13

엑소브레인 한국어 구어체 형태소 분석 및 개체명 인식 기술 v1.0

기술개요

› 한국어 구어체 텍스트에 대해서 형태소분석, 개체명인식을 수행하고, 언어분석된 결과를 전달해 주는 기술

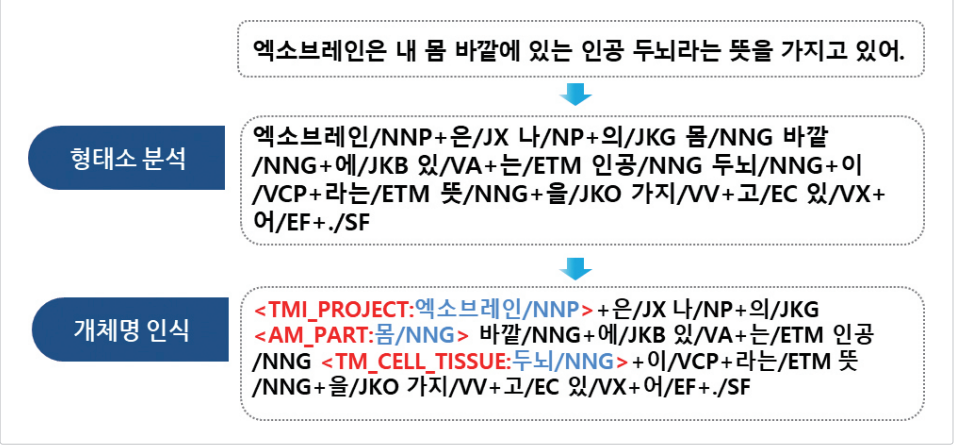
기술특성

› 구어체 형태소 분석 기술

- 문장이 입력되면 형태소를 찾고 형태소의 품사(세종 태그셋, 45개)를 자동으로 분석
- 규칙기반 방법과 기계학습 방법을 하이브리드한 형태
- 전처리/후처리 단계에 기분석 사전과 같은 대용량 형태소 사전을 활용하고, Smith-Waterman 알고리즘을 적용하여 구어체 텍스트에 적합한 원형복원 진행
- 음절 단위 품사태깅을 위한 sequence labeling이 가능한 기계학습 알고리즘(Structural SVMs) 사용

› 구어체 개체명 기술

- 형태소 분석된 문장이 입력되면 개체명을 인식하고 개체명 태그(146개)를 자동으로 분석
- 규칙기반 방법과 기계학습 방법을 하이브리드한 형태
- 전처리/후처리 단계에서 도메인에 적합한 개체명 사전 및 패턴을 적용 가능
- 형태소 단위로 sequence labeling이 가능한 기계학습 알고리즘(Structural SVMs) 사용
- 구어체가 가지는 특징인 축약어에 대한 처리를 효과적으로 할 수 있게 기계학습 및 전/후처리 개선



적용분야

- 정보검색, 질의응답, 정보추출

기술완성도 (TRL)

- 6단계 : 파일롯 규모 시작품 제작 및 성능 평가

기술이전 내용

- 음절단위 한국어 형태소분석 기술
- 세부분류 개체명인식 기술

지식재산권 현황

No.	출원·등록번호	특허명	상태
1	2413174	간접광고를 포함한 뉴스 기사 생성 시스템 및 방법	등록
2	2020-0179810	근거인식 기반 질의응답 시스템 및 방법	출원

기술이전 문의

- ETRI 연구성과확산실 | 042-860-4881 / etri_tco@etri.re.kr