Natural language processing Bible

6-7

서수원

Business Intelligence Lab.

산업경영공학과, 명지대학교

01

구문 분석

구문 분석 개요

- 구문 분석 : 자연어 문장에서 구성 요소들의 문법적 구조를 분석하는 것을 의미한다.
  - 구문 분석의 목표 : 자연어 문장의 문법적 구조를 '구문 문법'에 따라 자동으로 분석하는 것이 목표이다.

- 구문 문법 : 언어학에서 문법적 구성 요소들로부터 문장을 생성하기도 하고, 문장을 구성요소들로 분석하기도 한다.
  - 구문 문법을 정의 하는 것은 구문 분석에서 중요한 요소 중 하나이다.
  - 구문 분석 기술에서 활용되는 구문 문법은 구구조 문법과 의존 문법이 있다.

구구조 문법 (Phrase Structure Grammar)	의존 문법 (Dependency Grammar)
• 노암 촘스키(Noam Chomsky)가 제안	• 뤼시앵 테니에르(Lucien Tesnière)가 제안
• 구성소 관계(constituency relation)에 기반하여 문장 구조 분석	• 의존 관계(dependency relation)에 기반하여 문장 구조 분석
• 단어들이 모여 절을 구성하며, 절과 단어들의 계층적 관계에 따라 문장이 이루어진다고 분석	• 문장을 구성하는 단어들 간의 계층적인 의존 관계에 따라 문장이 이루어진다고 분석
• 문장 전체를 트리 구조로 분석할 때, 절과 단어들은 각각의 노드로 표현	• 문장 전체를 트리 구조로 분석할 때, 단어들은 각각의 노드로 표현되며 에지는 단어 간 의존 관계를 나타냄

구구조 구문 분석 - 규칙기반

- 구구조 구문 분석 : 문장 구성 요소의 구조가 비교적 고정적인 언어에 적합하다.
  - 규칙 기반, 통계 기반, 딥러닝 기반 구문 분석 방법이 존재한다.
- 문법 규칙의 형태로 '구구조 문법 ' 을 활용한다.
  - 구구조 문법 : 자연어 문장을 하위 '구성소' 들로 나눔으로써 문장 구조를 나타내는 문법을 말한다.
- 대표 알고리즘으론 CYK알고리즘이 있다.

■ 구구조 문법을 적용한 구문 분석 과정 예시

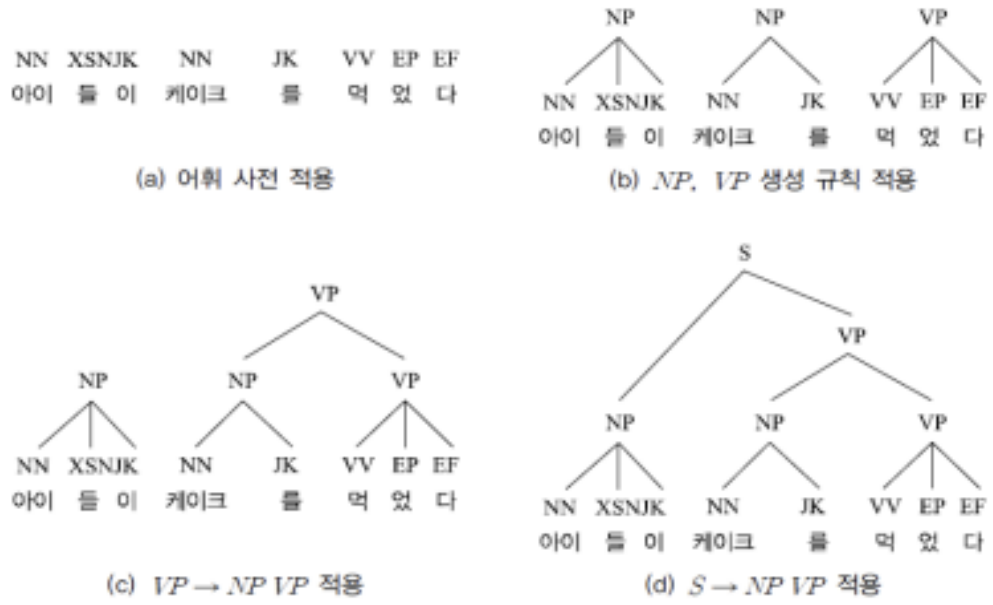


그림 6-4 구구조 문법을 적용한 의존 분석 과정

구구조 구문 분석 - 규칙기반

- CYK 알고리즘
  - Cocke-Younger-Kasami 알고리즘 이라고 한다.
  - 현재 모든 문맥 자유 문법을 파싱할 수 있는 가장 효율적인 알고리즘이다.
  - 문자열의 길이가 n일때 n^3의 시간 복잡도를 가진다.
  - 촘스키 정규형식으로 표현된 문법을 사용한다.
  - CYK알고리즘 보다 효율적인 알고리즘도 존재 하지만, 특정한 상황에만 이용 가능 하다는 단점이 있다.

구구조 구문 분석 - 통계기반

- 통계 기반 구구조 구문 분석 : 통계적으로 확률적 구구조 문법을 계산하여 구문 분석을 수행하는 방법이다.
- 확률적 구구조 문법 : 각 규칙에 대한 조건부 확률이 정의 된다.

‖ 표 6-2 ‖ 확률적 구구조 문법으로 나타낸 문법 규칙과 어휘 사전 예시

문법 규칙(The Grammar)	어휘 사전(The Lexicon)
$S \rightarrow NP\ VP\ [.99]$	$NN \rightarrow$ 아이[.60] | 케이크[.40]
$NP \rightarrow NN\ XSN\ JK\ [.20] \mid NN\ JK\ [.80]$	$XSN \rightarrow$ 들[.99]
$VP \rightarrow NP\ VP\ [.90] \mid VV\ EP\ EF\ [.10]$	$JK \rightarrow$ 이[.75] | 를[.25]
	$VV \rightarrow$ 먹[.99]
	$EP \rightarrow$ 었[.99]
	$EF \rightarrow$ 다[.99]

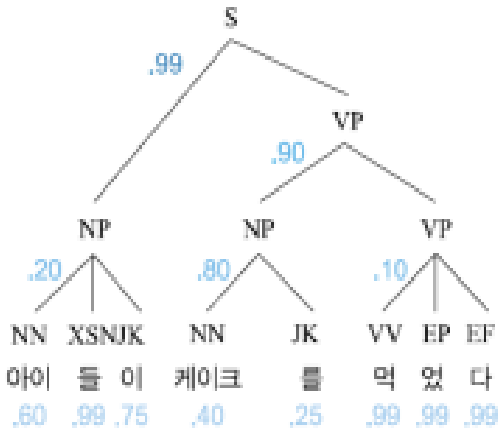
- A -> BC[p]

구구조 구문 분석 - 통계기반

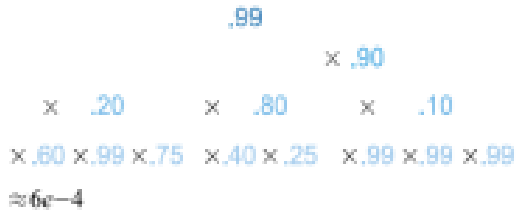
- 확률적 구구조 문법 규칙의 두가지 계산법
  - 인간이 직접 태깅한 구구조 구문 분석 코퍼스로부터 각 규칙이 나타나는 조건부 확률을 계산한다
  - 태깅되지 않은 자연어 문장에 구구조 구문 분석을 수행 해서, 문법 규칙의 조건부 확률을 조정한다.
    - ✓ Inside-outside 알고리즘 : 이러한 방식으로 확률적 구구조 문법 규칙을 계산하는 대표적 알고리즘 이다.

- 각 문법 규칙의 조건부 확률에 기반하여 가능한 구문 분석 결과 전체의 확률을 계산 할 수 있다
  - 한 문장에 대해 가능한 여러 구문 분석 트리 중 분석 결과의 전체 확률이 가장 높은 것을 결과로 제시한다.

$$P(T|S) = \prod_i P(\alpha_i \rightarrow \beta_i | a_i)$$



(a) 의존 분석 결과와 문법 규칙별 확률



(b) 의존 분석 결과 확률 계산식

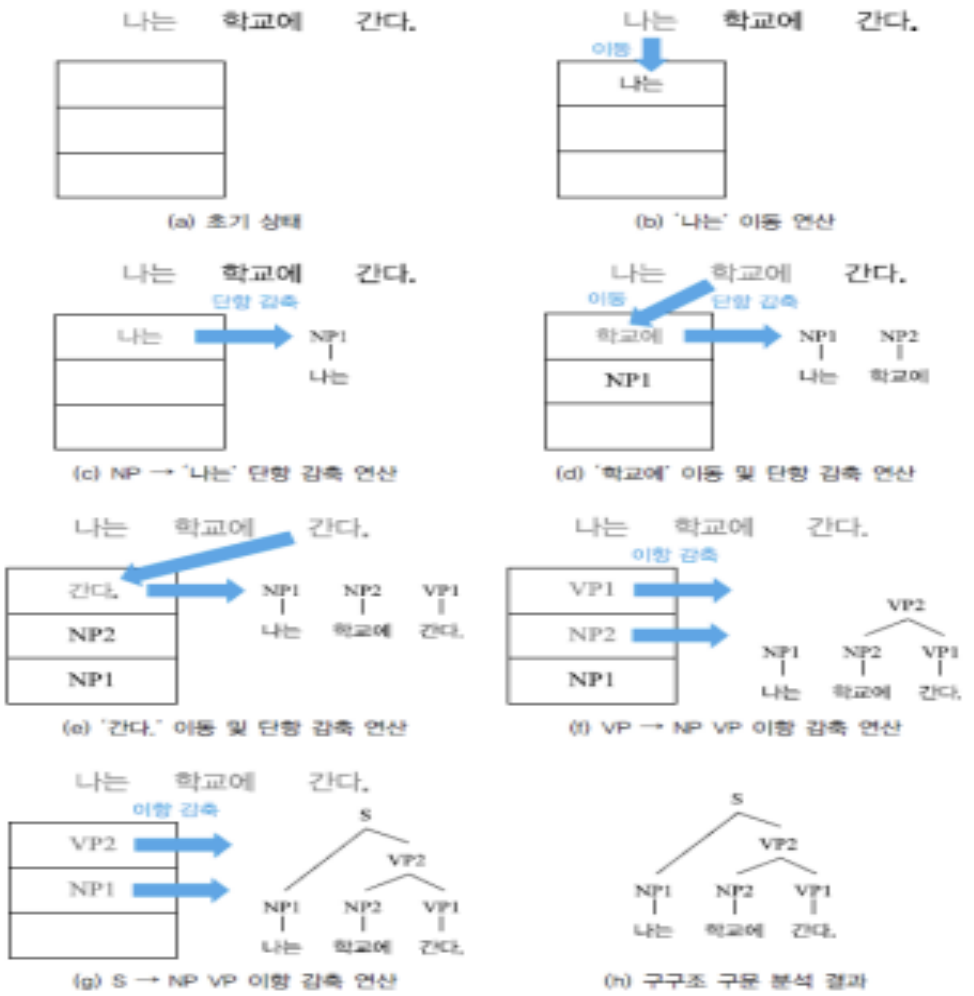
구구조 구문 분석 – 딥러닝 기반

- 딥러닝 기반 구구조 구문 분석 : 인간이 구축한 구구조 구문 분석 데이터셋으로부터 딥러닝 모델을 학습하여 구문 분석을 수행하는 방법이다.
- 대표적인 방법으로는 전이 기반 파싱이 있다.
  - 전이기반 파싱 : 자연어 문장을 한 단어씩 읽으며 현재 단계에서 수행할 액션을 선택하는 방식이다.
  - 이동-감축 파싱이 대표적이다.
    - ✓ 이동 연산 : 자연어 문장에 포함된 단어를 순차적으로 스택에 이동 시키는 연산을 의미한다.
    - ✓ 감축 연산 : 스택에 저장된 하나 또는 두개의 구성소를 꺼내 상위 구성소로 감축한 뒤 상위 구성소를 다시 스택에 이동 시킨다.

구구조 구문 분석 – 딥러닝 기반

- 이동 – 감축 파싱을 이용한 구구조 구문 분석 수행 예시

- 전이 기반 파싱
  - 장점 : 입력된 자연어 문장에 포함된 단어 수에 선형적인 전이 액션으로 구문분석이 가능하다.
  - 단점 : 각 전이 액션 선택시 문장 전체의 문법적 구조를 고려 하는 것이 어렵다. 오류 전파에 취약하다.



의존 구문 분석

- 의존 구문 분석
  - 자연어 문장에서 단어 간의 의존 관계를 분석함으로써 문장 전체의 문법적 구조를 분석하는 기술이다.
  - 대부분의 의존 관계 표현은 각 단어를 노드로 하고 단어간 의존 관계를 엣지로 나타내는 트리 구조를 따른다.
  - 비교적 문장 구조가 유연한 언어에 적합하다.

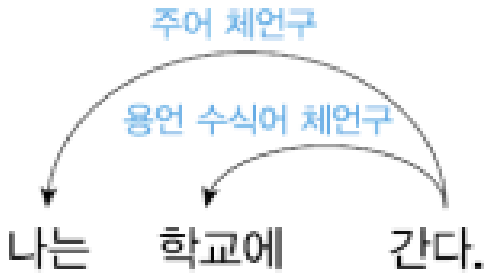


그림 6-7 의존 문법으로 표현한 구문 구조 예시

의존 구문 분석 – 규칙 및 통계 기반

- 규칙 기반 의존 구문 분석 : 의존 문법의 형태로 문법 규칙을 저장한 뒤 이를 적용한다.
  - 지배소 : 의존 관계 표현에서 절의 중심이 되는 단어이다.
  - 의존소 : 절 내에서 지배소에 의존하는 단어이다.
- 통계적 의존 구문 분석 : 문맥 의존 규칙의 조건부 확률을 통계적으로 계산하여 적용한다.

- ‘나는 학교에 간다.’라는 문장은 전체가 하나의 어절이다.
- 지배소 : ‘간다.‘
- 의존소 : ‘나는‘, ‘학교에’

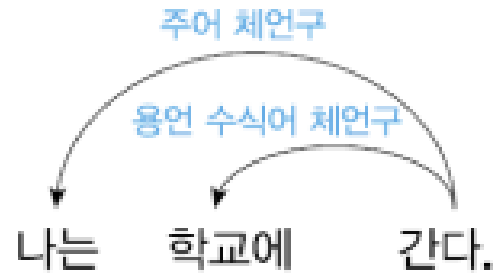


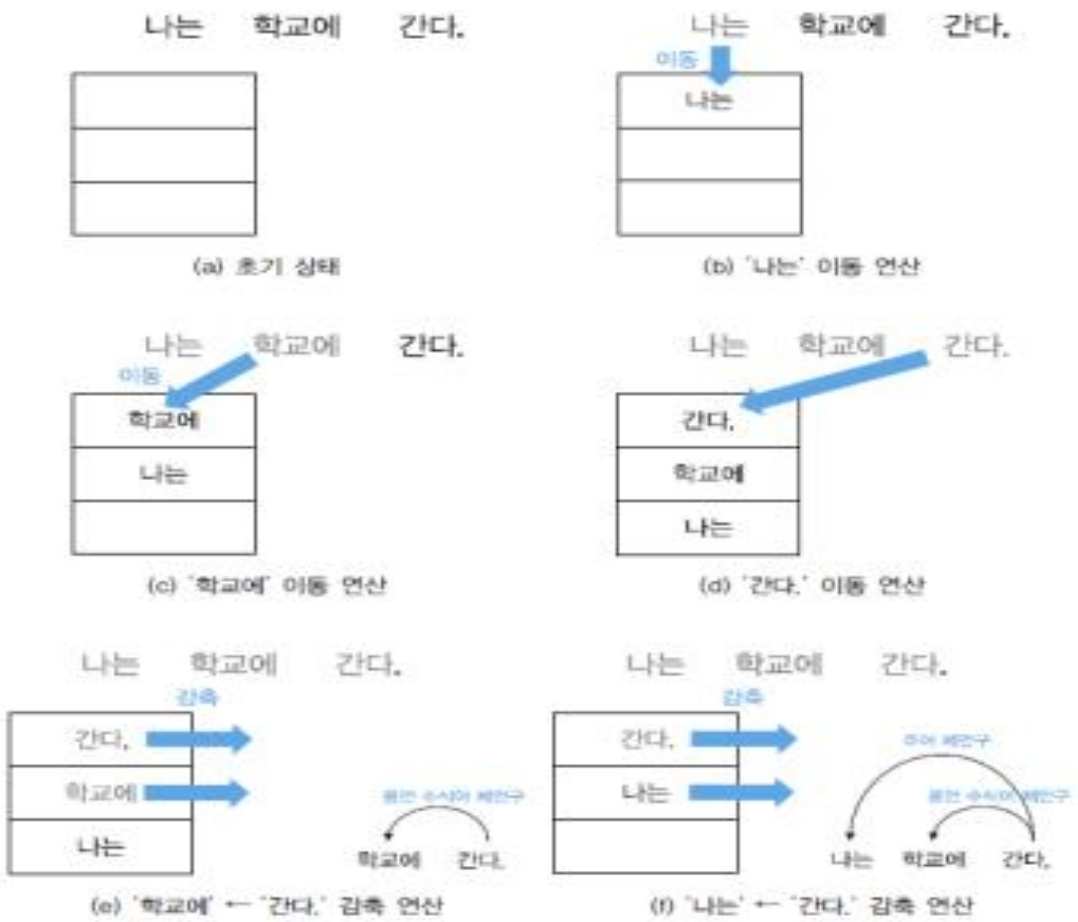
그림 6-7 의존 문법으로 표현한 구문 구조 예시

의존 구문 분석 – 딥러닝 기반

- 딥러닝 기반 의존 구문 분석
  - 인간이 구축한 의존 구문 분석 데이터셋으로부터 딥러닝 모델을 학습하여 구문 분석을 수행한다.
  - 딥러닝 모델을 이용해 가장 적합한 의존 분석 트리를 구축한다.
    - ✓ 전이 기반 파싱과 그래프 기반 파싱이 존재한다.

의존 구문 분석 – 딥러닝 기반(전이 기반 파싱)

- 전이 기반 파싱
  - 자연어 문장에 포함된 단어를 하나씩 의존 분석 트리에 포함시킴으로써 의존 구문 분석을 수행한다



의존 구문 분석 – 딥러닝 기반(그래프 기반 파싱)

- 그래프 기반 파싱
  - 자연어 문장에 포함된 단어 간의 가능한 모든 의존관계에 대한 점수를 계산한 뒤 가장 높은 점수를 갖는 트리를 선택하는 방법이다.

  - 장점 : 지역적 정보에 국한되는 전이 기반 파싱에 비해 문장 전체의 문법적 구조를 고려할 수 있다.
  - 단점 : 시간복잡도가 높아 실사용 단계에서 비효율성이 발생할 수 있다.

구문 분석 방법의 장단점

- 규칙 기반 구문 분석 방법의 장단점
  - 장점 : 미리 정의된 문법 규칙을 적용할 수 있는 문장에 대해서는 정확한 의존 분석이 가능하다.
  - 단점 : 문법 규칙을 미리 정의 하기위한 시간과 비용 문제가 발생한다.
    자연어 문장의 중의성 처리시 문제 발생한다.
- 통계 기반 구문 분석 방법의 장단점
  - 장점 : 구문 중의성을 갖는 문장에서도 확률적으로 계산하여 타당한 결과를 선택 가능하다.
  - 단점 : 장거리 의존 관계를 고려 하기가 어렵다.
- 딥러닝 기반 구문 분석 방법의 장단점
  - 장점 : 규칙 기반과 통계 기반 방법에서는 활용하기 어려운 문장 전체 구조 정보, 어휘의 하위 범주화 정보 등을 구문 분석에 활용할 수 있다.
  - 단점 : 대량의 파라미터가 학습 결과로 나타난다. 따라서 구문 분석 결과의 근거가 해석 불가능하다.

02

의미 분석

단어와 단어 의미 중의성

- 중의성
  - 둘 이상의 의미를 가지는 표현을 의미한다.
  - 어휘적 중의성
    - ✓ 다의어에 의한 중의성을 의미한다.
      - ❖ 예) 손 좀 보다.
        - » ‘신체 일부‘,’수리‘,’혼 내다 ' 의 의미를 가진다
  - 동음어에 의한 중의성
    - ✓ 동음이의어에 대한 중의성을 의미한다.
      - ❖ 예) 밤이 좋다.
        - » ‘시간’,’음식’을 의미한다.
  - 구조적 중의성
    - ✓ 수식어에 의한 중의성을 의미한다
      - ❖ 예) 부유한 철수와 영희가 명품 매장을 갔다.
        - » ‘부유한’의 수식어가 철수를 수식 하는지, 영희를 수식하는지를 모른다.

지식기반 방법

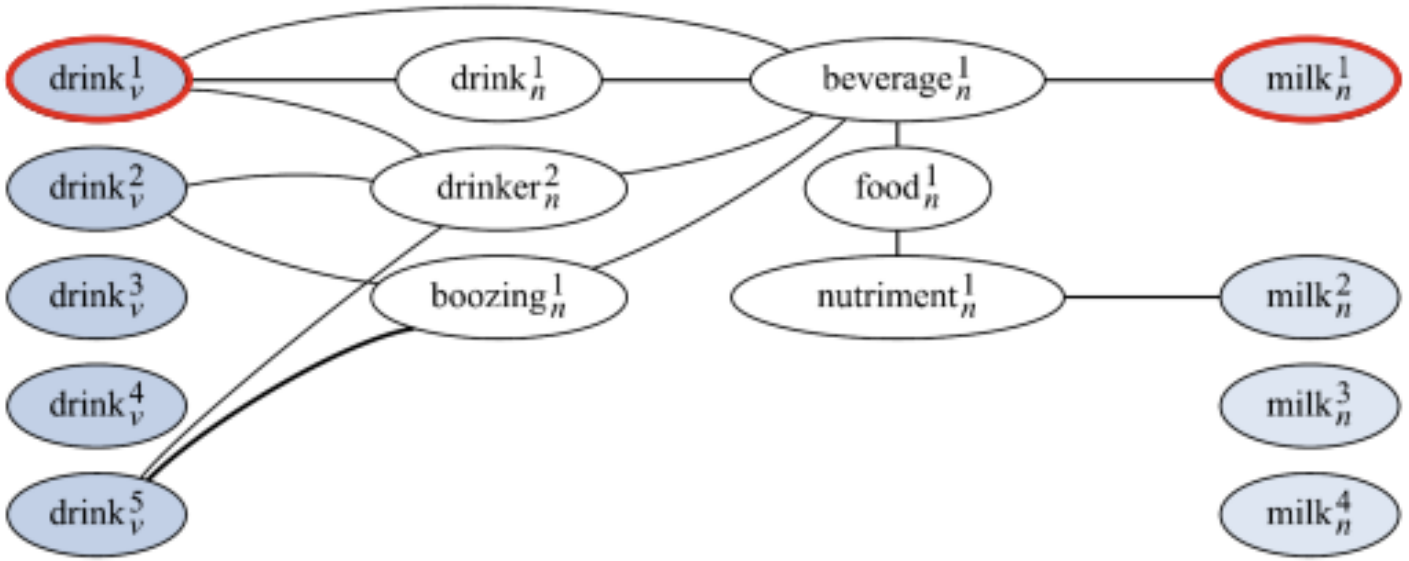
- 지식기반 방법
  - 문장에 등장한 단어들을 사전에 정의된 어휘 지식을 활용하여 예측하는 방법을 의미한다.
  - 사전 정의 기반 방법과 그래프 기반 방법이 있다.
- 사전 정의 기반 방법
  - Lesk 알고리즘을 사용한다.
    - ✓ 중의성 단어의 사전 뜻풀이에 쓰인 단어들과 주변 문맥에 나타난 단어의 사전 뜻풀이에 쓰인 단어들 사이에 중복되는 단어가 가장 많은 의미를 중의성 단어의 의미로 선택 한다.
    - ✓ 한계점 : 단어 간의 정확한 일치 기반이다.
      사전 정의에 의존적이다.

그 사람은 수술을 통해 불편한 다리를 고쳤다.

단어		사전 뜻풀이에 쓰인 단어
함께 나타난 단어	사람	생각, 언어, 만들다, 쓰다, 사회, 살다, 동물, …
	수술	피부, 점막, 조직, 기계, 병, 고치다, …
		…
중의성 단어	다리 01	사람, 동물, 몸통, 신체, …
	다리 02	물, 건너다, 시설물, …

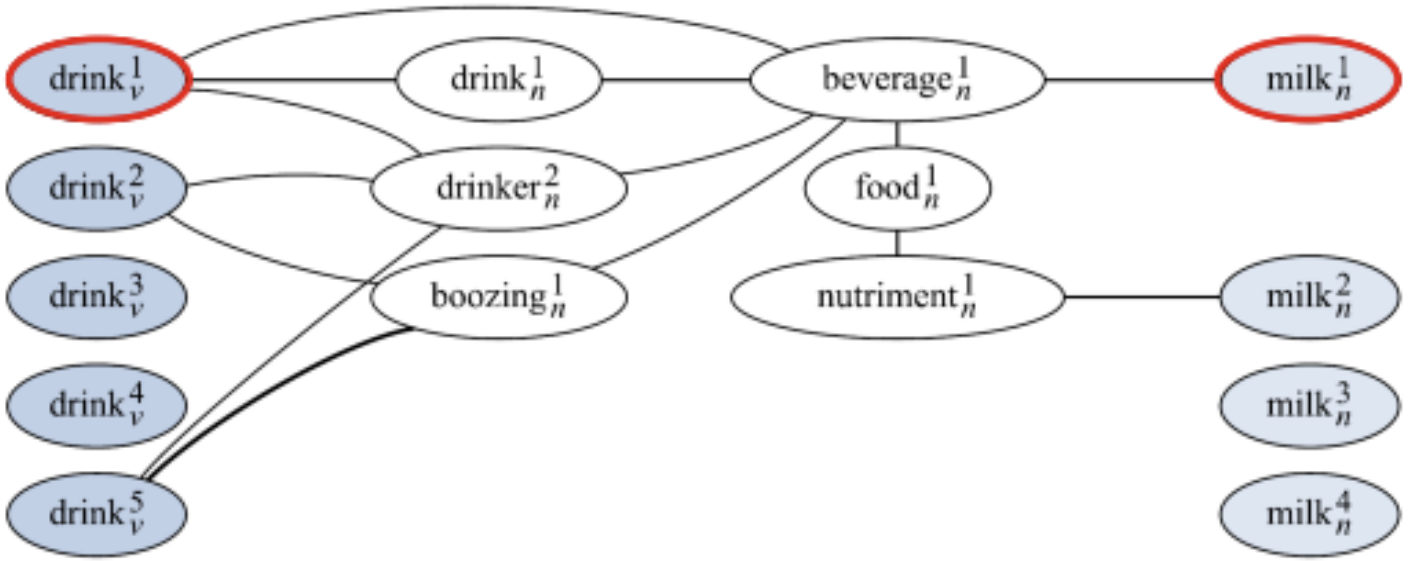
지식기반 방법

- 그래프 기반 방법
  - 단순 그래프 기반 방법
  - DFS,BFS알고리즘을 이용하여 검색되는 Edge를 추출한다.
  - 의미간 연결성 측정을 통해 가장 많이 연결된 의미를 선택한다.



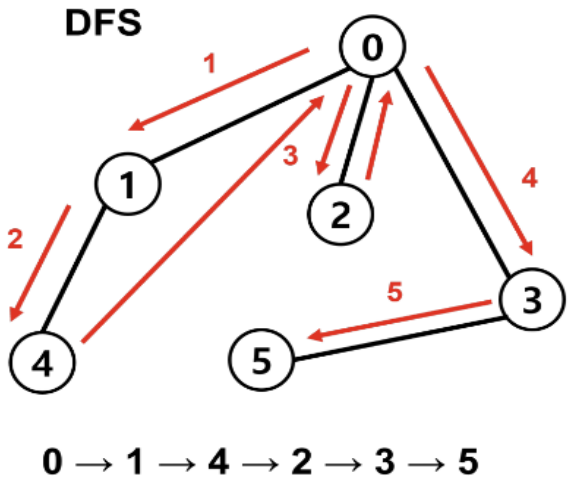
지식기반 방법

- 그래프 기반 방법
  - 단순 그래프 기반 방법
  - DFS(Depth-First Search),BFS알고리즘을 이용하여 검색되는 Edge를 추출한다.
  - 의미간 연결성 측정을 통해 가장 많이 연결된 의미를 선택한다.



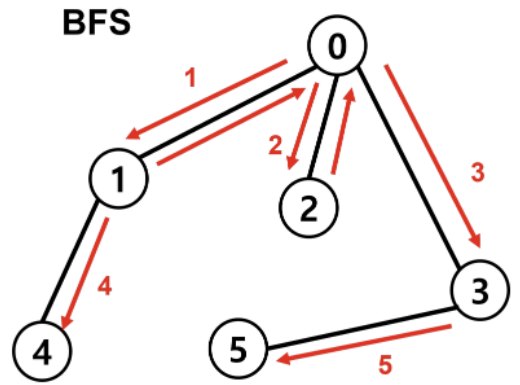
지식기반 방법

- DFS
  - 깊이 우선 탐색을 의미한다.
  - 하나의 정점부터 시작하여 아래로 깊게 내려가며 탐색한다.



0 → 1 → 4 → 2 → 3 → 5

- BFS
  - 너비우선탐색을 의미한다.
  - 시작 노드 인접 노드부터 확인하는 알고리즘이다.



0 → 1 → 2 → 3 → 4 → 5

지도학습 기반 방법

- 지도학습 기반 방법
  - 각종 기계 학습 알고리즘을 통해 단어 의미를 분석한다.
  - 기계학습 분류기 모델은 사용자가 정의한 규칙에 맞춰 성능을 높여왔다.
  - 학습한 특정 중의성 단어에 대해서만 해결 가능하다.
  - 나이브 베이지안 분류, K-NN,SVM등을 사용 가능하다.

지도학습 기반 방법

- 나이브 베이지안 분류
  - 베이지안 정리를 이용한 확률적 기계학습 알고리즘이다.
  - 사전확률에 기반을 두고 사후확률을 추론한다.
  - 모든 사건이 독립이라는 가정을 한다.

$$c = \arg\max_{c_k} P(c_k|\vec{x})$$

$$= \arg\max_{c_k} \frac{P(\vec{x}|c_k)}{P(\vec{x})} P(c_k)$$

$$= \arg\max_{c_k} [\log P(\vec{x}|c_K) + \log P(c_K)]$$

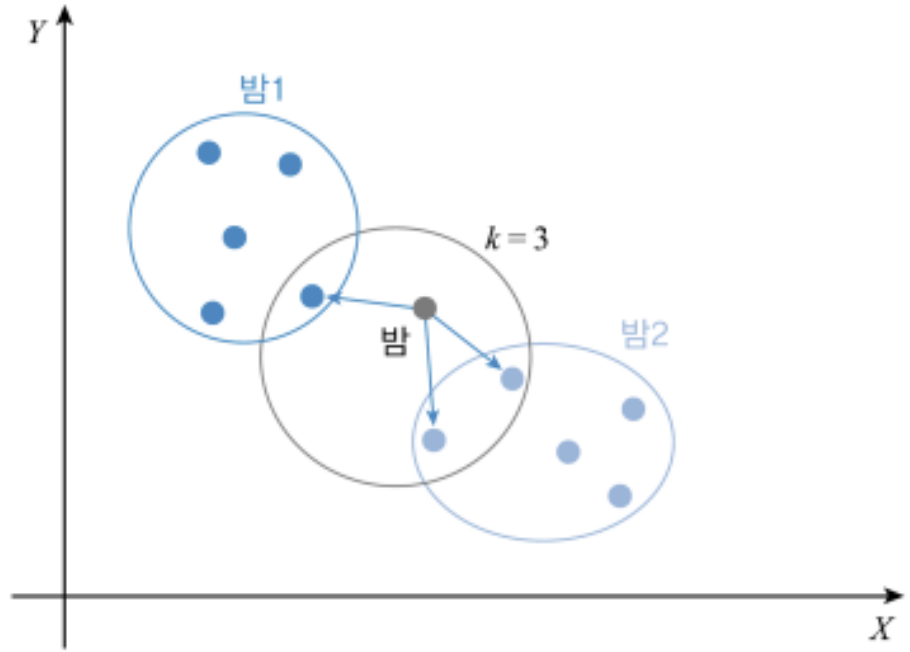
$$= \arg\max_{c_k} [\sum_{v_j \in x} \log P(v_j|c_K) + \log P(c_K)]$$

지도학습 기반 방법

- K-NN분류
  - 벡터 공간에서 유클리드 거리와 코사인 유사도를 활용해 가까운 데이터끼리 분류한다.

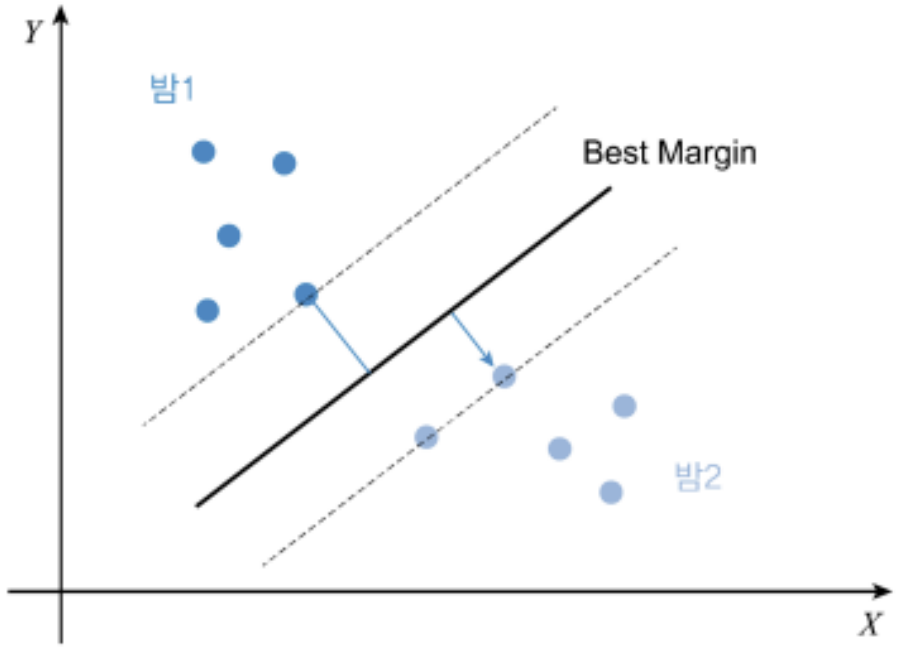
$$Euclidean\ Distance = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

$$Cosine\ \ similarity = \frac{A \cdot B}{\|A\|\,\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}}$$



지도학습 기반 방법

- SVM
  - 벡터 공간에 표현된 자질로부터 의미 클래스를 분류하기 위해 클래스 간에 가장 넓은 거리를 사용하는 방향으로 선을 그어 의미를 분류한다.



의미역 분석

- 의미역은 두가지로 구분된다.
  - 필수적 의미역
    - ✓ 서술어의 의미를 구성하는데 필수적으로 요구되는 것을 의미한다.
      - ❖ 행동주 ,도구,피동주,경험자 등으로 분류가 된다.
  - 수의적 의미역
    - ✓ 서술어의 의미를 보충해 주는 것을 의미한다.
      - ❖ 장소,위치 이유, 목적,시간 ,경로,방법으로 분류가 된다.

| 표 7-1 | 필수적 의미역

의미역	예시
행동주(agent)	철수(agent)가 돈을 낸다.
도구(instrument)	철수(agent)가 망치(instrument)로 못을 박는다.
피동주/수동자(patient)	철수가 민희(patient)를 사랑했다.
경험자(experiencer)	영희(experiencer)가 사랑에 빠졌다.
수혜자(benefactive)	내가 철수(benefactive)를 위해 밥을 사줬다.
출처/근원(source)	나는 식당(source)에서 밥을 주문했다.
도달점/목표(goal)	나는 책을 서랍(goal) 안에 보관했다.

지도학습 기반 의미역 분석

- 지도학습 기반 의미역 분석
  - 단어 의미 중의성 해소의 지도학습기반 방법과 마찬가지의 기계학습을 사용한다.
- 의미역 성능을 높이기 위해 문법적, 의미적 자질을 사용
  - 문법적 자질이란 형태소, 구문정보 등을 의미한다.
  - 의미적 자질이란 개체명과 같은 것을 의미한다. Ex)눈
  - 문법적, 의미적 자질을 사용하면 높은 성능을 보이지만, 자질을 추출하기 위해 대량의 학습데이터가 학습된 분류기를 따로 구축해야 한다는 단점이 존재한다.

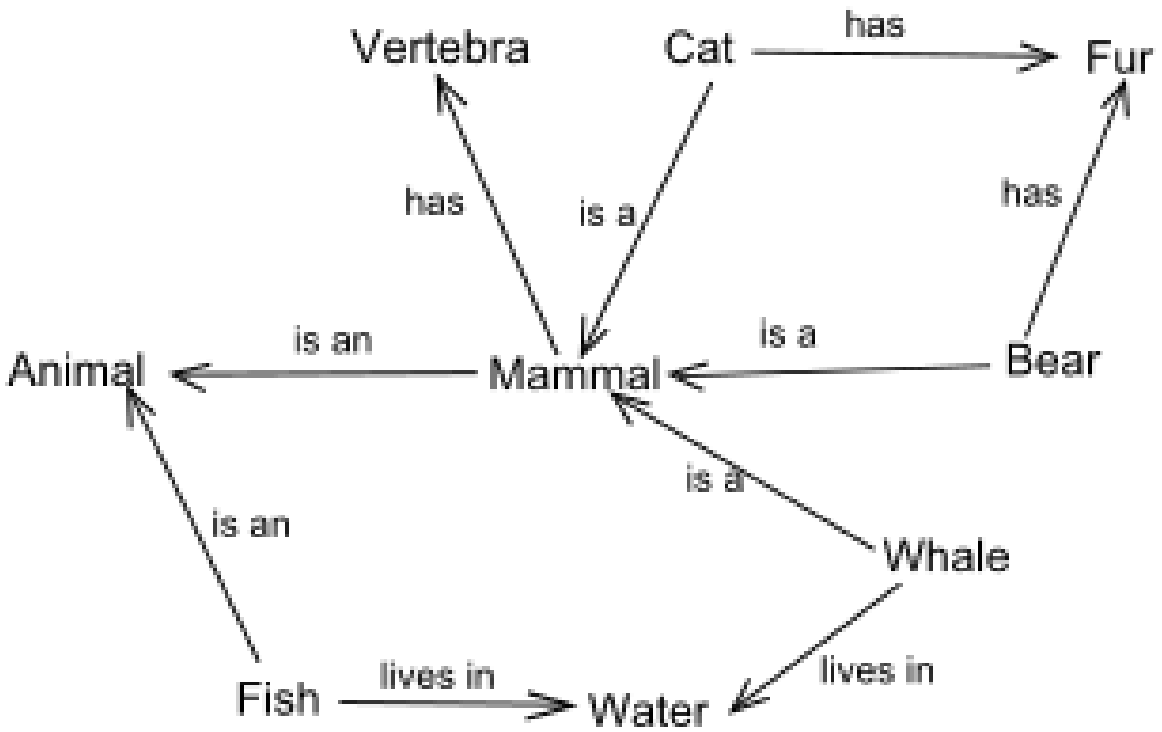
의미표현

- 의미표현
  - 대화자들은 의미 구조에 대한 지식으로 문장사이의 모순관계, 중의성 등의 관계를 정확하게 파악한다.
  - 의미표현이란 대화자들의 다양한 언어적 표현을 파악하는 단계를 의미한다.

- FOL (First-Order Logic)
- Conceptual Dependency Diagram
- Semantic Network
- Frame-based Representation
- Abstract Meaning Representation (AMR)

의미표현

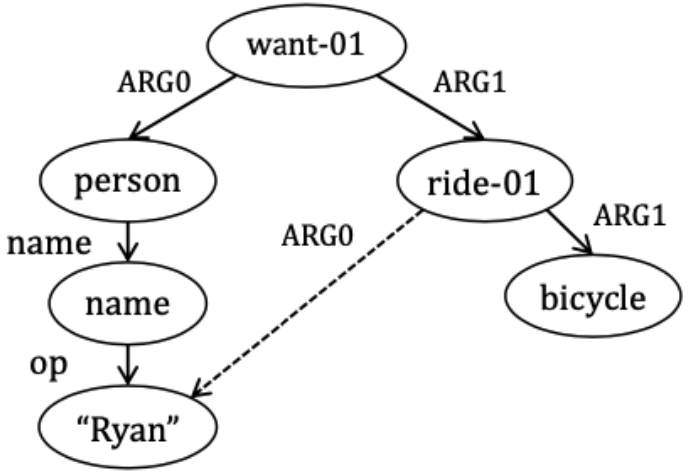
- Semantic Network
  - 매우 복잡한 분류나 인과관계를 갖는 추론에 자연스러운 표현이 가능하다.
  - 지식베이스의 크기가 커지면 너무 복잡해지므로 다루기가 힘들다는 단점이 있다.



의미표현

- AMR
  - 문장의 의미구조를 그래프로 표현한 것을 의미한다.
  - 실제 일어난 일과 미래,가정,소망을 구별하지 못한다
  - 시제 및 수 표현에 따른 접사 및 모음의 변화를 포함하지 않는다.

Ryan wants to ride bicycle.



Origin Text

AMR Graph