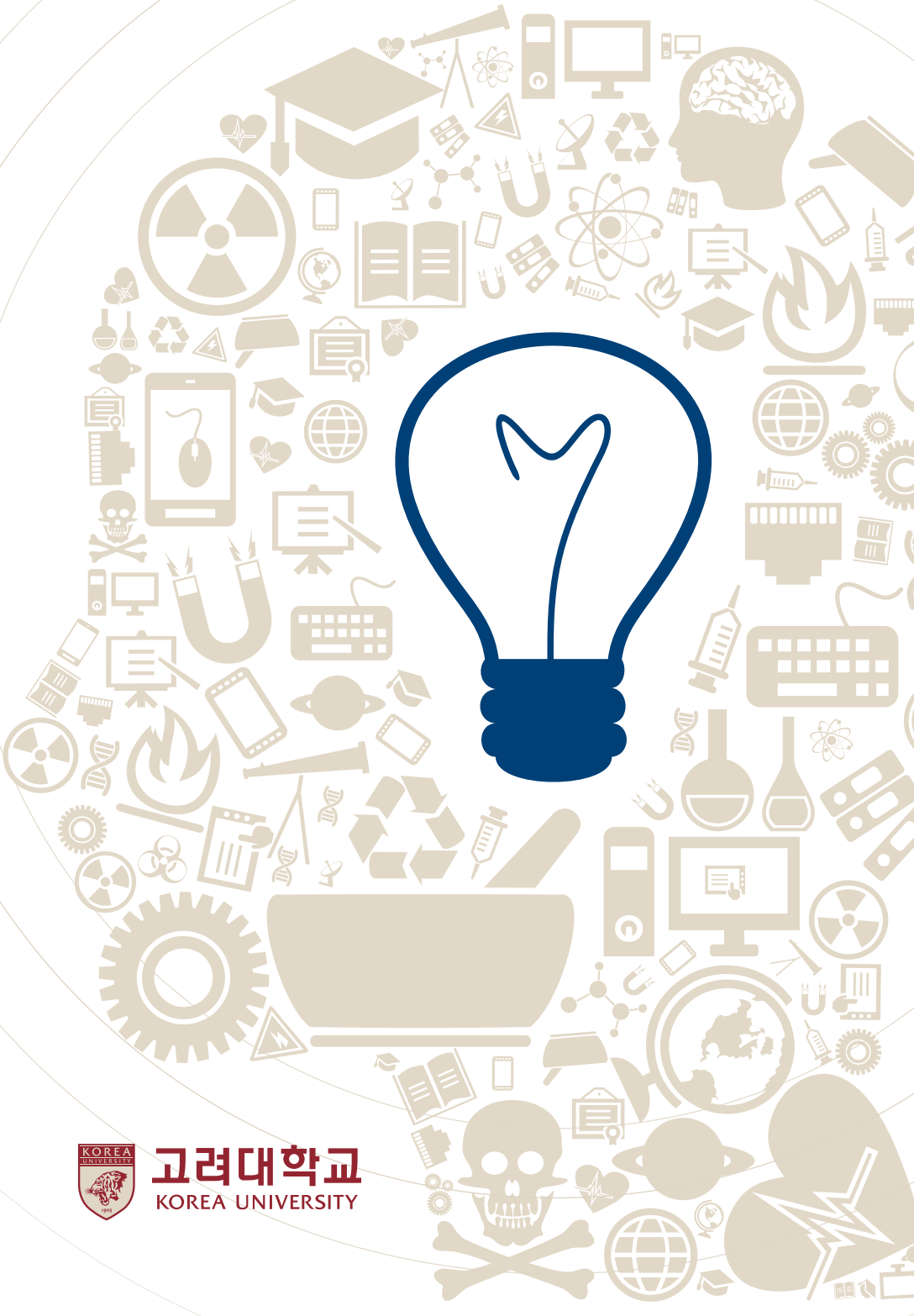
고려대학교

Human-Inspired AI

연구소 Vol.3, 2021



고려대학교
KOREA UNIVERSITY

인사말

임희석 교수/연구소장

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 Human-Inspired AI 연구소는 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 연구소에서는 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

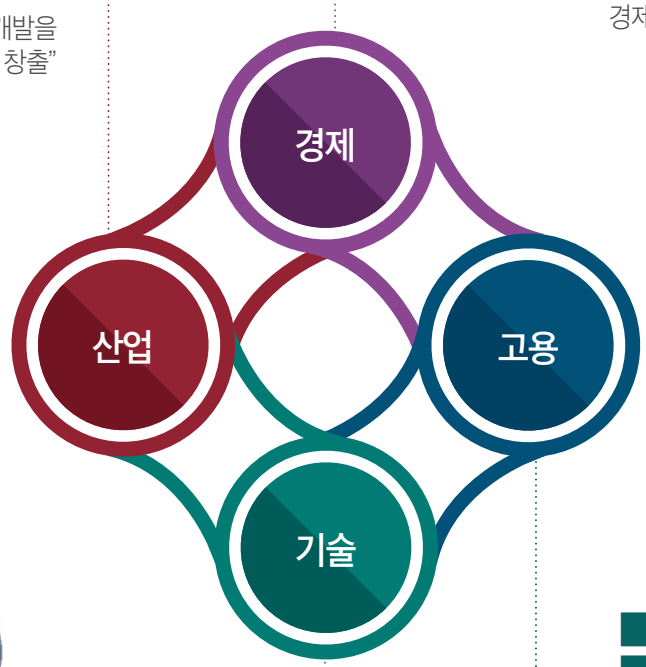
센 터 비 전

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"

Contents

목차

●○○○

자연어처리

한국어 띄어쓰기 자동 교정기

딥러닝을 이용한 영어 문법 오류 교정기

통계 및 확률 기반 형태소 분석 기술

딥러닝 기반 형태소 분석 기술

개체명 인식기 (Named Entity Recognition)

문서 자동 분류 기술

Bag of Characters를 응용한 Character-Level Word Representation 기술

병렬 코퍼스를 이용한 이중언어 워드 임베딩

Stack-Pointer Network를 이용한 한국어 의존 구문 분석

의존구문분석 (Dependency Parser)

Small Data의 한계를 극복하기 위한 전이 학습 모델

통계기반 한국어 뉴스 감정분석

대화속 화자의 감성 분석 (Emotion Recognition in Conversation)

자연어 추론에서의 교차 검증 앙상블 기법

Denoising Transformer기반 한국어 맞춤법 교정기

지식 임베딩 심층학습을 이용한 단어 의미 중의성 해소

Attentive Aggregation(주의적 종합)기반 크로스 모달 임베딩

Poly-encoder를 이용한 COVID-19 질의응답시스템

외부지식정보를 이용한 상식추론 질의응답시스템

사전 학습된 Transformer 언어 모델의 이종 언어 간 전이 학습을 통한 자원 희소성 문제 극복

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] 단순 규칙을 이용하여 띄어쓰기 교정이 가능한 경우

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

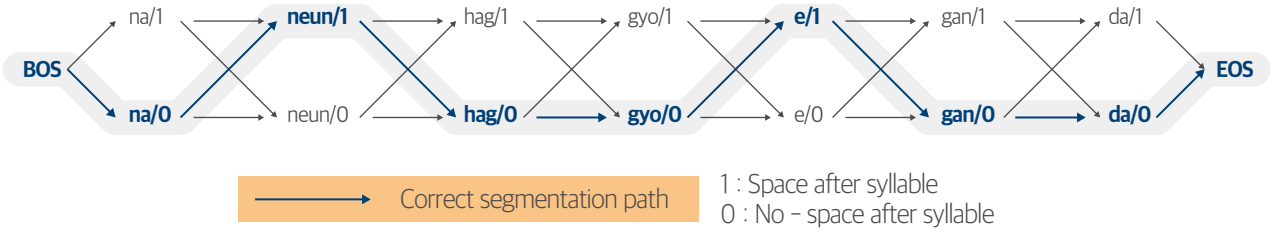
아버지가방에 들어가셨다 → 아버지가 방에 들어 가셨다

아버지가방에 들어가셨다 → 아버지 가방에 들어 가셨다

[그림] 단순 규칙으로 띄어쓰기 교정이 불가능한 경우 : 확률 모델의 적용 필요

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] 띄어쓰기 확률 경로 예시

3. 기술 활용 및 응용 분야

감정 분석, 자연어처리

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

4. 실험 (Only PDF)

본 기술은 띄어쓰기 문제를 품사 부착 문제와 같은 분류 문제(classification problem)로 간주한다. 은닉 마르코프 모델 (hidden Markov model; 이하 HMM)은 품사부착, 정보추출, 개체명 인식, 외래어 추출 등과 같은 자연어처리의 여러 문제를 해결하는 데에 많이 사용되는 모델이며 각 분야에서 높은 성능을 보이고 있다.

띄어쓰기 문제에서는 학습을 위해 따로 말뭉치를 구축할 필요가 없이 이미 존재하는 원시 말뭉치를 학습 말뭉치로 사용할 수 있다. 따라서 HMM이 띄어쓰기 문제에도 효과적으로 적용될 수 있으며 띄어쓰기 문제에 적합하도록 HMM을 일반화하여 확장된 문맥을 고려할 수 있는 통계적 모델을 사용한다.

1. 기술 설명

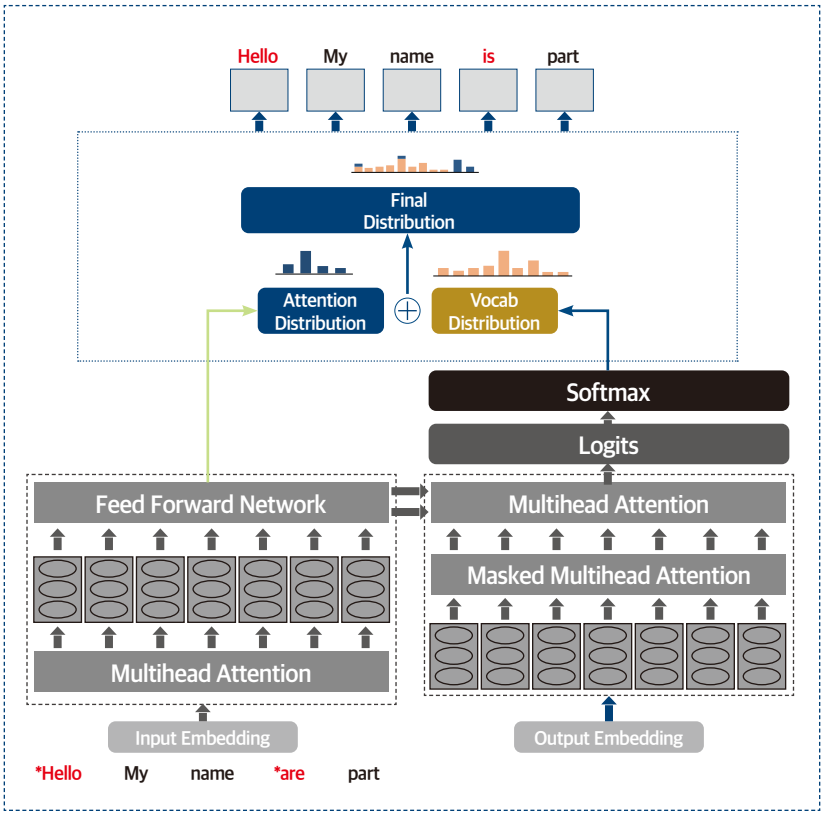
- 영어 문법 교정 시스템(Grammar Error Correction system)은 사용자가 입력한 영어 문장의 문법실수, 철자오류, 단어오용 등을 바로잡아 주는 인공지능 시스템이다.
- 영어 문법 교정 시스템에서 교정을 잘하는 것도 중요한 요소이나 옳은 문장이 들어왔을 때 옳은 문장을 그대로 교정없이 출력으로 내보내는 것 또한 매우 중요한 요소다.
- Overcorrection이란 입력으로 문법적으로 올바른 문장이 들어왔음에도 교정을 해야할 대상으로 간주하여 문장의 구조를 흐트러트리는 현상을 의미한다. NMT를 이용한 GEC 같은 경우 NMT의 고질적인 문제점인 반복번역, 생략, UNK(Unknown) 문제점 때문에 문장의 구조를 흐트러트리거나 Overcorrection 하는 경우가 존재한다.
- 현재 대부분의 논문들은 교정 성능을 높이는 것에만 집중하고 있지 옳은 문장이 입력으로 들어왔을 때 옳은 문장이 출력으로 나오는 것에는 집중하지 않고 있다. 실제 서비스를 했을 때 올바른 문장이 들어왔음에도 이상한 결과를 출력하거나 올바른 것도 고쳐버리는 오류가 발생하게 되면 좋은 교정성능을 가지고 있음에도 사용자들의 software에 대한 신뢰성이 떨어지게 된다.
- 본 연구는 교정 성능(Correction)과 과교정(Overcorrection) 성능을 포괄적으로 측정할 수 있는 새로운 Metrics 제안한다.

Input **Overcorrectoin**	Mr. Banks is aware that there are budget problems. Mr. Bank is aware that there are budget problems.(*Deleted*)
Input **Overcorrectoin**	I was iust going to cross the road when somebody shouted 'Stop!' I was iust going to cross the road when somebody shouted (*Deleted*)
Input **Overcorrectoin**	This knowledge may be relevant to them. This knowledge may be similar to them.(*Replaced*)
Input **Overcorrectoin**	Disposable income increased from 1999 to 2004. A good income increased from 1999 to 2004.(*Replaced*)
Input **Overcorrectoin**	Didnt you tell me that either Deborah or David has done his assignment? Did you tell me that either Deborah or David has done his assignment?(*Replaced*)
Input **Overcorrectoin**	In some countries you are not able to drink until you are 21. In some countries you cannot drink until you are 21.(*Replaced*)
Input **Overcorrectoin**	I will meet Jane, who is my sister. I will meet the Jane, who is my sister.(*Added*)
Input **Overcorrectoin**	One day last September, it rained for ten hours without stopping. One SeptemberOne day, it rained for ten hours without stopping.(*Replaced*)

Overcorrection 예시

2. 기술 방법

- 영어 문법 교정기 분야의 새로운 Metric인 covering grammar error and overcorrection performance (CGOP)를 제안함
- 해당 Metrics은 교정성능과 Overcorrection 성능을 포괄적으로 측정할 수 있는 최초의 지표임
- 교정성능은 Generalized Language Evaluation Understanding(GLUE) 점수를 이용하며 Overcorrection 성능은 Levenstein 알고리즘과 longest common substring(LCS) 알고리즘을 이용하여 성능 측정함



모델 구조

3. 기술 활용 및 응용 분야

• 본 기술은 Grammarly와 같은 상용화 문법교정시스템으로 응용 가능하며 더 나아가 어린이 영어교육 시장에 활용 가능함

4. 실험

4.1 실험 개요

• 대표적인 Sequence to Sequence 모델을 이용하여 Deep-learning based GEC의 교정 성능과 overcorrection 성능을 확인해본다. LSTM-Attention 그리고 Transformer 기반의 모델을 통하여 각각의 교정 성능과 overcorrection 성능을 검증하고 더 나아가 Copy Mechanism을 적용했을 때 성능 변화를 확인해본다.

4.2 실험 결과

• 기존 성능 측정 방법인 GLUE와 BLEU와 제안하는 Metrics인 CEOF의 모델 성능 순위가 뒤집힘

• Copy Mechanism이 Overcorrection 문제를 완화함을 발견

5. 데모

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

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 태그 집합>

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

4. 결과 화면 (Only PDF)

<형태소 분석 결과>

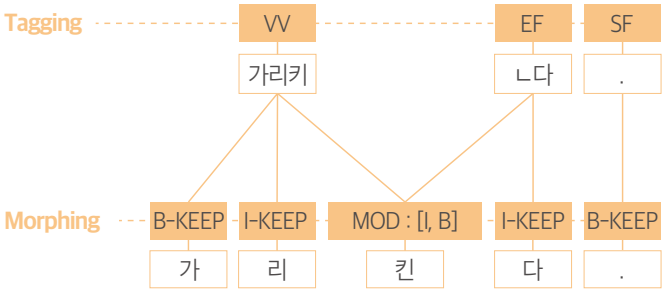
QUERY : 주택 문제의 경우 제 나이가 아직 젊으니까 가능성이 많지요.
기와나 슬레이트로 된 지붕들이 납작하게 펼쳐져 있는 것이 보인다.
내일이면 이제 모두 끝내고 조금 쉴 수 있을 거 같아.

RESULT:

주택 주택/NNG
문제의 문제/NNG+의/JKG
경우 경우/NNG
제 저/NP+의/JKG
나이가 나이/NNG+가/JKS
아직 아직/MAG
젊으니까 젊/VA+으니까/EM
가능성이 가능성/NNG+이/JKS
많지요. 많/VA+지요/EM+./SF
기와나 기와/NNG+나/JKB
슬레이트로 슬레이트/NNG+로/JKB
된 되/VV+ㄴ/ETM
지붕들이 지붕/NNG+들/XSN+이/JKS
납작하게 납작하/VA+게/EM
펼쳐져 펼쳐지/VV+어/EM
있는 있/VX+는/ETM
것이 것/NNB+이/JKS
보인다. 보이/VV+ㄴ다/EM+./SF
내일이면 내일/NNG+이/VCP+면/EM
이제 이제/MAG
모두 모두/MAG
끝내고 끝내/VV+고/EM
조금 조금/MAG
쉴 쉬/VV+ㄹ/ETM
수 수/NNB
있을 있/VV+을/ETM
거 거/NNB
같아. 같/VA+아/EM+./SF

1. 기술 설명

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

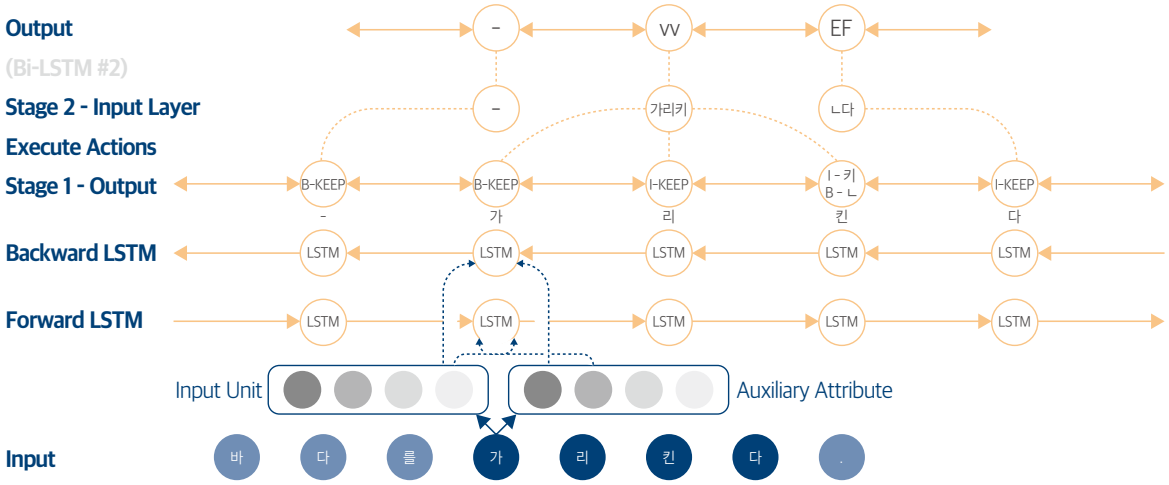


[그림] 형태소 분석 및 품사 부착 과정

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] 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 http://nlplab.iptime.org:32280/unitagger_demo/

4. 실험 (Only PDF)

제안된 방법을 적용하여 구현된 데이터 기반 양방향 LSTM 모델의 성능을 세종 말뭉치를 이용하여 정량적으로 평가한 결과, 언어학적 지식을 활용하지 않은 접근 방법들 중 가장 높은 단어 및 문장 단위 부착 정확도를 보임을 확인하였다.

Text Input

남북은 고위급회담을 13일 판문점 북측 통일각에서 개최할 예정이라고 통일부가 9일 밝혔다.

Analyze

Tagging Result

남북/NNP
은/JX
고위급/NNG
회담/NNG
을/JKO
13/SN
일/NNB
판문점/NNP
북/NNG
측/XSN
통일각/NNP
에서/JKB
개최/NNG
하/XSV
ㄹ/ETM
예정/NNG
이/VCP
라고/EC
통일부가/NF
9/SN
일/NNB
밝히/W
었/EP
다/EF
./SF

1. 기술 설명

- 개체명 인식기는 텍스트에서 인식시킬 개체를 정의하여 해당 개체를 인식시키는 기술로 본 개체명 인식기는 5개의 클래스[인물(PS), 장소(LC), 기관(OG), 시간(TI), 날짜(DT)]를 정의하였으며, 해당 개체에 한국 문화적 특성을 반영하였다.
- 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.

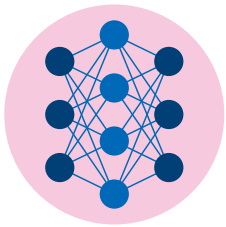
2. 기술 방법

- 한국어 기반으로 구축한 말뭉치의 전처리 과정을 통해 BI-LSTM-CNN-CRF 모델을 학습시킨다.

BI - LSTM - CNN - CRF 모델

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.



개체명 인식 결과

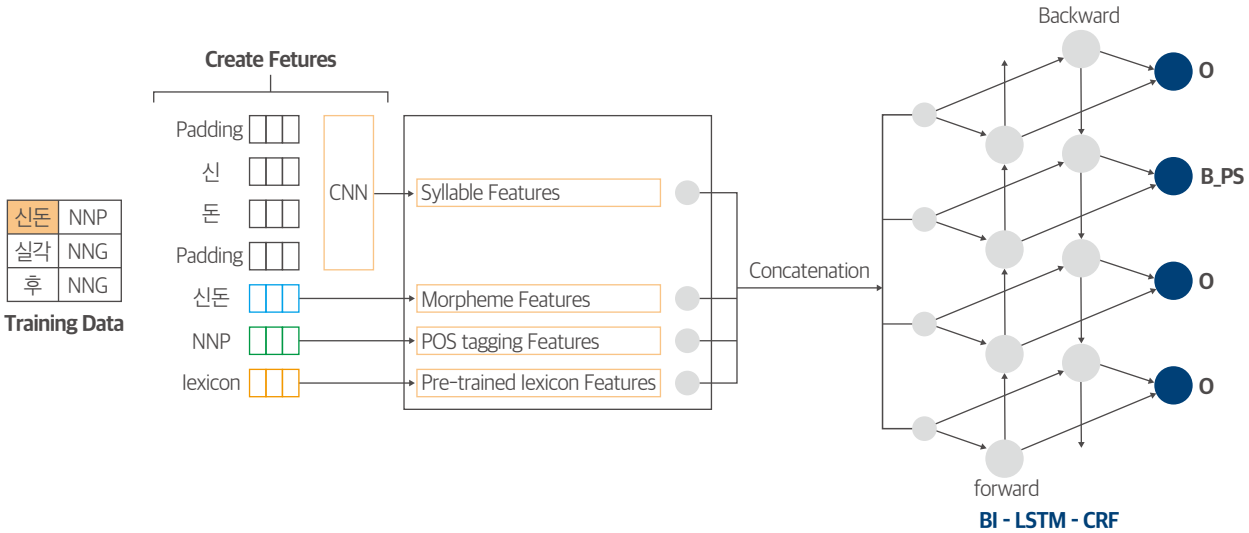
백제는 한국의 고대 국가 중 하나로, 고구려, 신라 와 함께 삼국 시대를 구성하였다. 시조는 부여·고구려에서 남하한 온조 집단으로 마한 54개 연맹체 중 하나인 백제국으로 시작해, 4세기 중엽 근초고왕 때 마한 전체를 통일했다.

- 학습된 모델에 텍스트를 입력으로 넣어 해당 문장에서 개체명으로 인식 가능한 개체를 확인할 수 있다.

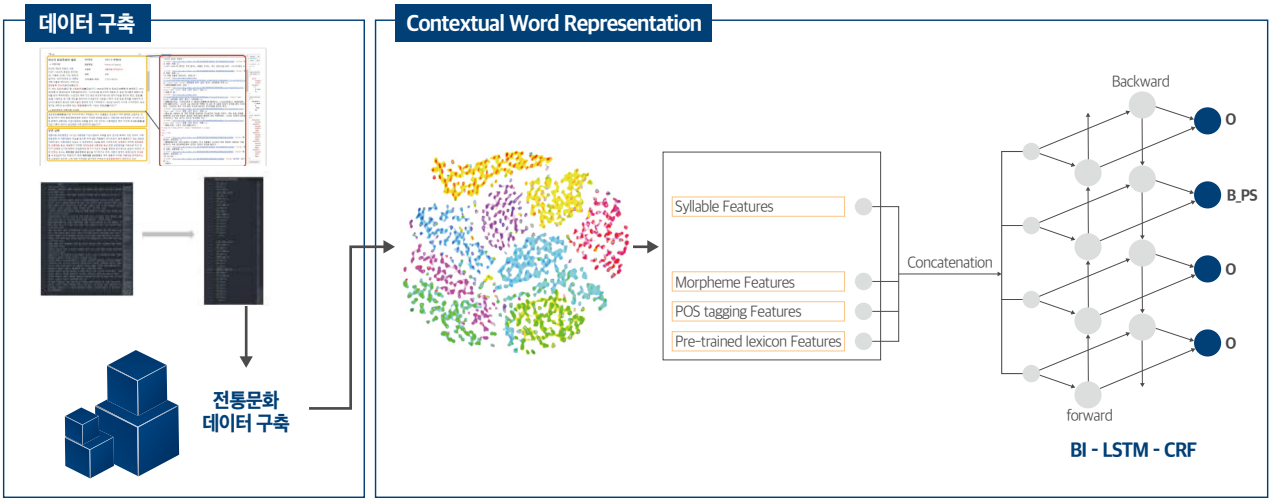
3. 기술 활용 및 응용 분야

- 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

4. 상세 기술 설명 및 실험(Only PDF)

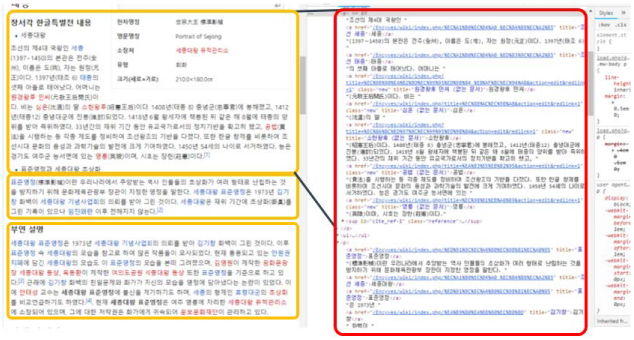


- 구축된 전통문화 데이터를 사용하여 모델을 학습시킨다.
- 자질 형성을 위해 첫 번째는 CNN을 통한 음절 단위 자질, 두 번째는 형태소 단위의 Glove vector 자질, 세 번째 품사 태깅 자질, 구축된 사전을 활용한 사전 자질을 BI-LSTM의 입력 데이터로 사용한다
- Hidden Layer를 통해 계산된 데이터는 최종적으로 CRF의 입력으로 사용하여 전이 확률값을 계산한 후 최종적으로 입력 값에 해당하는 개체명을 예측한다



입력문장	Input : 조선 세종은 조선의 제4대 군주이며 언어학자이다. 그의 업적에 대한 존경의 의미를 담은 명칭인 세종대왕으로 자주 일컬어진다. 성은 이, 휘는 도 본관은 전주, 자는 원정 아명은 막동이다
형태소 단위 개체명 분석 결과	형태소 단위 개체명 분석 결과 : {"조선" : "B_LC", "세종" : "B_PS", "은" : "O", "4" : "O", "군주 " : "O", "이" : "O", "며" : "O", "언어학자" : "O", "다" : "O", "." : "O", "그" : "O", "업적" : "O", "에" : "O", "대하" : "O", "ㄴ" : "O", "존경" : "O", "의미" : "O", "를" : "O", "담" : "O", "명칭" : "O", "대왕" : "O", "으로" : "O", "자주" : "O", "일컫" : "O", "어" : "O", "지" : "O", "ㄴ" : "O", "ㄴ다" : "O", "성" : "O", "," : "O", "휘" : "O", "는" : "O", "도" : "O", "본관" : "O", "전주" : "B_LC", "자" : "O", "원정" : "O", "아명" : "O", "막" : "O", "동이" : "O", }
개체명 인식 결과	개체명 인식 결과 : "<조선 세종 : PS>은 <조선 : LS>의 제4대 군주이며 언어학자이다. 그의 업적에 대한 존경의 의미를 담은 명칭인 <세종 : PS>대왕으로 일컬어진다. 성은 이, 휘는 도, 본관은 <전주 : LS>, 자는 원정, 아명은 막동이다."
태그 결과	태그결과 : ('조선' : 'LC') ('세종' : 'PS') ('전주' : 'LC')

- 전체 흐름은 왼쪽 그림과 같다.
- 구축한 데이터를 바탕으로 다양한 자질들을 생성 및 모델의 입력으로 사용하여 모델을 학습시킨 후, 사용자가 입력한 입력값에 해당하는 개체명을 예측하여 결과로 나타낸다.



- 각 기사의 개요와 내용에 대한 크롤링 과정
- 전체 2351개의 기사로부터 4702개의 문장과 15만 형태소 단위의 말뭉치를 추출했다.
- 태깅 방식은 BIO(Begin, Inside, Outside)를 활용하고 각 태그명 앞에 'B_'를 붙여 태그의 시작을 표기하고 연결된 어미는 'I'로 앞단어와 연결성을 나타낸다.
- 각 태그 중 인물(PS)이 가장 많이 태그되었으며 날짜(DT), 장소(LC) 순서로 태그 개수가 많은 것을 확인할 수 있다.

Category	Count	Frequency
B_PS(Person)	4231	2.92%
B_DT(Date)	2399	2%
B_LC(Location)	2217	1.53%
B_OG(Organization)	740	0.51%
B_TI(Time)	53	0.04%
I (Tag I)	3765	2.6%

- 실험 결과 (+데모)

Feature Representation	Accuracy	F1-score	
morpheme	97.4	78.4	기존 모델
morpheme + grapheme	97.5	84.1	
morpheme + syllable	97.8	86.2	
morpheme + syllable + POS tagging	98.3	88.1	
morpheme + syllable + POS tagging + lexicon	98.9	89.4	제안 모델

- 본 모델은 음절, 형태소, 품사 태깅, 사전 기반 자질을 Feature로 활용하여 Accuracy 98.9%, F1-score 89.4%로 기존 모델에 비해 가장 높은 성능을 보였다.

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

4. 결과 화면 (Only PDF)

- 분류하고자 하는 문서를 입력하면 해당 문서의 분류 결과가 5순위까지 출력

< 문서 분류 시스템 >

본 문서 분류기는 kNN(k Nearest Neighboring) 학습 방법을 이용한 문서 분류기의 데모시스템입니다.

인터넷 문서 5,000여개에서 추출한 자질 중 실험적으로 가장 높은 성능을 보인 20,00개의 자질을 추출하여 정보 검색 기법에서 사용되는 TF/IDF 기법을 이용하여 자질의 Weight값을 만들었고, Nearest Neighbor을 추출하기 위하여 Cosine Measure를 사용하고 있습니다.

아래의 창에 분류하고자 하는 문서를 입력하고 분류하기 버튼을 누르시면 해당 문서의 분류 결과가 순위별로 나타납니다.(5순위까지 출력됩니다.) [분류표보기]

> 뇌졸중은 전 세계의 많은 사람들에게 영향을 미치는 질병으로, 뇌졸중에 걸린 사람은 대개 후유증으로 장애를 입게 된다. 그래서 환자 본인과 가족들의 부담을 덜기 위한 재활훈련 및 치료 과정이 크게 발전했다. 그러나 뇌졸중 재활을 위해서는 반복적인 연습이 필요하다. 뇌졸중 및 뇌 혈관 센터 물리 치료 및 재활 의학과의 장원혁 및 김윤희 연구원이 지적한 바에 따르면 뇌졸중 환자는 고도의 집중 훈련뿐만 아니라 특정한 기능적 업무를 수행해야 하며 이 과정은 상당히 노동 집약적이다. 두 사람은 로봇을 활용하는 치료법이 뇌졸중 재활 분야에서 잠재적인 가능성을 보일 수 있다고 말했다.

분류하기

QUERY : 뇌졸중은 전 세계의 많은 사람들에게 영향을 미치는 질병으로, 뇌졸중에 걸린 사람은 대개 후유증으로 장애를 입게 된다. 그래서 환자 본인과 가족들의 부담을 덜기 위한 재활훈련 및 치료 과정이 크게 발전했다. 그러나 뇌졸중 재활을 위해서는 반복적인 연습이 필요하다. 뇌졸중 및 뇌 혈관 센터 물리 치료 및 재활 의학과의 장원혁 및 김윤희 연구원이 지적한 바에 따르면 뇌졸중 환자는 고도의 집중 훈련뿐만 아니라 특정한 기능적 업무를 수행해야 하며 이 과정은 상당히 노동 집약적이다. 두 사람은 로봇을 활용하는 치료법이 뇌졸중 재활 분야에서 잠재적인 가능성을 보일 수 있다고 말했다.

RESULT : 순위 - 중분류 - 분류코드

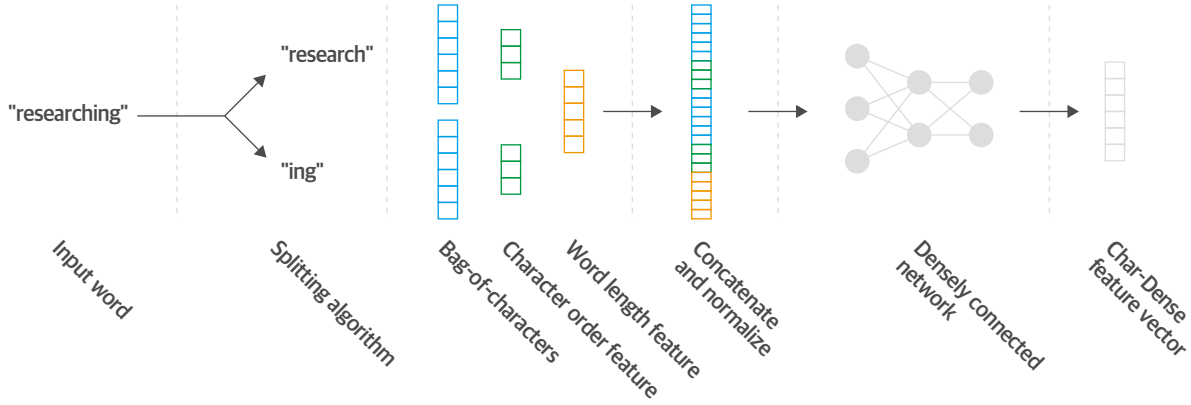
1. 질병/증상 0214
2. 대체의학 0203
3. 약/약학 0208
4. 응급처치 0209
5. 건강상식 0201

07 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.

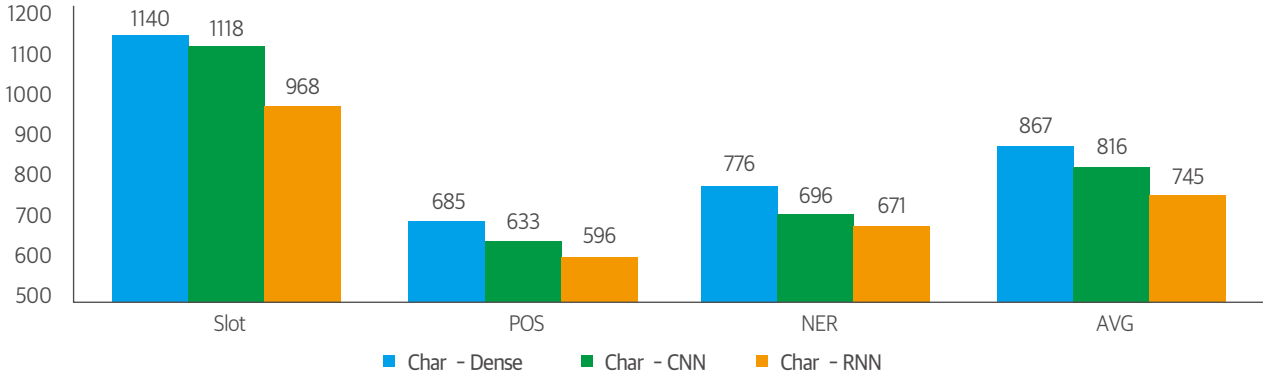
Sparse vector를 하나의 은닉층이 있는 완전연결 신경망의 입력으로 사용해서 최종적인 문자단위 자질 벡터를 생성한다.

3. 기술 활용 및 응용 분야

품사 부착, 개채명 인식, 자연어 처리

4. 실험 (Only PDF)

품사 부착, 개체명 인식, 슬롯 인식 실험을 통해 본 기술의 성능을 검증하였다. 실험 결과 슬롯 인식 정확도 96.62%, 품사 부착 정확도 97.73%, 개체명 인식 F-score 91.21을 기록하였다. 이는 기존 최신 기술보다 크게 앞서거나 비슷한 수준의 성능이다. 또한, 본 기술은 기존 기술 대비 문장 처리 속도가 빠른 것으로 나타났다.

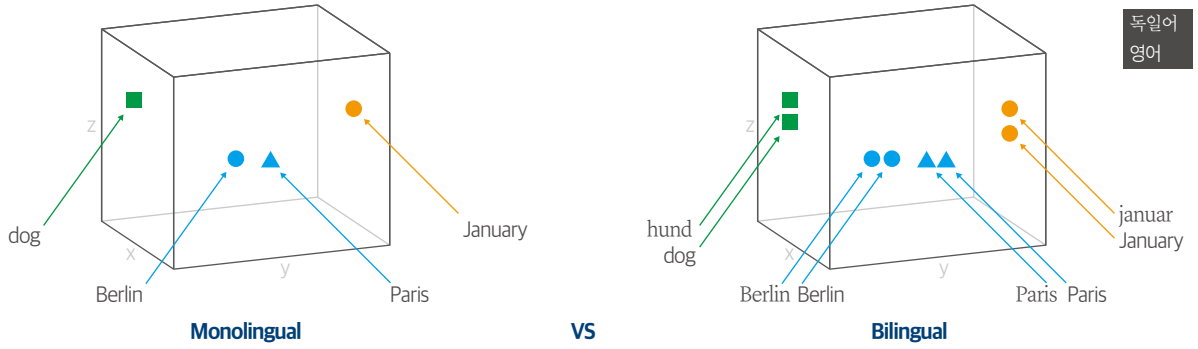


[그림] 초당 처리 문장 수. Char-Dense가 본 기술로, 경쟁 기술 대비 가장 빠른 것을 확인할 수 있다.

08 병렬 코퍼스를 이용한 이중언어 워드 임베딩

1. 기술 설명

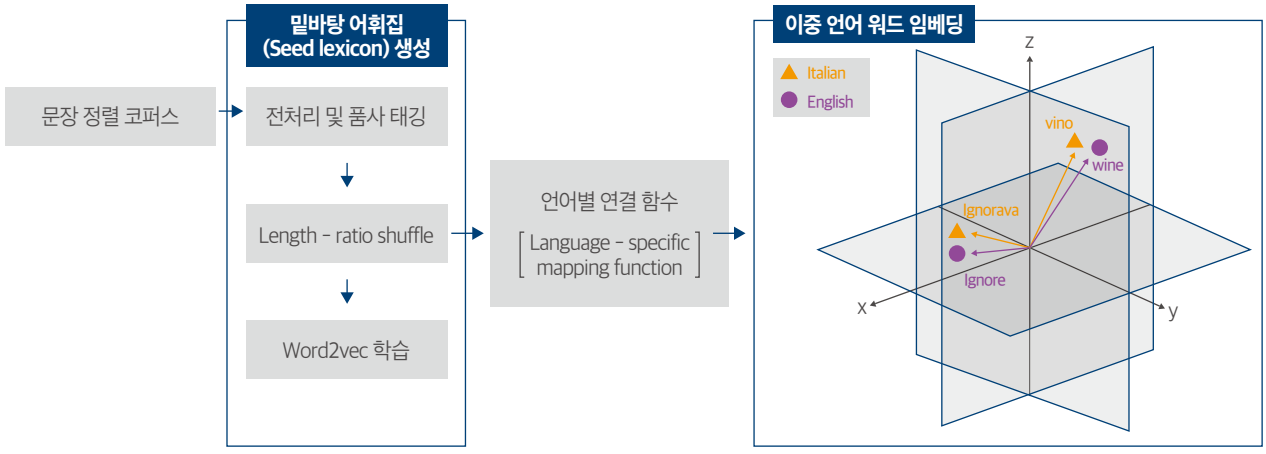
- 워드 임베딩이란 단어를 dense한 실수 벡터 공간에 매핑하되, 단어의 의미가 반영되도록 하는 방법
- 워드 임베딩의 활용방법 중인 하나인 이중 언어 워드 임베딩은 서로 다른 두 언어에서 유사한 의미를 가지는 단어가 유사한 공간에 매핑(mapping) 되도록 하는 것을 목표로 하 는데, 기계번역 분야에서 많은 연구가 이루어지고 있음



<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

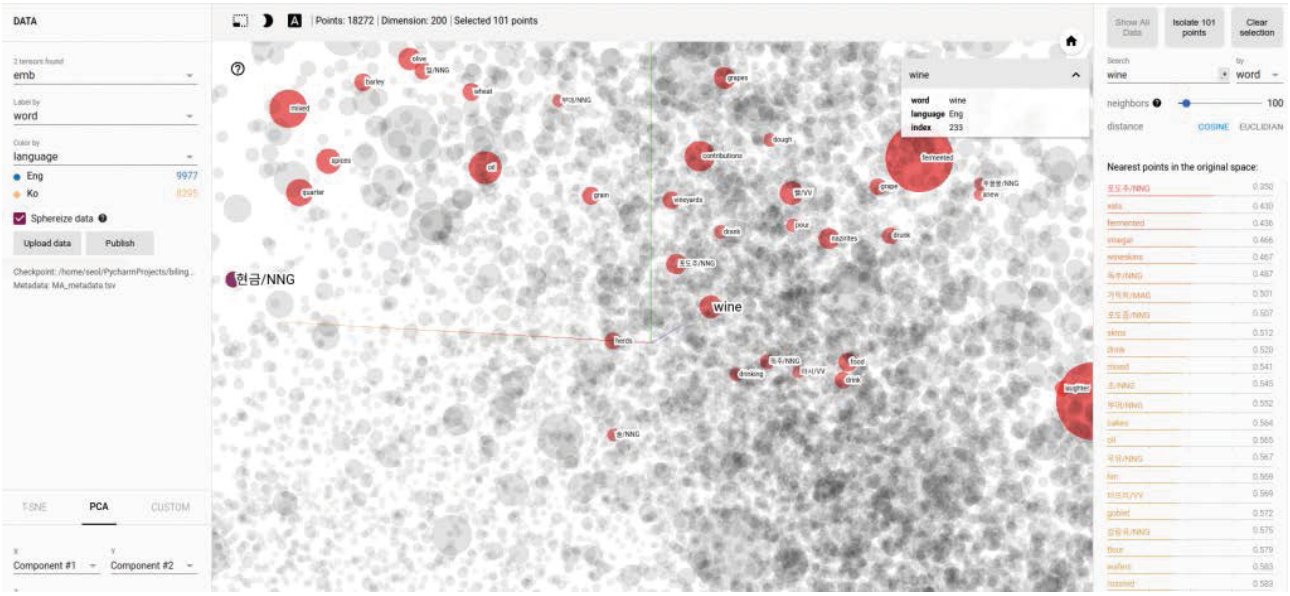
- 본 기술은 다중 언어에 대한 번역기에 활용될 수 있으며, 다중 언어 문서에서 정보검색 모델에서도 활용될 수 있다.
- 데모 http://nlplab.iptime.org:4321/seol2/mt/projector.html

4. 실험 (Only PDF)

4.1 실험 개요

• 영화 자막 코퍼스를 seed lexicon으로 이용하고, wikipedia를 통해 어휘를 확장하였다. 본 실험에서는 한국어-영어를 이용한 이중언어 임베딩을 수행하였다.

4.2 실험 결과



<Bilingual word embedding 시각화 결과 예시>

• 본 기술의 결과는 데모에서 확인가능하며, tensorboard를 이용하여 시각화하였다. 시각화 결과는 한국어와 영어에 대한 seed lexicon으로, 이중언어임베딩의 상위 5k 쌍을 가지고 시각화 하였다. 특정 단어를 검색하면 벡터공간에서 검색한 단어에 대해 제일 가까운 위치의 단어들을 시각화하여 보여준다.

1. 기술 설명

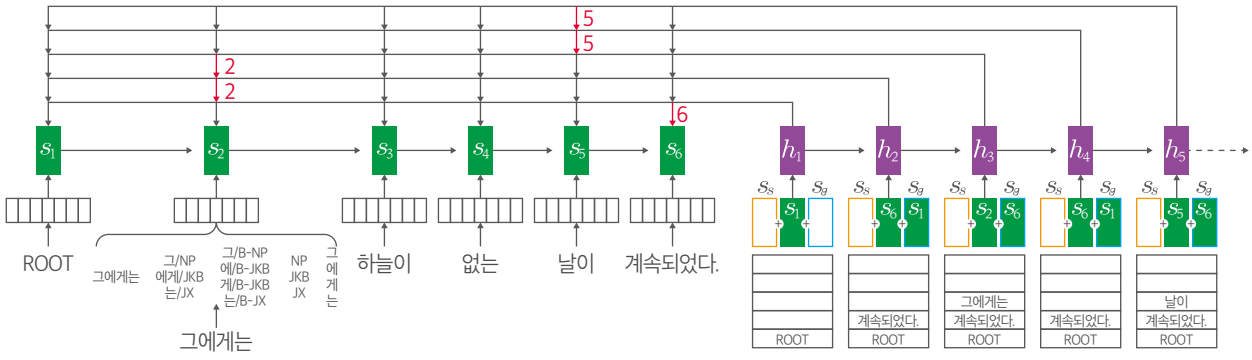
• 의존 구문 분석 기술은 자연어 문장에 포함된 단어들의 의존 관계를 분석하는 기술

• 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



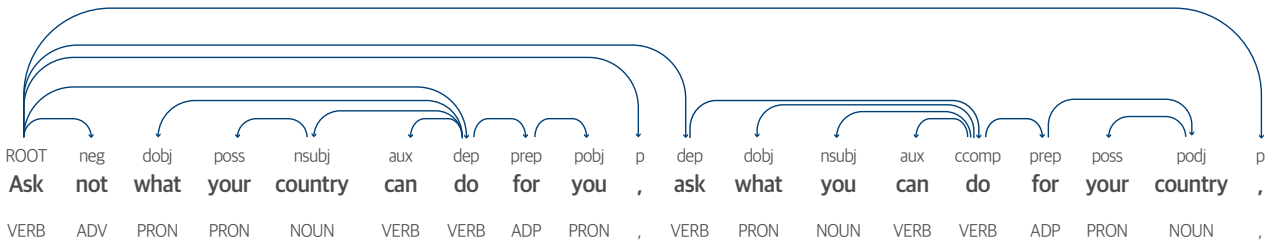
3. 기술 활용 및 응용 분야

• 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] “Ask not what your country can do for you, ask what you can do for your country.”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

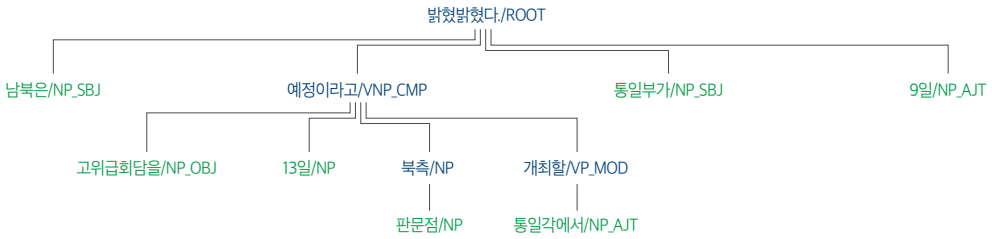
의존구문분석, 대화 시스템, 자연어처리

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

4. 실험 (Only PDF)

Sentence to Parse : 남북은 고위급회담을 13일 판문점 북측 통일각에서 개최할 예정이라고 통일부가 9일 밝혔다.

Visualization



CoNLL-U Output

Details

1 POS Tagging | 2 Dependency Parsing | 3 Reassembly | D Debug Logs

ID	FORM	LEMMA	UPOSTAG	XPOSTAG	FEATS	HEAD	DEPREL
1	남북은	남북/NNP + 은/JX	_	_	_	11	NP_SBJ
2	고위급회담을	고위급/NNG + 회담/NNG + 을/JKO	_	_	_	8	NP_OBJ
3	13일	13/SN + 일/NNB	_	_	_	8	NP
4	판문점	판문점/NNP	_	_	_	5	NP
5	북측	북/NNP + 측/NNB	_	_	_	8	NP
6	통일각에서	통일/NNG + 각/NNG + 에서/JKB	_	_	_	7	NP_AJT
7	개최할	개최/NNG + 하/XSV + ㄹ/ETM	_	_	_	8	VP_MOD
8	예정이라고	예정/NNG + 이/VCP + 라고/EC	_	_	_	11	VNP_CMP
9	통일부가	통일부/NNG + 가/JKS	_	_	_	11	NP_SBJ
10	9일	9/SN + 일/NNB	_	_	_	11	NP_AJT
11	밝혔밝혔다.	밝히/VV + 었/EP + 다/EF + ./SF	_	_	_	0	ROOT

[그림] 본 기술로 “남북은 고위급회담을 13일 판문점 북측 통일각에서 개최할 예정이라고 통일부가 9일 밝혔다”라는 문장의 의존구문분석을 진행한 결과

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

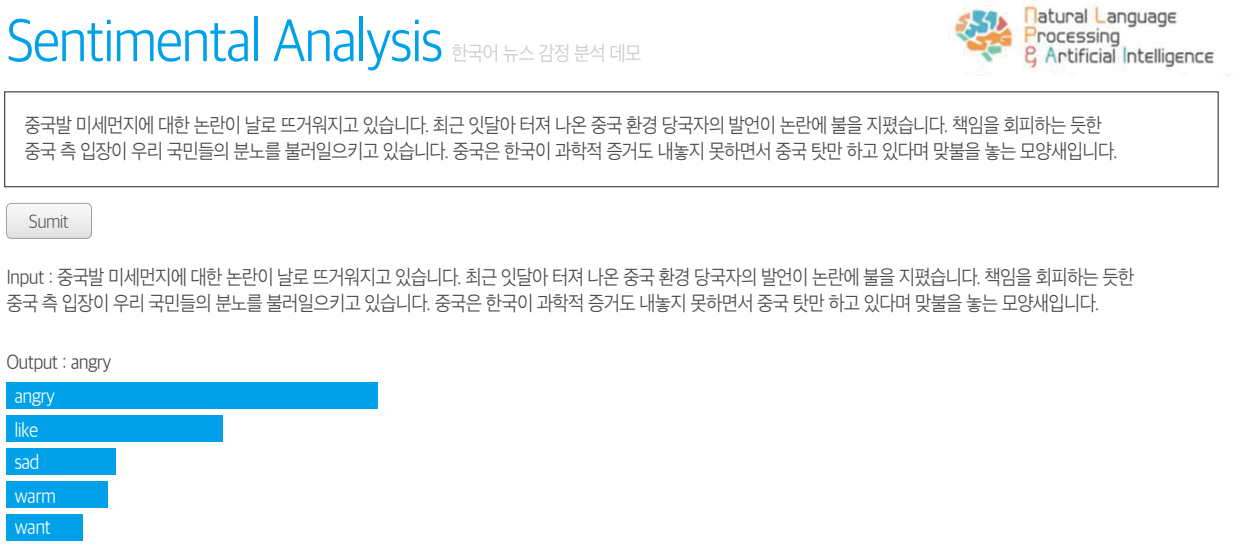
4. 실험 (Only PDF)

4.1 실험 개요

- slot filling 모델의 평가 방법으로는 f-1 score를 이용하였으며, TGT는 적대 학습을 적용하지 않고 slot filling 모델을 학습한 경우를 나타냄. 적대 학습을 적용한 도메인 분류 손실 함수를 얼마나 반영할지는 ! 계수의 정도에 따라 성능을 측정하였음
- 실험 결과 가중치를 부여하여 적대 학습 방법을 적용할 때가 기존의 적대 학습 방법을 적용하지 않은 경우보다 66.10에서 67.12로 약 1% 가량의 F-1 Score 뛰어난 향상이 있었음

Source \ Target		MIT Rest.
ATIS	TGT	66.10
	ADV(A = 1)	65.32
	ADV(A = 0.1)	**67.12**
	ADV(A = 0.01)	66.41

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)

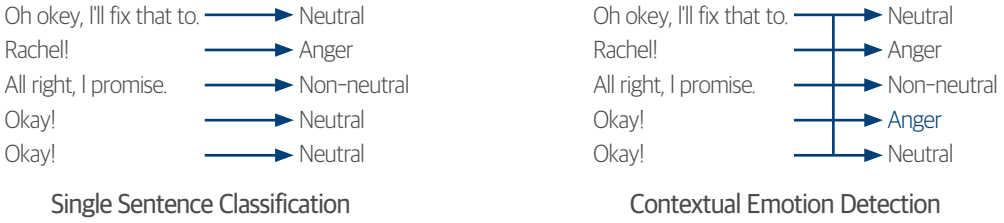
3. 기술 활용 및 응용 분야

- 본 기술은 적은 컴퓨팅 자원을 이용하며, 텍스트로부터 의미 있는 특징(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

13 대화속 화자의 감성 분석(Emotion Recognition in Conversation)

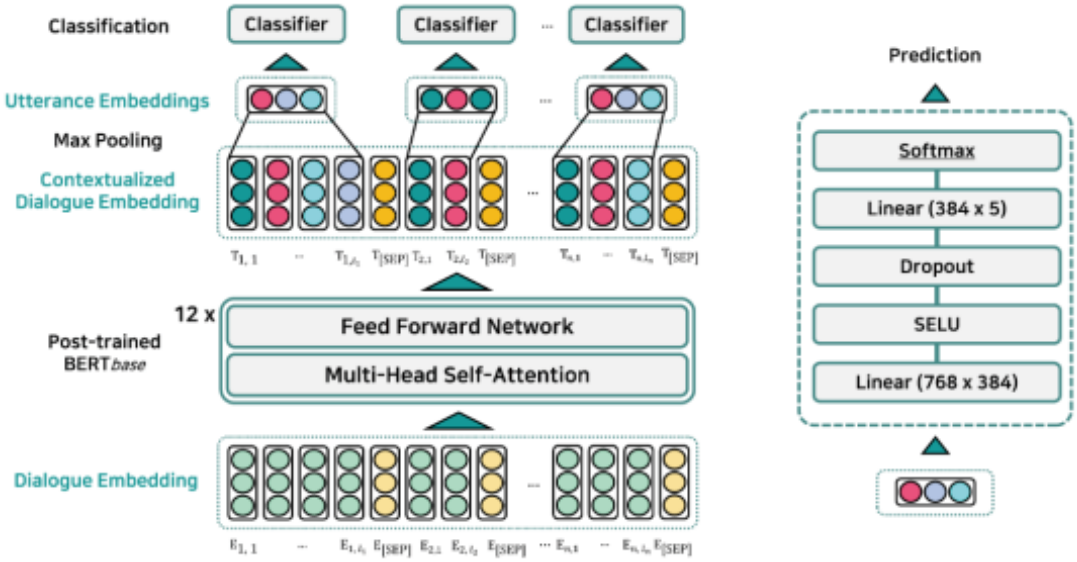
1. 기술 설명

대화속 화자의 감성 분석(Emotion Recognition in Conversation)



2. 기술 방법

- 대화(dialogue)를 BERT tokenizer로 분절화(tokenization)하고, 각 발화(utterance)가 끝나는 지점에 구분자로서 [SEP] token을 추가함
- 분절화된 대화 데이터를 BERT-base 모델로 인코딩하고, 각 발화에 해당하는 tokens를 max-pooling하여 deep contextualized utterance representations를 생성함
- 해당 utterance의 representations를 바탕으로 분류(classification)을 수행함



3. 기술 활용 및 응용 분야

- 본 기술은 AI 챗봇, 채팅 분석 등의 서비스에서 사용자 경험을 향상시키기 위한 감정 분석 모듈로 활용될 수 있다.
- 데모 http://nlplab.iptime.org:32290/

4. 실험 (Only PDF)

4.1 실험 개요

- 영어 일상 대화 데이터인 Friends와 채팅 데이터인 EmotionPush을 이용함
- BERT-base, BERT-large, RoBERTa-base, RoBERTa-large 모델이 대화 내 감정 분석을 수행하도록 학습함

4.2 실험 결과

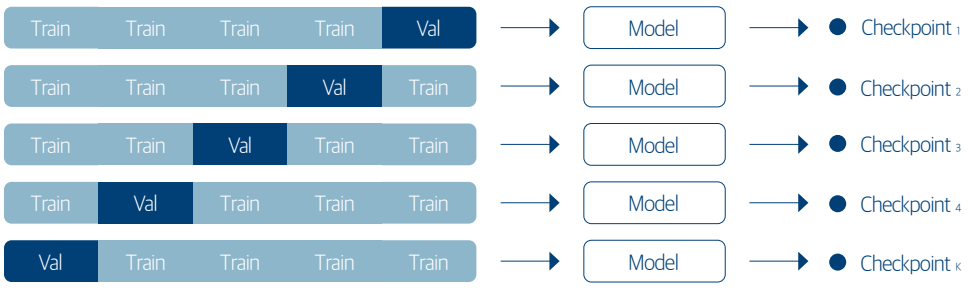
• RoBERTa-large-cased 모델을 사용했을 때 RNN 기반, GCN 기반의 이전 모델들보다 공통적으로 약 5% 높은 성능을 보였다.

Dataset	Model	meutral	joy	sadness	anger	surprise	disgust	fear	non-meutral	w-avg
Friends	DialogueRNN[4]	73.52	55.62	32.35	37.90	46.11	19.92	5.31	32.46	55.25
	DialogueGCN[5]	73.63	58.44	37.31	37.19	53.18	21.94	4.35	31.88	56.77
	BERT-base*	77.08	61.36	40.48	37.25	53.68	21.74	8.11	33.62	58.69
	BERT-large*	77.37	61.51	43.79	44.14	52.43	25.69	12.50	34.81	59.75
	RoBERTa-base*	77.26	**67.56**	42.25	46.11	51.93	26.55	11.76	35.10	60.26
	RoBERTa-large*	**77.48**	67.17	**45.03**	**51.57**	**55.77**	**29.82**	**15.87**	**37.76**	**61.17**
Emotion Push	DialogueRNN[4]	82.44	63.85	31.22	15.56	35.15	8.33	0.00	17.89	69.56
	DialogueGCN[5]	83.63	64.07	38.91	27.16	35.36	10.00	0.00	13.18	70.41
	BERT-base*	85.90	63.71	47.37	31.75	45.77	10.00	0.00	20.77	73.25
	BERT-large*	86.41	**69.60**	44.30	**41.27**	45.00	**40.00**	0.00	21.08	74.41
	RoBERTa-base*	**86.87**	69.22	49.33	35.09	47.67	22.22	0.00	22.51	75.29
	RoBERTa-large*	86.27	69.24	**50.91**	26.67	**54.82**	33.34	0.00	**28.11**	**75.97**

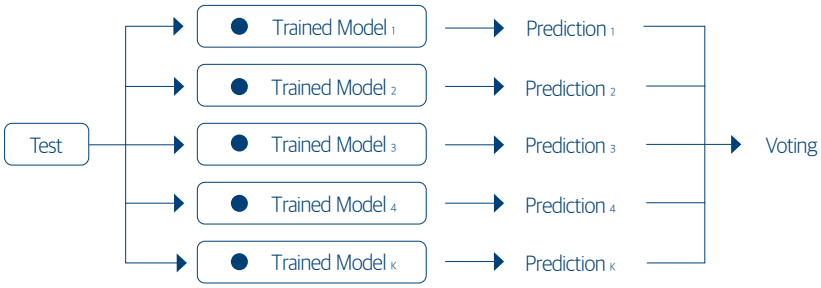
1. 기술 설명

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

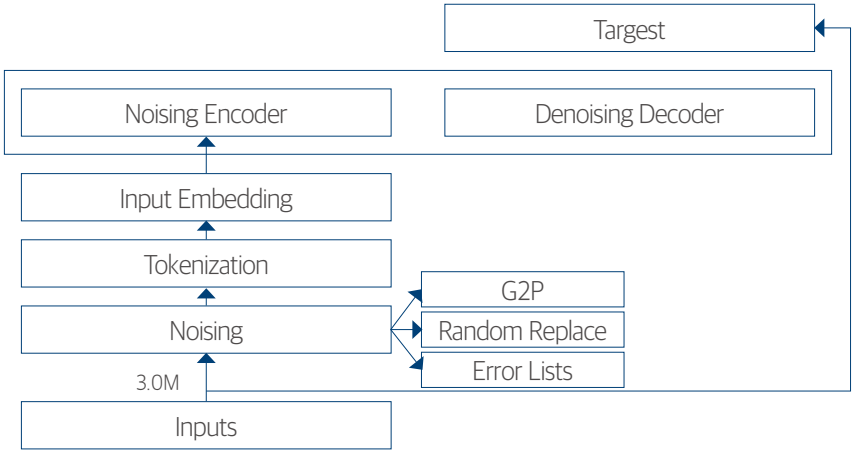
본 기술은 별도의 연산을 피하면서 분산 경감 면에서도 강점을 가지는 교차 검증 앙상블(Cross-Validated Ensenble, CVE)기법이다. 이는 Bagging처럼 여러 샘플을 추출해 학습하는 효과를 얻는 동시에 교차 검증시 기록된 checkpoints로 앙상블하므로 별도의 연산이 요구되지 않는다. 교차 검증 앙상블 기법은 다음과 같은 단계로 진행된다.

- 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.

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

• 데모 http://nlplab.iptime.org:32288/

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

지금 어디가세용

Translate

맞춤법 교정 결과

지금 어디가세요?

1. 기술 설명

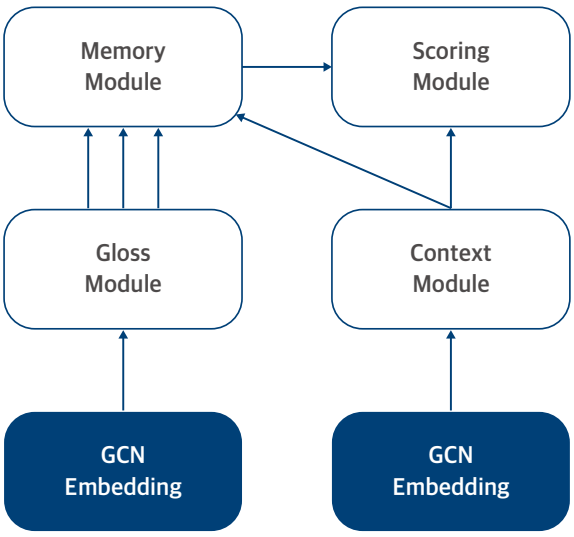
단어 중의성 해소란 두 개 이상의 의미를 가진 단어를 문장의 쓰임에 따라 정확하게 분석하는 것이다. 본 기술은 단어의 중의성을 해소하는 기술로 단어의 표상에 구문 정보와 의미 관계를 반영할 수 있도록 그래프 임베딩을 활용하였다.

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

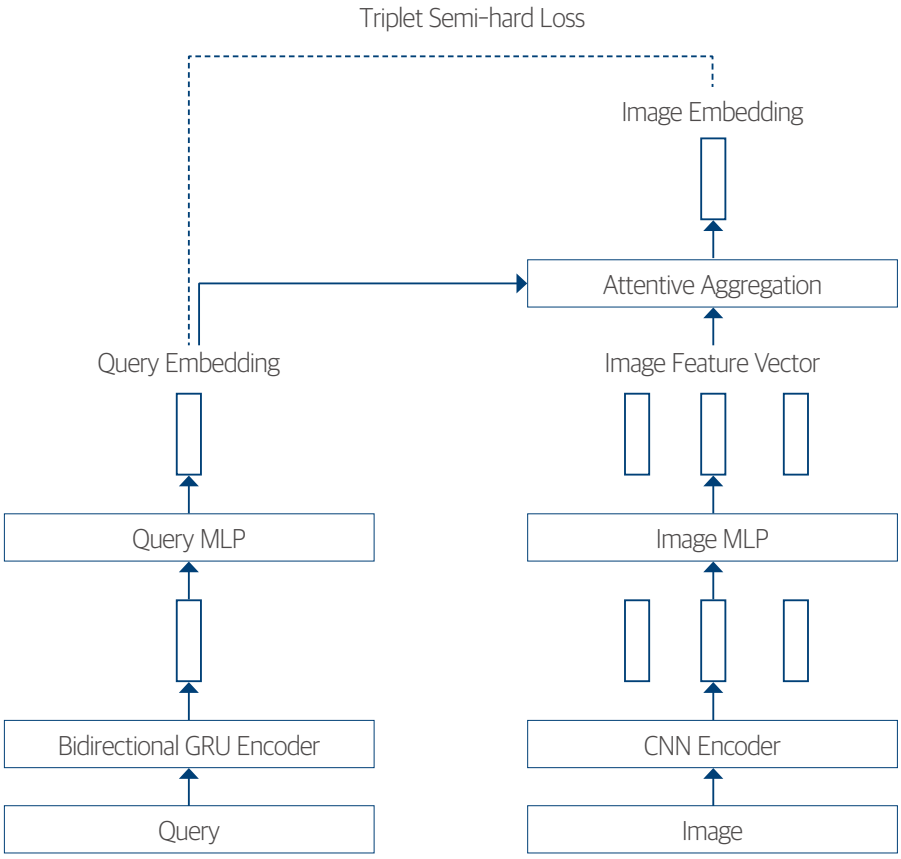
17 Attentive Aggregation(주의적 종합)기반 크로스 모달 임베딩

1. 기술 설명

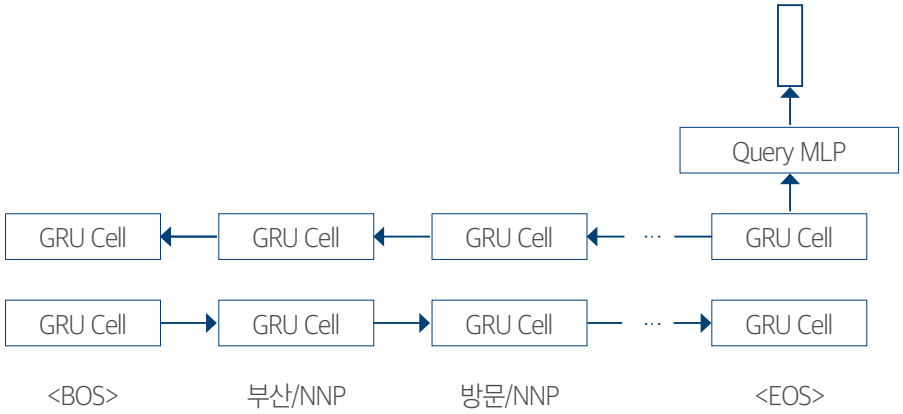
본 기술은 사진 검색을 위한 주의적 종합(Attentive Aggregation)기반의 언어-시각 크로스 모달 임베딩 모델로서 자연어 질의로부터의 사진 검색 과제를 해결할 수 있다. 본 기술은 사진으로부터 여러 개의 특징 벡터를 계산한 뒤 자연어 질의의 임베딩에 따라 Attentive Aggregation을 적용한다. 이는 이미지의 다양한 특징에 선별적으로 집중하여 질의와 사진 간의 유사도를 평가함으로써 언어와 시각 모달 간의 의미적 간극을 크게 줄일 수 있다.

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.

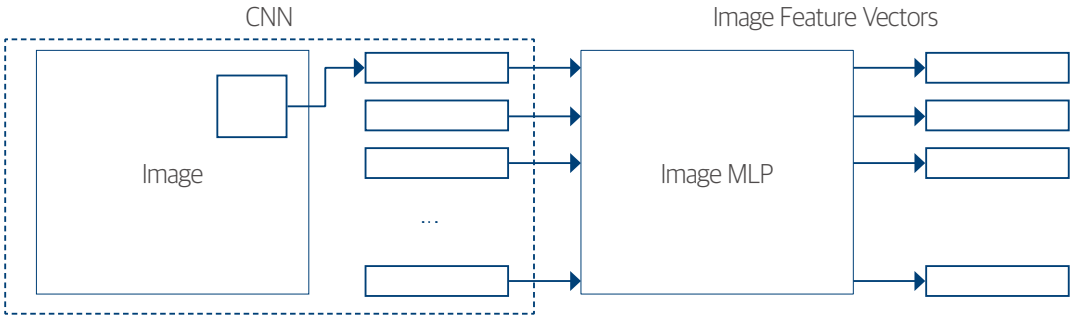


[질의 인코더] 양방향 GRU와 MLP구조로 구성되며, 자연어 질의로부터 하나의 질의 임베딩 벡터를 계산함. 입력으로는 자연어 질의를 분할한 토큰들의 임베딩을 사용하였으며, 본 모델에는 형태소 분석기를 통해 분할한 형태소들 중 질의의 핵심 정보를 나타낼 것으로 예상되는 명사와 동사 형태소를 사용함. 양방향 GRU Layer에서는 토큰들의 임베딩을 입력으로 받아 질의 전체의 정보를 반영한 특징 벡터를 계산하고, MLP Layer에서는 이를 사진 임베딩과 공통벡터 공간에 매핑되는 질의 임베딩으로 변환함



[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

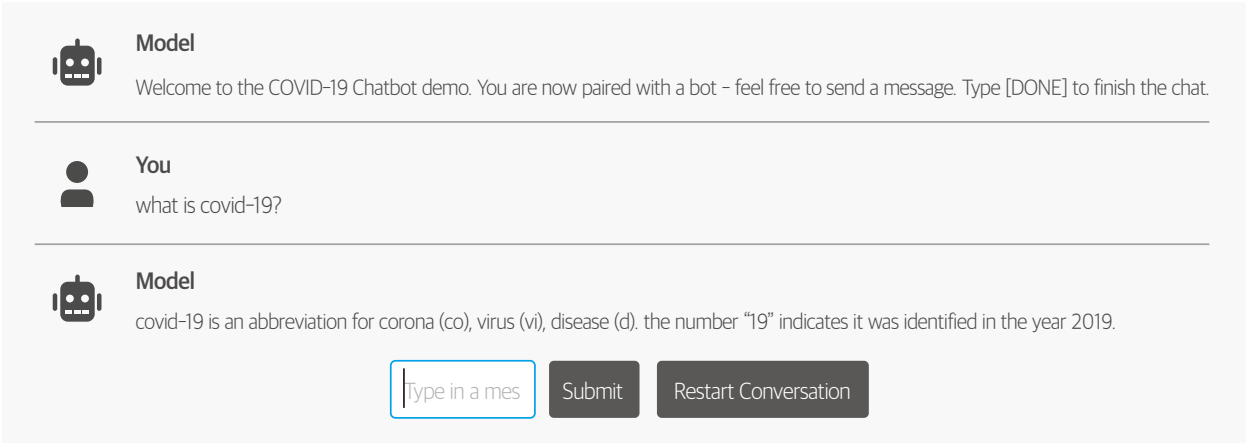


[그림] 사진 인코더 구조

[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.

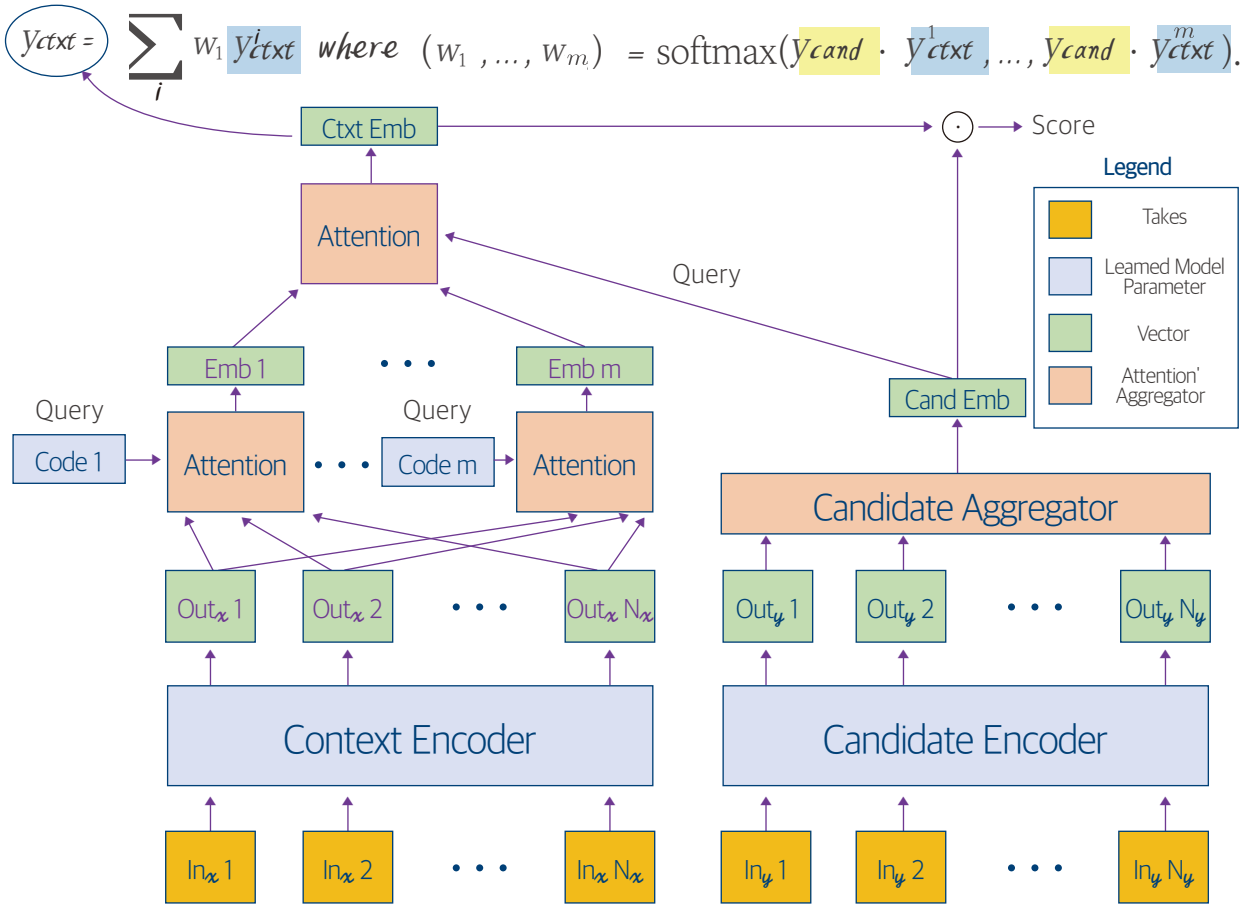
1. 기술 설명

- 사람들에게 질의 응답 시스템을 이용한 손쉬운 접근방법을 통해 COVID-19에 대한 믿을 만한 최신 정보 제공을 할 필요성이 있음
- 질의 응답 시스템은 질문에 대한 빠른 응답 속도와 응답의 높은 정확성을 필요로 함



2. 기술 방법

- 본 기술은 검색 기반의 질의 응답 서비스에서 모델의 응답 속도 및 정확성을 잘 반영할 수 있도록 Poly-encoder 모델을 기반으로 fine-tuning을 수행하였음



3. 기술 활용 및 응용 분야

• 본 기술은 검색기반의 질의 응답 시스템이 가능한 모든 분야에서 사용될 수 있다,

4. 실험 (Only PDF)

4.1 실험 개요

• 크롤링한 COVID-19관련 데이터 (Q-Q / Q-A)를 이용하여 모델 fine-tuning 및 검증실험을 수행하였다. 여기서 Q-Q는 질의와 유사한 질의를 찾아내는 것이며, Q-A는 질의에 대응하는 응답을 찾아내는 것이다.

4.2 실험 결과

Q-A					
	Candidates	Accuracy	F1	BLEU-4	MRR
Poly-encoder (Reddit)	20	0.36	0.46	0.37	0.54
Poly-encoder (Reddit)	10	0.45	0.54	0.46	0.63
JHU-COVID-QA@20(ft) **(ours)**	20	**0.79**	**0.83**	**0.79**	**0.87**
JHU-COVID-QA@20(ft) **(ours)**	10	**0.99**	**0.99**	**0.99**	**0.99**
Q-Q					
	Candidates	Accuracy	F1	BLEU-4	MRR
Poly-encoder (Reddit)	5	0.72	0.79	0.69	0.58
JHU-COVID-QA@20(ft) **(ours)**	5	0.72	**0.81**	0.69	**0.59**

• 검증 메트릭은 다음과 같다. Accuracy, F1, BLEU-4, MRR (Mean Reciprocal Rank)

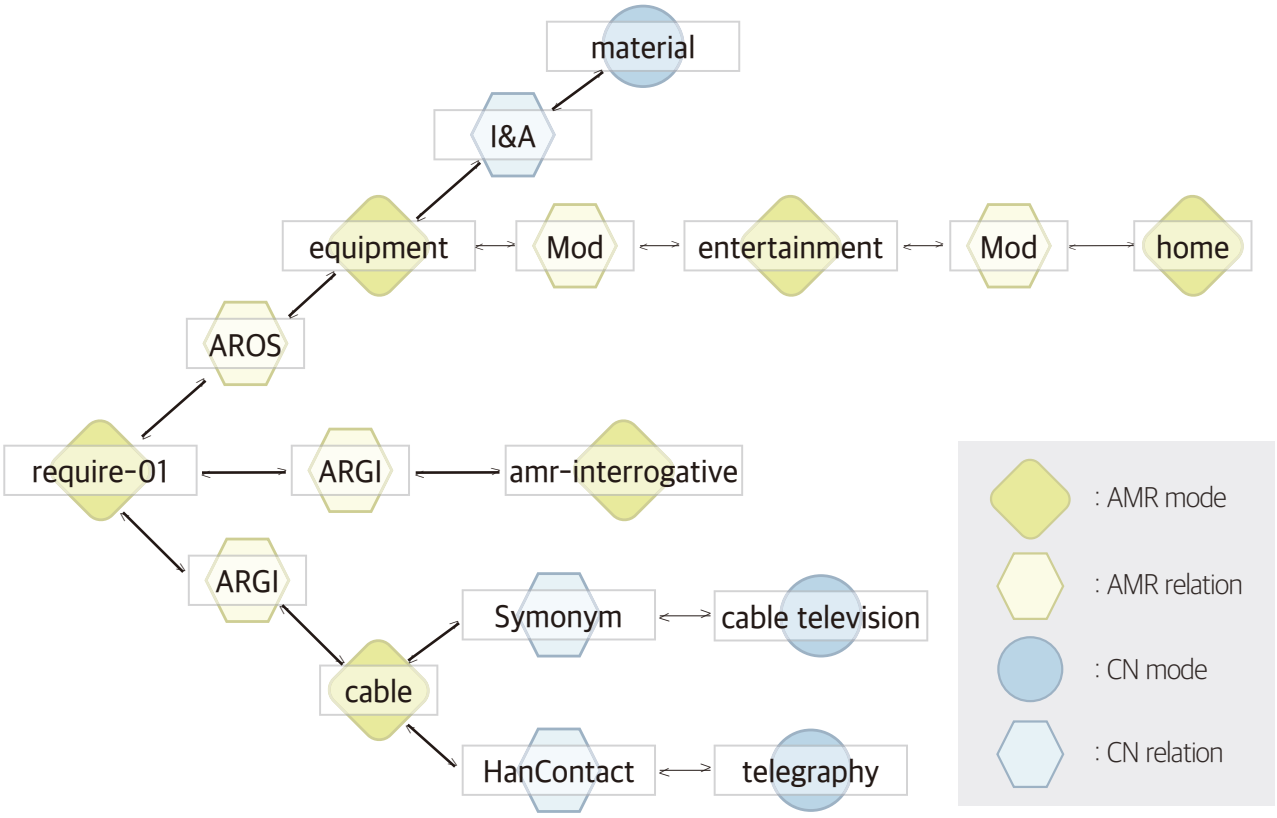
• 실험 결과 제안한 모델 JHU-COVID-QA (OURS)이 기존 베이스라인 (poly-encoder (Reddit)) 모델보다 높은 성능을 보였다.

1. 기술 설명

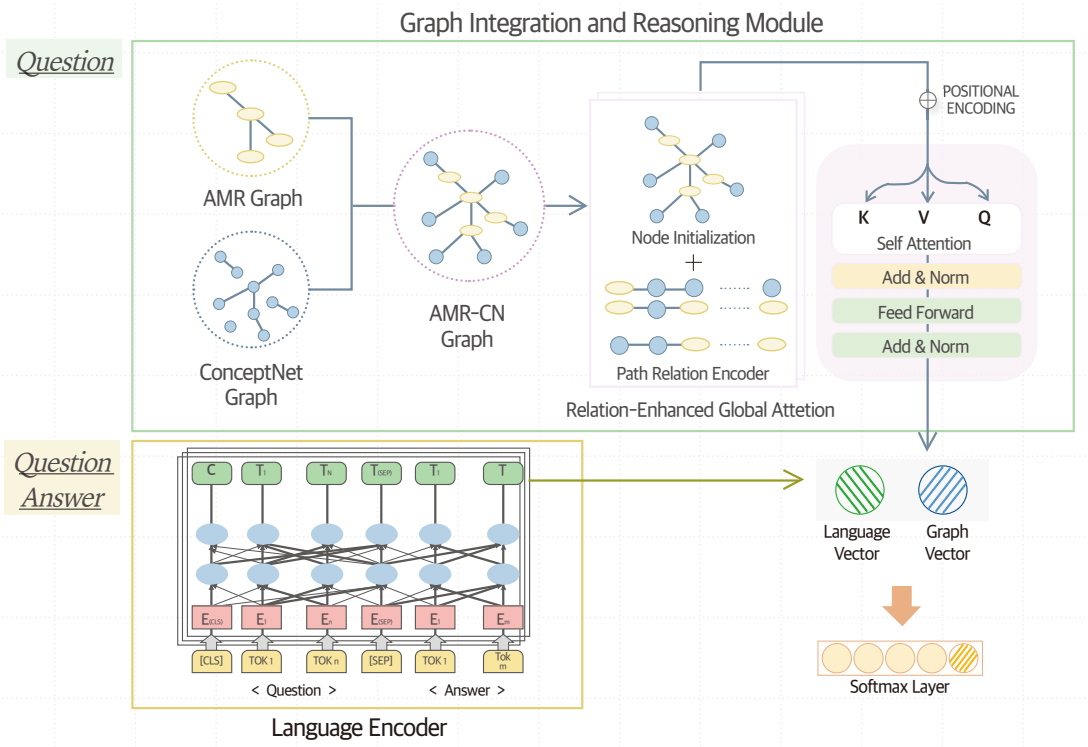
- 상식이란 사회의 사람들, 그리고 일상에서 얻어질 수 있는 지식들을 말함. 상식추론 이란, 이러한 상식 정보들을 이용하여 추론하는 논리적인 과정을 의미함
- 본 기술은 질의에 적합한 상식 그래프 추출하기 위하여 질의를 AMR(Abstract Meaning Representation) 그래프로 변환하고 이를 이용하여 상식 질의응답을 수행하는 기술임
- AMR(Abstract Meaning Representation) 그래프는 주어진 질의의 의미를 그래프 구조로 표현하고, 해석을 용이하게 만들어 줌
- AMR그래프가 가지고 있는 relation 중 ARG0과 ARG1은 프레임 논항 (Frame Argument)으로, 문장 내부에서 핵심적인 역할을 하는 중요 노드들과 연결되어 있음. AMR 구조를 이용하게 된다면, 질의에 대해서 꼭 필요한 상식 그래프만 추출할 수 있음. 효과적으로 상식 그래프(ConceptNet)를 추출하기 위해, 문장 내부에서 핵심적인 역할을 하는 ARG0와 ARG1에만 상식 그래프 확장. 기존의 단어 기반 해석에서 더 나아가, 그래프 경로기반으로 기계의 상식 추론을 해석할 수 있음

2. 기술 방법

- 본 기술이 사용하는 확장 그래프



- 본 기술은 그래프의 경로를 모델에 임베딩하기 위해 AMR-CN 확장 그래프를 relation을 node로 취급하는 Levi Graph 로 변환. Cai (2019)의 Graph Transformer의 encoder 부분을 재구성하여 경로 학습 모델을 구성한 후, 언어모델에서 나온 벡터값을 이용하여 오지선다 문제를 품



3. 기술 활용 및 응용 분야

• 본 기술은 상식 분야에 대한 질의응답을 수행할 수 있음

4. 실험

4.1 실험 개요

• AMR구조를 사용하지 않았을 때의 성능과 사용하였을 때의 성능 비교를 수행함

• 다양한 언어모델에 대한 실험을 수행함

4.2 실험 결과

Language Encoder	Graph type	Ndev-Acc.(%)	Ntest-Acc.(%)
BERT-base-cased	-	51.81	51.59
	AMR-original	52.82	52.78
	CN-full *(CF)*	53.80	53.10
	CN-pruned *(CP)*	52.61	52.53
	AMR-CN-full *(ACF)*	52.98	52.94
	AMR-CN-pruned *(ACP)*	**53.97**	**53.58**

• AMR구조를 사용하지 않았을 때보다 AMR 구조와 상식그래프(ConceptNet)를 통합하였을 때의 성능이 가장 높음

Language Encoder	Ndev-Acc.(%)	Ntest-Acc.(%)
BERT-base-cased	51.81	51.59
XLNet-base-cased	57.98	57.05
ALBERT-base	50.12	49.22
ELECTRA-base	71.25	70.19
BERT-base-cased w/ AMR-CN-pruned *(ACP)*	**53.97**	**53.58**
XLNet-base-cased w/ AMR-CN-pruned *(ACP)*	**61.01**	**60.35**
ALBERT-base w/ AMR-CN-pruned *(ACP)*	**51.51**	**51.08**
ELECTRA-base w/ AMR-CN-pruned *(ACP)*	**71.99**	**70.91**

• 다양한 언어모델에 대해서도 성능이 높아지는 것을 볼 수 있음

20 사전 학습된 Transformer 언어 모델의 이종 언어 간 전이 학습을 통한 자원 희소성 문제 극복 ●○○○

1. 기술 설명

- 방대한 양의 말뭉치와 언어 모델링 태스크를 통해 사전 학습된 Transformer 모델은 자연어처리 시스템의 뼈대로 활용될 시 광범위한 도메인 및 태스크에 걸쳐 큰 폭의 성능 향상을 보임
- 동일한 모델을 사용했을 때, 학습 데이터의 양은 언어 모델 및 하위 자연어처리 시스템의 성능에 가장 큰 영향을 미치는 요소이므로, 언어 자원의 불균형은 이러한 최신 자연어처리 기술이 다양한 언어로 확대되는 과정에 있어 큰 걸림돌
- 본 기술은 언어 모델의 학습 시 이종 언어 간 전이 학습을 사용하여 성능을 향상시킴

2. 기술 방법

- 언어 자원이 풍부한 언어에서 학습된 Transformer 기반 언어 모델에서 얻은 파라미터 중 재활용 가능한 부분을 이용하여 목표 언어의 모델을 초기화한 후 학습을 진행함
- 기존 언어와 신규 언어의 차이를 학습하는 역할을 하는 적응 층들을 추가하여 이종 언어 간 전이 학습을 도움

(a) 원시 언어 모델 (b) TLM (c) TLM+

[그림] 본 기술의 구조도. 학습은 a에서 c순으로 진행됨.

3. 기술 활용 및 응용 분야

본 기술은 사전 학습된 언어 모델을 기반으로 하는 모든 자연어처리 시스템에 적용될 수 있으며, 언어 모델을 사전 학습 시키기 위한 언어 자원이 부족한 상황에서 특히 효과적임

4. 실험 (Only PDF)

4.1 실험 개요

RoBERTa 모델에 본 기술을 적용하고 언어 자원이 희귀한 상황을 가정하여 영어로부터 한국어로의 전이 학습을 실험해본 결과, 전이 학습을 사용하지 않은 기준 모델 대비 perplexity는 47.6% 감소하고, 단어 예측 정확도는 18.0% 향상됨을 확인하였다.

4.2 실험 결과

	Perplexity	단어 예측 정확도 (%)
기준 모델	40.3	42.8
TLM	23.5	48.4
TLM+	**21.1**	**50.5**

[표] 이종 언어 간 전이 학습 실험의 정량적 성능 비교.

○●○○

대화 시스템

대화 시스템에서의 자연스러운 대화를 위한
Memory Attention 기반 Breakdown Detection

검색 기반 대화 시스템에서의 정답 예측 기술

딥러닝 기반 자동 질의응답 시스템

딥러닝 방법을 이용한 발화의 공손함 판단

기계 독해를 이용한 COVID-19 뉴스 도메인의 한국어 질의응답 챗봇

일상대화 생성 모델

시각 질의응답 시스템

화자의 페르소나를 반영한 대화 모델

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-term memory)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

4. 실험

- 본 연구에서 제안한 모델은 다음과 같다.
- TU: memory attention을 적용하지 않은 모델
- TU+S: system memory attention을 적용한 모델
- TU+U: user memory attention을 적용한 모델
- TU+S+U: user and system memory attention을 적용한 모델
- 본 모델에서 정량적 평가는 TU+S와 TU+S+U에서 기존 모델보다 뛰어난 성능을 보였음

		Model								
		Proposed model				CRF Baseline	Majority Baseline	KTH run2	PLECO run1	RSL17BD run2
		TU	TU+S	TU+U	TU+S+U					
Accuracy		0.458	0.464	0.467	**0.47**	0.4285	0.3720	**0.4415**	0.2950	0.4310
F1	F1(B)	0.5146	0.532	0.533	**0.556**	0.3543	0.3343	0.2949	**0.3636**	0.3201
	F1(PB+B)	0.6737	0.6906	0.6679	0.7441	0.76722	**0.8927**	0.7440	0.8744	0.8400

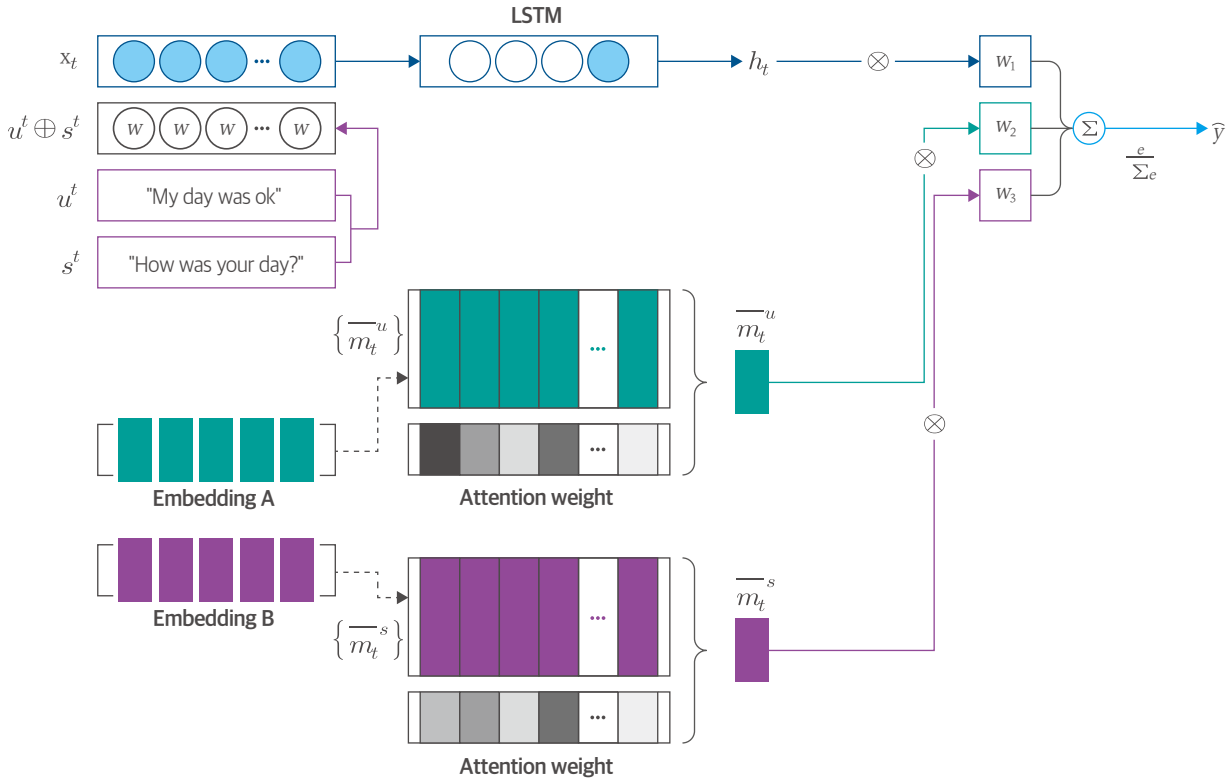
<제안한 모델의 정량적 성능 결과>

다음은 정성적 평가에 대한 것이다. TU+S+U에 대한 정성적 평가 결과이며, 하단 표는 한 다이얼로그에서 발화가 발생할 때, breakdown이 되기까지 attention의 변화를 시각화한 것이다. 실제 마지막에 breakdown이 발생하기까지 문제가 되는 문장들에 대해 모델에서 많은 attention weight를 사용한 것을 확인할 수 있다.

Role	Uttreances	Memory Attention Distribution user memory system memory
User1	Hello how are you?	0.004
System1	Hello, I really like movies. how about we talk about movies?	0.063
User2	That sounds good i like movies. what is your favorite movie?	0.0365
System2	Hello, I really like movies. how about we talk about movies?	0.063
User3	That sounds good i like movies. what is your favorite movie?	0.0365
System3	Ok. Let me know something you're interested in.	0.178
User4	I would like to talk about movies.	0.0366
System4	Ok. Let me know something you're interested in.	0.178
User5	I would like to talk about movies.	0.0366
System5	I like pop music.	0.258
User6	Do you like avril lavigne?	0.849
System6	I like pop music. (Dialogue break down)	0.258

<Memory attention distribution을 통한 모델의 정성적 결과>

5. 모델 개요 (Only PDF)

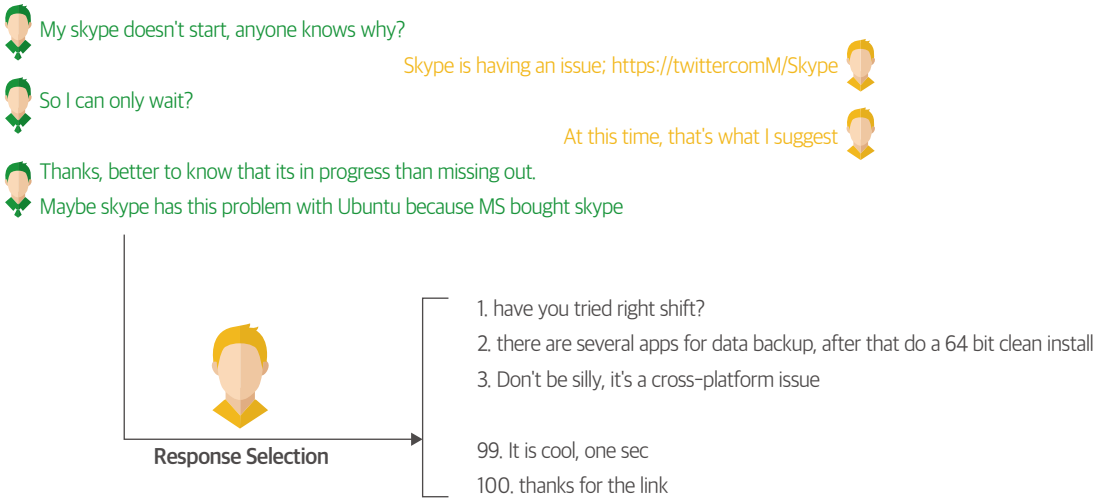


<Memory Attention 기반 Breakdown Detection 모델 개요>

• 위의 그림에서 $\oplus$, $\otimes$, Σ, w는 각각 concatenation, 매트릭스 multiplication, summation, 문장을 구성하는 단어를 의미한다. 본 모델의 학습 과정은 다음과 같다. (1) 사용자 발화 및 시스템 발화에 대한 sentence representation을 수행한다. (2) 현재 시점 t에서 시스템 $s^t = \{w_1^t, w^{t_2}, \ldots, w_n^t\}$ 및 사용자 $u^t = \{w_1^t, w^{t_2}, \ldots, w_n^t\}$의 발화를 인코딩 (Encoding)하기 위해 LSTM을 이용하여 h_t를 도출한다. (3) LSTM으로부터 획득한 인코딩 벡터와 현재 시점에서 모든 이전 시스템 발화에 대한 memory를 저장하여 attention을 이용한 attention weight 값을 도출한다. ($\overline{m_t}^u$, $\overline{m_t}^s$는 각각 사용자, 시스템 발화에 대한 memory context 벡터이다.) (4) 마지막으로 대화 시스템내의 발화에서 breakdown을 예측한다.

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

4. 실험 (Only PDF)

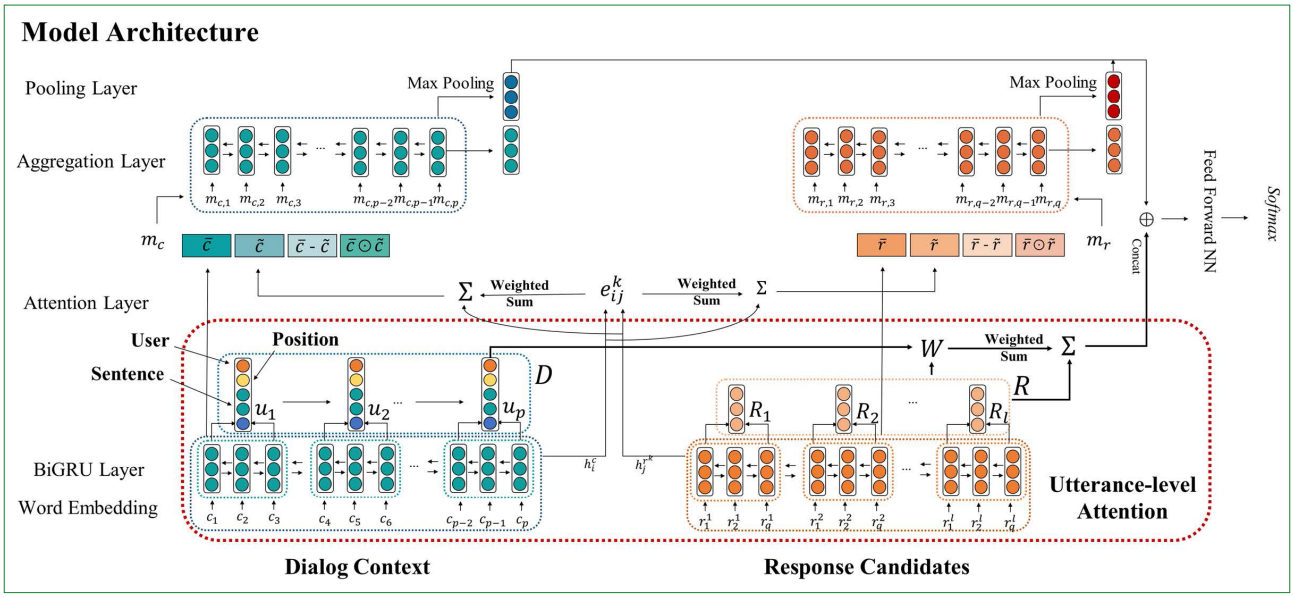
4.1 실험 개요

- DSTC7에서 제공한 Ubuntu Dialog Corpus와 Adviisng Dataset을 사용하여 response selection task에 대해 실험을 진행한 결과는 아래와 같음
- 본 기술은 DSTC7에서 제공한 Ubuntu와 Advising 데이터 셋에 대해서 실험을 진행하였으며, ESIM+SE+PE+UE(ELMO) 모델이 기존 Baseline 모델들의 성능보다 좋은 성능을 보여주었음

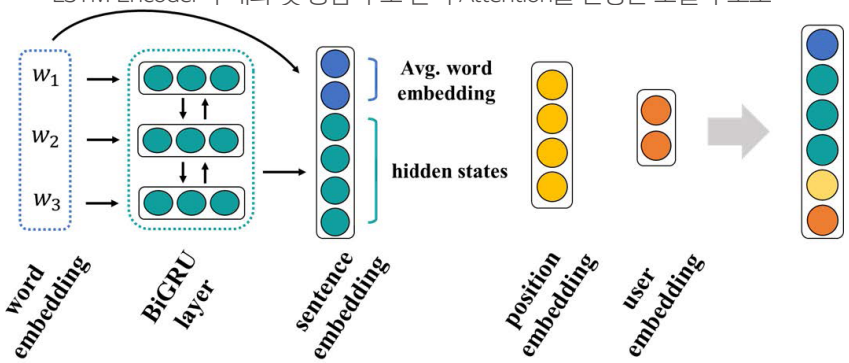
4.2 실험 결과

Task 1	Ubtmtu						Advising					
	R@1	R@2	R@5	R@10	R@50	MRR	R@1	R@2	R@5	R@10	R@50	MRR
(Lowe et al. 2015)	0.211	0.307	0.446	0.569	0.921	0320	0.074	0.108	0.210	0.342	0.802	0.162
(Dong and Huang 2018)	0.367	0.452	0.558	0.651	0.917	0.465	0.086	0.156	0.256	0.376	0.834	0.187
ESIM + SE (GloVe)	0.377	0.460	0.568	0.657	0.929	0.473	0.098	0.160	0294	0.430	0.834	0204
ESIM + SE + PE + UE (GloVe)	0.384	0.464	0.575	0.662	**0.921**	0.480	**0.112**	**0.166**	0.298	0.438	**0.859**	**0.214**
ESIM + SE + PE + UE (ELMo)	**0.406**	**0.493**	**0.606**	**0.691**	0.928	**0505**	0.106	0.160	**0.306**	**0.460**	0.858	0213

• 아래의 그림은 본 기술의 전체 모델 구조도 및 발화 임베딩의 구성을 도식화한 것임



LSTM Encoder와 대화 및 응답 후보 간의 Attention을 반영한 모델 구조도



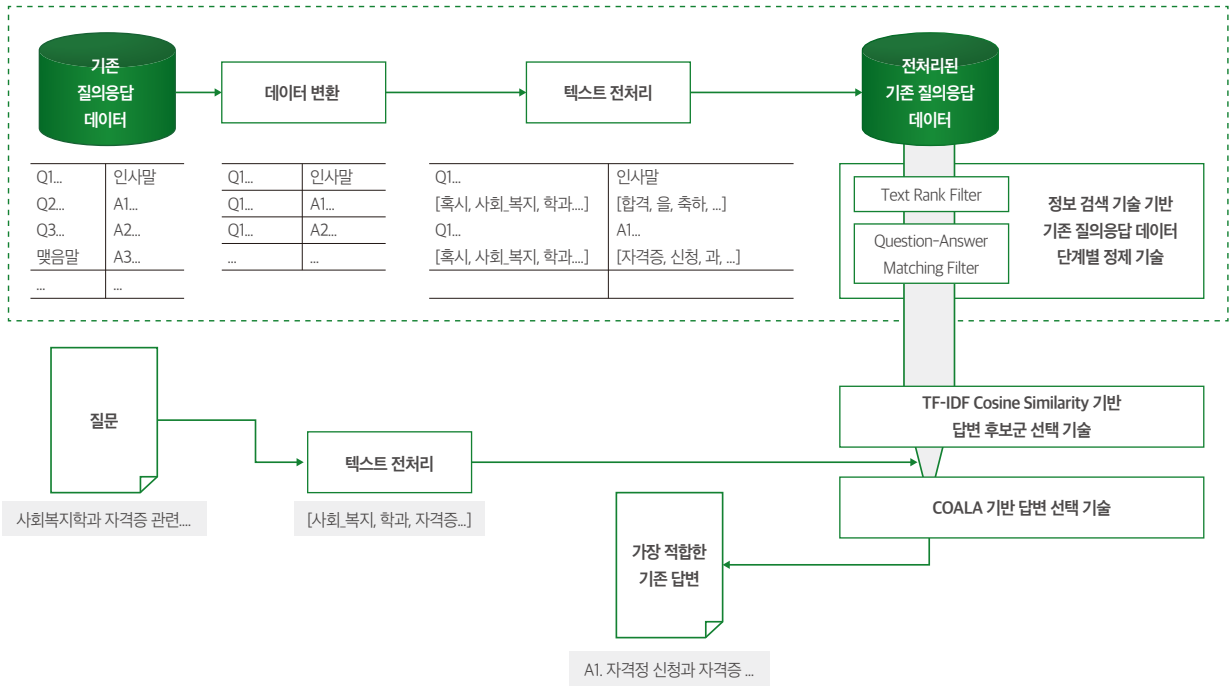
대화의 발화 정보들을 이용한 임베딩 구성 방법

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/

4. 실험 (Only PDF)

4.1 실험 개요

• 본 기술을 적용하여 서울사이버대학교 입학상담 게시판 질의응답 데이터 1248건으로부터 자동 질의응답 시스템 구축

4.2 실험 결과

• 본 기술을 적용하여 구축한 자동 질의응답 시스템의 결과 예시

학사편입 전형 기준이 궁금합니다.

학사편입 전형은 국내외 4년제 대학을 졸업(예정)하고 학사학의를 취득하거나 또는 법령에 의하여 이와 동등이상의 학력이 있다고 인정되면 지원자격에 학사편입전형으로 지원하실 수 있습니다.(2016.12.06) 더보기>>

심리상담학부 수업 온라인인가요

2. 서울사이버대학교는 시험과 수업이 모두 온라인으로 진행됩니다. 다만, 학과별 특성에 따라 대면수업, 실습수업, 오프라인 특강 및 학과활동, 동아리 활동, 스터디모임 등 다양한 오프라인 콘텐츠가 진행되고 있습니다. 특히 상담심리학과의 경우 대면수업 및 실습 수업이 오프라인으로 진행되며, 본교 서울사이버대학교는 전국에 심리상담센터를 두고 있어, 심리상담학부 재학생과 졸업생의 실무능력 강화를 위한 실습 및 수련기관 역할을 하고 있습니다. 본교 상담심리학과 홈페이지에 상담심리학과의 수업 내용 및 프로그램 확인해보시길 바랍니다.
http://counsel.iscu.ac.kr/(2017.02.10) 더보기>>

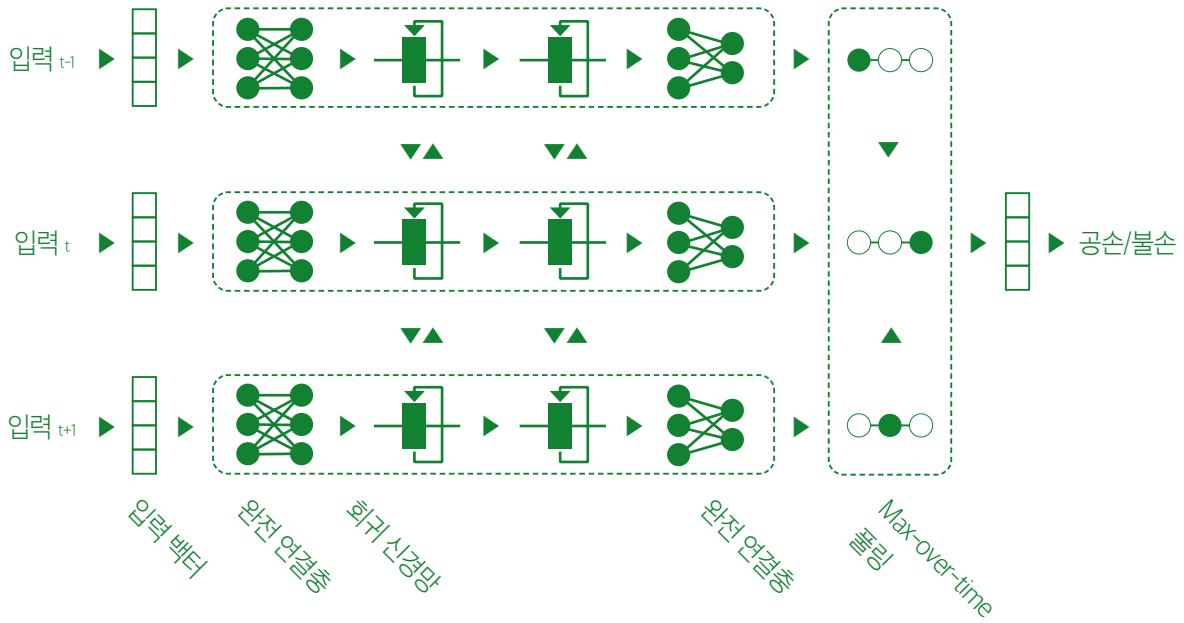
1. 기술 설명

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.

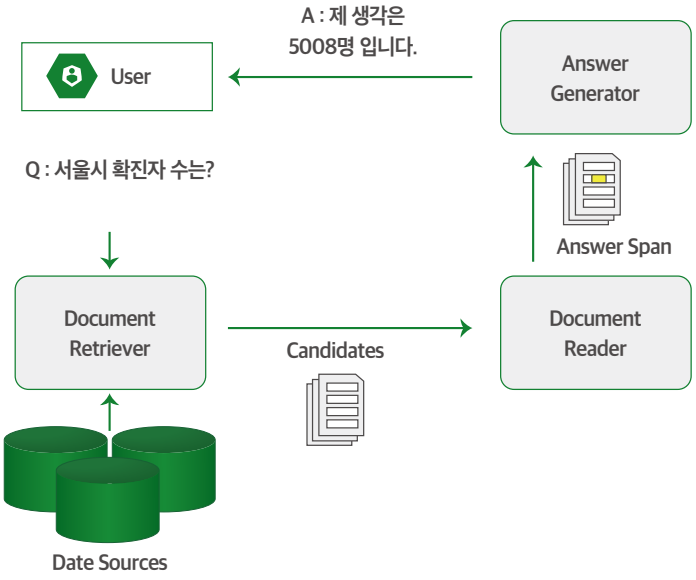
• 양방향 RNN을 이용한 문장분류
양방향 회귀 신경망(Recurrent Neural Network, RNN)은 단어를 순차적으로 입력받아 내부의 기억 구조를 활용하여 문맥 정보가 반영된 단어 표상을 생성한다. 본 연구에서는 RNN의 기억 구조를 보강하여 장거리 의존성 문제를 해소한 LSTM을 기반으로 모델을 구성하였다.



• 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.

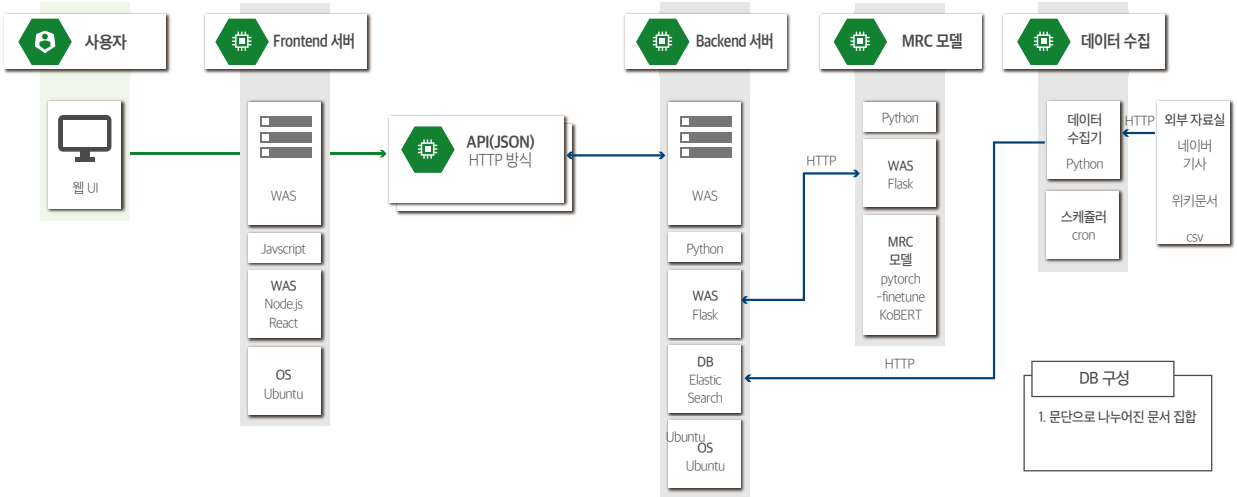
1. 기술 설명

- 기계독해(Machine Reading Comprehension; MRC)는 문단과 질의문이 주어졌을 때 정답에 해당하는 부분을 찾는 기술임
- 정답의 위치정보를 알기 위해 토큰화된 문장을 인코딩하고 이를 이용해 근거 문맥 내 토큰의 정답 확률을 구함. Transformer 아키텍처는 단어와 문장에 대한 문맥이 반영된 인코딩을 가능하게 하였다. Transformer 계열의 PLM인 BERT를 이용하여 MRC 연구가 수행되었으며, F1 점수 기준으로 사람 수준의 MRC 수준을 보여주고 있음
- 근거 문맥과 질문의 쌍이 사전에 주어지지 않는 경우, 질문에 알맞은 근거 문맥을 찾고 그 안의 정답 위치를 찾는 과정을 거치게 됨



2. 기술 방법

- 챗봇 구축 시 데이터 수집 단계에서 챗봇 기술에 적합하도록 전처리하여, 정보검색(Elastic Search) 기술을 이용해 무의미한 질의응답 데이터 제거
- 답변 후보군 선택 기술에서 BM25의 코사인 유사도를 기준으로 가능한 답변 후보군을 선택하고, 딥러닝 기반 최신 답변 선택 모델(BERT-based MRC)를 적용하여 답안 추출
- 정제한 답안을 JSON 형식으로 가공하여 사용자에게 제공



3. 기술 활용 및 응용 분야

- 본 기술은 도메인 특화 챗봇에 활용될 수 있으며, MRC 기반 정보검색 모델에서 활용될 수 있다.
- 본 기술은 신문기사, 게시판 글 등 정제되지 않은 데이터를 딥러닝 PLM BERT 기반 MRC 기술을 통합하여 자동으로 질의응답하는 시스템을 구축할 수 있다.
- 데모 (web) http://nlplab.iptime.org:36200/mrcv2
- 데모 (카카오톡) https://pf.kakao.com/_xoKUCK

4. 실험 (Only PDF)

4.1 실험 개요

- 문서 검색기의 경우 전처리(analysis) 과정에 따라 검색 결과가 달라지게되므로, 전처리 방법에 따른 top-k 정확도를 측정하였음

4.2 실험 결과

- KorQuAD 1.0 데이터 집합의 근거 문장(context)을 인덱싱하고 질문(question)을 질의문으로 검색하여 정답(answer)의 포함 여부를 기준으로 top-k 정확도를 측정

	top-1 정확도	top-5 정확도
공백 분절	71.68%	88.92%
형태소 분석 + 명사 추출	88.92%	97.26%

- 형태소 분석을 하여 명사를 추출하고 이를 이용하여 인덱싱할 경우 유의미한 성능 향상을 보임

1. 기술 설명

- 일상대화(Chitchat)는 개방형 질문(open-ended question)을 다루는 대화이며 도메인이 정해지지 않은 일반적인 대화를 다룸
- 일상대화 생성 기술은 일상대화에서 나타나는 단어의 순차 생성 확률을 학습하여, 자연스러운 문장을 생성하도록 함됨

2. 기술 방법

- 일상대화의 경우 사용자 입력 문장에 대응되는 문장을 생성해야 하므로, 대응되는 문장 쌍을 학습 데이터로 사용함
- 본 기술은 auto regressive 언어 모델로 구성되며, multi-layer transformer에 기반한 아키텍처를 가진다. 언어 모형은 사전 학습 모형(PLM)을 이용하며 이를 전이학습(transfer learning) 하여 일상대화를 생성함

3. 기술 활용 및 응용 분야

- 본 기술은 챗봇에서 대화 상황의 응답 생성에 사용될 수 있다.
- 사용자 친화적 UX 개발에 응용될 수 있다.
- 데모 http://nlplab.iptime.org:36200/dialo

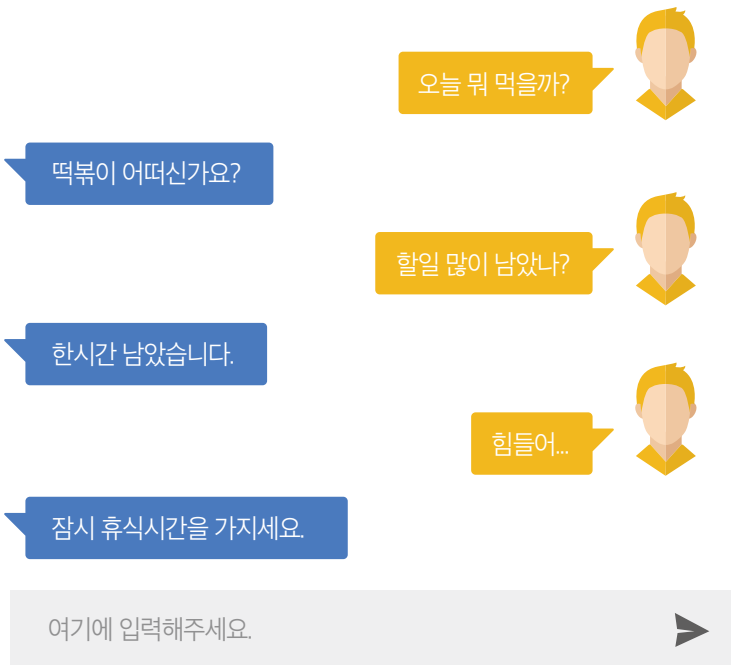
4. 실험 (Only PDF)

4.1 실험 개요

- ai-hub의 오피스 일상대화 데이터 집합을 이용하여 일상대화 생성 모델을 학습하고 이를 이용하여 문장을 생성하도록 하였음. 데이터 집합은 1,325개의 single-turn 대화로 구성되어 있음
- ko-gpt2 PLM을 전이 학습하여 일상대화 생성 모델을 학습하며, top-p 샘플링하여 디코딩을 하도록 하였음

4.2 실험 결과

- 학습은 1 gpu 환경에서 수렴까지 4시간 가량 소요되며, 수렴되었을 때의 loss와 ppl은 다음과 같음
- mean_loss : 0.24970679059624673, mean_ppl : 1.2984058797359466
- 학습 결과 학습 데이터의 single-turn 대화에 대한 복원이 되는 것을 볼 수 있으며, 학습 데이터에 없더라도 ko-gpt2의 PLM이 학습한 확률 분포로 표현 가능한 입력 문장에 대해 sensible한 답변 문장을 생성할 수 있음을 확인하였음



1. 기술 설명

- 주어진 이미지에 근거한 질의에 대해 알맞은 대답을 하는 기술
- VisDial v1.0 데이터를 활용함

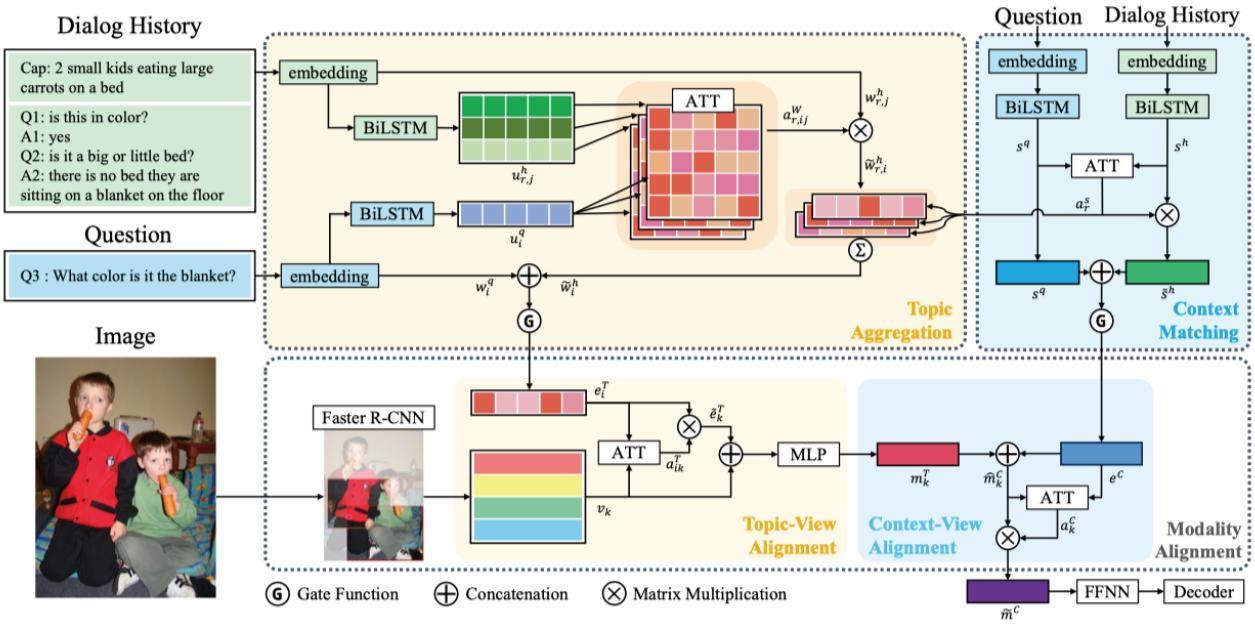
Dialog Topics

People Food household goods

Cap: 2 small kids eating large carrots on a bed
Q1: is this in color?
A1: yes
Q2: is it a big or little bed?
A2: there is no bed they are sitting on a blanket on the floor
Q3: what color is it the blanket?
A3: multicolored blues
Q4: are the kids boys or girls?
A4: boys
Q5: how old do they look?
A5: 7-9
Q6: do they have any other snacks?
A6: no

2. 기술 방법

- Faster-RCNN을 통해 이미지 안의 객체(object)들을 추출(extract)하고, 질의-응답 텍스트는 Bi-LSTM으로 임베딩(embedding)함.
- 서로 이질적인 모달리티의 입력값들을 융합하기 위해 단어-단위(word-level), 문장-단위(sentence-level)를 고려하여, 어텐션(attention)을 기반으로 연속적인 정렬(alignment)를 진행함.
- 질문의 의미적 의도를 파악하기 위한 문맥-객체 간의 연결과 단어-객체 간의 연결을 모두 고려함.



3. 기술 활용 및 응용 분야

• 본 기술은 시각장애인을 보조하는 수단으로 활용할 수 있으며, 텍스트로 이루어진 챗봇이 아니라, 이미지까지 이해하는 AI 챗봇으로 활용 가능함.

4. 실험 (Only PDF)

4.1 실험 개요

• 시각적 질의응답 데이터 셋인 VisDial v1.0을 이용함

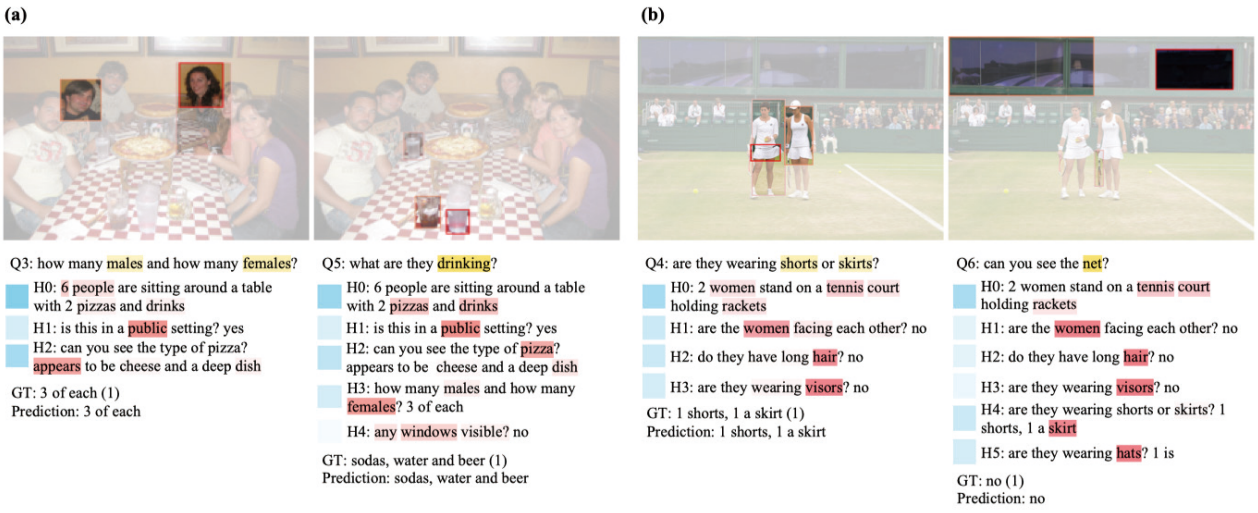
• MVAN(Multi-View Attention Network)를 제안함, MVAN는 Topic-Aggregation 모듈, Context-Matching 모듈, Modality-Alignment 모듈로 구성됨

4.2 실험 결과

• 기존 모델들 보다 우수한 성능을 보였음.

Model	AVG	NDCG	MRR	R@1	R@5	R@10	Mean
LF[5]	12	45.31(13)	55.42(12)	40.95	72.45	82.83	5.95
HRE[5]	12	45.46(12)	54.16(13)	39.93	70.45	81.50	6.41
MN[5]	11	47.50(11)	55.49(11)	40.98	72.30	83.30	5.92
GNN[34]	10	52.82(10)	61.37(10)	47.33	77.98	87.83	4.57
CorefNMN[15]	9	54.70(9)	61.50(9)	47.55	78.10	88.80	4.40
RVA[21]	8	55.59(8)	63.03(7)	49.03	80.40	89.83	4.18
DualVD[11]	7	56.32(7)	63.23(5)	49.25	80.23	89.70	4.11
Synergistic[8]	6	57.32(3)	62.20(8)	47.90	80.43	89.95	4.17
CAG[9]	5	56.64(6)	63.49(4)	49.85	80.63	90.15	4.11
DAN[12]	4	57.59(2)	63.20(6)	49.63	79.75	89.35	4.30
HACAN[32]	3	57.17(4)	64.22(3)	50.88	80.63	89.45	4.20
FGA[26]	2	56.90(5)	**66.20**(1)	**52.75**	**82.92**	**91.07**	**3.80**
MVAN(ours)	1	**59.37**(1)	64.84(2)	51.45	81.12	90.65	3.97
Synergistic† [8]	5	57.88(4)	63.42(5)	49.30	80.77	90.68	3.97
CDF† [13]	2	59.49(2)	64.4(4)	50.90	81.18	90.40	3.99
DAN† [12]	2	59.36(3)	64.92(3)	51.28	81.60	90.88	3.92
FGA† [26]	2	57.20(5)	**69.30**(1)	**55.65**	**86.73**	**94.05**	**3.14**
MVAN† (ours)	1	**60.92**(1)	66.38(2)	53.20	82.45	91.85	3.68

• 시각화를 통해 정답을 추론할 때, 주어진 입력값의 어느 부분에 집중하는지 나타냄.



5. 참고

• 논문: https://arxiv.org/abs/2004.14025

• 코드: https://github.com/taesunwhang/MVAN-VisDial

• 데모: http://nlplab.iptime.org:34242/

1. 기술 설명

- 기존 칫챗(chit-chat) 대화 시스템에서 모델이 일관성 없는 답변을 하거나, 재미가 없는 답변을 만드는 등의 문제점을 해결하기 위하여 페르소나 대화 데이터(PERSONA-CHAT)와 이를 활용한 태스크가 만들어졌음
- 페르소나 대화 데이터에서는 페르소나를 프로필 정보로 지니고 있는 두 명의 화자가 서로의 페르소나를 기반으로 대화를 주고받음
- 기계가 이와 같이 페르소나 정보를 가지게 되면 조금 더 일관성 있고 사람과 같이 재치 있는 답변을 할 수 있음
- 페르소나 대화 데이터를 사전학습된 언어모델에 미세조정하여 답변 선택을 잘할 수 있는 모델임

Persona of [PERSON1]

My mom is my best friend
I have four sisters
I believe that mermaids are real
I love iced tea

[PERSON2]: Hi, how are you doing today?
[PERSON1]: I am spending time with my **4 sisters**, what are you up to?
[PERSON2]: Wow, four sisters. Just watching Gane of Thrones.
[PERSON1]: That is a good show. I watch that **while drinking iced tea.**
[PERSON2]: I agree. What do you do for a living?
[PERSON1]: I'm a research. I'm researching the fact that mermaids are real.
[PERSON2]: Interesting. **I'm a website designer. Preety much spend all my time on the computer.**
[PERSON1]: That's cool. My mom does the same thing.
[PERSON2]: That's awesome. I have always had a love for technology.
[PERSON1]: Tell me more about yourself.
[PERSON2]: I really enjoy free diving, how about you, have any hobbies?
[PERSON1]: **I enjoy hanging with my mother. She's my best friend.**

페르소나 대화 데이터의 예시

2. 기술 방법

- 본 기술은 사전학습된 언어모델 BERT의 미세조정을 이용하여 페르소나 태스크에 맞게 학습함
- 학습 시 아래 그림과 같은 추가적인 보조 태스크를 설정하여 multi-task learning으로 학습, 페르소나 기반 답변 선택 학습에 도움이 될 수 있도록 함
- 보조 태스크 1은 대화가 페르소나에 기반하여 이루어지는 것에서 착안하여, Sentence Transformer를 이용하여 페르소나-발화 쌍을 찾고, 발화에 대한 distractor 두 개를 추가한 후 그 중에서 올바른 답을 찾을 수 있도록 학습
- 보조 태스크 2는 보조 태스크 1과 같은 방법으로 페르소나-발화 쌍을 찾고, 발화와 페르소나 문장에 대한 distractor를 각각 두 개씩 추가한 후 올바를 쌍에 1을 라벨, 아닌 후보들에 0을 라벨하여 학습

1 persona sentence + consistent question + consistent answer

1 persona sentence + consistent question + distracting answer1

1 persona sentence + consistent question + distracting answer2

보조 태스크 1 (Multiple choice)

0 0 1 1 0 0

<1 persona snt> + <1 persona snt> + <1 persona snt> + <turn> + <turn> + <turn>

보조 태스크 2 (Sentence labeling)

3. 기술 활용 및 응용 분야

• 본 기술은 페르소나를 반영하여 개인 맞춤 대화 시스템에 활용될 수 있으며, 다중 언어에 대한 번역기에 활용될 수 있으며, 다중 언어 문서에서 정보검색 모델에서도 활용될 수 있다.

4. 실험 (Only PDF)

4.1 실험 개요

• 보조 태스크의 효과를 검증하기 위하여, 미세조정만 진행한 BERT와 보조 태스크를 추가한 경우를 비교 실험하였음

4.2 실험 결과

Methods	Accuracy(Hits@1)
BERT	82.94
BERT with post-training 1	84.16
BERT with post-training 2	84.42

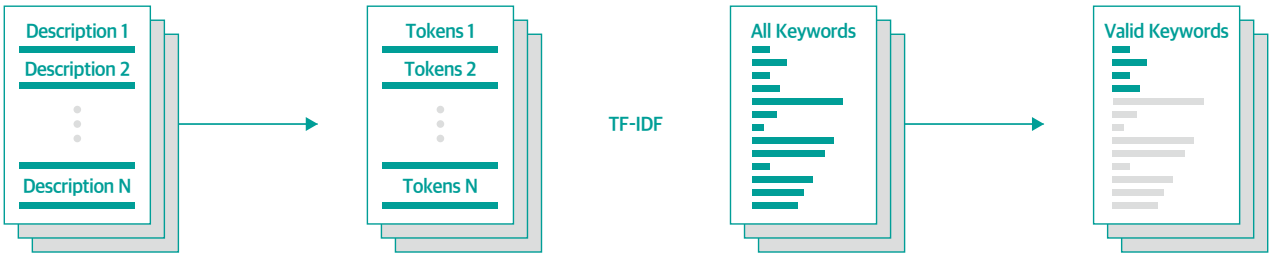
• 실험 결과, 미세조정만 진행한 BERT와 비교했을 때, 보조 태스크와 함께 multi-task learning을 한 경우 약 1-2%씩 성능이 오른 것을 확인할 수 있었음. 이는 효과적인 보조 태스크를 선정하여 multi-task learning으로 학습시 주 태스크에도 좋은 영향을 줄 수 있는 것으로 해석할 수 있음

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
Narrative기반 자동 비디오 분할
비지도 학습 알고리즘을 이용한 보고서 자동 분석 및 토픽 자동 추출 기술
순차 정보를 이용한 콘텐츠 추천 시스템 개발
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
아이들 교육을 위한 나오 로봇
GPT2를 활용한 유사 뉴스 기사 추천 시스템
나오 로봇을 활용한 이중 언어 교육
나오 로봇을 활용한 동화 추천 및 읽기
Virtual-Try-On Model for Fashion AI

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/

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

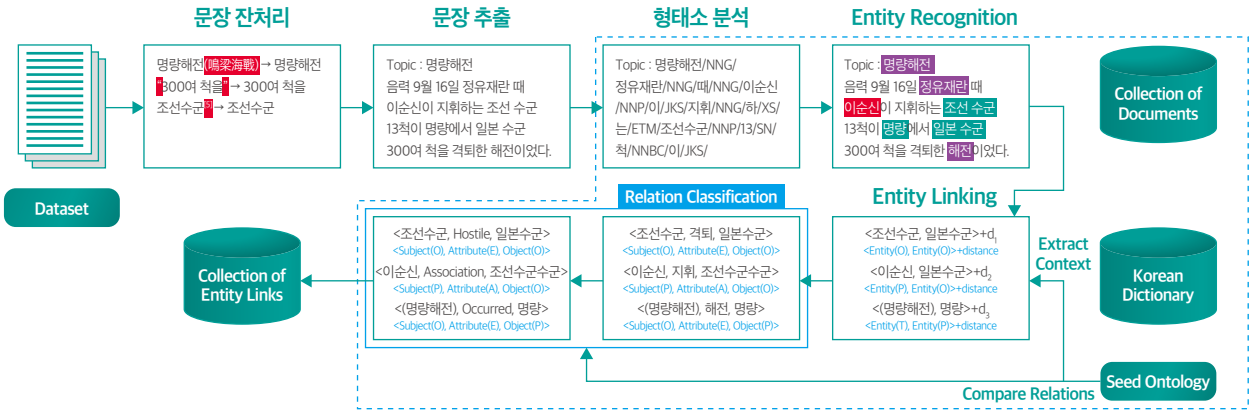
Lucy is going to the park and she is taking the dog for a walk

- 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 http://nlplab.iptime.org:32277/

4. 실험 (Only PDF)

4.1 실험 개요

- 단편적 및 전역적 관계 추출의 정확도를 평가하기 위하여 타 모델들과 함께 문서 단위의 평가 데이터로부터 관계 추출을 실행하여 Precision, Recall, F1 Score를 측정함

4.2 실험 결과

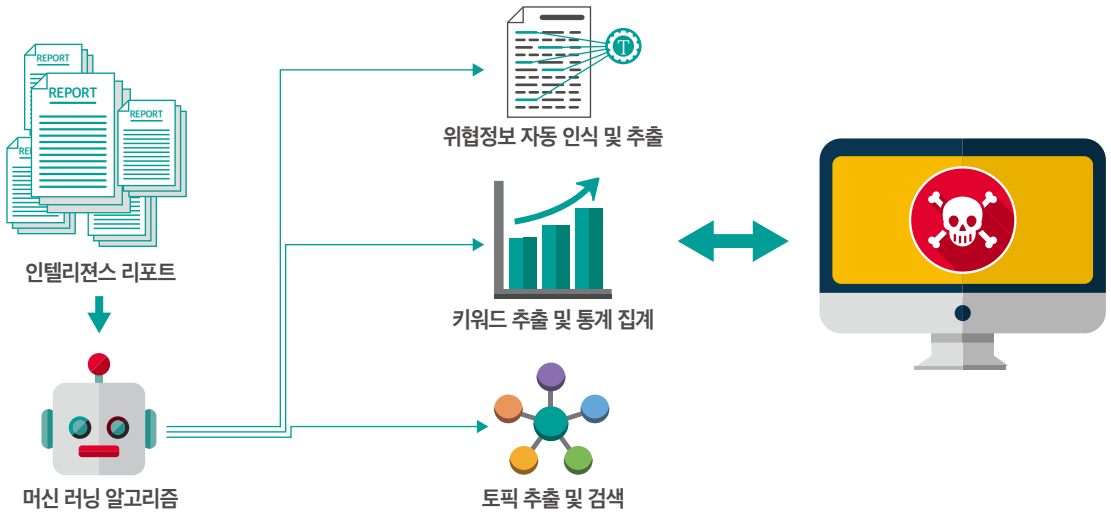
- 제안한 모델인 Augmented External Memory Model(AEMM)은 전체적으로 다른 모델들에 비하여 단편적 관계 분류 성능을 비교하면 더 낮은 성능을 보이는데, 이는 외부 메모리 신경망의 전역 관계 분류의 결과에 따라 영향을 받는 것이 오히려 단편적 관계 분류에 악영향을 끼치는 것으로 보임
- 전역적 관계 추출의 비교에서는 AEMM은 타 모델에 비하여 확연히 높은 Global Precision을 보여준다. 이 때문에 비록 Global Recall에서는 타 모델들과 비슷한 성능을 보임에도 F1 score에서 더 높은 성능을 보이는 것을 관측할 수 있음

	CNN	LSTM	한글모델	AEMM
Local Precision	0.327	0.341	0.390	0.269
Local Recall	0.315	0.347	0.259	0.307
Local F1 Score	0.321	0.344	0.311	0.287
Global Precision	0.194	0.183	0.198	0.383
Global Recall	0.313	0.332	0.262	0.287
Global F1 Score	0.240	0.236	0.226	0.328

[표] 단편/전역 관계 추출 모델간의 성능비교

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.

- **PDF를 HTML로 변환**
  - 이로부터 글자 크기 등 부가적인 정보를 얻고 이를 추후 프로세스에 활용
- **불필요한 메타텍스트 제거**
  - 주기적으로 반복되는 문자열 정보를 이용하여 제거
- **특수문자 정규화**
  - 동일한 기능을 하는 다양한 특수문자를 하나로 통일합
- **연속된 줄 파악**
  - 타 말뭉치로부터 수집한 다양한 통계를 바탕으로 연속된 단어 파악
- **문강 경계 구분**
  - 타 말뭉치를 이용하여 비지도학습 방법으로 훈련시킨 문잠경계 인식 기계학습 모델 사용
- **단어 토큰화**
  - 규칙 기반 토큰화 모델을 이용하여 각 단어를 토큰화합

[그림] PDF2TXT 과정.

3. 기술 활용 및 응용 분야

리포트 자동 분석 (타 분야 문서로 적용 가능)

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

4. 실험 (Only PDF)

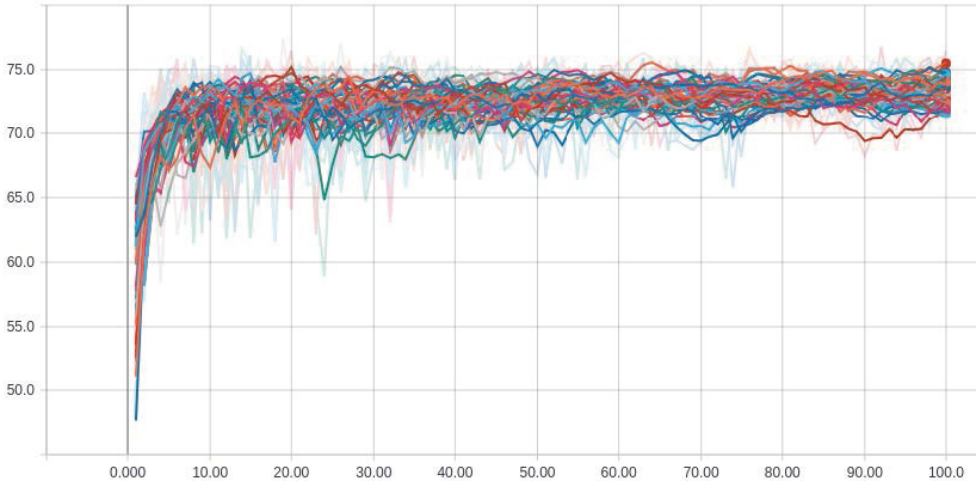
위협정보 개체명 인식 말뭉치 구축 과정

- 원시 말뭉치 수집 : 608건의 PDF 형식의 인텔리젼스 리포트 수집. 이 중 가장 라인 빈도수가 높은 파일(1500~5000라인)을 선정하여 배포 (국내외 영문 인텔리전스 리포트 .pdf file 608개, 인텔리전스 리포트를 토대로 전처리 작업한 .text file 608개, 데이터 구축에 관한 가이드라인). 태깅 데이터 구축을 위해 보안학과에 재학 중인 5명의 연구원 참여

비정형 위협 정보 자동 인식 및 추출 기술의 성능 파악에는 개체명 인식 기술에서 가장 널리 사용되는 정량적 평가 방식인 F-score를 이용한다. 이는 여러 단어로 구성될 수 있는 위협 정보의 특성상 accuracy만으로 평가하기 어려운 점을 반영한 지표이다. F-score는 precision과 recall의 조화평균 값으로, 아래 식에 따라 계산한다.

$$F_1 = \left(\frac{\mathrm{recall}^{-1} + \mathrm{precision}^{-1}}{2} \right)^{-1} = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}$$

본 기술의 성능을 명확히 검증하기 위해 동일한 시스템을 50회 반복 학습시키고, 학습된 모델의 최종 성능을 통계적으로 비교하였다. 각 모델은 100 epoch동안 학습시키며, 이는 전체 학습 데이터에 대해 100회 훈련되었음을 의미한다. 시간의 흐름에 따른 성능의 변화는 아래 그래프와 같다.

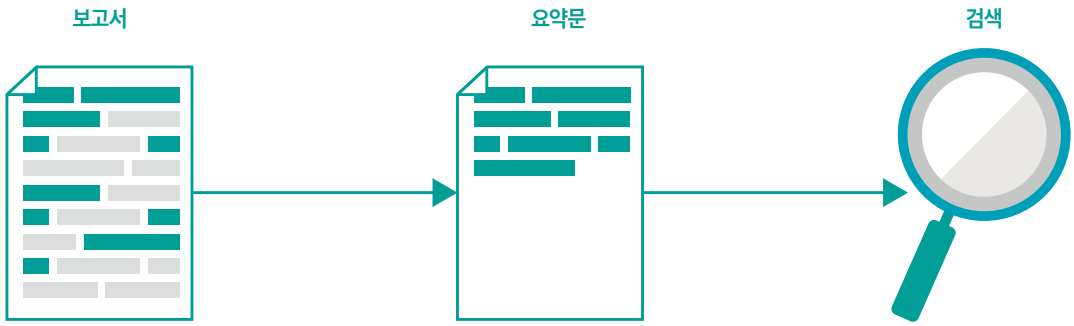


[그림] 전체 50개 모델의 학습 과정을 나타낸 그래프. X축은 epoch, Y축은 F-score를 나타낸다.

이러한 방법으로 총 50개 모델의 성능을 측정한 결과, 평균 F-score는 73.31, 표준 편차는 1.16으로 확인되었다.

1. 기술 설명

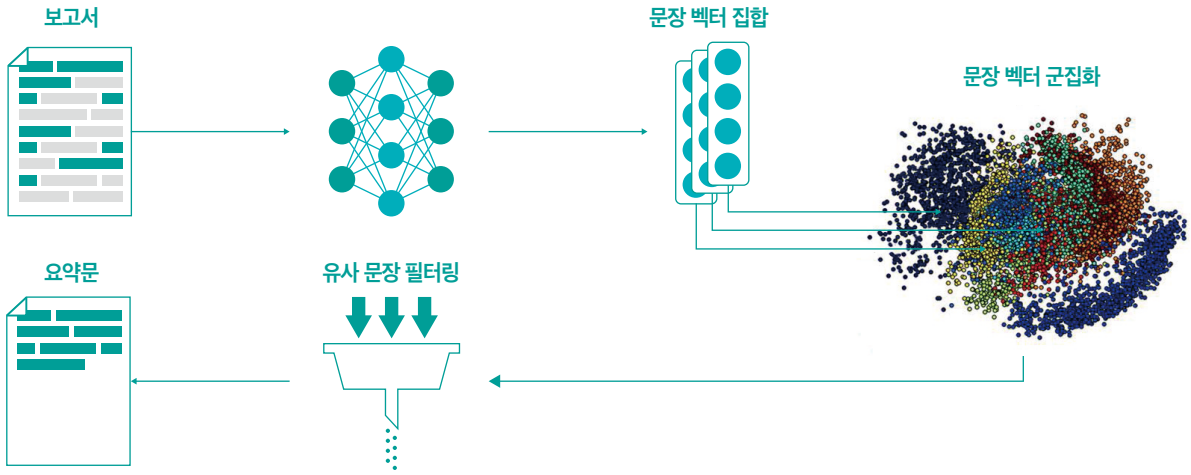
본 기술은 비지도 학습 알고리즘을 바탕으로 문장 추출에 의한 자동 문서 요약 방법이다. 특히, 본 기술은 특정 언어나 문서 특징에 의존하지 않으므로 확장이 용이하다.



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 http://nlplab.iptime.org:32270

Summary

"머신러닝을 이용한 문서 자동 요약 기술"의 데모입니다.

There has been significant open source reporting which has documented the alignment between apparent information collection efforts of China-based threat actors and the strategic emerging industries documented in China 's Five Year Plan (FYP) .

The 13th FYP was released in March 2016 and the sectors and organisations known to be targeted by APT10 are broadly in line with the strategic aims documented in this plan .

We have observed the threat actor copying malware over to systems in a compromised environment , which did not have any outbound internet access .

Systems sharing credentials across the client and the MSP are of particular interest to APT10 , and are commonly used by the threat actor in order to gain access to new areas of the network APT10 simultaneously targets both low profile and high value systems to gain network persistence and a high level of access respectively .

The threat actor 's known working hours align to Chinese Standard Time (CST) and its targeting corresponds to that of other known China-based threat actors , which supports our assessment that these campaigns are conducted by APT10 .

[그림] 본 기술을 이용하여 PDF 형식의 인텔리젼스 리포트를 요약한 결과물.

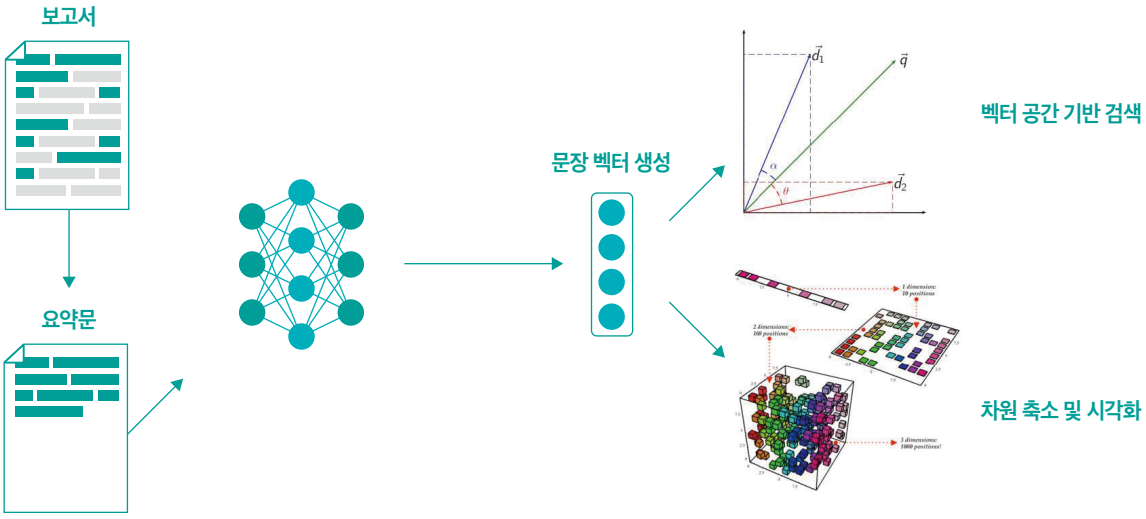
1. 기술 설명

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.

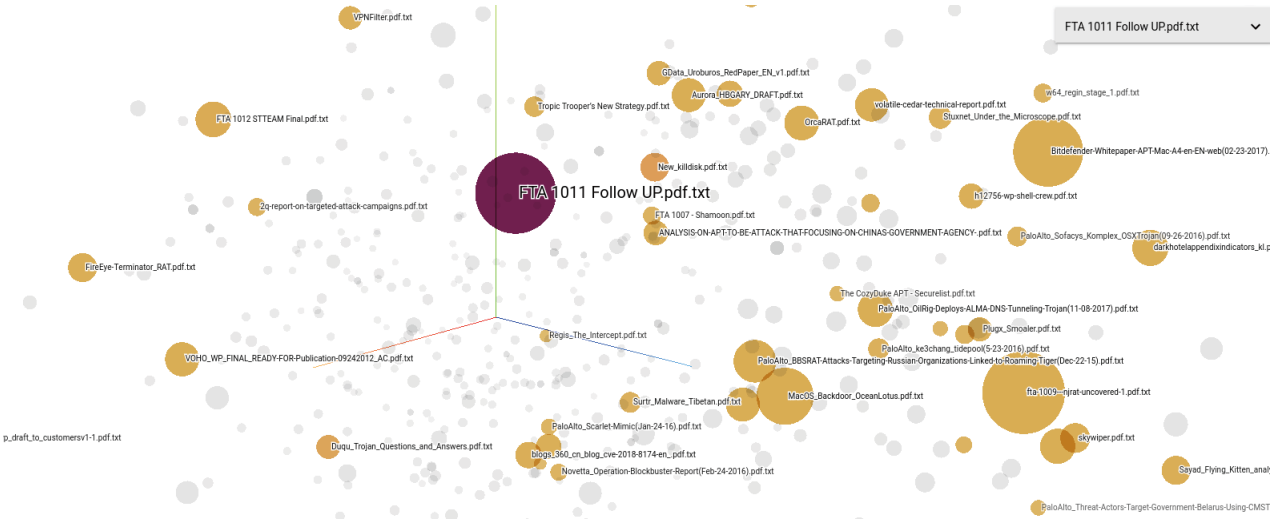
여기서 더 나아가, 문서가 투사된 벡터 공간을 t-distributed Stochastic Neighbor Embedding(t-SNE)와 같은 차원 축소 기법을 이용하면 이를 인간이 시각적으로 인지할 수 있는 공간인 3차원 이하로 변형할 수 있고, 이를 시각화하여 검색 인터페이스로 응용 가능하다. 이를 위해 Tensorboard를 활용하였다.



3. 기술 활용 및 응용 분야

정보 검색, 문서 분류

4. 실험 (Only PDF)



[그림] 문서 임베딩 공간을 시각화한 결과.

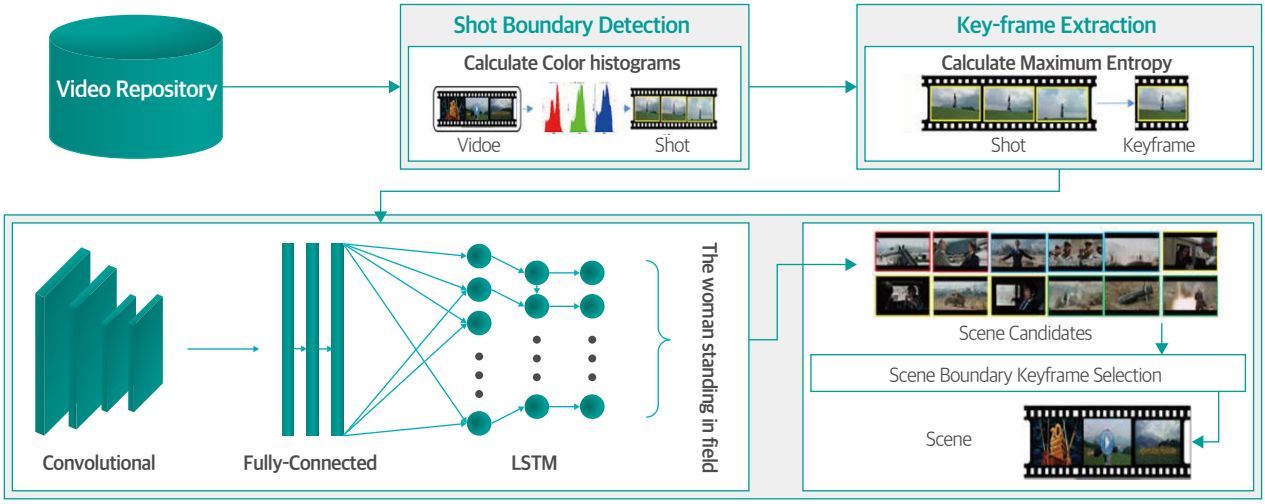
[그림] 선택한 문서와 유사한 문서들의 목록.

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] Video Scene Detection 모델 구조도

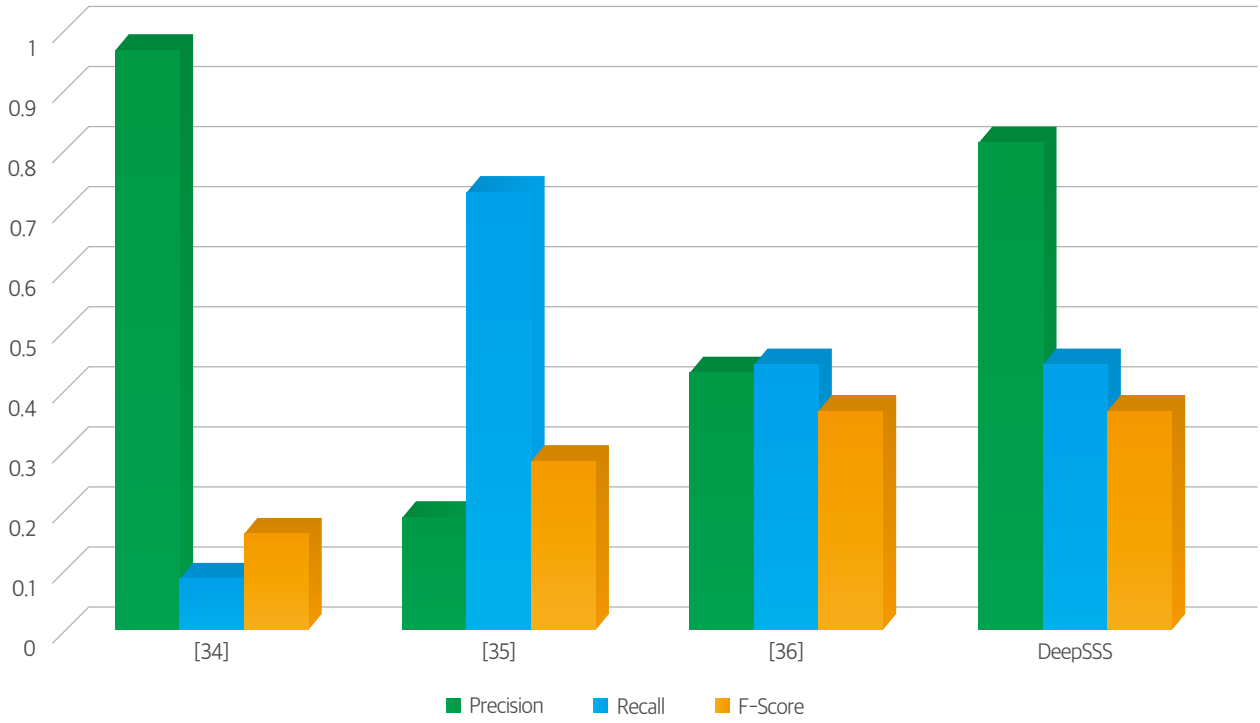
3. 기술 활용 및 응용 분야

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

4. 실험 (Only PDF)

4.1 실험개요

• TRECVid 2016 데이터 세트로부터 무작위로 10개의 영상을 사용하여 수작업으로 영상분할 정답 세트를 제작한 뒤 제안된 모델과 타 영상분할 모델 3개의 Precsion, Recall, F-score 점수를 통하여 성능을 비교함



4.2 실험 결과

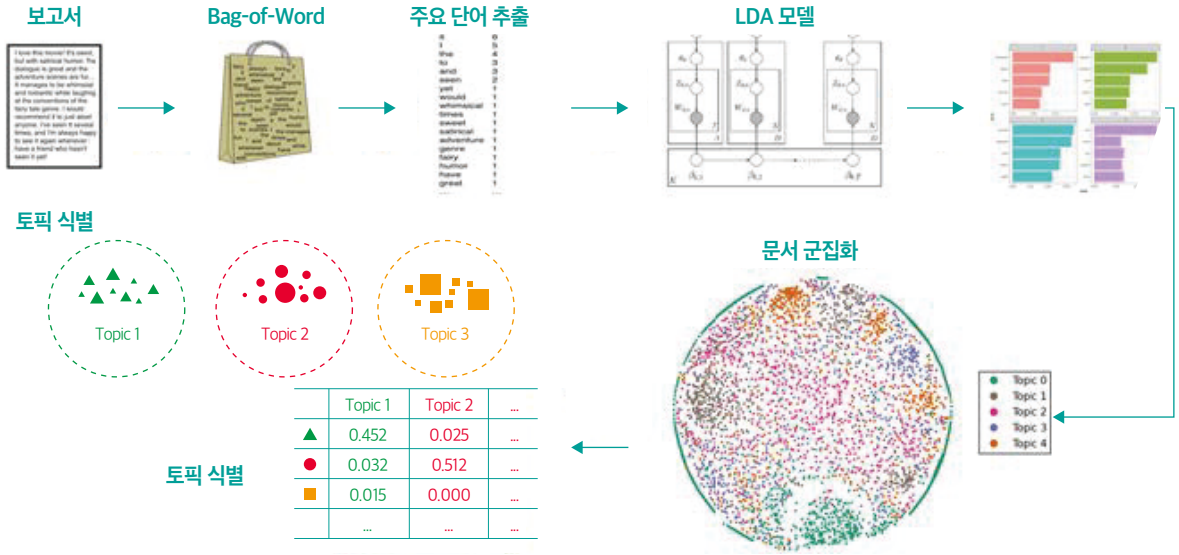
• 해당 모델은 Precision과 Recall에서는 각각 Color Histogram 모델와 Transition Detection 모델에 비하여 성능이 떨어지나 종합적으로 성능간 균형이 제일 균일하며 가장 높은 F-score 점수를 보여줌

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 단어 의분포, 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/

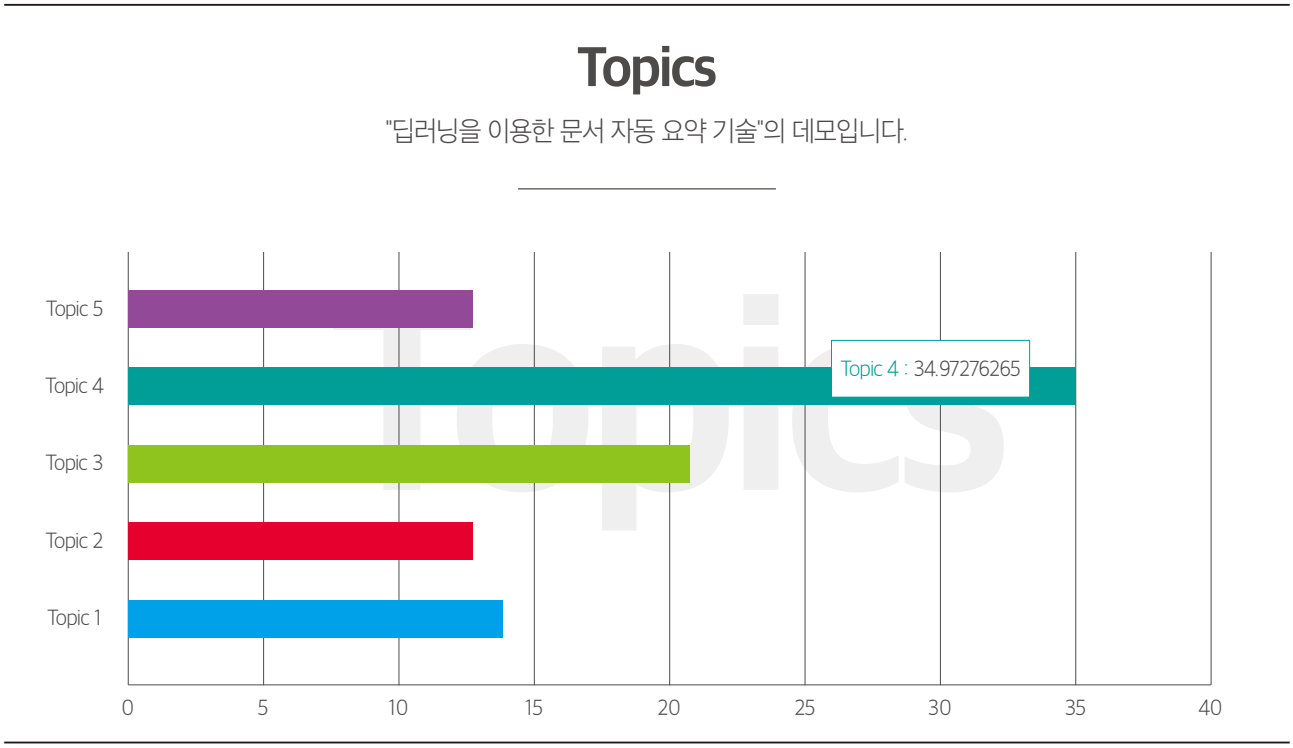
4. 실험

4.1 실험 개요

• 비정형 보고서 문서에서 주제를 찾기 위해 토픽 모델링인 LDA를 활용하여 실험을 진행하였음. 실험을 진행한 결과는 다음과 같음

4.2 실험 결과

• 본 기술의 결과는 보고서를 입력하였을 때 보고서와 관련된 주제가 어디에 들어가며 다른 주제보다 얼마나 가까운지 확률인지 확인할 수 있음

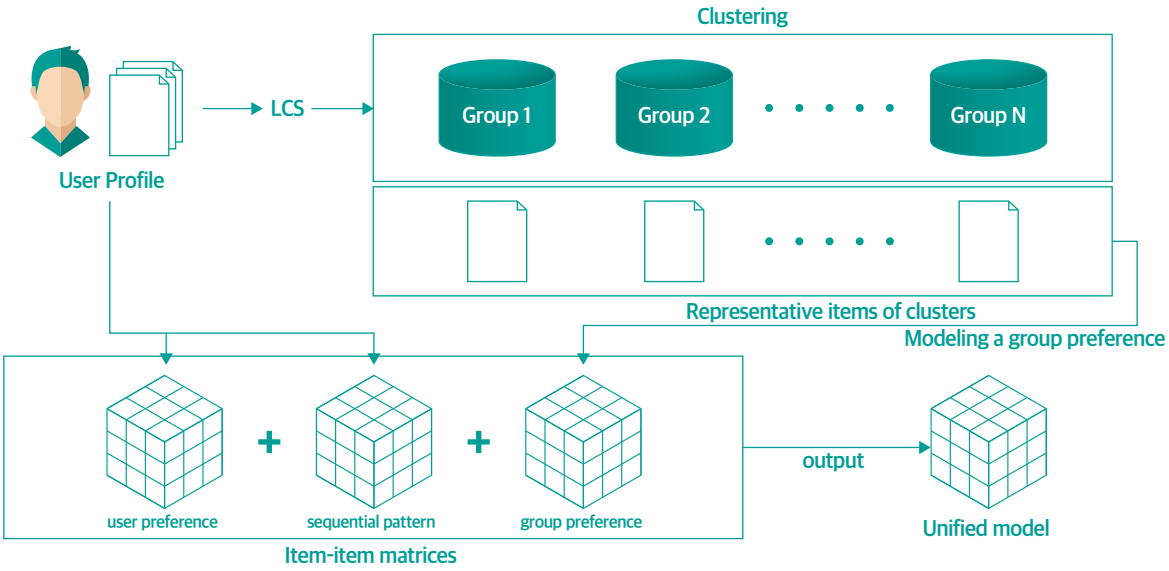


<보고서에 대한 각 토픽별 확률>

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/

4. 실험

4.1 실험 개요

- 아마존 데이터 및 Epinion, Foursqure 데이터를 통하여 기존의 모델들과의 비교실험을 진행한다.
- 3가지 평가 방법을 사용하여 성능을 측정하였으며, 가장 좋은 추천 성능을 보인다.

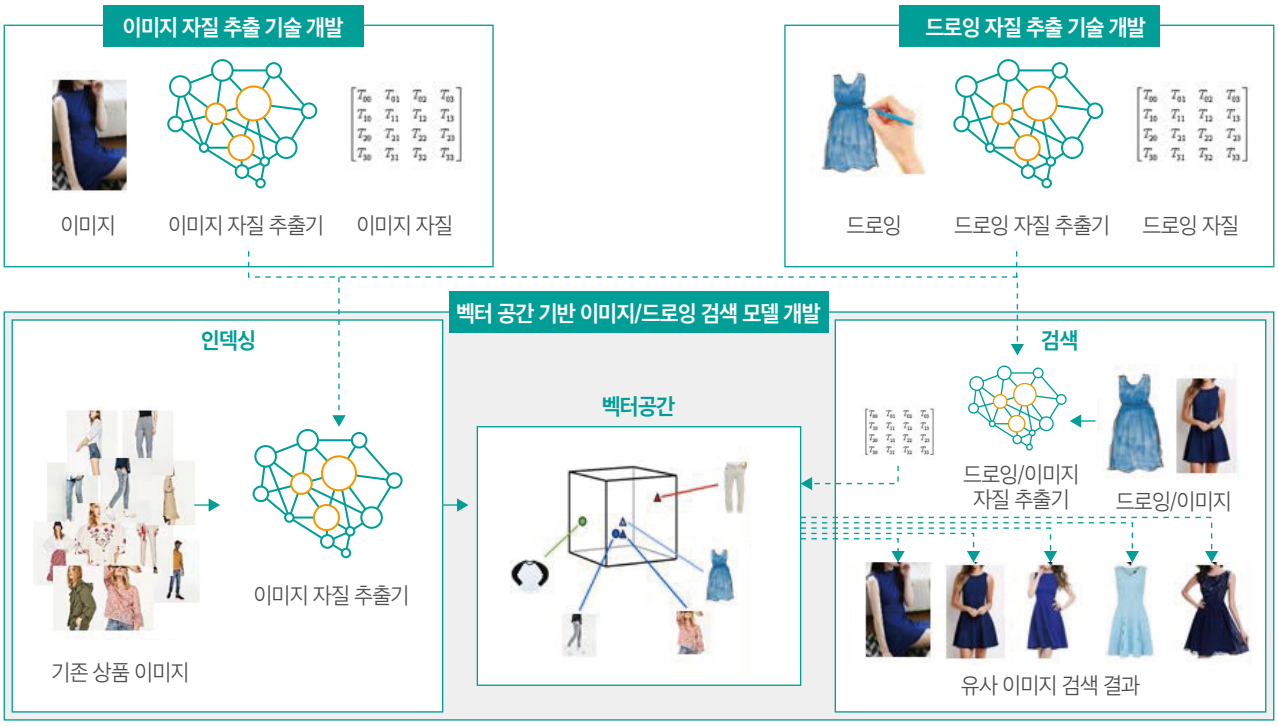
4.2 실험 결과

- 해당 추천 모델인 GPS가 다른 모델에 비해 높은 성능을 보임을 알 수 있다.

Results (SPS).										
Datasets	method	BPR-MF (a)	FISM (b)	FPMC (c)	Fossil (d)	GPS (e)	improvement			
							d vs a	e vs b	e vs d	e vs best
A-Auto	sps@30	0.0384	0.0882	0.0275	0.0863	0.1012	0.048	0.013	0.015	0.013
A-Video	sps@30	0.0327	0.1072	0.0399	0.0875	0.1493	0.055	0.042	0.062	0.042
A-Elec	sps@30	0.0411	0.0421	0.0309	0.0428	0.0511	0.002	0.009	0.008	0.008
A-Office	sps@30	0.0386	0.1003	0.0630	0.1390	0.1461	0.100	0.046	0.007	0.007
Epinions	sps@30	0.1184	0.1147	0.0789	0.1184	0.1974	0.000	0.083	0.079	0.079
Foursquare	sps@30	02555	0.2622	0.2516	0.3162	0.3262	0.061	0.064	0.010	0.010
avg.(k=100)	sps@30	0.0919	0.1185	0.0815	0.1298	0.1669	0.038	0.048	0.037	0.034
Results (Recall)										
Datasets	method	BPR-MF (a)	FISM (b)	FPMC (c)	Fossil (d)	GPS (e)	improvement			
							d vs a	e vs b	e vs d	e vs best
A-Auto	recall@30	0.0386	0.0834	0.0263	0.0821	0.0954	0.044	0.012	0.013	0.012
A-Video	recall@30	0.0334	0.1009	0.0387	0.0831	0.1456	0.050	0.045	0.063	0.045
A-Elec	recall@30	0.0436	0.0437	0.0309	0.0442	0.0509	0.001	0.007	0.007	0.007
A-Office	recall@30	0.0380	0.0756	0.0436	0.0750	0.0830	0.037	0.007	0.008	0.007
Epinions	recall@30	0.0727	0.0902	0.0370	0.0848	0.1390	0.012	0.049	0.054	0.049
Foursquare	recall@30	0.2382	0.221	0.2314	0.2517	0.2634	0.014	0.042	0.012	0.012
avg.(k=100)	recall@30	0.0767	0.1007	0.0636	0.1010	0.1309	0.024	0.030	0.030	0.026
Results (NDCG).										
Datasets	method	BPR-MF (a)	FISM (b)	FPMC (c)	Fossil (d)	GPS (e)	improvement			
							d vs a	e vs b	e vs d	e vs best
A-Auto	ndcg @30	0.0169	0.0479	0.0136	0.0397	0.0504	0.0228	0.0025	0.0107	0.0025
A-Video	ndcg @30	0.0292	0.0830	0.0321	0.0679	0.0888	0.0387	0.0058	0.0209	0.0058
A-Elec	ndcg @30	0.0262	0.0265	0.0189	0.0268	0.0402	0.0006	0.0137	0.0134	0.0134
A-Office	ndcg @30	0.0202	0.0498	0.0237	0.0456	0.0549	0.0254	0.0051	0.0093	0.0051
Epinions	ndcg @30	0.0727	0.0902	0.0370	0.0848	0.1390	0.0121	0.0488	0.0542	0.0488
Foursquare	ndcg @30	0.1367	0.1399	0.1294	0.1589	0.1973	0.0222	0.0574	0.0384	0.0384
avg.(k=100)	ndcg @30	0.0503	0.0729	0.0425	0.0706	0.0951	0.0203	0.0222	0.0245	0.0190

1. 기술 설명

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



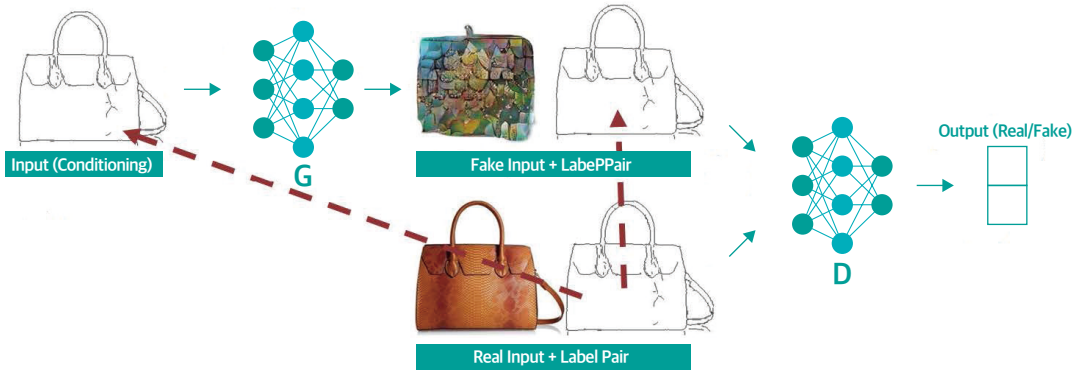
[그림] 벡터 공간 기반 이미지/드로잉 검색 모델의 구조도

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

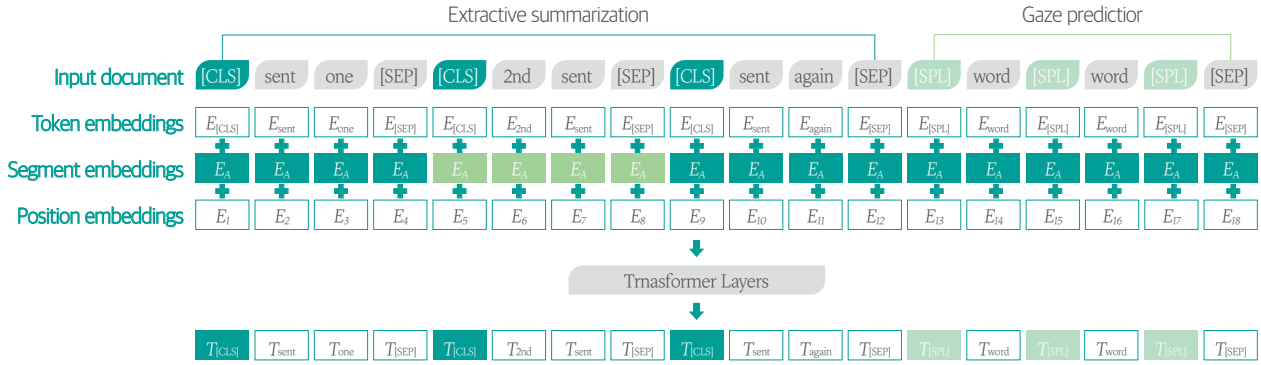
4. 실험 (Only PDF)

[그림] 스케치 업샘플링 모듈을 이용해 업샘플링된 결과물의 예

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.

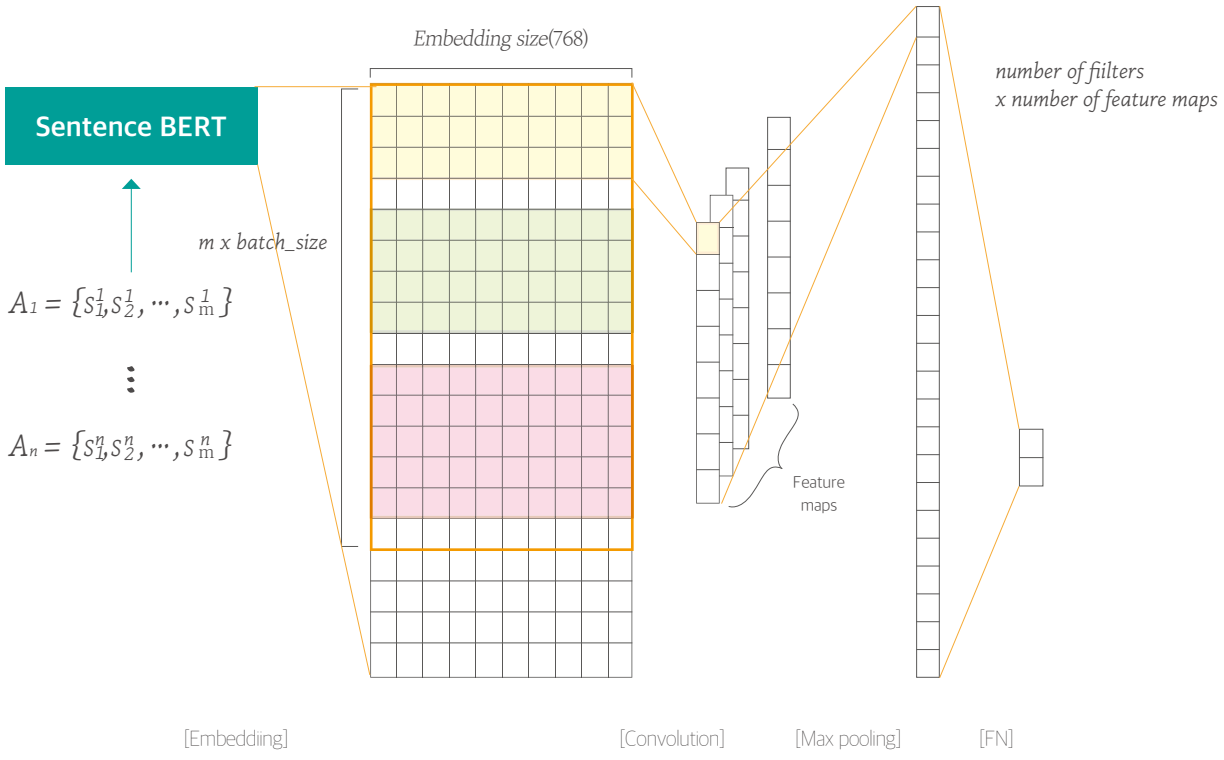
1. 기술 설명

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.



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

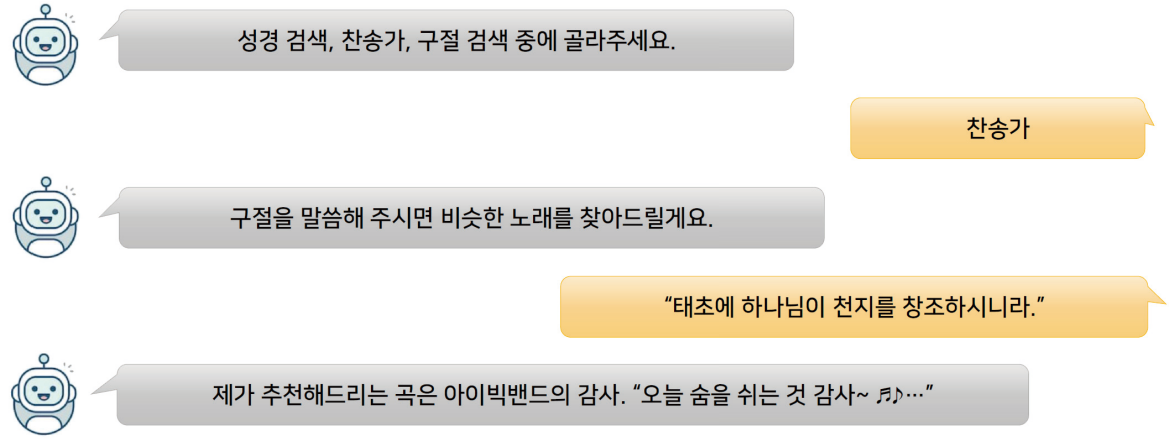
어떤 구절을 듣고 싶으십니까?

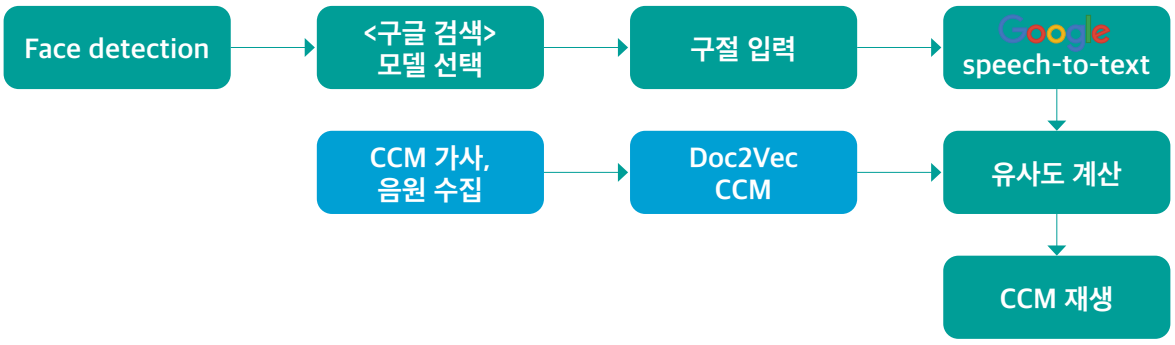
창세기 1장 1절부터 5절까지 들려줘.

"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

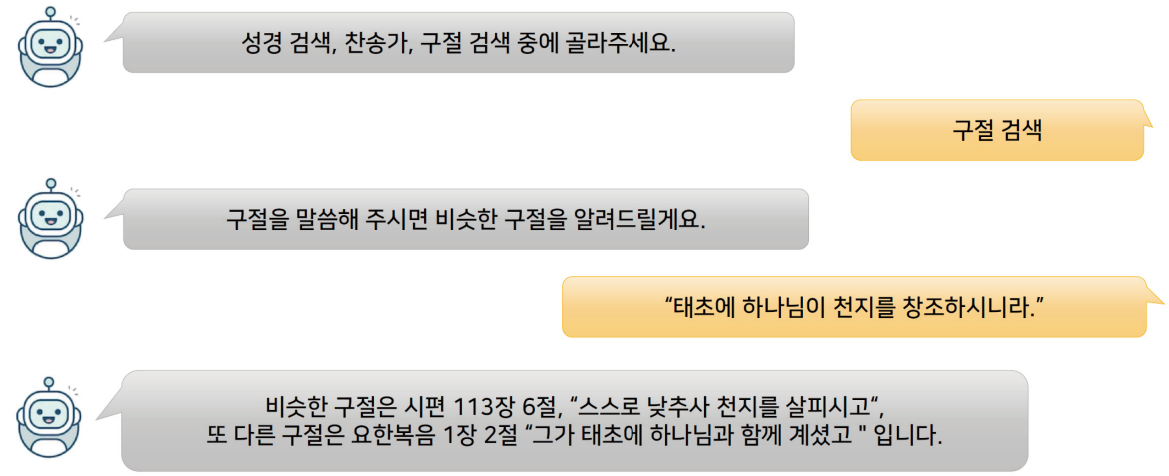


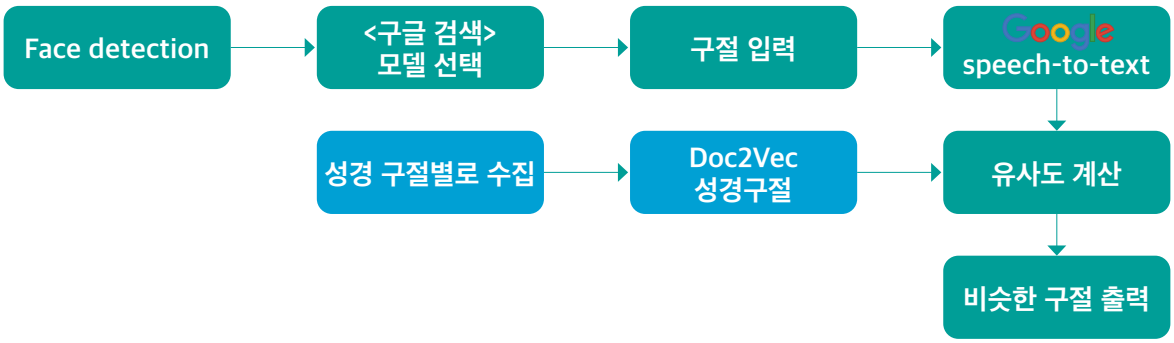


• 비슷한 구절 검색 모델

- 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

다른 종교에도 적용 가능

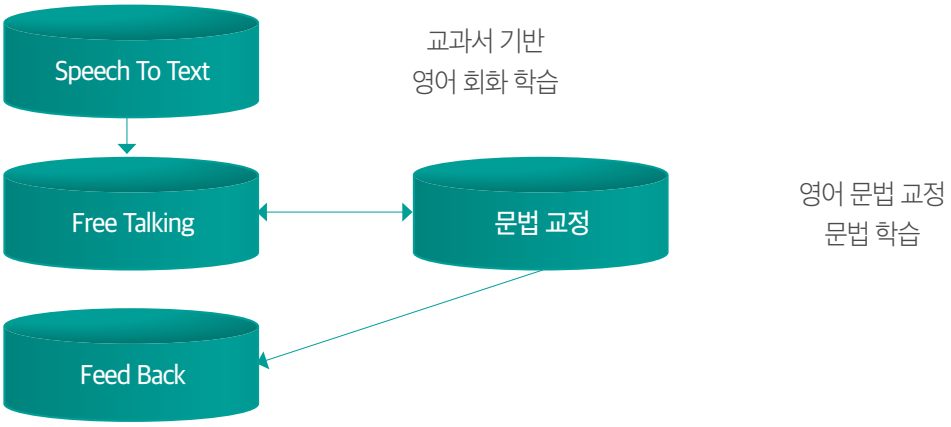
1. 기술 설명

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

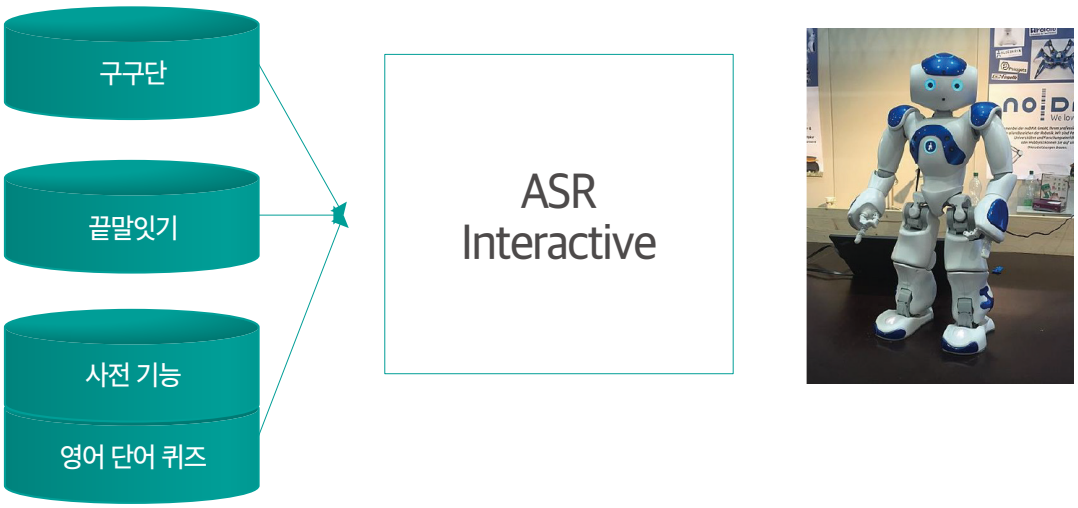
• 시나리오 기반 Free Talking

- 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

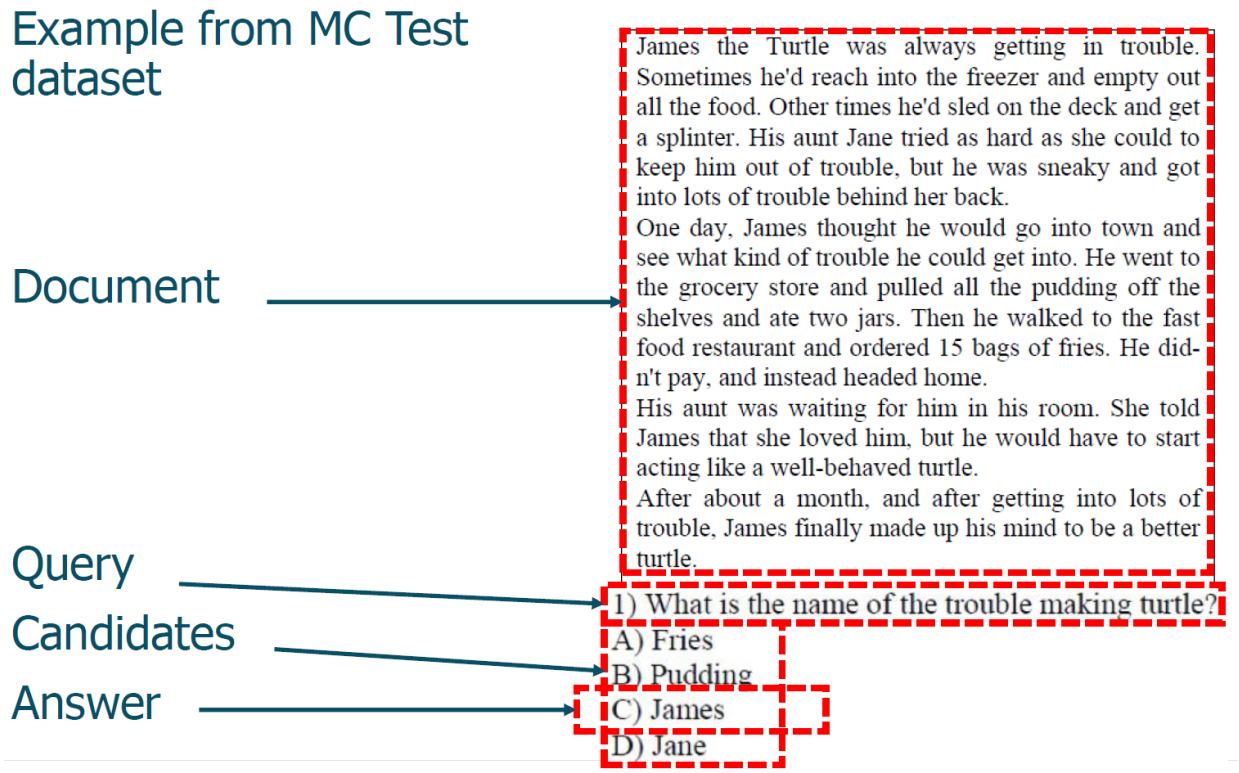
- 한국어 단어로 끝말잇기 게임을 할 수 있으며, 한국어 학습에 도움을 줌
- 게임 옵션: 중도취소, 다시 듣기, 게임횟수 설정

• 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

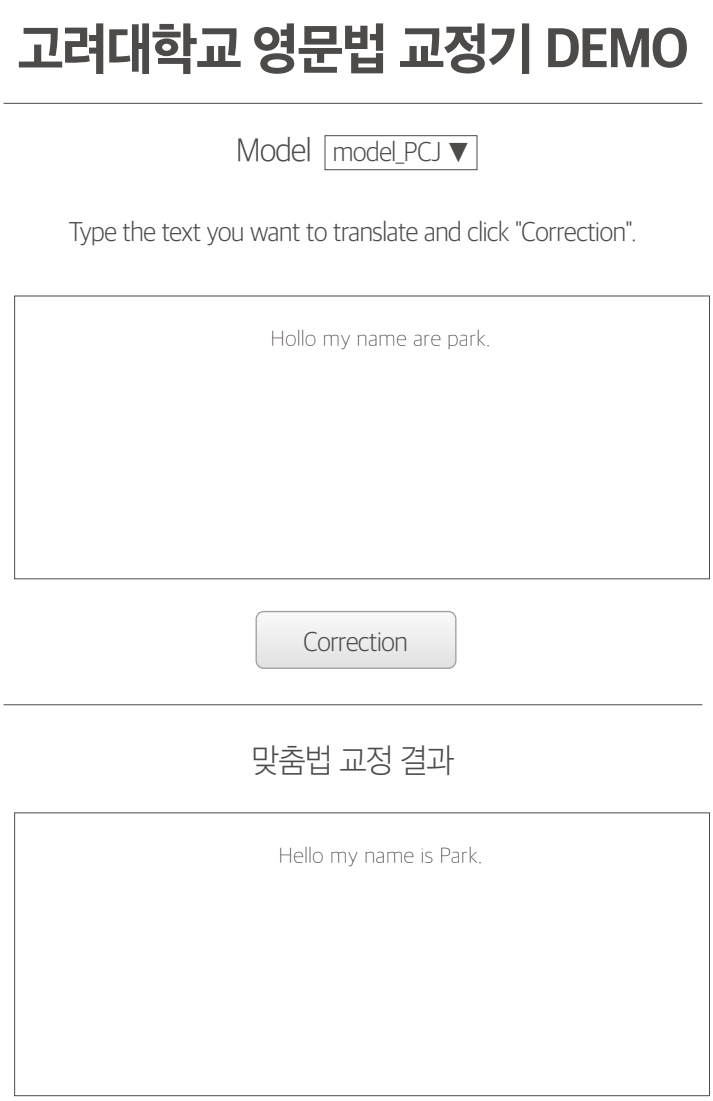
- 기계독해(MRC, Machine Reading Comprehension)란 인공지능 알고리즘이 스스로 문제를 분석하고 질문에 최적화된 답안을 찾아내는 기술이다.
- 본 기술은 사용자-로봇의 대화를 통하여 기계독해가 가능하도록 개발하였으며, 10초동안 사용자가 로봇에게 이야기를 들려주고, 로봇에게 이야기와 관련된 질문을 하면 로봇이 사용자의 이야기에서 정답을 추론하여 답을 한다.



3. 실행 결과

• 딥러닝 기반 영어 문법 교정기

• 데모 http://nlplab.iptime.org:32292/




뉴스 기사 추천 시스템 DEMO

verson 2.2

최신뉴스 데이터베이스, 2021.01.12 ~ 2021.01.14

created by Jaehyung Seo, NLP & AI LAB(Korea University)

연합뉴스로 가기

뉴스 기사 URL을 입력하세요!

여기에 입력하세요!

https://www.yna.co.kr/view/AKR20210118165000530?section=safe/news&site=topnews01

제출

추천 기사 목록을 아래에서 확인하세요!!

현재 열람 중인 뉴스 기사 제목

오늘도 400명 밀물듯…'코로나19 1년' 하루앞 일단 감소세 지속

현재 열람 중인 뉴스 기사 본문

(서울=연합뉴스) 김서영 기자 = 국내에서 신종 코로나바이러스 감염증(코로나19) 첫 확진자가 나온 지 20일로 꼭 1년째가 된다. [그래픽] 국내 코로나19 신규 확진자 (서울=연합뉴스) 김영은 기자 = 중앙방역대책본부는 18일 0시 기준으로 국내 신종 코로나바이러스 감염증(코로나19) 신규 확진자가 389명 늘어 누적 7만2천729명이라고 밝혔다. 이 가운데 '사회적 거리두기' 단계 결정에 있어 주요 지표가 되는 지역발생 확진자 수는 일평균 491명이다.

추천 뉴스 기사 1순위

[코로나 1년] 위기에 빛난 k방역, 3차 대유행에 흔들…성공적 마무리 어떻게: (서울=연합뉴스) 신선미 강애란 김서영 기자 = 국내에서는 지난해 1월 20일 첫 신종 코로나바이러스 감염증(코로나19) 확진자가 발생한 뒤 1년간 약 7만명이 양성 판정을 받았다. 사회적 거리두기는 사람 간 접촉을 최소화해 전파를 막는 전통적인 감염병 대응 방식으로, 정부는 이 조치의 실효성을 높이기 위해 지난해 3월 말 다중이용시설의 영업을 중단시키는 '집합금지' 행정명령까지 동원했다.

추천 뉴스 기사 1순위의 점수

3.7338595

뉴스 기사 추천 시스템 DEMO

verson 2.2

최신뉴스 데이터베이스, 2021.01.12 ~ 2021.01.14

created by Jaehyung Seo, NLP & AI LAB(Korea University)

GPT2를 활용한 뉴스 기사 추천 시스템

- 2020.10.03 (1차 업데이트)
- 2020.11.10 (2차 업데이트)
- 2002.11.19 (2-1차 업데이트)

간략한 소개

GPT2 언어 모델을 활용해서 제목 및 본문을 바탕으로

현재 열람하고 있는 기사와 유사한 문맥과 어휘를 지니고 있는 뉴스 기사를 추천하는 시스템입니

유사도 점수는 0점에서 5점까지 설정되어 있으며, 기본 값은 2.5점으로 설정했습니다.

(수 차례 실험으로 2.5점을 설정했습니다!)

새로운 뉴스 기사는 '직접 데이터 넣어보기' 탭에서 뉴스 기사가 작성된 URL을 입력하면 됩니다.

(단, 파싱 규칙이 깔끔한 연합뉴스 URL을 사용하는 것을 권장)

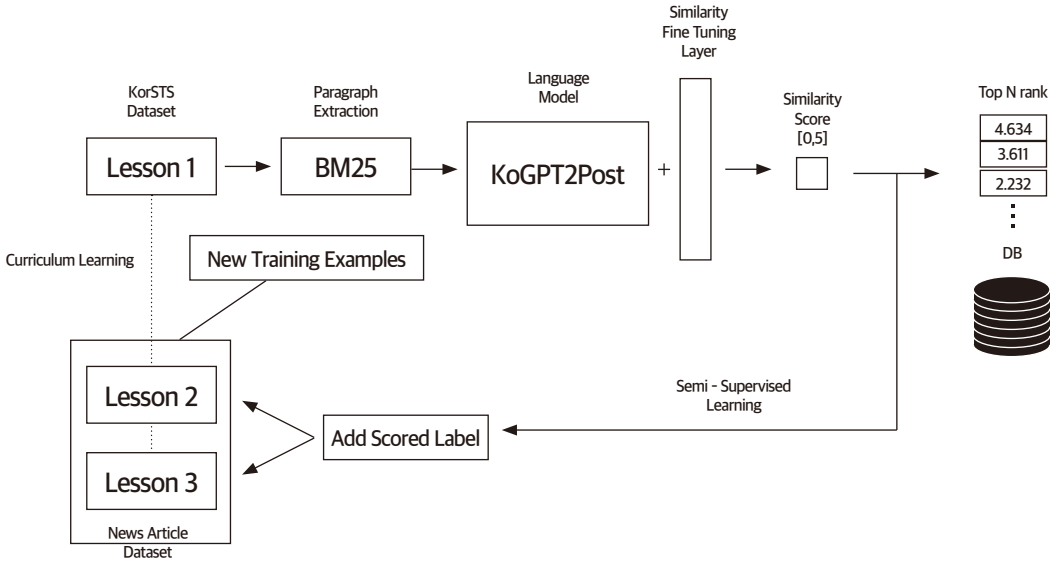
추천 뉴스기사 목록은 매 버전마다 최신 뉴스 기사로 업데이트합니다.

Created by: 서재형(Jaehyung Seo), NLP & AI Lab (고려대학교)

email: wolhalang@gmail.com

1. 기술 설명

- GPT2 언어 모델을 활용해서 두 개의 유사한 문서 사이의 유사도를 측정하는 방법.
- 뉴스 기사 문서 추출을 위해서 Newspaper3K를 활용하여 기사 제목 및 본문에 대해서 크롤링을 진행했으며, BM25를 통해 핵심 문장을 추출하여 활용.
- 문장 단위 유사도 비교에 그쳤던 기존 연구의 한계를 극복하기 위해 커리큘럼 학습과 준 지도 학습을 통해서 문단과 문서 단위에서도 문맥적 정보를 반영한 유사도 비교가 가능.



[그림] KoGPT2Post: 모델 학습 및 기사 추천 과정

2. 기술 방법

- 본 기술은 문서 단위의 유사도 비교를 통해서 실제 뉴스 기사와 데이터베이스 상에 저장되어있는 뉴스 기사 사이의 관계를 추론하여 사용자에게 유사한 기사를 추천하는 기술
- 기술의 흐름은 문장 단위의 유사도 비교 학습을 통한 KoGPT2 언어 모델의 미세 조정 훈련과 이후 BM25를 통한 문서 핵심 추출 문단과 문서에 대해서 유사도를 점진적으로 파악할 수 있도록 함.

3. 기술 활용 및 응용 분야

- 본 기술은 온라인 뉴스 기사를 제공하는 플랫폼에서 사용자가 현재 열람하고 있는 뉴스 기사와 유사한 뉴스 기사를 자동으로 피드에 노출하여 추천하는데 사용할 수 있음.

4. 실험

4.1 실험 개요

- 문장 유사도 측정을 위한 KorSTS 데이터셋을 활용하여, 스피어만 상관 계수를 통해 문장 간의 의미적 유사도를 예측하도록 훈련을 진행. 문장 이상의 데이터에 경우에도 동일한 방법을 사용했으며, 커리큘럼 학습 규칙에 따라서 점차 문장의 길이가 길어지도록 설정. 실제 뉴스 데이터를 학습 데이터로 활용하기 위해서 준 지도 학습을 통해 이전 단계의 평가 데이터를 다음 단계의 훈련 데이터로 사용.

4.2 실험 결과

[표1] 2,3단계 학습 성능 평가-뉴스데이터(512, 1,042)

모델	2단계 r_s	3단계 r_s
TF-IDF + Cosine	58.01	57.75
Doc2Vec + Cosine	30.68	29.01
KoGPT2 (SKT-AI)	87.81	92.55
KoGPT2Post (Ours)	89.43	94.14

[표2] 유사도 기반 추천 시스템 결과

현재 기사	추천 기사	유사도 점수
빅뱅 탑, 코로나19 극복 위해 1억원 기부, 그룹 빅뱅의 탑이 신종 코로나바이러스 감염증 피해 극복을 위해 ... 과거에 탑은 2018년 11월 4일 ... 용산 복지 재단에 이웃돕기성금 1104만원을 기부한 바 있다.	Rank 1. 신민아 기부, 의료진 및 취약 계층 위해 1억원 쾌척 ... 25일 신민아의 소속사 에임엔터테인먼트 측은 신민아가 코로나19 확산 방지에 노력하는 ... 사랑의 열매 측에 1억원을 기부했다고 밝혔다.	3.8666
빅뱅 탑, 코로나19 극복 위해 1억원 기부, 그룹 빅뱅의 탑이 신종 코로나바이러스 감염증 피해 극복을 위해 ... 과거에 탑은 2018년 11월 4일 ... 용산 복지 재단에 이웃돕기성금 1104만원을 기부한 바 있다.	Rank 2. 북구, 명의천사 기부챌린지 성황리에 마무리 ... 북구는 지난해 11월부터 올해 지난달 말까지 기부문화 확산을 유도하기 위해 ... 이웃을 돕는데 소중하게 사용될 예정이다.	3.6776
카카오톡 오류에 사용자들 발동동 ... 서버 불안정, 원인 파악 중 ... 카카오톡이 2일 오전 장애를 일으키며 많은 사용자가 불편을 ... 서비스 이용에 불편을 드려 죄송하다고 안내했다.	Rank 1. 카카오스토리도 오류? ... 오후 2시까지 시스템 점검 카카오스토리 측은 이날 홈페이지를 통해 보다 안정적인 서비스 이용을 위해 ... 앞서 이날 오전 9시께 카카오톡에 오류가 발생했다.	3.7187
카카오톡 오류에 사용자들 발동동 ... 서버 불안정, 원인 파악 중 ... 카카오톡이 2일 오전 장애를 일으키며 많은 사용자가 불편을 ... 서비스 이용에 불편을 드려 죄송하다고 안내했다.	Rank 2. 재택 근무 많은데 ... 카카오톡 먹통, 출시 10주년을 맞이하는 ... 이용자들이 큰 혼란과 불편을 겪었다 ... 카카오 측은 반복되지 않도록 최선을 노력을 다하겠다고 밝혔다.	3.4060

- 본 기술의 결과는 데모에서 확인 가능하며, 단어 빈도수를 통한 유사도 비교보다 문장의 길이가 점차 길어질수록 더 우수한 성능을 보임. 특히 오른쪽 표의 정성적 평가 결과를 통해, 단순히 어휘에 해당하는 반복뿐만 아니라, 문맥적인 정보를 고려해서 유사도 점수를 예측하는 것을 확인할 수 있음.

5. 데모

http://nlplab.iptime.org:32272/

15 나오 로봇을 활용한 이중 언어 교육

1. 기술 설명

• 최신 딥러닝 기반 기계번역 및 대화시스템을 이용하여 로봇을 이용한 스마트 자동통역 시스템뿐만 아니라 이를 확장하여 교육용 대화시스템을 개발하고자 함. 본 연구의 차별점으로 기존의 단일 언어 교육용 대화 시스템을 응용하여 이중 언어 대화 시스템을 개발하고자 함. 즉 외국어 학습을 위한 자동통역 소프트웨어를 NAO에 적재하고자 함. 사용자의 외국어 학습 동기 유발 및 학습 성취도 향상을 기대하며 인공지능과 이중언어를 사용하여 직접 대화하는 신개념 외국어 학습법을 제안함.

2. 기술 방법

• 자연어처리 분야 중 기계번역, 음성인식, 대화시스템 기술을 융합할 것이며 더 나아가 서비스적 관점, 사용자 관점을 반영하여 사용자 친화적 및 실질적으로 사용자에게 도움이 될 수 있는 외국어 학습을 위한 로봇을 이용한 스마트 자동통역 시스템 및 교육용 대화 시스템 (이중언어 대화 시스템)을 제작함.

• 기술적인 부분만 개발하지 않고 직접 대화 시나리오를 제작하여 인문학적인 요소와 기술적인 요소를 융합할 것입니다. 기술적인 요소로는 최신 딥러닝 기술 기반 기계번역, 음성인식, 음성 합성, 대화시스템 기술을 이용할 것이며 이를 각각 분리해서 서비스하는 것이 하나의 서비스로 융합하여 개발을 함.

1. 음성인식기술: NAO에서 제공하는 기본 API를 이용
2. 음성합성기술: NAO에서 제공하는 기본 API를 이용
3. 기계번역기술: 한-영, 영-한 기계번역을 직접 데이터 수집부터 번역 서버 및 웹서버까지 제작하여 REST API 형태로 제작. NAO는 REST API를 직접 이용하는 형태
4. 시나리오 제작: 최대의 학습 효과를 위하여 초등학교 교과서 기반 영어 Free Talking 시나리오를 제작

전체 프로세스

a. 대화를 시작할 준비를 한다

b. 사용자는 한국어로 대화를 시작한다.

c. 사용자가 발화를 하게 되면 STT(Speech To Text)기술을 이용하여 음성인식 결과가 도출된다.

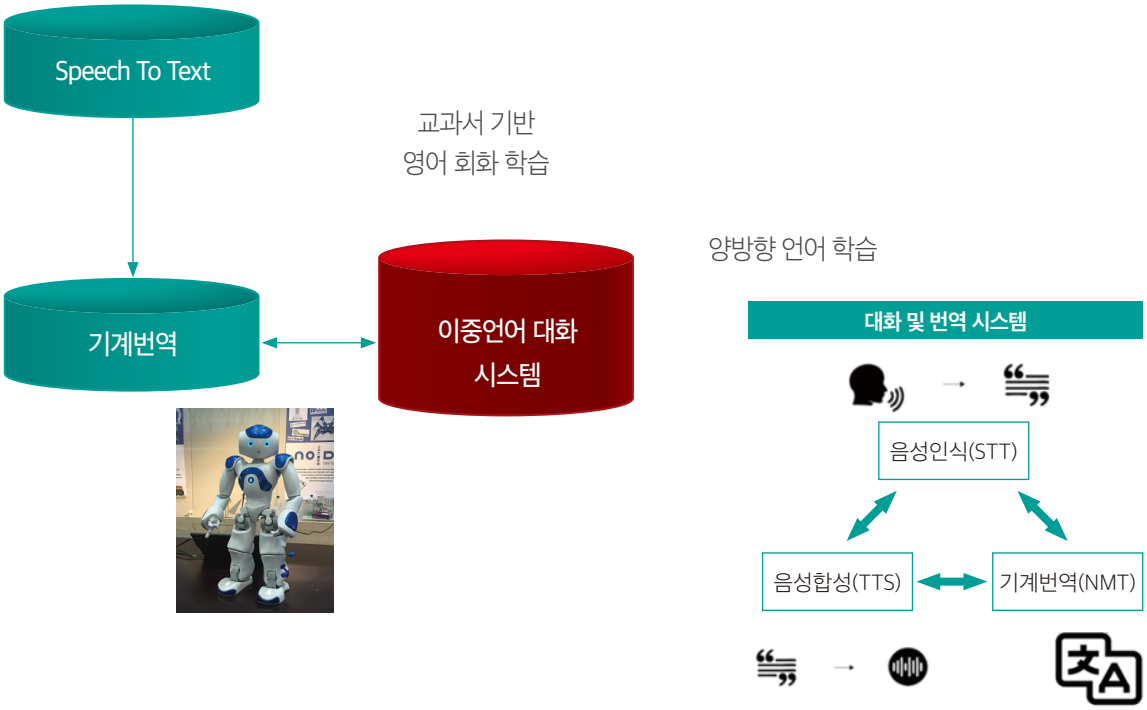
d. STT결과 값을 기계번역의 입력으로 넣는다. (필요시 음성인식 후처리를 위한 NLU 기술 도입)

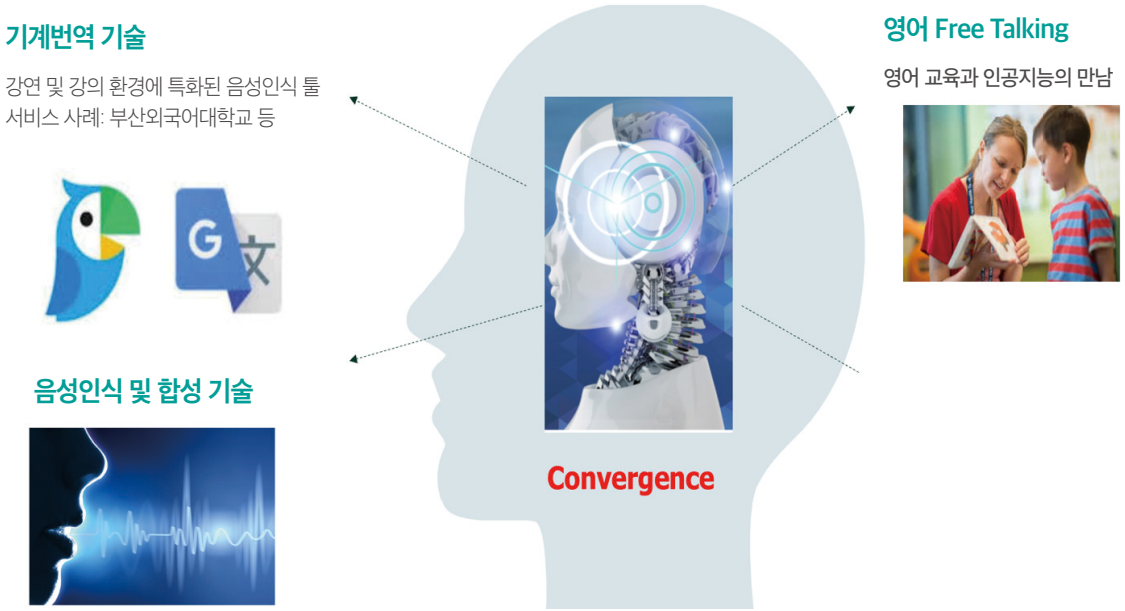
e. 번역 된 문장을 바탕으로 TTS (Text to Speech)기술을 이용하여 NAO는 영어로 발화를 진행한다. (신개념 이중언어 교육용 대화 시스템.

f. 이를 통해 자동통역 기능이 탑재된 신개념 교육용 대화 시스템을 이용한 언어 학습이 실현된다.

3. 기술 활용 및 응용 분야

본 아이디어와 유사한 제품이 중국에서 상용화하여 성공한 사례가 존재합니다. (Lilly) Lilly는 Single Turn 대화 방식으로 실제 외국어 학습에 도움이 되는지 의문입니다. 그러나 저희 팀이 제작할 제품은 Multi Turn 대화를 진행한 후 Free Talking이 가능할 뿐만 아니라 이중언어로 대화가 가능합니다.



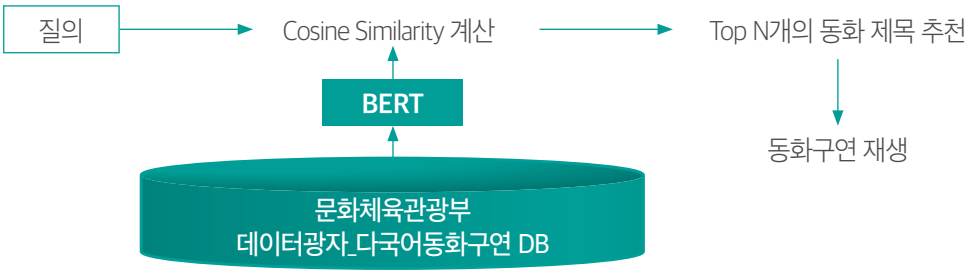


4. 데모

https://www.youtube.com/watch?v=4kCMMa62LDY&t=9s

1. 기술 설명

- BERT 언어 모델을 활용해서 사용자의 발화 내용과 유사한 제목을 지닌 동화책 본문 사이의 유사도를 측정하는 기술.
- 데이터베이스의 경우 문화체육관광부 데이터광장의 다국어동화구연 API를 활용. 제목, 본문 그리고 동화책을 읽어줄 수 있는 녹음 파일을 담고 있음.
- 사용자와의 연속적인 발화를 통해 유사도를 기반으로 한 N개의 동화를 추천하고, 나오는 사용자에게 동화를 재생.



2. 기술 방법

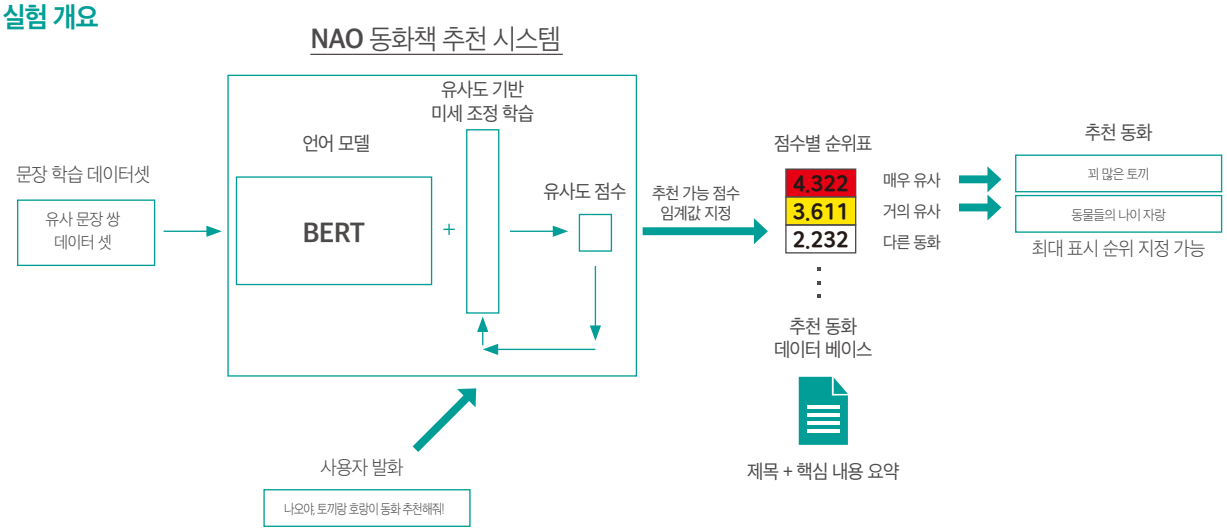
- 제목과 본문을 합친 문단 사이의 유사도는 BERT 언어 모델을 KorSTS 데이터셋을 통해 미세 조정 훈련을 진행.
- 미세 조정 단계에서 BERT는 문장 쌍에 얼마나 유사함을 지니고 있는지 양방향 인코더를 통과한 결과 값을 반환한 [CLS] 토큰이 지니고 있는 은닉 벡터 값과 사전에 제시한 점수 값 사이의 관계를 회귀 식을 통해 해석하여, 평균 제곱의 오차를 줄여나가는 방향으로 역전파를 통한 최적화를 진행.
- 아이들의 발화에 대한 로봇의 음성 인식 및 합성기술은 NAO에서 제공하는 기본 API를 이용.
- 음성 인식 이후 텍스트로 변환된 데이터는 질의에 대한 IR(Information Retrieval) 기술을 적용.
- 해당 질의는 미세 조정을 완료한 BERT를 통해서 문서요약을 통해서 핵심 주제를 담고 있는 동화의 초록 사이의 언어 모델이 반환하는 유사도 점수를 통해 상위 N개의 동화 목록을 추천함. 추천 유사도 점수에 대한 임계값과 목록의 개수는 사용자가 직접 조정할 수 있도록 제공할 예정.

3. 기술 활용 및 응용 분야

• 추천 시스템용 대화 시스템을 응용하여 인공지능 로봇 NAO에 적재함. 해당 시스템은 아이들의 질의를 바탕으로 동화책을 추천하고, 선택한 동화에 대해서 로봇이 특정한 동작과 함께 동화책을 읽도록 제작. 이러한 방식은 독서의 첫 단추를 강압과 의무에 의한 숙제와 같은 요소가 아닌, 로봇과 함께할 수 있는 하나의 놀이처럼 인식하도록 함.

4. 실험

4.1 실험 개요



4.2 전반적인 기술 진행도

(1) 나오의 전원을 연결하여 대화를 준비함.

(2) 사용자는 한국어로 발화 (최초 대화에서 '나오야'를 통해서 호출한 이후에 대화를 이어나가기).

(3) 사용자 발화는 나오의 STT(Speech To Text) 기술을 통해 음성인식 결과를 반환함.

(4) 반환 값은 추천 동화책 검색의 입력으로 사용함(다양한 상황에 따른 실험을 통해서 필요하다면 반환 값에 대한 후처리 NLU 기술 도입. 현재는 언어 모델을 활용한 만큼, 어휘 정보뿐만 아닌 발화 문맥을 고려해서 추천).

(5) 반환된 발화 문장과 유사도 평가에 대한 미세 조정을 마친 BERT를 통해 임베딩 요약 동화책 정보 간의 유사도 점수를 통해서 상위 3개의 동화책을 추천함.

(6) 추가적인 발화를 통해서 추천 동화책 중 1권을 선택하고, 선택된 동화책 정보를 문화체육관광부의 다국어동화구현DB의 동화책 내용을 불러옴.

(7) DB에서 반환된 값은 TTS(Text to Speech) 기술을 통해서 나오는 한국어로 동화책의 내용을 재생함.

(8) 구축 데이터베이스 상의 동화책 읽기 기능이 탑재된 로봇 나오를 통해 아이들은 동화책을 통해 독서에 대한 첫걸음을 즐거운 기억으로 인식하며, 대화를 통한 독서 학습을 실현함.

• 본 기술의 결과는 영상에서 확인 가능하며, 단순히 어휘에 해당하는 반복뿐만 아니라, 발화 문맥적인 정보를 고려해서 유사도 점수를 예측하는 것을 확인할 수 있음.

5. 동영상 링크

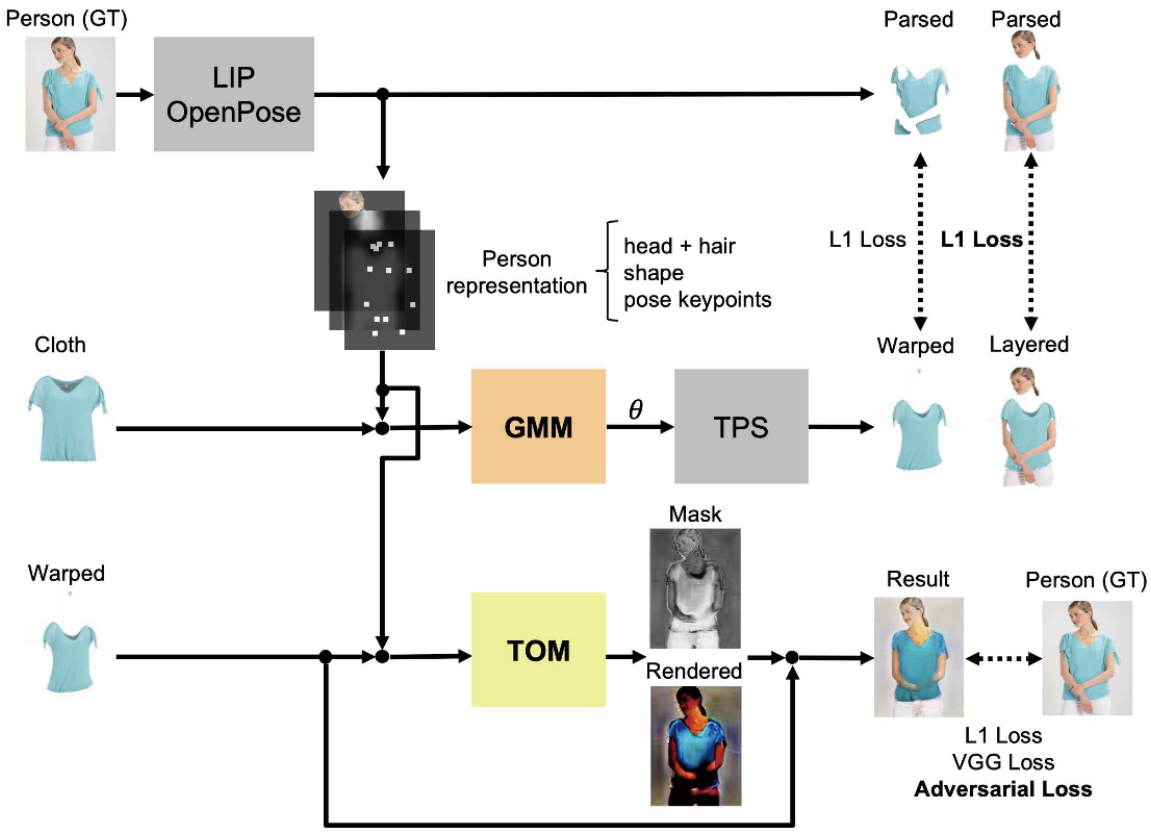
https://www.youtube.com/watch?v=rI8ep18-ISE

1. 기술 설명

- VITON-GAN과 같은 Deep Learning 기반 AI 모델은 매장 내 의류 및 모형 모델을 사용하여 시범적으로 볼 수 있는 가상 이미지를 생성할 수 있음
- 3D 정보 외에 다른 정보를 사용하지 않고 영상 기반의 Virtual Try-On Generator(VITON-GAN)을 사용하여 세밀한 부분까지 보완하여 원하는 의류 아이템을 모형 모델에 맞게 제공함
- 본 모델의 generator는 CP-VTON에서 구현된 GMM(geometry matching module)과 TOM(try-on module)로 구성되며, blockade를 해결하기 위해 TOM에 adversarial loss를 추가하였음

2. 기술 방법

- CP-VTON에는 아래 그림과 같이 크게 세 단계가 있음



- 첫 번째로 TOM은 TOM 결과 이미지, 매장 내 의류 이미지 및 사람 표현을 입력으로 사용하는 판별자에 대해 적대적으로 훈련되고 결과가 실제인지 가짜인지 판단함
- 두 번째, GMM의 loss function은 신체에 겹쳐진 옷의 생성 이미지와 실제 이미지 사이의 L1 distance를 포함함
- 마지막으로, 데이터를 확장시키기 위해 무작위로 수평 뒤집기 하여 사용됨

3. 실험

- 데이터 세트에는 16,253개의 여성 모델 이미지와 상위 의류 이미지 쌍이 포함되어 있으며, 이 쌍은 각각 training(13,221쌍), validation(1,000쌍), test(2,032쌍) set으로 구성되어 있음
- VITON-GAN은 blockade 문제를 해결하기 위해 CP-VTON보다 더 명확하게 손과 팔을 생성함

4. 데모

http://nlplab.iptime.org:32299/

기계번역

고려대학교 다국어 신경망 기계번역기

딥러닝 기반 한국어 고전번역기

PicTalky: Text to Pictogram

COVID-19 도메인특화 기계번역기

인간의 인지과정을 반영한 도메인 특화 번역기

1. 기술 설명

- 모델의 변경 없이 각종 pre-processing 및 post-processing을 통해 모델의 성능을 향상시킬 수 있다는 연구의 움직임을 기반으로 low-resource 언어인 Korean-English NMT에 다양한 decoding strategies를 적용하여 모델의 변경 없이 번역 성능이 향상됨을 비교 실험을 통해 증명함
- Beam size에 따른 성능 변화 실험, n-gram blocking에 따른 성능 변화 실험, length penalty를 적용하였을 때 성능 향상 여부 등의 실험을 진행하였고, 실험결과 다양한 decoding strategies가 성능 향상에 도움이 됨을 알 수 있었으며 기존 Korean-English NMT 연구들에 비해 비교적 좋은 성능을 보임

7 고려대학교, Machine Translation 한-영 기계번역 모델 - 한국어·영어 번역 말뭉치 AI데이터 활용

- **(연구 개요)** 고려대학교 박찬준 학생이 한·영 번역 말뭉치 AI 데이터를 활용하여 개발한 기계번역 모델의 성능 향상

〈 고려대학교 Machine Translation 〉

- IWSLT*에서 기계독해 모델의 성능평가를 위해 사용하는 데이터셋 Test2016와 Test2017을 활용하여 테스트 진행

* The International Workshop on Spoken Language Translation : AI를 활용해 통역, 번역의 정확성을 겨루는 대회

- **(연구 결과)** BLEU* 점수에서 16.38(Test2016 기준), 14.03(Test2017 기준)로 기존 타 대학의 기계독해 모델에 비해 상대적으로 높은 연구성과 창출

* BLEU Bilingual Evaluation Understudy : 기계 번역 결과와 사람이 직접 번역한 결과가 얼마나 유사한지 비교하여 번역에 대한 성능을 측정하는 방법

https://www.aihub.or.kr/node/4525

2. 기술 방법

• 본 연구에서 실험을 진행한 다양한 decoding strategies은 크게 3가지로 beam Size에 따른 성능 변화 실험, n-gram blocking에 따른 성능 변화 실험, length penalty와 stepwise penalty에 따른 성능 변화 실험을 진행하였다. 해당 strategies들을 독립적으로 적용하는 것이 아닌 점층적인 pipelining 형태로 적용하여 가장 최적의 성능을 도출해내었다.

3. 기술 활용 및 응용 분야

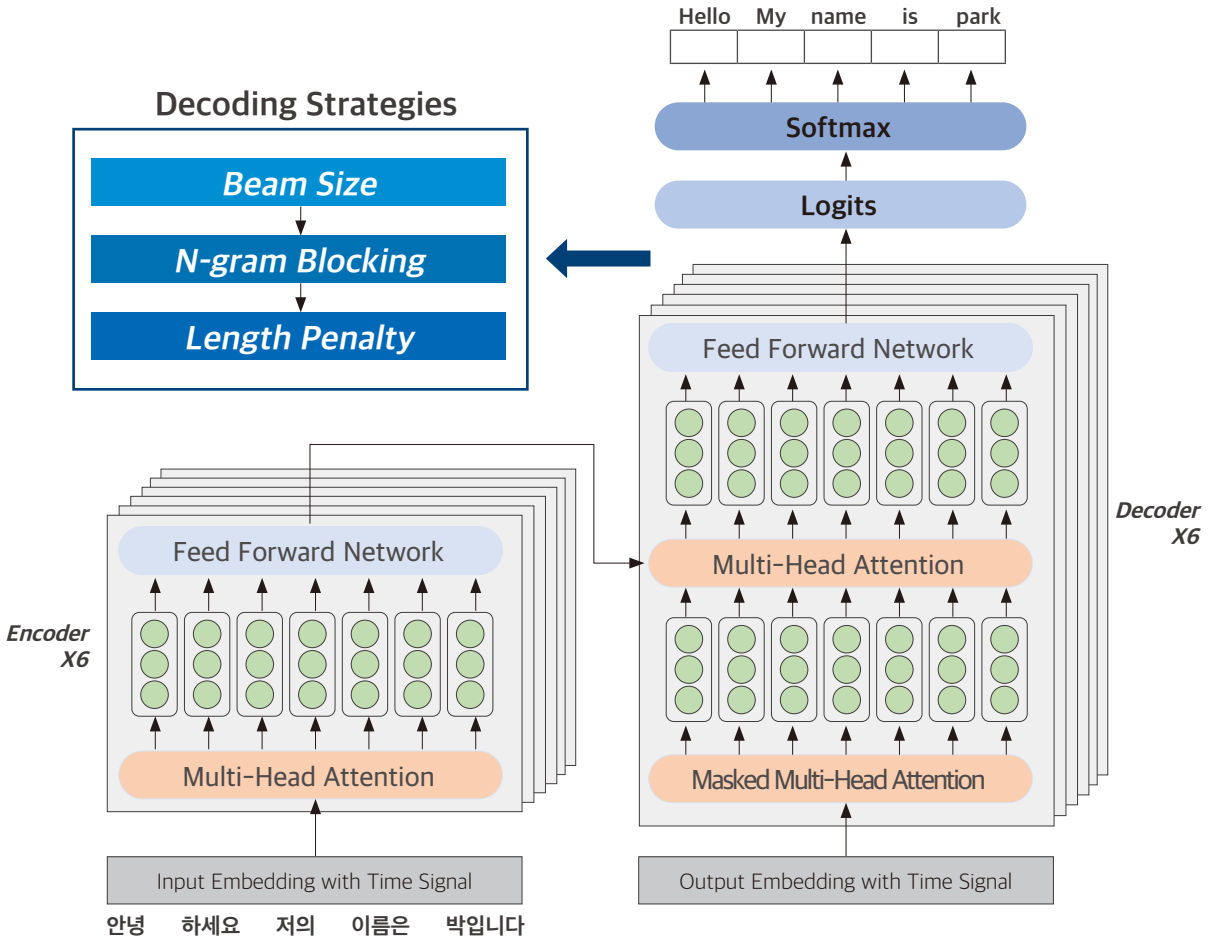
• 다양한 이종 언어들과의 번역

• 번역 사업

4. 실험 (Only PDF)

4.1 실험 개요

• 본 연구는 Transformer를 기반으로 한 Korea University(KU) model을 baseline으로 하여 이를 기반으로 다양한 decoding strategies을 적용하여 비교 실험을 진행하였다.



4.2 실험 결과

• Post Processing 적용시 성능이 향상됨을 확인함

Beam size	Iwslt 16	Iwslt 17
Beam 1	17.27	14.84
Beam 2	**17.77(+0.50)**	**15.19**
Beam 3	17.51	14.99
Beam 4	17.49	14.83
Beam 5	17.34	14.75
Beam 6	16.97	14.49
Beam 7	16.81	14.41
Beam 8	16.78	14.31
Beam 9	16.67	14.29
Beam 10	16.46	14.23

Beam size	Total Time	Average	Token per /s
Beam 1	13.929	0.012	1609.359
Beam 2	14.667	0.012	1477.046
Beam 3	15.711	0.013	1353.141
Beam 4	16.241	0.014	1292.145
Beam 5	17.683	0.015	1175.981
Beam 6	18.565	0.016	1101.098
Beam 7	19.679	0.017	1026.473
Beam 8	20.949	0.018	960.227
Beam 9	22.693	0.019	881.692
Beam 10	23.907	0.020	828.938

N-gram Blocking	Iwslt 16	Iwslt 17
Uni-gram	5.14	4.98
Bi-gram	15.98	14.43
Tri-gram	17.62	15.09
4-gram	17.65	**15.24**
5-gram	17.74	15.22
6-gram	17.75	15.21
7-gram	17.72	15.20
8-gram	**17.77**	15.20
9-gram	**17.77**	15.20
10-gram	**17.77**	15.20

Penalty	Iwslt 16	Iwslt 17
Average Length Penalty	17.94	**15.42(+0.08)**
Step Wise Lenght Penalty	17.79	14.95
(Average+Step Wise)Length Penalty	**17.98(+0.71)**	15.22

5. 데모

http://nlplab.iptime.org:32296

1. 기술 설명

- 고전번역: 조선왕조실록, 승전원일기와 같은 고어를 번역하는 것을 의미함
- 기계번역: 소스문장(Source Sentence)을 타겟문장(Target Sentence)으로 컴퓨터가 번역하는 시스템을 의미하며 이를 고전번역에 적용할 경우 소스문장에 고어 타겟 문장에 한국어가 적용될 수 있음

기존 방식의 고전번역의 한계

- 사람이 아무것도 안하고 고전번역만 하는데 80년이 걸림
- 고전번역 전문가 양성의 어려움이 있고 제한된 인력 구조
- 현재 고전번역 전문가는 200여명 수준이며 고전번역자 양성 기간은 관련학과 졸업자기준으로 10년이상 소요됨(한국 고전번역의 현황과 과제, 2015년 국정감사 정책 자료집)
- 고전번역을 위한 관련 지식 및 실력에서 개인별 편차가 있음. 이에 따라 번역결과물의 품질편차가 발생하게 됨.

인공지능 기술의 발전

- 딥러닝의 등장으로 기존 RBMT,SMT보다 좋은 성능의 기계번역기를 개발할 수 있음
- 기계번역 기술을 활용하여 고전문자를 복구하려는 시도가 최근에 여러 논문에서 연구됨. (일본의 Kuronet, 그리스 고어, Decipher)

NMT기반 고전번역의 장점은?

- 기존 고전번역사들의 업무 효율성 강화
- 빠른 시간에 번역 가능
- 플랫폼을 통한 번역결과물의 DB화 및 지식증강형 Infinite Training모델 구축
- 품질 편차를 최소화하고 일관된 번역 품질을 만들어 낼 수 있음.
- 미번역된 문서에 대한 번역도 가능하다. (규장각 도서 등)

2. 기술 방법

- 본 연구는 고전번역에 특화된 서브워드 분리기법을 적용하면 모델의 성능을 획기적으로 올릴 수 있다고 판단하여 동일한 모델의 다양한 Subword Tokenization 방법을 적용하여 실험을 진행하였다. 고전번역에서 중요한 요소 중 하나로 Entity를 얼마나 잘 번역하는 것이냐이다. 고전번역의 데이터를 보면 사람의 이름, 장소, 기관 등이 문장의 대부분을 차지한다. 그 당시에는 기록을 남기는 것이 중요한 문제였기에 Entity의 대한 정보가 상당히 중요하다. 이러한 고전번역에 특징의 기반하여 본 논문에서 Entity 정보를 서브워드 분리 작업에서 Restrict를 진행하였다. 즉 Entity Based Vocabulary Restriction 방법론을 제안한다.
- 즉 만약 “이순신”이라는 인명 정보가 나오게 된다면 해당 정보는 Subword Tokenization은 진행하지 않고 그대로 유지하게 된다. 즉 Entity정보를 분리하지 않고 학습 데이터의 이용하는 방법론이다.

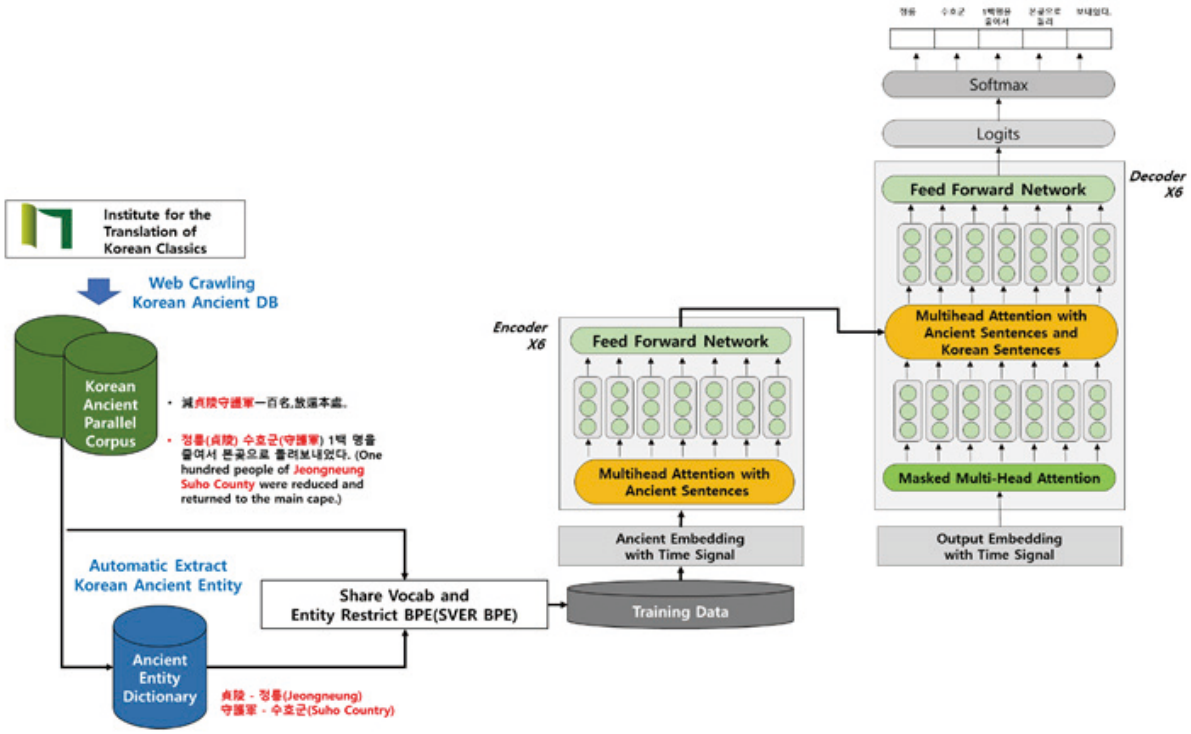
3. 기술 활용 및 응용 분야

- 고전 문서의 현대적 풀이
- 과거의 일상적인 삶의 모습과 당대 생활 자체 복원

4. 실험 (Only PDF)

4.1 실험 개요

- 인공신경망 기계번역 기술을 고전 문헌 번역에 활용한 'AI 기반 고전문헌 자동번역시스템'을 구축했다. 대표적인 Sequence to Sequence 모델인 LSTM-Attention 그리고 Transformer기반의 모델을 이용하여 고전번역기의 성능과 Subword Tokenization을 어떻게 하느냐에 따라 성능이 어떻게 달라지는지 확인해본다.



4.2 실험 결과

- 서브워드 분리를 어떻게 하느냐에 따라 다양한 실험을 진행하였다. Char단위, B.P.E, Sentencepiece Unigram 방법과 제안하는 Entity and Vocab Restrict 방법을 통해 서브워드 분리를 진행한 후 실험결과를 비교하였다. 추가적으로 Vocab은 그대로 놔두고 Entity만을 분리하였을 때 성능이 어떻게 변화하는지도 살펴보았다.

Model	BLEU	Token Per Second
Scntcnccpicce-LSTM-Attcntion	24.39	2758
Sentcnccpiccc-Transformcr	22.69	982
BPE-LSTM-/₩ttcntion	25.18	2029
BPE-Transformcr	24.43	1122
Char- LSTM-/₩ttention	23.66	8785
Char- Transformer	16.24	1466
Entity Restrict- LSTM-Attention	14.74	3013
Entity Restrict- Transformer	15.12	1174
(Our) Share Vocab and Entity Restrict BPE - LSTM Attention	29.40	5004
(Our) Share Vocab and Entity Restrict BPE - Transformcr	**29.68**	1379

5. 데모

http://nlplab.iptime.org:32242/

1. 기술 설명

- 언어발달 장애를 가진 아동들은 일상생활 및 사회생활에서 많은 어려움을 겪으며 이는 생애 전반을 걸쳐 지속됨
- Augmentative and Alternative Communication(AAC, 보완대체 의사소통)는 언어장애를 앓는 이들에게 실질적인 의사소통 수단으로 사용될 수 있음
- 본 연구는 픽토그램을 AAC의 수단으로써 최대한 활용하여 언어발달 장애 아동이 타인과 의사소통하고 언어 이해 능력을 향상시킬 수 있도록 돕는 딥러닝 기반 인공지능 서비스임

2. 기술 방법

- 픽토그램은 대표적인 보완 대체 의사소통 수단으로 언어의 어려움이 있는 사람들에게 도움이 된다. 픽토그램과 같은 전달 매체는 규칙 및 기호체계를 이해해야만 하는 언어와 다르게 보다 직관적으로 빠르게 의미를 전달할 수 있으며 이로 인하여 픽토그램은 의사소통 장애를 치료하고 개선하는 데에 보조적으로 사용될 뿐만 아니라 정보 전달 수단으로도 널리 사용된다.
- 픽토그램은 그림 교환 의사소통 체계(PECS, Picture Exchange Communication System)에도 적극 활용되며 이를 언어 재활 분야에도 응용하고 있다. 의사소통판(Communication Board)에 그려진 그림을 이용하여 타인과 의사소통하는 법을 픽토그램을 통해 터득할 수 있는 것이 대표적인 사례이다. 또한 그림을 통해 문장을 만들고 대상 식별과제를 수행하는 등 아동의 언어능력과 인지능력을 동시에 향상시킬 수 있다. 이와 같은 방법은 언어 체계를 배우지 못한 아동들의 언어 이해력 증진과 구어 발화에 실질적인 도움을 준다.

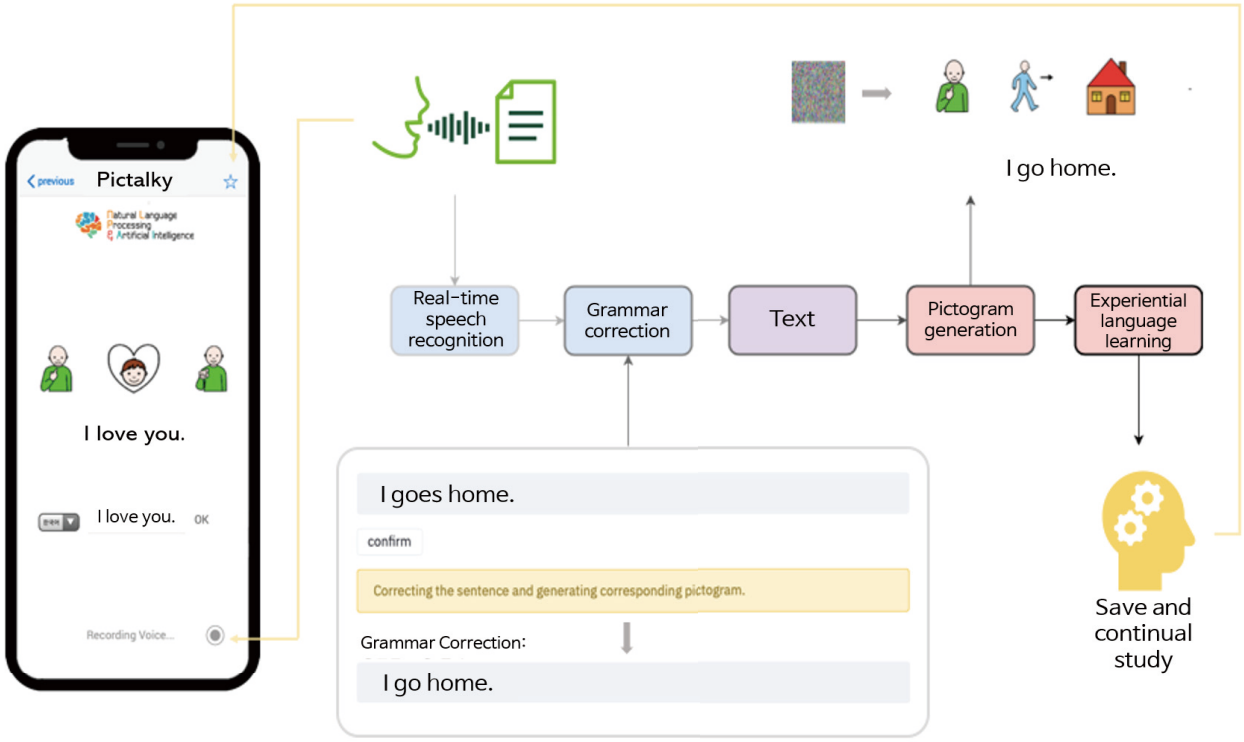
3. 기술 활용 및 응용 분야

- 지적장애와 자폐성 장애에 기인한 발달장애인 중 0세~14세에 해당하는 발달 장애 아동들의 의사 소통 수단으로 사용가능

4. 실험 (Only PDF)

4.1 실험 개요

- 본 연구에서 제안하는 서비스는 발달 장애 아동의 의사소통을 돕고 언어 이해를 증진시키는 데에 목적이 있다. 발화 내용을 청각 및 시각적으로 동시에 인코딩하여 전달하므로 사용자가 언어를 잘 알지 못하더라도 화자의 의도를 직관적으로 이해할 수 있다. 또한 텍스트와 이미지가 함께 전달되기에 언어의 요소들을 직접적으로 가르쳐주지 않아도 스스로 추론하여 언어를 배울 수 있는 암묵적 학습 또한 가능하다. 따라서 제안하는 서비스는 발달 장애 아동을 대상으로 제작되지만, 전반적인 언어에 관하여 재활 치료, 특수교육 정보 전달의 목적으로도 두루 적용될 수도 있다.



[그림] The architecture of proposed Deep learning-based AAC service

4.2 실험 결과

입력으로 I love danceing이라는 오류문장이 들어가면 딥러닝 기반 영문법 교정기를 통해 I love dancing이라는 문장으로 교정을 진행한다. 교정을 진행한 문장을 Text to Pictogram 모듈을 통해 텍스트를 픽토그램으로 변경해주게 된다.

5. 데모

http://nlplab.iptime.org:32257/

1. 기술 설명

- 최근 세계보건기구(WHO)의 Coronavirus Disease-19(COVID-19)에 대한 팬데믹 선언으로 COVID-19는 세계적인 관심사이며 많은 사망자가 속출하고 있다. 이를 극복하기 위하여 국가 간 정보 교환과 COVID-19 관련 대응 방안 등의 공유에 대한 필요성이 증대되고 있다.
- 이러한 요구에 맞춰 우리 연구소에서는 COVID-19 도메인에 특화된 인공신경망 기반 기계번역(Neural Machine Translation(NMT)) 모델을 개발하였다.
- 이 모델은 영어를 중심으로 프랑스어, 스페인어, 독일어, 이탈리아어, 러시아어, 중국어 지원이 가능한 Transformer 기반 양방향 모델이다.
- 실험결과 BLEU 점수를 기준으로 상용화 시스템과 비교하여 모든 언어 쌍에서 유의미한 높은 성능을 보였다.

2. 기술 방법

- COVID-19 도메인에 특화된 번역기를 위한 특화 방법은 다음과 같은 단계로 이루어진다:

  1) COVID-19 관련 데이터 수집

  2)해당 도메인에 특화된 전처리 기법(Subword Tokenization 모델 제작 시 해당 도메인의 데이터로만 모델 제작

  3) COVID-19 도메인에 특화된 Vocab 추출

  4) Sequence to Sequence 모델을 이용한 도메인 특화 모델 제작

  5) 특화된 번역기와 기보유 된 번역 엔진과의 성능 비교 평가

- 도메인 특화에서 무엇보다 중요한 요소는 해당 도메인에 특화된 데이터를 구축하는 일이며 이는 시간과 비용이 많이 드는 작업이다. 그러나 본 논문에서 사용한 Corona Crisis Corpus같은 경우 TAUS에서 모든 사람들에게 무료로 오픈되어 사용되고 있으며 이로 인하여 데이터 구축에 대한 시간과 비용을 절약할 수 있다.

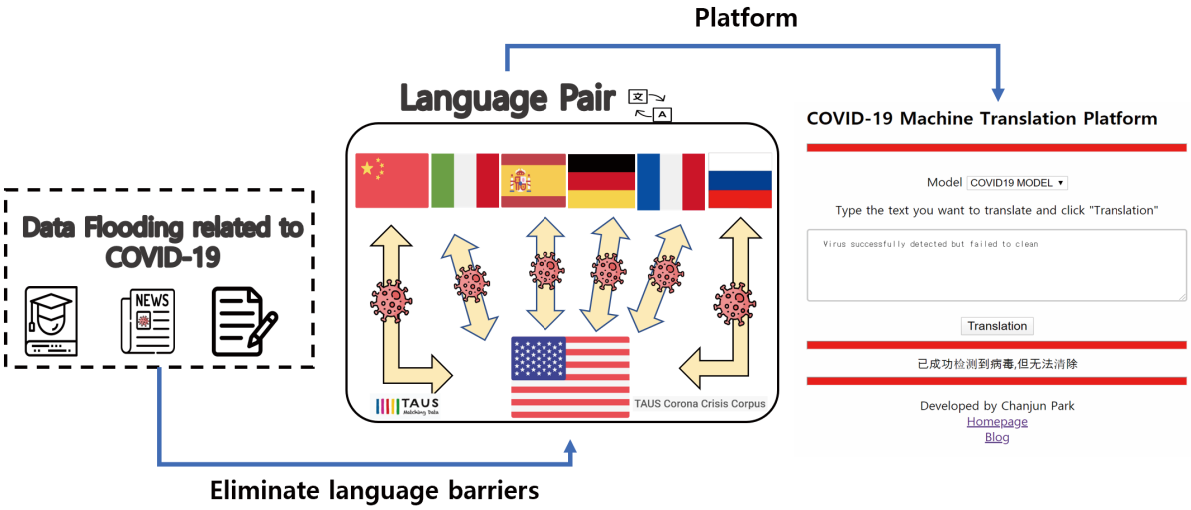
3. 기술 활용 및 응용 분야

- 특화된 도메인 용어에 대한 올바른 번역에 응용

4. 실험 (Only PDF)

4.1 실험 개요

- 본 연구에서 실험을 위한 데이터로 TAUS에서 공개한 Corona Crisis Corpus를 이용하였다. 해당 코퍼스는 영어를 중심으로 스페인어, 이탈리아어, 프랑스어, 독일어, 러시아어, 중국어의 병렬 말뭉치를 제공해준다.



[그림] Concept of COVID-19 Neural Machine Translation Platform

4.2 실험 결과

• 실험결과 본 논문에서 제안한 번역 모델이 상용화 시스템인 구글 번역기와 비교하여 모든 언어쌍에 대해여 BLEU 점수와 BLEU1, BLEU2, BLEU3, BLEU4까지 모든 수치에서 높은 성능을 보였다.

Experimental Results of COVID-19 Model versus Google Translation

Model	BLEU	BLEU1	BLEU2	BLEU3	BLEU4
(Our) English-Chinese	26.23	53.70	32.10	23.90	19.80
(Google) English-Chinese	15.36	47.40	21.10	10.90	6.10
(Our) Chinese-English	36.28	65.80	44.10	34.30	28.00
(Google) Chinese-English	29.49	59.70	35.30	23.60	16.20
(Our) English-French	46.10	71.60	53.70	42.30	33.80
(Google) English-French	43.21	68.30	50.20	38.30	29.50
(Our) French-English	48.62	74.20	54.60	42.50	33.60
(Google) French-English	44.61	69.20	50.30	38.50	29.50
(Our) English-German	35.21	64.00	42.00	31.10	24.10
(Google) English-German	26.03	53.10	31.40	20.30	13.50
(Our) German-English	41.89	71.20	49.80	38.10	30.10
(Google) German-English	36.00	64.70	43.00	30.50	22.20

(Our) English-Italian	44.80	70.20	51.10	40.00	32.10
(Google) English-Italian	39.64	64.50	45.40	34.00	26.00
(Our) Italian-English	50.21	75.50	56.00	44.30	35.90
(Google) Italian-English	47.75	72.90	54.30	42.80	34.10
(Our) English-Spanish	44.40	71.50	51.80	40.20	32.00
(Google) English-Spanish	40.44	66.30	46.30	34.50	26.20
(Our) Spanish-English	46.69	74.30	54.00	42.50	34.20
(Google) Spanish-English	42.89	68.20	48.50	36.50	28.00
(Our) English-Russian	28.09	56.50	35.20	25.30	18.90
(Google) English-Russian	26.08	53.40	33.30	22.50	15.50
(Our) Russian-English	34.35	65.10	41.00	29.70	22.30
(Google) Russian-English	31.09	58.70	36.50	24.90	17.50

5. 데모

http://nlplab.iptime.org:32250/

1. 기술 설명

- 도메인특화 NMT를 만들기 위한 기존 방법들은 대부분 general corpora에 대한 pretrain을 거친 후 domain-specialized corpora에 대한 finetuning을 하는 방식으로 진행되었다.
- 해당 기술은 cross language speech perception과 관련한 인지과학적 이론을 바탕으로 기존의 방법들을 재해석하였고, 인간의 인지과정에서 모티브를 얻은 새로운 도메인특화 방법론인 Cross Communication Method(CCM) 방법론이다. 실험결과 기존의 방법론들과 비교하여 양적으로나 질적으로나 더 우수한 성능을 거두었다.
- CCM: Cross Communication Method for Domain Specialized Neural Machine Translation

2. 기술 방법

- CCM에서는 Primary mapping으로 인한 secondary mapping의 제약을 없애기 위해 mapping 과정을 직렬화하지 않았다. 그리고 general corpora와 domain specialized corpora가 배치 내에서 소통할 수 있도록 배치 구성면에서 기존 방법과의 차별점을 두었다. 더 나아가 일반 코퍼스는 source language와 target language에 대한 일반적인 번역을 학습하고, 도메인 특화 코퍼스는 도메인에 특화된 용어들과 표현들을 학습한다는 점에서 각각 성격이 구별된다는 점을 감안하여 배치 구성 시 비율을 고려했다.
- 본 연구는 cross language speech perception과 관련한 해석들을 바탕으로 기존 방식에 대한 의문을 가지게 되었다. 영유아는 이중 언어 음성을 인식 및 구별할 때 primary mapping의 영향 없이 phoneme들을 구별해낼 수 있다. 그러나 어른의 경우 특정 언어에 대한 mapping이 고정되어 있기 때문에 새로운 언어의 음성을 구별하고자 할 때 initial mapping에 의해 새로운 mapping의 학습을 제한받게 된다. 이에 대해 우리는 도메인 특화 기계번역에서 PFA technique를 활용하는 것이 과연 옳은지에 대한 의문을 가지게 되었다. 따라서 본 연구에서는 기존의 방법에서 탈피하여 새로운 도메인 특화 기계 번역인 Cross Communication Method(CCM)을 제안한다.

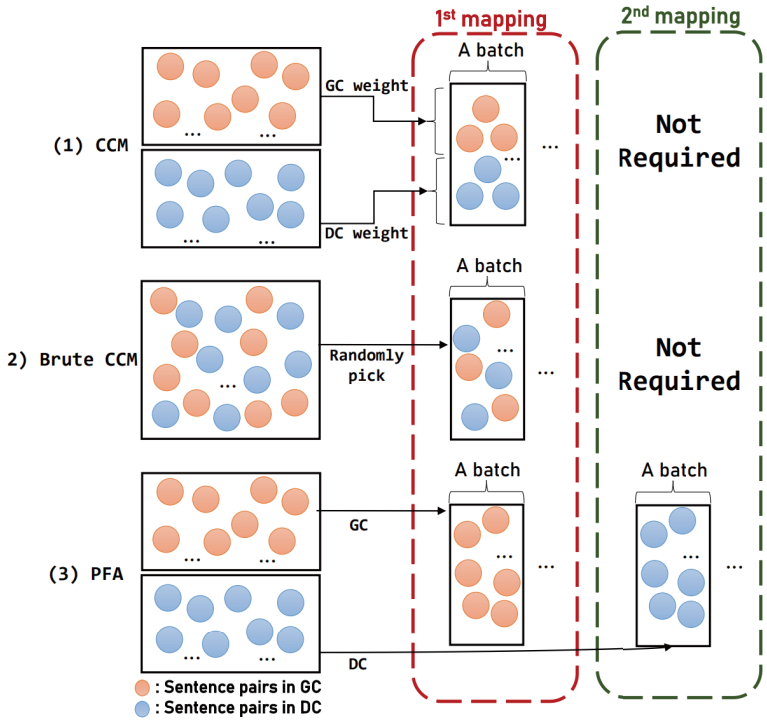
3. 기술 활용 및 응용 분야

- 다양한 도메인 특화 기계번역 분야에 활용 가능

4. 실험 (Only PDF)

4.1 실험 개요

- 본 연구는 CCM과 PFA(Pretrain-Finetuning-Approach), Brute CCM을 각각 비교해봄으로써 도메인 특화 기계 번역에서 최적의 성능을 내는 방법론을 찾는다. Brute CCM에서는 corpora에 대한 구분 없이 combined corpora를 활용하여 번역을 학습한다. 이 방법론들에 대해 성능을 비교할 뿐만 아니라 질적 분석도 수행함으로써 visible하게 방법론들에 대한 결과도 비교하였다.



4.2 실험 결과

실험결과 본 연구에서 제안하는 CCM이 가장 좋은 성능을 보임을 알 수 있었다. Brute CCM 방법과 비교했을 때, CCM은 16.13 BLEU score로 더 높은 성능을 냈다. 우리는 이를 통해 단순히 GC와 DC를 한데 합치는 것 이상으로, 추가적인 tuning이 있어야만 도메인특화 번역에서 optimal한 성능을 보일 수 있다는 것을 보였다.

CCM은 PFA(Pretrain-Finetuning-Approach)에 비해서도 1.19 BLEU score가 더 높은 모습을 보여주었다. 이는 PFA를 진행할 시 DC에 대한 학습을 진행하면서 이전에 학습된 정보를 잃기에 발생하는, catastrophic forgetting문제와 관련 지어 해석할 수 있다.

Training Method	BLEU
General Model	27.06
Google Translation	55.68
Random batch training	75.40
Incremental Trining	90.34
CCM(ours)	**91.53**

[표] Domain specialization NMT performance. General model means

Corpus weight	BLEU
1.00	**91.53**
0.50	91.26
0.33	91.30
0.25	90.85
0.20	90.58
0.10	88.87
0.03	80.79
0.02	77.91
0.01	74.21

[표] Domain specialization NMT performance by corpus weight.

자연어처리와 인공지능 (교육과정)

교육 과정 개요

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

기술의 융합을 통해 비약적인 기술 발전이 가속화되고, 공유경제, 온디멘드 경제의 기본이 되는 디지털 플랫폼 기반의 기술 및 기업들이 성장세를 이룰 것으로 예측됩니다.

모든 산업에서 인공지능 기술은 필수입니다. 특히 컴퓨터가 경험을 통해 인간처럼 스스로 학습할 수 있게 하는 기계학습(Machine Learning)은 인공지능에서 핵심적인 기술입니다. 이러한 이유로 제약, 의학, 건설, 디자인, 교육 분야 등 많은 산업 분야의 기업들이 기계학습을 도입하기를 원하고 있습니다.

기술의 내재화를 위해서는 전문 인력이 가장 중요한데, 인공지능 분야의 전문가들을 중소기업과 중견기업에서 고용하기가 쉽지 않은 것이 사실입니다. 이에 고려대학교 Human-Inspired AI 연구소에서는 "인공지능 기초 교육과정"을 개최하고자 합니다.

본 교육과정은 인공지능의 개념에서부터 기계학습의 기초이론, 딥러닝 알고리즘, 기계학습 Tool kit 학습, 그리고 실무적용을 위한 예제 실습 등으로 내실 있게 구성하였습니다.

짧은 기간이지만 본 교육 과정을 통해서 인공지능 및 기계학습의 이론과 실무기술을 학습한 수강생들이 본인이 속한 기관에서 기계학습의 지평을 열고 회사의 인공지능 기술의 내재화를 위한 교두보가 될 수 있음을 확신합니다.

많은 분들이 본 교육 과정에 참가하여 기계학습 기반의 인공지능에 대한 이해와 이를 바탕으로 현장에 적용하거나 새로운 비즈니스 창출의 기회가 될 수 있기를 기대합니다.

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)

단기 과정

<PART 1. 기계학습 기초이론>

주 제	학습목표
인공지능 개념 이해-I	∨인공지능의 개념을 학습한다. ∨신경망의 기원이 되는 퍼셉트론에 대해 학습한다.
인공지능 개념 이해-II	**[Supervised, Unsupervised learning]** ∨퍼셉트론의 한계를 극복하는 신경망의 개념을 학습한다. ∨최적의 손실 함수를 찾는 경사법을 학습한다.
인공지능 개념 이해-III	**[신경망, 딥러닝 이해]** ∨층을 깊게 쌓은 심층 신경망(딥러닝)의 특징, 풀어야할 과제, 가능성을 이해한다.
신경망 학습 원리-I	**[오차역전파 개념이해]** ∨가중치 매개변수의 기울기를 효율적으로 계산하는 오차역전파법을 학습한다.
신경망 학습 원리-II	**[신경망 학습 관련 기술의 이해-I]** ∨매개변수 갱신, 가중치의 초깃값, 배치 등의 기술을 학습한다.
신경망 학습 원리-III	**[신경망 학습 관련 기술의 이해-II]** ∨정규화, 과대적합(오버피팅), 드롭아웃, 하이퍼파라미터 최적화 등의 기술을 학습한다.

<PART 2. 자연어처리 이론 및 응용시스템>

주 제	학습목표
자연어처리의 기본	∨자연언어 처리의 개념을 이해한다.
개체명 인식(Named Entity Recognition)	∨이름을 가진 개체(Named Entity)를 인식하는 개체명 인식 기술을 학습한다.
언어모델(Language Model)	∨가장 자연스러운 단어 시퀀스를 찾아내기 위해 다음 단어 시퀀스의 확률을 할당(assign)하는 언어모델을 학습한다.
정보추출(Information Extraction)	∨비정형 텍스트로부터 유용한 정보를 자동으로 추출하는 정보 추출을 학습한다.
질의응답(Question & Answering)	∨사용자가 필요한 정보를 자연어 질문으로 입력하였을 때, 시스템이 질문에 부합하는 정보를 찾아 제시하는 기술을 학습한다.
기계번역(Machiine Translation)	∨하나의 언어로 쓰인 글을 같은 의미를 나타내는 다른 언어의 글로 변환하는 기계번역에 대해 학습한다.
대화 시스템(Dialog System)	∨자연어를 사용해 인간과 대화하는 대화시스템에 대해 학습한다.

장기 과정

<PART 1. 인공지능 개념 및 이해>

학습 내용		세부 내용
인공지능 개요	인간의 정보처리 원리를 모사한 지능형 시스템의 개념에 대해 학습한다.	인공지능이란
		인공지능의 특징
		인공지능 연구분야
기계학습 개념 및 활용	기계학습의 기본 개념과 원리를 소개하고, 종류와 활용방법을 알아본다.	기계학습이란
		기계학습의 원리
		기계학습의 종류 및 활용
기계학습 기초 알고리즘	기계학습 알고리즘 유형에 따른 기초 알고리즘 개념, 데이터 표상, 기계학습에서의 데이터에 대해 학습한다.	Concept Learning
		Decision Tree
		Linear Logistic Regression
		Neural Network
		Bayesian Learning
		Instance based learning and LR
		Genetic Algorithm
		Analytical Learning
		SVM
		HMM
		Supervised learning
		Unsupervised learning

<PART 2. 인공지능 개발 준비>

학습 내용		세부 내용
딥러닝 개발 환경	딥러닝 개발에 많이 사용되는 프로그래밍 언어 및 프레임워크를 학습한다.	Python 기초/고급
		Colab 실습환경 및 데이터 전처리
		Tensorflow tutorial
Term Project 1	본 교육과정을 통해 적용해 볼 수 있는 도메인을 선정하고, 팀을 구성하여 수행한다. 교수자의 조언을 통해 도메인 및 주제를 선정한다.	

<PART 3. 딥러닝 기초 이론>

학습 내용		세부 내용
인공신경망 개념과 원리	퍼셉트론(Perceptron)의 동작 원리와 MLP(Multi-Layered-Perceptron)에 대해 학습한다.	신경망 개념과 구조
		신경망 동작원리
		MLP 구조
		MLP 동작원리
딥러닝 개요	딥러닝에 대한 기본 개념을 하고, 딥러닝 알고리즘 유형 및 활용 방안을 소개한다.	딥러닝이란
		딥러닝 모델의 핵심
		딥러닝 시스템 구축을 위한 고려사항
		딥러닝 모델의 뼈대
		비선형 결정 경계와 활성함수
		딥러닝 모델의 학습
딥러닝 기초 실습	딥러닝 환경 설정을 바탕으로 간단한 알고리즘 구동에 초점을 맞춰 실습한다.	backpropagation, ReLU, Weight 초기화, Dropout 등
		MNIST 실습
Term Project 2	주제 및 팀 구성에 따른 기획안 발표	

<PART 4. 딥러닝 알고리즘>

학습 내용		세부 내용
CNN (Convolutional Neural Network)	CNN 알고리즘의 개념과 동작 원리를 학습하며, CNN 알고리즘을 바탕으로 영상 분류, 물체 위치 추정 및 검출 등의 시각 인식 문제에 응용하는 방법을 소개한다.	CNN이란
		CNN의 구조
		CNN 활용
RNN (Recurrent Neural Network)	RNN 알고리즘의 구성과 동작 원리를 학습하며, RNN 알고리즘을 바탕으로 언어 모델링, 자동 번역, 이미지 캡셔닝 등 응용 방법을 소개한다.	RNN이란
		RNN의 구조
		RNN 활용
CNN, RNN 실습	CNN과 RNN을 활용하여 감성분석, 언어모델 등을 실습한다.	CNN for Sentiment Analysis
		Language Model and RNN
Term Project 3	진행 과정 점검 및 애로사항 체크	

<PART 5. 언어지능 구현>

학습 내용		세부 내용
자연어처리 개요	자연어처리에 대한 기본 개념 및 자연어처리 절차에 대해 학습하고, 최신 연구 동향을 소개한다.	자연어처리란
		자연어처리의 응용 분야
		자연어처리는 왜 어려운가?
		자연어처리 연구의 패러다임
		딥러닝을 사용하는 자연어처리 연구
언어모델 (Language Model)	언어모델의 개념과 종류, 일반화에 대하여 학습하고, 언어모델 평가방법 및 퍼플렉서티에 대하여 소개한다.	언어모델이란
		통계적 언어모델
		일반화(Generalization)
		모델 평가와 퍼플렉서티(Perplexity)
질의응답 (Question&Answering)	질의응답 시스템의 과거부터 현재까지의 변화에 대하여 학습하고, 최근 딥러닝을 적용한 질의응답 시스템에 대하여 소개한다.	질의응답이란
		정보검색 기반 질의응답
		딥러닝 기반 질의응답
		딥러닝 기반 질의응답 모델
		시각 질의응답(Visual Question&Answering)
Term Project 4	진행 과정 점검 및 애로사항 체크	

<PART 6. 시각지능 구현>

학습 내용		세부 내용
시각지능의 이해	이미지/동영상 등의 데이터를 기반으로 시각 이해 지능 및 시각 분석 지능 등의 개념에 대해 학습한다.	GAN이란
		GAN의 구조
		GAN 활용
시각지능 구현	GAN(Generative Adversarial Network) 등과 같은 시각지능 구현 알고리즘에 대해 학습한다.	
Term Project 5	본 교육과정을 통해 학습한 내용을 바탕으로 팀별 산출물을 발표하고 교수자가 조언함으로써 학습 능률을 높이도록 한다.	

* 본 기초교육과정은 3주로 진행됩니다.

* 한 주에 2회씩 총 6회로 구성되어 있습니다. (1회, 3시간)

* 프로젝트가 포함되어 있으며, 마지막 수업에 프로젝트 발표가 있습니다.

* 세부 내용은 변동될 수 있습니다.

* Human-inspired AI 연구소 인공지능 기초교육과정은 앞서 진행한 수요조사 결과를 토대로 작성되었습니다.

*** 프로젝트**

- 1주차: 팀 구성 및 주제선정
- 2주차: 프로젝트 진행 및 질의응답
- 3주차: 프로젝트 결과 발표

특허 등록

특허등록

특허명	등록번호	등록일
스마트 시니어 인지반응 기반의 모델링 방법 및 장치	10-2092633	2020. 03. 18.
방송 표준을 위한 개인 맞춤형 UX/UI서비스를 제공하는 장치 및 방법	10-2014475	2019. 08. 20.
사물인터넷에 기반한 경험 공유 방법 및 장치	10-1909646	2018. 10. 12.
음식 배달 중개 방법 및 장치	10-1896408	2018. 09. 03.
집단지성을 이용한 뉴스 판단 방법 및 장치	10-1869815	2018. 06. 15.
집단감성을 이용한 맞춤형 영화 상영 방법 및 그 장치	10-1858120	2018. 05. 09.
사물인터넷 기반의 스마트 의자 및 착석자세 분석 방법, 스마트 의자 관리 장치 및 그 방법	10-1816711	2018. 01. 03.
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.
온라인 학습자를 위한 주의집중 판단 시스템 및 그 방법	10-1770817	2017. 08. 17.
인문학 정보를 자동으로 구성하는 방법	10-1760478	2017. 07. 17.
집단지성을 이용한 꿈 해몽 방법 및 장치	10-1748411	2017. 06. 12.
학습코스 자동 생성 방법 및 시스템	10-1745874	2017. 06. 05.
사용자 참여 기반의 정책 발굴 방법	10-1739925	2017. 05. 19.
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
- 지능형 패션 이미지 검색 시스템
- 한국어 개체명 인식기 및 의존 구문 분석기
- 지능형 분류기술
- 자연어-사진 크로스모달 임베딩 및 검색 기술
- 지능형 치매재활훈련기술
- 파라미터의 계수를 이용한 신경망 축소 기술
- 학습기반 질의 처리 기술
- 딥러닝을 이용하여 이미지를 검색하는 단말 장치 및 방법
- 기사 유사도 추천 및 문서 내 핵심키워드 추출 기술
- 딥러닝 기반 자동 질의응답 시스템 원천기술
- 딥러닝 기반 자동 질의응답 시스템 기술
- 딥러닝을 이용한 유사문서 검색기술
- 뇌혈류 영상이미지 인식 및 판독 결과 다이얼로그 원천기술
- 영문법 교정기 원천기술
- 생애검진 자연어 챗봇 서비스

본 책자는 과학기술정보통신부 및 정보통신기획평가원의
대학ICT연구센터지원사업(IITP-2018-0-01405)의 지원을 받아 수행된 결과임.

고려대학교

Human-Inspired AI
연구소