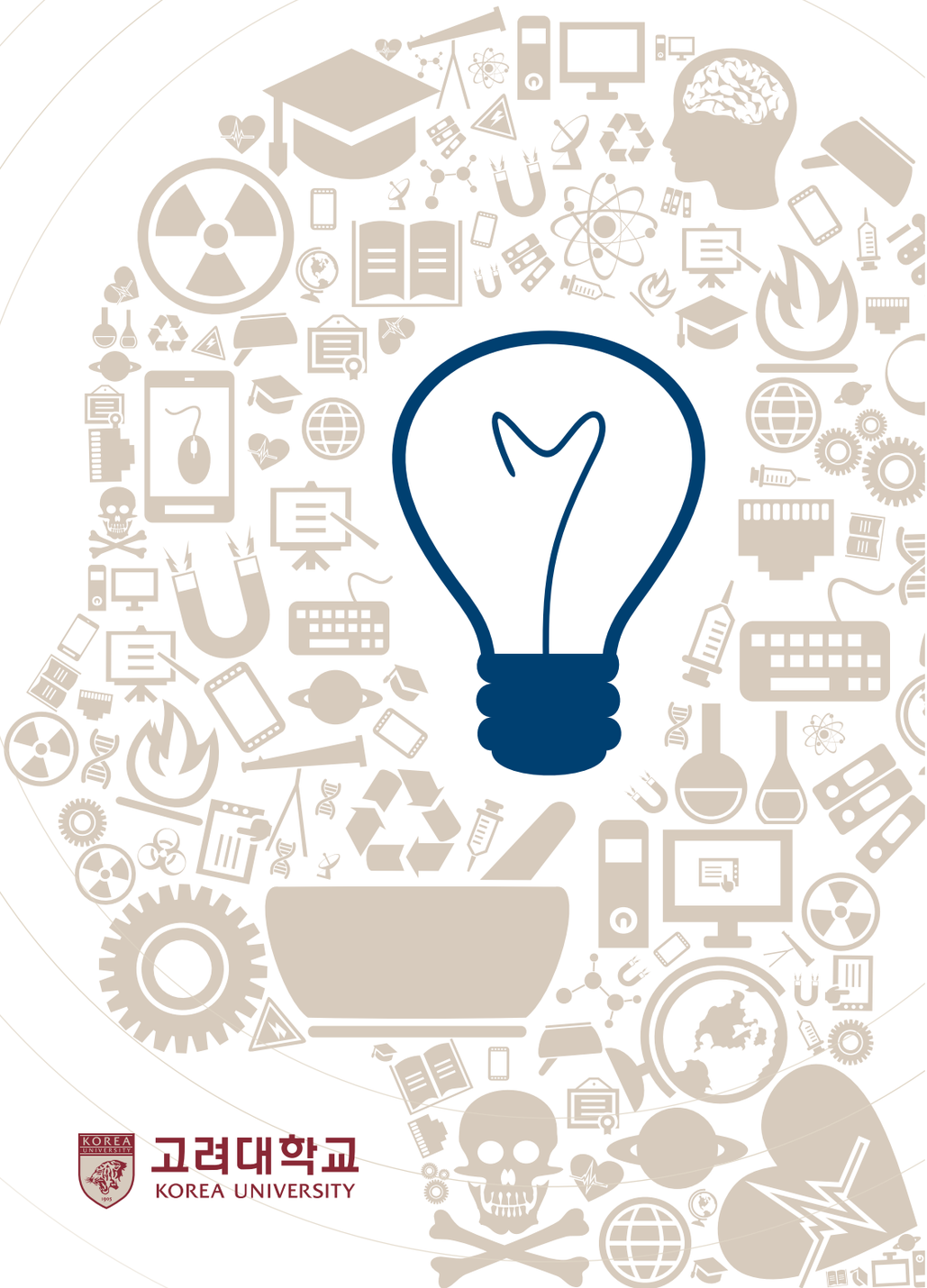
고려대학교

Human-Inspired AI & Computing 연구센터



고려대학교
KOREA UNIVERSITY

센터장 인사말

임희석 교수/센터장

현 시대는 스마트 디지털 시대라고 할 수 있습니다. 아날로그 시대에서 단순하게 디지털로의 전환이 아닌 스마트한 디지털 세상으로의 전환이 요구되고 있습니다. 모든 산업과 비지니스는 스마트라고 할 수 있는 인공지능 기술이 접목되어야 경쟁력을 가질 수 있으며 가치를 창출할 수 있습니다.

반면 스마트한 변화에 실패하는 어떤 국가나 산업도 과거의 번영을 지속할 수 없다고 예측됩니다. 모든 산업과 비지니스는 그들의 전통적인 결과물을 스마트라는 함수를 통하여 지능형 결과물을 만들 수 있어야 경쟁력을 가질 수 있습니다. 가치를 창출할 수 있는 스마트 함수를 만드는데 기여할 수 있는 인공지능 기술은 이제 모든 세계와 산업 현장에서 절실히 요구되는 핵심 성장 동력입니다.

최근 딥러닝 기술의 발전에 힘입어 인공지능 기술의 성능이 향상되었습니다. 하지만 사회는 인간 수준의 지능을 갖는 인공지능 기술을 요구하고 있으며, 그러한 요구를 충족시키기 위해서는 많은 연구와 노력이 필요합니다. 고려대학교 HI AI & Computing 센터는 이러한 요구에 부응하기 위하여 설립되었습니다. 가장 지능적인 인간의 뇌신경정보처리 원리와 인간 지능을 가능케하는 핵심 능력을 모델링하여 인간을 닮은 지능 기술을 개발하는 것이 본 센터의 핵심 방향이라 할 수 있습니다. 최근 인공지능 분야와 기계학습 분야에서 최고의 성능을 내고 있는 강화학습, 딥러닝, attention mechanism 등이 인간의 정보처리 원리를 반영한 기술들의 예라 할 수 있습니다.

본 센터에서는 강화학습과 딥러닝 모델처럼 사용하게 될 최고의 새로운 인공지능 기술을 개발하기 위하여 노력할 것입니다. 이를 통한 산업 발전, 국가의 경쟁력 강화, 그리고 인류의 행복한 삶에 기여할 수 있으리라 기대하며, 많은 분들의 성원과 응원을 부탁드립니다.

센 터 목 표

Human-inspired Machine Learning

- 인간 지능의 기본 요소를 반영한 기계학습 방법 연구
- 인간의 고차원적 인지 기능을 모방한 기계학습 방법 연구
- 인간 지능의 요소들을 융합한 멀티모달 기계학습 방법 연구
- 현실세계에 대한 지식을 바탕으로 한 능동적 기계학습 방법 연구

Human-inspired Rapid Learning

- 효율적인 학습을 위한 인간의 학습 원리를 반영한 AI 개발
- 데이터 부족 문제를 극복하기 위한 최적화 AI 기술 개발
- 학습 모델을 위한 데이터 구축 및 변환 기술 개발
- 실세계 적응 및 의사결정이 가능한 AI 기술 개발

멀티모달 기반의 지식 표현, 획득 및 추론 기술 융합

- 지식획득 및 지식정제기술 개발
- 지식 추론 및 변형 기술의 개발과 지식 생성을 위한 데이터셋 구축
- 지식 표현 방법의 개발 및 획득 · 추론 융합모델의 성능평가 및 검증
- Situation Recognition 및 이를 이용한 능동적 지식 추천 기술 개발

지능 정보 응용 서비스 개발 - A.I. Nurse

- 환자 및 병실 상태 정보의 관심영역 분할 기술 개발
- IOT 정보 기반 환자 및 병실 상태 정보 분석 기술 개발
- 환자 및 병실 상태의 이상 징후 예측 및 판단 지원 기술 개발
- 환자 및 병실 상태의 이상 징후 예측 및 지원 기술 임상 검증

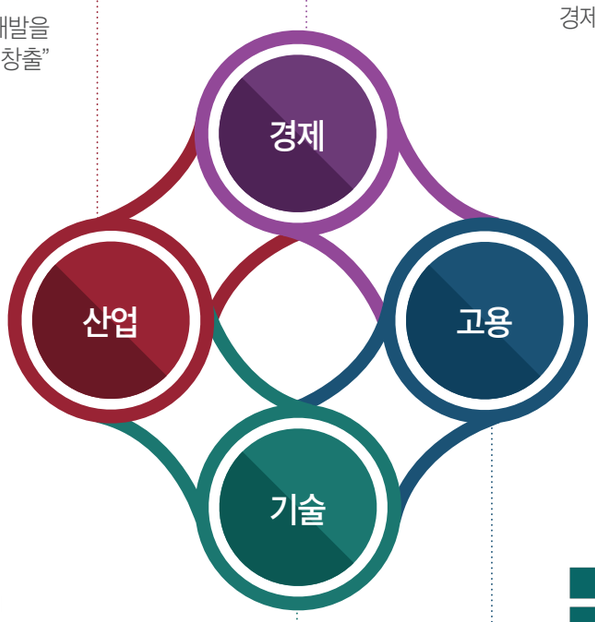
센 터 비 전

지식 패키지 산업 창출

"지식 획득 및 추론 기술개발을
통한 지식 패키지 산업 창출"

국가적 소모비용 절감

"AI 기술의 국가적인
경제 소모비용 절감 기여"



국제적 위상 제고 기여

"국제적 AI 선도 기술 보유국 구현"

국내 AI 인력 부족문제 해결

"세계적인 연구소 진입을 통한
인력양성 및 국가 차세대 리더 배출"

센 터 조 직


연구소장
임희석 교수
부연구소장
국내 저널 편집위원회
해외 저널 편집위원회
운영위원회
자문위원회
경영혁신본부
AI 원천기술
연구기획단
지능형
산업혁신단
AI융합 학술/
교육 사업단
AI사업
대외협력단

산학협력후원정책

후원 정책	정책내용
AI 공동연구 협력	기업이 AI 원천기술 개발 요청 시 적극적으로 협력
A I 공동과제 협력	기업이 공동 연구과제 요청 시 적극적으로 협력
공동 브랜딩 협력	기업이 요청하는 공동 브랜딩 관련 적극적으로 협력 (MOU, 공동 행사 개최, 언론 홍보 등)
A I TECH DAY	AI TECH DAY 행사 초대
	기업의 비즈니스모델에 부합하는 기업 전용 AI 기술교류회 추진
A I 자문회의	기업이 AI 자문회의를 요청하여 자문 수행
	기업의 문제해결을 위해서 국내외 AI 전문가를 초빙하여 AI 컨퍼런스 추진
국내 저널 발간	국내 저널 발간 국내 저널 발간 협력
국제 학술대회	1인 학술대회 등록비 면제
	1인 저널 심사비 면제
	1인 저널 게재비 면제
	1 인 항공료 지원
	1인 숙박비 지원
	기업 세션 지원(기업 전용 포스터 발표, 패널 토의)
	기업 자문 세션 지원
	기업 특별 세션 지원(데모/투자 데이, 리크루팅 등)

Contents

목차

연구실 개요

●○○○

자연어처리

띄어쓰기 자동 교정기

형태소 분석 기술

Korean Morphological Analyzer

개체명 인식기 (Named Entity Recognition)

문서 자동 분류 기술

Bag of Characters를 응용한 Character-Level Word Representation 기술

병렬 코퍼스를 이용한 bilingual word embedding

Stack-Pointer Network를 이용한 한국어 의존 구문 분석

Dependency Parser

Small Data의 한계를 극복하기 위한 전이 학습 모델

통계기반 한국어 뉴스 감정분석자연어 추론에서의 교차 검증 앙상블 기법

딥러닝 방식을 이용한 환유 해소

Denoising Teansformer기반 한국어 맞춤법 교정기

자식 임베딩 심층학습을 이용한 단어 의미 중의성 해소

Attentive Aggregation(주의적 종합)기반 크로스 모달 임베딩

단문 데이터를 활용한 다차원 감성 분석 서비스

딥러닝 작문 서비스와 응용 서비스 개발

01 띄어쓰기 자동 교정기

1. 기술 설명

본 기술은 기계학습을 이용하여 문장에서 띄어쓰기 오류가 있는 부분을 자동으로 파악하고 이를 올바르게 수정하는 방법이다.

Correction Rule
들어가셨다 → 들어+Space+가셨다

아버지가방에 들어가셨다 → 아버지가 방에 들어 가셨다

[그림 1] 단순 규칙을 이용하여 띄어쓰기 교정이 가능한 경우.

Correction Rules
가방 → 가+Space+방
non_space+가방 → Space+가방

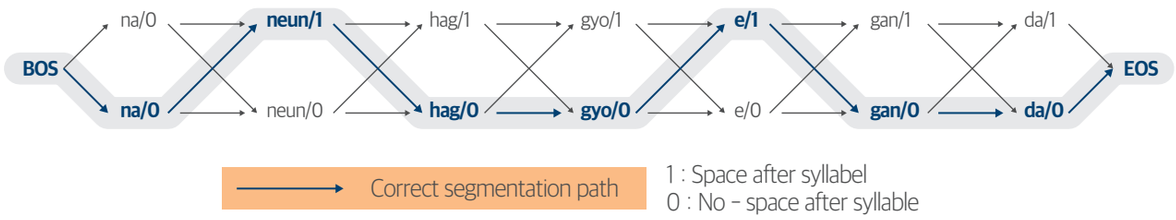
아버지가방에 들어가셨다 → 아버지가 방에 들어 가셨다

아버지가방에 들어가셨다 → 아버지 가방에 들어 가셨다

[그림 2] 단순 규칙으로 띄어쓰기 교정이 불가능한 경우. 확률 모델의 적용이 필요하다.

2. 기술 방법

한국어의 경우, 띄어쓰기는 독자에게 글의 가독성을 높이고 문장의 뜻을 정확히 전달하기 위해 매우 중요하다. 자동 띄어쓰기 시스템은 자연어처리 응용 시스템의 가장 기본이 되는 형태소 분석기의 전처리기, 문자인식기가 인식한 문서의 줄 경계를 복원하기 위한 후처리기, 음성인식기로부터 생성된 연속 음절 문장을 올바르게 띄어쓰기 위한 후처리기, 맞춤법 검사기의 한 모듈로서도 중요한 역할을 하고 있다.



[그림 3] 띄어쓰기 확률 경로 예시.

3. 기술 활용 및 응용 분야

감정 분석, 자연어처리

데모 시스템 : http://blpdemo.korea.ac.kr/autospacing/

02 형태소 분석 기술

1. 기술 설명

- 형태소 분석은 표층형 (surface level form)인 어절로부터 의미가 있는 최소 단위인 형태소 (morpheme)를 추출하는 작업
- 형태소 분석을 위해서는 어절을 분석하여 형태소의 결합으로 분리하고, 각 형태소에 품사정보를 할당하고, 형태소 결합 시 발생하는 음운 변화를 원형 (root form)으로 복원하는 것이 필요

<형태소 분석의 예>

예: 나는 나는 새를 보았다.

나는
나 / 대명사 + 는 / 조사
나 / 동사 + 는 / 관형형 어미
날 / 동사 + 는 / 는 / 관형형어미

2. 기술 방법

- 코퍼스의 통계적 특성과 확률 모델을 기반으로 한 전통적인 방식의 형태소 분석과 품사 태거임
- 품사부착 말뭉치 (POS tagged corpus)로부터 자동으로 획득한 통계 정보만으로 분석을 수행하였으며 3가지 언어 단위 (어절, 형태소, 음절)에 따른 분석 모델을 사용
- 어절, 형태소, 음절 단위 모델을 순차적으로 적용

<품사 태깅표>

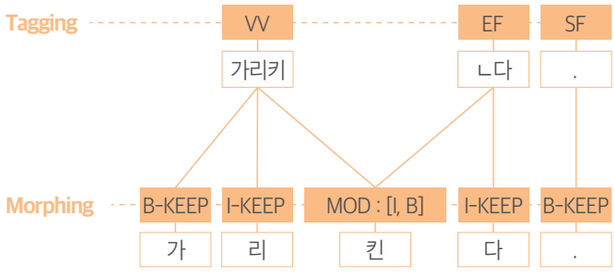
NNG :일반명사	JKS :주격조사	XSV :동사파생접미사
NNP :고유명사	JKG :관형격조사	XSA :형용사파생접미사
NNB :의존명사	JKO :목적격조사	SF :마침표,물음표,느낌표
NP :대명사	JKB :부사격조사	SP :쉼표, 가운뎃점, 콜론, 빗금, 줄표, 물결
NR :수사	JKV :호격조사	SS :따옴표,괄호표
VV :동사	JKQ :인용격조사	SE :줄임표
VA :형용사	JX :보조사	SO :붙임표(숨김,빠짐)
VX :보조용언	EP :선어말어미	SL :외국어
VCP :지정사	EM :어말어미	SH :한자
MM :관형사	ETN :명사형전성어미	SW :기타기호
MAG :일반부사	ETM :관형형전성어미	SN :숫자
MAJ :접속부사	XPN :명사파생접두사	NA :분석불능범주
IC :감탄사	XSN :명사파생접미사	

3. 기술 활용 및 응용 분야

- 본 기술은 번역기, 자연어 이해 및 생성 등 언어처리 분야의 핵심기술
- 데모 http://blpdemo.korea.ac.kr/MA

1. 기술 설명

본 기술은 어떠한 언어 단위도 입력으로 사용할 수 있으며 다단계 변형을 기반으로 형태소 분석 및 품사 부착을 수행하는 방법이다.

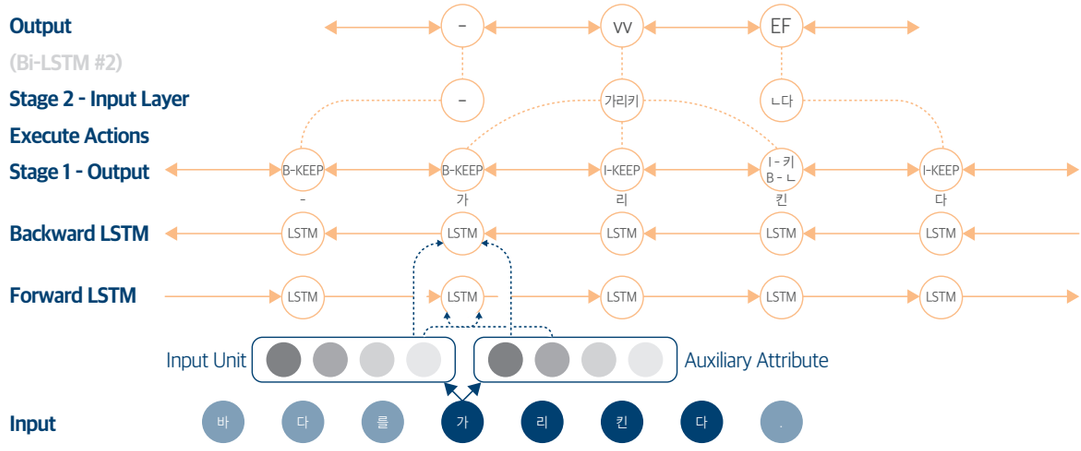


[그림 4] 형태소 분석 및 품사 부착 과정

2. 기술 방법

본 기술은 형태소 분석과 품사 부착의 두 단계를 거친다. 문장에 대해 형태소 분석이 우선 이루어지고, 형태소 분석 결과에서 각 형태소에 대해 품사를 부착한다. 모든 과정은 데이터 기반 종단 시스템으로, 사람의 개입 없이 학습 데이터만으로 모델을 훈련시킬 수 있다.

전체 모델은 양방향 Long Short-Term Memory(LSTM)-Conditional Random Field(CRF) 딥러닝 구조를 이용한다.



[그림 5] 본 기술을 바탕으로 “가리킨다”는 문자열이 형태소 단위인 “가리키”와 “ㄴ다”로 분할되고, 각각에 품사가 부착되는 과정

3. 기술 활용 및 응용 분야

형태소 분석, 자연어처리

데모 시스템 : http://nlplab.iptime.org:32280/unitagger_demo/

04 개체명 인식기 (Named Entity Recognition)

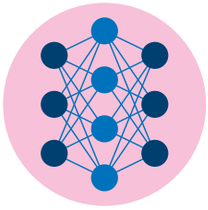
1. 기술 설명

- 정의: 개체명 인식기는 텍스트에서 인식시킬 개체를 정의하여 해당 개체를 인식시키는 기술을 말한다. 본 개체명 인식기는 5개의 클래스[인물(PS), 장소(LC), 기관(OG), 시간(TI), 날짜(DT)]를 정의하였으며, 해당 개체에 한국 문화적 특성을 반영한 개체명 인식기이다.
- 말뭉치 구축 : 학습에 필요한 말뭉치 구축을 위해 한국학중앙연구원 디지털 인문학 웹사이트의 백과사전 기사에서 전통문화와 관련된 기획기사 및 중심기사로부터 각 기사의 개요와 내용에 대한 문장들을 크롤링하였다.

텍스트 입력

백제는 한국의 고대 국가 중 하나로, 고구려, 신라 와 함께 삼국 시대를 구성하였다. 시조는 부여·고구려에서 남하한 온조 집단으로 마한 54개 연맹체 중 하나인 백제국으로 시작해, 4세기 중엽 근초고왕 때 마한 전체를 통일했다.

BI - LSTM - CNN - CRF 모델



개체명 인식 결과

백제는 한국의 고대 국가 중 하나로, 고구려, 신라 와 함께 삼국 시대를 구성하였다. 시조는 부여·고구려에서 남하한 온조 집단으로 마한 54개 연맹체 중 하나인 백제국으로 시작해, 4세기 중엽 근초고왕 때 마한 전체를 통일했다.

2. 기술 방법

- 한국어 기반으로 구축한 말뭉치의 전처리 과정을 통해 BI-LSTM-CNN-CRF 모델을 학습시킨다.
- 학습된 모델에 텍스트를 입력으로 넣어 해당 문장에서 개체명으로 인식 가능한 개체를 확인할 수 있다.

3. 기술 활용 및 응용 분야

- 본 모델을 영어 데이터로 학습시킬 경우 영어 기반의 개체명 인식기로 활용할 수 있다.
- 구축한 말뭉치를 다른 모델에 활용할 수 있다.
- 데모 : http://nlplab.iptime.org:32280/ner_demo/index.html

05 문서 자동 분류 기술

1. 기술 설명

- 문서가 어떤 카테고리에 해당하는지 자동으로 분류
- 본 기술은 kNN (k-nearest neighbors algorithm) 학습 방법을 이용

2. 기술 방법

- 인터넷 문서 5,000여개에서 추출한 자질 중 실험적으로 가장 높은 성능을 보인 2,000개의 자질을 추출
- 정보 검색 기법에서 사용되는 TF/IDF 기법을 이용하여 자질의 가중치 (Weight) 값 계산
- Nearest Neighbor를 추출하기 위하여 Cosine Measure를 사용

3. 기술 활용 및 응용 분야

• 본 기술은 정보 분류(대/중/소), 검색, 추천, 광고 등 언어처리 분야의 활용기술

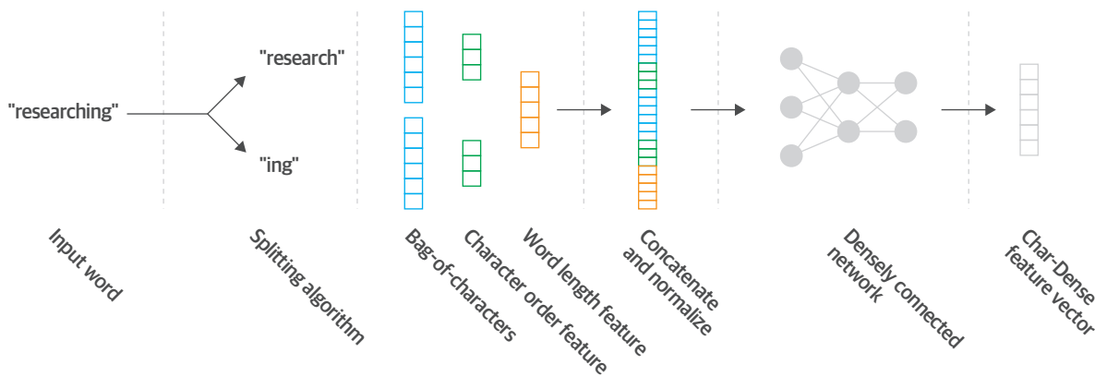
• 데모 http://blpdemo.korea.ac.kr/DocuCate/doccat.htm

06 Bag of Characters를 응용한 Character-Level Word Representation 기술

1. 기술 설명

본 기술은 완전연결 신경망을 이용하여 빠른 시간 안에 효과적인 문자 단위 자질을 자동적으로 추출할 수 있도록 하는 것이다. 자연어처리 시스템은 문자 단위 자질을 잘 반영할 수 있어야 한다. 이는 신조어 등 학습 시 존재하지 않았던 단어 등의 처리에 매우 효과적이다.

2. 기술 방법



본 기술은 Bag-of-Characters (BOC)를 바탕으로 한다. 문자 BOC, 문자 순서 정보 자질, 단어 길이 자질을 concatenate 하여 sparse vector를 생성한다. 이 sparse vector는 단어마다 유일하고 변하지 않으므로 속도 향상을 위해 캐싱이 가능하다.

Sparse vector를 하나의 은닉층이 있는 완전연결 신경망의 입력으로 사용해서 최종적인 문자단위 자질 벡터를 생성한다.

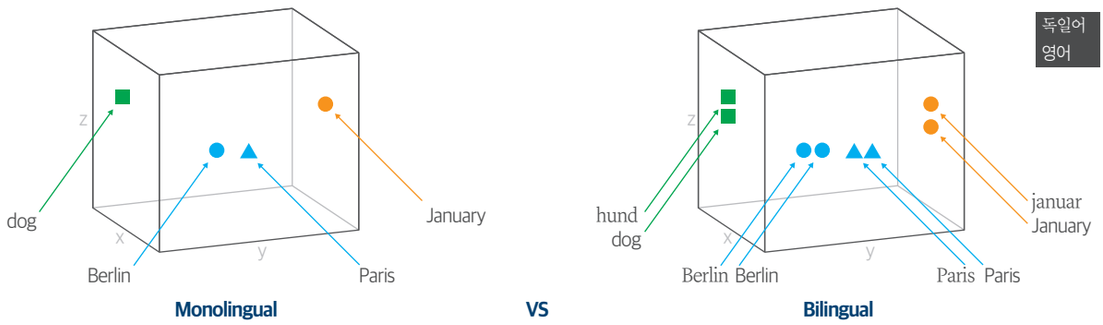
3. 기술 활용 및 응용 분야

품사 부착, 개채명 인식, 자연어 처리

07 병렬 코퍼스를 이용한 bilingual word embedding

1. 기술 설명

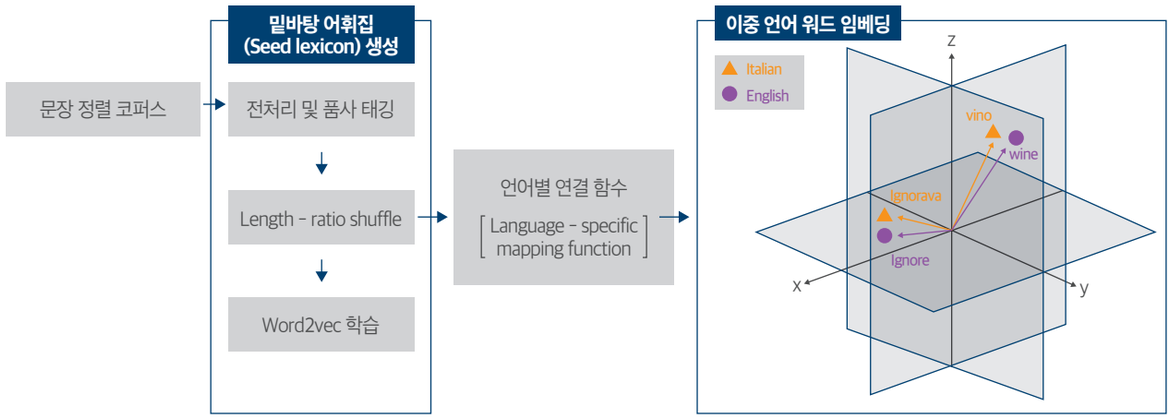
- 워드 임베딩이란 단어를 dense한 실수 벡터 공간에 매핑하되, 단어의 의미가 반영되도록 하는 방법
- 워드 임베딩의 활용방법 중인 하나인 이중 언어 워드 임베딩은 서로 다른 두 언어에서 유사한 의미를 가지는 단어가 유사한 공간에 매핑(mapping) 되도록 하는 것을 목표로 하 는데, 기계번역 분야에서 많은 연구가 이루어지고 있음



<Monolingual vs Bilingual 예시>

2. 기술 방법

- 본 기술은 문서 정렬 코퍼스보다는 언어 간의 연결고리(bilingual signal)가 강한 문장정렬 영화자막 데이터를 이용한 이중 언어 워드 임베딩 모델 개발
- 개발한 모델은 영화자막 데이터를 강력한 언어 간의 연결고리로써 밑바탕 어휘집으로 사용하여 서로 다른 두 언어를 동일한 공간의 벡터 공간으로 매핑



<Bilingual word embedding 모델 개요>

3. 기술 활용 및 응용 분야

- 본 기술은 다중 언어에 대한 번역기에 활용될 수 있으며, 다중 언어 문서에서 정보검색 모델에서도 활용될 수 있다.
- 데모 http://nlplab.iptime.org:4321/seol2/mt/projector.html

1. 기술 설명

• 의존 구문 분석 기술은 자연어 문장에 포함된 단어들의 의존 관계를 분석하는 기술

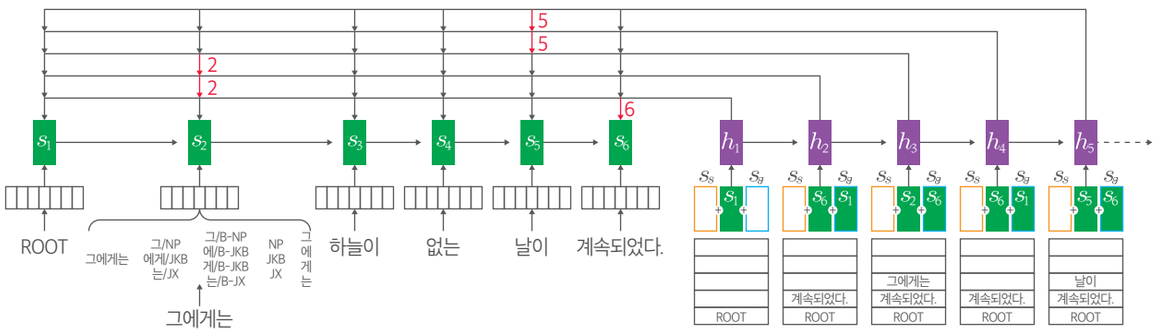
• 그림과 같이 단어들의 의존 관계과 각 의존 관계의 유형을 나타내는 의존 분석 트리 구축

(예: '학교에'는 '가서'에 의존하는 부사어)

2. 기술 방법

• 최신 딥러닝 기반 의존 분석 모델인 Stack-Pointer Network를 한국어 의존 구문 분석에 적합하도록 확장

• 양방향 LSTM-CNN 구조의 인코더에서 각 어절의 단어 표상 생성에 형태소, 형태소 품사 정보가 포함된 음절, 형태소 품사, 음절 정보를 추가 활용



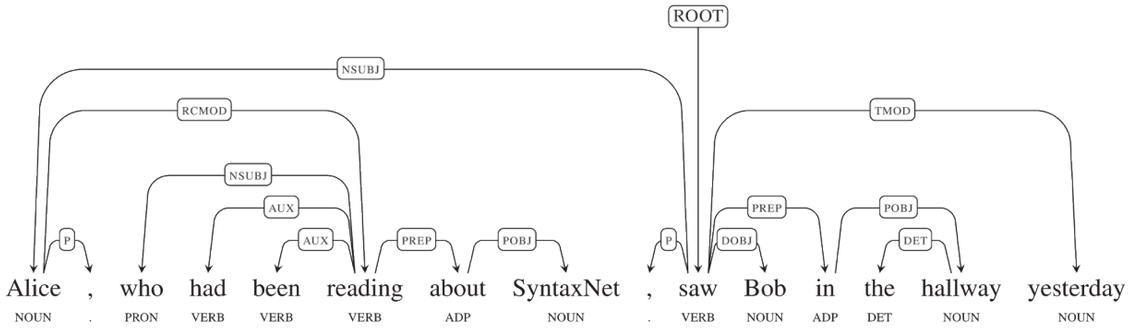
3. 기술 활용 및 응용 분야

• 본 기술은 대용어 참조 해소, 기계 번역 등의 다양한 자연어 이해 기술에 세부기술로 활용 될 수 있음

• 데모 http://nlplab.iptime.org:32281/kr-stack-pointer/index.py

1. 기술 설명

본 기술은 영어를 대상으로 하는 SyntaxNet 시스템을 한국어에 사용할 수 있도록 한 것이다. SyntaxNet은 구글에서 개발한 의존구분분석 기술로, 데이터 기반 종단간 시스템으로 동작한다. SyntaxNet의 의존구문분석 정확도는 94% 이상으로, 인간의 수준인 96~97%에 가까운 성능을 보인다.



[그림] "Alice, who had been reading about SyntaxNet, saw Bob in the hallway yesterday"라는 문장에 대한 의존구분분석 예시

2. 기술 방법

의존구분분석은 상위 레벨 자연어처리 작업 중 하나로, 수많은 가능한 의존 트리에서 최적의 트리를 찾아내야 한다. SyntaxNet은 품사 정보가 입력으로 필요하다. 이에 추가로 한국어에 적용하기 위해서는 형태소 분석이 우선적으로 진행되어야 한다. SyntaxNet 모델에 의해 의존 구분분석이 완료된 결과에 대하여, 원래의 어절 형태로 형태소들을 재결합하는 과정도 요구된다.

3. 기술 활용 및 응용 분야

의존구문분석, 대화 시스템, 자연어처리

데모 시스템 : http://andrewmatteson.name/psg_tree.htm

10 Small Data의 한계를 극복하기 위한 전이 학습 모델

1. 기술 설명

- 전이 학습은 특정 환경에서 만들어진 모델을 다른 비슷한 task에 적용하는 것으로, 이는 데이터가 부족한 분야에도 적용할 수 있음
- 풍부한 데이터로 먼저 모델을 학습하고 데이터가 부족한 비슷한 task에 대해 모델의 전이를 진행하는 것임. Small Data의 한계를 극복한다는 점에서 큰 장점이 있음
- 아래는 항공권 예약을 위한 ATIS 데이터와 식당 예약을 위한 MIT 데이터임. 각각의 slot들은 조금씩 다르지만, 예약을 위한 대화 데이터라는 점이 유사하며, ATIS의 city와 MIT의 Location이 특징이 위치라는 점에서 매우 유사함

ATIS UTTERANCE EXAMPLE IOB REPRESENTATION

Sentence	*show*	*flights*	*from*	*Boston*	*To*	*New*	*York*	*today*
Slots/Concepts	O	O	O	B-dept	O	B-arr	I-arr	B-date
Named Entity	O	O	O	B-city	O	B-city	I-city	O
Intent	*Find_flight*							
Domain	*Airline Travel*							

ATIS 항공권 예약 데이터에 대한 Slot Filling의 예시

Are	there	any	French
O	O	O	B-Cuisine
restaurants	**in**	**downtown**	**Toronto**
O	O	B-Location	I-Location

MIT 식당 예약 데이터에 대한 Slot Filling의 예시

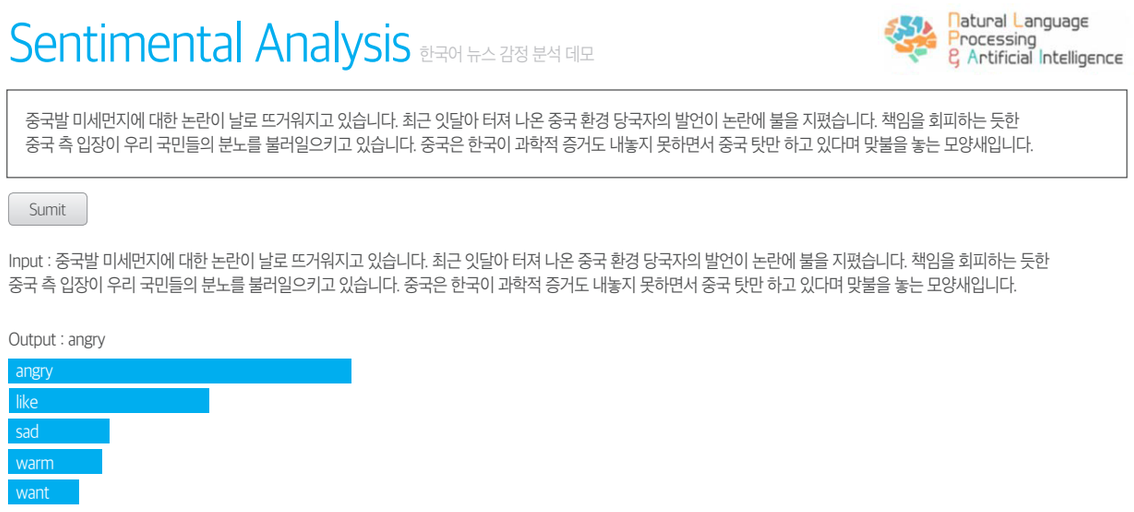
2. 기술 방법

- 자연어 이해 시스템을 학습하기 위해서는 많은 양의 라벨링 된 데이터가 필요하며 새로운 도메인으로 시스템을 확장할 때, 새롭게 데이터 라벨링을 진행해야 하는 한계점이 존재한다. 본 연구는 적대 학습 방법을 이용하여 풍부한 양으로 구성된 기존(source) 도메인의 데이터부터 적은 양으로 라벨링 된 데이터로 구성된 대상(target) 도메인을 위한 슬롯 채우기(slot filling) 모델 학습 방법이다.
- 본 연구에서는 슬롯 채우기(Bi-directional LSTM 기반), 도메인 분류를 위한 적대 학습, Orthogonality Loss 등을 적용하여, 도메인 고유 및 공유 자질을 서로 상호 배타적으로 학습하였다.
- 대화 데이터 중 항공권 예약 도메인 데이터인 ATIS 데이터와 식당 예약 도메인 데이터인 MIT 식당 예약데이터를 이용하여 실험을 진행하였으며, 적대 학습 방법을 이용한 슬롯 채우기 모델 성능을 확인하였다.

3. 기술 활용 및 응용 분야

- 본 기술은 도메인 간 전이 학습이 가능하기에 데이터가 부족한 목적 지향 대화 데이터 시스템의 학습에 활용될 수 있음

1. 기술 설명



<Sentiment Analysis Demo 결과 화면>

- Text Sentiment Analysis는 텍스트로부터 예상되는 감정과 반응을 예측하는 기술
- 데이터는 5개의 감정이 태깅된 10만 개 이상의 뉴스 기사를 이용함. 최소한의 전처리 과정을 거쳐 감정을 예측하는 통계기반 알고리즘을 제안함

2. 기술 방법

- 뉴스 기사에 등장한 단어들을 vocabulary에 추가함
- 뉴스 기사에 대한 vocabulary 내 단어의 tf-idf* 값을 구하고, 뉴스 기사에 태깅된 감정을 참조하여 각 단어들을 5차원 벡터로 표현함
- 입력된 텍스트에서 vocabulary에 포함된 단어를 찾아 미리 계산된 벡터값으로 변환하고, 모든 단어의 벡터값을 합산하여 가장 높은 confidence를 가진 감정을 출력함
- (*tf-idf: 해당 단어의 출현 빈도와 희귀성을 고려하여, 해당 단어가 해당 문서에 대해 얼마나 가치 있는 단어인지 나타내는 값)

3. 기술 활용 및 응용 분야

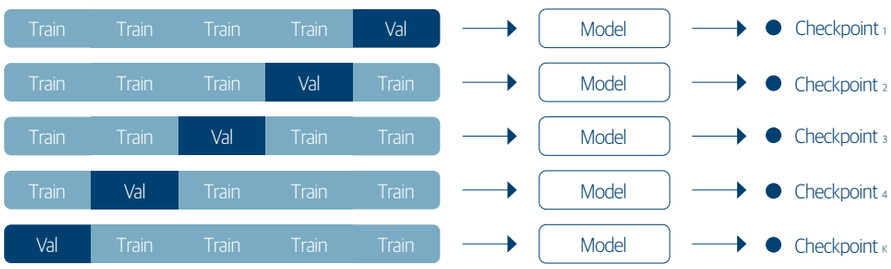
- 본 기술은 적은 컴퓨팅 자원을 이용하며, 텍스트로부터 의미 있는 특질(feature)을 추출함. 따라서 음성 인식, 자연어 이해 등 다른 자연어처리 모델에 적은 비용으로 의미 있는 특질을 제공 가능함
- 데모 http://nlplab.iptime.org:32280/sentiment_demo/index.py

12 자연어 추론에서의 교차 검증 앙상블 기법

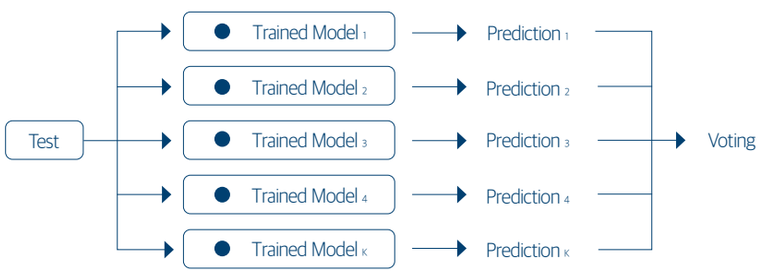
1. 기술 설명

앙상블(Ensemble)은 여러 모델들의 예측값을 종합하여 최종 판단을 내리는 기계학습 기법이다. 대표적인 앙상블 기법으로는 Bagging(Bootstrap Aggregating)이 있으며, 이는 다양한 샘플로 모델을 학습시키기 위한 반복과정이 필요하여 앙상블기법만을 위한 별도의 연산이 요구된다. 이러한 문제를 해소하기 위하여 Checkpoint Ensemble(CE) 기법이 제안되었으나 학습 소요 시간이 경감되어 데이터의 분포가 고르지 않을 경우 높은 분산을 보일 수 있다는 한계가 있다. 본 기술은 앙상블 기법을 교차검증 방법과 결합하여 앙상블 연산을 위한 비용을 줄이며 일반화 성능을 높인다.

2. 기술 방법



Training Phase



Test Phase

본 기술은 별도의 연산을 피하면서 분산 경감 면에서도 강점을 가지는 교차 검증 앙상블(Cross-Validated Ensenble, CVE)기법이다. 이는 Bagging처럼 여러 샘플을 추출해 학습하는 효과를 얻는 동시에 교차 검증시 기록된 checkpoints로 앙상블하므로 별도의 연산이 요구되지 않는다. 교차 검증 앙상블 기법은 다음과 같은 단계로 진행된다.

- 전체 학습 데이터를 k-fold로 나누고, 선정 모델을 k개의 샘플 데이터로 개별 학습 시킨다. 이때 validation score가 가장 높은 지점을 미리 기록한다.
- 교차 검증 데이터로 학습을 마친 뒤, 학습한 모델들과 테스트셋을 입력 받는다.
- 각 fold별로 validation score가 가장 높은 checkpoint를 찾아 k개의 모델을 준비한다.
- 선정된 k개의 모델이 예측한 labels를 평균내어 최종 예측 값을 반환한다.

13 딥러닝 방식을 이용한 환유 해소

1. 기술 설명

환유법이란 대유법의 일종으로 표현하려는 대상을 그와 관련된 다른 사물 또는 속성으로 대신 나타내는 방법이다. 최근 대부분의 환유 해소 연구는 환유의 특정한 유형에 대한 분류 문제로 진행되며, 일반적으로 LOC(location) 및 ORG(organization)유형에 대하여 연구가 진행되고 있다. 아래는 환유 유형(LOC/ORG)의 예이다.

[Literal]

(1) 올해 G20은 대한민국 **서울**에서 개최한다.

(2) **삼성**은 2019년 CES에서 신제품을 공개하였다.

[Metonymic]

(1) **서울시**는 오늘 오전 공식 입장을 발표하였다.

(2) 우리 가족들은 **삼성**만 쓴다.

[Literal]의 문장들은 각각 지리적 위치인 서울과 기관을 지칭하는 삼성을 나타내고 있다.

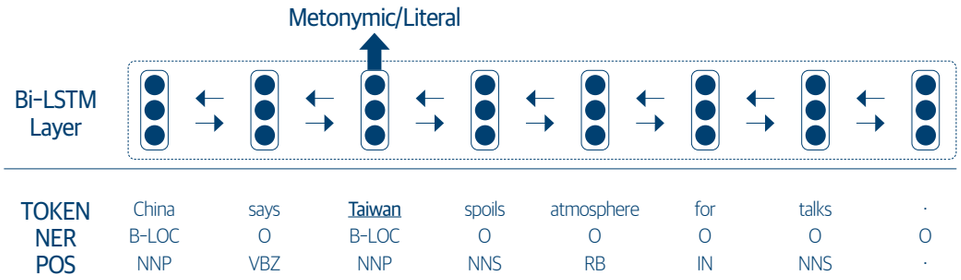
[Metonymic]에서는 본연의 의미가 아닌 환유적 표현을 담고 있다. (1)의 경우, 대한민국 수도 서울이 아닌 서울시 관계자 또는 서울 시장이라고 해석할 수 있으며, (2)의 경우, 삼성 기업이 아닌 삼성의 제품들을 대신 표현한 것이라고 해석할 수 있다.

본 기술은 언어학적 자질 정보를 최소화환 딥러닝을 이용한 환유 해소 모델로서 주어진 엔티티의 환유 여부를 구분하기 위하여 앞뒤 단어와의 연관성, 문장 전체의 문맥 정보를 고려하였다.

2. 기술 방법

• Feature-based 방법

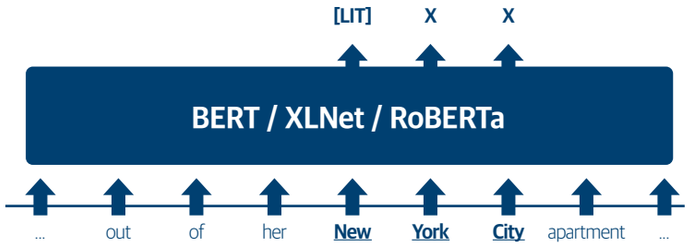
GLoVE 및 ELMo와 같은 단어 임베딩을 기반으로 하며, LSTM(long-short term memory)모델을 통하여 입력 문장의 문백 정보를 표현하고, 환유가 나타나는 엔티티의 단방향 또는 양방향 LSTM(BiLSTM) 출력 값을 통해 모델 학습을 진행한다. 아래 그림은 양방향 LSTM을 기반으로 하는 환유 해소 모델을 나타낸 것으로 모델의 입력으로는 각 토큰의 단어 임베딩, 개체명 인식 임베딩, 품사 임베딩이 결합되어 들어간다.



[그림] 양방향 LSTM기반의 환유 해소 모델 아키텍처

• Feature-tuning 방법

대용량 코퍼스에 대해 비지도 학습을 진행하여 미리 학습시킨 contextual language model(BERT, XLNet, RoBERTa)모델로 학습을 진행하였다. BERT의 경우 순방향과 역방향 LSTM은닉 출력값을 결합하여 임베딩을 생성한 ELMo와 달리 attention기반의 Transformer인코더를 통해 완전한 양방향 학습을 진행하였다. Masked LM(MLM)과 Next Sentence Prediction(NSP) 두 개의 목적함수로 비지도 학습을 진행으로 문맥을 고려한 언어 표현을 가능하게 하였다. 환유 엔티티 토큰의 BERT, XLNet, RoBERTa 모델 출력 값은 최종적으로 해당 토큰이 Metonymic인지 Literal인지 분류하는데 사용되며, 해당 엔티티가 여러개의 subtoken으로 분리되었을 때에는 엔티티의 가장 앞에 있는 토큰을 분류를 위한 다층 퍼셉트론의 입력으로 사용한다. 다층 퍼셉트론은 tanh활성화 함수와 softmax출력 층으로 구성되며, 모델 전체는 end-to-end방식으로 학습이 진행된다. 모델 학습은 Cross-entropy loss를 사용하여 진행하였다.



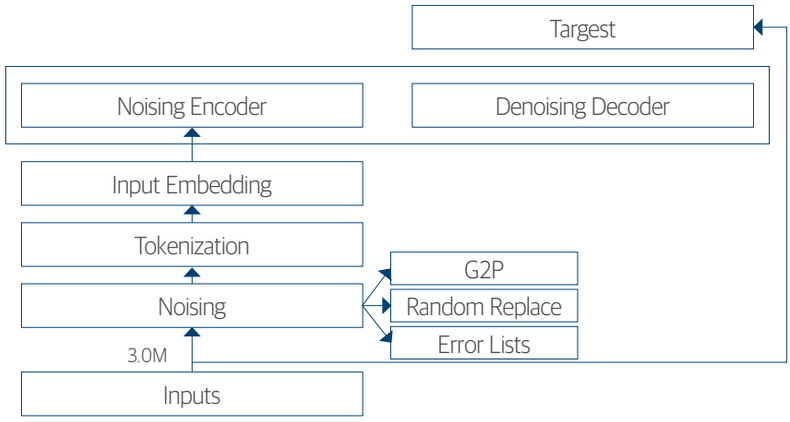
[그림] Fine-tuning방식의 환유 해소 모델

14 Denoising Transformer기반 한국어 맞춤법 교정기

1. 기술 설명

맞춤법 교정이란 주어진 문장에서 나타나는 철자 및 맞춤법 오류들을 올바르게 교정하는 것이다. 본 기술은 기존의 맞춤`법 교정기술과 달리 소스 문장에 맞춤법 오류문장, 타겟 문장에 올바른 문장을 넣어 학습시키는 기계번역 관점에서의 맞춤법 교정기술이다.

2. 기술 방법



기계번역이란 소스문장(Source Sentence)을 타겟문장(Target Sentence)으로 번역하는 시스템으로 이를 맞춤법 교정 시스템에 적용하여 소스문장으로는 오류문장을, 타겟 문장으로는 교정문장으로 사용하였다. 본 기술은 기존의 규칙기반 맞춤법 교정방식, 통계기반 맞춤법 교정방식과 달리 고품질의 병렬 말뭉치가 존재할 경우 별도의 규칙을 구축하지 않아도 다양한 양상의 맞춤법 오류를 수정할 수 있는 Transformer방식으로 개발하였다.

Transformer방식은 Convolution과 Recurrence 없이 오직 Attention만을 이용한 기계번역 모델로 Query, Key, Value를 기반으로 하는 Multi Head Attention을 기반으로 한다. 이는 입력과 출력에 대해 각각 Self Attention을 학습하고 이후 입력과 출력사이의 Attention을 학습하는 구조를 가진다.

3. 실행결과

• URL: http://nlplab.iptime.org:32288/

Type the text you went to translate and click "Translate".

지금 어디가세용

Translate

맞춤법 교정 결과

지금 어디가세요

15 지식 임베딩 심층학습을 이용한 단어 의미 중의성 해소

1. 기술 설명

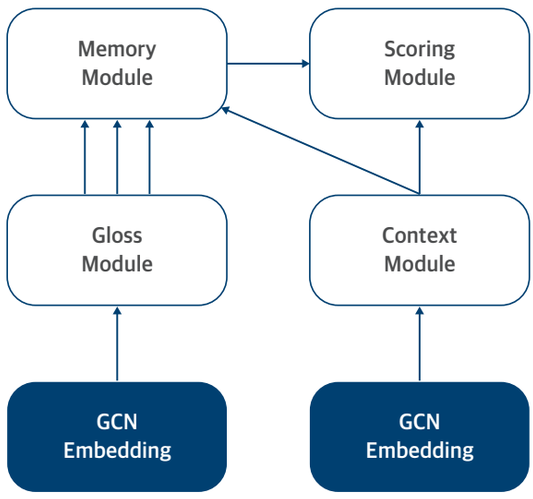
단어 중의성 해소란 두 개 이상의 의미를 가진 단어를 문장의 쓰임에 따라 정확하게 분석하는 것이다. 본 기술은 단어의 중의성을 해소하는 기술로 단어의 표상에 구문 정보와 의미 관계를 반영할 수 있도록 그래프 임베딩을 활용하였다.

2. 기술 방법

본 기술은 단어 표상에 구문 정보와 의미 관계 정보를 반영하기 위하여 GCN(Graph Convolution Network)를 사용하였으며, 구문 정보를 반영하기 위하여 Stanford CoreNLP parser에서 표현되는 의존 관계 정보를 활용하였다. 또한 의미 관계 정보를 나타내기 위해 WordNet정보를 활용하였다.

[단어 중의성 해소 모델]은 Context, Gloss, Memory, Scoring 4개의 모듈로 구성되어 있으며, 모든 단어 벡터는 SemGCN 단어 표상 결과를 사용하였다.

- Context Module: 중의성 단어를 가지는 단어의 문장을 Bi-LSTM을 통해 순방향, 역방향으로부터 나온 벡터값을 concatenate하여 표현함
- Gloss Module: 중의성 단어의 의미설명(Gloss)정보를 같은 방법으로 Bi-LSTM을 통하여 표현하며, Gloss Expansion방법을 사용함. 동시에 명사품사를 가지는 상위어, 하위어의 모든 의미설명 정보들도 Bi-LSTM으로 표현함. 상위어, 하위어 정보는 BFS(Breadth First Search)를 통하여 깊이 K만큼 추출하여 관련된 Gloss정보를 Context Module과 같이 표현함. 이러한 Gloss정보들은 Relation Fusion Layer을 통해 상위어는 순방향 LSTM에 나열하고, 하위어는 역방향 LSTM에 나열하여 벡터로 표현한 뒤, concatenate하여 표현함
- Memory Module: Context Module의 벡터결과와 Gloss Expantion 모듈에서의 벡터 결과를 Attention을 통해 계산 후 메모리를 업데이트함
- Scoring Module: Context Module의 벡터결과와 Memory 모듈의 마지막 Attention 결과값을 사용하여 중의성 단어의 의미를 선택함



[그림] 단어 중의성 해소 모델

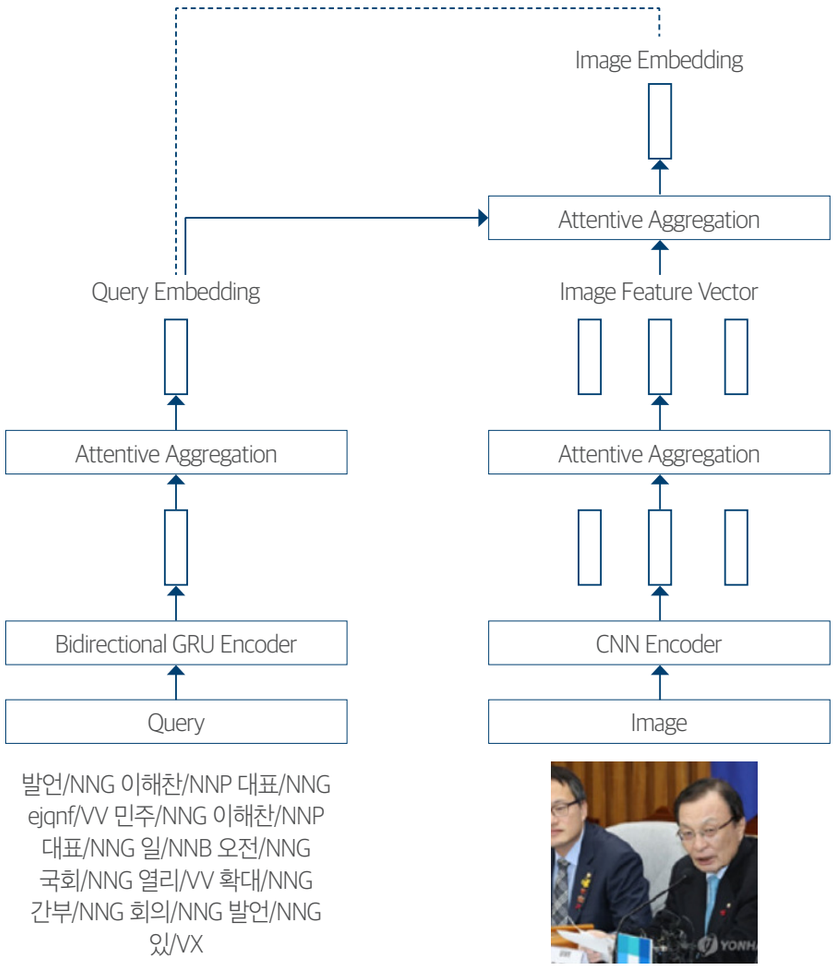
16 Attentive Aggregation(주의적 종합)기반 크로스 모달 임베딩

1. 기술 설명

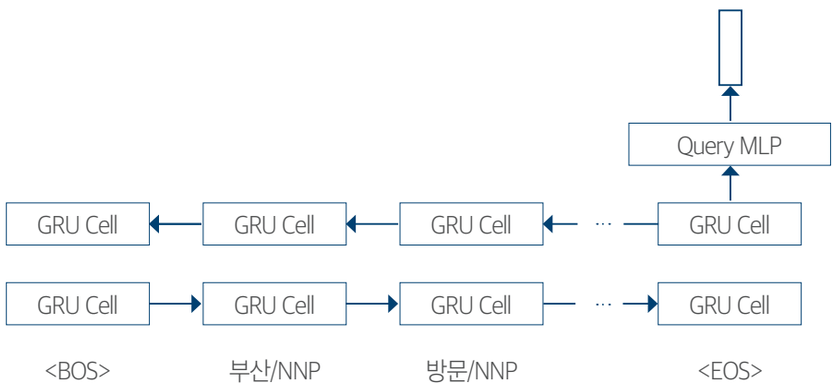
본 기술은 사진 검색을 위한 주의적 종합(Attentive Aggregation)기반의 언어-시각 크로스 모달 임베딩 모델로서 자연어 질의로부터의 사진 검색 과제를 해결할 수 있다. 본 기술은 사진으로부터 여러 개의 특징 벡터를 계산한 뒤 자연어 질의의 임베딩에 따라 Attentive Aggregation을 적용한다. 이는 이미지의 다양한 특징에 선별적으로 집중하여 질의와 사진 간의 유사도를 평가함으로써 언어와 시각 모달 간의 의미적 간극을 크게 줄일 수 있다.

2. 기술 방법

본 기술은 질의 기반 종합 검색 대상 임베딩 방법에 기반하여 질의 인코더, 사진 인코더, Attentive Aggregation Layer로 구성되어 있다. 질의 인코더와 사진 인코더에서는 자연어 질의와 사진으로부터 의미적 특징들을 추출하며, 서로 다른 형태의 데이터인 질의와 사진을 공통의 벡터 공간에 매핑하는 것을 목표로 한다. 계산된 사진 임베딩과 질의 임베딩 간의 Triplet Semi-hard Loss를 최소화하여 의미적으로 유사한 사진과 질의의 임베딩 간 거리를 최소화하였다.

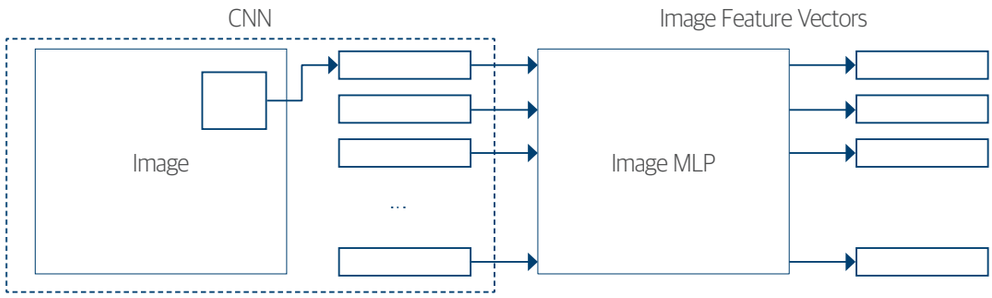


[질의 인코더] 양방향 GRU와 MLP구조로 구성되며, 자연어 질의로부터 하나의 질의 임베딩 벡터를 계산함. 입력으로는 자연어 질의를 분할한 토큰들의 임베딩을 사용하였으며, 본 모델에는 형태소 분석기를 통해 분할한 형태소들 중 질의의 핵심 정보를 나타낼 것으로 예상되는 명사와 동사 형태소를 사용함. 양방향 GRU Layer에서는 토큰들의 임베딩을 입력으로 받아 질의 전체의 정보를 반영한 특징 벡터를 계산하고, MLP Layer에서는 이를 사진 임베딩과 공통벡터 공간에 매핑되는 질의 임베딩으로 변환함



[그림] 질의 인코더 구조

[사진 인코더] CNN과 MLP구조로 구성되며, 사진의 여러 영역을 각각의 사진 특징 벡터들로 인코딩함. CNN Layer에서는 각 픽셀의 RGB색상 값을 0~1 범위의 실수로 변환된 값을 입력으로 받아 사진의 각 영역에 대한 특징 벡터들을 계산함. 이후 MLP Layer에서 이를 부다 상위의 의미정보를 반영하는 사진 특징 벡터들로 변환함



[그림] 사진 인코더 구조

[Attentive Aggregation Layer] 질의 임베딩 벡터에 따라 여러개의 사진 특징 벡터들을 가중합하여 사진 임베딩을 계산함. Attentive Aggregation은 질의 기반 종합 검색 대상 임베딩의 종합 방법으로 활용되었으며, 이는 질의 임베딩에 따라 정보량이 많은 사진으로부터 다양한 정보를 추출하여 선택적으로 활용할 수 있게 함.

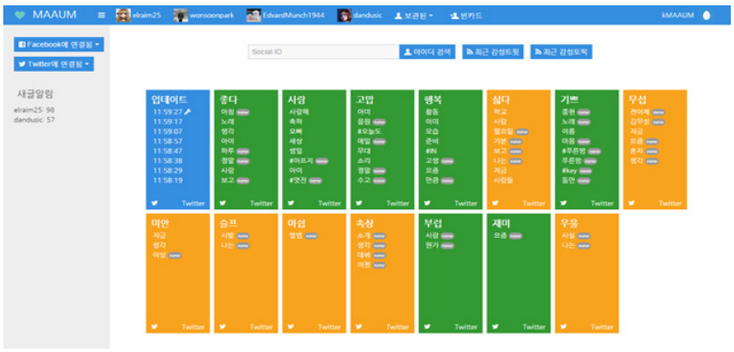
17 단문 데이터를 활용한 다차원 감성 분석 서비스

1. 기술 설명

- 소셜미디어 상의 단문 데이터를 활용하여 의미 있는 정보를 찾고, 조직화함으로써 정보 간의 관계나 트렌드 등을 분석하는 서비스를 제공함
- 검색 키워드에 대한 결과 제공시 해당 키워드에 대한 대중들의 감정을 분석하여 제공함
- 감정 분석은 긍/부정이 아닌 세분화된 25개의 감정으로 분류함

2. 기술 방법

- 전문가에 의한 감성 코퍼스를 구축하고 이를 교차검증하여 감성 코퍼스에 대한 신뢰도를 확보함



- 기계학습 기반의 토픽/감정 키워드 간의 관계분석 서비스를 제공함

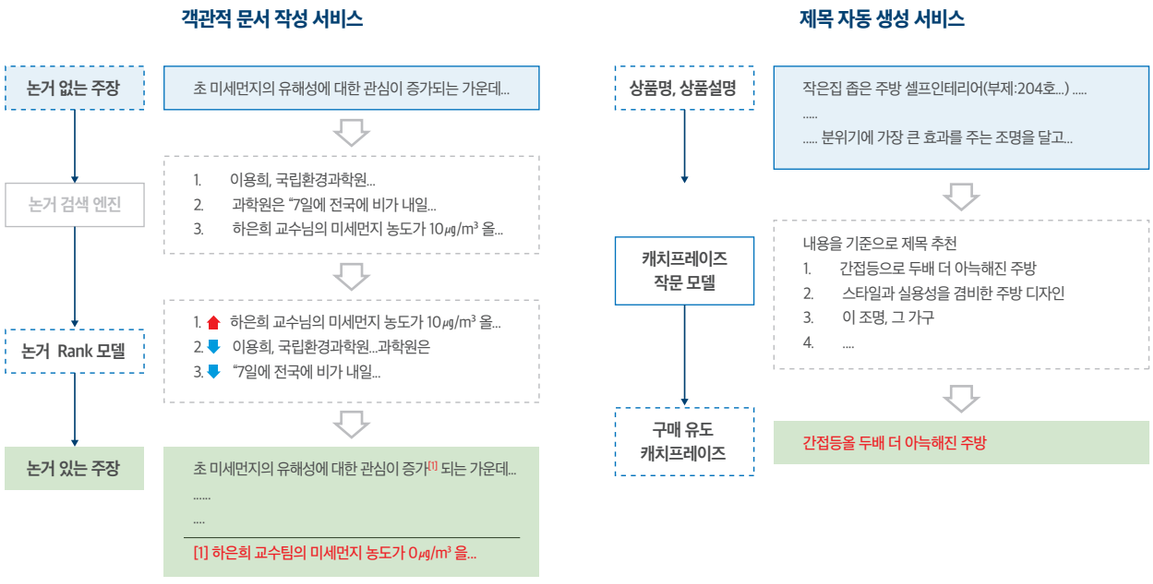
3. 기술 활용 및 응용 분야

- 트렌드 감정 분석 및 여론 분석에 활용 가능함

18 딥러닝 작문 서비스와 응용 서비스 개발

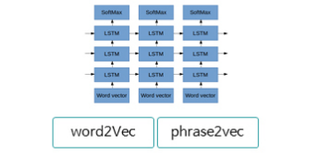
1. 기술 설명

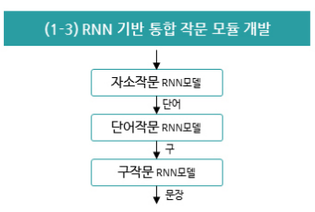
- 데이터를 활용하여 비즈니스 문서 작성에 대한 템플릿 및 예문을 자동으로 제시하고 작성 내용에 대한 객관적 근거 자료 제공을 자동화 하여 문서 작성에 소요되는 시간과 비용을 절감하기 위한 서비스를 제공함
- 문서 작성 시 정확한 근거 자료(연구소, 협회, 정부기관 등)를 제시함으로써 객관적인 내용 작성이 가능하도록 지원하며, 작성된내용(문서, 상품설명 등)을 학습하여 트렌드와 회사 및 제품의 특성에 맞는 제목을 자동 생성하여 제공함

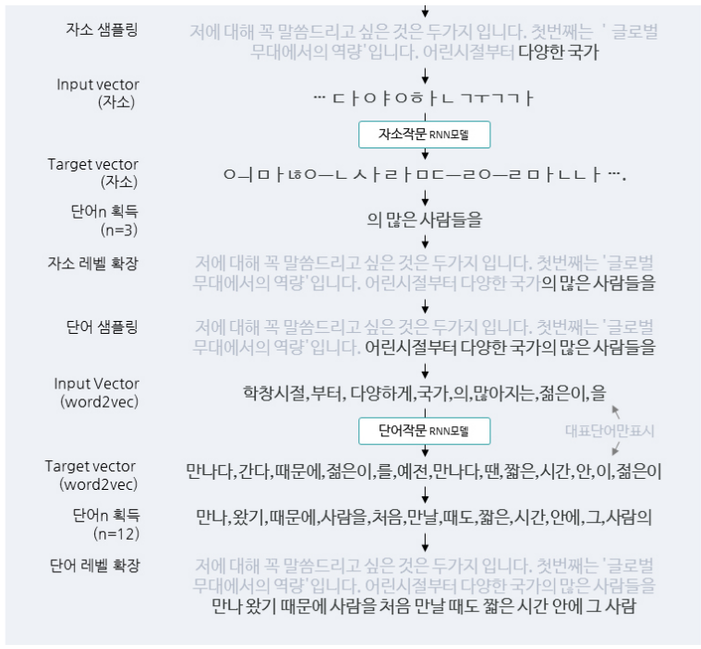


2. 기술 방법

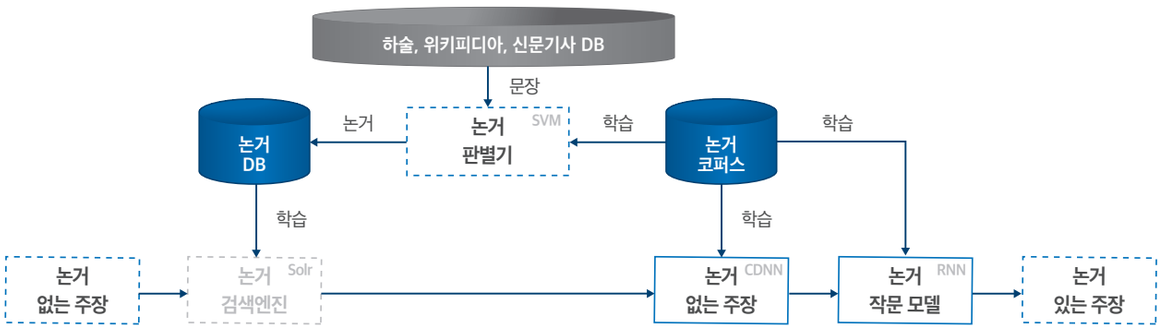
- 문맥을 고려한 자연스러운 문장 완성을 위한 작문 모듈 개발







• 객관적 문서 작성을 위해 Word2Vec+SVM과 딥러닝 기반의 논거 검색엔진을 구축하여 문장 장성 시 해당 내용에 맞는 논거(학술 논문, 위키피디아, 신문기사)를 제공함



• 자동 제목 작성을 위해 포털 사이트로부터 학습 데이터를 구축하고, 자체 알고리즘을 적용함

3. 기술 활용 및 응용 분야

• 대화 시스템에서 문맥정보를 고려한 자연스러운 문장 생성이 가능함

○●○○

대화 시스템

대화 시스템에서의 자연스러운 대화를 위한
Memory Attention 기반 Breakdown Detection

듀얼 메모리 네트워크를 이용한 대화 시스템

검색 기반 대화 시스템에서의 정답 예측 기술

한국어 특성을 반영한 시스템 액션 템플릿 기반의 대화 시스템

딥러닝 기반 자동 질의응답 시스템

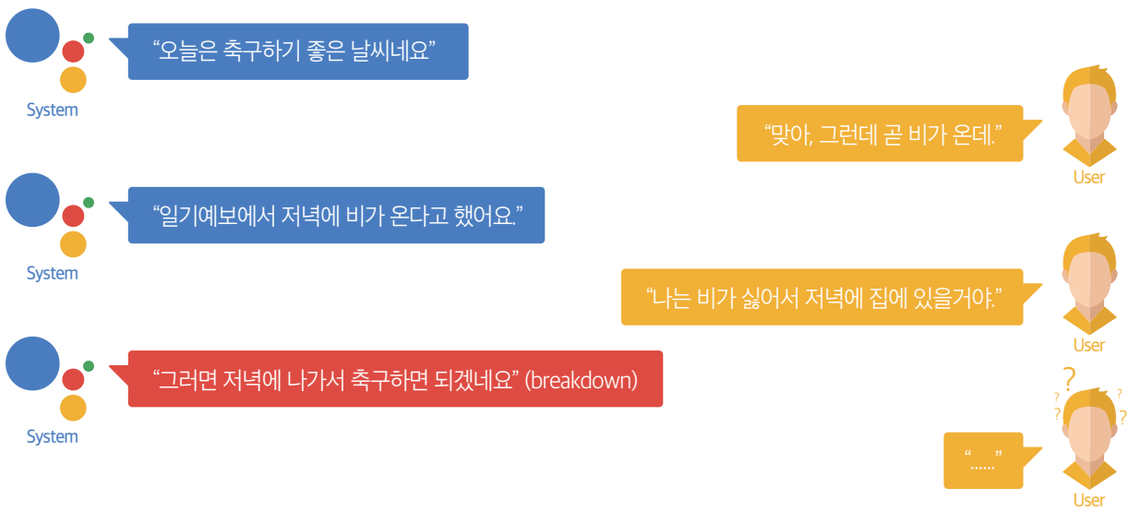
딥러닝 방법을 이용한 발화의 공손함 판단

01 대화 시스템에서의 자연스러운 대화를 위한 Memory Attention 기반 Breakdown Detection

1. 기술 설명

- 대화 시스템에서 Breakdown detection이란 사람과 시스템간의 자연스러운 대화의 흐름이 끊어지는 현상을 탐지하는 것임
- 대화 시스템을 이용하는 사용자 입장에서는 자연스러운 대화가 이루어져야 시스템에 대한 만족을 통해 원활한 서비스를 이용할 수 있음
- 아래 그림은 대화 시스템에서 breakdown이 발생하는 예시를 보여준 것임. 시스템-사람 간의 대화를 보면 마지막에 사람이 "나는 비가 싫어서 저녁에 집에 있을 거야."라고 하였으나, 시스템은 문맥에 맞지 않는 발화(빨간색)를 하여 자연스러운 대화의 흐름이 끊김을 알 수 있음

<대화 시스템에서 시스템-사람간의 대화에서 breakdown 발생 예시>



2. 기술 방법

- 본 기술은 end-to-end 기반의 breakdown detection 모델이며, LSTM(Long short-termmemory)을 이용하여 대화내에 사용자와 시스템의 발화를 인코딩하고 시스템 발화에 대해 memory network기반의 attention 기법을 이용하여 breakdown detection을 수행하는 구조를 가지고 있다.

3. 기술 활용 및 응용 분야

- 대화 시스템을 지원하고 있는 기기의 소프트웨어서 활용 가능하며, 기존의 인공지능 스피커 서비스인 NUGU, kakao mini 등에서 활용가능함

02 듀얼 메모리 네트워크를 이용한 대화 시스템

1. 기술 설명

- 목적 지향 대화 시스템은 식당 예약과 같은 특정한 목적을 수행할 수 있는 대화 시스템으로, 다양한 도메인에서 사용될 수 있음
- 시스템 액션 템플릿을 통해 시스템이 응답할 수 있는 답변을 한정하고, 도메인 지식을 반영했기 때문에 적은 양의 데이터로도 대화 모델의 학습이 가능함

System Action Templates
hello what can I help you with today any preference on a type of cuisine
api_call <price> <number of people> <cuisine> <location>
great let me do the reservation
here it is <in f o_addr e s s>
here it is <in f o_phone>
how many people would be in your party
I'm on it is there anything i can help you with
ok let me look into some options for you
sure is there anything else to update
sure let me find an other option for you
what do you think of this option: <re s tau r ant >
where should it be which price range are looking for
you're welcome

식상 예약 관련 시스템 액션 템플릿의 예

2. 기술 방법

- 듀얼 메모리를 이용한 대화 이력 추적 : 사용자 발화와 시스템 응답 발화 표현을 구성하고 이를 각각 사용자 발화 메모리, 시스템 응답 메모리로 구성하였음.
- 주어진 사용자 발화에 대한 응답 생성 시, 이중 메모리에서 Attention Mechanism을 통해 어떤 발화에 각각 집중할지 고려하여 시스템 응답 생성한다.

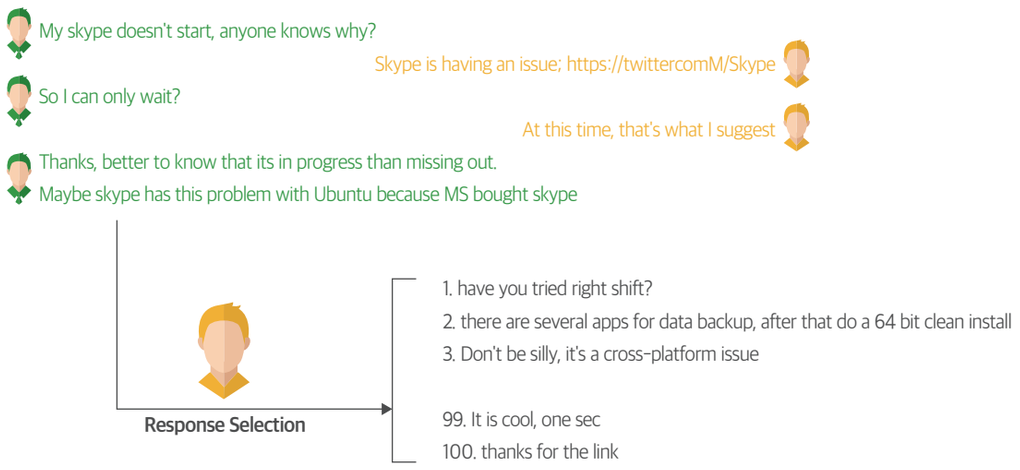
3. 기술 활용 및 응용 분야

- 본 기술은 특정 도메인(e.g. 예약 시스템)의 대화 시스템 구축에 활용될 수 있음

03 검색 기반 대화 시스템에서의 정답 예측 기술

1. 기술 설명

- 검색 기반 대화 시스템이란 대화의 마지막 응답을 후보들(candiates) 중에서 찾아 제공하는 대화 시스템
- 대화 문맥 정보를 활용하여 가장 관련 있는 응답을 찾아 사용자에게 답변을 제공해 주는 것을 목표로 하며, 검색 기반 대화 시스템은 챗봇을 위한 대화 시스템 분야에서 많은 연구가 진행되고 있음



Ubuntu troubleshoot과 관련된 대화와 이에 대한 응답 예측하는 예

2. 기술 방법

- 본 기술은 문장을 효과적으로 표현할 수 있는 LSTM Encoder와 또한 대화의 문맥에서 중요한 부분에 대해 집중적으로 모델에 반영하기 위해 단어 단위의 Attention mechanism을 사용하여 모델을 개발하였음
- 대화 내 발화의 중요 특징(사용자 정보, 발화의 순서, 문장 임베딩)들을 반영하여, 대화 문맥 정보를 더욱 잘 표현할 수 있도록 모델 개발

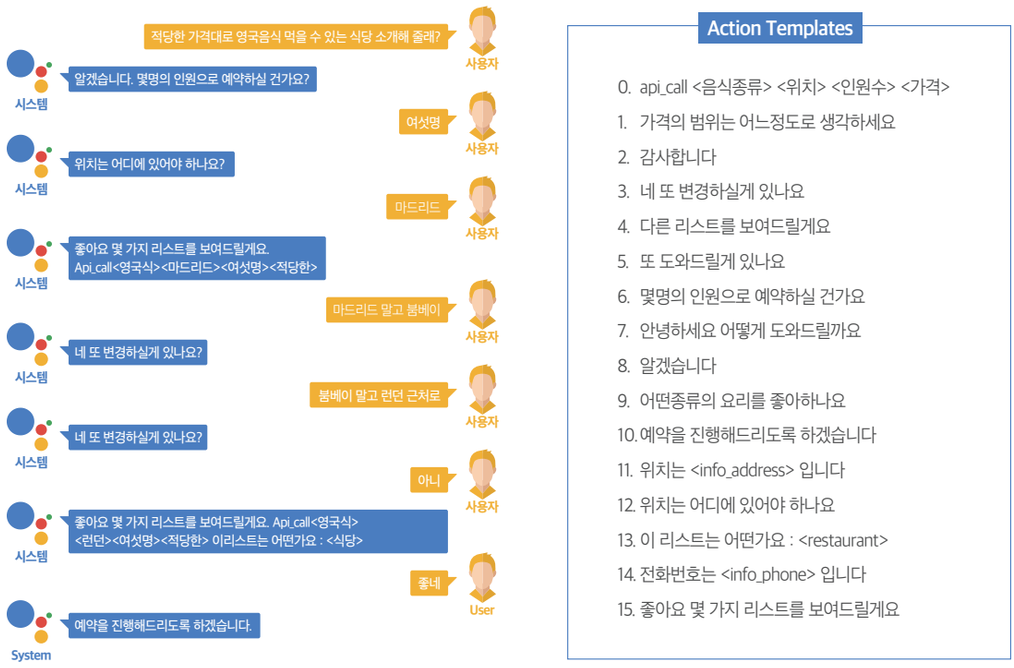
3. 기술 활용 및 응용 분야

- 본 기술은 검색을 기반으로 하는 챗봇 시스템 구축 및 학습에 활용될 수 있으며, 도메인 영역에 관련 없이 활용될 수 있음

04 한국어 특성을 반영한 시스템 액션 템플릿 기반의 대화 시스템

1. 기술 설명

- 목적 지향 대화 시스템은 식당 예약과 같은 특정한 목적을 수행할 수 있는 대화 시스템으로, 다양한 도메인에서 사용될 수 있음
- 한국어 식당 예약 대화 대이터로 모델 학습을 진행하였으며, 입력 부분에서 한국어 특성을 반영한 feature extraction 모듈을 가짐
- 액션 템플릿을 통해 시스템이 응답할 수 있는 답변을 한정하고, 도메인 지식을 반영했기 때문에 적은 양의 데이터로도 대화 모델의 학습이 가능함



Action Templates

0. api_call <음식종류> <위치> <인원수> <가격>
1. 가격의 범위는 어느정도로 생각하세요
2. 감사합니다
3. 네 또 변경하실게 있나요
4. 다른 리스트를 보여드릴게요
5. 또 도와드릴게 있나요
6. 몇명의 인원으로 예약하실 건가요
7. 안녕하세요 어떻게 도와드릴까요
8. 알겠습니다
9. 어떤종류의 요리를 좋아하나요
10. 예약을 진행해드리도록 하겠습니다
11. 위치는 <info_address> 입니다
12. 위치는 어디에 있어야 하나요
13. 이 리스트는 어떤가요 : <restaurant>
14. 전화번호는 <info_phone> 입니다
15. 좋아요 몇 가지 리스트를 보여드릴게요

한국어 식당 예약 시스템과 시스템 액션 템플릿의 예

2. 기술 방법

- Entity extraction, utterance embedding, bag of words 와 같이 3개의 feature를 반영하여 발화 자질을 구성하였고 LSTM으로 인코더를 구성하였음.
- 도메인 특정 지식을 반영하기 위해 시스템 응답 액션 템플릿을 정의하였음. 따라서 일반적인 end-to-end learning 방식의 모델들보다 상대적으로 적은 양의 학습 데이터로 모델의 학습이 가능한 장점이 있음
- 시스템의 응답은 시스템 액션 템플릿 내에 모두 정의되어 있으며, 총 16개의 응답 후보중 가장 적절한 응답을 softmax 확률 분포로 찾아냄

3. 기술 활용 및 응용 분야

- 본 기술은 식당 예약과 같은 특정 도메인의 대화 시스템 구축에 활용될 수 있음. 한국어 대화 데이터로 학습이 가능한 모델로서 다른 도메인으로의 확장이 용이함
- 데모 : http://nlplab.iptime.org:8886/

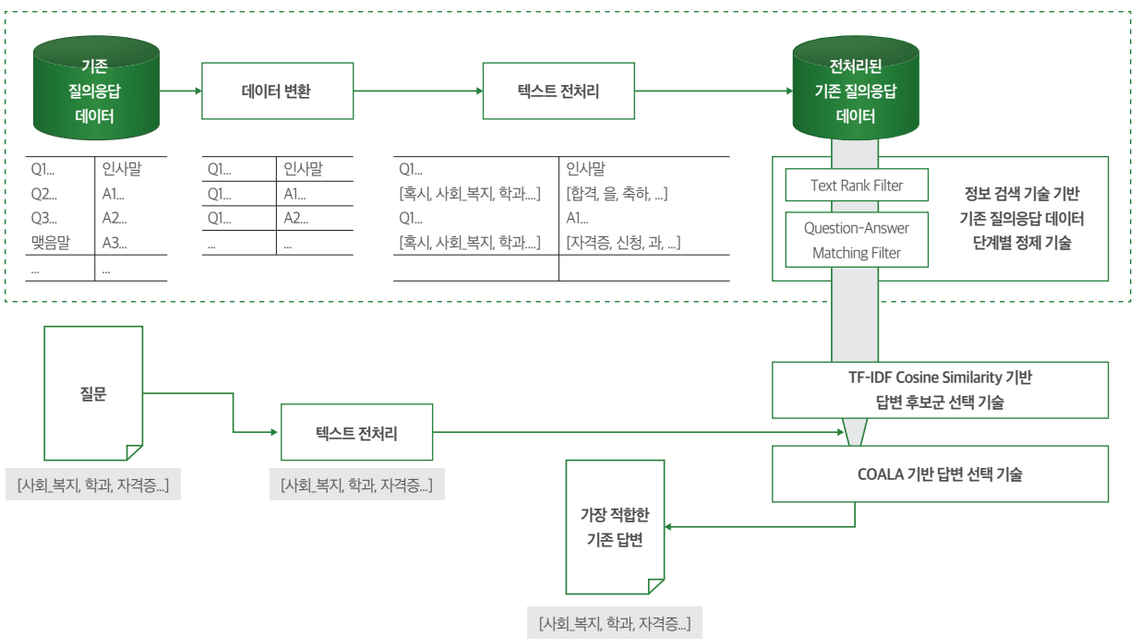
05 딥러닝 기반 자동 질의응답 시스템

1. 기술 설명

- 자동 질의응답 시스템 (챗봇)이란 주어진 질문에 대한 적절한 답변을 자동으로 제시하는 시스템
- 질의응답 방법 중 검색 기반 방법은 기존 질의응답 데이터에서 주어진 질문에 가장 적절한 기존 답변을 선택하여 답변을 제시하는 방법

2. 기술 방법

- 본 기술은 Q&A 게시판 데이터 등 소량의 정제되지 않은 데이터로부터 검색 기반 방법을 적용한 딥러닝 기반 자동 질의응답 시스템 구축



- 챗봇 구축 시 '데이터 전처리 기술'에서 주어진 데이터를 챗봇 기술에 적합하도록 전처리하고, '기존 질의응답 데이터 단계별 정제 기술'에서 정보검색 기술을 적용해 무의미한 질의응답 데이터 제거
- 챗봇 서비스 시 '답변 후보군 선택 기술'에서 TF-IDF feature의 코사인 유사도를 기준으로 가능한 답변 후보군을 선택하고, '답변 선택 기술'에서 딥러닝 기반 최신 답변 선택 모델 COALA를 적용하여 최종 답변 선택

3. 기술 활용 및 응용 분야

- 본 기술은 중소기업 및 개인사업자 등 기존 챗봇 기술에 대한 접근성이 낮은 사용자들에게 최신 챗봇 기술을 보급하고 소비자 상담 효율을 높일 수 있음
- 데모 http://nlplab.iptime.org:32283/

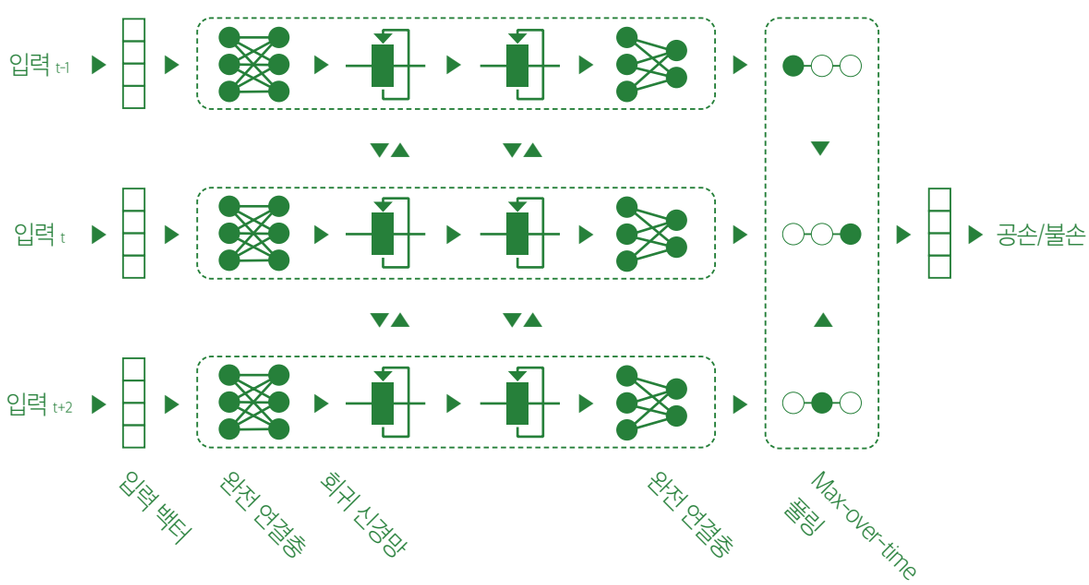
1. 기술 설명

• 본 기술은 인간의 발화가 주어졌을 때, 이의 공손함을 판단하는 시스템이다. 공손함은 언어학에서 광범위하게 탐구된 주제 중 하나로 인간의 언어를 구성하는 핵심적인 요소이며, 전 세계 다양한 문화권에 걸쳐 광범위하게 나타나는 인간 언어의 공통적인 요소 중 하나이다.

2. 기술 방법

기존 연구들은 사용된 기계학습 모델이 단어의 순서와 문맥 정보를 반영하지 못한다는 한계점을 가지고 있다. 본 기술은 각 단어와 그 단어의 문맥 정보를 동시에 반영할 수 있도록 양방향 LSTM(Long Short-Term Memory) 모델과 최근 자연어처리 분야에서 각광받고 있는 BERT 모델을 바탕으로 개발하였다.

• 양방향 RNN을 이용한 문장분류
양방향 회귀 신경망(Recurrent Neural Network, RNN)은 단어를 순차적으로 입력받아 내부의 기억 구조를 활용하여 문맥 정보가 반영된 단어 표상을 생성한다. 본 연구에서는 RNN의 기억 구조를 보강하여 장거리 의존성 문제를 해소한 LSTM을 기반으로 모델을 구성하였다.



• BERT를 이용한 문장분류
BERT(Bidirectional Encoder Representations form Transformers)는 사전 훈련된 모델로, 광범위한 자연어처리 시스템에서 매우 효과적인 모델이다. 기존 연구들에서 공개한 데이터는 딥러닝 모델을 훈련 시키기에 부족하여 데이터가 부족한 상황에서도 효과적으로 동장하는 BERT모델을 사용하였다.

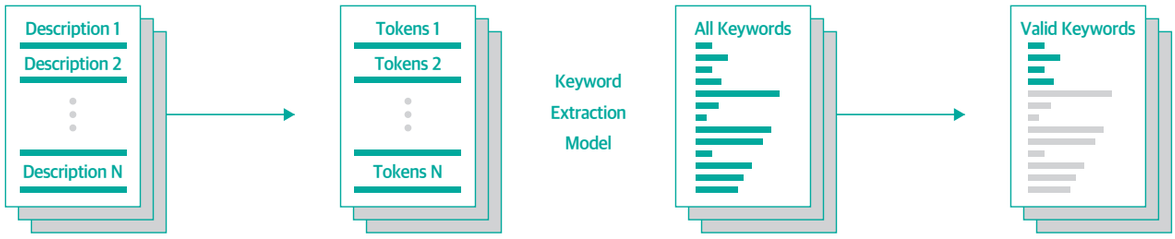
정보 검색/분류/추출/요약 기술

머신러닝 기반 보고서 자동 분석 및 키워드 추출 기술
메타러닝을 응용한 문서 단위의 관계 추출
비정형 위협정보 자동 인식 및 추출
머신러닝을 이용한 문서 자동 요약
딥러닝을 이용한 유사 문서 검색 및 시각화
Automatic Video segmentation based on Narrative
비지도 학습 알고리즘을 이용한 보고서 자동 분석 및 토픽 자동 추출 기술
순차 정보를 이용한 콘텐츠 추천 시스템 개발
지능형(암묵적) 프로파일링 및 추천기술
스케치를 이용한 패션 의류 검색 시스템
Eye tracking 기반의 휴먼 리딩을 반영한 추출 요약 기법
Sentence BERT 임베딩을 이용한 과편향 뉴스 판별
종교활동을 위한 휴머노이드 질의응답 로봇
NAO for Kids Education
온라인 패션 데이터를 활용한 트랜드 분석 및 시각화

01 머신러닝 기반 보고서 자동 분석 및 키워드 추출 기술

1. 기술 설명

- 보고서가 증가함에 따라 사용자가 원하고자 하는 문서를 짧은 시간 내에 판단하여 찾기는 쉽지 않음
- 이러한 문제점을 해결하기 위해 보고서에 대한 핵심 키워드를 자동으로 추출하여 사용자가 선택적으로 볼 수 있으며, 이를 통해 사용자가 효율적으로 원하는 문서를 찾을 수 있도록 키워드 추출알고리즘을 이용함



키워드 추출 알고리즘 모델

2. 기술 방법

보고서 자동 분석 및 키워드 추출 모델

- 본 기술은 정답 셋이 없는 Unsupervised Learning으로 진행되었으며, 보고서에 대해 중요 키워드를 추출하는 것으로 전체 문서를 단어 단위로 추출한 후 단어의 빈도수 계산을 하는 키워드 알고리즘을 통해 중요 단어를 추출함
- 개발한 모델은 각 단어의 가중치를 계산한 후 집단 간 텍스트 특성의 차이나 토큰 사이의 관계 등을 분석하여 상위 적당 K개수의 가중치를 가지는 키워드를 선정하는 연구임

3. 기술 활용 및 응용 분야

- 본 기술은 문서에 대한 정보를 간단한 단어로 추출하므로 키워드 별 문서 검색, 문서 분류, 문서간 유사도에 활용될 수 있음
- 데모 http://nlplab.iptime.org:32270/

02 메타러닝을 응용한 문서 단위의 관계 추출

1. 기술 설명

- 관계 추출의 목적은 구조화되지 않은 정보에서 구조화된 정보를 추출함으로써 입력받은 정보에 있을 수 있는 중의성을 줄이고, 해당 정보를 처리하는데 있어 그 과정을 단순화 하여 처리를 더욱 빠르고 정확하게 분석할 수 있도록 하는 것
- 관계 추출은 크게 2가지 종류로 나뉘는데 전역 수준의 관계 추출(Global Level Relation Extraction)과 문장 수준의 관계 추출(Mention Level Relation Extraction)로 나눌 수 있음
- 해당 연구에서의 목표는 전역 수준의 관계 추출을 하되, 언급 수준의 관계 추출을 병행함으로써 정보의 누락을 최대한 방지하여 성능과 완성도를 유지함

전역관계추출
Global level RE

- 문서와 같은 다량의 정보를 입력 받음
- 연구 사례/데이터양이 적음
- 낮은 성능/높은 가격
- 지식베이스 구축, 문서 요약, 세균-질병 관계 분석 등에 활용됨

단편적관계추출
Mention level RE

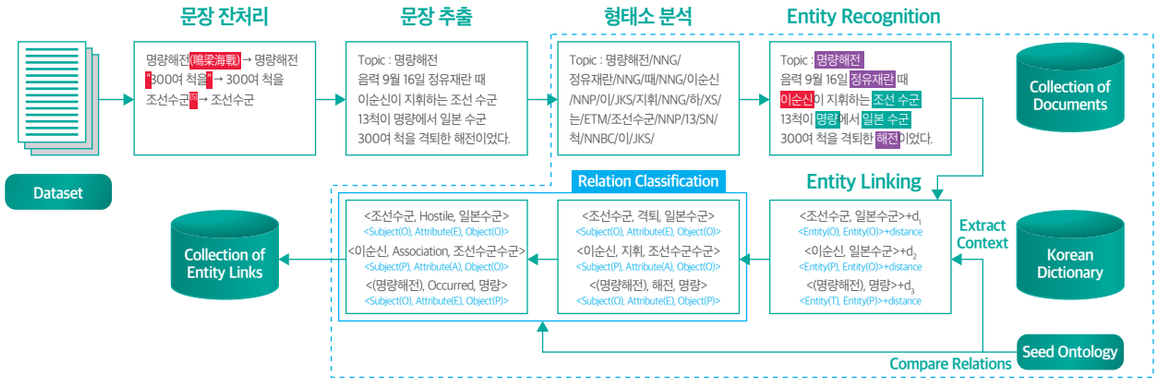
Lucy is going to the park and she is taking the dog for a walk

- 문장 단위의 단편적 정보를 입력 받음
- 연구 사례/데이터양이 풍부함
- 높은 성능/낮은 가격
- Q&A 시스템, 기계학습, 분자 상호작용 분석 등에 활용됨

<관계추출의 종류 및 특징>

2. 기술 방법

- 기존 관계추출 방법은 한국어처럼 주어나 목적어가 자주 생략되는 언어를 다룰 경우에는 추출한 결과가 생략된 주어나 목적어에 해당되는 개체들의 관계를 제대로 표현하지 못한다는 약점도 존재함
- 각 개체 간 관계를 외부 메모리에 저장하고 분석하여 여러 문장에 걸쳐 표현되는 개체간 상호관계를 추출하는 관계추출 모델을 제시함



<관계 추출을 통해 자연어 정보를 구조화되지 않는 정보로 바꾸는 과정>

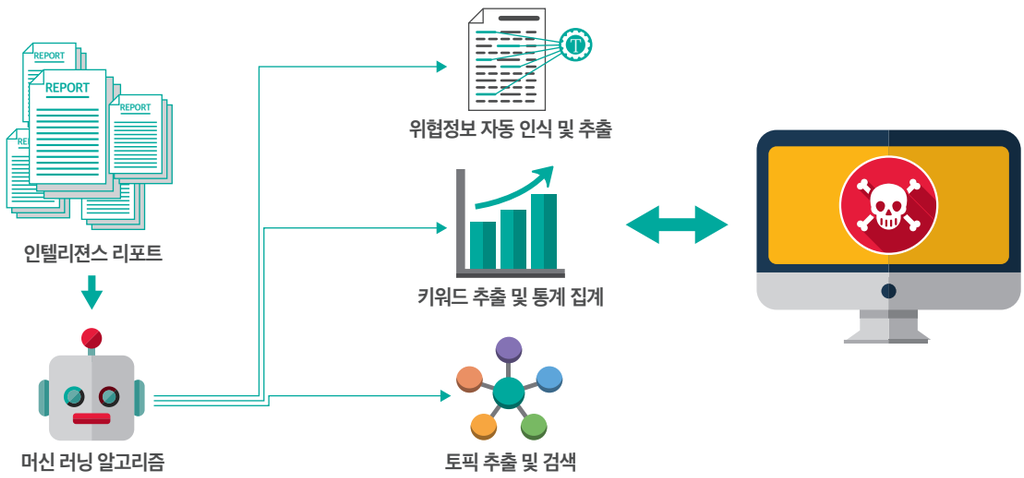
- 모델은 단편적 관계 추출 모델과 외부 메모리 신경망으로 이루어져 있음
- 훈련은 각각 단편적 관계 추출 모델의 훈련, 전역 관계를 위한 메모리 증강 신경망 훈련, 마지막으로 메모리 증강 신경망 훈련의 결과를 반영한 관계 추출 모델의 재훈련으로 총 3단계가 존재함

3. 기술 활용 및 응용 분야

- 기술 활용 및 응용분야로는 Knowledge Base 및 Ontology 자동 구축과 텍스트 문서 및 분자간 관계 정보 요약 및 추출이 존재함
- 본 기술의 단편적 관계추출에 한해서는 데모에서 확인이 가능함
- 데모 nlplab.iptime.org:32277

1. 기술 설명

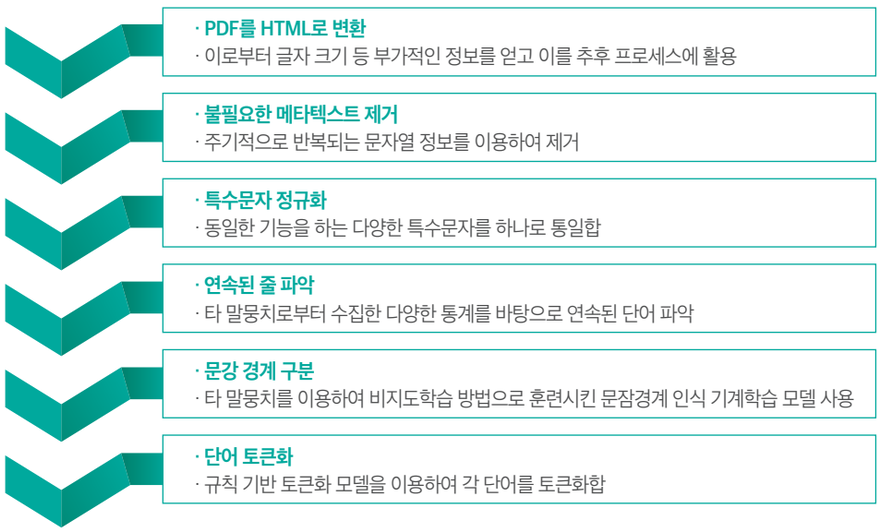
본 기술은 딥러닝 기술인 Long Short-Term Memory(LSTM)-Conditional Random Field(CRF)를 이용하여 인텔리전스 보고서 등 문서 파일 내의 비정형 위협정보를 모델링하고 정형화된 형태로 마이닝하기 위한 것이다.



2. 기술 방법

PDF 문서들을 분석하기 위해서는 문자열로 이루어진 본문을 파일로부터 추출하는 과정이 선행되어야 한다. 하지만 PDF 문서는 단락, 문장, 본문 등의 구분이 없으며, 각 글자의 글씨체, 크기와 위치 정보만 담겨 있다. 따라서 PDF 문서를 분석하여 텍스트를 일관성 있게 추출하고, 기계학습 모델에 사용할 수 있도록 이를 문장 단위로 구분하고 토큰화하는 과정이 선행되어야 한다. 이를 위해 기계학습, 정규표현식, 위키피디아 문서 통계를 활용한 하이브리드 문장경계 인식 기술을 개발하여 사용하였다.

추출된 텍스트에 대해서 양방향 LSTM-CRF 모델을 이용하여 위협정보를 추출한다. 해당 모델의 훈련은 지도학습 방법을 이용하였으며, 이를 위해 수백 건의 인텔리전스 리포트를 수집하여 이 중 백여 건의 리포트에 대해 수작업 태깅으로 학습 말뭉치를 구축하였다.



[그림 52] PDF2TXT 과정.

3. 기술 활용 및 응용 분야

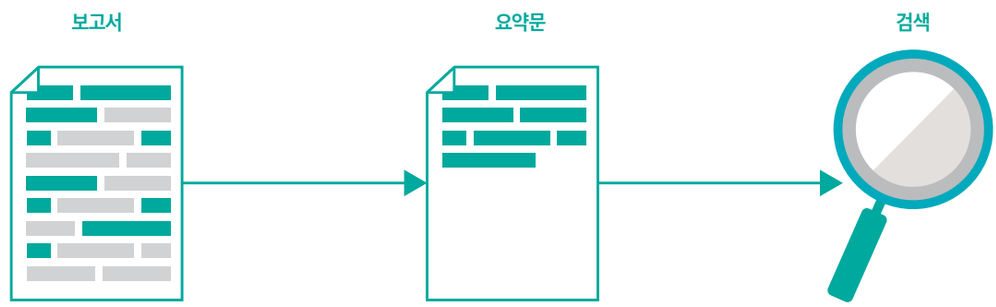
리포트 자동 분석 (타 분야 문서로 적용 가능)

데모 시스템 : http://nlplab.iptime.org:32270/kisa_demo

04 머신러닝을 이용한 문서 자동 요약

1. 기술 설명

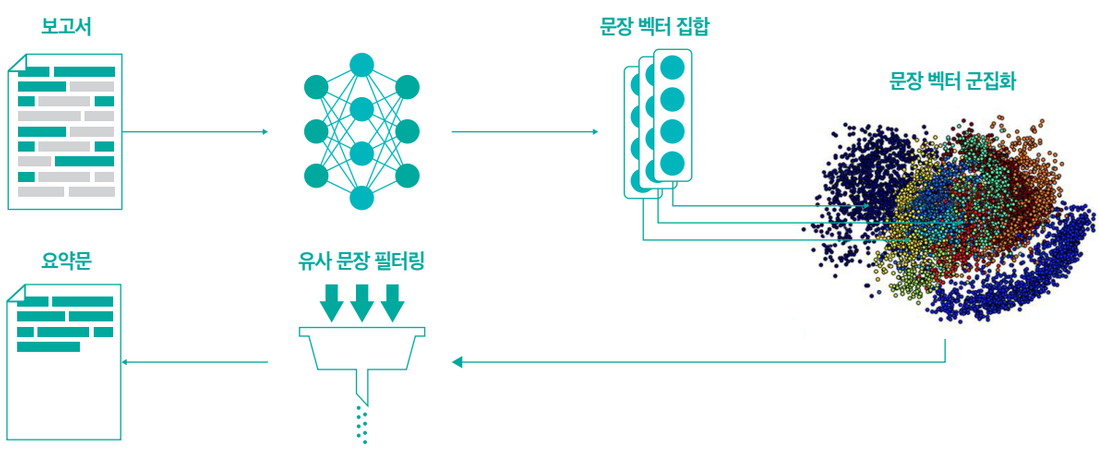
본 기술은 비지도 학습 알고리즘을 바탕으로 문장 추출에 의한 자동 문서 요약 방법이다. 특히, 본 기술은 특정 언어나 문서 특징에 의존하지 않으므로 확장이 용이하다.



2. 기술 방법

본 기술은 비지도 학습 알고리즘인 K-means clustering을 사용한다. 기본 가정은 비지도 학습 알고리즘을 이용하여 비슷한 아이디어(문장)를 클러스터링할 수 있다는 것이다. 이후 요약을 생성하기 위해 가장 대표적인 문장이 각 클러스터에서 선택된다. 또한, 이 방법을 사용하면 생성된 요약의 단어 수를 어느 정도 제어할 수 있다는 장점이 있다.

본 기술의 문서 요약 시스템은 문장 벡터 생성 시 기존의 TF-IDF 방법을 이용한 벡터 생성이 아닌, 딥러닝 방법을 사용한다. 이는 단어 불일치 문제 등을 해결할 수 있다는 장점이 있다. 문장 벡터 생성 후 요약 기술은 클러스터링 기반 추출 요약 방법을 사용한다.



3. 기술 활용 및 응용 분야

정보 검색, 자동 요약

데모 시스템 : http://nlplab.iptime.org:32270

05 딥러닝을 이용한 유사 문서 검색 및 시각화

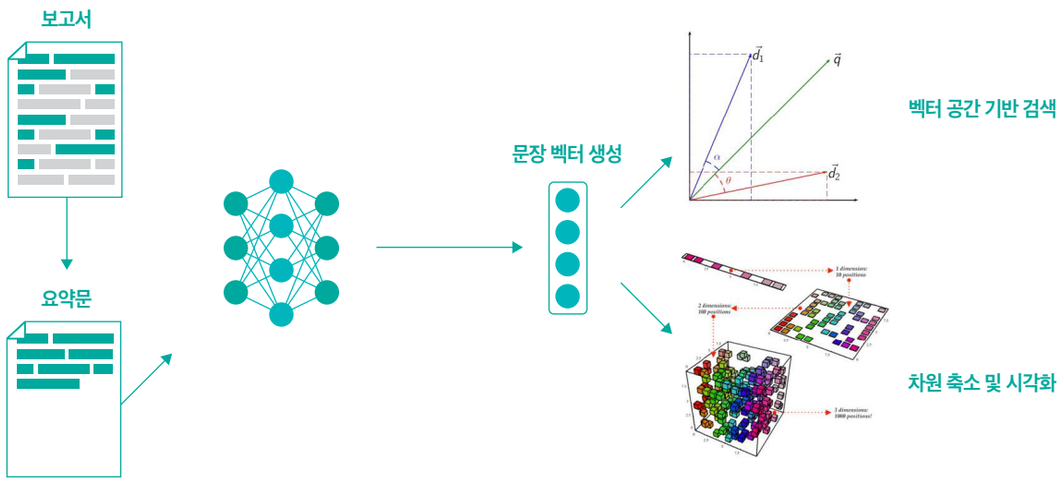
1. 기술 설명

본 기술은 문서를 가상의 벡터 공간에 투사하고 그 차원을 축소한 후, 이를 시각화하여 지능적으로 유사 문서를 탐색할 수 있는 방법이다.

2. 기술 방법

문서를 가상의 벡터 공간에 투사하면, 벡터 공간 모델을 이용하여 문서 간의 유사도를 수치화 할 수 있고, 이로부터 유사 문서 검색이 가능해진다. 문서를 벡터 공간에 임베딩하고 검색 등을 수행하기 위해서는 문서를 고정 길이의 벡터로 표현할 수 있어야 한다. 본 기술에서는 문서 임베딩을 생성하기 위해 본 연구실이 보유 중인 문장 임베딩 기술과 문서 자동 요약 기술을 응용하였다.

여기서 더 나아가, 문서가 투사된 벡터 공간을 t-distributed Stochastic Neighbor Embedding(t-SNE)와 같은 차원 축소 기법을 이용하면 이를 인간이 시각적으로 인지할 수 있는 공간인 3차원 이하로 변형할 수 있고, 이를 시각화하여 검색 인터페이스로 응용 가능하다. 이를 위해 Tensorboard를 활용하였다.



3. 기술 활용 및 응용 분야

정보 검색, 문서 분류

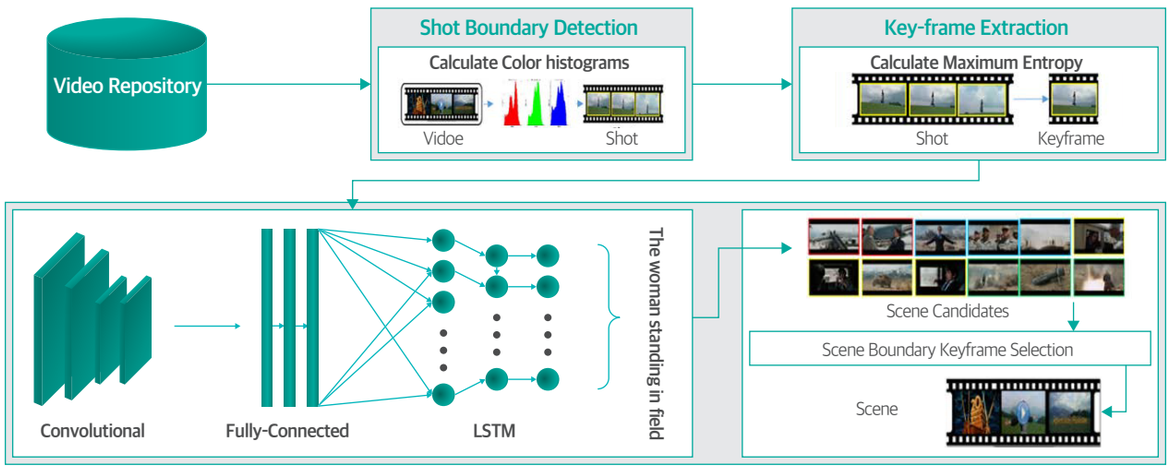
1. 기술 설명

• 최근 동영상 이해(Video Understanding)에 대한 연구는 다양한 분야에서 이루어지고 있다. 해당 연구에서는 이러한 비디오 이해의 전처리 과정으로써 입력받은 비디오를 의미적으로 통일성을 지니는 단편적인 영상으로 나누는 것을 목표로 함

2. 기술 방법

• 의미적으로 통일성을 지니는 단편적인 영상 감지를 진행하기 위해서는 먼저 비디오를 장면 단위로 나눔

• 실질적으로 영상을 장면단위로 모두 처리하는 것은 실질적으로 너무나 많은 연산과 비용을 소요하기 때문에 장면단위로 나눈 영상을 각각 분석하여 해당 장면을 대표할 이미지를 찾음



• 이미지로부터 정보를 추출하여 의미적으로 연결된 shot들을 판별하여 의미적으로 통일된 Scene들의 집합으로 다시 조합함

[그림 62] Video Scene Detection 모델 구조도

3. 기술 활용 및 응용 분야

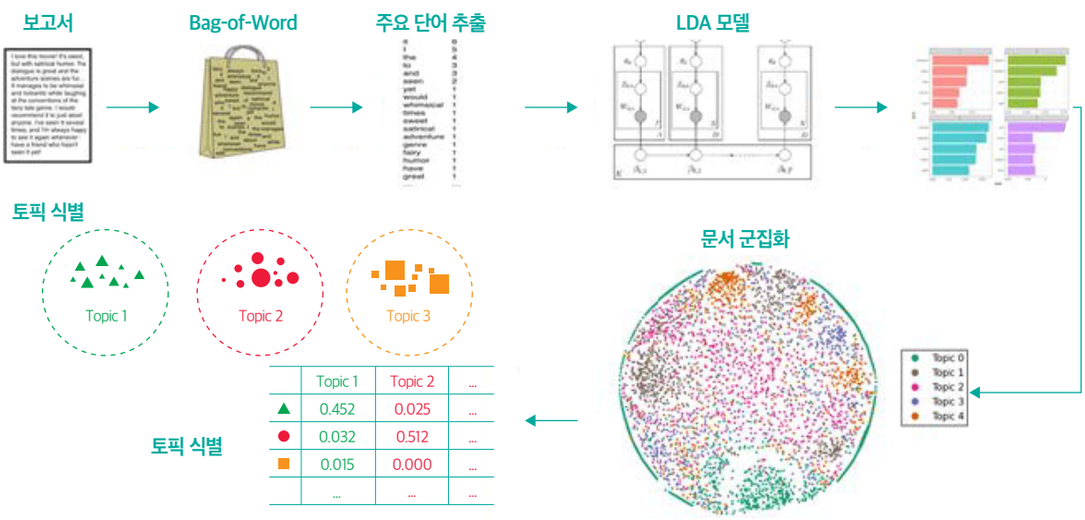
o 동영상 이해를 위한 자동적인 전처리 과정으로 동영상 자동 분할 시스템을 이용하여 자동적인 영상분할을 통하여 야구, 축구와 같은 동영상으로부터 하이라이트를 분리하여 추출할 수 있음

07 비지도 학습 알고리즘을 이용한 보고서 자동 분석 및 토픽 자동 추출 기술

1. 기술 설명

- 토픽 모델(Topic model)이란 문서 집합의 추상적인 "주제"를 발견하기 위한 통계적 모델중 하나로, 텍스트 본문의 숨겨진 의미구조를 발견하기 위해 사용되는 텍스트 마이닝 기법임
- 본 기술은 해당 보고서가 어떤 토픽에 적합한지 파악하기 위해 토픽 모델링 기법 가운데 하나인 잠재 디리클레할당(Latent Dirichlet Allocation, LDA)를 이용함. LDA는 주어진 문서에 대하여 각 문서에 어떤 주제들이 존재하는지에 대한 확률모형이며, 토픽별 단어의분포, 문서별 토픽의 분포를 모두 추정함

2. 기술 방법



보고서 자동 토픽 추출 기술 모델

- 본 기술은 보고서 PDF 파일을 넣으면 분석이 쉽도록 txt로 전환하고, Bag-of-words를 이용하여 전체 보고서에서 중요한 단어 최소 5000개를 사전으로 생성함
- 만들어진 사전을 바탕으로 새로운 문서가 들어왔을 때 토픽 모델 알고리즘인 LDA를 활용하여 문서별 토픽 분포 확률을 계산함

3. 기술 활용 및 응용 분야

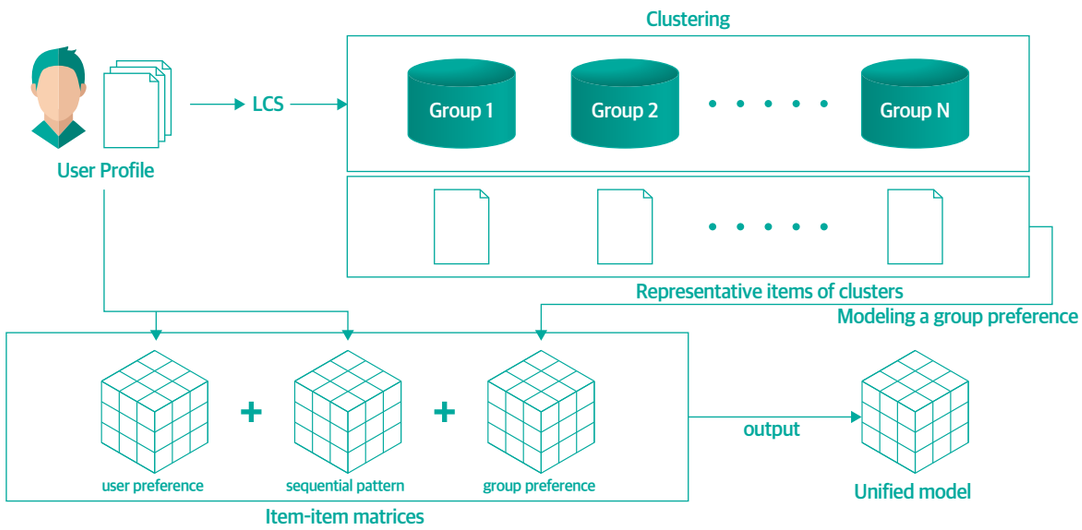
- 본 기술은 방대한 자료에서 자동으로 비정형 텍스트 집합을 이해하기 쉽도록 정리할 수 있으므로 텍스트마이닝 분야 외에도 유전자 정보, 이미지, 네트워크와 같은 자료에서 유의미한 구조를 발견하는데에도 유용하게 사용될 수 있음
- 데모 http://nlplab.iptime.org:32270/

08 순차 정보를 이용한 콘텐츠 추천 시스템 개발

1. 기술 설명

- 추천 시스템은 사용자가 소비할 만한 콘텐츠 또는 아이템을 예측하여 사용자에게 콘텐츠를 제시해주는 시스템을 말함
- 해당 기술은 사용자의 소비 순서 정보를 통하여 순차 패턴을 모델링하고, 사용자들의 유사도를 통해 그룹 선호도 모델을 모델링함으로써 사용자들에게 순차적인 콘텐츠 또는 아이템을 추천해주는 기술임
- 기존 연구와의 차이점은 그룹 선호도를 유사도 모델로 정의하고, 사용자의 선호도와 순차패턴, 그룹 선호도를 하나의 단일 모델로 통합하여 모델의 차원을 축소하여 기존 연구들의 추천 성능보다 더 향상된 추천 모델을 제안하였음

2. 기술 방법



- 사용자와 사용자가 소비한 정보가 주어졌을 때, 사용자가 소비한 콘텐츠 또는 아이템의 순서 정보와 그 유사도를 통하여 사용자들의 그룹을 추출하고, 그룹들의 대표 아이템 셋을 정의하여 그룹의 선호도 모델을 하나의 행렬로 모델링함
- 사용자가 소비한 정보를 통하여 특정 사용자의 선호도 모델과 순차 패턴을 각각을 행렬로 모델링함
- 사용자 선호도, 순차패턴, 그룹 선호도를 통합하여 하나의 행렬로 모델링하고, 해당 모델을 기계학습 방법론으로 학습하여 사용자에게 순차적인 소비가 가능하도록 아이템 또는 콘텐츠를 예측하여 제시함

3. 기술 활용 응용 분야

- 해당 기술은 사용자들에게 영화를 추천해주는 시스템, e-커머스 환경에서의 상품 추천, 사용자 선호에 맞는 음악 추천 등 다양한 도메인에 적용하는 것이 가능하다.
- 인공지능 서비스의 대다수의 마지막 단계는 추천으로 인공지능 서비스와 연계하여 활용하는 것이 가능하다.
- 데모 http://nlplab.iptime.org:32280/rec_demo/

09 지능형(암묵적) 프로파일링 및 추천기술

1. 기술 설명

• 방대해지고 있는 온라인 시장에서는 소비자도 자신이 원하는 니즈에 대해 키워드 검색으로 원하는 것을 일일이 찾기란 쉽지 않은 일이다. 이를 해소해줄 수 있는 것은 소비자의 니즈를 반영한 지능형 추천이다. 기존 온라인 구매 사이트는 소비자의 니즈를 파악하고추천하기 위하여 설문조사 형식으로 소비자의 선호 상품 니즈를 파악하는 것이 대부분이었다. 본 기술에서는 기존 방법의 한계점을 해소하고자 지능형(암묵적) 프로파일링 방법을 통하여 소비자들의 니즈와 선호하는 것에 대해 간편하고 효과적으로 파악할 수 있는모델을 제안하였다. 또한 이렇게 수집된 데이터로 학습한 딥러닝기반의 지능형(암묵적)추천 모델을 통하여 이미지 자체에 대한 특성을 반영하도록 학습하는 방법을 개발하였다.

2. 기술 방법

• 본 기술에서는 사용자 프로파일링을 통하여 사용자가 선호하는 상품 이미지에 대해 지능형(암묵적) 프로파일링 방법을 통해 선호 상품 데이터를 수집하며, 이미지 자체에 대해 특성이 반영된 이미지 벡터값을 이용하여 Deep neural network기반의 선호 상품과 비선호 상품에 대한 각각의 $score^{+}$, $score^{-}$ 값을 도출하여 maximization loss를 적용하여 학습을 진행한다.

3. 기술 활용 및 응용 분야

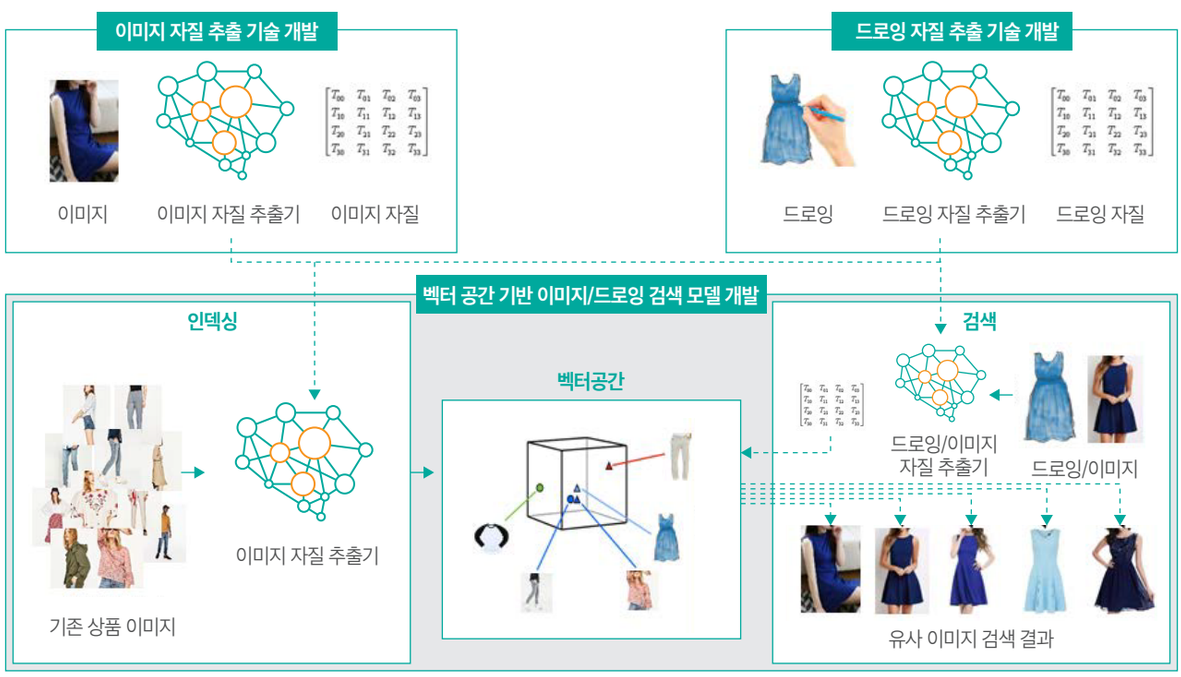
• 본 기술은 이미지가 존재하는 상품을 대상으로 추천기술을 제공할 수 있다. 이를 기반으로 의류 추천, 얼굴형에 맞는 안경 또는 선글라스 추천 등의 기술에 응용될 수 있다.

• 본 기술을 이용한 실적용 예시는 남성 의류 추천에 해당 기술을 응용하였다. 해당 기술은 크게 두 가지의 추천 모델로 구성된다. 첫 번째는 상의 및 하의에 대한 코디 추천 기술이며, 두 번째는 사용자가 업로드하는 이미지(상의 또는 하의)에 대해 어울리는 하의 또는 상의를 추전해주는 기술이다.

• 소개 영상 https://drive.google.com/file/d/1NccHDRAr_yM6XHTyPLdF1oBAVxLGjIjv/view?usp=sharing

1. 기술 설명

본 기술은 사용자가 원하는 상품의 스케치를 그리면, 이를 바탕으로 유사한 시각적 특성을 가진 상품을 검색하는 방법이다.



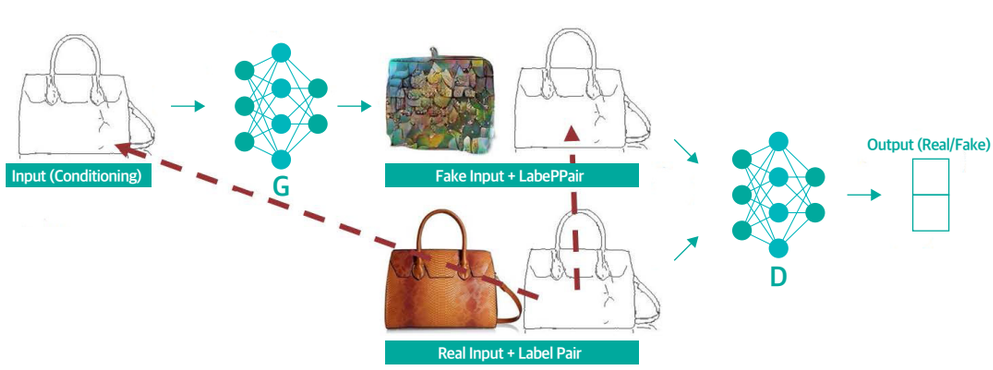
[그림] 벡터 공간 기반 이미지/드로잉 검색 모델의 구조도

2. 기술 방법

스케치 기반 상품 검색 시스템은 사용자가 원하는 상품의 스케치를 그리면 딥러닝 기술을 이용하여 이를 이미지 수준으로 업샘플링하고, 업샘플링된 이미지로부터 얻은 자질 벡터로 벡터공간 기반 유사 이미지 검색을 수행하는 방법을 사용한다.

사진 기반 상품 검색을 위해 이미지 자질 벡터를 추출할 수 있는 CNN(convolutional neural network) 모델을 훈련시켜야 한다. 이를 위해 패션 상품의 카테고리를 분류할 수 있는 이미지 분류기를 훈련시켜 활용한다.

스케치 기반 상품 검색을 위한 스캐치 업샘플링은 GAN(Generative Adversarial Network)을 이용한다. GAN은 상호 대립되는 두 신경망을 교차로 훈련시키는 생성 모델로, 이미지 생성분야에서 기존의 방법보다 선명한 결과물을 얻을 수 있어 최근 각광받고 있다.



<그림> Generative Adversarial Network을 이용한 스케치 업샘플링 모델의 구조도

3. 기술 활용 및 응용 분야

정보 검색, 유사 상품 검색, 스케치를 이용한 모조 상품 검색

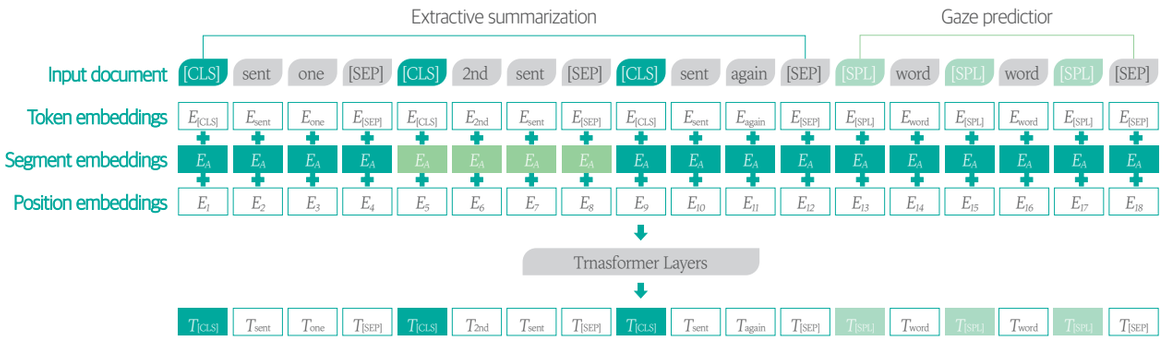
데모 시스템 : http://nlplab.iptime.org:32280/fashion_demo/

11 Eye tracking 기반의 휴먼 리딩을 반영한 추출 요약 기법

1. 기술 설명

- 추출요약이란 문서내에 주요한 요약정보가 되는 문장 또는 단어를 추출하여 요약을 생성하는 기법을 의미한다. 본 기술은 휴먼 리딩(Human reading)을 위한 인지처리과정을 위해 아이트래킹(Eye tracking) 데이터 기반의 추출 요약(Extractive summarization) 기술로서 기존의 귀납적 편향을 해소하기 위하여 아이트래킹 데이터 기반의 새로운 추출 요약 모델이다.

2. 기술 방법



본 기술은 사전학습 언어 모델인 BERT를 기반으로 문장과 단어 정보를 모두 반영하는 구조이다. 또한 본 모델은 텍스트 요약을 수행할 때 사람의 인지처리 과정을 모방하여, 아이트래킹 데이터를 기반으로 사람의 사전지식을 귀납적 편향으로 사용하여 기존의 문제점을 해소하였다.

본 모델은 요약 문서의 문장 데이터와 아이트래커를 통하여 실험한 문장 데이터로 서로 다른 독립적인 태스크를 수행하기 때문에 다중 도메인 학습(Multi domain learning)으로 정의할 수 있으며, 아래와 같은 구조를 가진다.

- 다중 단어 및 문장 인코딩 : 문장과 단어에 대한 인코딩 정보를 동시에 사용하여 각 문장에 대한 문맥 임베딩(Contextual embedding)을 반영하고, 단어에 대한 아이트래킹 정보를 활용한다.
- Segment embeddings : 문서내에 있는 다중 문장들을 구문한다.
- Fine-tuning with multi-domain unified layer: 서로 대른 두 개의 태스크(task)를 수행할 수 있도록 통합된 다중 도메인 레이어로 구성되며, 첫 번째 요약(Summarization)파트에서는 추출 요약을 수행하며, 두 번째 시선 예측(Gaze)파트에서는 토큰에 대한 first pass prediction을 수행한다.

12 Sentence BERT 임베딩을 이용한 과편향 뉴스 판별

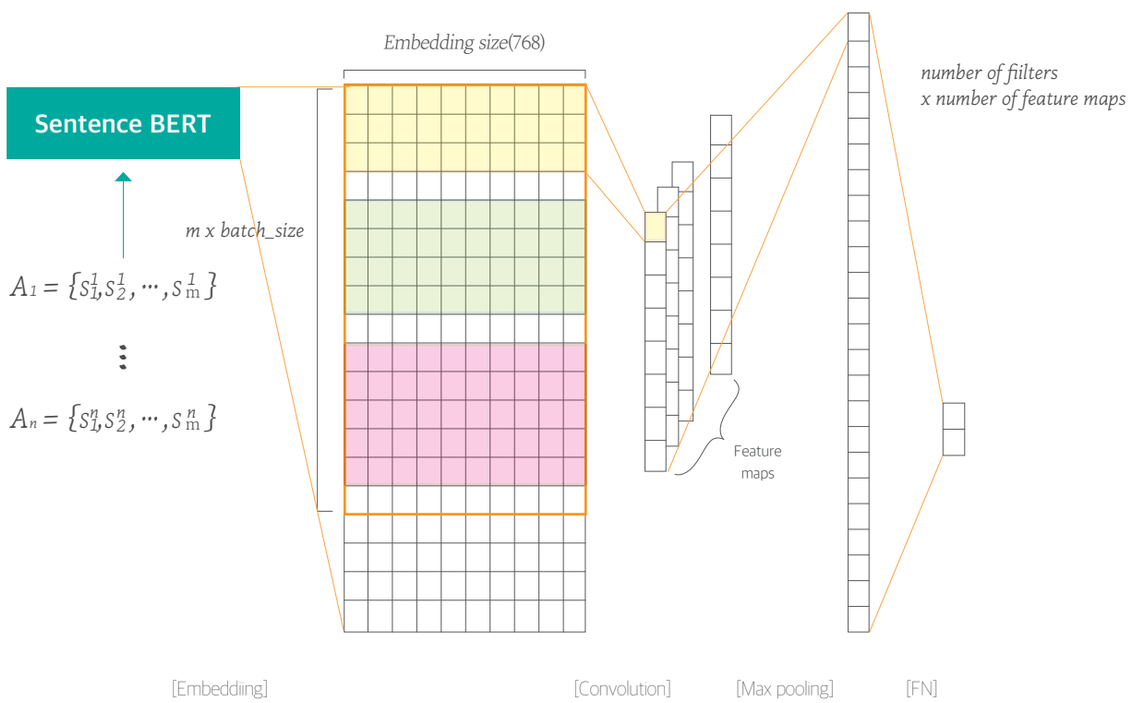
1. 기술 설명

과편향 뉴스는 주어진 기사 내용이 비논리적이거나 특정한 사람이나 정당에 편향되어 있는 뉴스를 의미한다. 본 기술은 과편향 뉴스 판별(hyperpartisan news detection)모델로서 뉴스 기사가 특정 인물 또는 정당에 편향되었는지 판단하는 모델이다. 기존 연구들은 feature-based ELMo, CNN 모델이 사용되었으나 이는 문서 임베딩이 아닌 단어 임베딩의 평균을 사용하는 한계가 있다. 따라서 feature-based 접근법을 따르며 Sentence-BERT(SentBERT)의 문서 임베딩을 이용한 feature-based SentBERT기반의 과편향 뉴스 판별 모델을 개발하였으며, 본 모델은 기존 state-of-the-art 모델보다 f1-score 기준 1.3%높은 성능을 보였다.

2. 기술 방법

기존의 BERT 임베딩 대신 pre-trained BERT로부터 의미적으로 유의한 문장 임베딩을 추출할 수 있도록 수정된 모델인 SentBERT모델을 사용한다. SentBERT모델을 통하여 추출된 문장 임베딩은 코사인 유사도를 통해 비교가 가능하며, 고정된 사이즈의 문장 임베딩을 얻기 위해 다음과 같이 학습된다.

BERT output 벡터의 평균값을 구한 뒤, 생성된 문장 임베딩의 의미적 유의성을 코사인 유사도로 계산한다. 그 후 siamese network 혹은 triplet network가 임베딩의 weight를 업데이트 시킨다. 이에 따라 산출된 임베딩은 기존의 BERT임베딩과 다르게 의미적으로 유사한 문장들은 벡터 스페이스 안에서 그 거리가 가까워져 기존의 BERT임베딩보다 의미적 정보를 잘 담을 수 있다.



13 종교활동을 위한 휴머노이드 질의응답 로봇

1. 기술 설명

본 기술은 시각 장애인, 노인 등 텍스트에 접근하기 어려운 사람들에게 로봇의 음성으로 도움을 제공하기 위하여 개발되었으며, 한국어/영어가 지원된다.

• 종교 개인 비서 로봇의 역할
- 여러 가지 이유로 경전을 읽을 수 없는 사람들에게 음성으로 내용 제공 가능
- 복음, 장, 절 단위에 구애받지 않고, 듣고 싶은 부분 검색 가능
- 집에서 종교음악을 듣고 싶어도 여러 이유에 의뢰 할 수 없는 사람들에게 도움
- 비슷한 구절을 기반으로 추천하여 관련된 노래와 또 다른 구절 검색 가능
- 전문 종교인이 아닌 일반 신자들에게 편리한 접근성 제공
- 이를 통하여 종교인들의 심리적 웰빙과 긍정적 정서 함양에 도움

2. 기술 방법

• 성경검색 모델

- 사용자가 듣고 싶은 성경의 범위를 로봇에게 질의, 로봇이 해당 범위를 낭독함

- Rule-based로 구현하였으며, 질의로 들어온 성경의 범위 인덱스를 추출하여 성경을 낭독함

- 검색예시

• 요한복음
• 마태복음 1장
• 출애굽기 들려줘
• 마태복음이랑 마가복음 들려줘
• 시편 1장 2절 들려줘
• 창세기 1장 1절부터 2장까지 들려줘
• 잠언 2장 1절부터 3절까지 들려줘
• 누가복음 1장 1절부터 1장 15절까지 들려줘

성경 검색, 찬송가, 구절 검색 중에 골라주세요.

성경 검색

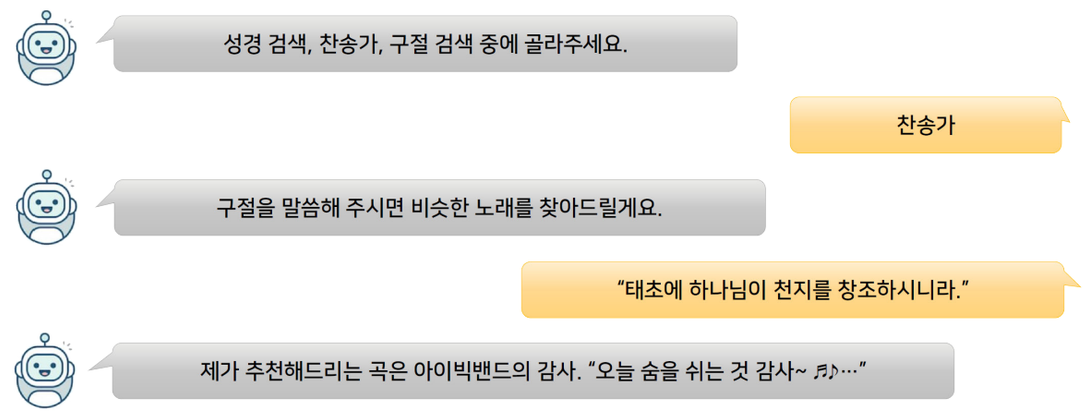
어떤 구절을 듣고 싶으십니까?

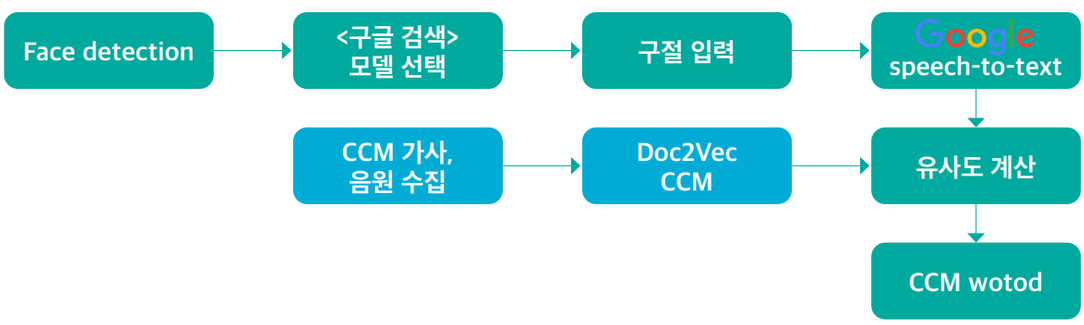
창세기 1장 1절부터 5절까지 들려줘.

"태초에 하나님이 천지를 창조하시니라…<중략>…"

• CCM 추천 모델

- 사용자가 특정한 구절을 로봇에게 질의하면, 해당 구절과 비슷한 내용의 CCM을 검색 및 추천
- Gensim의 Doc2Vec모델을 이용하여 하나의 CCM 가사를 하나의 문서로 분류하고, 분류된 문서를 300차원 벡터로 변환함
- 로봇이 입력값으로 하나의 구절을 받으면 문서간 유사도 계산을 통하여 입력과 가장 유사한 곡으로 추천 및 재생함

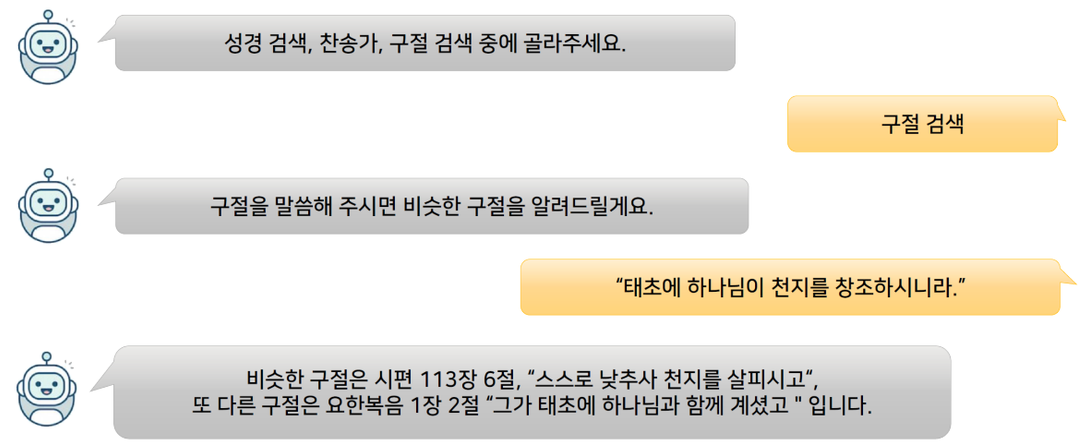


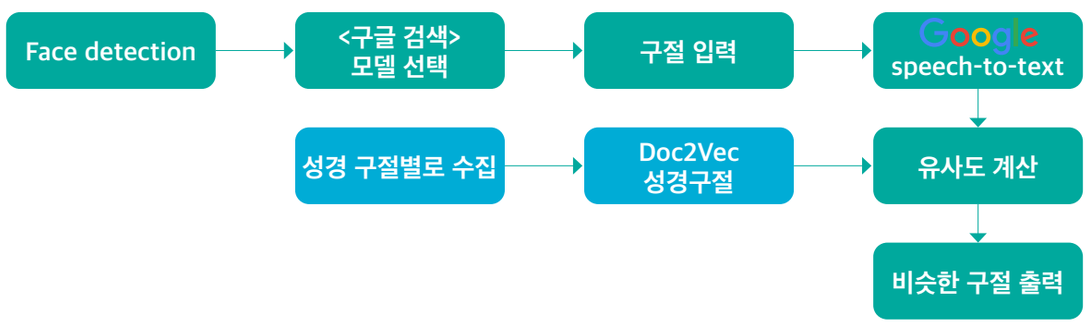


• 비슷한 구절 검색 모델

- 사용자가 특정한 구절을 로봇에게 질의하면 해당 구절과 비슷한 내용의 또 다른 구절을 검색 및 추천함

- 알고리즘은 CCM 추천 모델과 동일함





3. 기술 활용 및 응용 분야

교회 예배 후 포럼 활동 / 개인 예배 활동 보조 가능

종교에도 적용이 가능.

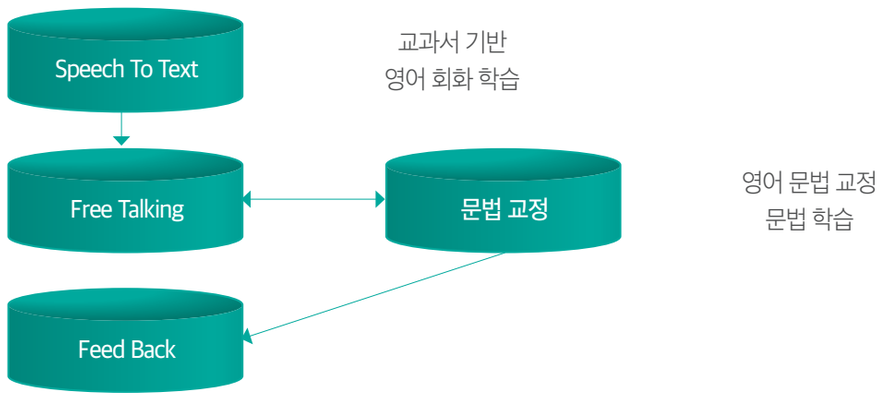
1. 기술 설명

본 기술은 외국어 학습을 목적으로 개발하였으며, 시나리오 기반 Free Talking, 영어문법 교정 피드백, 사용자들의 흥미 유발을 위한 언어지능, 외국어 지능, 수리지능 게임을 개발하였다.

2. 기술 방법

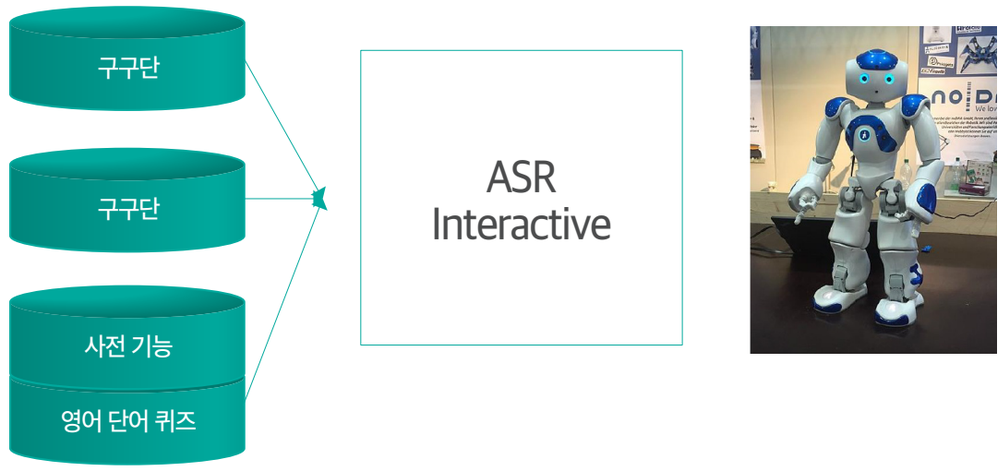
• 시나리오 기반 Free Talking

- 초등학교 저학년 대상 교육용 로봇으로 초등학교 교과서 기반으로 20개의 시나리오를 생성함
- 딥러닝 기반 영어 문법 교정기를 개발 및 적용하여 사용자와 로봇이 대화를 나눈 뒤, 로봇이 사용자의 영어 문법을 교정하여 알려줌



• Intelligent games

- 한국어 기초 사전을 기반으로 자체 한영사전을 제작하였으며, 자체 한영사전을 바탕으로 영어사전과 학습용 미니게임을 개발함
- 한영사전은 파이썬 딕셔너리 형태로 제작되었으며, 학습 대상의 수준을 고려하여 초급, 중급 어휘로 구성함. 또한 원활한 음성인식을 위하여 동음이의어를 다의어로 취급함
- 예) {'먹다': 'eat', 'be deaf', '가격':'hitting','price'}



· 수리지능을 위한 구구단 게임

- 영어로 구구단 게임을 진행할 수 있으며, 영어와 수학을 동시에 학습하는 효과를 가짐
- 게임 옵션: 중도취소, 다시 듣기, 게임횟수 설정, 점수 산정

· 언어지능을 위한 끝말잇기 게임

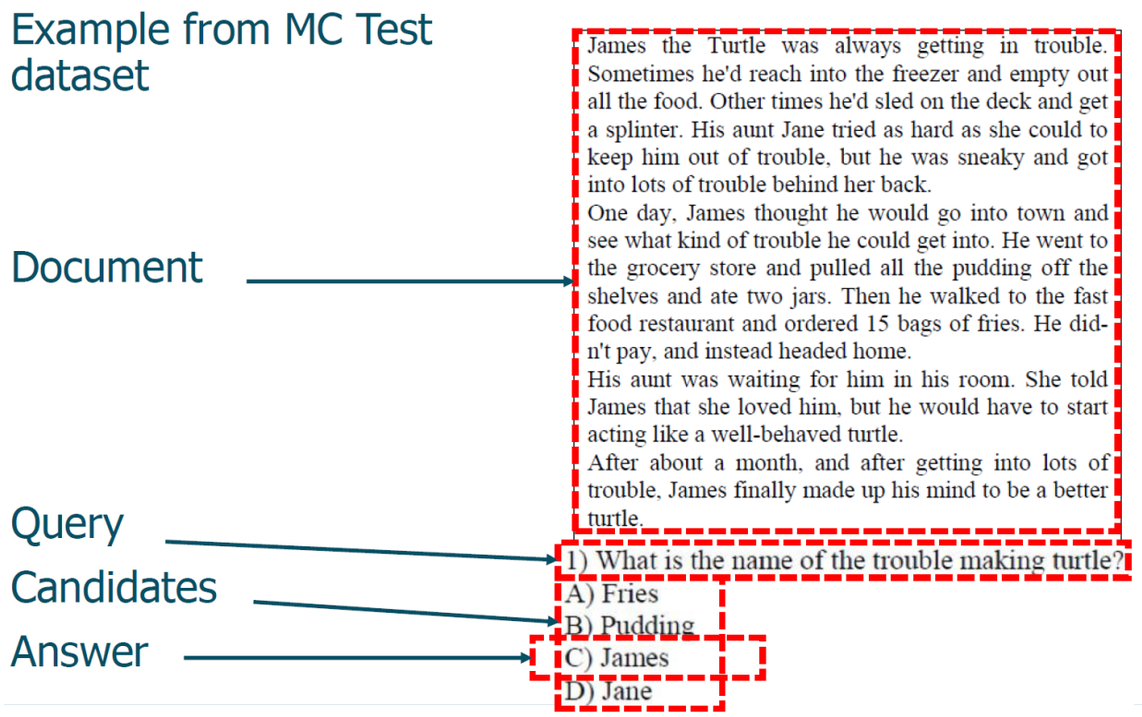
- 한국어 단어로 끝말잇기 게임을 할 수 있으며, 한국어 학습에 도움을 줌
- 게임 옵션: 중도취소, 다시 듣기, 게임횟수 설정

· 외국어 지능을 위한 영어 단어 게임

- 로봇이 한국어 단어와 영어 보기를 제시하면, 사용자가 보기 중 알맞은 영어 단어를 맞추는 게임으로 영어 단어 학습에 도움을 줌
- 게임 옵션: 중도취소, 다시 듣기, 게임횟수, 객관식 항목 수 설정

• Interactive Machine Reading Comprehension

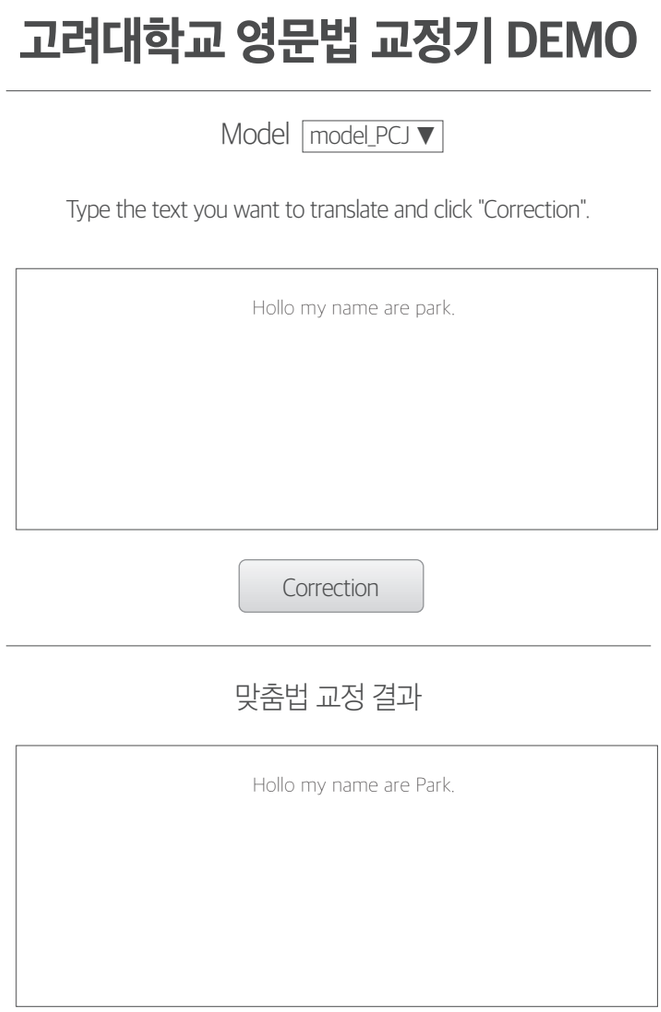
- 기계독해(MRC, Machine Reading Comprehension)란 인공지능 알고리즘이 스스로 문제를 분석하고 질문에 최적화된 답안을 찾아내는 기술이다.
- 본 기술은 사용자-로봇의 대화를 통하여 기계독해가 가능하도록 개발하였으며, 10초동안 사용자가 로봇에게 이야기를 들려주고, 로봇에게 이야기와 관련된 질문을 하면 로봇이 사용자의 이야기에서 정답을 추론하여 답을 한다.



3. 실행 결과

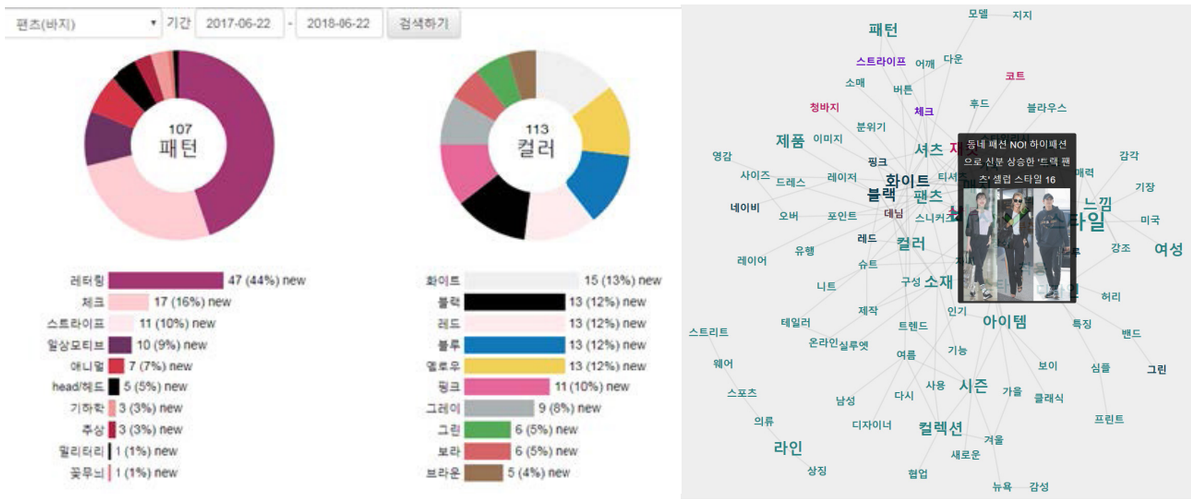
• 딥러닝 기반 영어 문법 교정기

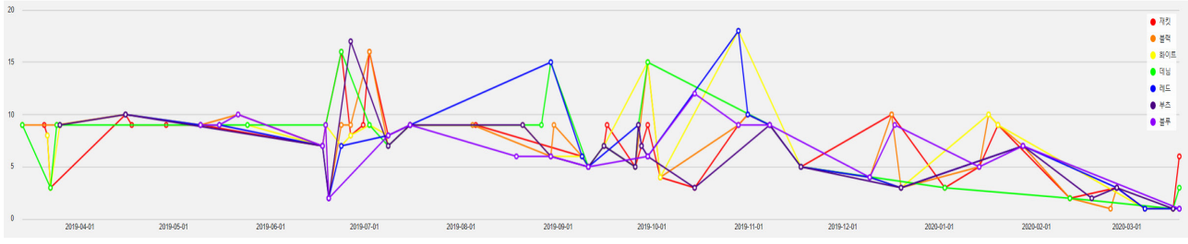
• URL: http://nlplab.iptime.org:32292/



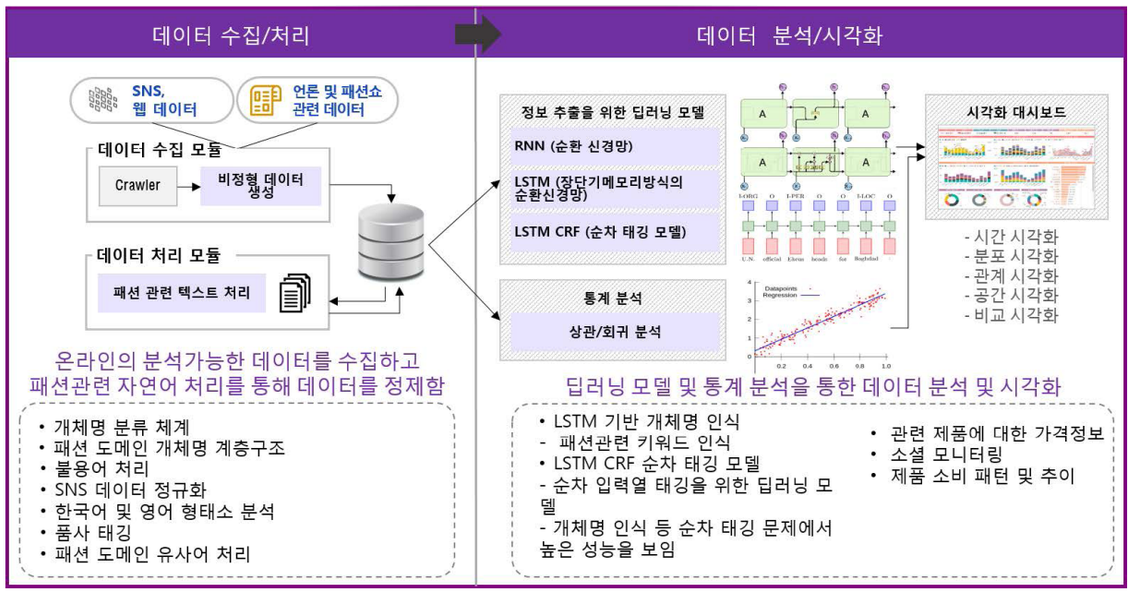
1. 기술 설명

- 소셜미디어 및 패션관련 데이터 수집 및 분석을 통해 최신 이슈 키워드와 트렌드를 다양한 방법으로 제공함
- 이를 위해 전문가에 의한 패션 도메인 개체명 계층구조를 구축하고 이를 활용하여 패션 데이터 수집 및 개체명 인식에 활용함으로써 정확성을 높임





2. 기술 방법



- 소셜미디어 및 패션관련 데이터 수집 및 분석을 통해 최신 이슈 키워드와 트렌드를 다양한 방법으로 제공함
- 수집된 텍스트 데이터 분석을 위한 데이터 전처리 기술 및 데이터 분석을 위한 통계분석 기술을 적용함
- 패션 상품에 대한 연관 관계 분석 기술 및 기계학습 기반 추천 기술 개발을 통해 상품에 대한 이해, 소비자에 대한 이해, 연관 상품 및 개인화 추천 서비스 개발함
- 패션 관련 제품에 대한 연관관계 분석은 쇼핑몰 사용자들의 검색어 입력 패턴을 수집하여 검색어의 관계를 수식에 의해 분석한 후, 연관 검색어를 제공함
- 패션 관련 키워드를 통한 토픽을 추출하고 토픽 단어사이의 연관성 분석을 통해 시각화함
- 상품, 소비자, 연관 상품에 대한 모니터링 및 시각화 서비스 개발함

토픽	패션에 관련 키워드	
바지	레깅스, 스키니팬츠, 스트레이트팬츠, 와이드팬츠, 죠드퍼팬츠, 치노팬츠 칠부팬츠, 카고팬츠, 크롭팬츠, 테링러링팬츠, 테이퍼드팬츠, 트랙팬츠 포멀팬츠, 플레어팬츠, 하렘팬츠, 하이웨이스트팬츠, Leggings & Knit Bottoms, Skinny, Straight Leg, Wide Leg Pants Jodhpurs, Chino, 3/4 Lengths, Cargo Trousers, Crop Top, Tapered Track & Sweatpants, Formal, Flares, Harem Pants 스키니, 일자 바지, 통바지, 승마 바지, 치노 바지, 칠부 바지 카고 바지, 트랙 바지, 포멀 바지, 플레바지, 하렘 바지, 츄리닝 바지	

3. 기술 활용 및 응용 분야

• 빅데이터 분석을 통해 패션 트랜드 예측 및 트랜드 기반 수요예측 분야에 활용 가능

사용자 모델링

MOOT (Massive Open Online Textbook) 학습자 분석 및 시각화 기술

온라인교육 환경 기반의 mind-wandering 판단 기술

스마트 시니어 인지 측정 및 예측 모델

스마트 시니어 맞춤형 프로파일링 시스템

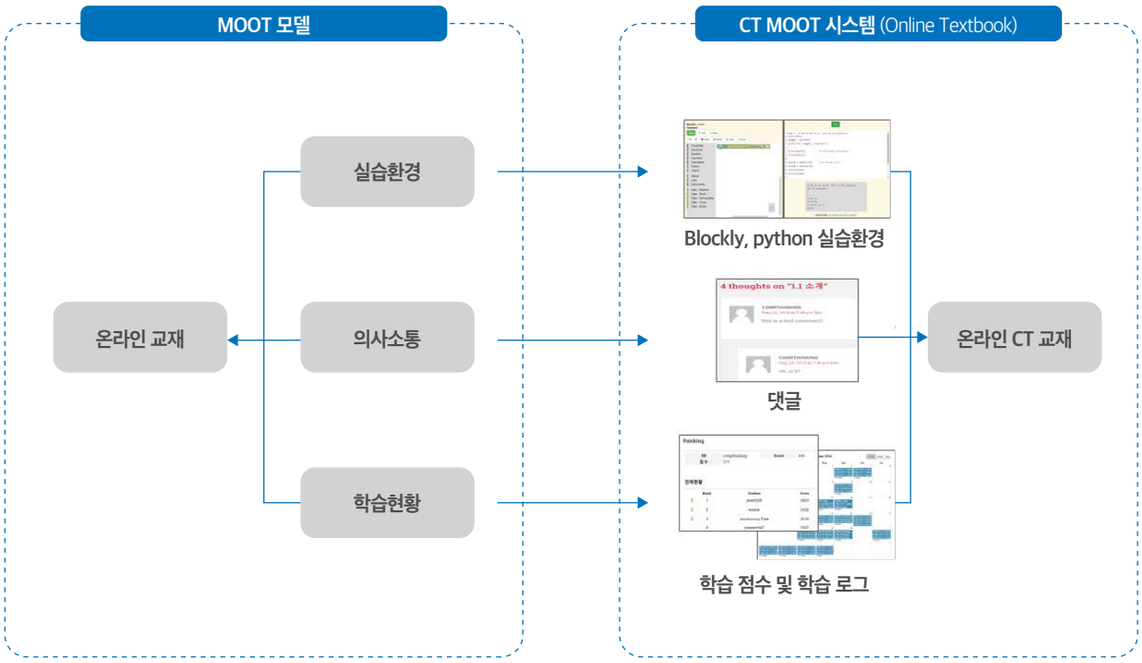
언어 및 인지재활을 위한 온라인 평가·훈련 서비스 플랫폼

법률 코디네이터 서비스

01 MOOT (Massive Open Online Textbook) 학습자 분석 및 시각화 기술

1. 기술 설명

- MOOT는 Massive Open Online Textbook의 약자로 대규모의 사용자에게 제공되는 온라인 교재임
- MOOT는 단순한 온라인 교재가 아닌 실시간 상호작용을 할 수 있으며 단계적인 학습에 따른 실습, 퀴즈, 과제 등 능동적인 학습이 가능한 플랫폼임. 교재 내에서 제공하는 실습과 퀴즈를 통해 자기 주도 학습에 집중하여 좀 더 나은 온라인 교육 시스템임



Massive Open Online Textbook 플랫폼을 기반으로 한 CT MOOT 시스템 개발

2. 기술 방법

- 본 기술은 Massive Open Online Textbook 플랫폼을 개발하여 Computational Thinking의 과목을 적용하였음
- 개발한 CT MOOT 안에서 로그인하여 학습 활동을 진행하면 출석체크와 같이 언제 어디서 어디까지 공부를 했는지 기록이 되며, 텍스트로 학습하면서 실습(Blockly, Python)을 할 수 있고, 퀴즈를 풀 수 있음. 이와 같은 학습 활동에 대한 점수를 주어 랭킹에 적용함. 적용된 랭킹 점수를 보게 되면 사용자들이 학습에 대한 의지를 더 키워갈 수 있음

3. 기술 활용 및 응용 분야

- 이 기술은 text를 통한 공부 뿐만 아니라 실습도 함께할 수 있으므로, MOOT 플랫폼을 기반으로 실습이 포함된 과목에 적용하여 활용할 수 있음
- 데모 http://www.kucomputationalthink.org/

02 온라인교육 환경 기반의 mind-wandering 판단 기술

1. 기술 설명

- 본 기술은 인터넷 강의 시청 후, mind-wandering 여부를 자동으로 판단
- mind-wandering은 강의 내용 중에 있는 단어를 자동으로 추출해서 이를 기반으로 판단
- 교수자는 문제은행 등 퀴즈 제작의 부담이 없고 학습자는 평가에 대한 부담이 없음

최소학습 자동 판단

최소한의 학습 노력 여부를 자동으로 판단 할 수 있는 시스템

쉽고 빠른 시작

학습 판단을 위해 교수자는 퀴즈를 생성할 필요 없고 학습자는 퀴즈가 없는 빠른 시작

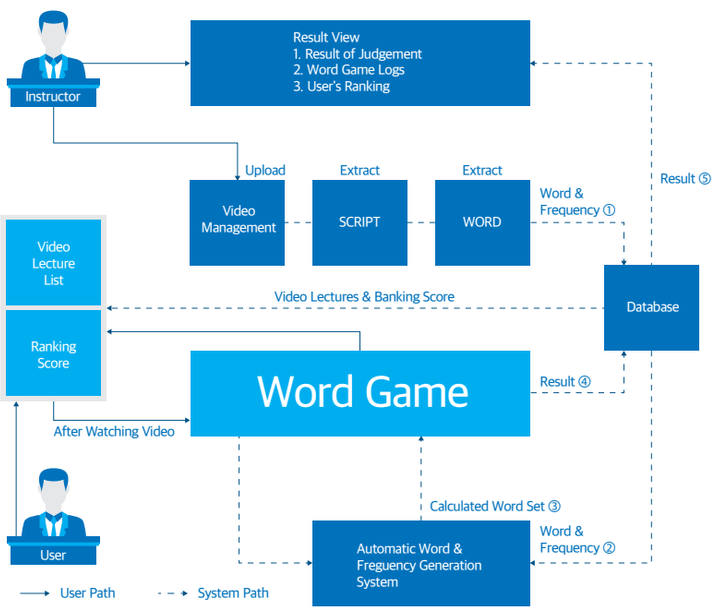
단어게임 자동생성

비디오 콘텐츠를 업로드하면 시스템이 자동으로 단어게임을 생성

게임화

단어 게임과 랭킹 시스템이 적용되어 학습자들의 지속적인 학습 참여와 학습 동기 향상

2. 기술 방법



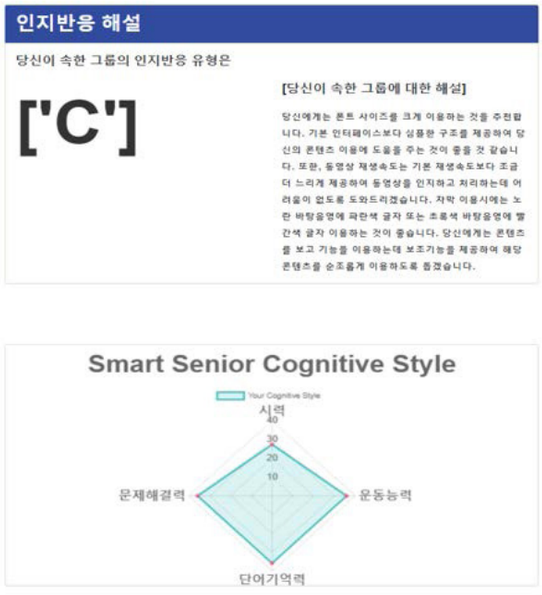
- 업로드된 인터넷 강의에서 스크립트가 추출되고 단어와 빈도수가 데이터베이스에 저장
- 학습자가 인터넷 강의를 시청하고 단어게임을 시작하면, 학습자의 이전 학습 이력과 현재 학습 정보를 반영하여 각 단어의 가중치 값을 계산
- 단어게임이 자동으로 생성

3. 기술 활용 및 응용 분야

- 본 기술은 MOOC, 거꾸로교실 등 온라인 강의 환경에 활용
- 데모 http://mljs.org/

1. 기술 설명

- 본 연구는 스마트 시니어를 대상으로 온라인 인지측정 검사 9종을 개발 및 수집한 데이터를 이용하여 기계학습기반의 인지기능 예측 알고리즘을 개발하였다. 개발한 모델은 해당 검사에 대한 수치를 입력하면 인지검사 결과에 대한 해설을 확인할 수 있는 예측모델이다.
- 하단의 그림은 온라인 인지측정 데이터에 대한 인지 결과를 보여주는 화면이다. 인지그룹(A, B, C) 중에 어느 그룹에 속하는지 알려주며, 사용자의 인지에 대한 전반적인 설명을 텍스트로 보여주며, 시력, 문제해결력, 운동능력, 단어기억력에 대한 방사형 그래프로 어느 능력이 상대적으로 우수한지 보여준다.



<스마트 시니어 인지반응 예측 모델 결과>

2. 기술 방법

- 스마트 시니어에 대한 정의를 위하여 인지반응을 기반으로 시니어를 정의하는 모델을 만든다. 이를 위하여 클러스터링 기법을 통해 시니어의 유형을 군집화하고 각 군집별 도출된 값을 이용하여 semi-supervised learning을 이용한 시니어 인지반응 예측 정확도를 검증한다.

3. 기술 활용 및 응용 분야

- 시니어를 대상으로 하는 의료기관, 복지센터 등에서 시니어를 위한 인지측정에 활용될 수 있으며, 온라인 인지 측정 도구는 PC, 스마트폰, 태블릿 모드에서 사용가능하여 기기에 국한되지 않음
- 또한 인지반응 측정 결과를 통하여 시니어에게 맞는 UI/UX 요소(글자 크기, 동영상 자막색, GUI 구조 등)를 추천할 수 있음
- 과제 소개 영상 https://drive.google.com/open?id=1cL0O5XbkDXh9auNsU4goXFHOPTPBIEKC
- 모델 데모 영상 https://drive.google.com/open?id=1IJCnYzU_HVbAcpMDLhsH1HMyPvNcL6mR

04 스마트 시니어 맞춤형 프로파일링 시스템

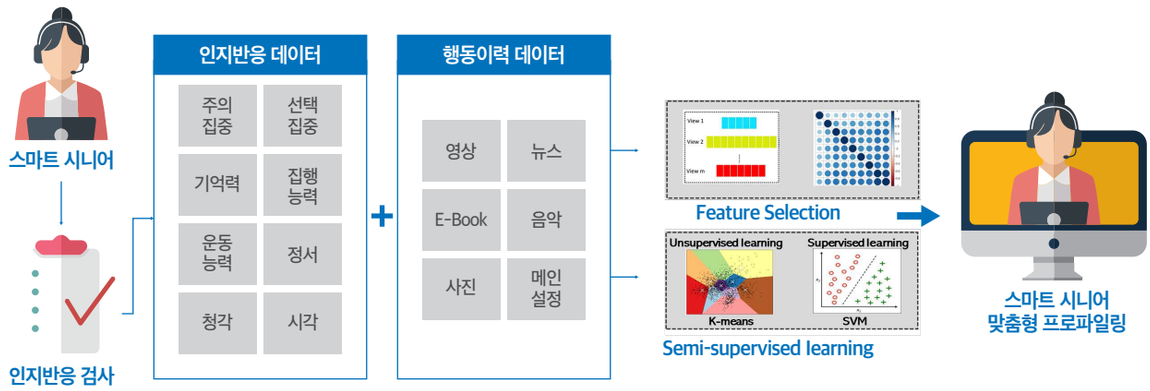
1. 기술 설명

- 스마트 시니어 세대의 신체적 특성을 고려하기 위해 인지적 특성 저하 요소를 파악하였으며, 신체적 특성에 맞게 글자 크기나 화면 비율 등 적용된 온라인 플랫폼을 제공함
- 스마트 시니어 맞춤형 프로파일링은 스마트 시니어의 인지반응 검사지에서 유의미한 검사지만 추출하기 위해 Feature Selection을 이용하였으며, 인지반응 검사 결과와 온라인 활동하면서 수정된 결과를 통해 Semi-Supervised Learning 기법을 활용하여 최종 스마트 시니어 프로파일링 시스템을 개발함

[기술 활용 및 응용 분야]

다른 분야에서 데이터(정답 set)가 있다면 의미있는 feature를 추출하고, 추출된 feature를 통해 어떤 분류에 속하는지 결과를 확인할 수 있다.

2. 기술 방법



스마트 시니어 인지 및 온라인 활동 기반 프로파일링

- 인지반응 검사를 거친 스마트 시니어의 온라인 활동 이력을 바탕으로 180여 개의 인지반응 검사지에서 유의미한 검사 지문을 추출하기 위해 Feature Selection 알고리즘을 적용함
- 또한, 인지반응 검사 결과와 온라인 활동 이력을 바탕으로 SVM(Support Vector Machine)을 활용하여 새로운 사용자가 들어왔을 때 인지반응 검사 결과만으로도 상세 맞춤형 UI/UX를 서비스하기 위해 스마트 시니어 인지 및 온라인 활동 기반 프로파일링 모델을 개발함

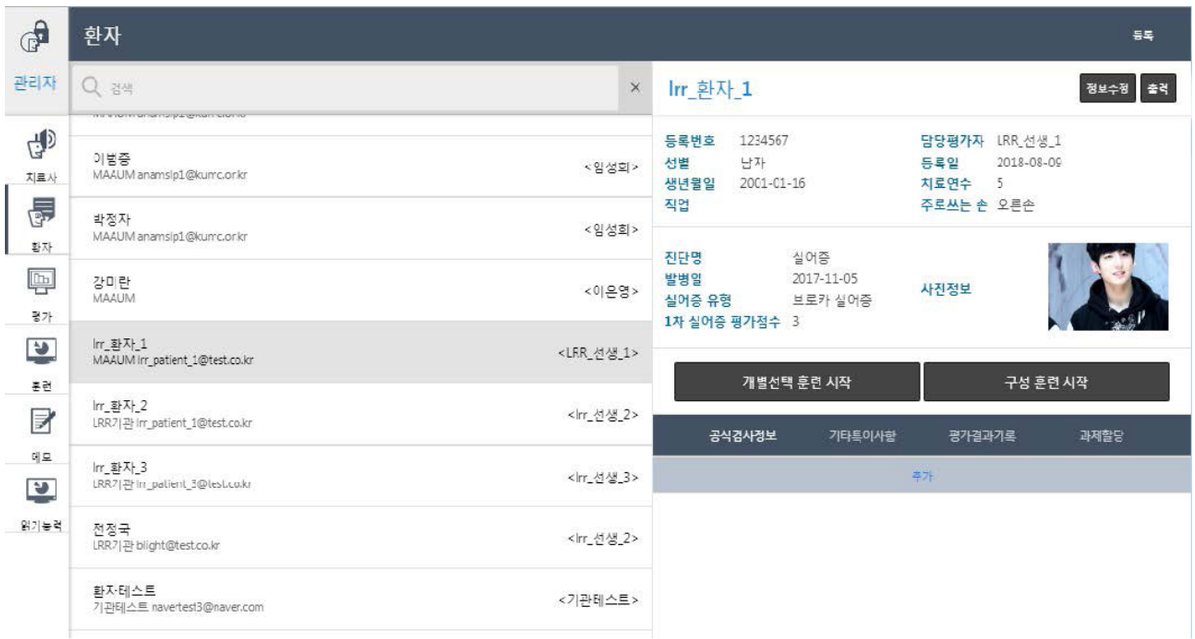
3. 기술 활용 및 응용 분야

- 본 기술은 새로운 사용자의 인지검사 결과만으로도 상세 맞춤형 UI/UX를 서비할 수 있는 스마트 시니어 맞춤형 프로파일링 시스템을 개발하여 스마트 시니어에게 서비스를 제공할 수 있음
- 데모 http://senior.ontheit.com

05 언어 및 인지재활을 위한 온라인 평가·훈련 서비스 플랫폼

1. 기술 설명

- 국내 실어증 및 인지기능 저하의 가장 큰 원인인 뇌혈관 질환이 사망원인 상위를 차지하고 있으며, 인구 고령화로 인해 지속적인 증가세를 보이는 추세임. 이러한 문제를 해결하기 위해 언어 및 인지 재활이 치료 방안으로 권고되고 있음. 효과적인 언어 재활을 위해서는 즉시적이며 장기적인 재활 방안 및 프로그램이 시급함
- 언어 및 인지재활을 위한 온라인 평가·훈련 서비스 플랫폼은 언어처리 및 인지모형 연구를 바탕으로 평가, 훈련 과제가 탑재된 시스템을 의미하며 동적으로 과제를 추가 삭제 구성할 수 있는 플랫폼 개념을 담고 있음.
- 환자 및 사용자는 다양한 언어 및 인지재활 서비스를 활용할 수 있으며 맞춤형 서비스 또한 활용 가능함



2. 기술 방법

- 언어·인지재활에 최적화된 온라인 평가 프로토콜 및 콘텐츠
- 언어 및 인지재활에 최적화된 온라인 평가 콘텐츠 제작을 위하여 선별검사를 포함한 11개 영역(선별검사, 인지선별검사, 언어선별검사, 구성능력, 주의, 지각, 기억, 언어기능, 집행기능, 실행기능, 정서)
- 선별검사(2개 과제), 인지선별검사(2개 과제), 언어선별검사(2개 과제), 구성능력(2개 과제), 주의(5개 과제), 지각(2개 과제), 기억(6개 과제), 언어기능(10개 과제), 집행기능(3개 과제), 실행기능(5개 과제), 정서(3개 과제) 프로토콜 및 콘텐츠 구성
- 언어·인지재활에 최적화된 온라인 맞춤형 훈련 프로토콜 및 콘텐츠
- 언어 및 인지재활에 최적화된 온라인 맞춤형 훈련 콘텐츠 제작을 위하여 언어기능 훈련(단어 수준의 이해 산출 능력, 의미범주 수준의 이해 산출 능력, 문장 수준의 이해 산출 능력, 일반적 사실 및 담화 이해 산출 능력 훈련)과 인지기능훈련(집행기능, 실행기능, 구성능력, 기억력, 주의)으로 프로토콜 구성
- 단어 수준의 이해 산출 능력(6개 과제), 의미범주 수준의 이해 산출 능력 훈련(10개 과제), 문장 수준의 이해 산출 능력(9개 과제), 일반적 사실 및 담화 이해 산출 능력(7개 과제), 집행기능(2개 과제), 실행기능(3개 과제), 구성능력(2개 과제), 기억력(2개 과제), 주의 (1개 과제)로 프로토콜 및 콘텐츠 구성

• 온라인 평가 및 맞춤형 훈련 통합 시스템 및 콘텐츠 저작 시스템

• 뇌질환 환자 혹은 사용자, 언어 및 인지기능 평가, 훈련, 결과 통합 데이터베이스

• 온라인 평가 및 맞춤형 훈련 콘텐츠 저작 시스템 및 통합 서비스 시스템

• Deep learning 기반의 콘텐츠 추천 모델 연구 및 기술

• 읽기 능력 향상을 위한 음성인식 프로토콜 및 서비스 시스템

3. 기술 활용 및 응용 분야

• 본 기술은 언어 및 인지 재활을 위한 솔류션으로 활용될 수 있음

• 데모 http://aphasia.co.kr

06 법률 코디네이터 서비스

1. 기술 설명

- 의뢰인이 원하는 변호사의 특징을 검색 키워드로 입력하면 원하는 변호사를 매칭시켜주는 서비스를 제공함
- 의뢰인이 기본적인 상담 후 변호사를 선임해야 할 경우, 변호사가 의뢰인들의 구인공고를 보고 역경매방식으로 지원하는 서비스를 제공함
- 대부분의 집단소송이 카페 등 폐쇄적으로 진행되었다면 집단소송에 대한 정보를 알리고, 해당 집단소송의 절차를 간소화시켜 일반인들의 접근성을 높일 수 있도록 지원하는 서비스를 제공함

2. 기술 방법

- 상담후기와 변호사가 작성한 소개 글을 토대로 변호사들의 특징을 추출하고, 의뢰인이 원하는 키워드와 추출한 특징들을 매칭하는 알고리즘을 적용함
- 역경매방식의 변호사선임 서비스는 의뢰인들이 변호사 선임을 위한 공고를 내면, 변호사들이 역경매방식으로 소송비용을 제출하고 의뢰인은 여러 소송비용과 예측 승소율을 토대로 원하는 변호사 선입을 지원하는 플랫폼을 구축함
- 변호사가 집단소송에 대한 아이디어를 제시하고, 클라우드 공모 형태로 집단소송 의뢰인들을 모집하며 진행상황 뿐만 아니라 새로운 집단소송에 대한 정보제공 서비스를 구축함

자연어처리와 인공지능 (교육과정)

교육 과정 개요

최근 4차 산업혁명은 인간과 기계의 잠재적 능력을 극대화시키는 제반 기술혁신이 경제·사회 전반의 시스템에 큰 변화를 가져올 것으로 전망되고 있습니다.

기술의 융합을 통해 비약적인 기술발전이 가속화되고 있는 산업에서 인공지능 기술은 필수입니다. 특히 컴퓨터가 경험을 통해 인간처럼 스스로 학습할 수 있게 하는 기계학습(Machine learning)은 인공지능에서 핵심적인 기술입니다.

우리 연구소는 Human-inspired Machine learning, Human-inspired Rapid learning,지식 표현, 획득,추론 기술의 융합 및 지능정보 등의 원천기술을 보유하고 있으며 이에 대한 기술활용과 비즈니스 창출 역량 배양을 위한 교육을 실시하고 있습니다.

자연어 처리와 인공지능분야의 전문인력 양성을 위한 초급,중급의 수준별 교육과 기간별 교육, 기업연계 회사맞춤형 교육 등을 실시하고 있으며,인공지능 프로젝트를 진행하기 위한 딥러닝 심화, 로젝트 핵심기술, 고리즘 구현, 듈 설계 및 구현에 관련된 미니프로젝트 중심의 실습교육을 운영하고 있습니다.

부디 본 교육이 인공지능 전반의 기술이해와 산업현장에 적용이 되어 나아가 우리나라 AI분야의 초석이 되고 우리 모두에게 도움이 되었으면 합니다.

교육 프로그램

- S그룹 언어지능 교육과정(단기)
- L그룹 중급 언어지능과정(3-4주)
- 하계/동계 자연어처리와 언어지능(기초교육과정)
- AI 산업전반 및 활용사례, 실무에 적용할 수 있는 프로젝트 연구(회사 맞춤형 교육)
- AI 기초 프로그래밍 및 심화프로그램(자연어처리, 음성인식, 영상처리)
- AI와 빅테이터 분야 인력양성을 위한 교육

세부 교육 과정

1.자연어처리 소개, 프로그래밍 및 자연어처리의 기본 원리

자연어처리 개요: 자연어처리에 대한 정의 및 자연어처리 절차, 최신 동향

딥러닝의 소개: 자연어처리의 핵심기술인 딥러닝 기법인 CNN,RNN

언어를 이해하는 컴퓨터: 언어를 이해하는 자연어처리 기술

언어를 생성하는 컴퓨터: 언어를 생성하는 자연어처리 기술

자연어처리의 다양한 응용 분야: 문서분류, 자동정보추출, 기계독해, 문서요약, 기계번역, 자동질의응답, 대화 시스템 등

Python 기초: Python 기초 문법 및 함수, Python을 이용한 뉴스기사 분석 및 시각화

자연어처리를 위한 전처리 프로그래밍 방법: 텍스트 데이터 분석 및 시각화

Python을 이용한 뉴스기사 분석 및 시각화:

2. 자연어처리, 기계학습 및 데이터마이닝

자연어처리 기초: 자연어처리의 정의 및 절차, 최신 동향

텍스트 전처리: 텍스트 데이터를 사용하고자 하는 목적에 맞게 가공하기 위한 토큰화, 어간 추출, 불용어 제거, 텍스트 분리

어휘 분석, 문장 분석, 의미 분석: 텍스트 데이터를 의미의 최소 단위인 어휘로 분리하고 적합한 품사 정보를 할당하기 위한 형태소 분석, 문장 구조분석, 문장의 의미 해석방법

문맥 분석: 하나 이상의 문장으로 구성된 텍스트 데이터를 진술, 주장, 추측, 명령, 요청 등 발화의 의도를 분석하고 구분하는 방법

구문 분석: 주어진 텍스트를 일련의 구문과 토큰으로 분해하여 해당 토큰의 언어적 정보를 제공하는 방법

화행 분석: 대화 속에서 문장의 화행을 알아내는 방법

개체명 인식: 텍스트 데이터에서 객체를 표현하는 단어들을 구분하고, 그 단어에 해당 객체를 의미하는 라벨을 할당하는 기법

형태소 분석: 형태소 분석이란 형태소를 비롯하여 어근, 접두사/접미사, 품사(part of speech)등 다양한 언어적 속성을 파악하는 방법

웹 스크래핑: 웹 사이트 상에서 원하는 부분에 위치한 정보를 자동으로 추출하여 수집하는 기술

웹 크롤링: 자동화 봇(bot)인 웹 크롤러가 정해진 규칙에 따라 복수 개의 웹 페이지를 브라우징 하는 행위

토큰화: 데이터를 사용하고자 하는 용도에 맞게 토큰이라 불리는 단위로 나누는 작업

과거에 대한 이해, 미래에 대한 예측 선택: 기계학습과 데이터베이스 소개 및 기계학습의 원리

미래에 대한 예측을 위한 다양한 기계학습 방법 습득: 다양한 기계학습 모델 및 인공신경망, 딥러닝 소개

기계학습 도구 실습 및 기계학습을 이용한 문제해결: 언어모델, 기계번역, 영상주석 생성 등 기계학습 방법을 이용한 문제해결 소개

3. 여러가지 자연어처리 응용분야

Named Entity Recognition: 텍스트 데이터에서 객체를 표현하는 단어를 구분하고 그 단어에 해당하는 객체를 의미하는 라벨을 할당하는 기법

Language model: 일련의 순서를 가진 텍스트 데이터가 주어졌을 때 다음에 위치할 텍스트 데이터를 확률적으로 예측하는 언어 모델과 통계적 기법과 기계학습 기반의 방법론

Information Extraction: 비정형 텍스트 데이터에서 목적에 맞는 정형화된 텍스트 정보를 추출하는 방법과 개체명 인식과 개체간의 관계를 표현하는 등의 방법론

Question & Answering: 질문이 주어졌을 때 그에 해당하는 답변을 자동으로 선택, 생성하는 방법과 이를 구현하기 위한 규칙 기반, 기계학습 기반의 방법론

Machine Translation: 입력된 단어를 다른 단어로 바꿔서 출력해주는 방법을 설명하고 전통적인 기계번역 방법 및 통계 기반, 기계학습 기반의 번역방법론

Text Generation: 주어진 상황 및 입력 텍스트에 적절한 문장을 생성하는 방법을 설명하고 기계학습 기반의 방법 및 강화학습 기반의 방법

Machine Reading Comprehension: 주어진 텍스트 데이터의 문법적,의미적 맥락을 이해하여 상황에 맞는 답변방법을 설명하고 MRC를 위한 자연어처리 기술 및 평가방법

Dialogue System: 사용자와 컴퓨터가 정보를 주고받는 시스템에 대한 설명과 대화시스템의 종류와 구축방법

Text Summarization:텍스트 데이터의 정보를 컴퓨터가 압축된 문장으로 표현해주는 방법과 자동요약의 종류 및 기법

Text Categorizarion & Sentiment Analysis: 문서에 포함된 텍스트 데이터를 분석하여 정해진 카테고리에 따라서 분류하는 방법과 텍스트 데이터에서 작성자의 주관적인 의견을 텍스트로부터 분석해내는 방법과 구현방법

4. 딥러닝 기반 자연어처리 (실습, 응용 개발 프로젝트)

Colab 툴킷 사용: Colab은 구글에서 공개한 웹기반의 Python 개발 환경으로 기본적이 사용법과 특징

단어 임베딩: 단어 임베딩은 단어를 벡터로 표현하는 것으로 임베딩 기법의 종류를 설명하고 기본적 기법 활용

딥러닝 기반의 Language 모델링: 여러 가지 자연어처리의 응용에서 학습한 언어모델의 일부를 Colab을 통해 구현

어절 자동생성기 개발 프로젝트: RNN을 이용

딥러닝 기반의 한국어 문장 및 문서, 감정 분석: Text Categorizarion & Sentiment Analysis 방법을 Colab을 통해 일부 구현

감정분석 또는 문서분석기 개발 프로젝트: CNN을 이용

인공신경망과 기계학습: 인공신경망과 기계학습의 이론 및 실습

CNN, RNN, 언어표현: CNN, RNN등 딥러닝 이론 및 실습

한국어 언어표현 실습: 한국어 자연어처리 이론 및 실습

5. 시각지능

컴퓨터비전 구현,영상의 이해 및 CNN활용

Open CV for python3, Open CV 활용

Segmentation,Transfer Learning, Auto Encoder

시각지능 프로그램(차량번호판 인식 등)

세부 교육 과정

예시 : 단기교육 과정

주차	과목	내용
1	기초 통계에서 고급 통계 및 프로젝트	• 과거에 대한 이해 : 통계를 위한 tool 사용 및 기초통계 에서 고급통계 • 분석을 위한 데이터 설계까지 (문제의 원인을 분석하는 과정)
2		
3		
4		
5	기계학습 및 데이터 마이닝 기초	• 과거에 대한 이해 / 미래에 대한 예측 선택 • 미래에 대한 예측을 위한 다양한기계 학습 방법 습득 • 기계학습 도구 실습 및 기계학습을 이용한 문제해결
6		
7		
8		
9	시각 지능을 위한 다양한 알고리즘	• 시각 지능의 이해 • 시각 지능의 구현 및 실습
10		
11		
12	딥러닝 이론	• 딥러닝의 원리 이해 및 해부 • CNN, RNN, LSTM, GRU, RNN with Attention • 딥러닝을 이용한 문제 해결 및 성능 향상 - 실습
13		
14		
15		
16	자연어처리의 기본 원리, 딥러닝 기반 자연어처리, 딥러닝을 이용한 자연어처리 응용 개발 프로젝트	• 언어를 이해하는 컴퓨터 • 언어를 생성하는 컴퓨터 • 자연어처리의 다양한 응용 분야 • 딥러닝을 이용한 자연어처리 프로젝트
17		
18		
19		
20		

예시 : 중급과정

- 교육목표 : AI · Big Data 분야별 Project를 Leading하며 문제를 해결하기 위해 필요한 핵심 이론 및 분석 방법론 습득
- 교육대상 - AI · Big Data 직무 또는 Project 담당자
  - 언어분야의 AI · Big Data 적용에 대한 기본기가 필요한 인원
- 사전과정 : Machine Learning과 Python 기본 지식 미 보유 시 LG AIB 입문 先수강 필요
- 이수기준 - Machine Learning Project 수행 및 Output 도출
  - 출석률 70% 이상, 종합평가점수 70점 이상
- 교수진 : 고려대학교 자연어처리 연구실 및 HI AI & Computing 연구소 (http://nlp.korea.ac.kr/and http://hiai.kr)
- 주요학습내용

과목	주요내용
자연어처리 프로그래밍 기초	• 자연어처리를 위한 전처리 방법 • 텍스트 데이터 분석 및 시각화
자연어처리와 기계학습	• 자연어 처리 기초 - 전처리 - 어휘 분석 - 구문 분석 - 의미 분석 - 화행 분석 • 기계학습 - 언어 모델 - CNN 및 RNN • 기계학습을 이용한 텍스트 데이터 처리 실습
자연어처리 Project	• 자연어처리 Project • Project 수행 및 발표/Feedback

예시 : 프로그램 전체일정

<table>
<tr><th>구분</th><th colspan="4">1주차</th><th colspan="3">2주차</th><th colspan="3">3주차</th></tr>
<tr><td>09:00-
10:00</td><td>과정 안내</td><td rowspan="3">자연어 처리
기초(1)
-자연어 처리
전 처리,
형태소 분석</td><td rowspan="3">자연어처리
응용 I
-언어모델, NER,
기계번역 등-</td><td rowspan="3">자연어처리
응용II
-대화시스템, 문
서요약 등</td><td rowspan="3">기계학습의
이해
(2)</td><td rowspan="3">딥러닝(1)
CNN</td><td rowspan="3">딥러닝(3)
RNN</td><td rowspan="3">Project 수행(1)
Sentiment
analysis</td><td rowspan="3">Project 수행(2)
Language
Model</td></tr>
<tr><td>10:00-
11:00</td><td rowspan="2">자연처리기초
프로그래밍(1)
-nltk기반 텍
스트
데이터 전처리
01-01</td></tr>
<tr><td>11:00-
12:00</td></tr>
<tr><td>12:00-
13:00</td><td colspan="9">중식</td></tr>
<tr><td>13:00-
14:00</td><td rowspan="4">자연어처리기초
프로그래밍(2)
-nltk기반
텍스트
데이터 전처리</td><td rowspan="4">자연어처리
기초(2)
-어휘분석
-구문분석,
의미분석
-화행분석</td><td rowspan="2">형태소분석
기 실
습</td><td rowspan="2">기계학습의
이해
(1)</td><td rowspan="4">신경망 기초</td><td>딥러닝(2)
언어 표현
Word2Vec
Glove,
FastText,
ELMO, BERT</td><td rowspan="4">딥러닝을 이
용한
NER실습</td><td rowspan="3">Project 수행
(계속)</td><td rowspan="2">수행(2)
프로젝트
발표(1시간)</td></tr>
<tr><td>14:00-
15:00</td></tr>
<tr><td>15:00-
16:00</td><td rowspan="2">의존분석기실
03-03의존
분석.pptx</td><td rowspan="2">-MNIST
(숫자이해)
및 텐서플로어</td><td rowspan="2">딥러닝 실습
(word
embedding)</td></tr>
<tr><td>16:00-
17:00</td><td rowspan="2">프로젝트 발표</td><td>프로젝트
발표(1시간)</td></tr>
<tr><td>17:00-
17:30</td><td>차시 안내</td><td>과정 마무리</td></tr>
</table>

예시 : Lesson Plan

Danial Hooshyar	주요내용	방식
자연어처리 프로그래밍 기초	1 .자연어처리 프로그래밍 기초 - 파이썬 프로그래밍 기초 - 자연어처리 전처리 프로그래밍 - 파이썬을 이용한 뉴스기사 분석 및 시각화 (형태소분석, 토큰화, 테그 클라우드)	실습 강의
자연어처리와 기계학습	1. 자연어처리의 기초 - 전처리, 어휘 분석, 구문 분석, 화행 분석 2. 자연어처리의 응용 - 개체명 인식, 언어 모델, 기계 번역 3. 기계학습의 이해 - 최적화, 일반화, 정규화, 일반적인 문제점 - 자료 분류와 군집화	실습 강의
신경망과 딥러닝	1. 신경망 - 인공신경망과 기계학습 2. 딥러닝 - CNN, RNN, 언어 표현 3. 한국어 언어 표현 실습	실습 강의
딥러닝 기반 자연어처리 프로젝트	1. Sentiment Analysis 프로젝트 2. 언어 모델 프로젝트 3. 코칭 및 토론	실습 발표 코칭 토론

특허 등록

특허등록

특허명	등록번호	등록일
집단지성을 이용한 뉴스 판단 방법 및 장치	10-1869815	2018. 06. 15.
음식 배달 중개 방법 및 장치	10-1896408	2018. 09. 03.
사물인터넷에 기반한 경험 공유 방법 및 장치	10-1909646	2018. 10. 12.
집단감성을 이용한 맞춤형 영화 상영 방법 및 그 장치	10-1858120	2018. 05. 09.
사물 인터넷 기반의 대출 관리 방법 및 그 장치	10-1795462	2017. 11. 02.
사물 인터넷 기반 스마트 화분 및 그 관리 시스템	10-1789165	2017. 10. 17.
사물인터넷 기반의 스마트 의자 및 착석자세 분석 방법		
스마트 의자 관리 장치 및 그 방법	10-1816711	2018. 01. 03.
집단지성을 이용한 꿈 해몽 방법 및 장치	10-1748411	2017. 06. 12.
학습코스 자동 생성 방법 및 시스템	10-1745874	2017. 06. 05.
온라인 학습자를 위한 주의집중 판단 시스템 및 그 방법	10-1770817	2017. 08. 17.
사용자 참여 기반의 정책 발굴 방법	10-1739925	2017. 05. 19.
인문학 정보를 자동으로 구성하는 방법	10-1760478	2017. 07. 17.
전자 책상을 이용한 멘토 추천 방법	10-1653620	2016. 08. 29.
지능형 학습 관리 방법	10-1693592	2017. 01. 02.
인지능력 측정 장치 및 방법	10-1222210	2013. 01. 08.
학습자 인지능력 기반의 외국어 학습 시스템 및 방법	10-1136415	2012. 04. 06.
외국어 학습자용 인지능력 진단 시스템 및 방법	10-1113908	2012. 02. 01.

기술 이전

기술이전

- 딥러닝기반 고유명사 개체명 인식기술
- 딥러닝 기법을 이용한 온라인 콘텐츠 추천 기술
- 딥러닝 기법을 이용한 한국어 개체명 인식 시스템
- 딥러닝 기법을 이용한 콘텐츠 추천 시스템
- 외국어 학습자용 학습 과제 수행 시스템 및 방법
- 동영상 내의 멀티모달 정보 색인 기술
- 사용자 콘텐츠 소비 정보를 이용한 추천 시스템
- 은닉 마르코프 모델을 이용한 시계열적 추천 모델
- 온라인 협력 학습 플랫폼
- 디지털 콘텐츠 전용 검색 기술
- 반응형 웹기반의 소셜 러닝 서비스 플랫폼
- Intelligent fasion image retrieval system
- 한국어 개체명 인식기 및 의존 구문 분석기
- 지능형 분류기술

본 책자는 과학기술정보통신부 및 정보통신기획평가원의 대학ICT연구센터지원사업(IITP-2018-0-01405)과
정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 결과임.
(NRF-2016R1A2B2015912,NRF-2017M3C4A7068189)

고려대학교

Human-Inspired AI & Computing
연구센터