자연어 처리

위키백과, 우리 모두의 백과사전.

자연어 처리(自然語處理) 또는 자연 언어 처리(自然言語處理)는 인간의 언어 현상을 컴퓨터와₩ 같은 기계를 이용해서 묘사할 수 있도록 연구하고 이를 구현하는 인공지능의 주요 분야 중 하나다. 자연 언어 처리는 연구 대상이 언어 이기 때문에 당연하게도 언어 자체를 연구하는 언어학과 언어 현상의 내적 기재를 탐구하는 언어 인지 과학과 연관이 깊다. 구현을 위해 수학적 통계적 도구를 많이 활용하며 특히 기계학습 도구를 많이 사용하는 대표적인 분야이다. 정보검색, QA 시스템, 문서 자동 분류, 신문기사 클러스터링, 대화형 Agent 등 다양한 응용이 이루어지고 있다.

형태소 분석

자연 언어 처리에서 말하는 형태소 분석이란 어떤 대상 어절을 최소의 의미 단위인 '형태소'로 분석하는 것을 의미한다. (형태소는 단어 그 자체가 될 수도 있고, 일반적으로는 단어보다 작은 단위이다.) 정보 검색 엔진에서 한국어의 색인어 추출에 많이 사용한다. 형태소 분석 단계에서 문제가 되는 부분은 미등록어, 오탈자, 띄어쓰기 오류 등에 의한 형태소 분석의 오류, 중의성이나 신조어 처리 등이 있는데, 이들은 형태소 분석에 치명적인 약점이라 할 수 있다. 복합 명사 분해도 형태소 분석의 어려운 문제 중 하나이다. 복합 명사란 하나 이상의 단어가 합쳐서 새로운 의미를 생성해 낸 단어로 '봄바람' 정보검색' '종합정보시스템' 등을 그 예로 들 수 있다. 이러한 단어는 한국어에서 띄어쓰기에 따른 형식도 불분명할 뿐만 아니라 다양한 복합 유형 등에 따라 의미의 통합이나 분해가 다양한 양상을 보이기 때문에 이들 형태소를 분석하는 것은 매우 어려운 문제이다. 기계적으로 복합명사를 처리하는 방식 중의 하나는, 음절 단위를 기반으로 하는 bi-gram 이 있다. 예를 들어, '복합 명사'는 음절 단위로 '복합+명사', '복+합명사', '복합명+사' 의 세 가지 형태로 쪼갤 수 있고, 이 중 가장 적합한 분해 결과를 문서 내에서 출현하는 빈도 등의 추가 정보를 통해 선택하는 알고리즘이 있을 수 있다. 일반적으로, 다양하게 쪼개지는 분석 결과들 중에서 적합한 결과를 선택하기 위해, 테이블 파싱이라는 동적 프로그래밍 방법을 사용한다.