NLP Overview

정윤경

성균관대학교

Outline

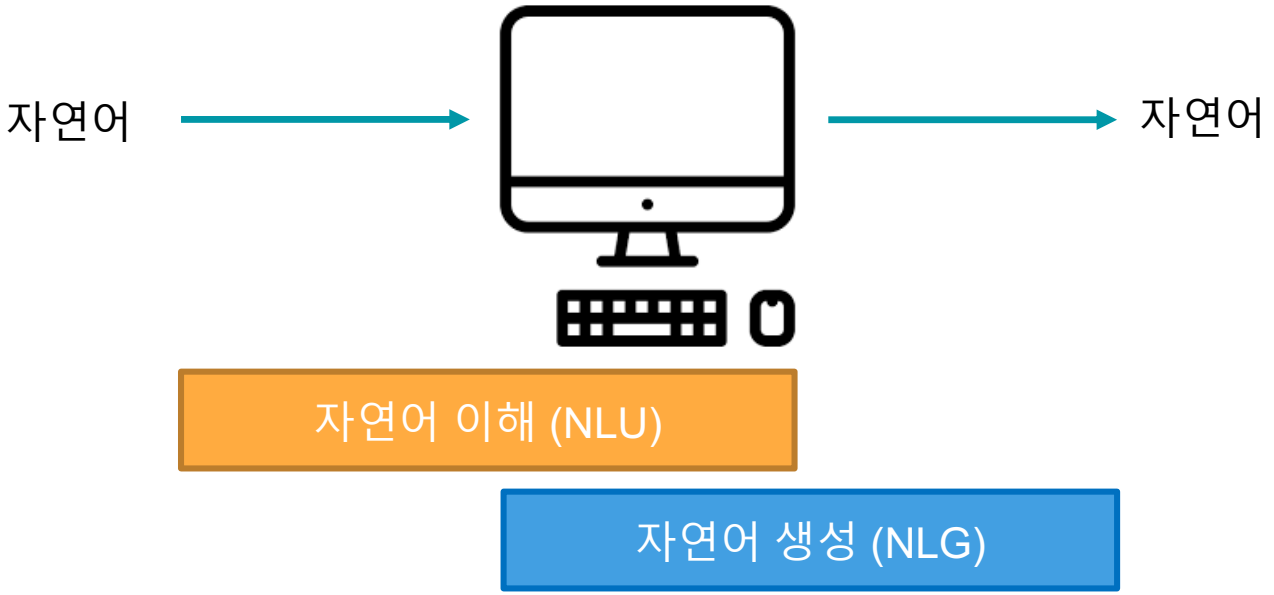
- 자연어처리 소개
- 수업 운영 방식
- 자연어처리 기술의 트렌드

Natural Language is

- 컴퓨터에서 사용하는 프로그램 작성 언어 또는 기계어와 구분하기 위해 인간이 일상생활에서 의사 소통을 위해 사용하는 언어 (출처: IT 용어사전)

Natural Language Processing (NLP)

- 컴퓨터를 이용해 사람의 자연어를 분석하고 처리하는 기술 (출처: IT 용어사전)
  - e.g., 단어의 수 세기, 맞춤법 검사, 작문 스타일 구분, 대화의 의도 파악, 기사 작성, 소설 창작



자연어처리 (출처: 위키피디아)

- 자연어 처리는 인간의 언어 현상을 컴퓨터와 같은 기계를 이용해서 묘사할 수 있도록 연구하고 이를 구현하는 인공지능의 주요 분야 중 하나다.
- 자연 언어 처리는 연구 대상이 언어 이기 때문에 당연하게도 언어 자체를 연구하는 언어학과 언어 현상의 내적 기재를 탐구하는 언어 인지 과학과 연관이 깊다.
- 구현을 위해 수학적 통계적 도구를 많이 활용하며 특히 기계학습 도구를 많이 사용하는 대표적인 분야이다.
- 정보검색, QA 시스템, 문서 자동 분류, 신문기사 클러스터링, 대화형 Agent 등 다양한 응용이 이루어지고 있다.

NLP Applications

- Document analysis
- Text classification
- Sentiment analysis, Opinion mining
- Text Summarization
- Machine Translation
- Text Generation
- Chatbot, Conversational AI
- Intelligent Assistant
- Question Answering

ELIZA (MIT AI Lab, 1966)

```
  Welcome to
                 EEEEEE  LL      IIII   ZZZZZZ   AAAAA
                 EE      LL       II       ZZ   AA   AA
                 EEEEE   LL       II     ZZZ    AAAAAAA
                 EE      LL       II    ZZ      AA   AA
                 EEEEEE  LLLLLL  IIII ZZZZZZ    AA   AA

  Eliza is a mock Rogerian psychotherapist.
  The original program was described by Joseph Weizenbaum in 1966.
  This implementation by Norbert Landsteiner 2005.


ELIZA: Is something troubling you ?
YOU:   Men are all alike.
ELIZA: What is the connection, do you suppose ?
YOU:   They're always bugging us about something or other.
ELIZA: Can you think of a specific example ?
YOU:   Well, my boyfriend made me come here.
ELIZA: Is it important to you that your boyfriend made you come here ?
YOU:   He says I'm depressed much of the time.
ELIZA: I am sorry to hear that you are depressed.
YOU:   It's true. I am unhappy.
ELIZA: Can you explain what made you unhappy ?
YOU:
```

Siri (October, 12, 2011)

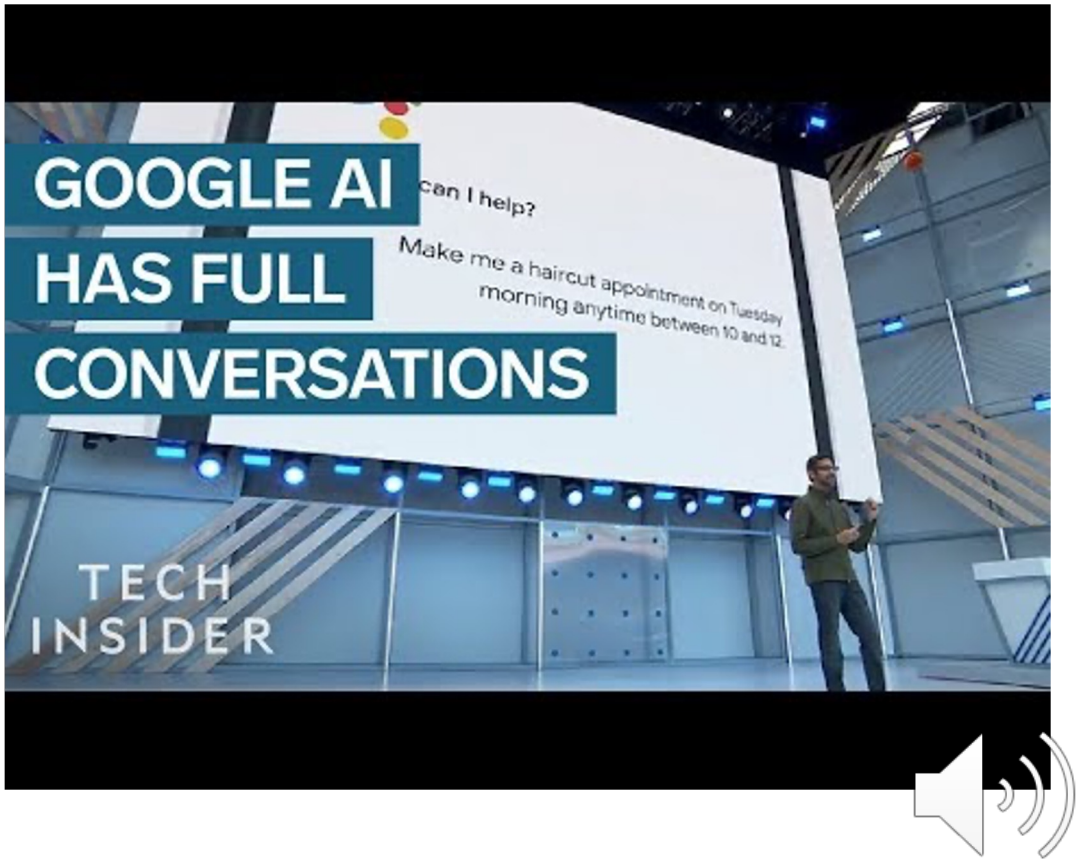
- Tasks that Siri can perform
- https://www.apple.com/kr/siri/

IBM Watson at Jeopardy (2011)

Q&A system

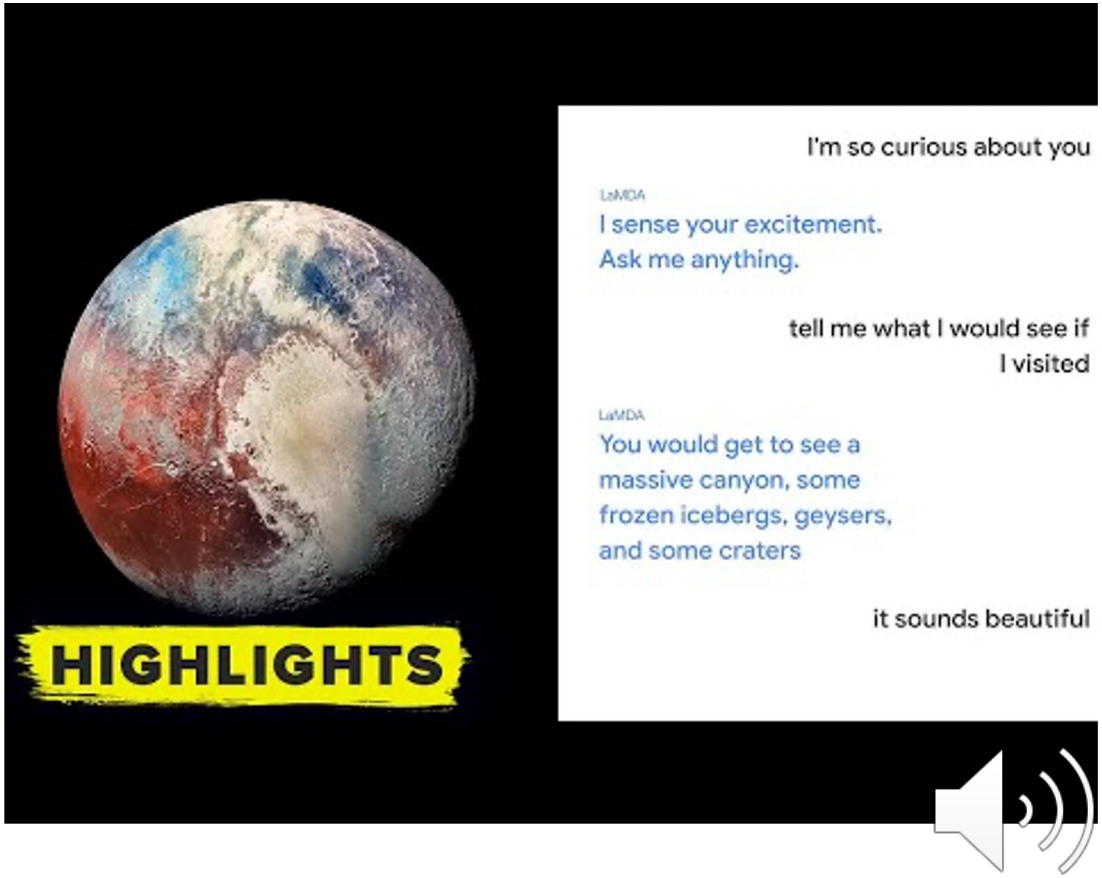
Google Duplex (2018)

- Task-oriented
  - making an appointment for the user
- Understands the user's preference
- Can handle exceptional situations



Google LaMDA (2021)

- Open domain conversation
- Able to learn concept
- Built on Transformer trained on dialogue
- Will be used for Google Assistant, search, and workspace



DALL-E (openAI, 2021)

- Creates images from text captions for a wide range of concepts expressible in natural language
- Trained a 12-billion parameter version of GPT-3 to generate images from text descriptions

TEXT PROMPT

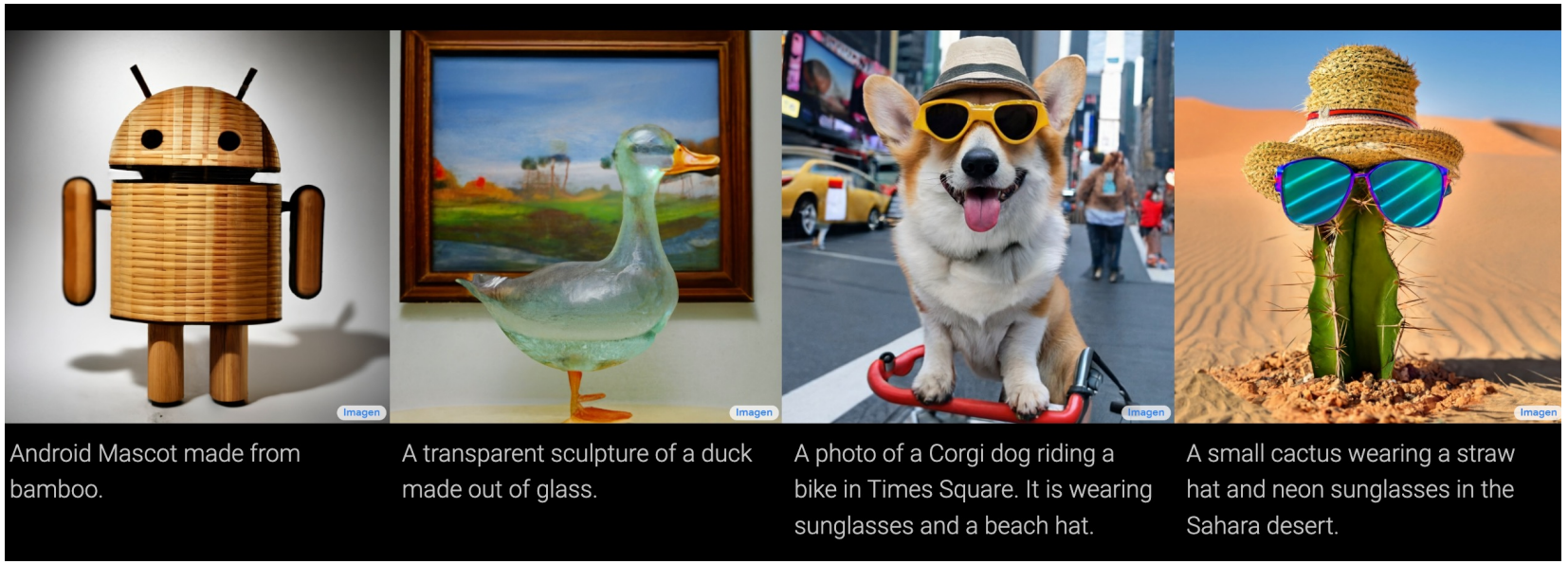
an armchair in the shape of an avocado. . . .

AI-GENERATED IMAGES

Edit prompt or view more images↓

IMAGEN (Google, 2022)

- a text-to-image diffusion model with an unprecedented degree of photorealism and a deep level of language understanding
- builds on the power of large transformer language models (e.g. T5) in understanding text and hinges on the strength of diffusion models in high-fidelity image generation



Goals

- 자연어처리기술의 주요 개념을 이해하고 설명할 수 있다
- 텍스트를 수집하고 통계 분석 및 전처리를 수행하는 파이썬 프로그램을 작성할 수 있다
- 텍스트를 분류, 클러스터링, 생성하는 프로그램을 설계하고 작성할 수 있다
- 텍스트를 처리하는 프로그램의 기능을 평가할 수 있다
- 기계학습과 딥러닝 기술이 자연어 문제를 처리하기 위해 어떻게 활용하는지 이해하고 사용할 수 있다

Schedule

- 텍스트 데이터셋 (corpus)
- 텍스트 분석 및 통계 기반 분류
- 텍스트 정규화, 표준화
- 문서 벡터화
- 단어 품사 태깅, Chunking & Named Entity Recognition
- 기계학습, 딥러닝 기반 분류
- 문장 구조 분석
- 토픽 모델링과 요약
- 문서 유사도 분석
- 뉴럴 언어 모델
- 임베딩

Prerequisites

- 파이썬 프로그래밍
- 기계학습
- 딥러닝

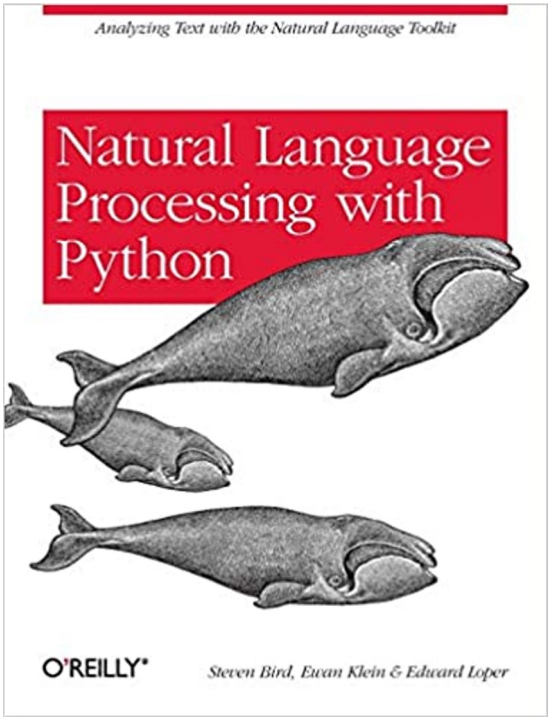
Development Tools

- Google Colaboratory
- NLTK, KoNLPy
- Scikitlearn, Gensim
- Keras, tensorflow
- Numpy, pandas, matplotlib, Beautiful Soup

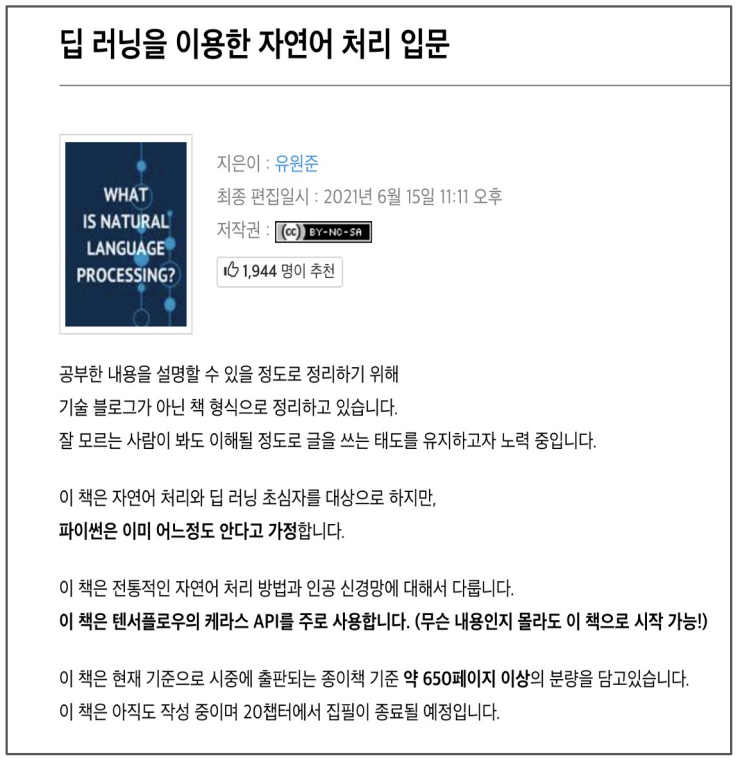
Grading

- 과제 및 평소 학습: 50%
  - 업로드한 파일 반드시 확인
- 기말 과제: 35%
- 발표: 15%
- **결석이 4주차 이상인 경우 학칙상 F**

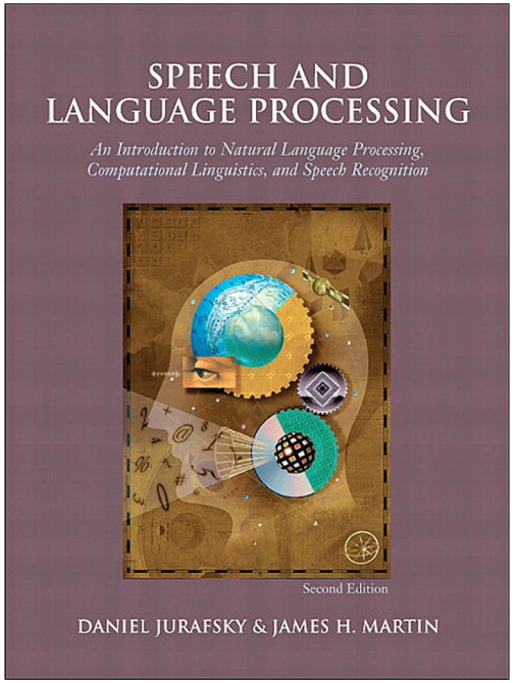
Reference



http://www.nltk.org/book/

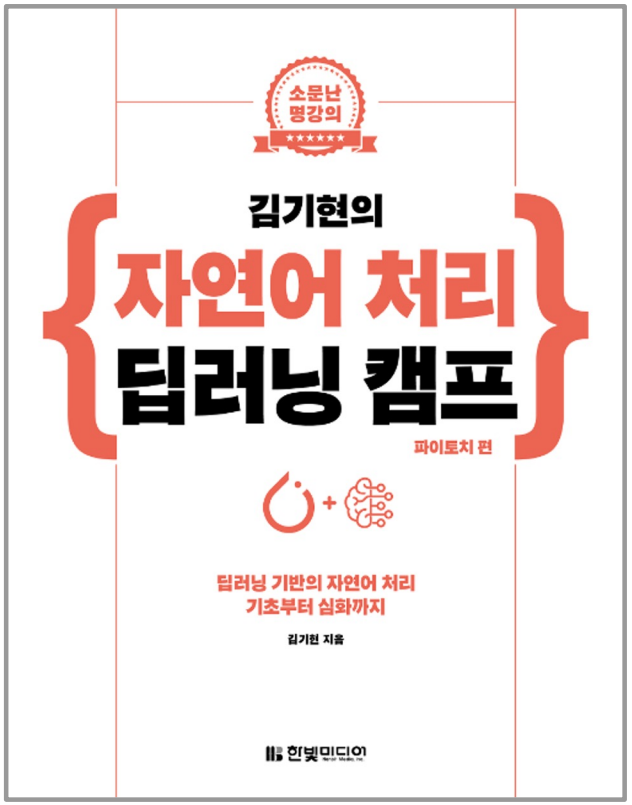


https://wikidocs.net/book/2155

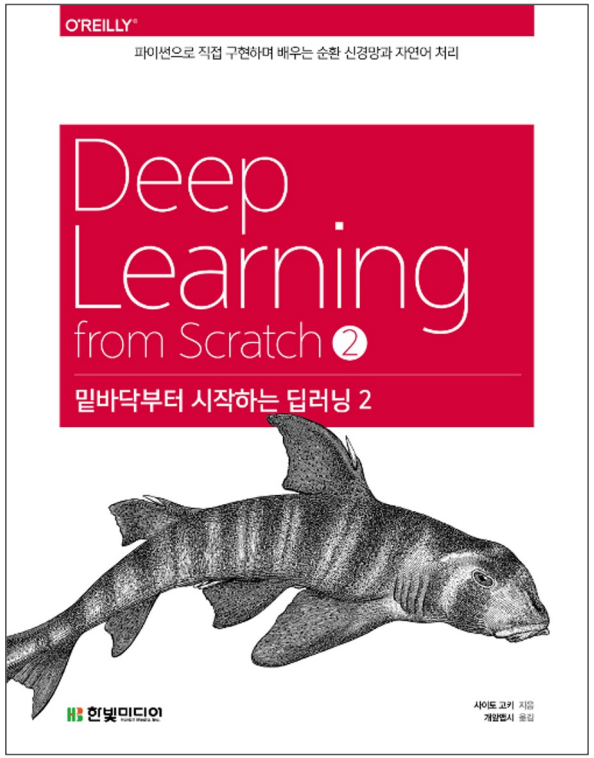


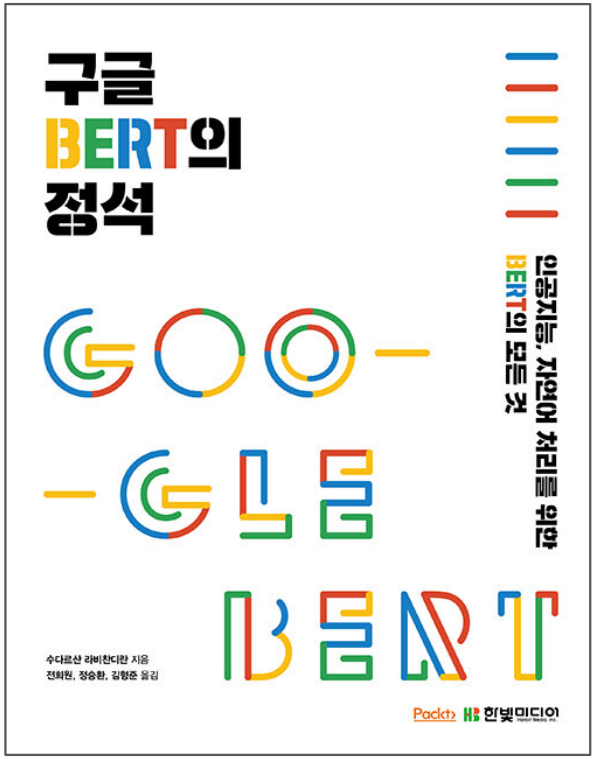
https://web.stanford.edu/~jurafsky/slp3/

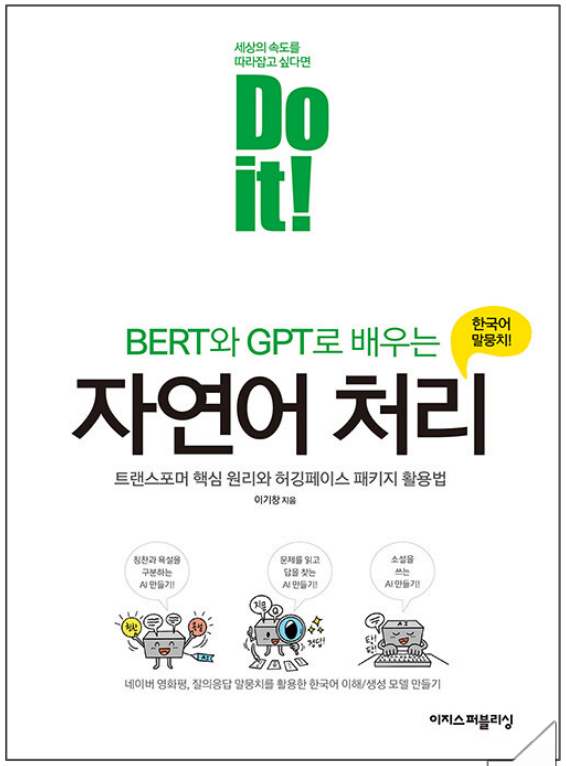
Recommendations



Recommendations







Natural Language Processing

NLP is AI-complete

- The most difficult problems in AI
- Language is **ambiguous**
- Language is dynamic
- Requires world knowledge and logical reasoning

Why is NLP so hard..?

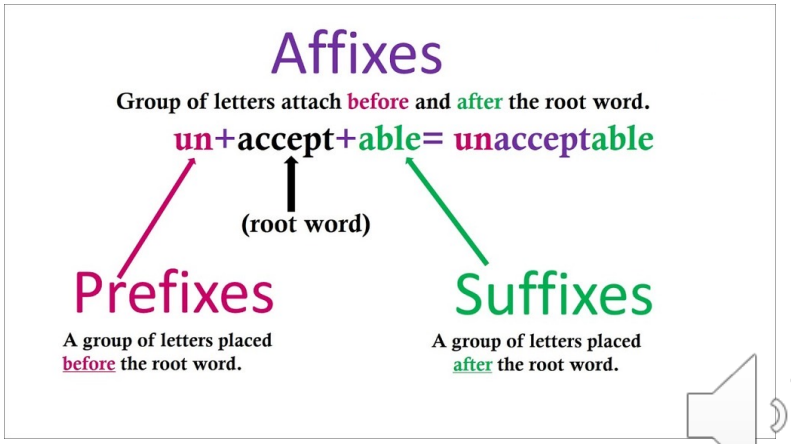
because of ***ambiguity***

“Get the cat with the gloves.”

Why is NLP so hard..?

Morphology

- The study of how words are composed of morphemes (the smallest meaning-bearing units of a language)
- Two broad classes of morphemes
  - **Stems**: “main” morpheme of the word, supplying meaning
  - Affixes: Bits and pieces that combine with stems to modify their meanings and grammatical functions (prefixes, suffixes, infixes)

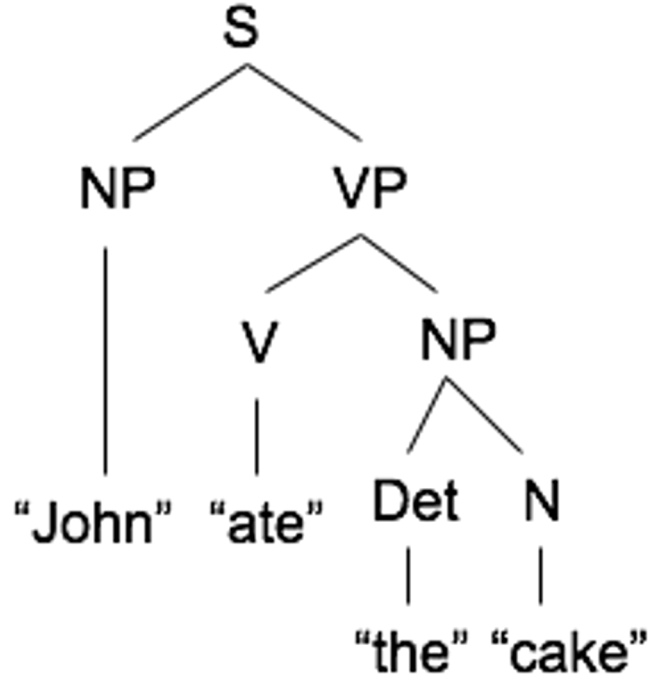


Syntactic Analysis

"John ate the cake"

Grammar

R0: S → NP VP
R1: NP → Det N
R2: VP → VG NP
R3: VG → V
R4: NP → "John"
R5: V → "ate"
R6: Det → "the"
R7: N → "cake"



Semantic Analysis

- Derive the meaning of a sentence.
- Often applied on the result of syntactic analysis.

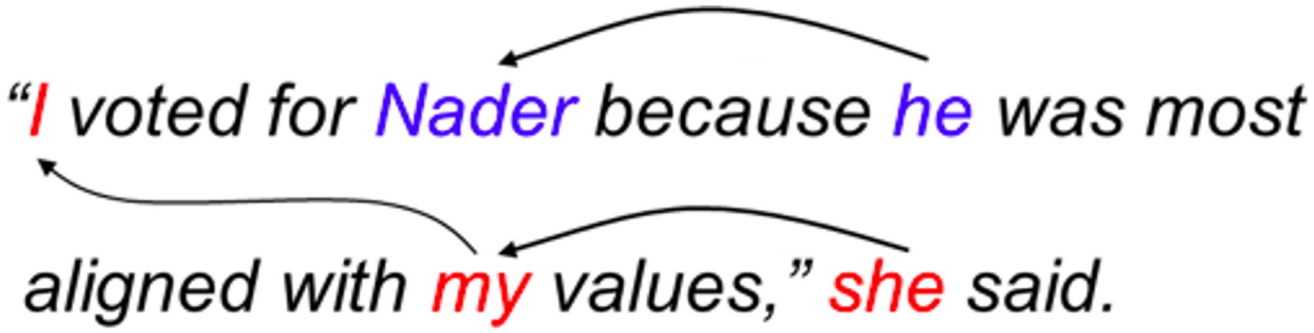
  “John ate the cake.”
  NP V NP
  ((action INGEST) ; syntactic verb
  (actor JOHN-01) ; syntactic subj
  (object FOOD)) ; syntactic obj

- To do semantic analysis, we need a (semantic) dictionary (e.g. WordNet, http://www.cogsci.princeton.edu/~wn/).

Discourse Analysis

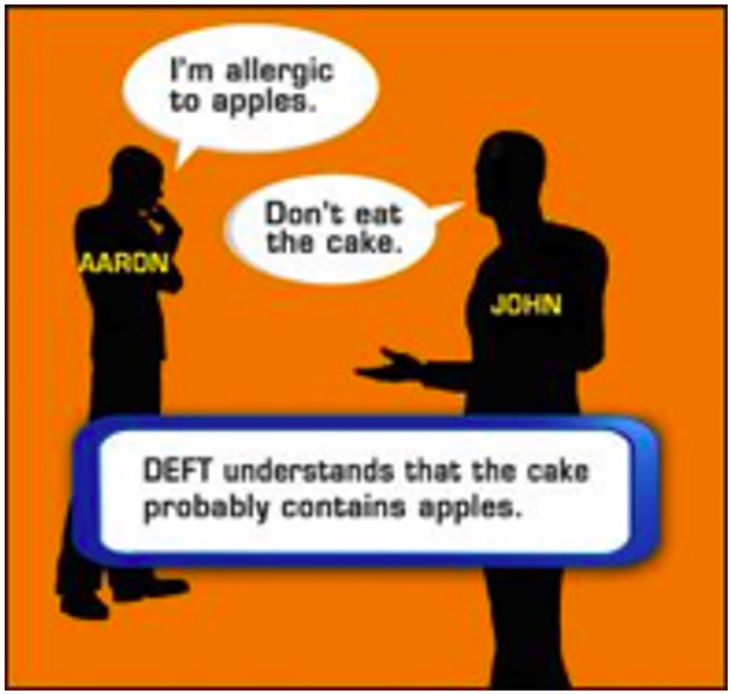
- Analyze relations between sentences, including:

  Anaphora (e.g. “he”, “she”, “it”, “they”) resolution

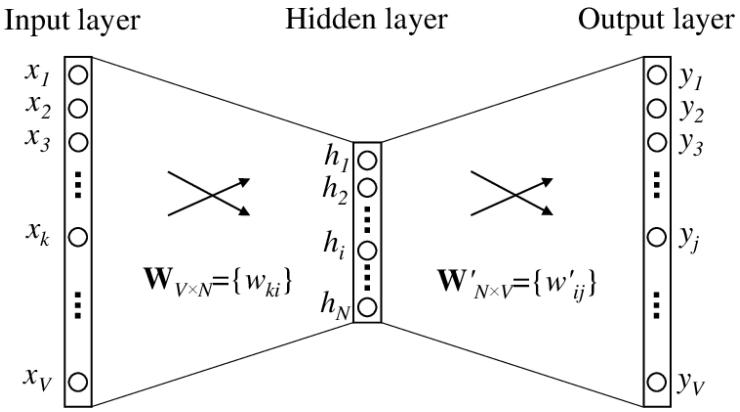


Pragmatics Analysis

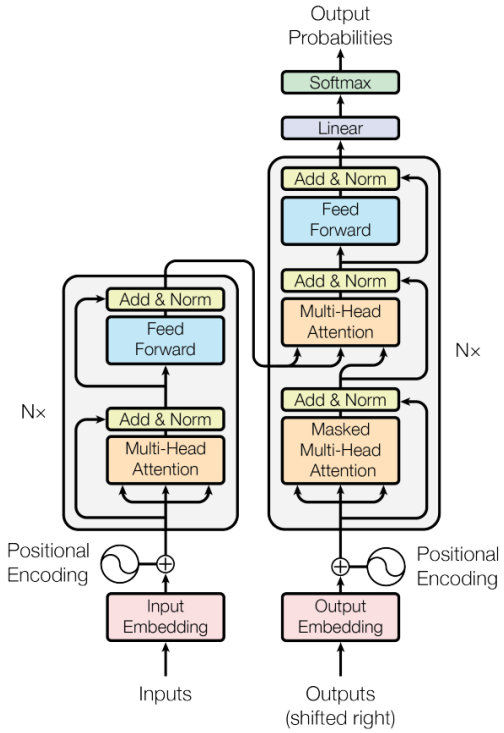
- Language as social interaction between agents rather than descriptive texts.
- Analyses include:
  - World knowledge
  - Speech act: an utterance that has performative function
    Speaker X: "We should leave for the show or else we'll be late." (request, suggestion)
    Speaker Y: "I am not ready yet." (statement, rejection)



딥러닝 & Transformers



Distributed Representations of Words and Phrases and their Compositionality (NIPS 2013)



Attention is all you need (NIPS 2017)

Summary

- 정의
- 응용 분야
- NLP의 주요 사건들
- 수업 운영
- 자연어처리를 위한 연구 분야