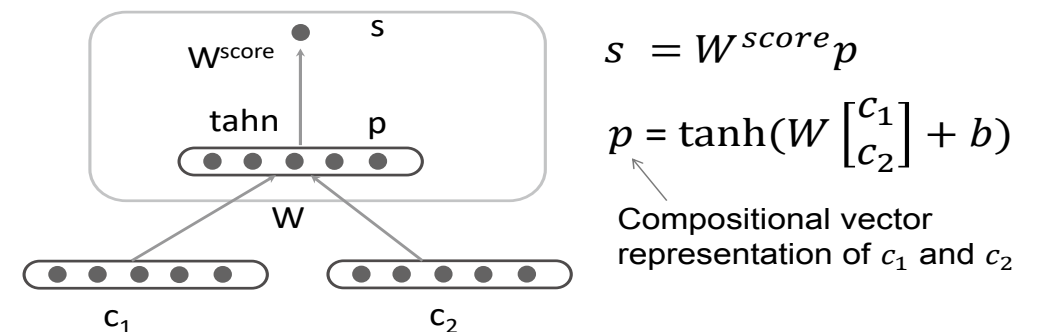
순환 언어 모델  
(Recurrent Language Model)

- 출력부분에 RNN을 적용
- 출력열을 RNN으로 생성  $\rightarrow$  응답 생성, 기계 번역
- 현재 단어 생성시 이전까지 생성된 모든 단어를 참조하여 생성
  - N-gram의 경우에는 (n-1)개의 이전 단어만을 참조



## 재귀 신경망 (Recursive Neural Net)

- 부분표상으로부터 단일 표상을 합성



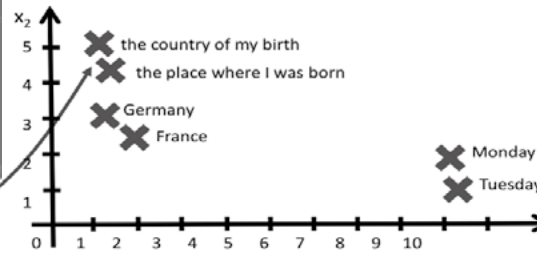


## 재귀 신경망: 구 임베딩

Use principle of compositionality

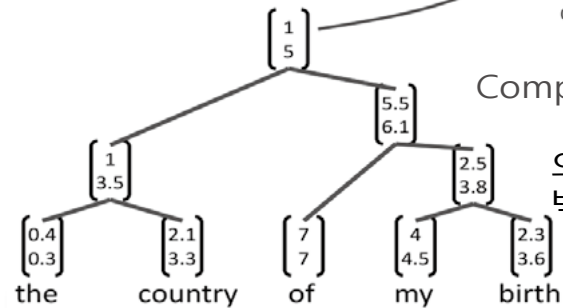
The meaning (vector) of a sentence is determined by

- (1) the meanings of its words and
- (2) the rules that combine them.



Compositional Vector Representation

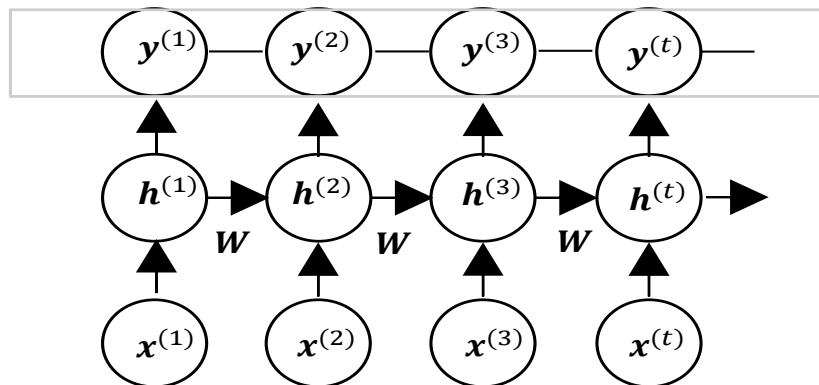
의미적으로 유사한 phrases들이 벡터 공간상에서 근접하게 배치



Slide Credit: Socher

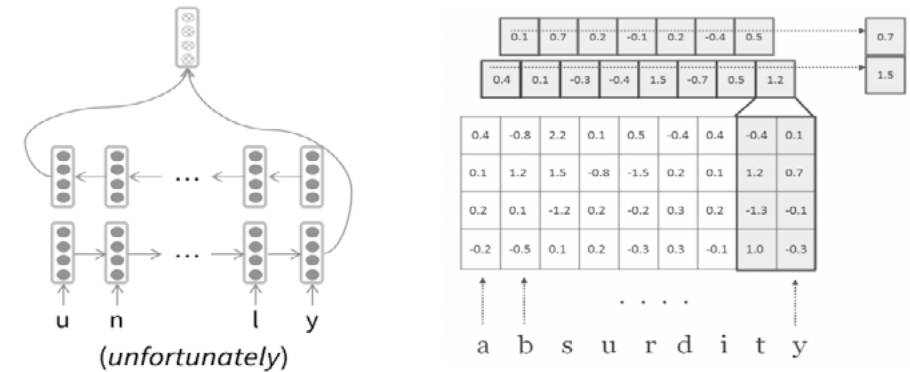
## LSTM CRF

- 순차열 태깅에서 가장 효과적인 딥러닝 방법
- 출력층에 레이블간의 의존성을 모델링하는 output dependency를 추가 (Deep CRF)



## 문자 기반 표상 (Character-Based Rep)

- 한국어의 경우 단어가 복수개의 형태소로 구성
  - 단어 임베딩 벡터를 직접 학습하는 것은 비효율적
- 합성성 원리 적용
  - 문자 임베딩 벡터 → 합성을 통해 단어 임베딩 벡터 유도
    - » 단어에서 문장벡터를 합성하는 것과 유사



LSTM 기반 [Ling et al' 15]

ConvNet 기반

## 뉴럴 기계 번역: Neural Encoder-Decoder 조건적 순환 언어 모델 (Conditional Recurrent Language Model)

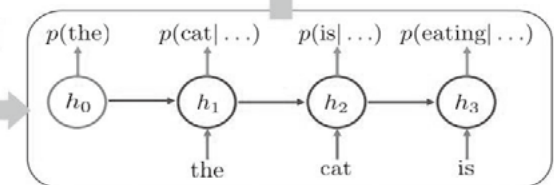
Le chat assis sur le tapis.



The cat sat on the mat.

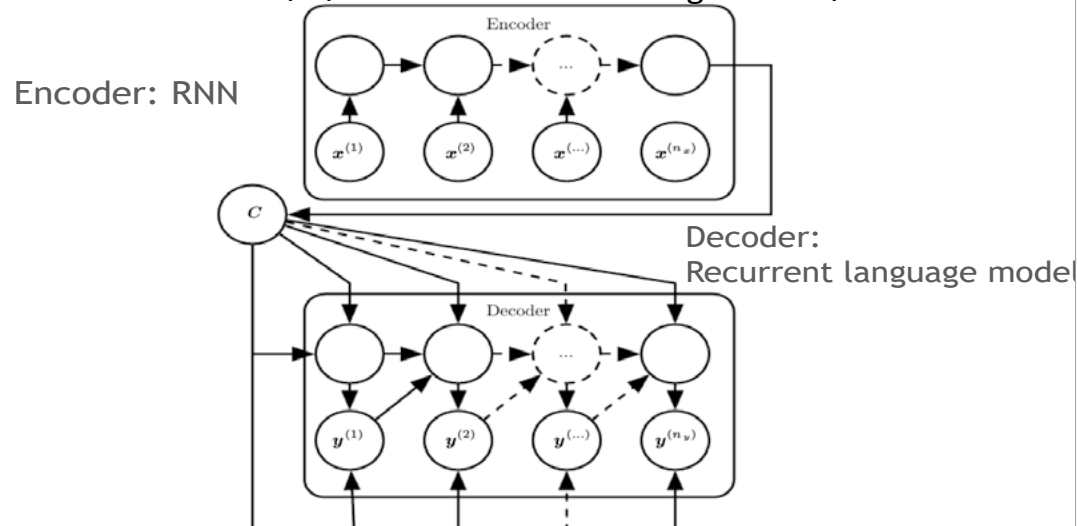
Encoder

Y

Credit: <http://nlp.stanford.edu/projects/nmt/Luong-Cho-Manning-NMT-ACL2016-v4.pdf>

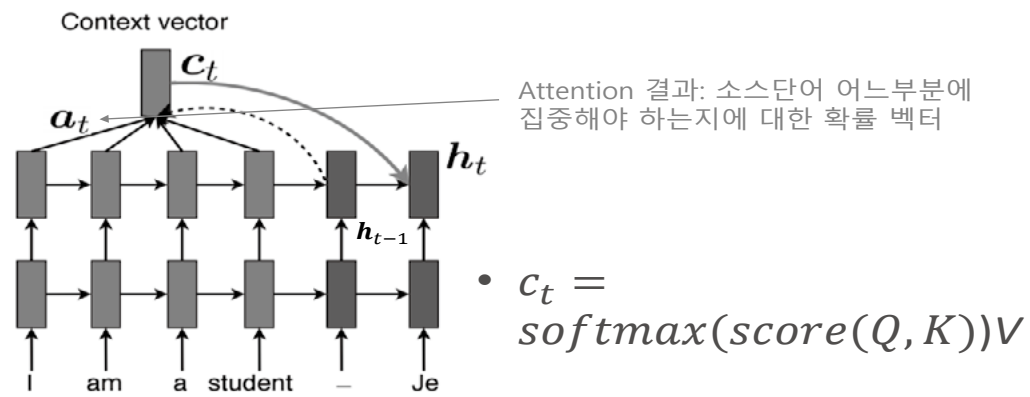
## 뉴럴 기계 번역: Neural Encoder-Decoder [Cho et al '14]

- Encoder: 소스문을 RNN으로 encoding → 입력문 저장
- Decoder: 목적문을 RNN으로 decoding → 출력문 생성



## Neural Encoder-Decoder: 주의 집중 메커니즘 (Attention Mechanism)

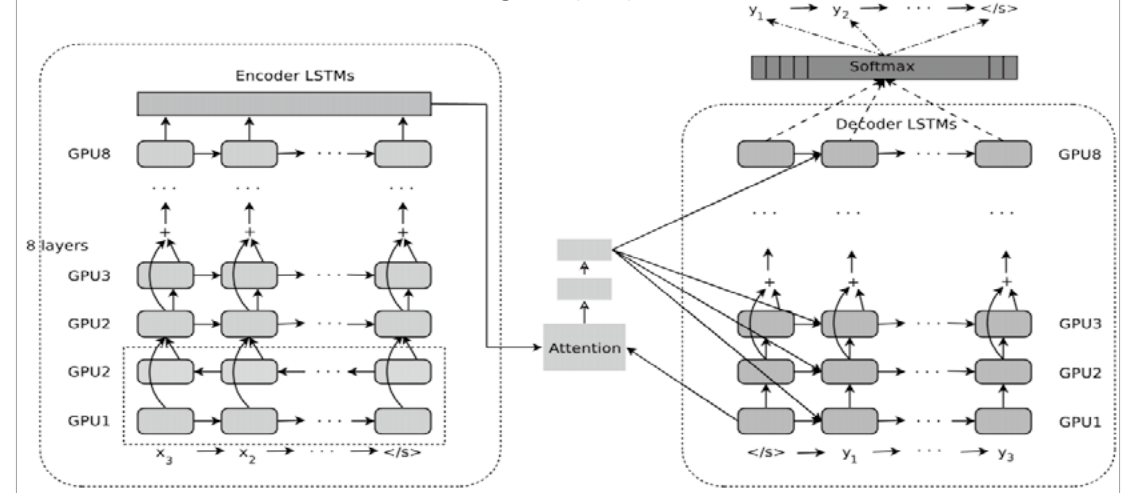
- Context vector를 고정시키지 않고, Decoding시 context vector를 attention을 통해 동적으로 획득 [Bahdanau et al '15]



## 뉴럴 기계 번역: Google's NMT [Wu et al '16]

8개의 인코더/디코더를 갖는 다층 LSTM 신경망

구글의 Tensor Processing Unit (TPU)로 학습됨



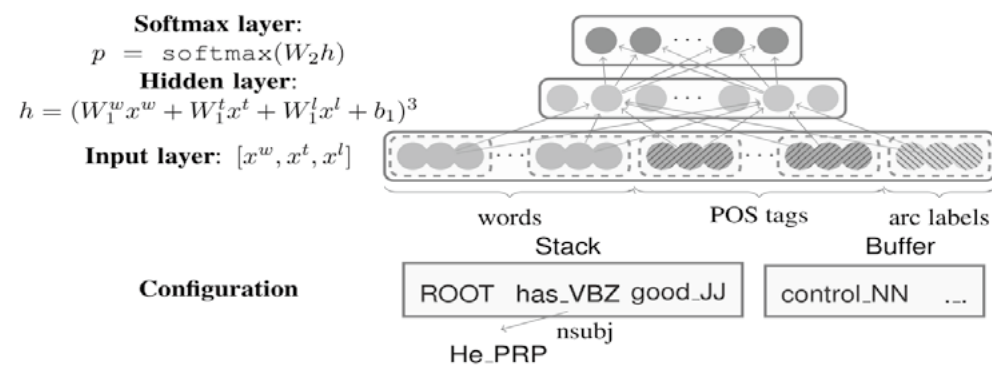
## 뉴럴 기계 번역: GNMT 성능 [Wu et al '16]

Mean of side-by-side scores on production data

	PBMT	GNMT	Human	Relative Improvement
English → Spanish	4.885	5.428	5.504	87%
English → French	4.932	5.295	5.496	64%
English → Chinese	4.035	4.594	4.987	58%
Spanish → English	4.872	5.187	5.372	63%
French → English	5.046	5.343	5.404	83%
Chinese → English	3.694	4.263	4.636	60%

구글의 기존 구기반 SMT에 비해 약 60%의 error 감소율을 보임

## 딥러닝 기반 전이 기반 의존 파싱 [Chen and Manning '14]



## 언어처리 모듈: 딥러닝 vs. 기계 학습

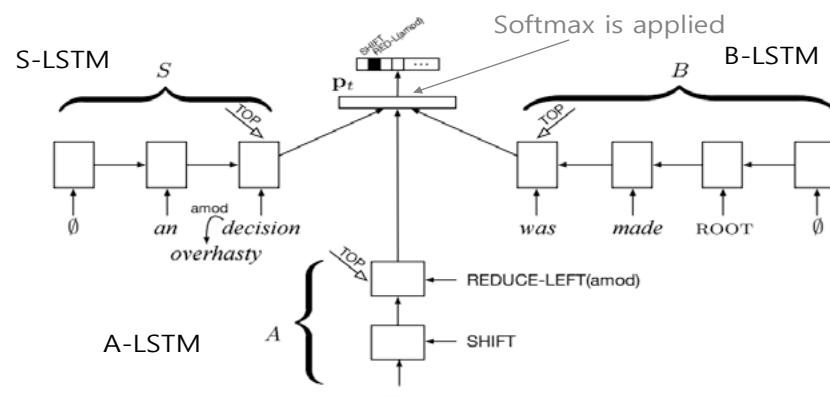
- 한국어 형태소 분석
  - Sejong dataset에서 실험

	F1 (morph)	어절 정확도
CRF [나승훈 '12; '15]	97.61%	96.14%
CRF (BIES표기)	97.75%	N/A
구기반+CRF	97.74%	96.35%
Bi-LSTM CRF*	97.03%	N/A
Encoder-Decoder	96.83%	95.38%

- 딥러닝의 성능은 CRF에 비해 제한적
  - 중국어의 경우에는 딥러닝 성능이 기존 CRF의 성능을 향상시킴

## LSTM기반 전이 기반 의존 파싱 [Dyer et al '15]

- 스택, 버퍼, 액션 히스토리를 LSTM으로 인코딩



## 언어처리 모듈: 딥러닝 vs. 기계 학습

- 한국어 의존파싱
  - [SPMRL '14 dataset]

		UAS	LAS
기존 기계학습 기반 방법		89.10%	87.27%
최고 성능 [Björkelund et al 2014]			
전이기반	Stack LSTM [나승훈 KCC 16']	89.10%	87.34%
	Stack LSTM + 컨트롤러 [나승훈 KCC 16']	89.94%	88.36%
	SyntaxNet [민진우 KCC 17']	90.33%	88.69%
	Tagging&Parsing 통합 모델 [민진우 HCLT '17]	90.48%	88.87%
그래프기반	Deep biaffine [나승훈 KCC '17]	90.85%	89.31%

- 딥러닝 모델은 기존 기계 학습에 비해 성능 증가를 가져옴 (LAS 2%이상)

## 언어처리 모듈: 딥러닝 vs. 기계 학습

### • 한국어 개체명 인식

방법	Dev	Test
CRF	84.70%	84.82%
LSTM-CRF (베이스라인)	85.57%	82.98%
LSTM-CRF (morph, 문자기반 LSTM)	86.72%	84.49%
LSTM-CRF (morph, 제안 문자기반)	87.01%	85.87%
LSTM-CRF (morphtag, 문자기반 LSTM)	87.32%	85.80%
<b>LSTM-CRF (morphtag, 제안 문자기반)</b>	<b>88.60%</b>	<b>86.53%</b>

딥러닝 모델은 개체명 인식의 성능을 비약적으로 향상시킴 - 영어권 개체명 인식도 유사

## 언어처리 모듈: 딥러닝 vs. 기계 학습

### • 영어 관계 분류 실험

- SemEval-2010 Task 8

Classifier	Additional Information	$F_1$
SVM (Rink and Harabagiu, 2010)	POS, WordNet, Prefixes and other morphological features, dependency parse, Levin classed, PropBank, FanmeNet, NomLex-Plus, Google $n$ -gram, paraphrases, TextRunner	82.2
RNN (Socher et al., 2011)	Word embeddings + POS, NER, WordNet	74.8 77.6
MVRNN (Socher et al., 2012)	Word embeddings + POS, NER, WordNet	79.1 82.4
CNN (Zeng et al., 2014)	Word embeddings + word position embeddings, WordNet	69.7 82.7
FCM (Yu et al., 2014)	Word embeddings + dependency parsing, NER	80.6 83.0
CR-CNN (dos Santos et al., 2015)	Word embeddings + word position embeddings	82.8 84.1
SDP-LSTM (Xu et al., 2015b)	Word embeddings + POS + GR + WordNet embeddings	82.4 83.7
DepNN (Liu et al., 2015)	Word embeddings, WordNet Word embeddings, NER	83.0 83.6
depLCNN (Xu et al., 2015a)	Word embeddings, WordNet, word around nominals + negative sampling from NYT dataset	83.7 85.6
BRCNN (Our Model)	Word embeddings + POS, NER, WordNet embeddings	85.4 <b>86.3</b>

## 언어처리 모듈: 딥러닝 vs. 기계 학습

### • 영어 Entity linking

[Nguyen '16] 논문에서 발췌

Systems	Wikipedia 2014				Wikipedia 2016			
	ACE	CoNLL	WP	WIKI	ACE	CoNLL	WP	WIKI
<i>DK2014</i> (Durrett and Klein, 2014)	79.6	-	-	-	-	-	-	-
<i>AIDA-LIGHT</i> (Nguyen et al., 2014b)	-	84.8	-	-	-	-	-	-
<i>Local CNN</i> (Francis-Landau et al., 2016)	<b>89.9</b>	85.5	90.7	82.2	86.1	84.5	90.4	81.4
<i>Global-RNN</i>	89.7	<b>87.2</b> †	<b>91.2</b> †	<b>83.7</b> †	<b>87.8</b> †	<b>86.5</b> †	<b>91.2</b> †	<b>81.7</b>

ConvNet, RNN 등 딥러닝 기반 방식은 기존 기계학습 기반 방법을 크게 향상 시킴

## 자연언어처리 - 딥러닝 vs. 기계 학습: 요약

### • 딥러닝 기반 방법

- 각종 NLP tasks 등에서 딥러닝 방법이 기존 기계학습 방법에 비해 성능 향상을 보여줌

- 언어분석, 정보추출, 기계 번역, 기계 독해, 텍스트 요약, 대화 처리, QA 등

- 정보검색 분야에서도 딥러닝 기반 방법이 성능 향상을 보이는 연구들이 발표중

### • 최근 딥러닝 연구 동향

- GAN (Generative Adversarial Networks)

- Reinforcement learning 등이 기계번역, 언어생성 연구에 적용



## 딥 러닝 기반 한국어 언어처리

- 형태소 분석: 중국어 segmentation/tagging과 유사
  - 중국어: 문자열, 한국어: 음절열
- 형태소 분석 이후의 tasks: 두가지 접근법
  - 1) 형태소 단위: 형태소 분석을 수행 후 형태소 단위로 처리
  - 2) 합성 기반 방법: 문자 기반 합성을 통해 단어의 임베딩 벡터 합성
    - 문자 단위로 형태소 또는 음절을 사용
      - 예) 형태소 embedding vectors → 단어 embedding vector
    - 합성 기반 방법은 현재 Morphologically rich languages에 적용되고 있는 common approach 중 하나

## 딥 러닝 기반 한국어 언어처리

- 한국어 특화 부분
  - 단어 임베딩
    - 합성 방법의 경우 형태소 또는 음절로부터 단어 임베딩을 유도
    - 형태소 임베딩은 자동형태소 분석기를 통해 기존 word2vec 등을 통해 얻음
  - 단어 임베딩을 합성방식으로 학습하는 방법 모색 필요

## 한국어 언어 처리: 전망

- 딥러닝 기반 언어처리 가속화
  - 주요 언어분석 tasks외에 의미 파싱, 질의 응답, 요약, 텍스트 분류, 기계 독해, 대화 처리 등으로 확장
- GAN, reinforcement learning 등을 언어분석에 적용
- 한국어외에도 영어권에도 적용될 수 있는 공통적인 독자적인 모델을 추구
  - 언어처리 모델의 다국어 확장성: 학습데이터가 주어지면 한 언어에서 적용되는 모델은 다른 언어에서 확장 적용
- 지식베이스와 연계를 통한 정보 시스템 응용
  - 의료, 법률 등 도메인 특화 정보 서비스

## 한국어 언어처리: 개선점

- 주요 tasks들의 데이터셋 표준화 문제
  - 주요 tasks인 형태소 분석, 구문분석, WSD, 의미역 태깅은 dataset은 존재하나 standard dataset이 부재
    - 연구자들마다 별개의 dataset으로 평가
- 확장 Tasks들에 대한 dataset 부재 또는 부족
  - 개체명 인식
  - 이미지 캡션
  - 의미 파싱, AMR 파싱, RST Discourse 파싱
  - Dialog Act 분류, Dialog state tracking
  - Large-scale 한영 병렬 코퍼스
  - Large scale QA 등

주제 5

# 우리말 자연 언어 처리 기술의 전망

· 토론자 차정원(창원대학교 컴퓨터공학과 교수)



# 우리말 자연어처리 기술의 전망

2017.11.10  
차정원  
창원대학교

Adaptive Intelligence Research

1

## 딥러닝 기반 자연언어처리

- 의미표현의 분류를 다음과 같이 할 때,
  - 분산 의미표현(Distributional meaning representation)
  - 합성 의미표현(Compositional meaning representation)
- 분산 의미표현이 실제 단어의 벡터화 표현 이외의 기능에 대해서 회의적인 시각도 있다.

2

## 미등록어 처리

- Pre-training으로 인해 FFNN보다 나아졌지만 여전히 미등록어에 대해서 성능이 좋지 못하다. 이것이 네트워크의 한계에 의한 것인가? 극복이 가능한 방법이 무엇이 있는가?

3

## 학습 데이터의 양

- Deep learning의 문제점 중에 가장 많이 언급되는 것이 학습 데이터의 크기에 대한 것이다. 일정 수준 이상의 성능을 달성하려면 요구되는 데이터의 양이 매우 크다. 이 부분을 극복하기 위해서 GAN, One-shot learning 등이 제안되었지만 여전히 레이블링 데이터를 활용한 학습에 비해 성능 낮다.

4

## Deep Neural Network?

- 현재의 deep neural network이 feature selection을 자동으로 하는 것과 다른 것이 무엇인가? 결국 특정 selection 기법이 아니라 모든 경우를 해보고 제일 성능이 좋은 것을 선택하는 것 이상이 있는가?

5

## 교육

- Deep learning을 통한 언어처리 방법에서는 언어처리 지식을 많이 사용하지 않고 원인에 대한 분석이 없는 경우도 많다. 이런 환경에서 학생들에게 어떤 것을 교육해야 하는가?

6



**from Waleed Kadous, Quora**

- 7

## 8

· 발표자 **이경님**(엔씨소프트 에이아이 센터 스피치 랩 음성인식팀)



## 1. 들어가기

사람이 의사소통을 하기 위해 사용하는 일반적이고 효과적인 수단은 언어(말과 글)이고, 음성은 인간의 가장 자연스러운 의사소통 방식이다. 말의 내용과 감정을 인식하고 의미를 이해하여, 상황에 따라 자연스러운 대화를 주고 받기 위해 필요한 음성언어처리 기술은 인간의 자연어 발화를 컴퓨터가 자동으로 이해하고 처리하는 알고리즘을 연구하는 분야로 대화형 개인비서 에이전트, 인공지능(AI) 스피커, 자동 통번역, 음성대화 질의응답(QA) 시스템 등 다양한 응용 서비스 사례를 들 수 있다.

이미 모바일 메신저 서비스를 통해 많은 양의 소통이 이루어지고 있고, 최근에는 채팅 봇이 콜센터 안내, 쇼핑 도우미 및 고객 상담 등의 다양한 서비스를 통해 대화형 상거래가 가속화되고 있다. 현재는 문자 기반으로 그 서비스가 제공되고 있으며 음성인식 성능이 기대 수준에 이르면 음성 인터페이스로 그 기능을 확장 연계 할 수 있을 것이다. IT 시장조사기관 가트너(Gartner)에 따르면 2019년에는 스마트폰과 사용자간의 상호 작용 중 20%가 가상개인비서(Virtual Personal Assistants)를 통해 이루어질 거라 것과 오는 2020년까지 개인용 기기는 70억 대, 웨어러블 기기는 13억 대, 그리고 IoT 기기는 57억 대로 늘어날 것으로 전망하고 있으며, 이 중 최소 20억 대의 기기 및 사물인터넷 장비가 누르지 않고 제어할 수 있는 제로터치 UI 기반으로 작동할 것으로 전망하고 있다. 성공적인 생태계가 활성화 된다면 앞으로 무한한 확장 가능성을 기대할 수 있을 것이다.

서비스의 선두적인 애플 'Siri'(2011)를 시작으로 마이크로소프트 'Cortana'(2014), 구글 'Assistant'(2016) 등 스마트폰 기반 대화형 개인비서와 아마존 'Echo'(2014)과 구글 'Home'(2016), 애플 'HomePod'(2016)을 비롯하여 SKT '누구'(2016), KT '기가지니'(2017), 카카오 '미니'(2017)와 네이버 '웨이브'(2017) 등 국내 기업들도 최근 음성인식 기반의 인공지능 스피커들을 출시하고 있다. 스마트폰을 비롯, 스마트 스피커, 스마트 홈 허브 기능을 가지는 셋톱박스, TV, 냉장고 등 스마트 가전으로 음성인식 기술이 급격히 확산되고 있는 점을 고려할 때, 터치와 버튼이 아닌 음성으로 기계를 제어하고 소통하는 새로운 세상이 가져올 미래 삶의 변화를 기대해 볼 만하다.

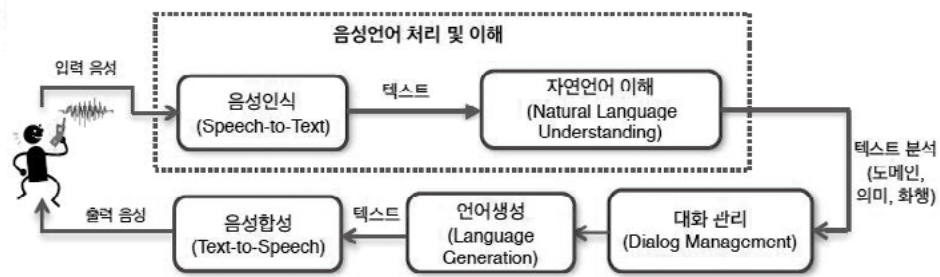
글로벌 기업들도 로컬 서비스에 국한하지 않고 다국어 서비스 지원 및 확장을 준비하고 있는 상황에서 특히, 한국어 음성언어처리 기술 및 응용 서비스 개발 동향과 성능 개선을 위해 진행되고 있는 연구들을 소개하고자 한다.

## 2. 음성대화 인터페이스 및 응용 서비스

음성대화 인터페이스는, 사람들 간의 또는 사람과 기계 간의 인간의 자연어 음성을 컴퓨터가 듣고 이해하여 주어진 상황에 따라 적절하게 대응하면서 대화를 나누는 음성대화시스템(Spoken Dialog System) 구축을 궁극적인 목표로 한다. 사용자의 음성이 입력되면 음성인식을 통해 나온 텍스트 결과로부터 자연언어 이해를 통해 텍스트 심층 분석 결과를 구하게 되고, 언어생성기는 적절한 응답 문장을 생성하고, 음성합성을 통해 스피커 출력으로 음성

을 재생한다. 해당 요소 기술들을 독립적으로 구성할 수도 있지만, 연속성을 갖는 선순환 구조이기 때문에 음성언어 처리 관점에서 보면 음성인식과 언어이해 분야를 밀결합(tightly-coupled)하여 음성신호처리 기술부터 언어처리 영역까지 포괄하는 ‘음성언어 이해’라는 기술 분야로 좀더 사람의 발성언어를 잘 반영하여 성능 향상의 한계를 넘어서려는 시도를 하고 있다.

한편 과거의 음성인식 기술은 아나운서가 책을 읽듯이 발성하는 음성을 대상으로 하는 낭독체 음성인식 기술이 주로 연구 대상이었으나, 딥러닝 및 잡음처리 기술의 발전으로 인해 현재는 사람 간의 자연스런 대화 음성을 대상으로 기술 고도화가 이루어지고 있다.



[그림 1] 음성대화 인터페이스 시스템 구성

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

[표 1] 인공지능 대화형 서비스 기능

	기기 제어 및 서비스 연동 기능	정보 검색 및 질의 응답	일반 채팅
기능	음성 명령 등을 통해 특정 기기 기능에 액세스하고 대화형으로 상호 작용	연계된 포털 사이트를 통해 사용자가 원하는 정보 검색	지능형 가상비서와의 일상 대화
발성 예	“엄마에게 늦다고 문자 보내줘” “캘린더 상에 내일 일정이 있는지 체크” “음악 소리 줄여줘”	“US달러 오늘 환율은 얼마야?” “US달러와 호주달러간 환율 알려줘”	“날씨 참 화창하고 좋네” “재밌는 이야기 좀 해봐?”

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

[표2] 국내 인공지능 스피커 서비스 모델 추진 동향

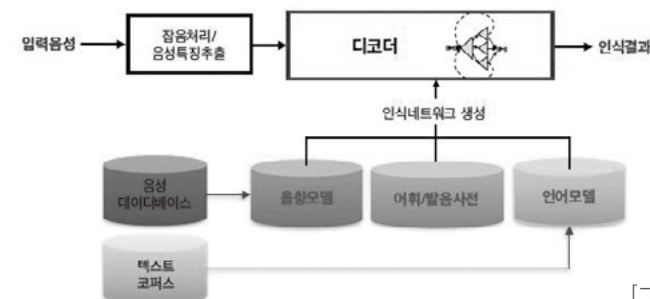
	SKT '누구'	카카오 '미니'	네이버 '웨이브'
주요 기능	음악 감상, 일정 관리, 날씨, 커머스, IPTV, 교통정보	검색, 뉴스 확인 음악 감상, 날씨 안내, 음성 명령으로 카카오톡 이용	지식 정보 검색 음악 추천, 팟캐스트, 음성 메모, 일정 알림, 뉴스 브리핑
연동 서비스	T맵, BTV	카카오톡 (카카오톡시, 카카오퍼블 등 확장 예정)	네이버 뮤직
음성 호출어	아리아, 팅커벨, 레베카, 크리스탈	헤이 카카오	샬리아

### 3. 음성인식 요소 기술

대화형 인공지능 서비스의 관문인 음성인식 기술에 대해 소개하고자 한다. 인공지능 기술과 마찬가지로 음성인식도 오랜 기간 발전을 거듭해 오면서 성능 개선의 한계로 인한 암흑기를 겪기도 했다. 최근 몇 년 동안 혁신적인 성능을 보이기 된 데에는 클라우드 서버 및 고성능 GPU와 같은 하드웨어의 눈부신 발전과 빅데이터 기반 대용량 분산 처리 기술 활용에 그 배경을 두고 있다.

음성인식을 위한 많은 데이터 및 다양한 지식은 음향학적 관점 및 언어학적 관점의 두 가지 방향에서 볼 수 있다. 음향학적 관점에서는 화자, 배경 잡음, 마이크론 등 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.

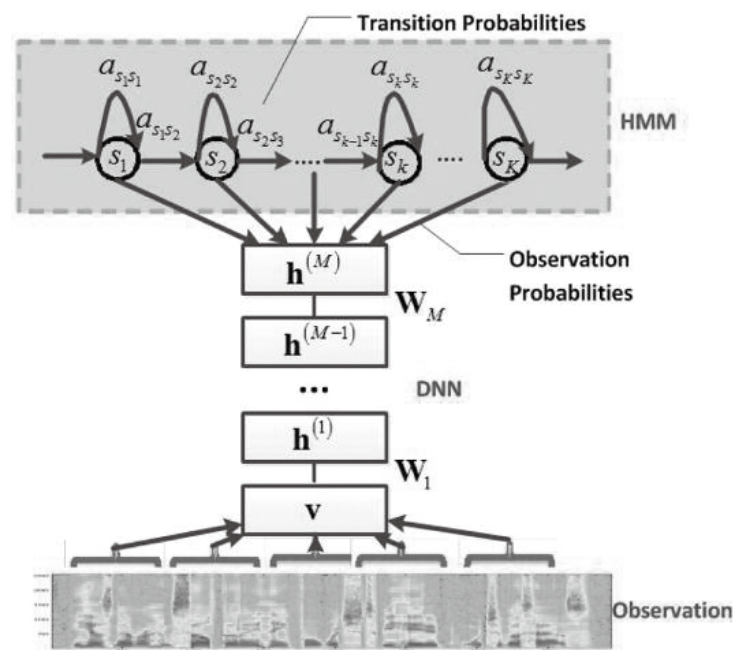
음성인식 시스템은 크게 음성/언어 데이터로부터 인식 네트워크 모델을 생성하는 오프라인 학습단계와 사용자가 발성한 음성을 인식하는 온라인 탐색 단계로 나뉘 볼 수 있다. 음성인식 엔진은 크게 음성과 언어 정보라는 중요한 사전 지식을 사용해 음성 신호로부터 문자 정보를 출력하게 되는데, 이때 개념적으로 음성 신호를 문자 심볼로 해석한다는 차원에서 음성인식 알고리즘을 디코더(decoder)라고 부르기도 한다. 디코딩 단계에서는 학습단계 결과인 음향모델(AM; Acoustic Model), 언어모델(LM; Language Model)과 발음사전(Pronunciation Lexicon)을 이용하여 입력된 특징벡터를 모델과 비교, 스코어링을 통해 단어열을 최종 결정하게 된다.



[그림2] 음성인식 시스템 구성도

음향모델링은 해당 언어의 음운 환경 별 발음의 음향적 특성을 확률 모델로 대표 패턴을 생성하는 과정이고, 언어모델링은 어휘 선택, 문장 단위 구문구조 등 해당 언어의 사용성 문제에 대해 문법 체계를 통계적으로 학습하는 과정이다. 또한 발음사전 구축을 위해서는 텍스트를 소리 나는 대로 변환하는 음소변환(G2P; Grapheme-to-Phoneme) 구현 과정이 필요하며, 표준발음을 대상으로 하는 발음변환 규칙만으로는 방언이나 사용자의 발화 습관과 어투에 따른 다양한 패턴을 반영하기 어려운 경우가 있어 별도의 사전 구축이 필요하게 된다.

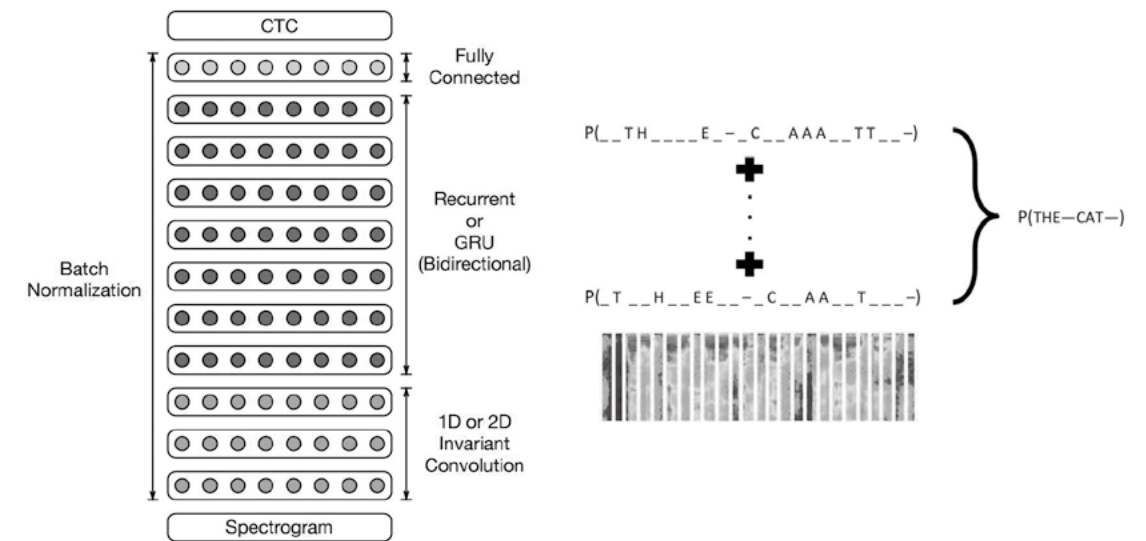
음성인식 성능은 음성 데이터베이스의 크기와 품질에 비례하여 성능 향상이 이루어진다. 상용 서비스에 적용되는 음향모델은 대부분 확률통계 방식인 HMM(Hidden Markov Model) 기반으로 이루어졌으며, 2010년대 들어서면서 딥러닝 기반의 HMM/DNN 방식으로 단어 인식 오류율 기준으로 약 20% 정도의 성능 향상을 이끌었다[7]. DNN과 HMM을 결합하는 방법은 [그림 3]과 같이 HMM의 각 상태 확률 분포를 모델링하는데 사용되는 GMM을 DNN으로 대체하는 것으로, 그 외의 모델 구분 단위, 단위별 학습자료 자동 생성 및 모델 결합을 통한 문장 인식 확장 등은 HMM에서의 방식을 다수 그대로 사용하는 반면 DNN을 추정해야 하는 파라미터가 많아 학습시간이 많이 소요된다.



[그림2] 음성인식 시스템 구성도

최근에는 sequence-to-sequence 방식의 RNN(Recurrent Neural Network) 기반으로 속도와 성능 면에서 좋은 결과를 내기 시작했다. 음성인식에서도 end-to-end 학습 방식의 발전으로 일련의 오디오 특징을 입력으로 일련의 글자(character) 또는 단어들을 출력으로 하는 단일 함수를 학습할 수 있게 되었다. 이 함수들은 중간에 음소 단

위 또는 발음 사전의 단위 변환을 거치지 않고도 긴 일련의 오디오에 대해 디코딩을 수행할 수 있다는 특징이 있다. 성공적인 예로 '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 9천600여 명의 7천 시간 길이 음성 샘플과 15가지 종류의 소음을 더해 10만 시간 가량의 샘플을 확보한 것으로 발표하였고, 중국어에 대해서도 적용하기 시작하였다[9].



[그림 4] RNN과 end-to-end 모델 설명[9]

마지막으로 음성인식 결과의 정확도를 높이기 위해서는 문법 구조를 잘 반영한 언어모델이 요구된다. 대표적인 언어 모델링 방법은 통계적인 방법에 따라 n개의 단어열에 대한 출현빈도를 확률값으로 나타내는 n-gram 기법이다. 정교한 n-gram을 생성하기 위해서는 다양한 코퍼스뿐만 아니라 실제 서비스에서 나타나는 언어 양상을 모델링할 필요가 있다. 이를 위해서는 대규모 코퍼스 기반의 언어 모델링 기술이 필수적이다. 음성인식 서비스의 경우 서비스 어휘의 수는 기하급수적으로 증가하며, 특정 도메인으로만 대상 영역을 한정할 수 없는 특징을 가지기 때문에 언어모델의 대용량화와 지속적인 확장 기능을 요구한다. 또한 미관측 어휘(OOV; Out-of-Vocabulary)에 대한 언어모델링의 한계를 해결할 수 있는 방법으로서 빅데이터로부터 대규모 텍스트 코퍼스를 얻을 수 있음을 전제로 하고 있다.

여전히 기존 방식의 한계로는 학습 코퍼스에서 관측되지 못한 어휘열에 대한 확률 값의 불안정한 추정 문제와 함께 n 값의 제약으로 인해 히스토리를 충분히 반영하기 어렵다는 점을 들 수 있다. 이를 해결하기 위해 위에 언급된 end-to-end 방식의 경우 RNN 구조를 이용한 글자(단어, 음절) 기반 언어모델을 적용하는 방법과 word2vec과 같은 단어 임베딩 연구도 활발히 진행되고 있다.



## 4. 언어학습을 위한 자유발화형 한국어 음성언어처리

한국어의 경우, 음성인식 디코딩 과정에서의 탐색공간 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 음성인식에서는 형태소 분석에 기반한 어휘 선정의 문제와 함께 인식대상 어휘의 제한으로 인한 미관측 어휘(OOV) 발생에 따른 인식 오류 발생이 성능에 영향을 끼치게 된다.

또한 방대한 양의 텍스트 데이터를 자동 형태소 분할하는 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 언어모델의 확장을 연계하여 고려해야 전반적으로 인식 성능을 개선할 수 있다. 따라서 다양한 어휘, 문법 등을 분석함으로써 무제한급 자연어 음성인식을 위한 언어 지식을 체계화할 수 있다. 다양한 환경에서 다양한 화자가 발생한 사용자 로그 정보들은 그 자체가 거대한 말뭉치 데이터로서 음성언어 처리 기술의 성능을 향상시키는 주된 리소스가 된다. 특히 음성인식의 성능 개선을 위해서는 음성 로그뿐만 아니라 방대한 분량의 텍스트 코퍼스 수집 및 한국어 음성언어처리 기술 확보가 필수적이다.

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

예상된 시나리오를 넘는 좀더 자연스러운 사람의 음성언어를 인식하고 이해하기 위해서는 대규모의 한국어 구어체 어휘 확보가 절실하다. 주로 모바일 메신저나 오픈 채팅창에서는 문자로 정보를 전달하고 있지만, 음성 메시지를 가정한 구어체에 매우 가깝다. 다만 좀더 빠르고 간략하게 하기 위해 축약어나 발음 나는 형태로 적기도 하는데 이러한 다양한 대화 현상을 반영할 수 있어야 한다. 물론 사람과 사람간의 대화와 사람 대 기계의 대화 내용은 서로 다를 수 있지만 인공지능 페르소나를 통해 인간과의 대화에 가깝다고 가정하고자 한다.

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

대화체 코퍼스로서 매우 유용한 것은 채팅에서 사용되는 문장들로 채팅의 형태에는 여러 사람이 함께 이야기하는 형태, 일대일로 이야기하는 형태, 쪽지를 주고 받는 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 타겟팅 하는 것이 매우 어렵다. 목적 기반(Task-oriented)의 시스템에서도 단순 정보 교환 및 일반 오픈 채팅이 이루어지기 때문에 수집 범위를 한정 짓기 보다는 대규모의 데이터를 이용하는 방안을 선택하고 있다.

게임 도메인 확장을 고려하여 인벤(inven) 사이트의 게시판 글과 답변뿐만 아니라 엔씨소프트의 PC 게임 ‘리니지’와 모바일 게임 ‘리니지M’ 채팅창에서 주고받은 대화내용을 모니터링 하면서 대화체 코퍼스를 보강하고 있다. 익명성에 기반한 전체 공개 대화 내용만을 대상으로 하며, 비속어나 금치어는 입력시 \*\*\* 표기 되기 때문에 그 원문은 알 수 없다. 전체 오픈 채팅 방에서는 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 축하인사 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 불구하고 비공식적, 비격식적 문서의 양이 절대적으로 늘어나면서 불완전한 문장 또는 문법에 어긋나는 표현들도 함께 늘어나게 된다. 표3과 같이 문어체 문장과는 달리 채팅 문장의 특성을 살펴볼 수 있다.

[표 3] 채팅 문장 특성 분석

채팅 특성	예
신조어, 도메인 특화 단어	혈창, 혈원, 잇섬, 유디, 드상, 축드상, 혈톡, 프리섬
발음 변형 및 축약	드가봐(들어가봐야), 강(그냥), 일루와(이리로 와), 젤루(제일로)
띄어쓰기 오류	단검이라스틴이음음, 이거88찍으세, 템제대로나오면
겹받침 탈락 및 오타	하겠어, 훔챜나, 햏쵸, 힘뽕어, 무엇이, 햏써
감정 표현 관련 기호	ㅋㅋㅋ, ㅌㅌ, ㅎㅎ, @@, —;
단음소 음절 및 낱자 표기	ㅇㅇ, ㅅㄱ, ㅍㅍㅍ, ㅅㅇㅅ, 그만 ㅈ ㅌ~~~~, ㅎㅇㅣ
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 발음이 같아 철자를 혼용하거나 의도적으로 발음을 예측해서 강조 문구로 사용하는 경우가 많이 발견되었다. 이런 현상은 특히 소통할 때 철자법을 체크하는 시간을 보내기보다 속도나 효율을 우선시 하기 때문이기도 하다.

- 많이[마니]: 마니(많이) 먹어라, 마니(많이) 아프냐
- 했따[헨따], 했지[헨찌] : 잘했지(잘했지), 말했지(말했지), 잘했따(잘했다), 망했따(망했다)
- 좋아[조아]: 조아짐(좋아짐), 조았어(좋았어), 아주 조아(좋아), 조아조아(좋아좋아)
- 조쿠먼(종구먼), 아랏어(알았어), 머찌다(멋지다), 마자요(맛아요)

맞춤법 및 철자 오류에 관해서는 언어가 사용되는 한 그 기준과 변화 수용에 대해 끊임없는 고민이 필요하다. 맞춤법 오류는 끊임 없이 발생 가능하고, 얼마나 희한한 철자가 입력되는지 보면 경이로울 정도이다. 자연스럽게 사



## 주제 6 음성 언어 처리, 어디까지 왔나?

용자들은 고빈도로 노출되는 문장을 정답으로 생각하게 되는데 마찬가지로 음성인식 결과가 내놓는 출력이 인식 오류뿐만 아니라 철자 오류 없이 결과를 내주는 것이 필요하다.

대화체 문장을 반영하기 위해 생각보다 많은 양의 입력 오류를 처리하고 필터링 해야 하며, 또한 게임을 주제로 한 대화가 주를 이루기 때문에 도메인 특화 작업에 필요한 신규 어휘 등록 및 분석 정교화 작업이 필요하다. 교과 서적인 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	저도침이라 저도침봐요 예형님	저/MM 도침/NNG 이/VCP 라/EC 저/NP 도/JX 침/MAG 봐요/VV+EC 예형/NNP 님/XSN	저/MP 도/JX 침/NNG 이/VCP 라/EC 저/NP 도/JX 침/MAG 봐요/VV+EC 예/IC 형/NNG 님/XSN
신조어/ 축약어	출책 푸귀	출/VV+ETM 책/NNP 푸/IC 귀/NNG ('푸른 귀걸이'의 약어)	출책/NNG 푸귀/NNG
어체 변형	안녕하세요 안녕하세요여 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 품사태그 사용: NNG(일반명사), NP(대명사), MM(관형사), VV(동사), VCP(궁정지정사), IC(감탄사), EC(연결어미), EF(종결어미), JX(보조사), EP(선어말 어미), XSV(동사파생접미사), XSN(명사파생접미사)  
(단, 형태소 분석기마다 결과가 다를 수 있음.)

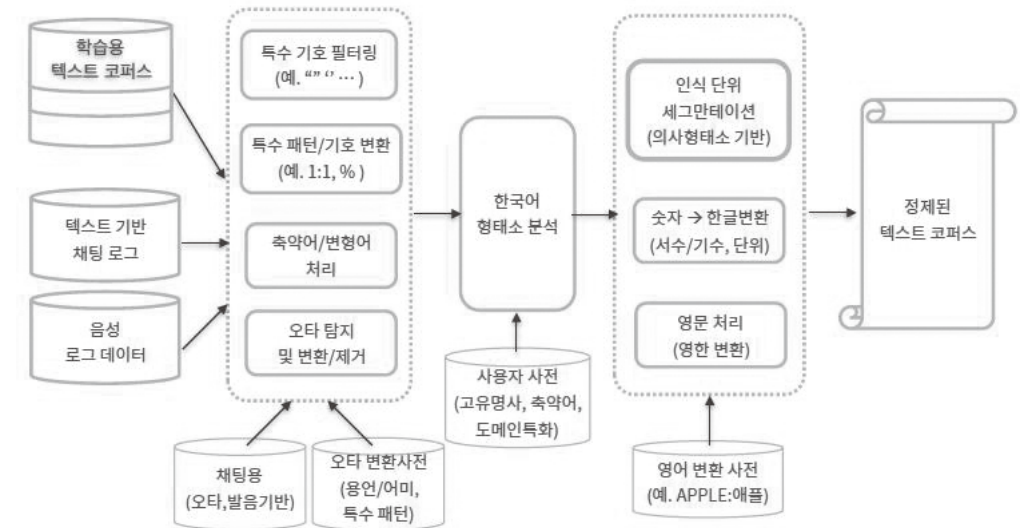
철자법에는 개인적인 감정이 포함되어 있지 않아 사용자의 감정과 상태를 구별하지 못한다. 감성 대화가 가능하게 하려면 양상(modality) 정보를 이해할 수 있어야 한다. 말투의 변화를 반영하려면 '안녕하세요'라는 인사말과 같은 경우 [표 5]와 같이 뒤에 용언의 활용어미로 “-하세여/-하세염/-하세움/-하세용/-하세웁/-하셈/” 등과 같이 구분할 수 있어야 한다.

[표 5] 발음 변형 기반의 이형태 표현

표준 입력	변형된 대체 표현
뭐하니	뭐하냐, 뭐하노, 뭐하냥, 뭐하나, 뭐하능, 뭐하누, 뭐하남, 모하니
반가워요	반가워용, 반가워여, 반가워움, 반가워요, 반가워용, 반가워여, 반가워유, 반가워염
안녕하세요	안녕하세여, 안녕하세용, 안녕하세움, 안녕하세염, 안녕하세유 안녕하세요, 안녕하세여, 안녕하세용, 안녕하세움, 안녕하세염 안냐세요, 안냐세여, 안냐세용, 안냐세움, 안냐세염

## 주제 6 음성 언어 처리, 어디까지 왔나?

기존 음성인식 시스템에서는 “-하세요”를 표준 어휘로 정하고, 발음사전에 허용 범위 내에서 발음의 다양성을 반영하는 방식으로 수행하였다. 이런 경우 명시적인 표출형 어휘와 암시적으로 발음 다양성을 발음사전에 넣어 처리하는 것 사이에 혼잡도를 어떻게 수용할 것인지 판단이 필요하게 된다. 등록 기준이 정해지면 발생 빈도에 기반한 통계 모델을 채택하는 것이 일반적일 것이다.



[그림 5] 언어모델 학습용 코퍼스 정제를 위한 텍스트 처리 프로세스

## 5. 앞으로의 발전 방향

한국어의 경우, 인식 단위로 의사 형태소를 사용하기 때문에 후처리 모듈에서 인식결과를 어절단위로 재구성하는 과정이 필요하며, 일반적으로 숫자나 영문의 경우 변환해주는 후처리 과정이 필요하다. 또한 음성인식 결과가 완벽하지 않기 때문에 오류 보정을 위한 노이즈 채널 모델과 같은 후처리 방식을 적용하여 그 정확도를 향상시킨다. 서비스 중에는 음성인식 엔진이 블랙박스이지만 선 순환적으로 언어모델에 후처리 보정 기술을 반영한다면 후처리에서 수행해야 할 부담을 줄이고 인식 성능의 향상도 가져갈 수 있을 것이다.

여전히 대화체 음성인식이 어려운 이유는 ‘그러니까’, ‘음’, ‘아침’ 등등 헤아릴 수 없이 많은 간투사가 수시로 사용되며, 더듬거림, 어휘의 도치 현상, 동일 어휘의 반복이나 어휘적 단락(끊어짐), 재발성 등으로 인한 비문법적인 비정형 발성이 빈발함에 기인하는데, 이와 같은 비정형 자연어(unstructured spontaneous speech) 처리는 학습을 통한 모델링으로는 여전히 한계가 있다.

## 주제 6 음성 언어 처리, 어디까지 왔나?

직관적으로 봐도 기존의 어휘적 쓰임새를 통계적 지식에 의존해서 처리해야 하는 메커니즘으로는 해결이 어렵기 때문에 새로운 방식의 언어 모델이 필연적으로 개발되어야 한다. 이러한 통계적 방식의 단점을 극복하고 비정형 자연어를 효과적으로 인식하기 위해 현재 다양한 딥러닝 기술이 활발하게 연구되고 있다.

### 참고문헌

- [1] 컴퓨터 정보용어대사전, 한국사전연구소 (언어정보처리, 음성정보처리)
- [2] 권오욱, 최승권, 노윤형, 김영길, 박전규, 이윤근, “자유발화형 음성대화처리 기술동향”, Electronics and Telecommunications Trends Vol. 30, No. 4, 26–35, Aug, 2015.
- [3] 박전규, “인공지능 기술 개발 어디까지 왔나? 딥러닝 기반의 음성인식 기술”, 컴퓨터월드, 2016.
- [4] 김상훈, 조재원, “말하는대로 통역에서 비서까지, 음성인식 기술”, 융합연구리뷰, Vol.3 No.6, 2017.
- [5] 과학기술정책연구원, “지능형 개인비서 시장 동향과 국내 산업 영향 전망”, 동향과 이슈 제 35호, 2017.
- [6] 김학수, “인공지능 음성언어 비서 시스템의 자연언어처리 기술들”, 정보과학회지 제35권 제8호, 2017.8.
- [7] Dahl, George E., et al. "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," IEEE Transactions on audio, speech, and language processing 20.1 (2012): 30–42.
- [8] Hinton, Geoffrey, et al. "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," IEEE Signal Processing Magazine 29.6 (2012): 82–97.
- [9] Amodei, Dario, et al. "Deep speech 2: End-to-end speech recognition in english and mandarin," International Conference on Machine Learning, 2016.

## 주제 6

# 음성 언어 처리, 어디까지 왔나?

· 토론자 정민화(서울대학교 언어학과 교수)



## 주제 6: 음성 언어 처리, 어디까지 왔나?

### 토론 자료

2017년 11월 10일

서울대학교 언어학과

정민화 교수

▪ Email: mchung@snu.ac.kr



## 최근 음성 언어 처리 기술의 발전 요인

### 1. 빅 & 리얼 데이터

- Broadcast News data
  - 128M word corpus, 143K word vocabulary
- Google's Web 1T 5-gram Version 1 (2006)
  - 1T word corpus from 95B sentences, 13.6M word vocabulary
- Big & Real Data
  - Big & Real Data from services on cloud computing platforms
  - From Interspeech 2017 Tutorial (Björn W. Schuller, Nicholas Cummins)

Big Data.



Volume	Velocity	Variety
13 TB / 8300 h	GB/h	video, audio
350 mio tweets/day	real-time	diff. resol. / format
1.3 bio users	crawled	social data feed
130 mio web pages	every 48 h	text



## 최근 음성 언어 처리 기술의 발전 요인

2. 빅 컴퓨팅 파워
  - Cloud Computing
  - GPGPU (General Purpose Graphic Processing Unit)
3. 딥 러닝 & 오픈 소스 환경
  - 각종 딥 러닝 구조
    - DNN(Deep Neural Networks), Deep RNNs(Recurrent Neural Networks), Deep CNNs(Convolutional Neural Networks), Deep LSTM (Long Short Term Memory) Networks, ...
  - End-to-End Approaches
    - 음성인식, 음성합성, 대화처리, 기계번역, ...
  - 지식의 역할에 대한 도전 & 기회

3

## 우리말 정보화를 위한 대응방안은?

1. 빅 & 리얼 데이터 ?
2. 빅 컴퓨팅 파워 ?
3. 딥 러닝 & 오픈 소스 환경 ?

4





