

### ch3 언어학의 기본 원리

#### why study?

자연어에 대한 원리에 대해 이해해야 자연어처리를 학습하고 연구를 수행할 수 있다.

#### #1 언어의 단위

##### -음절

언어를 말하고 들을 때, 하나의 덩어리로 여겨지는 가장 작은 발화의 단위. 말소리의 단위. 한 국어에서 음절은 초성(자음(C)), 중성(모음(V)), 종성(자음(C))으로 구성.

음절의 형태는 V, C+V, V+C, C+V+C로 있다. 각각의 예시로는 아, 약, 다, 말이 있다.

\*음절은 엄밀히 말해서 말소리의 단위이므로, 발음이 나는 대로 처리 해야 하지만, 자연어처리에서는 편의를 위해 문법에 맞춰 작성된 것을 사용한다.

##### -형태소

의미를 가지는 가장 작은 단위. ==> 형태소를 쪼개면 더 이상 기능이나 의미를 갖지 않는다.

실질 형태소: 실질적인 의미를 갖는 형태소, 명사, 동사, 형용사, 부사가 해당

형식 형태소: 문법적인 기능을 하는 형태소, 조사, 어미가 해당

자립 형태소: 문장에서 홀로 쓰일 수 있는 형태소. 명사, 대명사, 수사, 관형사, 부사, 감탄사 등

의존 형태소: 다른 형태소와 결합되어 사용한다. 조사, 어미, 형용사, 동사의 어간이 있다.

유일 형태소, 불구형태소, 특이형태소 : 적은 수의 형태소와만 결합하는 것. (예: 오솔길의 오솔, 가랑비의 가랑)

형태소의 이형태: 사용 환경에 따라 형태소의 모양이 달라지는 것.(예: 은/는, 이/가) 이런 이형태 중 가장 본래의 형태에 가까운 것을 기본형이라 한다.

##### -어절

한 개 이상의 형태소가 모여 구성된 단위. 글에서 띄어쓰기 단위와 거의 일치.

##### -품사

한국어에서는 5언(체언, 수식언, 관계언, 독립언, 용언) 9품사(명사, 대명사, 수사, 관형사, 부사, 조사, 감탄사, 동사, 형용사)가 있다.

가변어: 단어의 형태가 변하는 것, 용언이 해당

불변어: 형태의 변화 없이 쓰이는 것, 체언, 수식언, 관계언, 독립언.

체언: 문장에서 몸통, 중심이 되는 역할, 명사, 대명사, 수사

수식언: 꾸며주는 단어. 관형사, 부사가 속함.

--관형사: 체언 앞에 놓여 체언을 꾸며주는 역할. 성질이나 상태를 꾸며주는 성상 관형사, 어떤 대상을 가리켜 지시하는 지시 관형사, 사물의 양이나 수를 나타내는 수관형사가 있다.

활용형이 아니기 때문에 어간이나 어미로 나뉘지 않고, 조사가 붙지 않으며 시제도 없다. 반드시 체언 앞에 붙어 그 체언만을 꾸민다.

--부사: 용언 앞에서 그 내용을 꾸며주거나 문장 전체를 꾸민다. 서상 부사는 '매우'와 같이 정도를 나타내며, 지시 부사는 '내일', '이리'와 같이 시간과 처소 또는 특정한 대상을 가리킨다. 부정 부사는 '안', '못'처럼 용언의 뜻을 부정한다. 양태 부사는 말하는 이의 태도를 표현, '반드시', '아마'가 있다. 접속 부사는 단어와 단어를 이어주거나. 문장과 문장을 이어준다.

관계언: 자립형태소에 붙어 문법적 관계를 나타내는 의존형태소이다. 조사가 이에 해당.

--조사: 체언 또는 용언의 명사형 등의 뒤에 붙어 말의 뜻을 더해준다. 격조사, 접속 조사, 보조사가 있다.

독립언: 독립적으로 쓰이는 품사이다. 감탄사가 여기에 속한다. 감정 감탄사와 의지 감탄사, 호응 감탄사로 이루어져 있다.

용언: 독립된 뜻을 가지고 어미를 활용하여 문장 성분으로서 서술어의 기능을 하는 말이다. 동사, 형용사가 해당한다. 어간과 어미로 이루어져 있다. 이런 어간에 여러 어미가 붙어서 형태가 바뀌는 것을 활용이라 한다.

#2 언어의 구조-언어를 이해 할 때 두 가지 뜻으로 이해 할 수 있는 문장을 구분하기 위해서 필요함.

#### 구구조

문장을 구성하고 있는 요소들이 서로 끌어당겨 한 덩어리가 됨으로써 형성되는 일정한 구조.

문장을 구성하는 2개이상의 구성요소는 서로 선행관계, 지배관계를 맺고 있다.

한국어는 문장 안에서 단어의 배열이 비교적 자유로우므로 구구조는 영어와 같이 단어의 배열이 정해져있는 언어를 분석 할 때 많이 쓰인다. 보통 영어 문장은 크게 명사구와 그 뒤의 동사구로 이루어져있다. 이를 세분화 시켜 어떤 요소가 어떤 요소의 앞, 뒤에 위치하는지. 어떤 요소의 지배를 받는지 표현할 수 있다.

수형도나 괄호 표시로 나타낸다. 예를들어 The puppy sleeps under the table이라는 문장(S)에서 명사구(NP)[The puppy]와 동사구(VP)[sleeps under the table]로 나뉜다. NP는 the(Article)와 puppy(Noun)으로 구성, VP는 sleeps(Verb)와 전치사구(PP)로 구성, PP는 under(P)와 명사구(NP)로 다시 NP는 the와 table로 구성되어있다. 이렇게 구구조로 분석하면, 두 가지 의미를 가지는 문장을 선행관계와 지배관계를 이용해 서로 다른 구조로 구분할 수 있다.

#### 의존구조

문장에 포함된 단어들이 서로 의존 관계를 이루어 하나의 구문을 이룬다. 문장에는 각 단어의 지배소와 의존소로 이루어져 있고 그 의존 관계를 분석하는 것을 의존 구문 분석이라고 한다. 한국어는 비교적 어순이 자유로워 구문 분석 시 의존 구문 분석이 주로 이용된다.

이도 화살 표를 통해 의존관계를 표현함으로써 여러 가지 의미로 해석될 수 있는 문장을 구분해준다.

그래프 기반 의존 구문 분석: 비결정적인 방법으로 어떠한 문장이 가질 수 있는 모든 의존 분석 결과 중에서 가장 높은 점수의 의존 트리를 선택하는 방식. 즉, 가능한 모든 경우의 수를 그리고 그 중 의존구조에 대한 점수가 가장 높은 것을 선택한다. 모든 경우의 수를 계산하기 때문에 속도가 느리다.

전이 기반 의존 구문 분석: 선형적 탐색을 해 근거리 의존 관계를 찾는데 강하고 속도가 빠르지만, 문장의 의존구조를 전역적으로 확인하는 데는 약하다.

#### #3 의미론과 화용론

##### 의미론

단어, 문장, 발화에서 그 표현이 실제로 가리키는 지시체와의 의미 관계를 파악하는 것. 문법적으로 옳은 문장이라도 의미가 어색==> 언어 취급을 안 함

예: 사료가 개를 먹었습니다. 문법은 맞지만, 의미가 말이 되지 않다. 이를 방지하기 위해 사용하는 개념이 의미역(Thematic relations, Semantic roles)이다. 어떠한 행동을 의도를 가지고 수행하는 'Agent', 행동에 영향을 받는 'Patient', 행동이 일어나도록 하는데 사용되는 'Instrument', 어떠한 행동이 일어나는 장소인 'Location' 등 다양한 의미역이 단어의 의미를 파악하는데 사용

의외에도 어휘적 관계(동의 관계, 반의 관계, 상하 관계, 원형, 동음이의어, 다의어 등)로도 단어의 의미를 파악할 수 있다.(ch1의 단어 임베딩처럼)

#### 화용론

언어 사용자와 발화 맥락을 고려하는 연구. 언어 자체가 아닌 주변을 설명, 청자, 화자, 시간, 장소 등을 고려한다.

이를 위해서는 상대방이 가리켜 표현하는 것이 무엇인지 알아야한다. 즉, 상황을 직시해야하는데, 가장 중요한 것이 화자가 말하는 '저것', '이것'이 무엇인지 아는 것이다. 또는 대화 참여자들 사이에서 이미 보유하고 있는 정보가 있다는 것을 가정한 채 대화가 이루어질 수도 있어 이를 고려해야 한다.

화행: 언어를 통해서 약속, 사과, 충고 선언 등의 행위가 이루어지는 것 ex) 그럼 저희는 다음 주 이 시간에 다시 봅시다.에서 다시 만나자는 약속이라는 행위가 포함되어 있다.

직접 화행: 화행을 직접적으로 표시하는 것 ex)너 수영 할 줄 알아? 라는 발화를 수영을 할 줄 아는지 여부를 묻기 위해 질문으로 쓰는 것

간접 화행: 간접적으로 표현하는 것 ex) 바로 위의 예시를 친구에게 바다에 가자고 제안하기 위한 질문으로 쓰는 것

이들 모두 언어의 함축적인 의미를 알기 위해 사용된다.