



## 2020학년도 산학연계 SW프로젝트 수행계획서

신청자 인적 사항	학부(과)	컴퓨터정보공학부	성명	이원재
	학년	3학년	학번	2018202059
	연락처 (휴대번호)	010-4575-5160	E - mail	willy99624@naver.com
	공학인증프로그램 참여 (O)	전문 ( O )      일반 (   )      보류 (   )		
설계 내용	지도교수	이혁준 교수님		
	기업명	한화시스템		
	희망하는 주제	인공지능 자연어 처리 기반 개체명 인식기 고도화 기술 개발		
	팀명	Viva Pro.		
	팀원	5 명 이원재 (컴퓨터정보공학부, 학번 2018202059) 조우진 (컴퓨터정보공학부, 학번 2015722057) 송현우 (컴퓨터정보공학부, 학번 2016802026) 신규표 (컴퓨터정보공학부, 학번 2018202058) 손승현 (컴퓨터정보공학부, 학번 2018202064)		

교육 신청관련 개인정보 수집, 활용 및 제공에 대한 동의		동의여부
수집하는 개인정보 항목	• 소속, 성명, 학년, 학번, 연락처, E-mail	■
개인정보의 수집 및 이용목적	• “산학연계 SW프로젝트” 안내를 위함.	■
개인정보의 보유 및 이용기간	- 수집한 고객의 개인정보를 수요조사 기간까지만 보유하며 수요조사 완료 후 관련법규에 의거하여 안전하게 파기합니다. (개인정보보호법 시행령 제 16조) - 정보제공자가 개인정보 수집, 이용에 대한 동의를 철회할 경우 수집한 개인정보를 즉시 파기합니다.	■
개인정보 제공 동의 거부 권리 및 동의 거부 따른 불이익 내용 또는 제한사항	귀하는 개인정보 제공 동의를 거부할 권리가 있으며, 동의 거부에 따른 불이익은 없음. 다만, 추가적인 교육 서비스를 받을 수 없음.	■

※ 개인정보 제공자가 동의한 내용외의 다른 목적으로 활용하지 않으며, 제공된 개인정보의 이용을 거부하고자 할 때에는 개인정보 관리책임자를 통해 삭제를 요청 할 수 있음.

「개인정보보호법」등 관련 법규에 의거하여 상기 본인은 위와 같이 개인정보 수집 및 활용에 동의함.

2020년 7월 8일  
 성명 : 이원재 (인)

본인은 『산학연계 SW프로젝트』에 참가신청을 희망합니다.

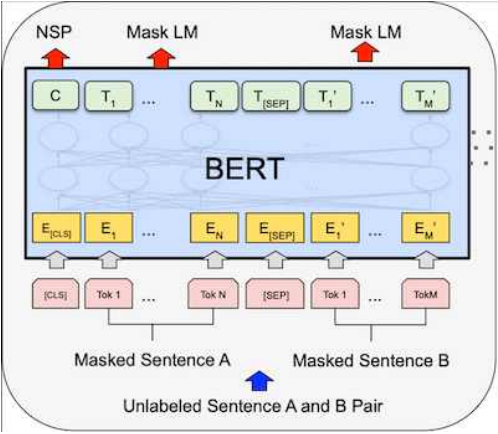
2020년 7월 8일  
 신청인 성명 이원재 (인)  
 지도교수 성명 이혁준 (인)

광운대학교 SW중심대학사업단장 귀하



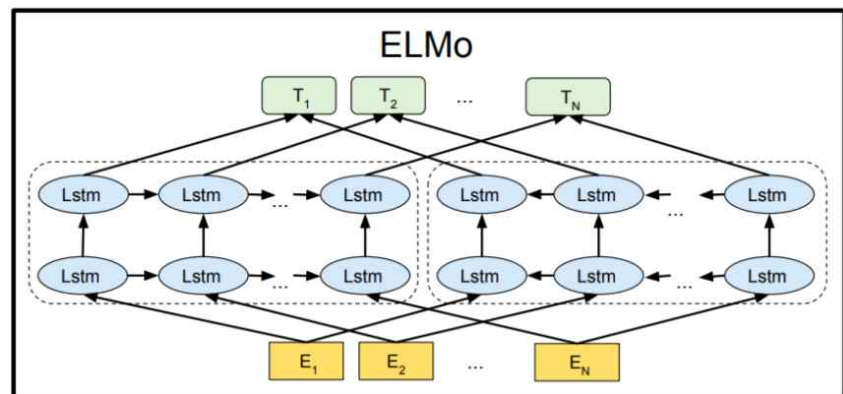
# 산학연계 SW프로젝트 수행계획서

<p>팀명/팀원</p>	<p>Viva Pro.</p> <p>이원재 (컴퓨터정보공학부, 학번 2018202059) 조우진 (컴퓨터정보공학부, 학번 2015722057) 송현우 (컴퓨터정보공학부, 학번 2016802026) 신규표 (컴퓨터정보공학부, 학번 2018202058) 손승현 (컴퓨터정보공학부, 학번 2018202064)</p>
<p>작품(과제)명</p>	<p>인공지능 자연어 처리 기반 개체명 인식기 고도화 기술 개발</p>
<p>1. 개발동기 및 필요성</p>	<div> <p>● 개발동기</p> <ul style="list-style-type: none"> <li>• 개체명 인식이 세계적으로 비약적인 발전중이지만, 한국어 개체명 인식은 그에 비해 많이 부진</li> <li>• 정보 검색에서 개체명은 주요 검색 대상</li> <li>• 질의 응답에서 개체명은 주요 질의 응답 대상</li> <li>• 개체명 인식 기술의 고도화로 인한 인간 생활의 편의성 증대</li> </ul> <p>● 문제점</p> <ul style="list-style-type: none"> <li>• 번역, 대화형 시스템 등의 부정확성</li> <li>• 공식 문서 검색 시의 내용 불일치 및 불필요한 결과 동시 출력</li> </ul> <p>● 필요성</p> <ul style="list-style-type: none"> <li>• 문장 내 개체명 인식률 상승 및 카테고리 세분화</li> <li>• 개체명 인식을 통한 정보의 가독성 증가</li> <li>• 개발 과정 중 태깅을 통한 문서 정리 용이</li> </ul> <p>● 국내 지적재산권 현황</p> <div> <div> <input type="checkbox"/> <div> <div>등록</div> <div>[5] 딥 러닝을 이용한 한국어 개체명 인식 장치 및 방법(Apparatus and method for recognizing Korean named entity using deep-learning)</div> </div> <div> <div>공보</div> </div> </div> <div> <div> </div> <div> <p>IPC : G06F 40/20 G06N 3/08</p> <p>출원번호 : 1020170165072</p> <p>등록번호 : 1020433530000</p> <p>공개번호 : 1020190065665</p> <p>대리인 : 특허법인충현</p> </div> <div> <p>출원인 : 주식회사 솔루게이트</p> <p>출원일자 : 2017.12.04</p> <p>등록일자 : 2019.11.05</p> <p>공개일자 : 2019.06.12</p> <p>발명자 : 양기주</p> </div> </div> </div> <p>(57) 요약</p> <p>본 발명은 딥 러닝을 이용한 개체명 인식 기술에 관한 것으로, 한국어 개체명 인식 방법은, 입력된 문장에 대한 한글의 자소를 기반으로 형태소를 분석하여 각각의 형태소에 대응하는 품사 태그(tag)를 매칭하고, 자소 기반의 형태소 및 품사 태그에 대하여 단어 벡터(word vector) 및 품사 태그 벡터(POS-tag vector)를 생성하고, 생성된 단어 벡터 및 품사 태그 벡터를 양방향 LSTM(bidirectional long short-term memory)에 입력하여 각각 단어 문맥 벡터 및 태그 문맥 벡터를 생성하며, 생성된 단어 문맥 벡터 및 태그 문맥 벡터를 결합(concatenate)하여 혼련된 문맥 벡터를 생성한다.</p> </div>

2. 과제 최종목표 및 개발내용	<p>● 과제 최종목표</p> <ul style="list-style-type: none"> <li>문장 내 단어를 정확하게 분석할 수 있도록 태그 세분화</li> <li>자연어를 이전 모델보다 더 높은 정확성으로 깔끔하게 처리할 수 있도록 모델 개발</li> </ul> <p>● 개발내용</p> <ol style="list-style-type: none"> <li>엑소브레인 말뭉치 및 기타 말뭉치 등의 데이터 확보</li> <li>한글 형태소 분석기로 Tokenizing 진행</li> <li>(3-1) 형태소별 분류 → 단어 임베딩 → 추가 분류 및 트레이닝</li> <li>(3-2) BERT를 사용하여 분류, 임베딩, 태깅 등의 작업을 한번에 효율적으로 진행</li> <li>여러 모델별 결과 확인 후, 효율성 높은 모델로 트레이닝 추가 진행</li> <li>개체명 인식 성능 평가</li> <li>개발 후에는 목적 지향 대화 시스템에 사용 가능</li> </ol>
3. 과제해결 방법	<p>● 인공지능 자연어 처리 기반 개체명 인식기를 고도화시키는 과정 (프로젝트 진행과정)</p> <ol style="list-style-type: none"> <li>개체명 인식을 하기 위해 Server에서 Client로부터 입력을 받으면 MeCab, KOMORAN, kahii등을 통해 문장을 tokenizing (주로 형태소 분석을 활용할 예정)</li> <li>1단계에서 tokenizing을 마친 결과는 Modeling 단계로 넘어감. 이 단계에서 사전 학습을 하기 위해서는 Pre-Trained Model이 필요하기 때문에 BERT, KoBERT, OpenAI GPT, ELMo 등의 모델을 사용</li> </ol> <p>1) BERT</p> <ul style="list-style-type: none"> <li>BERT는 언어표현 사전학습의 새로운 방법으로 그 의미는 '큰 텍스트 코퍼스'를 이용하여 범용목적의 '언어 이해(language understanding)' 모델을 훈련시키는 것과 그 모델에 관심 있는 실제의 자연 언어 처리 태스크(질문-응답 등)에 적용하는 것</li> <li>Encoder가 입력 문장들을 임베딩 하여 언어 모델링을 하면 이를 Fine-Tuning 하여 여러 자연어 처리 Task를 수행</li> </ul>  <p>The diagram illustrates the BERT architecture. At the bottom, an 'Unlabeled Sentence A and B Pair' is shown, consisting of 'Masked Sentence A' and 'Masked Sentence B'. These sentences are tokenized into '[CLS]', 'Tok 1', ..., 'Tok N', '[SEP]', 'Tok 1', ..., 'Tok M'. These tokens are then passed through an 'Encoder' (represented by yellow boxes labeled E<sub>[CLS]</sub>, E<sub>1</sub>, ..., E<sub>N</sub>, E<sub>[SEP]</sub>, E<sub>1'</sub>, ..., E<sub>M'</sub>) to produce 'BERT' embeddings (represented by blue boxes labeled T<sub>1</sub>, ..., T<sub>N</sub>, T<sub>[SEP]</sub>, T<sub>1'</sub>, ..., T<sub>M'</sub>). The BERT model is used for three tasks: 'NSP' (Next Sentence Prediction) on the [CLS] token, and 'Mask LM' (Masked Language Modeling) on the masked tokens (Tok N and Tok M).</p>

## 2) ELMo

- ELMo는 Embeddings from Language Model의 약자
- ELMo는 기학습된 언어 모델을 이용해 어휘 임베딩을 생성하는 방법
- 사전 훈련된 언어 모델(Pre-trained language model)을 사용한다는 점이 가장 큰 특징
- 동일한 단어가 문맥에 따라 다른 Vector로 표현될 수 있게 워드 임베딩을 하는 것이 특징



(3) 각 카테고리의 Fine-Tuning을 진행. Data Set(주로 엑소브레인 프로젝트의 말뭉치들을 활용할 예정)을 이용해 Tag(Category)를 다음과 같이 구분하고 마스킹 과정을 거침

- PS(PERSON) : 사람
- OG(ORGANIZATION) : 조직
- LC(LOCATION) : 위치
- DT(DATE) : 날짜
- TI(TIME) : 시간
- FD(FIELD) : 학술분야
- TR(THEORY) : 이론
- AF(ARTIFACT) : 인공물
- CV(CIVILIZATION) : 문명/문화
- EV(EVENT) : 사건
- AM(ANIMAL) : 동물
- PT(PLANT) : 식물
- O(OTHER) : 기타

(4) 3단계에서 여러 케이스로 나누어 진행한 후, 성능 측정(Perplexity, GPU Computing) 및 비교(Matplotlib)하여 Evaluation

(5) 4단계에서 도출한 결과값을 상용 클라우드와 AMP(Apache, Mysql, PHP)를 이용해 시각화

● 기존 KWBERT 연구와 다른 점 (개선점)

(1) 기존 KWBERT는 수많은 한국어들 중 person, organization, location, date, time 등만을 분류해내고 그 외의 단어들은 모두 'O' 항목으로 분류

→ 해결방법

- 개발한 모델이 더 많은 한국어(단어)들을 분류가 가능하도록 개선시키도록 하는 것에 초점을 맞추어 진행
- 이를 위해 여러 가지 BERT를 활용하여 엑소브레인 말뭉치 외의 여러 가지 말뭉치들을 분석하여 벡터 값으로 나타냄
- 결과값에 따라 작년의 'O' 에 해당하는 태그를 더 세분화시켜서 ETRI 대 분류 기준에 따라 단어를 추가한 후 분류를 하는 기술을 만들

(2) 카테고리를 추가하는 기술 말고도 기존의 연구에서 지적된 오류들을 해결하는 방향으로 나아감으로서 성능을 개선할 것이고, 기존의 연구에서 지적된 주요 오류들은 다음과 같음

※ 오류1) 개체명의 경계인식 (띄어쓰기)

- 특정한 개체명이 두 개 이상의 단어로 이루어진 경우 '단어 수 = 개체 수'로 인식하여 하나의 의미를 가지는 단어가 여러 단어로 다루어지는 경우 발생

• 예)

20세기 스튜디오 VS 20세기 스튜디오

→ 해결방법

- BIO 태그 사용
  - BIO 태그는 개체명의 시작에 해당하는 B, 개체명의 중간 또는 끝에 해당하는 I, 비개체명에 해당하는 O로 구성
  - 예시)
 

20세기 OG\_B  
스튜디오 OG\_I

※ 오류2) 다중 카테고리에 포함되는 단어의 중복 처리

• 예)

★★★  
**배**<sup>1</sup>

명사

1. 생명 사람이나 동물의 몸에서 위장, 창자, 콩팥 따위의 내장이 들어 있는 곳으로 가슴과 엉덩이 사이의 부위.

☐ 배가 나오다.

2. 동물 절족동물, 특히 곤충에서 머리와 가슴이 아닌 부분. 여러 마디로 되어 있으며 숨구멍, 항문 따위가 있다.

→ 해결방법

- 동일한 단어를 문맥에 따라 임베딩 단계에서 처리하는 기술을 활용

4. 추진체계 (역할 분담, 추진 일정 등)	<div>● 역할 분담</div> <ul style="list-style-type: none"><li>말뭉치자료확보 : 이원재, 조우진, 송현우, 신규표, 손승현</li><li>버트를 통한 워드 임베딩 : 이원재, 송현우, 신규표, 손승현</li><li>웹(PHP) : 이원재, 신규표, 송현우</li><li>토크나이징 및 태깅 : 이원재, 조우진, 손승현</li><li>신경망구축 : 이원재, 조우진, 송현우</li></ul> <div>● 추진 일정 (과제 개발 계획)</div> <table><tr><th></th><th>7월</th><th>8월</th><th>9월</th><th>10월</th><th>11월</th><th>12월</th><th>1월</th><th>2월</th><th>3월</th><th>4월</th><th>5월</th></tr><tr><td>프로젝트 제안 및 수행 계획</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>말뭉치 자료 확보</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>전처리 과정 및 정제 과정</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>기존 프로젝트 및 관련기술 학습</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>모델 구현 및 훈련</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>프로젝트 시각화</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>테스트 데이터를 통한 평가</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>최종 보완 및 피드백</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>프로젝트 최종 발표</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr></table>		7월	8월	9월	10월	11월	12월	1월	2월	3월	4월	5월	프로젝트 제안 및 수행 계획												말뭉치 자료 확보												전처리 과정 및 정제 과정												기존 프로젝트 및 관련기술 학습												모델 구현 및 훈련												프로젝트 시각화												테스트 데이터를 통한 평가												최종 보완 및 피드백												프로젝트 최종 발표											
	7월	8월	9월	10월	11월	12월	1월	2월	3월	4월	5월																																																																																																														
프로젝트 제안 및 수행 계획																																																																																																																									
말뭉치 자료 확보																																																																																																																									
전처리 과정 및 정제 과정																																																																																																																									
기존 프로젝트 및 관련기술 학습																																																																																																																									
모델 구현 및 훈련																																																																																																																									
프로젝트 시각화																																																																																																																									
테스트 데이터를 통한 평가																																																																																																																									
최종 보완 및 피드백																																																																																																																									
프로젝트 최종 발표																																																																																																																									
5. 기대효과	<div>● 기대효과</div> <ul style="list-style-type: none"><li>기업 및 다양한 서비스 분야에서 활용<ul style="list-style-type: none"><li>서비스 관련 분야에서 다른 사람과의 소통 및 상호작용에 대해 필요한 서비스를 빠르게 제공 → 효율적으로 접근 가능</li></ul></li><li>다양한 데이터 처리 분야에서의 활용<ul style="list-style-type: none"><li>자연어 처리가 필요한 의사결정 시스템, 질의응답 시스템, 정보검색 및 다양한 분야에서 활용 가능</li></ul></li><li>다른 기술과의 융합을 통한 활용<ul style="list-style-type: none"><li>기존에 존재하는 다양한 기술과의 응용을 통해 다양하게 활용 가능</li><li>예) 음성 및 아날로그 매체 인식 기술과의 응용을 통해 데이터 정리를 효율적으로 다루는 것이 가능</li></ul></li></ul>																																																																																																																								
6. 기타																																																																																																																									

## 7. 참고문헌

- [1] [http://kiise.or.kr/e\\_journal/2018/5/JOK/pdf/04.pdf](http://kiise.or.kr/e_journal/2018/5/JOK/pdf/04.pdf)
- [2] [http://aiopen.etri.re.kr/service\\_dataset.php](http://aiopen.etri.re.kr/service_dataset.php)
- [3] <https://github.com/SKTBrain/KoBERT>
- [4] <https://paul-hyun.github.io/bert-01/>
- [5] <https://c11.kr/gh7x>
- [6] <http://www.aitimes.kr/news/articleView.html?idxno=13117>
- [7] <https://wikidocs.net/33930>
- [8] <https://brunch.co.kr/@learning/12>
- [9] 한국어 특질을 고려한 단어 벡터의 Bi-LSTM 기반 개체명 모델 적용