INTRO TO NLP AND DEEP LEARNING

장예훈

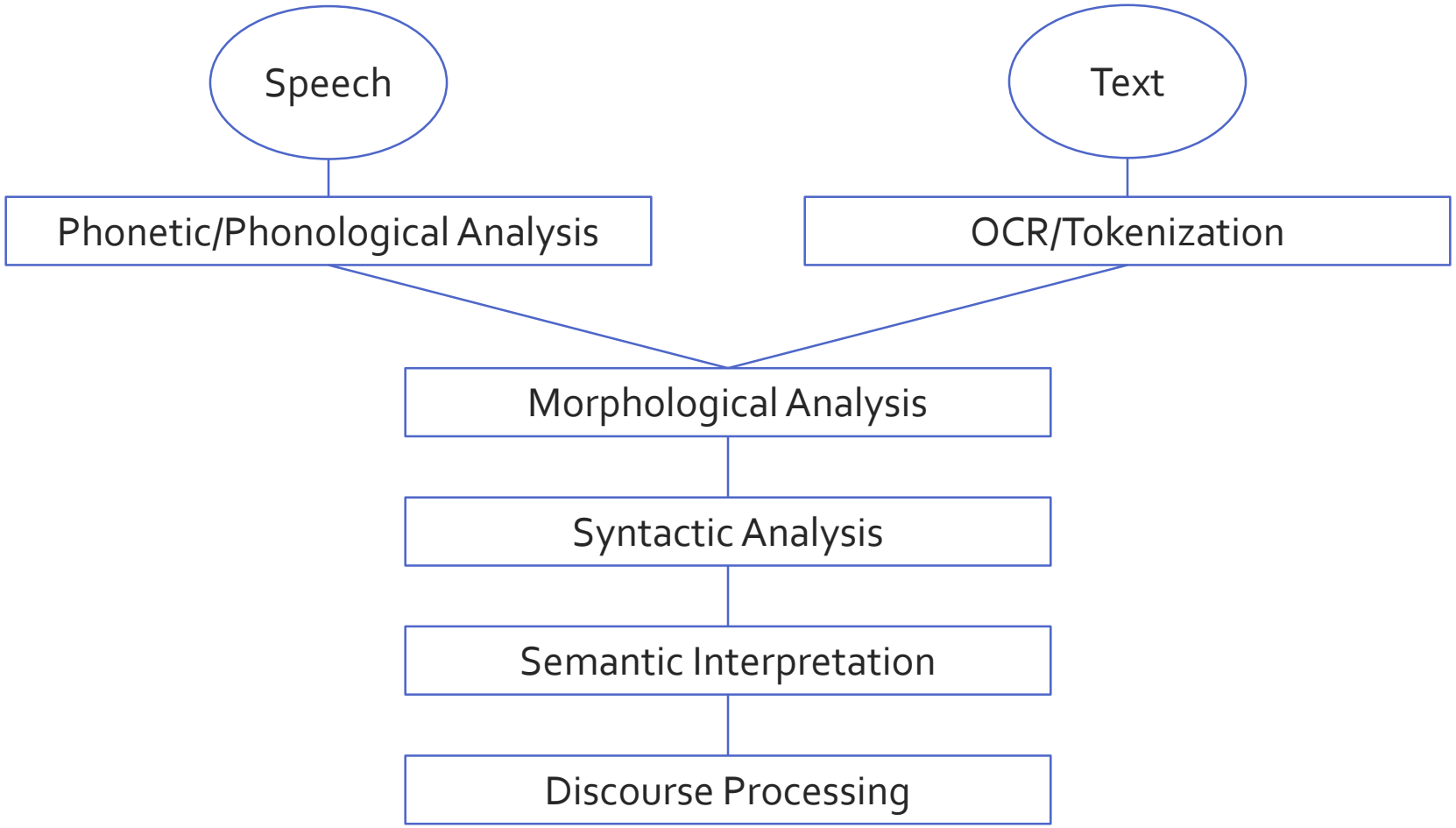
CS224D Lec 1

Natural Language Processing

- **Natural Language Processing** 이란 무엇인가?
  - 자연어 처리는 **컴퓨터 과학, 인공지능, 언어학**의 교차점에 있는 분야
- 목적
  - 컴퓨터가 자연어를 **"이해"**하고 처리하는 것

    e.g. Question Answering
- AI-Complete
  - 언어를 **완벽히** 이해하고 표현하는 것 (심지어는 정의하는 것)

NLP Level


Speech
Text
Phonetic/Phonological Analysis
OCR/Tokenization
Morphological Analysis
Syntactic Analysis
Semantic Interpretation
Discourse Processing

NLP Applications - range(Simple, Complex)

- Simple
  - 스펠링 체크
  - 키워드 검색 : Naver에서 한 두 글자만 쳐도 자동완성 되는 기능
  - 동의어 찾기
- Intermediate
  - 정보 추출 e.g. 웹사이트에서 상품 가격, 날짜, 장소, 사람이나 회사의 이름 등을 추출
  - Classifying
  - 긍정/부정 의미 찾기
- Complex
  - 기계번역 :구글번역, 파파고
  - ChatBot : 대화 시스템, 질의 응답

NLP in Industry

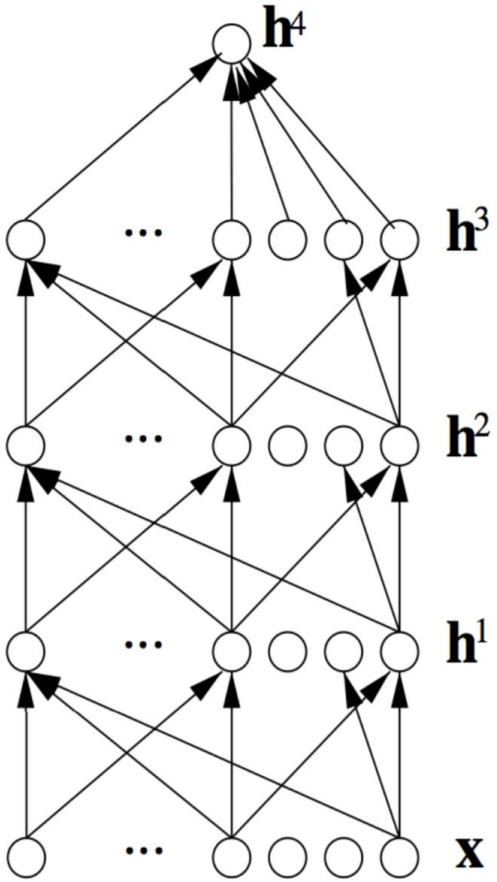
- 검색
- 온라인 광고
- 자동 번역
- 마케팅 또는 금융/거래에서의 감정 분석
- 음성 인식
- 고객 지원 자동화

Why is NLP hard?

- **Complexity**
  - 상황에 따라 의미가 달라짐
  - 다양한 해석의 여지 존재함
  - 예외 사항이 많음
  - Jane hit June and then **she** [fell/ran].
    - if she == Jane → [~~fell~~/**ran**] / if she == June → [**fell**/~~ran~~]
- **Ambiguity**
  - 모호성
  - I made her duck
    - 요리
    - 마법

What's Deep Learning (DL) ?

- Deep Learning은 Machine Learning의 Subfield
- **Representation Learning**
  - 좋은 feature와 representation을 자동으로 학습
- **Deep Learning Algorithms**
  - multiple level의 representation과 output을 학습
- "raw" 데이터를 input으로 받음



Reasons for Exploring Deep Learning

- Manually designed features
  - 지나치게 세분화 → over fitting 가능성
  - 불완전
  - 설계와 검증 시간 오래 걸림
- Learned features
  - 채택이 쉬움
  - 학습이 빠름

Reasons for Exploring Deep Learning

- Deep Learning은…
  - 유연하고 광범위하며 학습이 가능한 framework를 제공
  - 지도학습(supervised)과 비지도학습(unsupervised)이 가능
  - 방대한 양의 데이터
  - 빠른 속도의 H/W와 멀티코어의 CPU/GPU
  - 새로운 모델, 알고리즘, 아이디어

→ **IMPROVED PERFORMANCE!!**

Deep Learning + NLP = Deep NLP

- **NLP의 idea와 목표 + Representation Learning + Deep Learning methods**

- NLP를 통한 Improvements
  - 학습 단계: speech, morphology, syntax, semantics
  - 응용 분야: 기계 번역, 의미 분석, 질의 응답

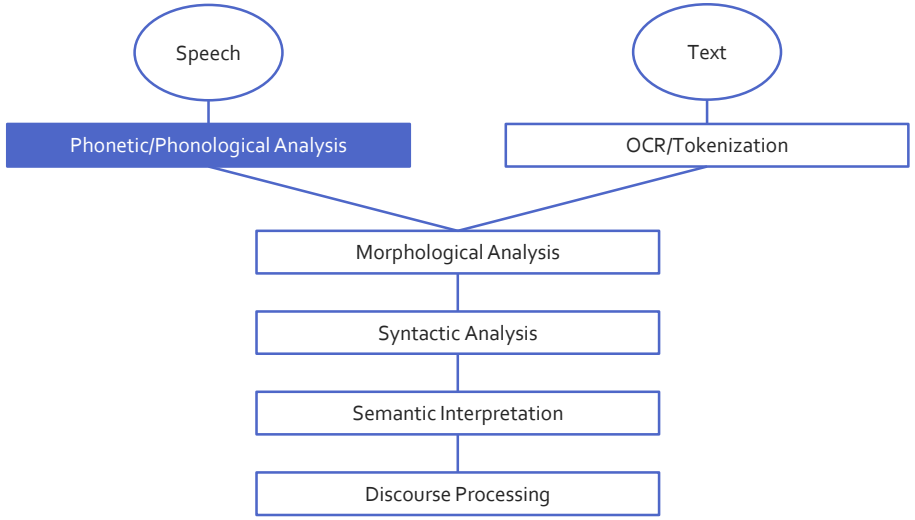
Representations at NLP Levels: Phonology

- 음성을 더 잘 이해하기 위한 목적

- **Traditional**
  - 음운을 하나하나 표현하려고 노력
  - 구강구조, 혀의 위치 등의 개인차

    → 음운 별로 미세한 발음차이, 소리의 고저, 장단, 억양 발생
  - 구분 시도 but, 제작시간 많이 소요

- **DL**
  - **음성 features로부터 음운을 예측 → 벡터로 표현**



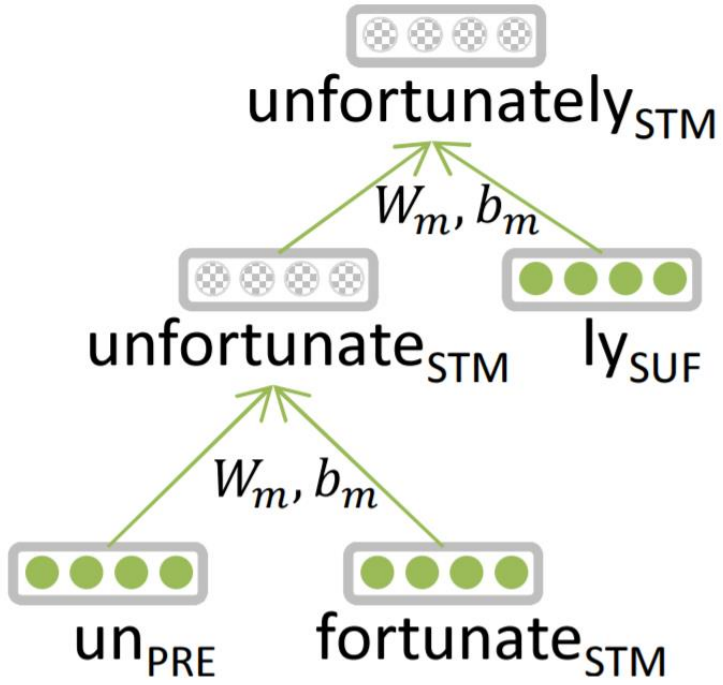
Representations at NLP Levels: Morphology

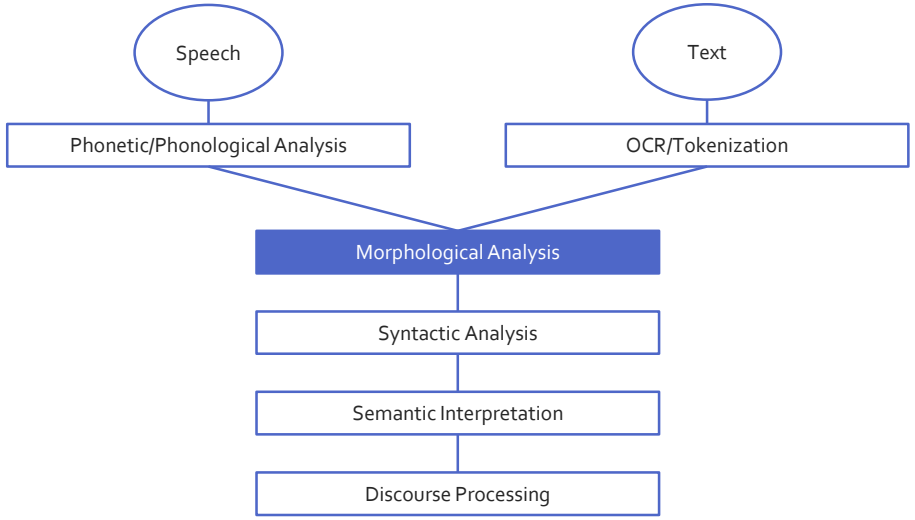
- 형태소를 분류하는 체계
- **Traditional**
  - 접두사, 접미사, 어간 등 세분화

prefix	stem	suffix
un	interest	ed

- **DL**
  - 모든 형태소는 vectors
  - Neural network를 통해

    두 벡터가 하나로 합쳐짐

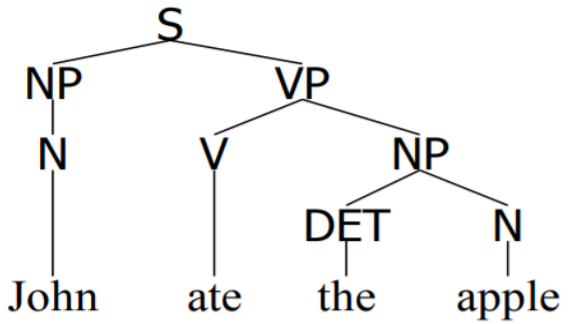


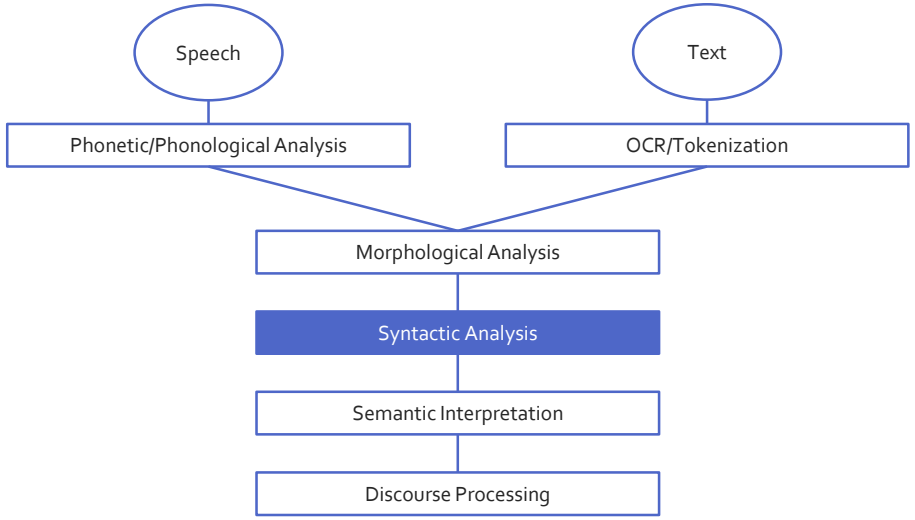


Representations at NLP Levels: Syntax

- 문법과 유사한 개념

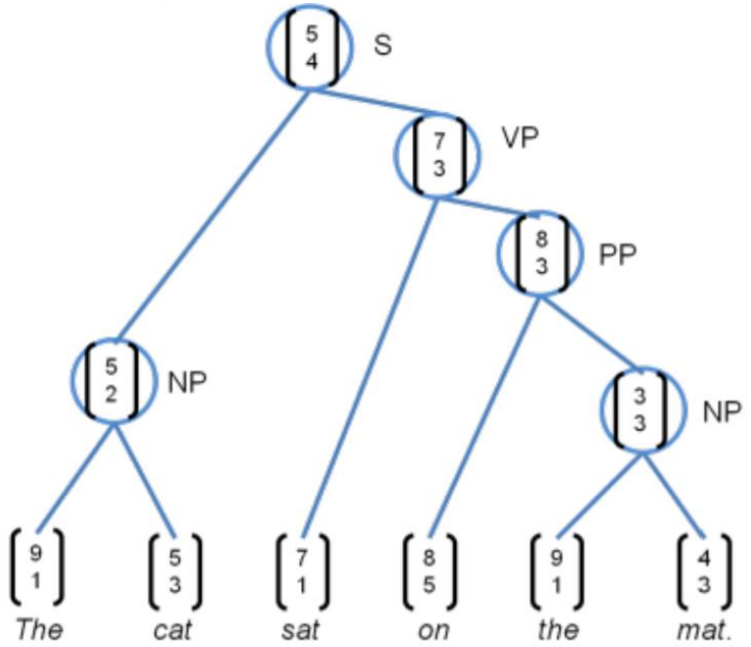
- **Traditional**
  - NP, VP와 같은 구문 별 카테고리화





- **DL**
  - 모든 단어와 문장은 vector
  - Neural network를 통해

    두 벡터가 하나로 합쳐짐



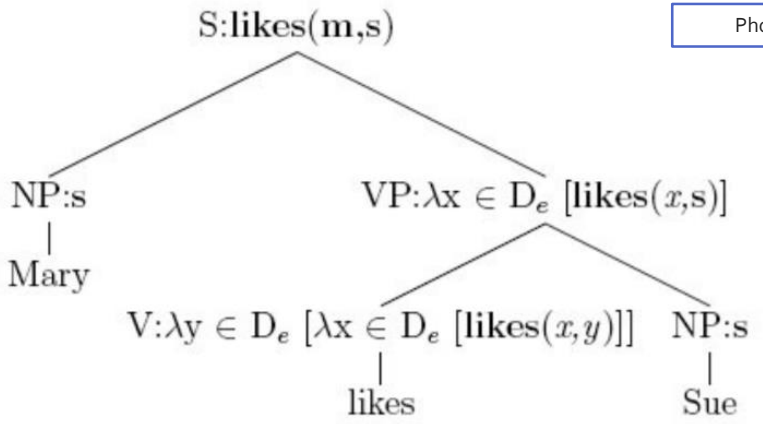
Representations at NLP Levels: Semantics

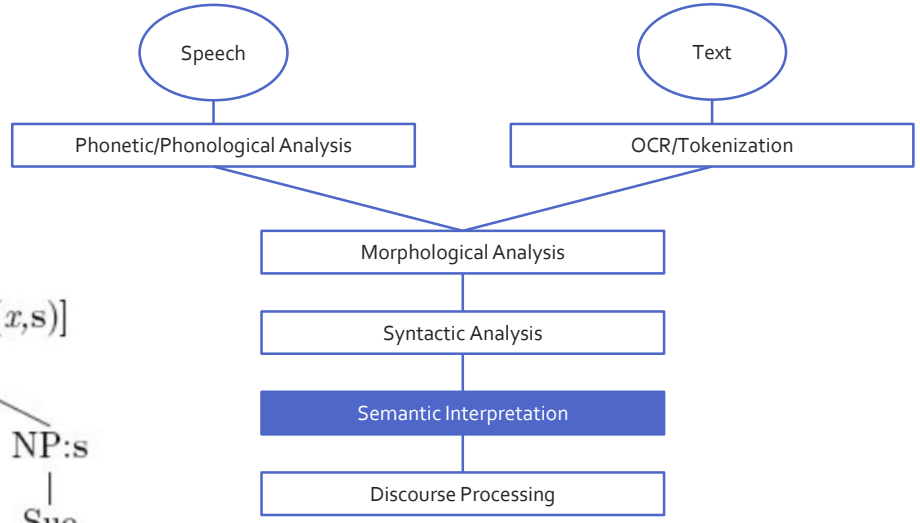
- Sentence의 의미를 해석
- **Traditional**
  - Sentence = 단어 + 구

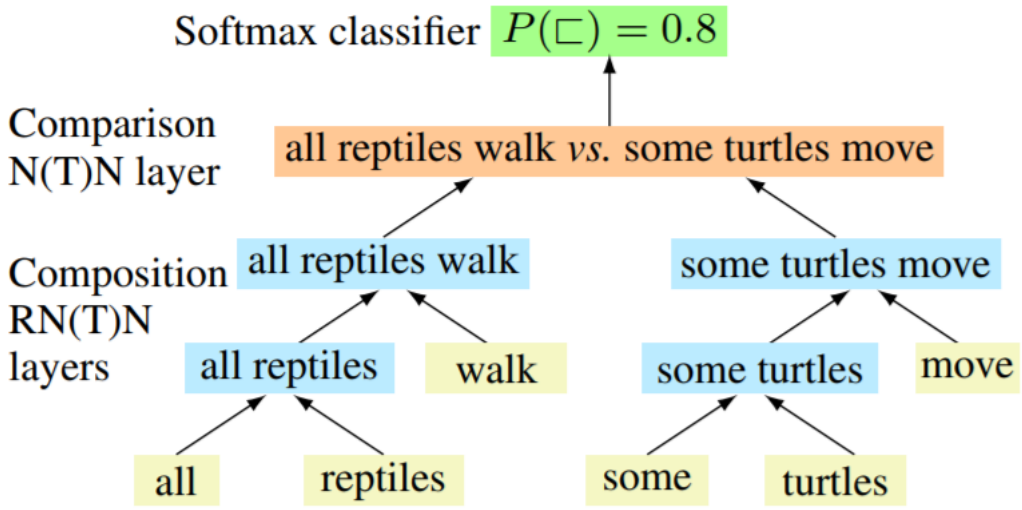
    → 특정한 λ(람다)값 필요
  - 언어의 유사성, 모호성 개념 X
- **DL**
  - 모든 단어와 문장과

    논리적 표현은 vector
  - Neural network를 통해

    두 벡터가 하나로 합쳐짐



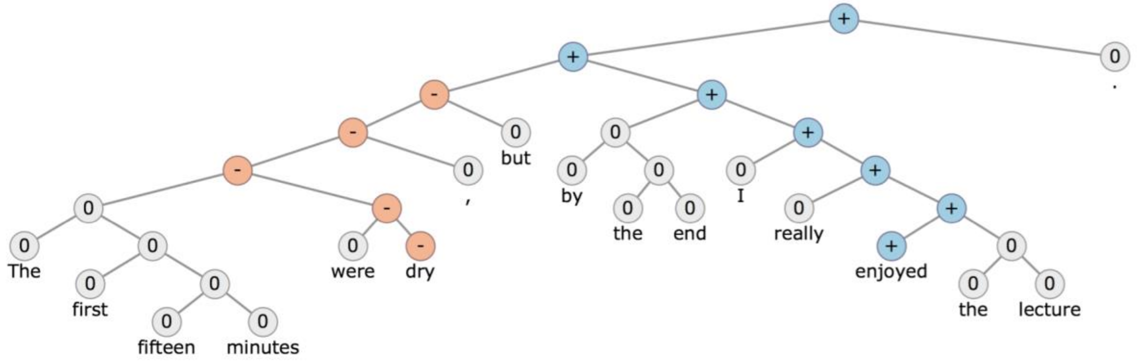




NLP Applications: Sentiment Analysis

- 감정 분석

- **Traditional**
  - Sentiment Dictionary를 만들고
    bag-of-word(단어 순서 무시) 또는
    hand-designed negations features(모든 단어 캡쳐 X)와 결합



- **DL**
  - 형태학적, 문법적, 논리적 의미 등에서 사용한
    동일한 Deep Learning Model 사용 가능 → **RNN**

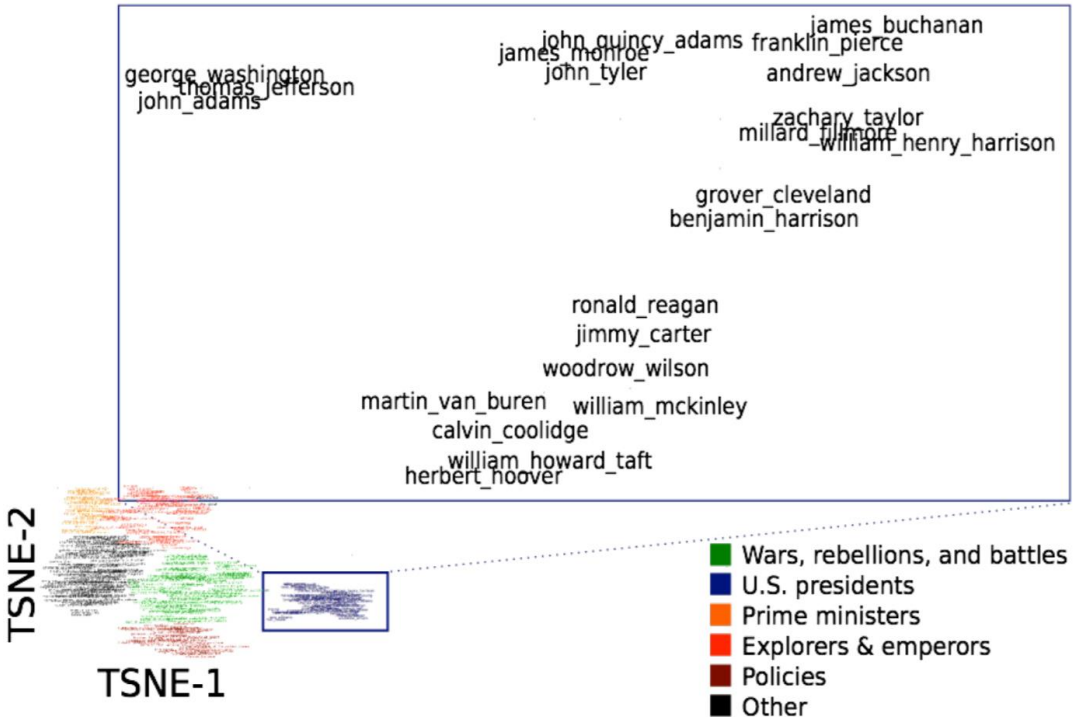
NLP Applications: Question Answering

- 질의 응답

- **Common**
  - 수 많은 Feature Engineering: 질문을 받는 분야와 그 분야의 지식을 확보

    e.g. 정규 표현식

- **DL**
  - 형태학적, 문법적, 논리적 의미, 감정 분석 등에서 사용한 동일한 Deep Learning Model 사용 가능
  - facts 들은 vector에 저장됨

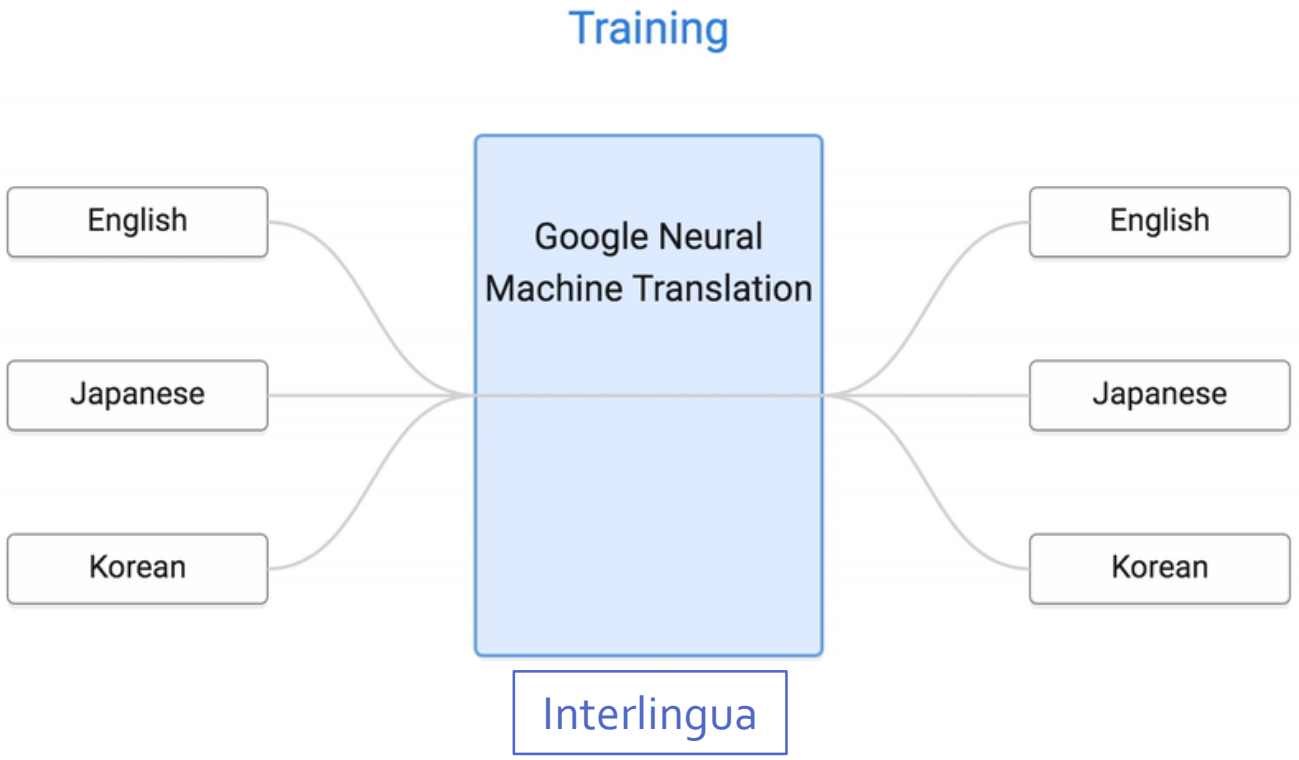


NLP Applications: Machine Translation

- 기계번역

- **Traditional**
  - 방법론
    - One Sentence → Other Language
    - 문법적 구조 이해
    - 문법적 의미 먼저 이해 → 다른 언어로 번역
  - 메모리 차지를 많이 함

- **DL**
  - **중간언어(Interlingua)** 사용



NLP Applications: Machine Translation

- **Sequence2Sequence**
  - Input Sentence가 Vector로 Mapping → Output Sentence가 생성