Ch5 형태소 분석

어휘 분석: 단어의 구조를 식별하고 분석을 통한 어휘의 의미와 품사에 관한 단어 수준의 연구

형태소 분석: 형태소를 자연어의 제약 조건과 문법 규칙에 맞춰 분석하는 것

#1.1 형태소 분석 절차

1. 형태소로 분리

- 처리 대상인 어절(단어)는 하나 이상의 형태소가 연결된 것
- 이를 형태소 열이라고 부르기도 함(한국어는 = 한국어 + 는)
- 형태소 연결 시, 형태소의 변형이 일어나므로 복원이 필요함(나는(flying) = 날 + 는)

2. 형태론적 변형이 일어난 형태소의 원형 찾기

- 하나의 형태소는 하나 이상의 형태소와 품사의 쌍으로 표현됨
- 형태소와 그 형태소의 품사를 쌍으로 나타낸 것을 형태소 품사쌍이라고 함
- 예) 나 - 대명사_나, 명사_나, 동사_나, 보조용언_나

3. 단어와 사전들 사이의 결합 조건에 따라 옳은 분석 후보 선택

- 형태소 품사 쌍열 후보군(나 - 대명사_나, 명사_나, 동사_나, 보조용언_나) 중 선택

#1.2 영어 형태소 분석

영어에서 최소 단위의 의미를 갖는 기본 단위는 단어이다. 따라서 어간 추출(stemming), 표제어 추출(lemmatization)을 통해 쉽게 형태소 파악 가능하다.

일반적으로 영어의 형태소는 접사이다. 접사는 접미사와 접두가로 나뉜다.

접사 제거 시 의미가 바뀌는 단어들이 존재하며, 최소한의 의미를 가진 형태소를 찾아 원형 분석 필요함.

#1.3 한국어 형태소 분석 라이브러리

한국어 형태소 분석기의 오픈 라이브러리

- KoNLPy- 한나눔, 코모란, 미캡(성능이 좋아 주로 씀), 꼬꼬마, 트위터
- Khiii(Kakao Hangul Analyzer |||) - 딥러닝(CNN)을 이용한 형태소 분석기

- 기준, 성능, 시간이 각각 다르므로 데이터에 맞는 분석기 활용

#2.1 품사 태깅이란?

-태깅: 같은 단어에 대해 의미가 다를 경우(중의성)를 해결하기 위해 부가적인 언어의 정보를 부착하는 것.

-품사 태깅: 문서 또는 문장을 이루고 있는 각 단어에 정확한 하나의 품사를 부여하는 것.

많은 단어가 형태론적 중의성을 가지기 때문에 품사태깅은 형태론적 중의성 해결이 필수적. 이를 위해선 문맥을 고려해야함

#2.2 형태론적 중의성 해결 방법

* 자동 품사 태깅 방법(지식 기반 태깅 방법, 통계 기반 태깅 방법)
    * 지식 기반 태깅 방법
        - 문맥틀(context frame) 형식으로 규칙을 기술하는 법
        - 제약 문법(constraint grammar)을 이용하여 규칙을 표현한 방법
        - 원시 말뭉치로부터 출현 빈도가 높은 중의적 단어를 처리하는 규칙과 휴리스틱 규칙 그리고 비문맥 규칙을 사용하는 방법
        - 패턴-처리 형태의 부정 지식을 나타내는 규칙, Finite-state intersection frammar를 사용하는 법
    * 통계적 품사 태깅 방법
        - 변형 마르코프 모형에 기반한 방법
        - 통계적 결정 트리에 기반한 방법
        - 최대 엔트로피 모형에 기반한 방법
        - 신경망에 기반한 방법
        - 베이지언 추론에 기반한 방법
        - 반복 알고리즘의 일종인 labelling 기법에 기반한 방법
        - 퍼지망에 기반한 방법
        - 분별 학습에 기반한 방법

#2.3 품사 태깅 접근법(규칙, 통계, 딥러닝 기반)

1) 규칙 기반의 접근법

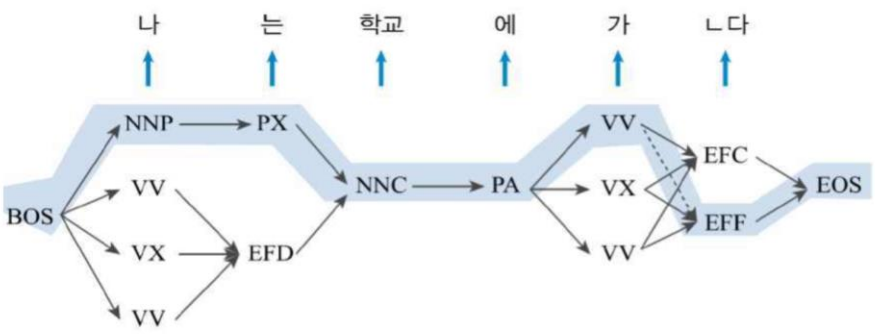
- 언어 정보에서 생성되는 규칙 형태로 표현, 이를 적용하여 태깅을 수행
- 품사 사이 관계 외에 어절에 대해서 높은 정확도를 나타내기 때문에 통계 기반 접근법으로 다루지 못하는 부분에 대해 교정 가능
- 언어 전문가가 완전히 수동으로 품사 태깅 데이터를 구축하거나 최소한의 규칙으로 자동 또는 반자동으로 구축 가능
- 수동으로 구축 시 정확성이 높지만, 시간과 노력이 많이 소요
- 자동이나 반자동일 시 규칙(코퍼스)에 의존적
- 기존의 접근법은 긍정 정보, 부정 정보, 수정 정보를 이용하여 중의성을 해결하고 태깅을 부착하는 방법이다.

2-1) 통계 기반의 접근법 hidden markov model(HMM)

- 태그가 부착된 대량의 코퍼스가 주어지면 적합한 모델을 선정하고 코퍼스에서 추출된 통계정보를 이용
- 대량의 코퍼스에 태그가 부착되어야 하는 단점이 있으나 주어지면 통계정보 추출이 용이, 자동 추출 가능
- 대표적으로 어휘 확률만을 이용하는 방법인 은닉 마르코프 모델 접근법(hidden markov model(HMM))이 존재(딥러닝 이전의 성능이 가장 좋은 접근 방법)

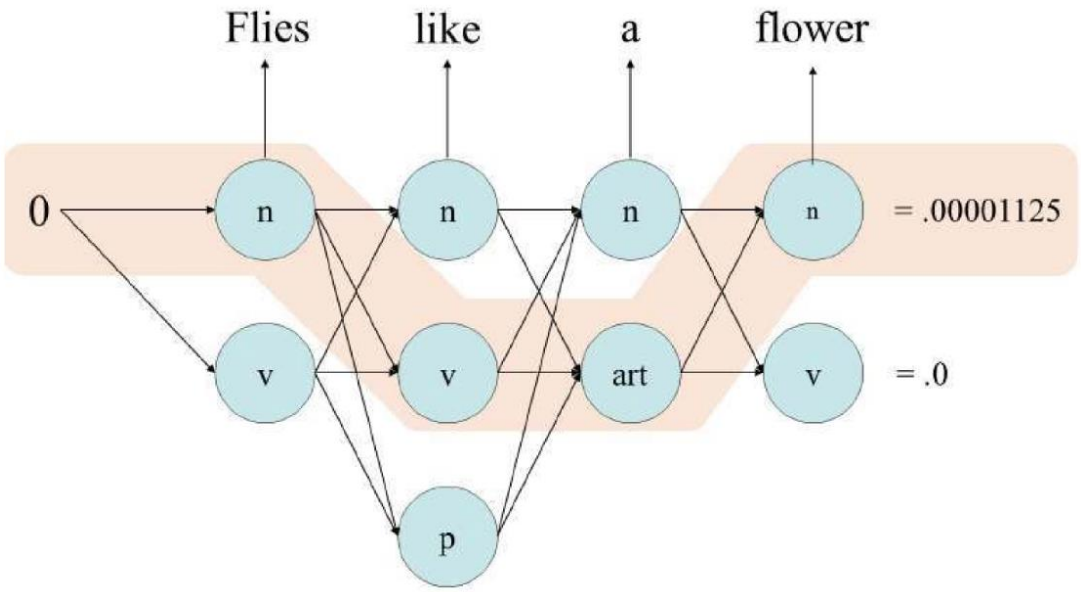
2-2) hidden markov model(HMM)

주어진 문장에서 형태소의 품사 태그 정보를 숨긴 채로 확률 정보를 이용하여 가장 가능성이 높은 경로를 찾음



품사 태깅을 위한 간단한 HMM 방법의 예

"Files like a flower" 예시 문장 분석



각 단계마다 확률 높은 것이 아닌, 총 확률이 제일 높은 path 고름

*요즘은 HMM을 잘 안씀

1. 딥러닝에서 attention방법을 현재는 주로 사용
2. 한국어에서는 잘 맞지 않음
3. 이해하기 어려운 개념임

3) 딥러닝 기반의 접근법(요즘 대세)

* 언어처리에 있어서 딥러닝의 효과

    1. 데이터로부터 특징을 자동으로 학습
    2. 폭넓은 문맥정보를 다룰 수 있음
    3. 모델에 적합한 출력을 다루기 간단함
    4. 언어가 아닌 이미지나 음성과 같은 모델들 간의 상호작용 가능, multi-modal 모델 구축 용이