AiDA Lab.
Artificial Intelligence & Data Analysis Laboratory

01. 자연어 처리 개요

Seokhwan Yang

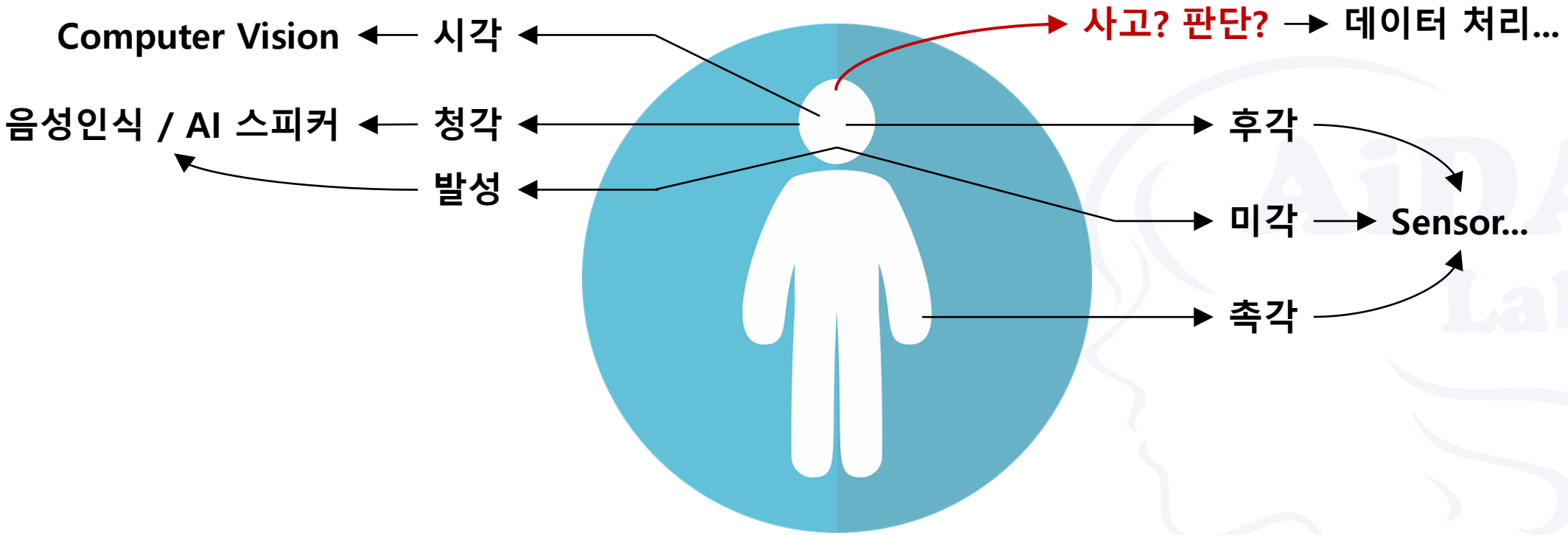
AI와 자연어 처리

자.. 그럼 여기서...

- 자연어 처리는 앞에서 살펴본 특징과 무엇이 다를까?
- 그리고 자연어 처리가 왜 중요할까?

현재, 일반적으로 볼 수 있는 AI 기술은..

- 하나의 개체(인간)가 가지는 생물학적인 기능을 흉내 낸 것



그런데... 인간은 사회적인 동물이다!!!

- 인간은 크고 작은 사회, 조직 안에서 지식과 지혜를 세대를 거쳐 누적 시킴으로써 문명을 이루었고, 본능과 다른 지성을 성립시켜왔다.

- 이 과정에서 사람과 사람 사이의 소통, 커뮤니케이션이 요구되었고

- 이런 커뮤니케이션을 위해 발생한 것이 언어, 즉 자연어이다.

자연어의 발생은...

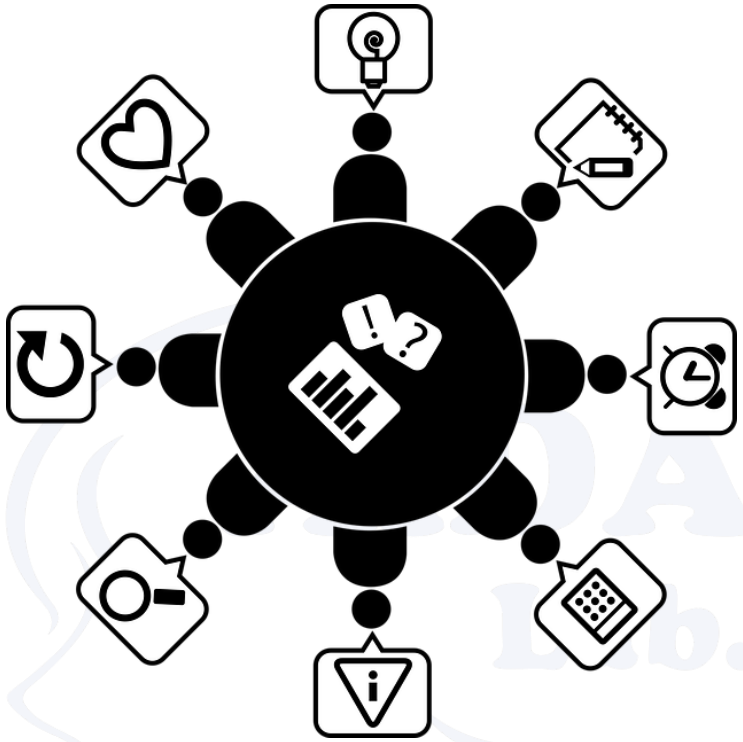
- 인간이 스스로의 지식을 구조화 할 수 있게 진화 시켰다.

  → 인간의 사고 행위는 언어(자연어)로 구성됨

- 언어의 발생
  - 인간의 진화는 매우 더디게 진행 → 언어의 발생과 함께 폭발적으로 진화
  - 언어 발생 이전의 사고 → 단순한 동물이 보이는 기계적인 반응에 그침
  - 언어를 사용하면서 → 지식의 구조화 성립, 현대적인 의사소통 능력 발생
- 인간의 사고 능력은 생물학적 기능의 모방으로 구현할 수 없으며 또한 사고 능력은 언어를 기반으로 성립함

AI와 자연어 처리

- 더불어...
- 인간의 지능, 지성은 복합적으로 구성됨
  - 한 사람, 하나의 개체에만 적용되는 지능, 지성
  - 사회를 구성하는 집단에서 발생하는 집단 지성



- 이러한 모든 것을 연결하는 핵심이 언어(자연어) 이다.

자연어 처리 개요

자연어(Natural Language)

- 프로그래밍 언어와 같이 사람이 인공적으로 만든 언어가 아닌,
- 사람이 일상생활과 의사소통에 사용해 온, 한국어, 영어와 같이 오랜 세월에 걸쳐 자연적으로 만들어진 언어라는 의미
  - 우리가 흔히 말하는 언어를 뜻함
- 자연어라고 부르는 까닭
  - 컴퓨터공학 등에서는 언어라고 하면 우선적으로 C 등의 프로그래밍 언어를 떠올리기 때문

자연어 처리(NLP)

- 컴퓨터가 인간의 언어를 알아들을 수 있도록 인간의 언어를 분석하고 해석하여 처리하는 인공지능의 한 분야
- 자연어를 컴퓨터로 해석하고, 의미를 분석하여 이해하고, 자동으로 생성하는 것 등에 관련된 분야

- 자연어 처리의 가장 기본적인 문제(? 목표?)는
  - "어떻게 자연어를 컴퓨터에게 인식시킬 수 있을까?" 라는 것이다.

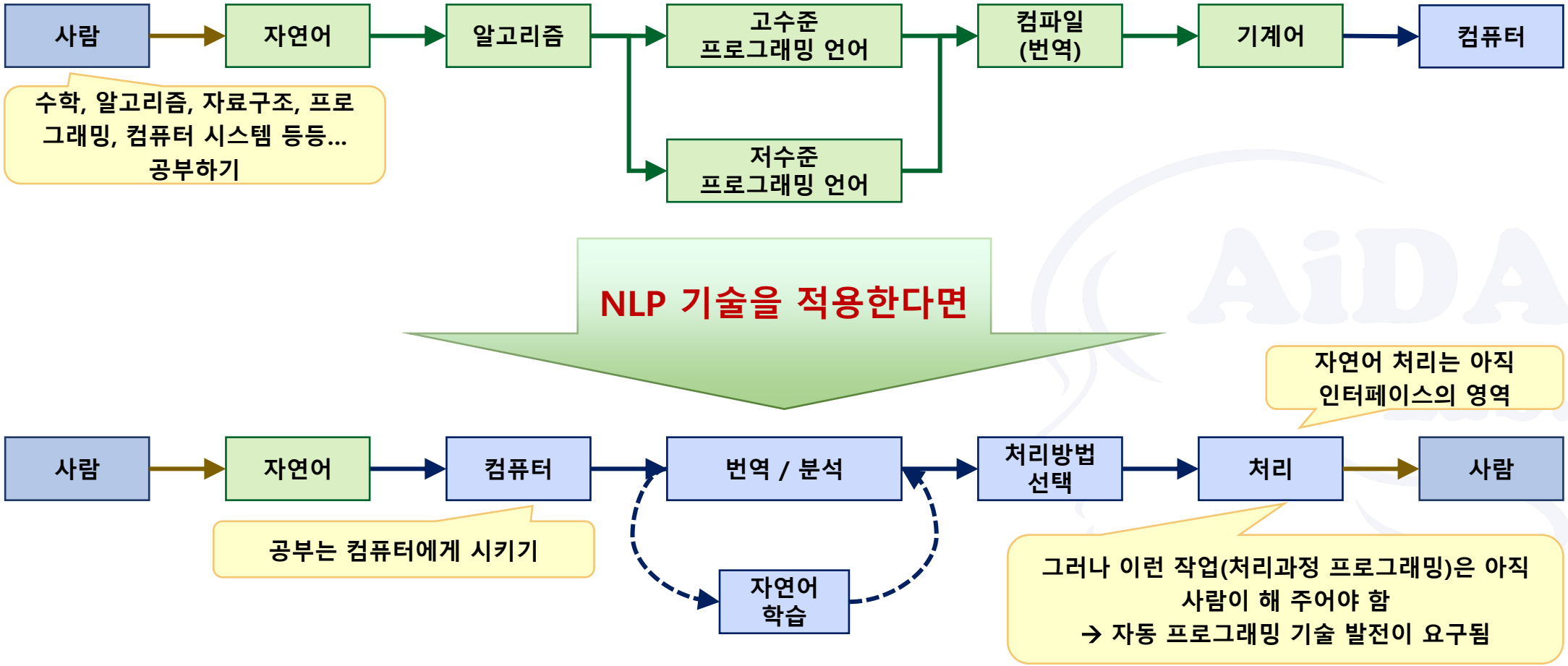
자연어 처리(NLP)

- 기존의 컴퓨터 관련 기술에서는
  - 컴퓨터에게 일을 시키기 위하여 사람이 컴퓨터의 언어를 학습함

- AI 기술을 사용하면...
  - 컴퓨터에게 일을 시키기 위하여 컴퓨터가 사람의 언어를 학습함

자연어 처리(NLP)

- 컴퓨터에게 일을 시키는 과정



자연어 처리(NLP)

- 자연어에 대한 연구는 오래전부터 이어져 오고 있음에도 불구하고 현재까지도 아직 컴퓨터가 자연어를 사람처럼 이해하지 못함
- 그러나 언어에 대한 깊은 이해없이 피상적인 확률 및 통계를 이용하여 대량의 정보를 처리하는 기술은 많이 발전한 상태임
- 대표사례
  - 검색 엔진: 인간의 언어를 깊이 이해하지 않고, 단어 간의 통계적 유사성에 바탕을 두고 문서를 검색함

자연어 처리(NLP)

- 자연어 처리에서 요구되는 기술들
  - 인공지능 기술
  - 컴퓨터 공학 기술
  - 언어학적 지식

광범위한 기술 분포로 쉽게 접근하기 어려움
그러나 제대로 기술을 익히고자 한다면..
또는 연구/개발에 투자하고자 한다면...

→ 필수

- 단순히 라이브러리를 사용한 개발만 한다면..
  - 다양한 자연어 처리용 라이브러리의 활용법 학습 요구

자연어 처리의 대표적인 주제

- 우리는 자연어 처리를 통해 어떤 문제를 해결하고자 하는가?
  - 텍스트 분류
  - 텍스트 유사도
  - 자연어 생성 (→ 텍스트 생성)
  - 기계 이해 (→ 텍스트 이해)

자연어 처리의 대표적인 주제

- 4대 문제를 이해하려면 먼저 "**단어 표현**"에 대하여 이해하여야 함

단어 표현

- 단어표현
  - 모든 자연어 처리 문제의 기본
  - 자연어를 어떻게 표현할 것인지 결정하는 것이 문제에 대한 해결의 출발점

단어 표현

- 컴퓨터의 텍스트 인식
  - 컴퓨터가 인식할 수 있는 것은 구성요소인 전자 소재의 “On/Of” 상태 뿐!!!
  - 이러한 “On/Off” 상태를 사람의 기준에서 좀 더 보기 쉽게, 그리고 산술적(수학적)으로 표현하기 쉽게 하기 위하여 0, 1의 값으로 대체(변환)하여 표현
  - 예
    - 언: 1100010110111000 (UTF-8 코드: C5B8)
    - 어: 1100010110110100 (UTF-8 코드: C5B4)

} 언어의 특징이 모두 사라짐

→ 자연어 처리 모델을 위한 데이터로 사용 불가

단어 표현

- 그러나…
  - 컴퓨터는 자연어를 이해하지 못하며
  - 전기 신호의 On/Off의 2진수 표현을 기반으로 한 수치 데이터만 사용 가능
- 따라서
  - 언어의 기본 단위가 되는 “단어”의 수치화가 필요함
  - 단어의 수치화 결과는 벡터로 표시 → 단어 임베딩, 단어 벡터 등으로 표현

단어 표현

- **단어의 수치화를 위한 대표적인 방법**
  - **원-핫 인코딩**
    - 각 단어에 인덱스 부여 → 각 단어의 벡터에서 해당 인덱스를 1, 나머지는 0
      - 아버지, 어머니, 삼촌, 이모, 고모: 길이 5의 벡터

        → (1,0,0,0,0), (0,1,0,0,0), (0,0,1,0,0), (0,0,0,1,0), (0,0,0,0,1)
  - **분포가설 기반의 인코딩**
    - 분포가설: 같은 문맥의 단어, 즉 비슷한 위치에 나오는 단어는 비슷한 의미를 가짐

      → 어떤 글에서 비슷한 위치에 존재하는 단어는 단어 간의 유사도가 높다고 판단
    - 카운트 기반 방법, 예측 방법

단어 표현

- **분포가설 기반의 인코딩**
  - 카운트 기반 방법
    - 특정 문맥 안에서 단어들이 동시에 등장하는 횟수를 직접 세는 방법

      → 동시에 등장한 횟수를 행렬로 나타내고, 그 행렬을 수치화하여 단어벡터를 만드는 방법
    - 종류: 특잇값 분해, 잠재 의미 분석, Hyperspace Analogue to Language(HAL), Hellinger PCA(Principal Component Analysis, 주성분 분석) 등
  - 예측 방법
    - 신경망 또는 통계모델 등을 통해서 문맥 안의 단어들을 예측하는 방법
    - 종류: Word2Vec, NNLM(Neural Network Language Model), RNNLM(Recurrent NNLM) 등
    - 대표적인 방법인 Word2Vec은 CBOW(Continuous Bag of Words), Skip-Gram 모델로 구분됨

단어 표현

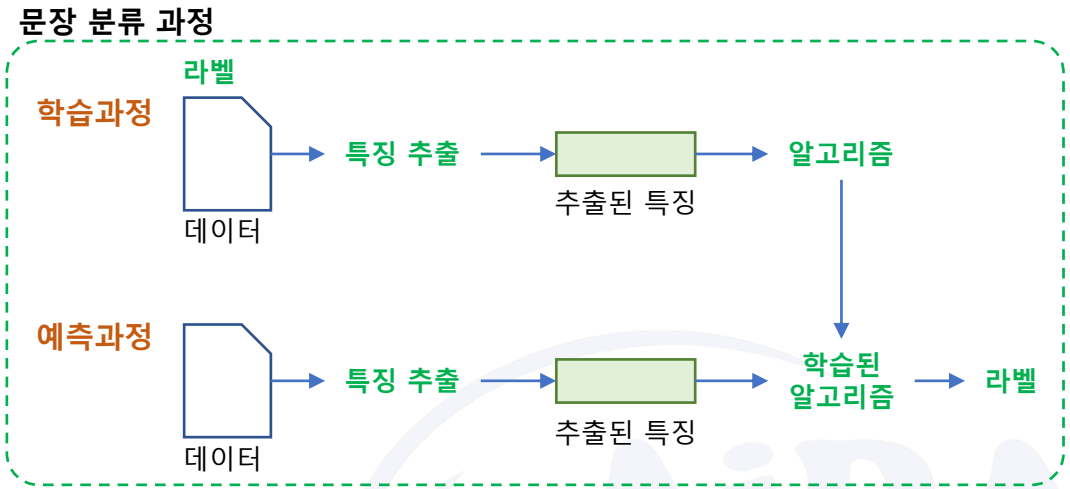
- **단어 벡터화 방법의 장단점**
  - **원-핫 인코딩**
    - 간편하고 이해하기 쉬움
    - 처리할 단어의 수가 많을 수록 벡터의 크기 급증 → 매우 비효율적(희소벡터)
    - 단어가 무엇인지만 확인 가능 → 단어의 의미, 특성, 관계 등의 정보는 표현되지 않음
  - **분포가설 기반의 인코딩**
    - 각 단어 사이의 유사도 및 관계성, 특징 등을 표현할 수 있음
    - 특히 예측 방법의 경우 성능이 뛰어나므로 가장 많이 활용됨

텍스트 분류

- 텍스트 분류 문제는?
  - 자연어 처리 문제 중 가장 대표적이고 많이 접하는 문제
  - 자연어 처리 기술을 활용해 특정 텍스트를 사람들이 정한 몇 가지 범주(Class) 중 어느 범주에 속하는지 분류하는 문제
- 텍스트 분류 예시
  - 스팸(SPAM) 분류, 감정 분류, 뉴스 기사 분류

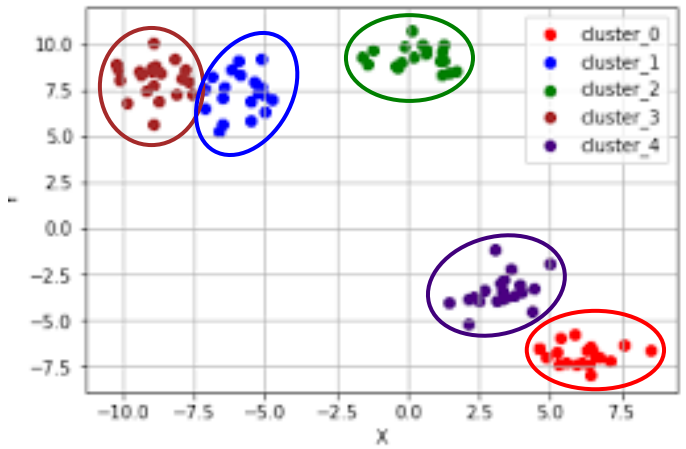
텍스트 분류

- **텍스트 분류 방법**
  - **지도학습을 통한 분류**
    - 선형 분류(Linear Classifier)
    - 신경망(Neural Network)
    - 나이브 베이즈 분류(Naive Bayes Classifier)
    - 서포트 벡터 머신(Support Vector Machine, SVM)
    - 로지스틱 분류(Logistic Classifier)
    - 랜덤 포레스트(Random Forest) 등.. 다수

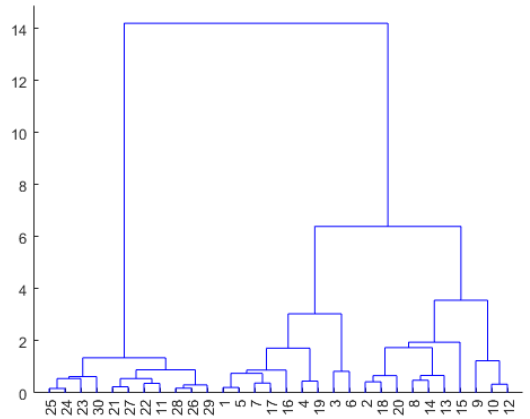


텍스트 분류

- **비지도학습을 통한 분류**
  - k-평균 군집화(k-means Clustering)
  - 계층적 군집화(Hierarchical Clustering)



k-평균 군집화



계층적 군집화

텍스트 유사도

- 텍스트가 얼마나 유사한지를 표현하는 방법
- 예시
  - 이 노래 누가 만들었어?
  - 지금 나오는 노래의 작곡가가 누구야?

사람은 두 문장이 같은 뜻이라는 것을 쉽게 이해하지만... 컴퓨터는 이해하지 못함

- 예시와 같은 두 문장이 얼마나 유사한지 정량화 하여 수치로 표현하는 모델을 만드는 것이 중요함

텍스트 유사도

- 유사도 측정을 위한 정량화 방법
  - 같은 단어의 개수를 사용하여 판단하는 방법
  - 형태소로 나누어 형태소를 비교하는 방법
  - 문장, 단어를 자소 단위로 나누어 각 단어를 비교하는 방법
  - 딥러닝을 기반으로 측정하는 방법 등...
- 많이 사용되는 유사도 측정 방법(4개)
  - 자카드 유사도, 유클리디언 유사도, 맨해튼 유사도, 코사인 유사도

자연어 생성

- 자연어 생성은..
  - 자연어 생성, 텍스트 생성, 문장 생성 등으로 표현되고 있음
  - 뉴스 기사 생성, 대화 생성, 챗봇 등 다양한 영역에서 활용됨
  - 인간과 컴퓨터(AI) 사이의 의사소통을 위한 가장 자연스러운 방식

- 다 같은 자연어 생성 기술이지만 굳이 분류하자면...
  - 언어모델 기반, 패턴 기반의 자연어 생성 등의 형태로 나누어 생각할 수 있음

자연어 생성

- **언어 모델을 이용한 방식**
  - **OpenAI의 GPT3, Naver의 HyperClova 등 다양한 언어모델 존재**
    - 최근 이슈가 되는 초거대 AI는 사실은 사전 학습된 거대한 언어모델을 말함
  - **수많은 데이터의 학습을 통하여 어떤 표현이 가장 적절한가 예측하여 문장 생성**
    - GPT3: 발표 당시(2020.06) 약 1,750억개의 파라미터 데이터를 이용하여 학습
    - HyperCLOVA: 발표 당시(2021.05.25) 약 2,040억개의 파라미터 데이터를 이용하여 학습
  - **동작 예시**
    - 어젯밤에 라면을 먹고 잤더니 얼굴이 (?)

      → 어떤 단어가 가장 적절할 것인지 예측하여 문장 구성

자연어 생성

- 언어 모델을 이용한 방식이 가지는 문제점
  - 가장 자연스러운 문장을 만들어 내는 것에 강점을 가지지만
  - 어떤 데이터를 기반으로 정확한 사실을 전달하는 문장을 만들지 못함

    → 뉴스 기사 생성, 장면 설명 등을 위한 자연어를 생성하기에 부적합
  - 뉴스 기사를 언어모델을 이용하여 작성한다면

    → 사람이 쓴 것 같은 자연스러운 기사를 쓸 수 있지만 데이터를 기반으로 하지 않기 때문에 가짜뉴스가 만들어짐

자연어 생성

- 패턴 기반의 자연어 생성 방식
  - 오래 전부터 자연어 생성 분야에서 가장 많이 사용되어 온 방식
  - 데이터가 주어졌을때 그 데이터를 표현하기에 가장 좋은 문장을 저장소에서 꺼내어 사용하는 방식
  - 수많은 패턴을 준비하고 가장 적절한 패턴이 무엇인지를 학습하여 적용

- 어떤 목적으로 문장을 생성할 것인지에 따라서 적용 기술을 선택

기계 이해 (→텍스트 이해)

- 기계가 어떤 텍스트에 대한 정보를 학습하고, 사용자가 질의를 던졌을 때 그에 대해 응답하는 문제
- 기계가 텍스트를 이해하고 논리적인 추론을 할 수 있는지를 데이터 학습을 통해 구현, 확인하는 것
- 앞서 설명한 텍스트 분류, 유사도, 자연어 생성 등 자연어 처리에 대한 전반적인 내용이 모두 포함되어 있다고 볼 수 있음

기계 이해 (→텍스트 이해)

- 다양한 기계 이해 모델 중 메모리 네트워크 모델 활용 방식의 예시

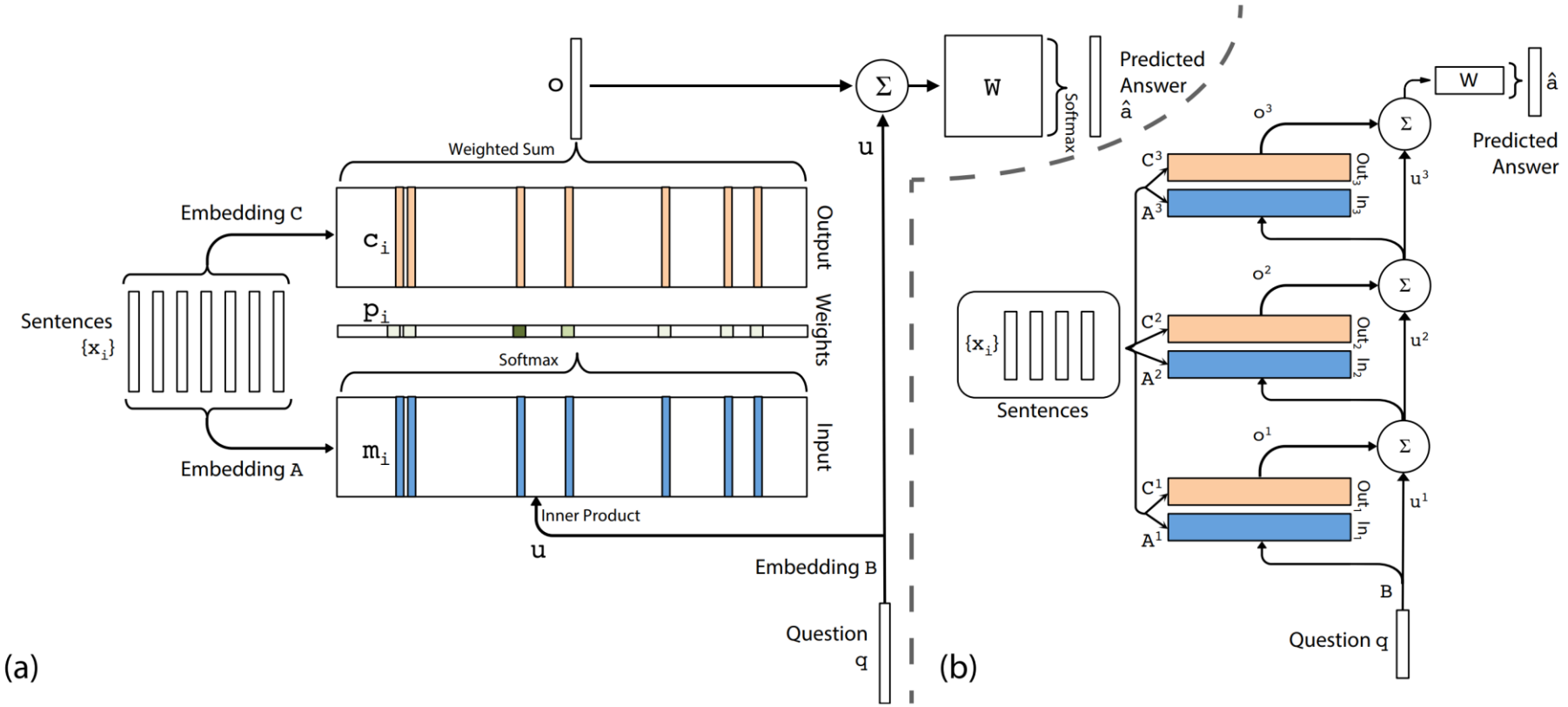


Figure 1: (a): A single layer version of our model. (b): A three layer version of our model. In practice, we can constrain several of the embedding matrices to be the same (see Section 2.2).

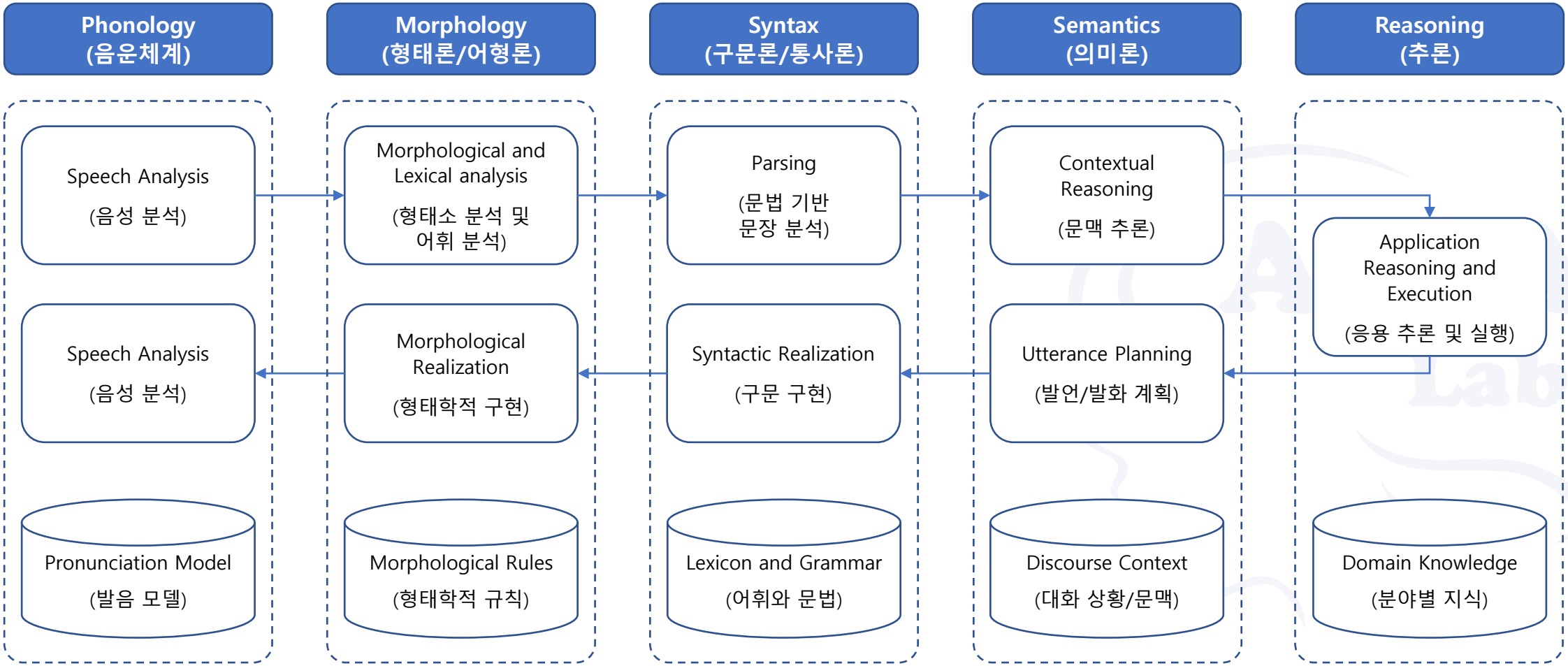
출처: Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston and Rob Fergus, "End-To-End Memory Networks," https://arxiv.org/pdf/1503.08895.pdf

자연어 처리 활용 분야

- **자연어 처리의 활용 분야**
  - **대량의 텍스트를 이해하고 수치화 하는 분야**
    - 감성분석(Sentiment Analysis), 문서분류, 문서요약 등
  - **사용자의 의도를 파악하고 대화하거나 도움을 주는 분야**
    - AI 스피커, Chatbot 등
  - **사용자의 입력을 받아 원하는 것을 검색하거나 답변을 제공하는 분야**
    - 기계번역, 질의응답 시스템 등

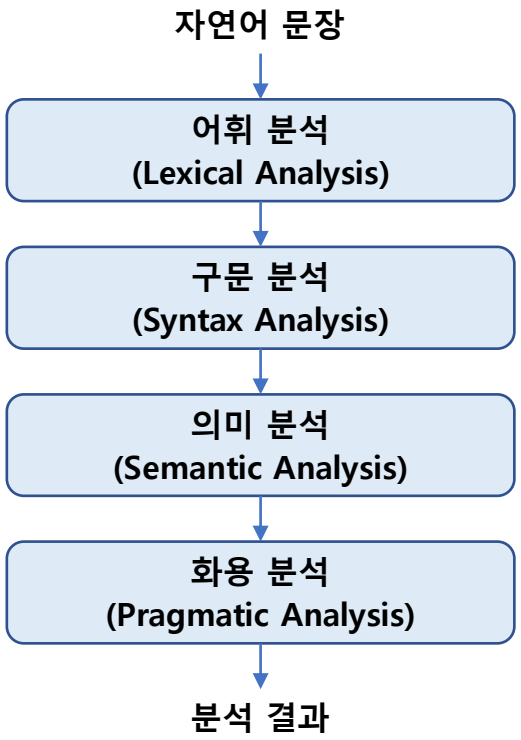
자연어 처리의 패러다임 변화

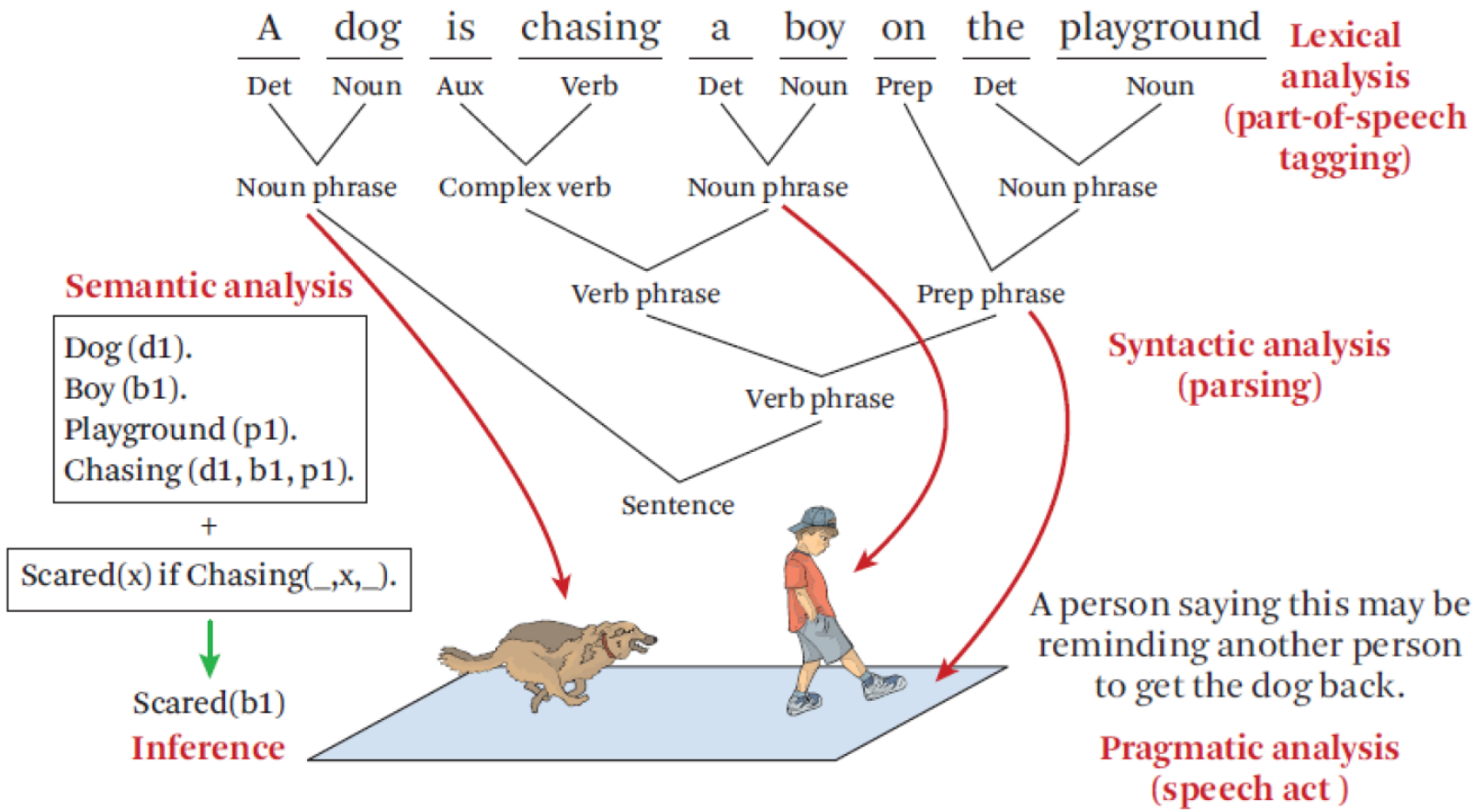
- 딥러닝 적용 이전의 자연어 처리 방식



자연어 처리의 패러다임 변화

- 자연어 처리(분석 중심)의 일반적인 단계



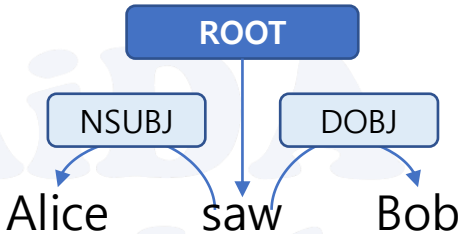


자연어 처리의 패러다임 변화

- **Lexical Analysis (어휘 분석)**
  - **기술 구성**
    - Sentence Splitting: 마침표, 느낌표, 물음표 등을 기준으로 분리
    - Tokenizing: 문서나 문장을 분석하기 좋도록 나눔(띄어쓰기 또는 형태소 단위로)
    - Morphological: 토큰들을 좀 더 일반적인 형태로 분석해 단어 수를 줄임으로써 분석의 효율성을 높임(가장 작은 의미 단위로 토큰화 함)
    - Stemming(어간 추출, 형태소 분석): cars, car → car
      = Lemmatization: 단어를 원형으로 표현
  - **필요성: 입력된 문장을 잘 분할해서 효율성을 높이기 위함**

자연어 처리의 패러다임 변화

- **Syntax Analysis (구문 분석)**
  - **내용**
    - 각각의 어절 단위로 구분, 해당 태그 부여 (Parsing Tree 이용)
    - 의존성 구문 분석 트리로 표현 (Dependency Parsing)
    - Saw(Root)를 기준으로 Alice(Subject), Bob(Object)의 관계를 나타냄
    - 이러한 문법적 관계는 서로 직접적인 연관이 있음
  - **필요성**
    - 언어는 문장의 구성에 규칙이 필요함
    - 문장 구성을 위한 규칙과 문법을 구성하기 위하여



Dependency Parsing

자연어 처리의 패러다임 변화

- **Semantic Analysis (의미 분석)**
  - **내용**
    - Syntactic + Meaning
    - 문장의 의미에 근거해서 그 문장을 해석하는 방법
    - 여러 의미 분석 방법과 다양한 유형의 문법 이용
    - 문장이 어떻게 구성되었는지 나타내는 규칙들로 구성된 일종의 형식 시스템

  - **필요성**
    - 규칙에 따라 문장은 만들었는데, 문장이 의미적으로도 올바른 것인지 확인해야 함
      - 예: 사람이 사과를 먹는다(O), 사람이 비행기를 먹는다(X)

자연어 처리의 패러다임 변화

- **Pragmatic Analysis (실용 분석)**
  - **실용주의**
    - 인간의 행동과 실제적인 측면에 대한 연구
    - 실제 상황에서의 언어, 표시, 단어 및 문장의 사용에 대한 연구
  - **실용적인 상호 작용 컨텍스트(문맥, 맥락)에서의 의미 연구를 가리킴**
    - 발언의 문자적 의미를 넘어 그 의미가 어떻게 구성되는지에 대한 것만 아니라 묵시적 의미에 초점을 맞춤 → 상호작용의 도구로 언어를 선택
    - 사람들이 언어를 사용할 때, "그 표현이 무엇을 의미하는지"와 같은 사람 사이에서 서로 통신하고 이해하는 방법

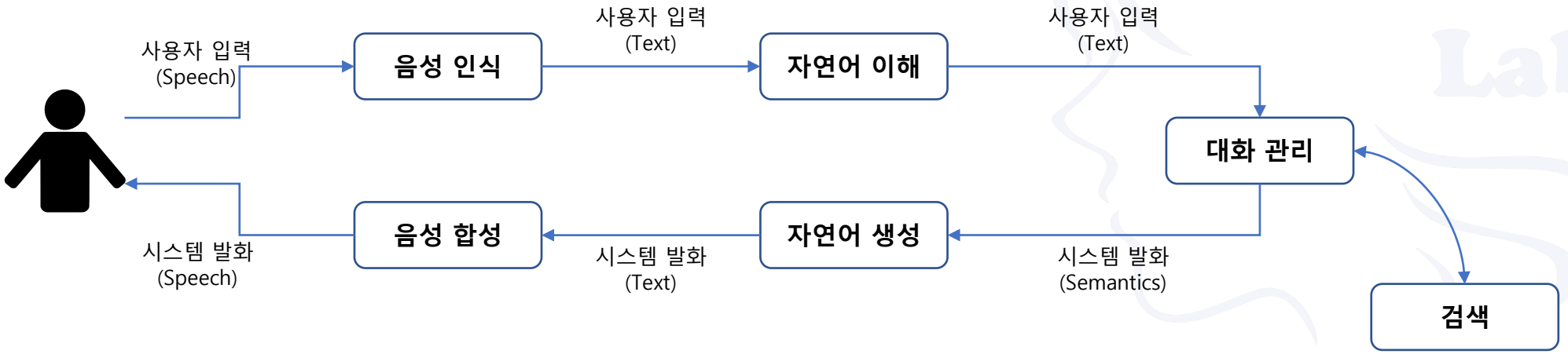
자연어 처리의 패러다임 변화

- **실용주의 관점에서 고려해야 할 사항**
  - 연사와 청취자 사이의 의미 전달, 협상
  - 발언에 대한 맥락
  - 발언이 가지는 의미 및 의미에 대한 잠재력

- **필요성**
  - 대화의 흐름을 파악하여 발화자의 의도에 맞도록 응답해야 함

자연어 처리의 패러다임 변화

- **내용**
  - 대화의 흐름 상 어떤 의미를 가지는지 탐색
  - 문맥의 구조 분석: 문장들 사이의 연관 관계 분석
  - 의도 분석: 전후 관계를 통한 실제 의도 분석
  - 대화 분석: 대표적인 담화 분석



자연어 처리의 패러다임 변화

- 딥러닝 적용 이전의 자연어 처리 방식의 문제점
  - 여러 단계의 모듈로 구성
    - 디자인이 복잡함
    - 상황에 따라 또 다른 서브 모듈 추가 빈번

    시스템이 매우 무겁고 복잡하여 구현 및 시스템의 구성이 어려움
  - 각각의 모듈이 완벽하게 동작하지 않음
    - 각기 발생한 오차가 중첩, 가중되어 뒤로 전파되는 **오차의 전파** 현상 발생
- 딥러닝의 적용화 함께 변화 시작
  - 각 하위 모듈의 딥러닝 모델화로 시작 → 점차 End-to-End 모델로 대체

End-to-End Deep Learning Model
종단간 기계학습. 입력부터 출력까지 파이프라인 네트워크 없이 한번에 처리하는 모델의 형태

자연어 처리의 패러다임 변화

- **전통적인 자연어 처리 VS 딥 러닝을 통한 자연어 처리**

전통적인 심볼릭 기반 접근 방법	딥러닝 기반 접근 방법
이산적(Discrete), 심볼릭 공간	연속적(Continuous), 신경망 공간
사람이 인지하기 쉬움	사람이 이해하기 어려움
디버깅 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

자연어 처리의 패러다임 변화

- **자연어 처리를 위한 심벌릭 데이터와 연속적인 데이터의 특징**

심볼릭(이산적, Discrete) 공간	신경망(연속적, Continuous) 공간
지식의 표현 방법 - 단어, 관계, 템플릿 - 고차원, 이산적, 희소 벡터 형태	**지식의 표현 방법** - 간소화 및 일반화 된 지식 그래프 - 저차원, 연속적, 꽉 찬 벡터 형태
추론(inference) - 거대한 지식 그래프로 인한 속도 감소 - 키워드 또는 템플릿 매칭에 민감함 (작은 차이에도 쉽게 결과가 바뀜)	**추론(Inference)** - 적은 메모리를 요구하며 빠르게 동작함 - 키워드 또는 템플릿 매칭에 강인함 (작은 차이에는 크게 영향을 받지 않음)
사람이 이해할 수 있으나 연산 효율성이 낮음	디버깅 어려움

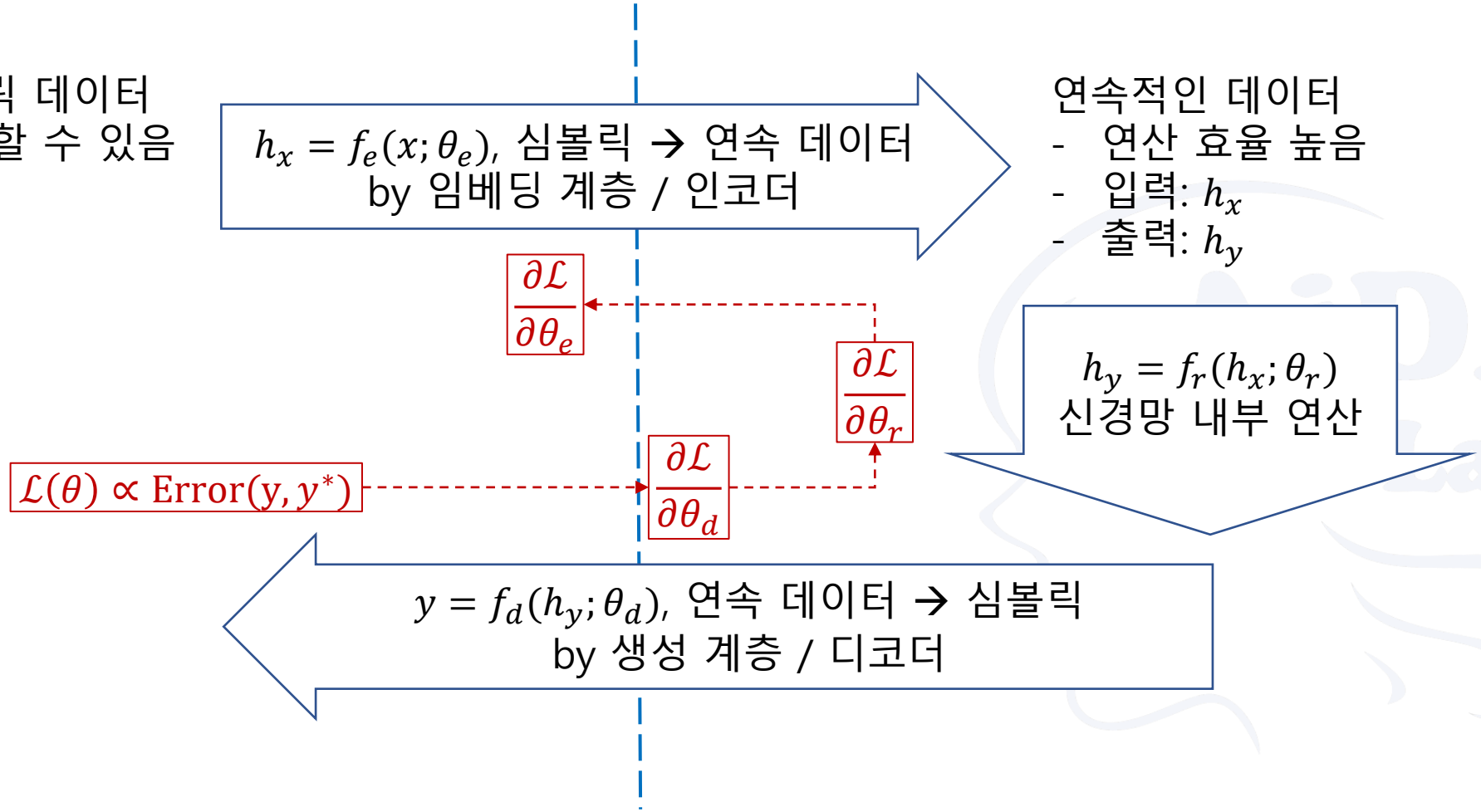
자연어 처리의 패러다임 변화

- Word2Vec 등의 임베딩 기술 적용
  - 단어(토큰)를 연속적인 벡터로 표현 가능 → 모호성과 유의성 문제 해결

- 딥러닝 적용
  - End-to-End 방식의 모델 적용 → 성능향상 유도
  - 개선된 RNN 계열 모델(LSTM, GRU) 활용 고도화
  - 주의모델(Attention) 적용 → 긴 길이의 시퀀셜 데이터를 이용한 학습 용이

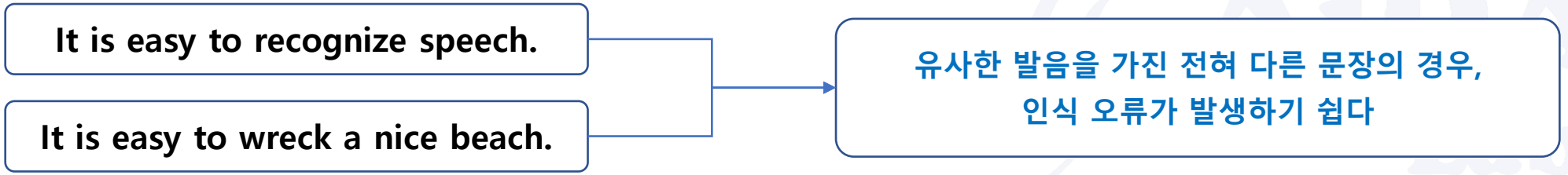
자연어 처리의 패러다임 변화

- **딥러닝 기반 자연어 처리 과정**



자연어 처리를 위한 언어학적 요소

- **Phonetics(음성학) & Phonology(음운론)**
  - 언어의 소리가 물리적으로 어떻게 형성되는지에 대한 이산적인 소리체계에 대한 연구 (예: disconnect → dis-k&-'nekt
  - 소리를 기준으로 분석하기 때문에 발생하기 쉬운 문제



  - 음성 인식
    - 음성의 파형을 기호화 하여 인식하는 음성학&음운론 기반의 연구 분야

음소: 더이상 작게 나눌 수 없는 음운론 상의 최소 단위

자연어 처리를 위한 언어학적 요소

- **형태론**
  - **어절**
    - 양쪽에 공백을 가지는 띄어쓰기 단위의 문자열
  - **단어**
    - 단일 품사를 가지는 단위
  - **형태소**
    - 의미를 가지는 언어 단위 중 가장 작은 단위
    - 의미 또는 문법적 기능의 최소 단위
    - 사전에 등록되어 있는 색인어의 집합

형태소의 예

나는 책을 읽었다. → 나, 는, 책, 을, 읽, 었다, .

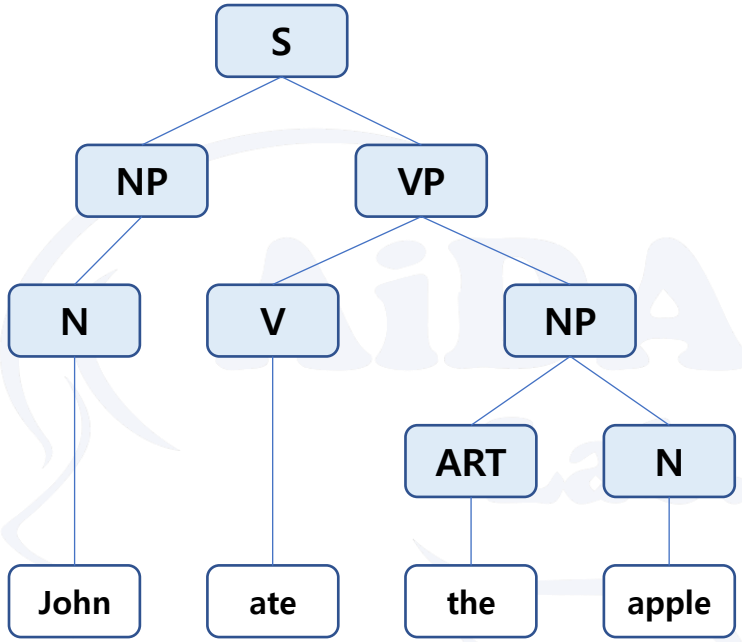
He tries to pass the exam.

자연어 처리를 위한 언어학적 요소

- **형태소 분석**
  - 형태소를 비롯하여 어근, 접두(미)사, 품사 등 다양한 언어적 속성의 구조를 파악하는 것
  - 입력된 문자열을 분석하여 형태소(Morpheme)라는 최소 의미 단위로 분리
  - 사전 정보와 형태소 결합 정보 이용
  - 정규 문법으로 분석 가능
  - 언어에 따라 난이도가 다름
    - 영어, 프랑스어 등: 쉬움
    - 한국어: 어려움

자연어 처리를 위한 언어학적 요소

- 구문론
  - 문법
    - 문장의 구조적 성질을 규칙으로 표현한 것
    - 각 규칙을 이용하여 문장을 생성, 분석할 수 있음
  - 구문 분석
    - 문법을 이용하여 문장의 구조를 찾아내는 과정
    - 문장의 구문 구조는 트리 형태로 표현 가능함
      - 몇 개의 형태소가 모여 구문 구조를 이루고
      - 구문 요소들 간의 결합구조를 트리형태로 구성



자연어 처리를 위한 언어학적 요소

- **의미 분석**
  - **개요**
    - 통사 분석 결과에 해석을 가하여 문장이 가진 의미를 분석함
    - 형태소가 가진 의미를 표현하는 지식 표현 기법이 요구됨
    - 통사적으로 옳으나 의미적으로 틀린 문장이 있을 수 있음
      - 돌이 걸어간다. (*cf.* 사람이 걸어간다.)
      - 바람이 달린다. (*cf.* 말이 달린다.)

그러나 '시적 표현' 등을 통하여 통용되는 경우도 많음

    - 모호성 등의 어려움이 있음
      - 말이 많다. (Horse? Speech? 해초?) (부피단위, 末(끝) 의 경우는 "많다"에 의해 대상 제외)

자연어 처리를 위한 언어학적 요소

- **실용 분석**
  - **문장과 실세계가 가지는 연관관계 분석**
  - **실세계의 지식과 상식의 표현이 요구됨**
  - **지시, 간접화법 등의 분석**
    - 지시
      - 대명사 등의 지시 대상
      - 상대방에게 행동을 요구하는 언어적 행위

자연어 처리가 어려운 이유

- **모호성**
  - **개념**
    - 단어의 중의성에 따라 발생하는 특징. 하나의 표현이 여러 의미를 가질 수 있는 성질
  - **발생원인**
    - 언어는 계속 진화하며, 특히 효율성을 극대화하는 방향으로 진화하기 때문
    - 최대한 짧은 문장 내에 많은 정보를 담고자 함(정보량이 낮은 내용은 생략)
    - 생략된 문맥(Context)을 인간은 여러 지식을 이용하여 효율적으로 채울 수 있지만, 기계는 이러한 작업에 매우 취약함

자연어 처리가 어려운 이유

- **모호성의 예시**
  - 단어의 중의성에 의한 예시

원문	차를 마시러 공원에 가던 차안에서 나는 그녀에게 차였다.
G사	I was **kicking** her in the **car** that wend to the park for **tea**.
M사	I was a **car** to her, in the **car** I had a **car** and went to the park.
N사	I got **dumped by** her on the way to the park for **tea**.
K사	I was in the **car** going to the **park** for **tea** and I was in her **car**.
S사	I got **dumped by** her in the **car** that was going to the park for a cup of **tea**.

업체별 한→영 기계번역 결과

  - "차"를 표현한 단어들: tea, car, kick, dumped
  - 일부는 표현을 빠뜨리거나 일부는 단어를 잘못 선택함 → 정확한 번역은 찾기 어려움

자연어 처리가 어려운 이유

- 문장 내 정보 부족에 따른 예시

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.
…	…

원문	선생님은 울면서 돌아오는 우리를 위로했다.
해석 1	선생님은 울면서 / 돌아오는 우리를 위로했다.
해석 2	선생님은 / 울면서 돌아오는 우리를 / 위로했다.
…	…

자연어 처리가 어려운 이유

- **다양한 표현**
  - **여러 상황을 표현, 묘사하기 위하여 다양한 표현이 사용되지만 그 의미는 동일한 경우가 매우 많음**

그런데 사람이 이해하는 단어의 의미가 미묘하게 다른 경우 또한 매우 많음

  - 골든 리트리버 한 마리가 잔디밭에서 공중의 공을 향해 달려가고 있습니다.
  - 공이 날아가는 방향으로 개가 뛰어가고 있습니다.
  - 개가 잔디밭에서 공을 쫓아가고 있습니다.
  - 잔디밭에서 강아지가 공을 향해 뛰어가고 있습니다.
  - 날아가는 공을 향해 멍멍이가 신나게 뛰어갑니다.
  - 밝은 갈색의 개가 공을 잡으러 뛰어가고 있습니다.

자연어 처리가 어려운 이유

- **불연속적 데이터**
  - **기존의 자연어 처리**
    - 이산 데이터를 대상으로 하였으므로 비교적 처리가 쉬운 편이었음
  - **딥러닝 기반의 자연어 처리**
    - 데이터를 딥러닝 모델에 적용하려면 연속적인 값으로 바꾸어 주어야 함
    - 벡터화, 단어 임베딩 등이 역할을 수행하고 있지만
    - 애초에 연속적인 값이 아니었기 때문에 딥러닝 모델의 구현에 다양한 제약이 존재함

자연어 처리가 어려운 이유

- **차원의 저주**
  - 불연속 데이터이므로 많은 종류의 데이터를 표현하려면 데이터 종류만큼의 대규모의 차원이 필요함. 어휘의 크기(규모)만큼의 차원이 요구됨.
  - 어휘 표현을 위한 방식을 희소 표현에서 분산 표현으로 변형하는 등 다양한 개선안 연구
  - 적절한 단어 임베딩 모델을 연구, 적용하여 차원 축소를 통해 해결 또는 개선
- **노이즈와 정규화**
  - 노이즈가 제대로 분리되지 않거나 영향력이 커지면 데이터의 원래 의미까지 손상 가능
  - 연속 데이터의 경우 일부가 약간 변경되어도 큰 영향이 없으나, 자연어 데이터는 불연속 데이터(심볼)이기 때문에 완전히 다른 의미가 되어버릴 수 있음
  - 띄어쓰기, 어순 차이 등으로 인한 정제(정규화) 이슈도 큰 어려움의 요소

한국어의 자연어 처리가 어려운 이유

- 교착어

종류	대표적인 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

  - 교착어의 수많은 파생 형태는 파싱, 형태소 분석, 언어 모델 등 전 분야에서 자연어처리를 어렵게 함
  - 접사가 붙어 다양한 단어가 생겨나므로 하나의 어근에서 비롯된 비슷한 의미의 단어가 매우 많이 생성됨

한국어의 자연어 처리가 어려운 이유

- 교착어의 형태변화 사례 (일부만 표기)

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
잡	+히		+었	+겠		+다	잡혔겠다
…							…
잡	+히	+시	+았(었)	+겠	+더라		잡히셨겠더라

한국어의 자연어 처리가 어려운 이유

- **접사에 따라 단어의 역할 결정 → 어순은 크게 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

번호	문장	정상여부
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

문장을 끊어서 사용한다면 문제 없는 경우도 발생함

한국어의 자연어 처리가 어려운 이유

- 띄어쓰기
  - 한국어의 띄어쓰기는 근대에 들어서 도입되었으므로 문장에서 띄어쓰기의 중요도가 낮음 → 영어의 경우, 띄어쓰기만으로 단어의 분리가 가능하지만 한국어는 불가능함

한국어의 자연어 처리가 어려운 이유

- 평서문과 의문문
  - 영어의 경우, 의문문과 평서문의 형태가 완전히 구별되지만 한국어는 의문문과 평서문이 같은 형태의 문장구조를 가지는 경우가 매우 많음
  - 물음표가 없으면 알 수 없는 경우가 많음

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

한국어의 자연어 처리가 어려운 이유

- **주어 생략**
  - **영어의 경우**
    - 기본적으로 명사가 매우 중요함
    - 매우 특별한 경우를 제외하고는 주어가 생략되는 경우가 없음
  - **한국어의 경우**
    - 동사를 중시하며 주어는 자주 생략됨
    - 문맥 정보를 이용하여 생략된 정보를 메꿈 → 컴퓨터가 취약한 부분

한국어의 자연어 처리가 어려운 이유

- 한자 기반의 단어
  - 한자를 조합하여 만들어진 단어가 많음
  - 한자 기반의 단어는 각 글자가 의미를 가지며, 그 의미들이 합쳐져서 하나의 단어의 뜻을 나타냄 (영어도 라틴어 기반의 단어는 조합형)

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

한국어의 자연어 처리가 어려운 이유

- **문제의 발생**
  - **한글이 한자를 대체하면서 문제가 발생함**
  - 한자는 표어문자, 한글은 표음문자
    - 표어문자: 하나 하나의 문자가 하나의 말, 단어, 형태소를 이루는 문자시스템
    - 표음문자: 문자에서 각 글자가 특정 의미를 가지지 않고, 단지 각 음성에 대응하여 발음을 나타내는 문자 시스템
  - 한자(표어문자) → 한글(표음문자) **➔ 정보의 손실 발생 ➔ 모호성 등의 문제 유발**
    - 인간은 정보의 손실로 발생한 모호성을 문맥의 이해를 통해 해석 가능. 기계는 어려움
  - 다른 언어보다 중의성에 따른 문제가 더욱 가중됨

문자의 발전 단계에서 가장 발전한 문자 체계는 표음문자임

한국어의 자연어 처리가 어려운 이유

- **서브워드 단위로 단어를 분절할 경우 가중되는 중의성 문제 사례**

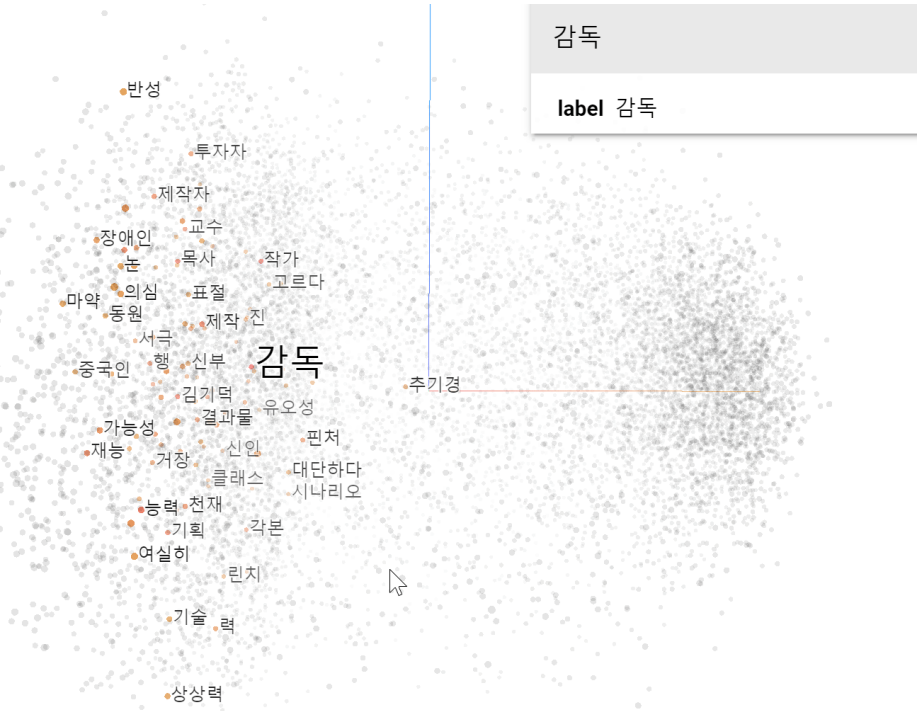
형태	텍스트
원문	저는 여기에 한 가지 문제점이 있다고 생각합니다.
형태소에 따른 분절	저/는/여기/에/한/가지/문제점/이/있/다고/생각/합니다/.
출현 빈도 기반의 서브워드 분절	_저 _는 _여기 _에 _한 _가지 _문 제 점 _이 _있 _다고 _생각 _합니다 _.

  - 문제점(問題點) → "문(問, 물을 문) 제(題, 제목 제) 점(點, 점 점)"으로 각각 분절됨
  - 제 → 결제(決濟)의 제(濟, 건널 제), 제공(提供)의 제(提, 끌 제) 등 다양한 "제"가 있음
  - 신경망 모델에서는 제는 각각의 의미를 가진 여러 개의 의미에 대하여 각각 임베딩 됨
    → 하나의 벡터가 아니기 때문에 여러 개의 벡터에 대한 평균 값으로 임베딩 됨
    → 애매한 처리 결과를 유발함

그런데 이런 경우는 사람도 헷갈림. 기계만의 문제가 아님

자연어 처리의 최근 추세

- **딥러닝의 자연어 처리 정복 과정**
  - **2010, RNN을 활용한 언어모델 시도**
    - 기존의 n-gram 기반 언어모델의 한계 극복 시도
  - **2013, 구글, word2vec 발표**
    - 단순한 구조의 신경망 사용, 단어를 효과적으로 잠재공간에 투사
    - 비슷한 의미의 단어일수록 저차원의 잠재공간에서 가깝게 위치함
    - 딥러닝 기반 자연어 처리 시, 네트워크 내부는 어떤 식으로 동작하는지에 대한 인사이트 획득



자연어 처리의 최근 추세

- **2014, Yoon Kim, CNN만을 활용한 텍스트 분류 모델 제시**
  - "문장은 단어의 시퀀셜 데이터 → RNN을 사용해야 한다"라는 고정관념을 깨뜨림
  - CNN 만들 활용해 기존의 텍스트 분류보다 성능을 끌어올린 텍스트 분류 모델 제시
  - 단어 임베딩 벡터와 결합하여 성능의 극대화를 이끌어냄

자연어 처리의 최근 추세

- 자연어 생성 시작
  - 2014, seq2seq 모델, Attention(주의) 기법 → 성공적인 기계번역 적용

    → 자연어 처리 분야에서 기존의 한정적인 적용 사례 탈피

    → 주어진 정보에 기반하여 자유롭게 문장을 생성하는 자연어 생성(NLG, Natural Language Generation)이 가능해짐

    → 기계번역 외에도 문장 요약, 챗봇 등 더 넓고 깊은 주제에 도입 시도 증가
  - 기계번역은 end-to-end을 활용하여 최초로 상용화에 성공

자연어 처리의 최근 추세

- **메모리를 활용한 심화 연구**
  - **Attention 기법의 성공 →연속적인 방식으로 정보를 읽고 쓰는 기법에 대한 관심 증가**
  - **메모리 증강 신경망(MANN) 확산**
    - 신경망을 통해 메모리를 활용하는 기법
      - 뉴럴 튜링 머신(NTM, Neural Turing Machine) 제시
      - 차별화 가능한 뉴럴 컴퓨터(DNC, Differentiable Neural Computer) 등장
    - 원하는 정보를 신경망을 통해 저장, 필요할 때 잘 조합해서 꺼내 사용하는 QA 등의 문제에 효율적으로 대응 가능 (Question Answering System)

MANN: Memory Augmented Neural Network

자연어 처리의 최근 추세

- **강화학습의 적용**
  - **딥러닝의 기계번역 분야 적용에 대한 문제점**
    - 딥러닝의 손실 함수와 기계번역에서의 목적 함수 사이의 괴리
    - 영상 처리 분야의 경우, 기존 MSE 손실 함수의 한계를 벗어나기 위하여 GAN 모델을 도입하여 생성모델 구현
    - 자연어 생성에도 GAN 모델에 강화학습 기법을 반영한 SeqGAN 등의 모델 제시
      → 강화학습의 폴리시 그래디언트(Policy Gradients) 기법을 자연어 생성에 적용 성공
  - **강화학습을 이용하여 실제 자연어 생성에서의 목적 함수로부터 보상을 받을 수 있게 됨 → 사람이 생성한 것과 유사한 문장 생성 능력의 극대화 유도**