Lecture 4

Topic	Dependency Parsing
날짜	@2025년 3월 21일

Stanford CS224N: Lecture 4

🔗 Link

https://youtu.be/KVKvde-_MYc?si=BRmnk0m7KJ282djM

1 The linguistic structure of sentences

[1] Constituency structure (구성 구조)

- 문장의 구성 요소를 통해 구조 파악

[2] Dependency structure (의존 구조)

- 단어 간의 의존 관계를 통해 구조 파악
- input → 문장, output → 트리 형태의 문장 구조

(1) Dependency parsing (의존 구문 분석)

- *표기 방법*
  - 단어 간의 연결: head (중심 단어)와 dependent (의존 단어)
  - 수식 받는 단어 (head)를 화살표의 시작 부분으로, 수식하는 단어 (dependent)를 화살표의 가리키는 부분으로 표기

- *why?*

  1 Phrase Attachment Ambiguity

  - Ex) Scientists observe whales from space
    - 수식 관계에 따라 과학자들은 우주에서 고래를 관측했다, 과학자들은 우주에서 온 고래를 관측했다 2가지 의미로 분석 가능
    - 구들이 어떤 단어를 수식하는지에 따라 문장의 의미가 달라지는 모호성이 존재

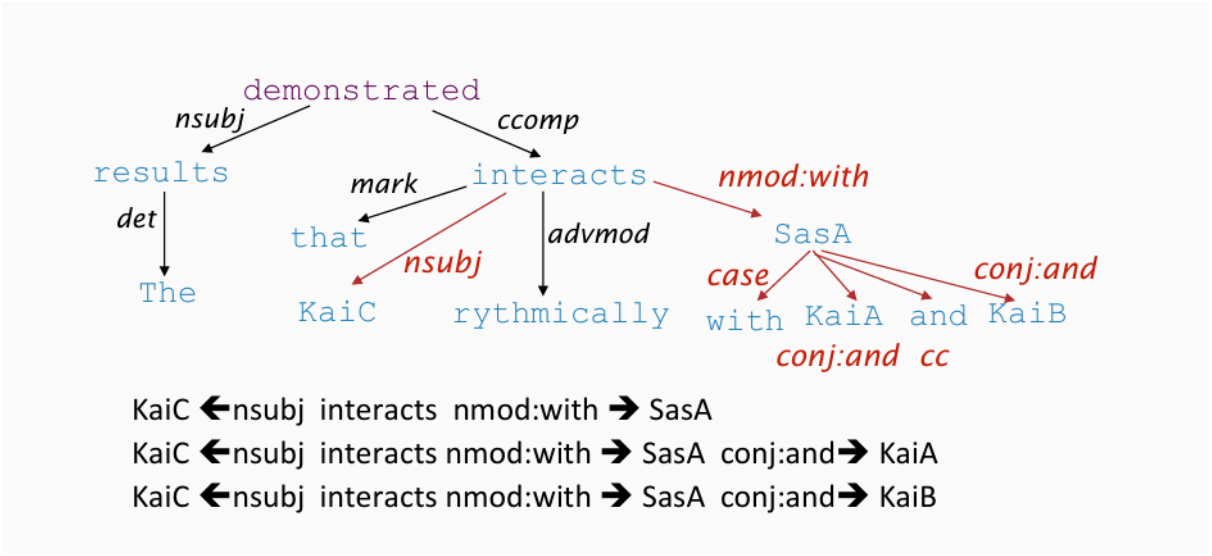
  - PP attatchment ambiguities multiply
    - 전치사구의 수식 관계에 따라 문장의 의미가 달라질 수 있다
    - Ex) The board approved [its acquisition] [by Royal Trustco Ltd.] [of Toronto] [for $27 a share] [at its monthly meeting].
      - 전치사구가 n개 존재할 때, 전치사구로 구성할 수 있는 이진 트리 구조의 개수:

        Catalan numbers= $\frac{(2n)!}{[(n+1)!n!]}$

  2 Coordination Attachment Ambiguity

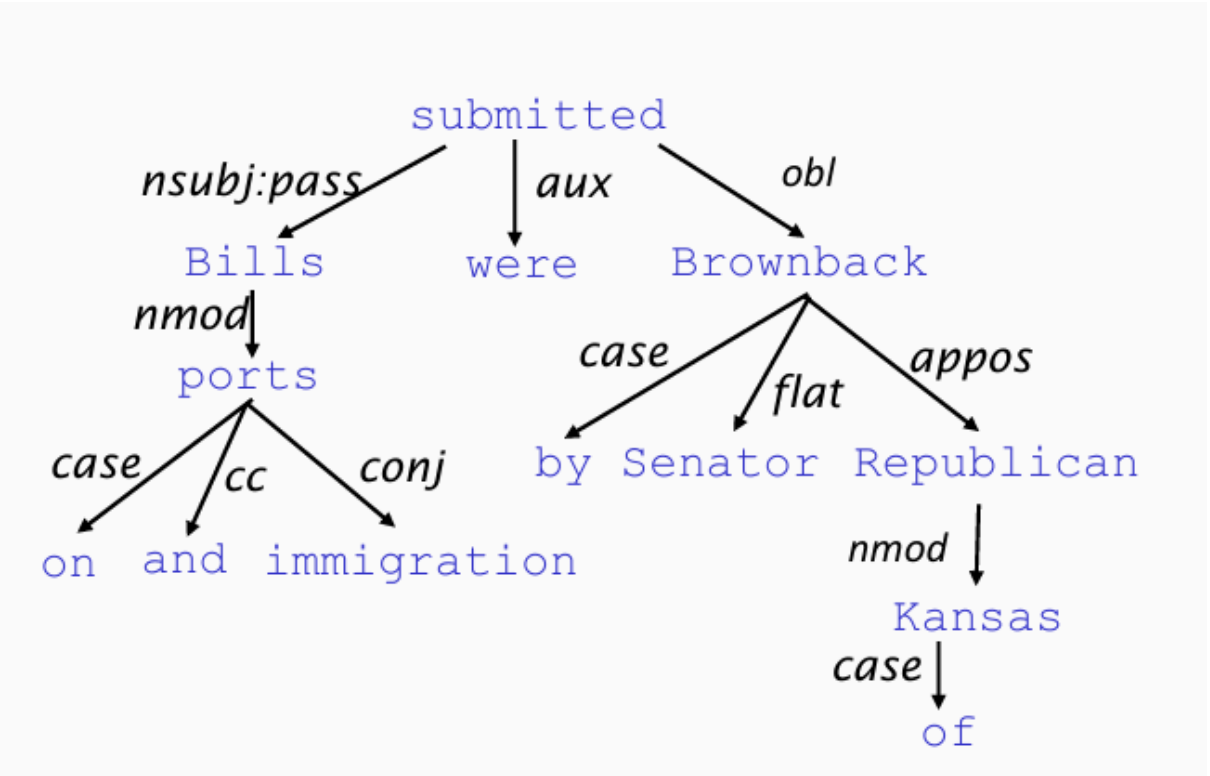
  - Ex) Shuttle veteran and longtime NASA executive Fred Gregory appointed to board
    - 수식 관계에 따라 우주선 베테랑이자 오랜 NASA의 임원인 Fred Gregory가 이사로 임명되었다, 우주선 베테랑과 오랜 NASA의 임원인 Fred Gregory가 이사로 임명되었다 2가지 의미로 분석 가능
    - 구들의 수식 범위에 따라 문장의 의미가 달라지는 모호성 존재

- **Extract semantic interpretation**



- nsubj (nominl subject): 문장의 주어 역할을 하는 명사와 동사 사이의 관계
  - ex) 'results'는 'demonstrated'의 주어
- det (determiner): 명사를 한정해주는 단어와 명사 사이의 관계
  - ex) 'The"는 'results'를 한정
- advmod (adverbial modifier): 동사, 형용사, 부사 등을 수식하는 부사와의 관계
  - 'rhythmically'는 'interacts'를 수식하는 부사

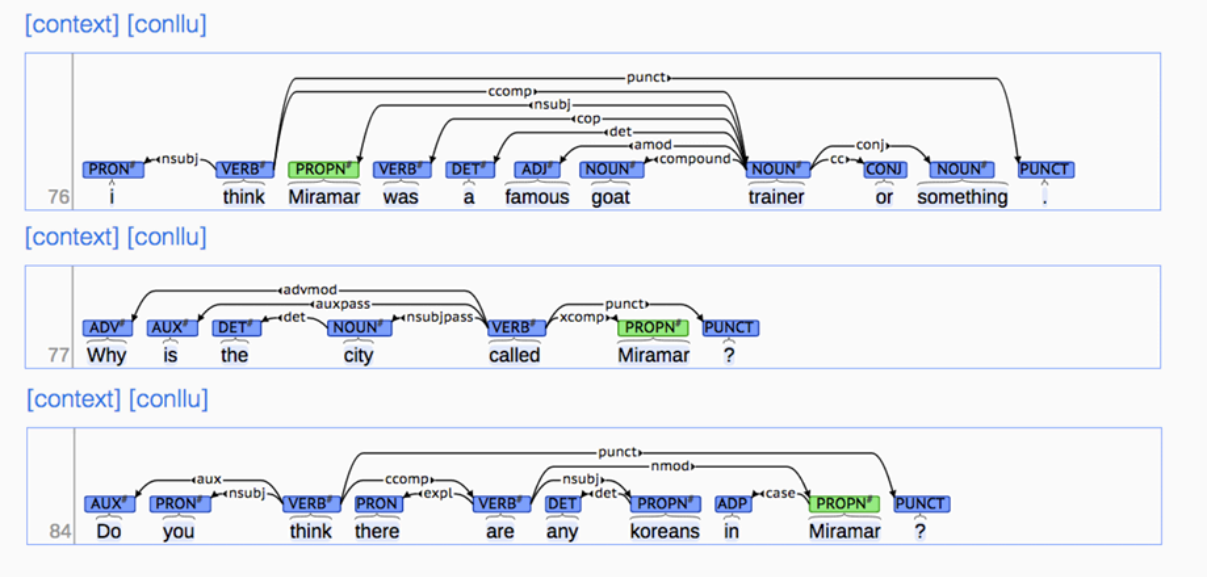
(2) Dependency Grammar and Dependency Structure



- 어휘 간의 관계를 화살표로 표시
- 의존 관계를 레이블로 표시 (ex: subject, prepositional object, apposition, etc.)
- Fake ROOT를 추가하여 모든 문장 성분이 최소 1개의 dependent가 되도록 한다

(3) The rise of annotated data & Universal Dependencies treebanks

- 문법 정보가 태깅된 데이터의 등장 (단어의 품사 정보, 단어 간 관계 정보 등)
- Universal Dependency Treebanks: 전세계의 여러 문장들에 대해, 문법 구조를 분석하고 주석 처리해둔 데이터의 모음, 전세계 언어에 대해 공통된 기준을 적용



- Treebank

1. Reusability of the labor: Treebank를 여러 분야에서 재활용할 수 있다
2. Board coverage: 소수의 직관에 의존하지 않고, 다양한 문장 구조를 다룰 수 있다
3. Frequencies and distributional information: 어떤 문법 구조가 얼마나 자주, 어떤 맥락에서 나타나는지 파악할 수 있다
4. A way to evaluate NLP systems: 자연어 처리 시스템의 평가 수단이 될 수 있다

[4] Dependency Parsing

1. **Constraint**

- 오직 한 단어만 ROOT에 대해 dependent할 수 있다
- A→B, B→A와 같은 순환 구조가 생기면 안된다

2. **Projectivity (투사성)**

- projective parse: 단어를 선형적으로 나타냈을 때, 화살표가 교차하지 않는 형태
- CFG 기반의 구조는 반드시 projective 해야 한다.

3. **Transition-based Parsing**

- Methods of Dependency Parsing
  - Dynamic programming

  :긴 문장을 여러 개로 나누어 하위 구간을 대상으로 트리를 만든다, 이후 해당 트리들을 합치는 방식으로 구문 분석 트리를 만든다

  - Graph Algorithms

  : 각 단어쌍의 의존 관계 후보에 점수를 부여하고, 확률이 높은 구문 분석 트리를 선택한다

  - Constraint Satisfaction

  : 문법적 제한 조건을 설정하고, 조건을 만족하는 단어에 대해서만 parsing하는 방식으로 구문 분석 트리를 만든다.

  - Transition-based parsing

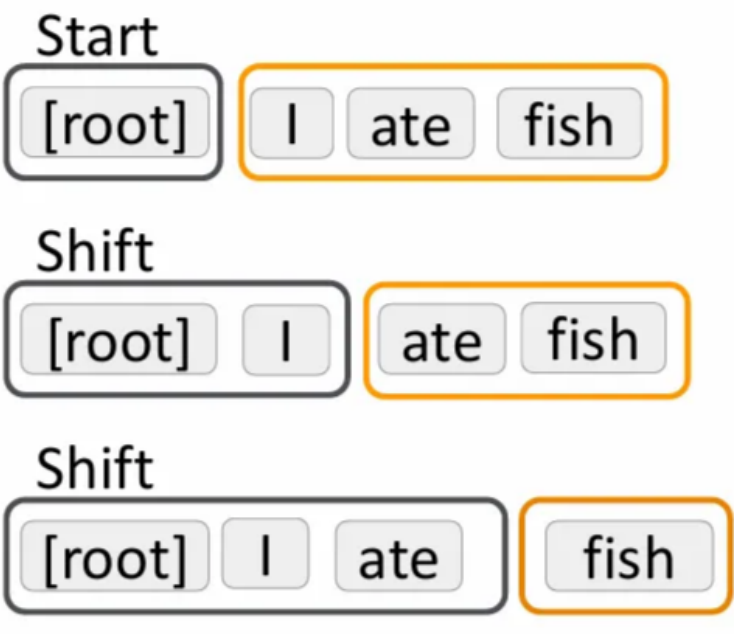
1 구성 요소

- Stack: 처리한 단어들을 저장하는 공간 (초기값: Root)
- Buffer: 아직 처리하지 않은 단어들을 저장하는 공간 (초기값: 입력 문장)
- Arcs: 현재까지 만들어진 의존 관계들의 집합 (초기값: X)
- Action: 문장 처리를 위해 사용하는 명령어 집합

2 Action

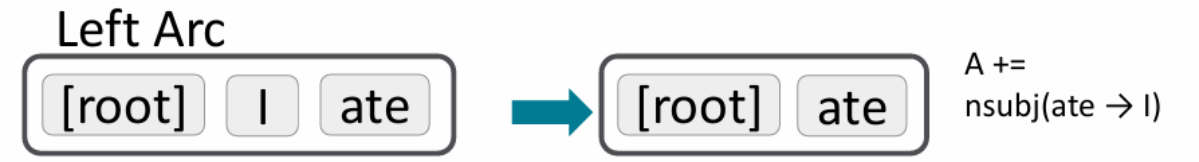
- Shift

버퍼의 맨 앞 단어를 Stack으로 이동

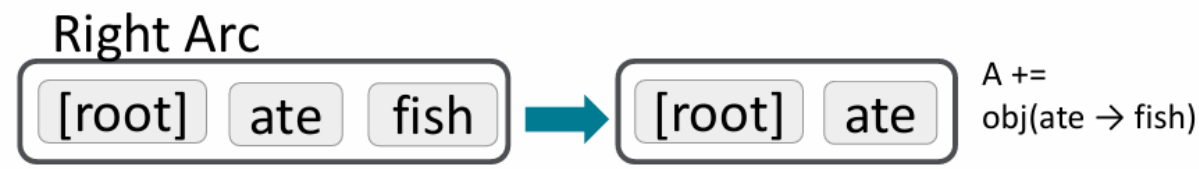


- Left-Arc

Stack의 Top 두 단어 중에서 오른쪽 단어가 왼쪽 단어의 head



- Right-Arc: Stack의 Top 두 단어 중에서 왼쪽 단어가 오른쪽 단어의 head



Start: σ = [ROOT], β = w_1, ..., w_n , A = ∅

1. Shift σ, w_i|β, A ➔ σ|w_i, β, A
2. Left-Arc$_r$ σ|w_i|w_j, β, A ➔ σ|w_j, β, A∪{$r(w_j,w_i)$}
3. Right-Arc$_r$ σ|w_i|w_j, β, A ➔ σ|w_i, β, A∪{$r(w_i,w_j)$}

Finish: σ = [w], β = ∅

Start 시기에는 Stack에 Root만, Buffer에 입력 문장이 존재한다

Shift로 첫 단어 w_i이 Buffer → Stack으로 이동

Left-Arc 연산으로 의존 관계 설정 후 Stack에서 w_i 제거

Right-Arc 연산으로 의존 관계 설정 후 Stack에서 w_j 제거

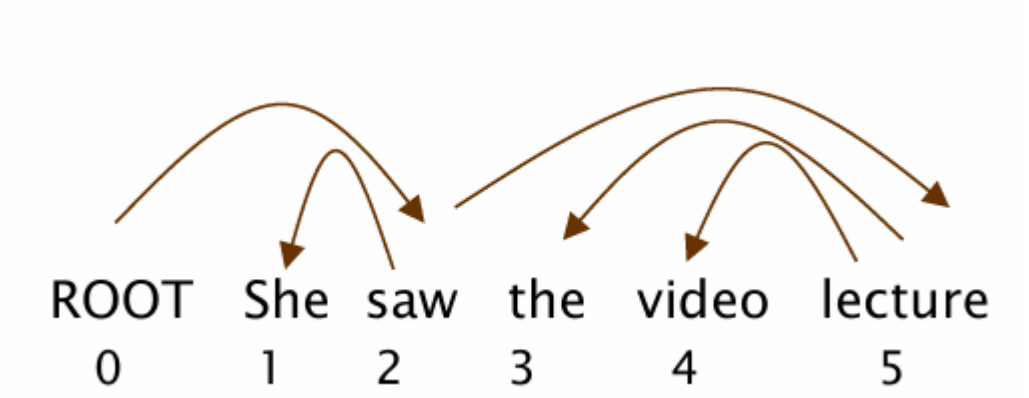
Finish 시기에 Buffer는 비어있다

3 MaltParser

- 다음에 어떤 Action을 취할 것인가에 대한 문제 → Machine Learning
  - 다음에 취할 Action을 discriminative classifier에 의해 결정
  - 선택지 = $|R| \times 2 + 1$ ($|R|$: 의존 관계 레이블 수)
  - Feature: top of stack word, POS (품사), first in buffer word, POS

4 Dependency Accuracy

ex)



Gold (정답) & Parsed (예측)

- [columns: Index/Head/word/label]

Gold

1	2	She	nsubj
2	0	saw	root
3	5	the	det
4	5	video	nn
5	2	lecture	obj

Parsed

1	2	She	nsubj
2	0	saw	root
3	4	the	det
4	5	video	nsubj
5	2	lecture	ccomp

1. UAS (Unlabeled Attachment Score): Head의 정답 여부만 평가 (80%)
2. LAS (Labeled Attachment Score): Head와 레이블의 정답 여부를 모두 평가 (40%)