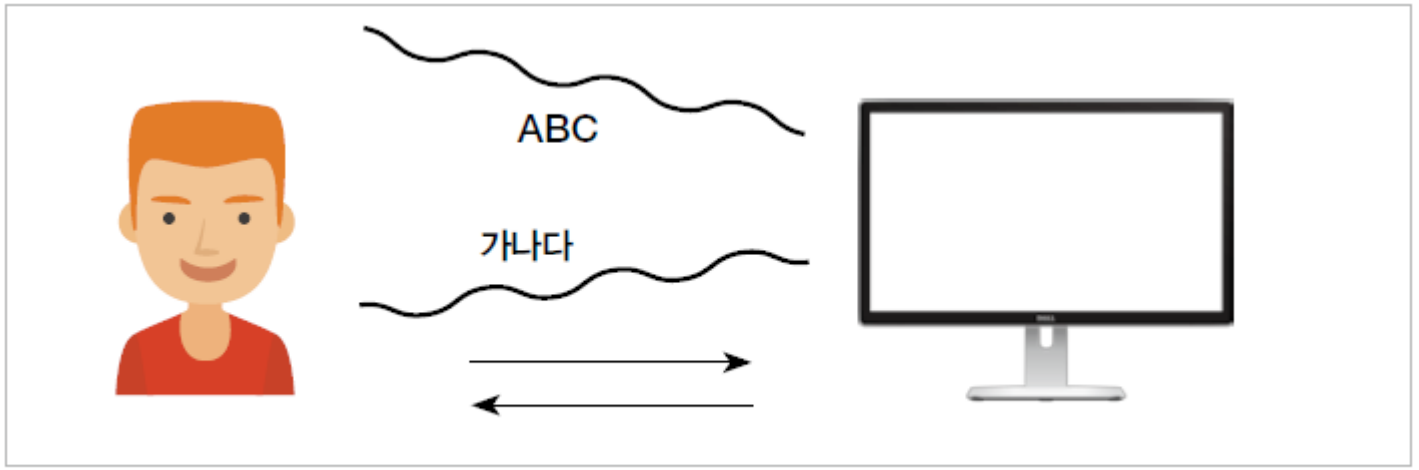
1장 딥러닝을 활용한 자연어 처리 개요

목 차

1.1 자연어 처리란 무엇일까?

- 자연어 처리 (Natural Language Processing)
  - 사람의 언어를 컴퓨터가 알아듣도록 처리하는 인터페이스
  - 컴퓨터공학 + 언어학 지식이 필요
- 자연어 처리의 최종 목표
  - 컴퓨터가 사람의 언어를 이해하고 여러 가지 문제를 수행
- 자연어 처리의 대표적인 과제 및 응용 분야
  - 대량의 텍스트를 이해하고 수치화 (감성 분석 – Sentiment Analysis)
  - 사용자의 의도를 파악하거나 대화하거나 도움 (애플의 시리, 아마존의 알렉사)
  - 요약 (summarization), 기계 번역 (machine Translation)
  - 사용자가 원하는 것을 검색 및 답변

1.1 자연어 처리란 무엇일까?



▶ 텍스트는 사람과 컴퓨터 사이의 가장 훌륭한 인터페이스

- 딥 러닝에 의해 비약적인 발전
  - 딥러닝의 기반이 되는 기술들을 이해
  - 이전 기술의 문제점 파악
  - 최신 기술이 어떤 돌파구를 마련했는지 파악
  - 딥러닝 이전의 주요 기술을 정리

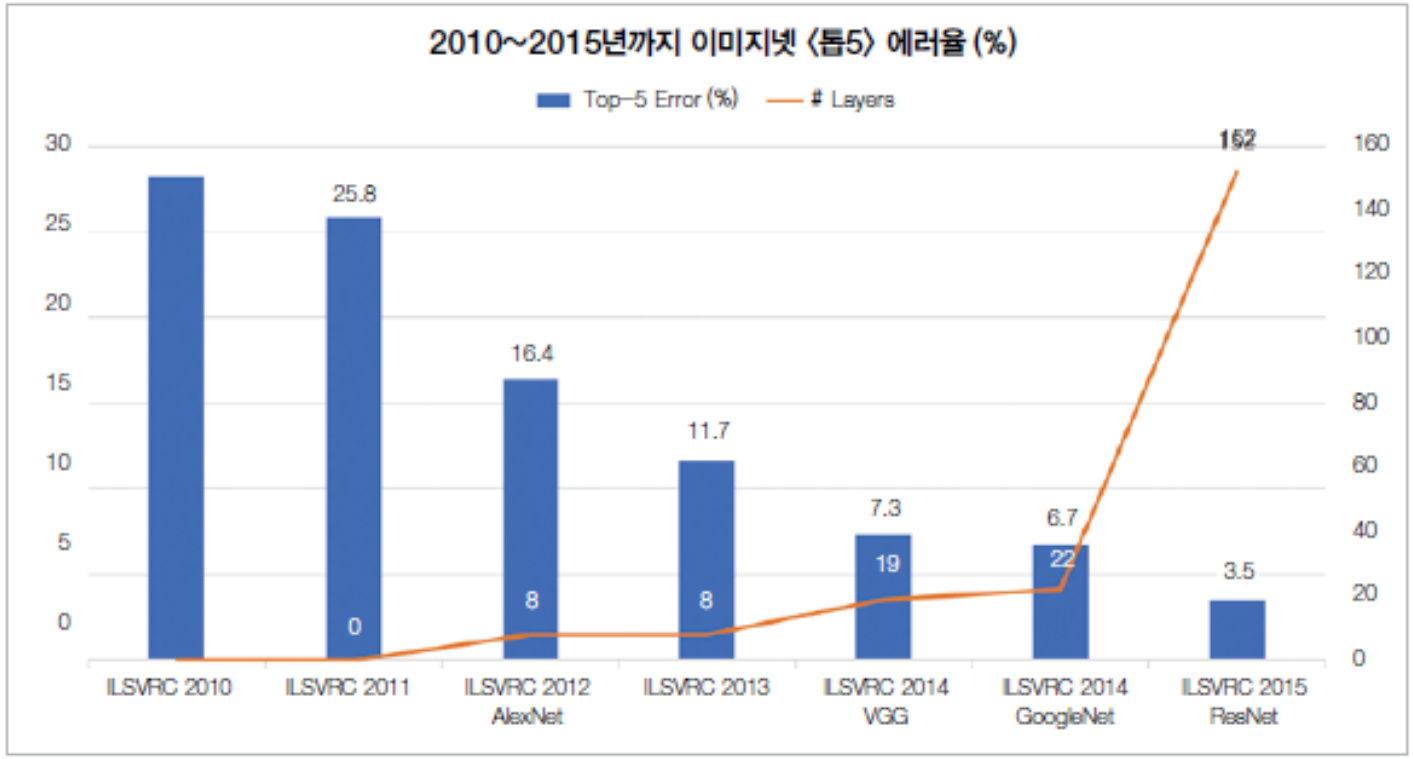
1.2 딥러닝 소개

- 이미지 분류 (이미지넷) 에서 가장 먼저 두각을 나타냄
- 음성 인식 분야에서 최초로 상용화
  - 가우시안 혼합 모델 (Gaussian Mixture Model, GMM) -> 심층 신경망 (Deep Neural Networks, DNN)
- 자연어 처리는 상대적으로 늦게 효과를 보임
  - 단어 간의 순서 및 상호 정보가 반영된 시퀀셜 데이터 (Sequential Data)
  - 어텐션 (Attention) 메커니즘으로 기계번역 (end-to-end 방식)

1.2.1 딥러닝의 역사

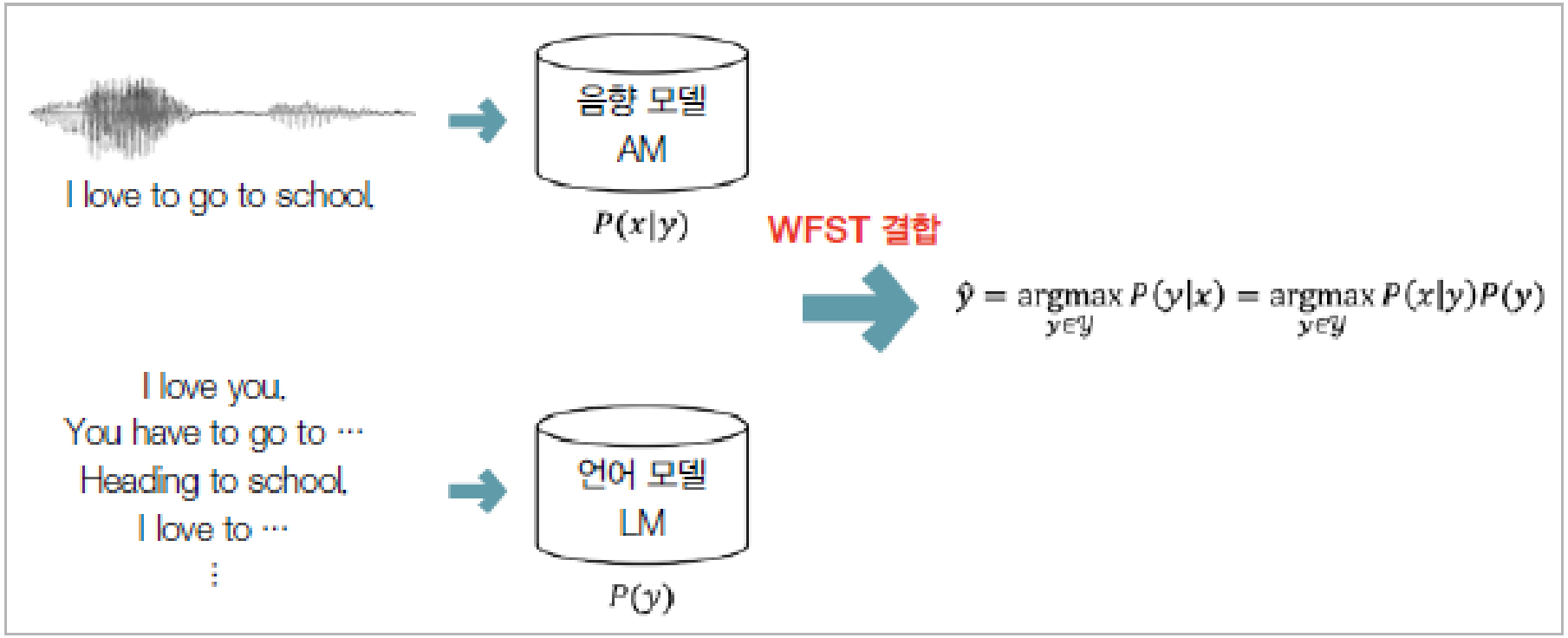
- 2010년 이전
  - 1950년대
    - 인공지능의 개념이 최초로 설립
    - 두번의 대유행과 두번의 빙하기
  - 1980년대
    - 역전파 알고리즘 개발로 신경망 학습이 가능해 짐
    - 데이터의 한계 및 컴퓨팅 파워의 제약으로 성능에 한계
    - Support Vector Machine 등의 기계 학습 알고리즘에 비해 낮은 성능
  - 2006년
    - 제프리 힌튼 (Geoffrey Hinton) 이 딥 빌리프 네트워크 (Deep Belief Networks, DBN) 개발
    - 은닉층을 효과적으로 사전 훈련 (pre-training)
    - 이후에도 오랜 세월을 기다려야 했음

1.2.1 딥 러닝의 역사 (이미지 분류)



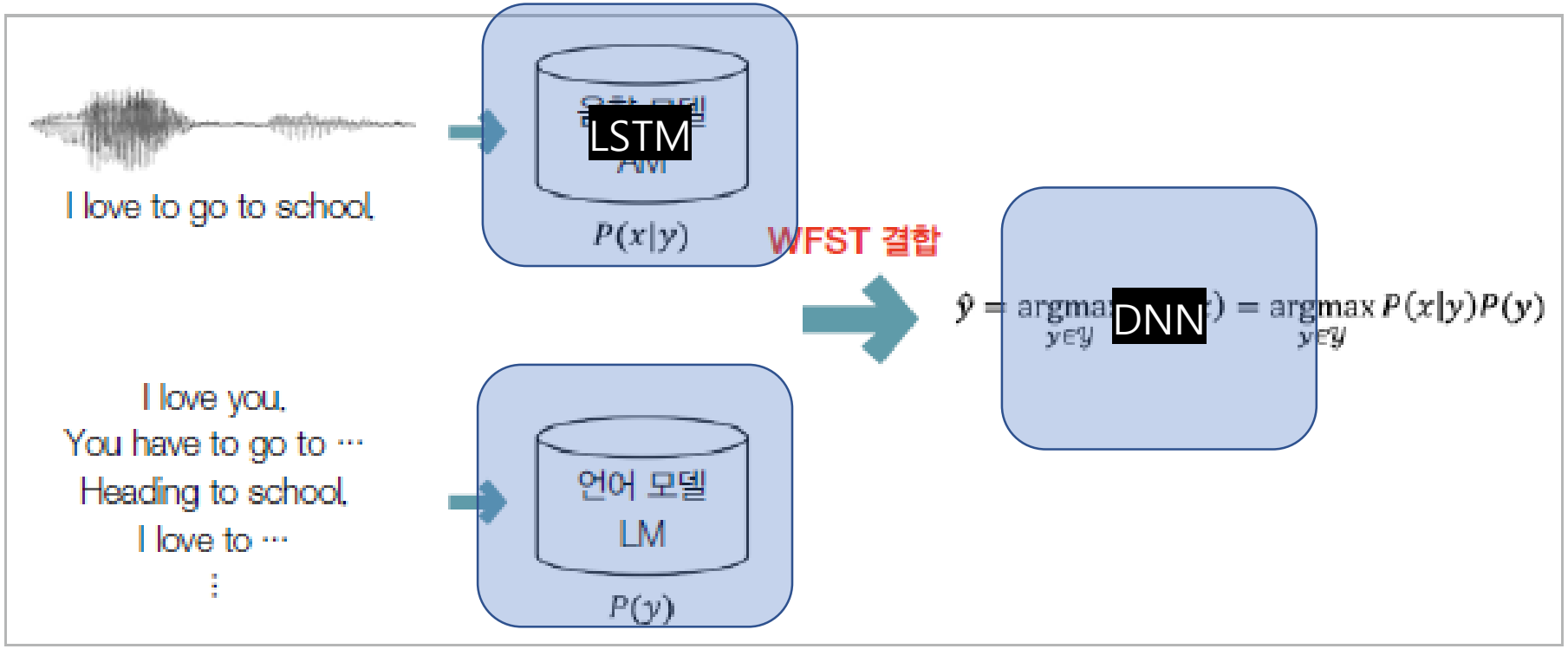
▶ 이미지넷의 최근 성능 변화

1.2.1 딥 러닝의 역사 (음성 인식)



▶ 전통적인 자동음성인식 시스템의 구성

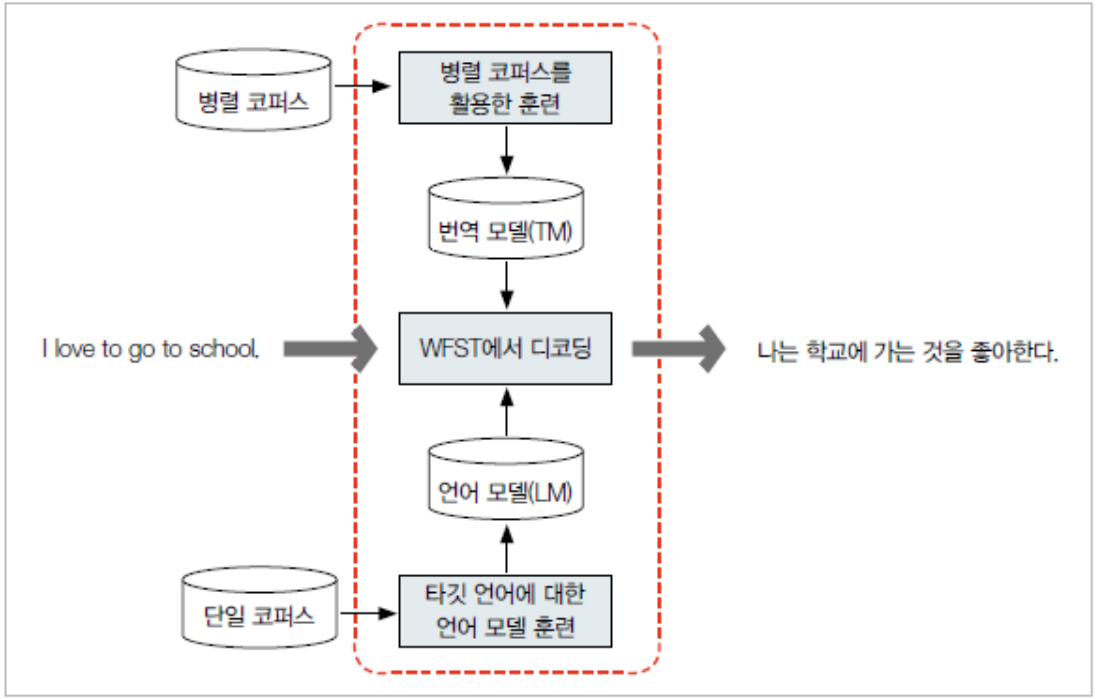
1.2.1 딥 러닝의 역사 (음성 인식)



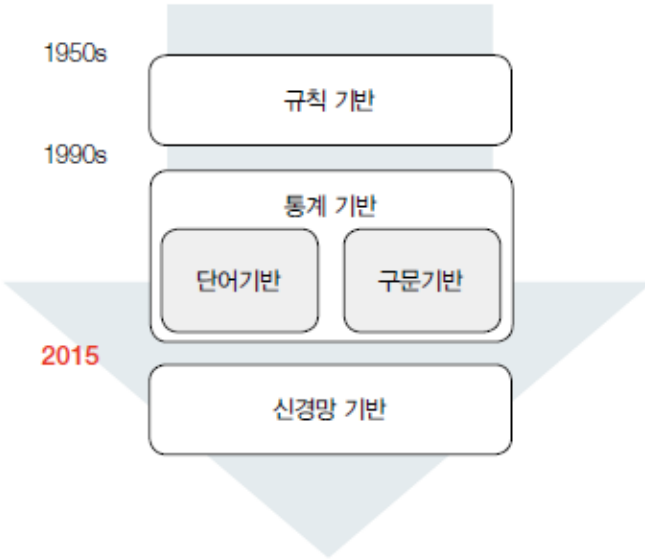
▶ 전통적인 자동음성인식 시스템의 구성

1.2.1 딥 러닝의 역사 (기계번역)

- 규칙 기반 기계 번역 (Rule-Based Machine Translation) -> 통계 기반 기계 번역 (Statistical Machine Translation)

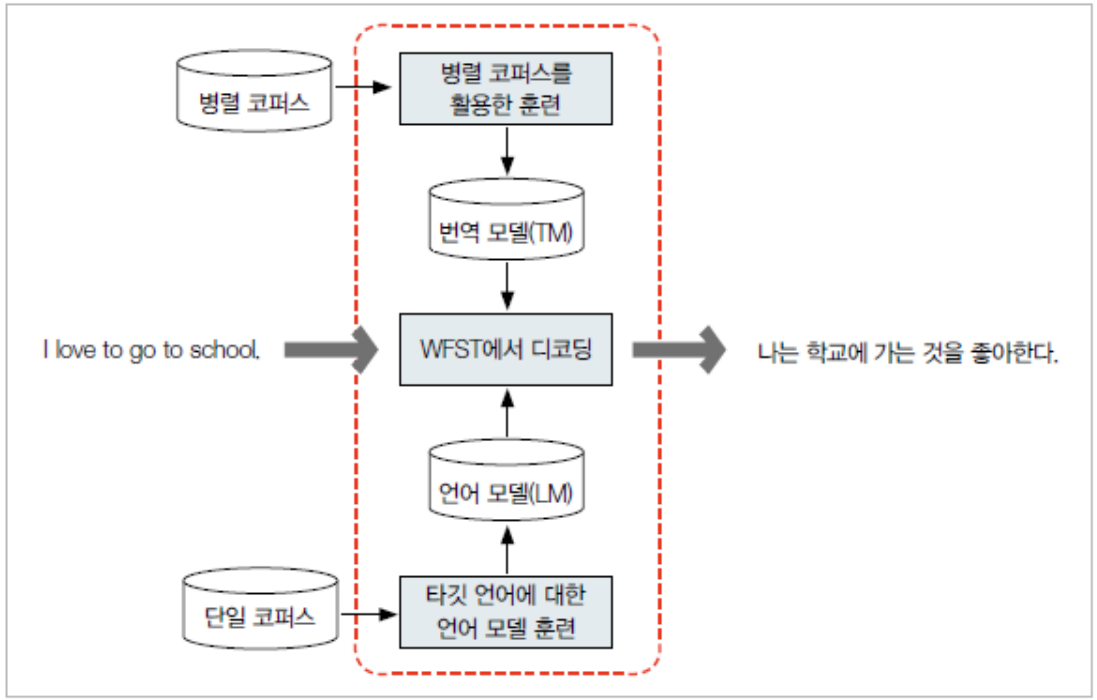


▶ 통계 기반 기계번역 시스템을 구성하는 서브 모듈

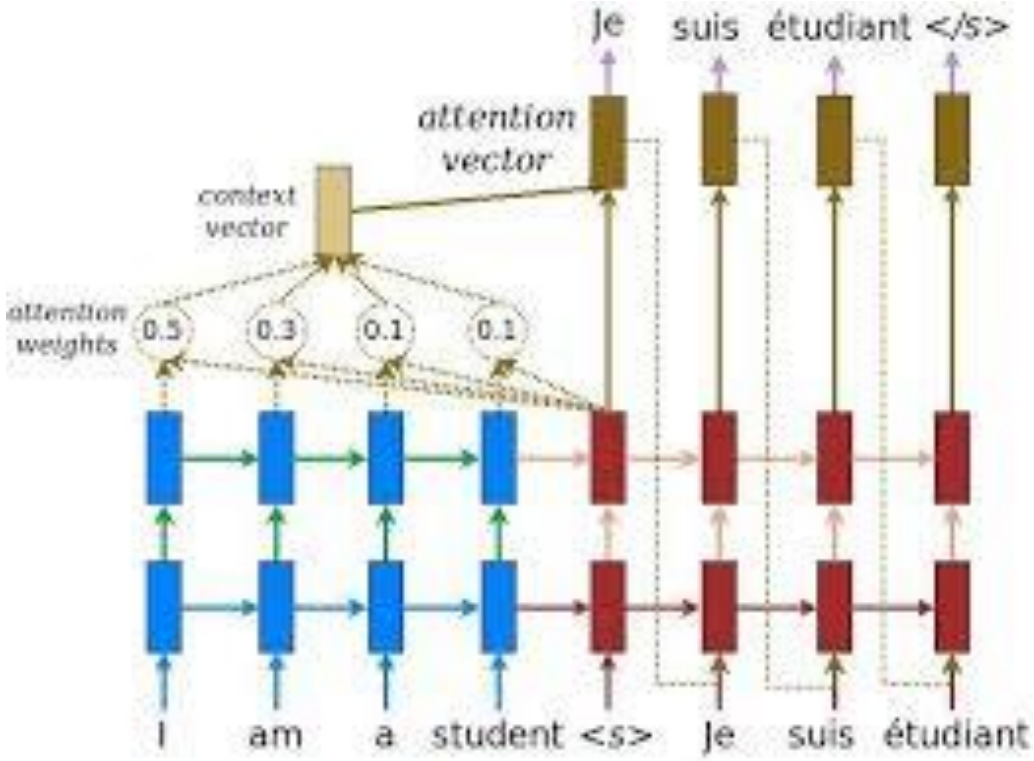


▶ 기계번역의 역사[2]

1.2.1 딥 러닝의 역사 (기계번역)



▶ 통계 기반 기계번역 시스템을 구성하는 서브 모듈



한림대학교 HALLYM UNIVERSITY

NLP Lab
Natural Language Processing

1.2.1 딥 러닝의 역사 (생성 모델 학습)

- 생성 모델 학습 (Generation Model Learning)

• 판별 모델 학습

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\ P(Y \mid X;\theta)$$

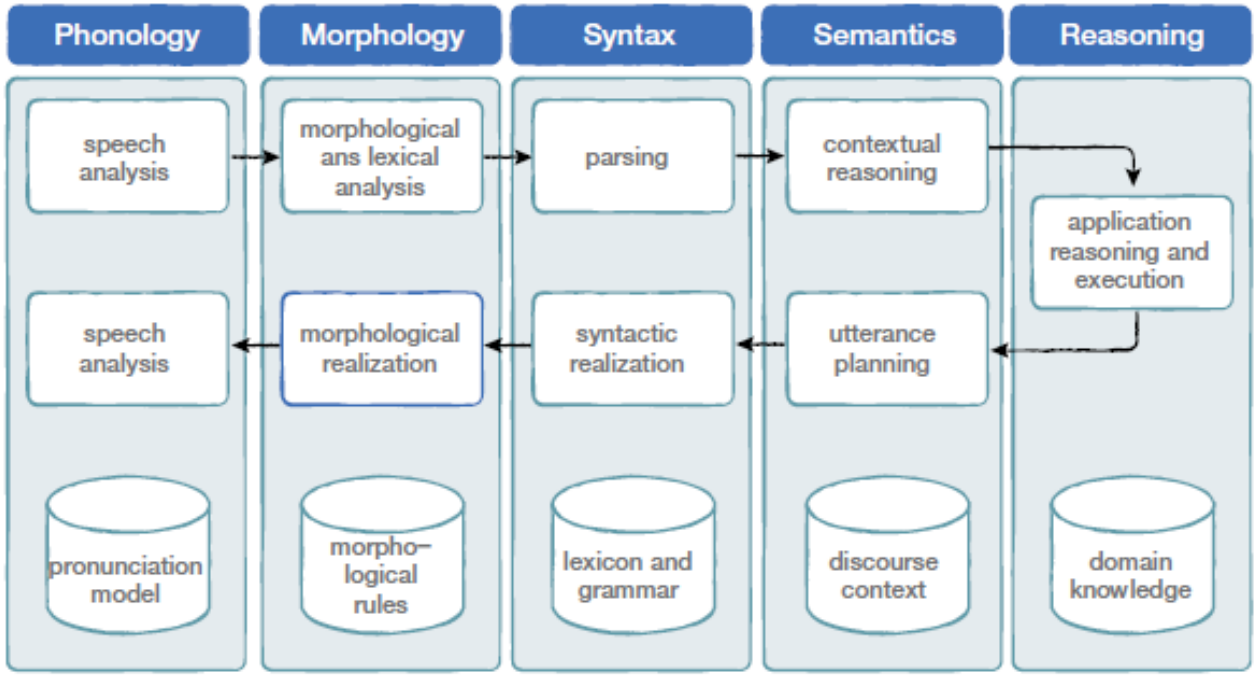
• 생성 모델 학습

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\ P(X;\theta)$$

적대적 학습 (Adversarial Learning)
변분 오토 인코더
(Variational Autoencoder, VAE)

▶ 2014년부터 2018년까지 생성적 적대 신경망(GAN)의 발전 사례[3]

1.2.2 자연어 처리의 패러다임 변화



▶ 자연어 처리를 위한 딥러닝 도입[13]

여러 가지 단계의 모듈로 구성
추가적인 서브 모듈 필요
매우 무겁고 구현 및
시스템 구성이 어려움
오차의 전파 (Error Propagation) 발생

1.2.2 자연어 처리의 패러다임 변화

- 딥 러닝의 도입
  - 하위 모듈 대체에서 end-to-end 모델로 전체 문제를 해결
  - 챗봇 등에서는 아직 end-to-end 도입이 어려움

전통적인 심볼릭 기반 접근 방법	딥러닝 기반 접근 방법
이산적(discrete), 심볼릭 공간	연속적(continuous), 신경망 공간
사람이 인지하기 쉬움	사람이 이해하기 어려움
디버그 용이	디버깅 어려움
연산 속도 느림	연산 속도 빠름
모호성과 유의성에 취약함	모호성과 유의성에 강인함
여러 서브 모듈이 폭포수 형태를 취하므로 특징 추출에 노력이 필요함	end-to-end 모델을 통한 성능 개선과 시스템 간소화 가능

▶ 전통적인 자연어 처리 접근 방식과 딥러닝을 통한 자연어 처리 접근 방식의 차이[13]

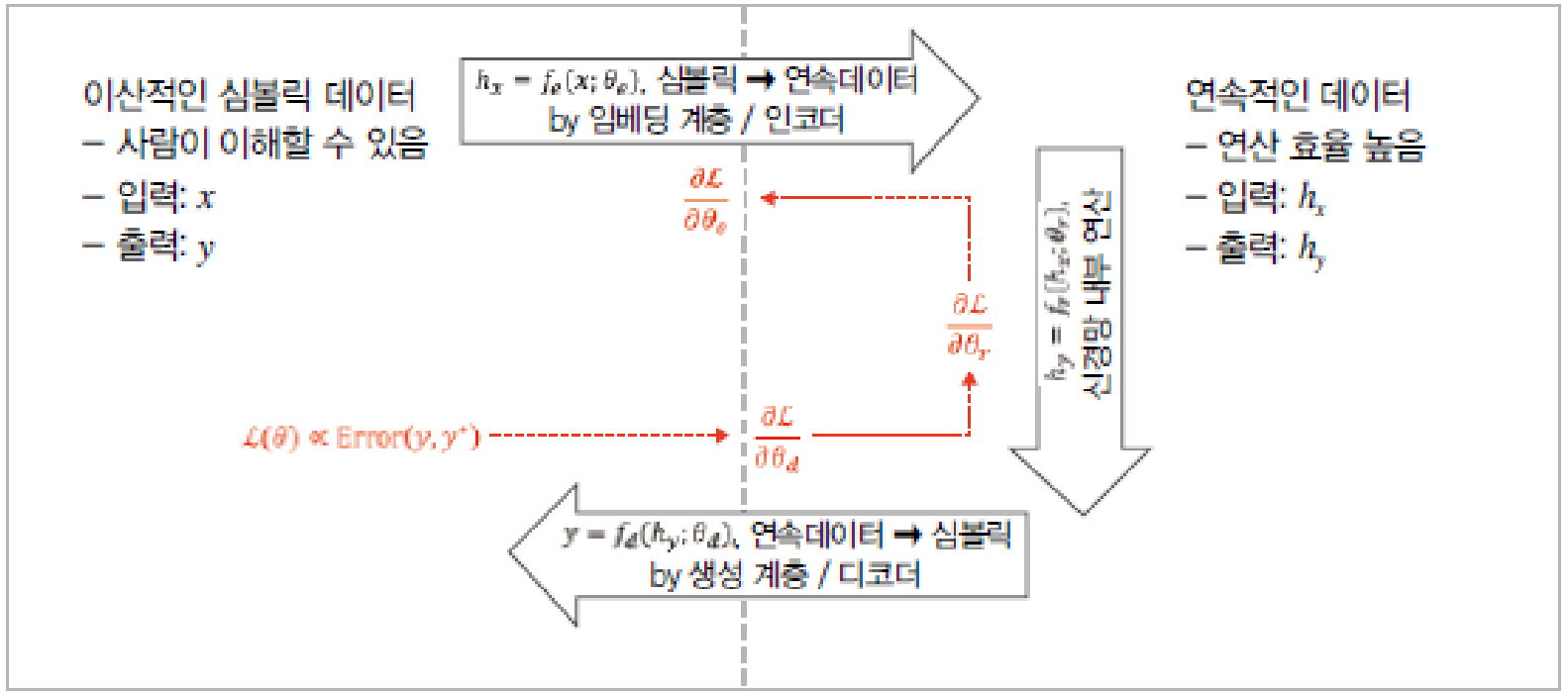
1.2.2 자연어 처리의 패러다임 변화

- RNN, LSTM, Attention 등이 도입

심볼릭(아산, discrete) 공간	신경망(연속, continuous) 공간
지식의 표현 방법 – 단어, 관계, 템플릿 – 고차원, 이산적, 희소 벡터 형태	지식의 표현 방법 – 간소화 및 일반화 된 지식 그래프 (Knowledge Graph, KG) – 저차원, 연속적, 꽉 찬 벡터 형태
추론(inference) – 거대한 지식 그래프(Knowledge Graph, KG)로 인한 속도 감소 – 키워드 또는 템플릿 매칭에 민감함	추론(inference) – 적은 메모리를 요구하며, 빠르게 동작함 – 키워드 또는 템플릿 매칭에 강인함
사람이 이해할 수 있으나, 연산 효율성 낮음	연산 효율이 높지만, 사람이 이해하기에 어려움

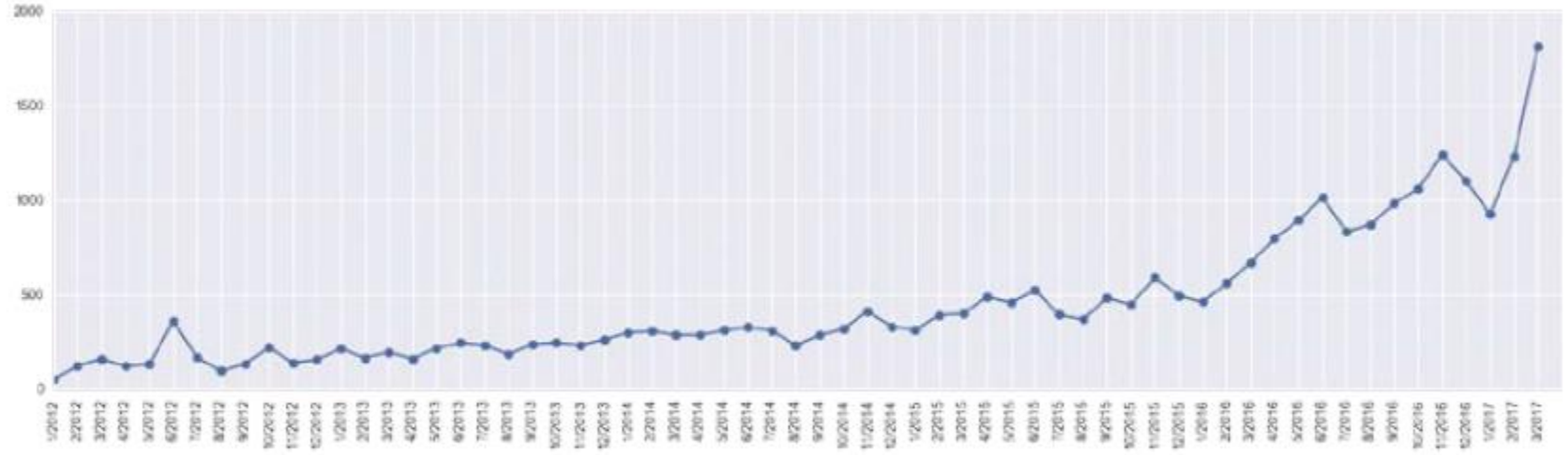
▶ 자연어 처리를 위한 심벌릭 데이터와 연속적인 데이터의 특징[13]

1.2.3 정리



▶ 딥러닝 자연어 처리 과정 요약[13]

1.2.3 정리



▶ 아카이브에 출판된 딥러닝 관련 논문 수의 급격한 증가를 볼 수 있습니다.[4]

1.3 왜 자연어 처리는 어려울까?

- 음성인식
  - 눈에 보이지 않는 신호를 다룸
  - 노이즈와 신호가 더해져서 나타남
  - 샘플링 주기가 짧아서 데이터의 길이가 매우 김
- 영상 처리 분야
  - 이미지 데이터는 너무 크고 다양함
  - 사람눈엔 다 똑같은 색도 컴퓨터에게 다를 수 있음
- 자연어처리
  - 불연속적인 단어들로 구성
  - 사람은 언어를 통해 타인과 교류하고 의견과 지식을 전달
  - 사람은 언어를 통해 지식을 축적
  - 언어는 사람의 생각과 지식을 내포
  - 컴퓨터가 사람의 언어를 이해한다면 컴퓨터에 사람의 지식과 의견 전달이 가능

1.3.1 모호성

원문	차를 마시러 공원에 가던 차 안에서 나는 그녀에게 차였다.
G 사	I was kicking her in the car that wend to the park for tea
M 사	I was a car to her, in the car I had a car and went to the park.
N 사	I got dumped by her on the way to the park for tea.
K 사	I was in the car going to the park for tea and I was in her car.
S 사	I got dumped by her in the car that was going to the park for a cup of tea.

▶ 업체별 한영 번역 사례 비교

G사 : In the car on the way to the park for tea, I was teased by her.
M사 : In the car I was going to the park for tea, I was a car to her.
N사 : I was dumped by her in the car on my way to the park for tea.
K사 : I was in the park for tea and I was in the car.

1.3.1 모호성

- 문장 내 정보의 부족으로 인한 모호성

원문	나는 철수를 안 때렸다.
해석# 1	철수는 맞았지만, 때린 사람이 나는 아니다.
해석# 2	나는 누군가를 때렸지만, 그게 철수는 아니다.
해석# 3	나는 누군가를 때린 적도 없고, 철수도 맞은 적이 없다.

▶ 문장 내 정보 부족에 따른 구조 해석 사례 1

원문	선생님은 울면서 돌아오는 우리를 위로했다.
해석# 1	(선생님은 울면서) 돌아오는 우리를 위로했다.
해석# 2	선생님은 (울면서 돌아오는 우리를) 위로했다.

▶ 문장 내 정보 부족에 따른 구조 해석 사례 2

1.3.2 다양한 표현

1. 골든 리트리버 한 마리가 잔디밭에서 공중의 원반을 향해 달려가고 있습니다.
2. 원반이 날아가는 방향으로 개가 뛰어가고 있습니다.
3. 개가 잔디밭에서 원반을 쫓아가고 있습니다.
4. 잔디밭에서 강아지가 프리스비를 향해 뛰어가고 있습니다.
5. 높이 던져진 원반을 향해 멍멍이가 신나게 뛰어갑니다.
6. 노란 개가 원반을 잡으러 뛰어가고 있습니다.

1.3.3 불연속적 데이터

- 딥러닝에 적용하기 위하여 연속적인 값으로 바꿔줘야 함
  - 단어 임베딩으로 바꿔줌
  - 여러 제약이 존재

- 차원의 저주
  - 많은 종류의 데이터를 표현하려면 데이터 종류만큼 엄청난 차원이 필요
  - 어휘의 크기만큼의 차원이 필요
  - 희소성 문제 발생
  - 단어 임베딩으로 적절한 수의 차원으로 축소

1.3.3 불연속적 데이터

- 노이즈와 정규화
  - 데이터에서 노이즈를 신호로부터 적절히 분리해해는 일은 매우 중요
  - 데이터가 살짝 바뀌어도 의미 변화가 크다
  - 단어가 바뀌면 전체 의미가 달라짐
  - 띄어쓰기, 어순 등의 문제

1.4 무엇이 한국어 자연어 처리를 더욱 어렵게 만들까?
1.4.1 교착어

- 교착어
  - 어근(의미) + 접사(문법)로 구성
  - 굴절어 : 어근의 형태가 바뀜
  - 고립어 : 어순이 중요

종류	대표적 언어	특징
교착어	한국어, 일본어, 몽골어	어간에 접사가 붙어 단어를 이루고 의미와 문법적 기능이 정해짐
굴절어	라틴어, 독일어, 러시아어	단어의 형태가 변함으로써 문법적 기능이 정해짐
고립어	영어, 중국어	어순에 따라 단어의 문법적 기능이 정해짐

▶ 교착어와 굴절어, 고립어의 특징

1.4.1 교착어

- 어근 + 접사가 다양한 형태로 결합
  - 파싱, 형태소 분석, 언어 모델링 등에서 문제를 어렵게 만듬
  - 접사로 인하여 비슷한 의미의 단어가 다수 발생
  - 분절을 통하여 어근과 접사를 분리

원형	피동	높임	과거	추측	전달		결과
잡						+다	잡다
잡	+히					+다	잡히다
잡	+히	+시				+다	잡히시다
잡	+히	+시	+었			+다	잡히셨다
잡			+았(었)			+다	잡았다
잡				+겠		+다	잡겠다
잡					+더라		잡더라
잡	+히		+었			+다	잡혔다

1.4.1 교착어

- 어순은 상대적으로 중요하지 않음

번호	문장	정상여부
1	나는 밥을 먹으러 간다.	O
2	간다 나는 밥을 먹으러.	O
3	먹으러 간다 나는 밥을.	O
4	밥을 먹으러 간다 나는.	O
5	나는 먹으러 간다 밥을.	O
6	나는 간다 밥을 먹으러.	O
7	간다 밥을 먹으러 나는.	O
8	간다 먹으러 나는 밥을.	O
9	먹으러 나는 밥을 간다.	X
10	먹으러 밥을 간다 나는.	X
11	밥을 간다 나는 먹으러.	X
12	밥을 나는 먹으러 간다.	O
13	나는 밥을 간다 먹으러.	X
14	간다 나는 먹으러 밥을.	O
15	먹으러 간다 밥을 나는.	O
16	밥을 먹으러 나는 간다.	O

1.4.2 띄어쓰기

- 동양권 언어에서 띄어쓰기는 근대에 들어서면서 도입됨
- 띄어쓰기에 대한 표준이 계속 변화
- 추가적인 분절을 통해 띄어쓰기를 정제해주는 과정이 필요

1.4.3 평서문과 의문문

- 의문문과 평서문이 같은 형태의 문장 구조를 가짐
  - 마침표나 물음표가 붙지 않으면 구분이 안됨
  - 음성 인식의 결과물로 나오는 텍스트는 더욱 어려움

언어	평서문	의문문
영어	I ate my lunch.	Did you have lunch?
한국어	점심 먹었어.	점심 먹었어?

▶ 평서문과 의문문 비교(영어 vs 한국어)

1.4.4 주어 생략

- 영어
  - 명사의 역할이 중요
  - 주어가 생략되는 경우가 없음
- 한국어
  - 동사의 역할이 중요
  - 주어가 자주 생략
- 컴퓨터의 한계
  - 문맥 정보를 잘 활용하여 생략된 정보를 메울 수 있는 능력이 없음
  - 주어가 생략되면 서술어의 주체가 누구인지 알 수 없음
  - 기계번역 등에서 문장의 정확한 의미 파악이 어려워짐

1.4.5 한자 기반의 언어

- 한자의 조합으로 이루어지는 단어가 많다.
  - 각 단어들의 의미가 합쳐져 하나의 단어의 의미를 구성
  - 영어에서 라틴어 기반의 단어들은 서브워드들이 함쳐져 하나의 단어의 의미를 구성

언어	단어	조합
영어	Concentrate	con(= together) + centr(= center) + ate(= make)
한국어	집중(集中)	集(모을 집) + 中(가운데 중)

▶ 서브워드의 조합으로 이루어지는 단어 사례

  - 한글이 한자를 대체하면서 문제 발생
    - 표의문자인 한자가 표음문자인 한글로 바뀌면서 정보의 손실 발생
    - 인간은 문맥을 통해 정보의 손실 문제를 해결
    - 다른 언어에 비해 모호성 문제가 심각

1.4.5 한자 기반의 언어

- 서브워드 단위의 분절이 중의성을 가중시킴

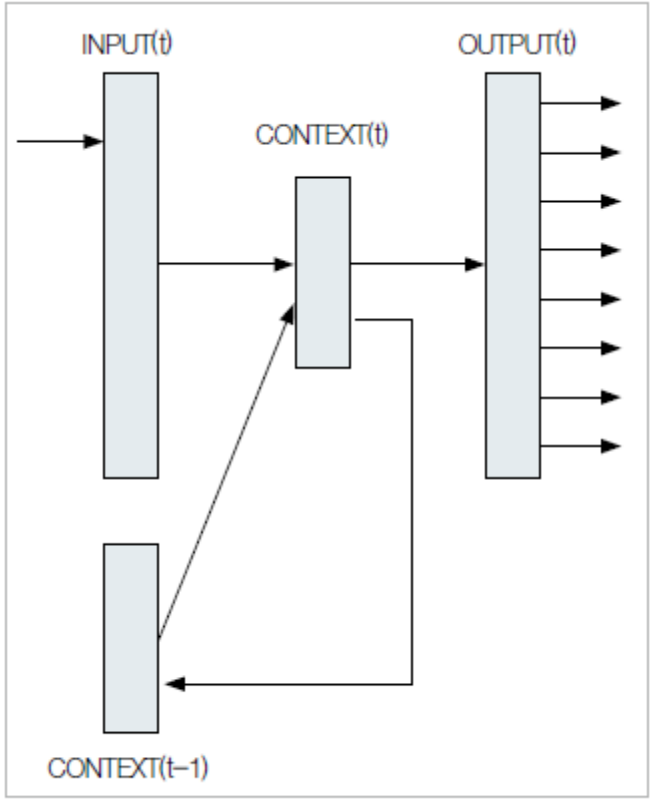
Type	Text
원문	저는 여기 한 가지 문제점이 있다고 생각합니다.
형태소에 따른 분절	저 는 여기 한 가지 문제점 이 있 다고 생각 합니다 .
출현 빈도 기반 서브워드 분절	_저 _는 _여기 _한 _가지 _문 제 점 _이 _있 _다고 _생각 _합니다 _.

▶ 서브워드 단위로 분절할 경우 가중되는 중의성 문제 사례

  - 문제점(問題點) 이 '문', '제', '점' 으로 분절될 경우
    - 제(題) 를 임베딩하는 경우 아래의 글자들과 구분될 수 없다. (표음문자니까)
      - 결제 (決濟) 의 제(濟)
      - 제공 (提供) 의 제(提)
    - '제' 라는 토큰이 임베딩 벡터로 변환시 문제 발생
      - 題, 濟, 提 를 하나로 임베딩
      - 각각의 의미의 평균값으로 임베딩

1.5 자연어 처리의 최근 추세
1.5.1 딥 러닝의 자연어 처리 정복 과정

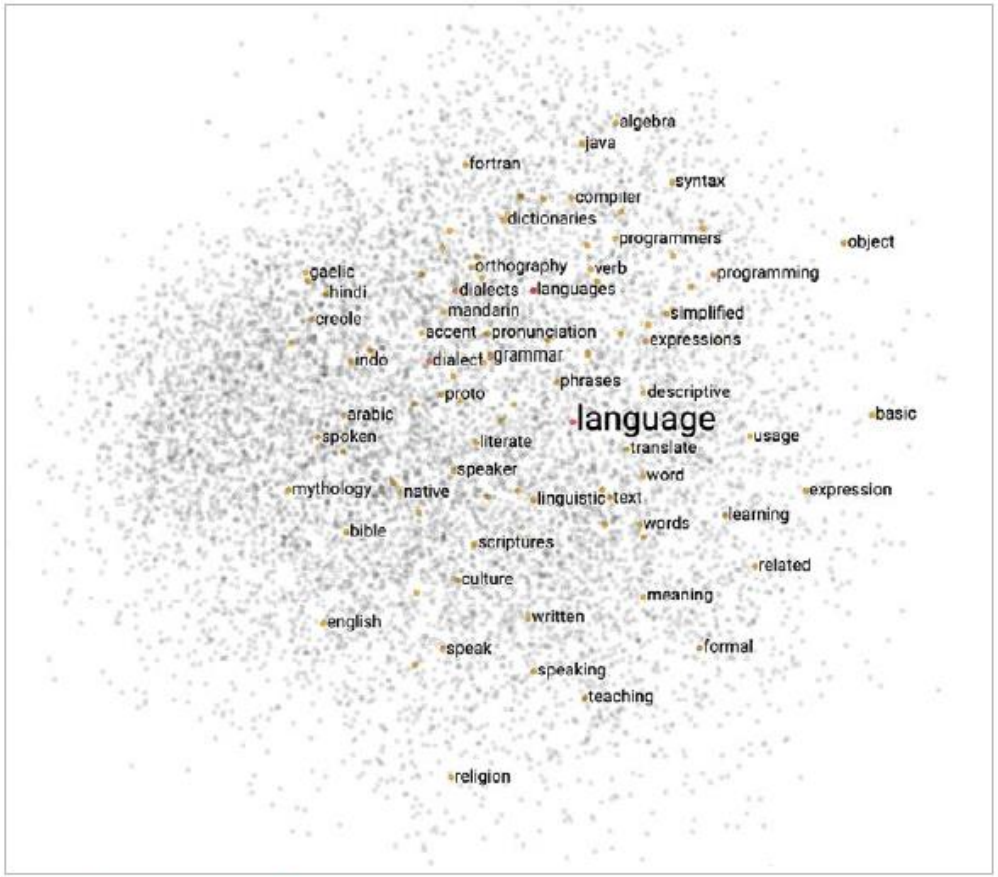
- RNN을 활용한 언어 모델링 시도 (2010)
  - n-gram 방식과의 결합하여 성능 향상
  - 음성 인식과 기계번역 분야에 적용시 구조적인 한계와 높은 연산량으로 성과가 저조



▶ 언어 모델링을 위한 초기 형태의 단순한 순환 신경망(RNN)[40]

1.5.1 딥 러닝의 자연어 처리 정복 과정

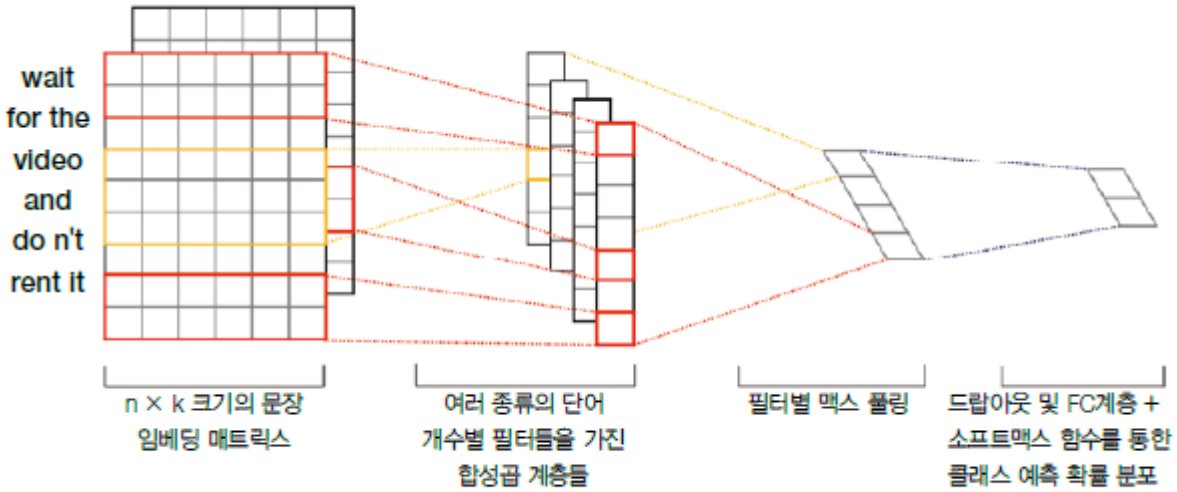
- word2vec 개발 (2013, Thomas Mikolov)
  - 단순한 구조의 신경망을 사용
  - 단어들을 잠재 공간 (latent space)에 투사
  - 비슷한 단어들은 가깝게 위치
  - 딥러닝 네트워크 내부 동작 원리 파악



▶ 단어 임베딩 벡터 시각화의 예[7]

1.5.1 딥 러닝의 자연어 처리 정복 과정

- CNN 으로 텍스트 분류 (Yoon Kim, 2014)

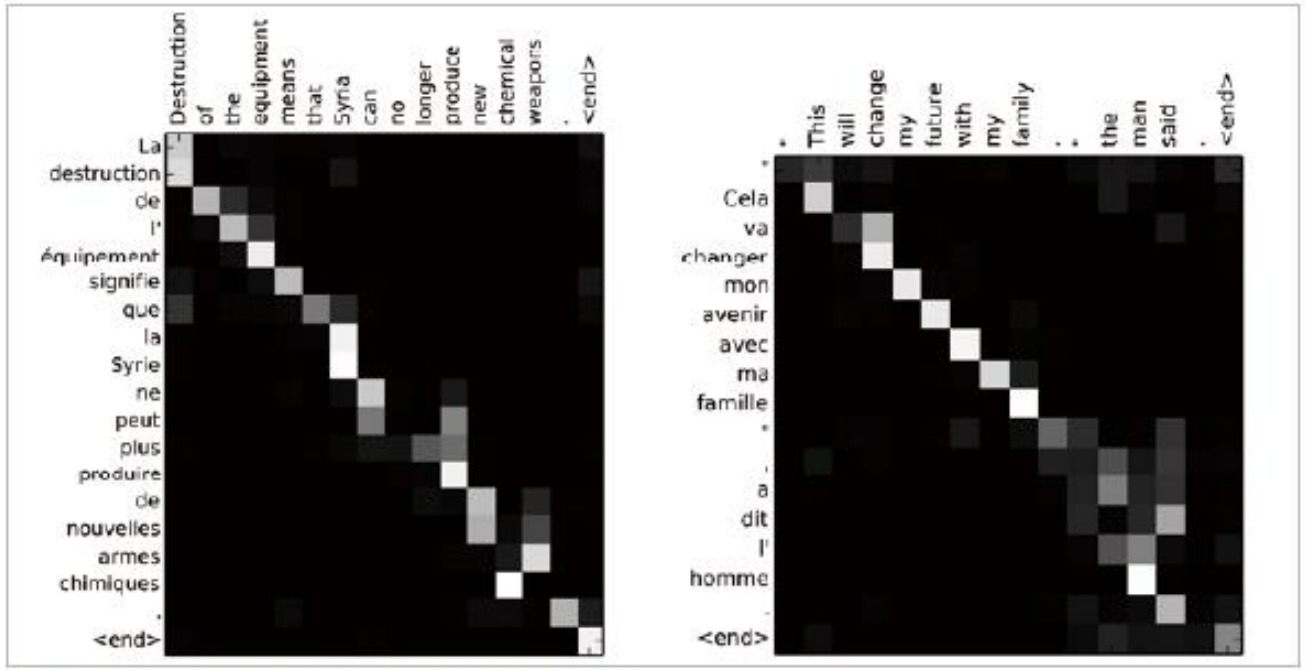


▶ CNN 텍스트 분류 모델 [30]

- 딥 러닝으로 형태소 분석, 문장 파싱, 개체명 인식, 의미역 결정 등의 언어처리 문제 해결
- 대부분의 문제가 end-to-end 모델로 해결책을 모색하는 방향으로 연구가 진행

1.5.2 자연어 생성의 시작

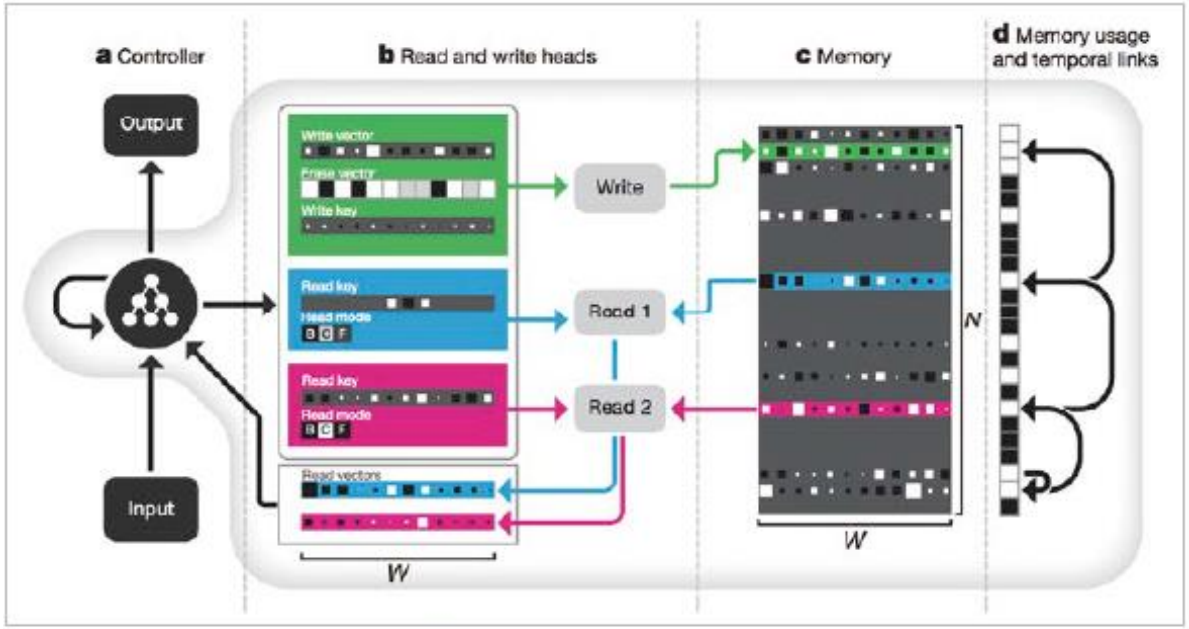
- 자연어 생성 (Natural Language Generation) 이 가능 (2014)
  - seq2seq + attention -> NMT (Neural Machine Translation)
  - 주어진 정보에 기반하여 자유롭게 문장을 생성 -> 자연어 생성
  - 기계 번역, 챗봇, 요약 등



▶ 어텐션을 통해 두 문장 사이의 단어간 정렬이 된 모습[3]

1.5.3 메모리를 활용한 심화 연구

- 뉴럴 튜링 머신 (Neural Turing Machine: NTM)
  - 여러 주소에서 연속적으로 정보를 읽고 쓰는 방법을 제시
- 디퍼런셜 뉴럴 컴퓨터 (Differential Neural Computer: DNC) (Google DeepMind)
  - NMT의 활용



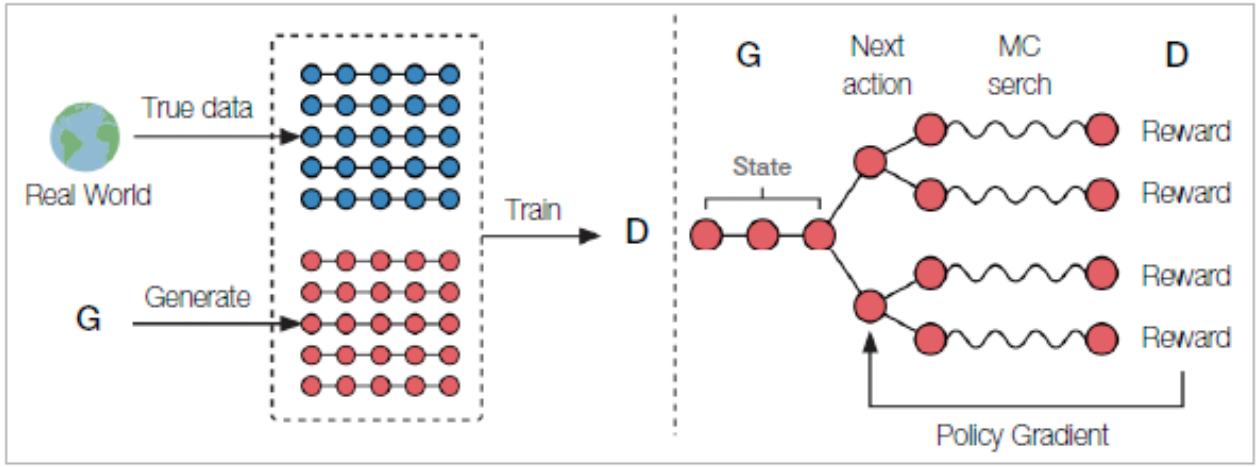
▶ 다이나믹 뉴럴 컴퓨팅(DNC)의 구조도[18]

1.5.3 메모리를 활용한 심화 연구

- Memory-Augmented Neural Network (메모리 증강 신경망)
  - 신경망을 통해 메모리를 활용
  - 원하는 정보를 신경망을 통해 저장하고 조합하여 활용 가능
  - QA 문제에 효과적

1.5.4 강화학습의 자연어 처리 분야에 대한 성공적인 적용

- 자연어 처리에서는 생성 모델 학습에 관심이 저조
  - 언어 모델 자체가 생성 모델
- 손실 함수 (loss function) 와 목적 함수 (object function) 의 괴리
  - 강화학습을 활용하여 SeqGAN 등 제시
  - 폴리시 그래디언트 (policy gradient) 적용



▶ 폴리시 그래디언트를 적용한 'seqGAN'의[61]