12

엑소브레인 일반 문서 기반 심층질의응답 기술 v1.0

기술개요

› 정해진 서식이 없는 일반 문서(매뉴얼, 지침, 규정 등)를 대상으로 단락 뿐만 아니라 테이블 정보에서도 검색 및 기계독해를 통해 정답 및 답을 설명할 수 있는 근거 정보를 제공하는 질의응답 기술

기술특성

› 딥러닝 언어모델

- 한국어 텍스트의 문맥(구문/의미)을 학습하여, 다양한 응용 태스크(언어분석/기계독해/문서분류 등)에 범용적으로 활용 가능한 딥러닝을 위한 언어모델
- 위키백과 및 신문기사 23.5 GB(약 15년 분량), 47억개 형태소 학습, 법률분야 텍스트 186MB 학습
- 한국어의 특성을 반영한 형태소 단위의 BPE 적용

› 자연어 질문분석 및 FAQ

- SVM 기반 기계학습 모델과 규칙 사전을 이용한 정답유형 인식 및 질문 분류
- 딥러닝 모델과 Lexico-Semantic 기반의 문장 유사도 계산

› 일반 문서 색인 및 검색

- 단락 형태로 변환된 JSON 형식의 가상문서로부터 단락과 테이블 정보를 색인하고 검색
- JSON 형식의 가상문서로부터 단락/테이블/FAQ 양식/FAQ 질문을 색인
- 형태소 및 문서 타입 기반의 심볼릭 검색 및 순위화

› 딥러닝 언어모델 기반 재순위화

- 딥러닝 언어모델 기반으로 일반 문서 대상으로 검색된 단락들을 정답이 포함될 확률이 높은 순위로 재순위화
- 정답이 포함되는 확률이 높은 순위로 검색 결과를 재순위화
- 약 209만개의 정답-오답 질문-단락 학습데이터를 이용한 fine-tunning

› 딥러닝 언어모델 기반 근거인식

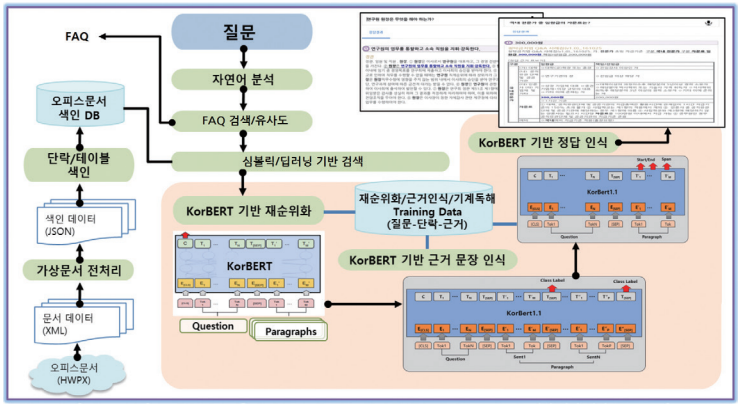
- 딥러닝 언어모델 기반으로 재순위화된 검색 단락에서 정답을 추론할 수 있는 근거 문장을 추론
- 재순위화 된 검색 단락에서 정답을 추론할 수 있는 근거가 되는 문장을 인식
- 약 8만1천개의 질문-단락-근거 학습데이터를 이용한 fine-tunning

› 딥러닝 언어모델 기반 기계 독해 모델

- 딥러닝 언어모델 기반으로 재순위화된 검색 단락과 근거인식 문장으로부터 각각 정답을 인식하여 하이브리드 기반으로 최종 정답을 추론
- 재순위화된 검색 단락으로부터 정답 경계 인식
- 근거인식된 문장을 결합한 단락으로부터 정답 경계 인식
- 하이브리드 기반의 최종 정답 경계 인식
- 약 8만1천개의 질문-단락-근거 학습데이터를 이용한 fine-tunning

› 분산처리 플랫폼

- 대용량 텍스트 대상 언어분석을 배치로 수행하여 색인하고, 심층질의응답 서버를 운용하기 위한 플랫폼
- 배치형 한국어 분석 기반 색인 및 시맨틱 검색
- 서버 및 쓰레드 풀 확장이 가능한 심층질의응답 시스템



일반문서 기반 심층 질의응답 구성도

적용분야

- 일반지식/전문분야 QA 시스템
- VOC(Voice of Customers) 분석 시스템
- 전문가 의사결정지원 시스템

기술완성도 (TRL)

- 6단계 : 파일롯 규모 시작품 제작 및 성능 평가

기술이전 내용

- 딥러닝 언어모델
- 자연어 질문분석 및 FAQ
- 일반 문서 전처리 및 색인/검색
- 딥러닝 언어모델 기반 재순위화/근거 인식/기계 독해
- 분산처리 플랫폼

지식재산권 현황

No.	출원·등록번호	특허명	상태
1	2020-0129497	기계 독해 학습 데이터 자동 생성 장치 및 그 방법	출원
2	2020-0130556	문장 의미 유사도 판단 방법 및 장치	출원
3	2020-011126	적대적 패러프레이즈 문장 자동 생성 시스템	출원
4	2020-0179810	근거인식 기반 질의응답 시스템 및 방법	출원

기술이전 문의

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