2018
NLP&AI Lab 연구실적

Natural Language Processing
& Artificial Intelligence

2018 연구실적

❖ 개인 연구

- 식당 예약 및 추천을 위한 한국어 대화 코퍼스 구축 ---------- 소아람
- Question Answer ---------- 좌희정
- 대화 시스템에서의 자연스러운 대화를 위한 Memory Attention기반 Breakdown Detection ---------- 이설화
- An Analysis on AI in CyberSecurity ---------- Titi
- 보고서 자동 분석 및 Topic 자동 추출 기술 연구 ---------- 허윤아
- Character-Level Feature Extraction with Densely Connected Networks ---------- 이찬희
- Structured “Image – Text” Mapping for Bidirectional “Image – Text” Transitions ---------- 김규경
- Utilizing Dual Memory Network with Domain Knowledge for Goal-Oriented Dialogues ---------- 이동엽
- 개체명 인식(Named Entity Recognition) 시스템 연구 및 개발 ---------- 김경민
- Enhanced Sequential Representation Augmented with Utterance-level Attention for Response Selection -- 황태선
- Answer Generation from Span Prediction in MRC Tasks ---------- 양기수

2018 연구실적 [과제]

❖ 스마트 시니어세대의 문화향유를 위한 인지반응 맞춤형 UI/UX기술 개발

- 인지반응 분석 시스템 개발 ---------- 이설화
- 스마트 시니어 맞춤형 프로파일 관리시스템 고도화 ---------- 허윤아
- 스마트 시니어 맞춤형 추천 시스템 개발 ---------- 황태선
- 표준안 및 기술 보고서 작성 ---------- 소아람

❖ 사용자 중심의 지능형 패션검색 및 맞춤형 코디네이션 제품 개발

- 벡터기반 사용자 선호 패션 프로파일링 · 코디 및 매칭 기술 개발 ---------- 이설화
- 드로잉 자질 추출 기술 ---------- 이찬희
- 딥러닝 기반의 패션 상품 유사 이미지 검색 모델 ---------- 이동엽

❖ 전통문화 융복합 지원을 위한 지능형 검색 플랫폼 구축

- Meta Learning based Global Relation Extraction ---------- 김규경
- 고유명사 개체인식 기술 개발 추가 연구 – 데이터 구축 ---------- 김경민

2018 연구실적 [과제]

❖ 머신러닝을 이용한 비정형 위협정도 자동추출 방안 연구 (KISA)

- 비정형 위협정보 자동 인식 및 추출 ---------- 이찬희
- 머신러닝을 이용한 문서 자동 요약 ---------- 이찬희
- 유사 문서 검색 및 시각화 ---------- 이찬희
- Keyword Extraction ---------- 허윤아
- 위협정보 개체명 인식 데이터 구축 ---------- 김경민
- OCR 프로그램 조사 및 성능 비교 ---------- 양기수

❖ 기계학습을 활용한 자연어 질의/응답 인공지능 모델 개발 (현대자동차)

- 기술 동향 보고서 및 기술 개발 전략 보고서 ---------- 김규경
- 자동차 법률 응답 대화 시스템 개발 ---------- 이동엽

식당 예약 및 추천을 위한 한국어 대화 코퍼스 구축

고려대학교 NLP&AI Lab.

소아람

연구 내용 - 대화 코퍼스 구축

❖ **대화 데이터 수집**

- 대화 데이터는 대화 상대에 따라 인간-기계 대화 데이터와 인간-인간 대화 데이터로 구분함
- 본 연구에서는 대화 코퍼스 수집을 위하여 인간-인간의 대화를 수집함
  - 시스템과 사용자 역할로 나누어 각 역할에 알맞게 대화를 나누어 대화를 수집함

연구 내용 - 대화 코퍼스 구축

❖ 대화 수집 가이드라인

- 시스템
  - 사용자가 원하는 식당을 알아내기 위하여 고정된 형식의 질문을 함

    예) 어느 종류의 식당을 원하시나요?

    원하시는 가격대를 말씀해 주세요.
  - 사용자가 원하는 식당 추천
  - 사용자가 선택한 식당에 대한 정보 제공 및 예약을 위한 대화 진행

연구 내용 - 대화 코퍼스 구축

❖ 대화 수집 가이드라인

- 사용자
  - 원하는 식당에 대한 정보를 얻기 위하여 시스템에 정보를 제공함
  - 다양한 발화를 수집할 수 있도록 최대한 다양한 형식으로 변화를 주어 시스템과 대화

    예) 5명, 다섯명, 오인, 5

연구 내용 - 대화 코퍼스 구축

❖ 수집 데이터

수집정보	참 고
위치	이태원 근처, 강남역 주변, 정자역, 판교 등
음식종류	한식, 양식, 중식, 일식, 이탈리안, 베트남, 기타
식당	실제 식당이름
날짜	월, 일, 오늘, 내일, 0요일
시간	00시 ~24시, 저녁, 오후, 점심
인원	0명
가격	1인당 15,000원, 5만원, 4명 10만원
예약도구	앱(app)으로 예약, 전화로 예약
전화번호	02-1234-5678, 031-777-8965
주소	서울시 서초구 서초동 143
주차장 유무	주차 가능 / 불가능
주차 서비스	가능 / 불가능

연구 내용 - 대화 코퍼스 구축

❖ 대화 시나리오

시스템: 어느 지역을 원하시나요?

사용자: 이태원 근처로 추천해줘

시스템: 어느 종류의 식당을 원하시나요?

사용자: 이탈리안 레스토랑이 좋을 것 같아

시스템: 원하시는 가격대를 말씀해 주세요.

사용자: 1인당 15000원~20000원

시스템: 언제 방문할 예정인가요?

사용자: 오늘 저녁 7시쯤 가려고

시스템: 이태원 근처의 당일 예약 가능한 식당은 이탈리안 레스토랑은 비스테까, 피제리아까사, 오레노 입니다.

사용자: 또 다른 식당은?

시스템: 하베스트, 레뒤플라, 라쿠치나, 라코르테 입니다.

사용자: 오레노 예약해줘

시스템: 몇 명인가요?

사용자: 음... 4명 예약해줘

시스템: 오늘 저녁 7시 오레노에 4명 예약하였습니다.

사용자: 또 다른 식당은?

시스템: 하베스트, 레뒤플라, 라쿠치나, 라코르테 입니다.

연구 내용 - 대화 코퍼스 구축

❖ 태깅(tagging)

사용자: [[오늘||날짜]] [[저녁||시간]] [[이태원||위치]]근처 식당 추천해줘
시스템: 어느 종류의 식당을 원하시나요?
사용자: [[이탈리안||종류]] 레스토랑
시스템: 원하시는 가격대를 말씀해 주세요.
사용자: [[1인당 15000원~20000원||가격]]
시스템: 언제 방문할 예정인가요?
사용자: [[오늘||날짜]] [[저녁 7시||시간]]
시스템: [[이태원||위치]] 근처의 예약 가능한 [[이탈리안||종류]] 레스토랑은 [[비스테까||식당]], [[피제리아까사||식당]], [[오레노||식당]]입니다.
사용자: 또 다른 식당은?
시스템: [[하베스트||식당]], [[레뒤플라||식당]], [[라쿠치나||식당]], [[라코르테||식당]]입니다.
사용자: [[오레노||식당]] 예약해줘
시스템: 몇 명인가요?
사용자: 음... [[4명||인원]] 예약해줘
시스템: [[오늘||날짜]] [[저녁 7시||시간]] [[오레노||식당]]에 [[4명||인원]] 예약하였습니다.
사용자: [[오레노||식당]] 전화번호 알려줘
시스템: [[02-794-0055||전화번호]]입니다.

연구 내용 - 대화 코퍼스 구축

❖ 데이터 통계

차시	Dialogue	slot											
		Slot수 (필수/추가)	위치	음식종류	식당	날짜	시간	인원	가격	예약도구	전화번호	주소	주차장 유무	주차 서비스
1차	192	7	561	374	686	531	530	483	201					
2차	306	12(6/6)	725	645	2242	720	721	581	429	40	244	268	117	21
3차	1390	12(6/6)	1387	2779	1339	1383	1383	1383	1392	1383	1387	1391	1384	1251
4차	210	12(6/6)												
5차	수집 중	12(6/6)												

- 총 2098개의 Dialogue를 수집함
- 사용자 발화 : 24262
- 시스템 발화 : 25757

Question Answer

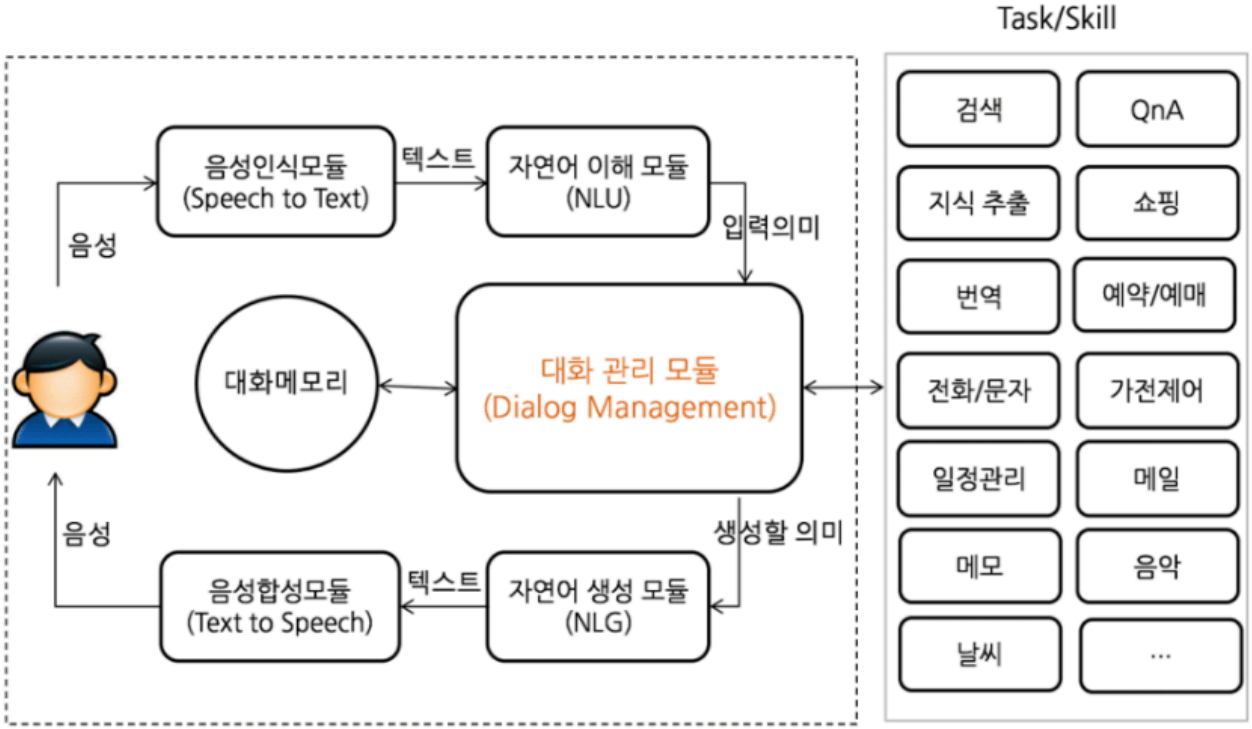
고려대학교 NLP&AI Lab.

좌희정

연구 계획

❖ 연구 내용 : task 지정을 위한 사용자 의도 파악

- Question & Answer 과정에서 사용자의 질문이 어떤 의도인지를 파악하기 위해 slot filling, slot tagging 을 통한 정확한 도메인 찾기



연구 계획

❖ 관련 논문

- Sequential Short-Text Classification with Recurrent and Convolutional Neural Networks
- A joint Model of Intent Determination and Slot Filling for Spoken Language Understanding
- ONENET: Joint Domain, Intent, Slot prediction for Spoken Language Understanding
- Attention-Based Recurrent Neural Network Models for Joint Intent Detection and Slot Filling

대화 시스템 에서의 자연스러운 대화를 위한 Memory Attention 기반 Breakdown Detection

고려대학교 NLP&AI Lab.

이설화

연구 내용 - Introduction

❖ **대화시스템에서의 이슈**

- 시스템과 사람 간의 자연스러운 대화

연구 내용 - Dialogue Breakdown

- 사람과 시스템 사이의 대화에서 문맥의 흐름이 끊겨 사람이 더 이상 대화 흐름에 참여할 수 없게 되는 상황
- Dialogue Breakdown Detection (Walker et al., 2002)
  - 시스템의 발화가 Dialogue Breakdown을 유발하는지 탐색하는 것 (Martinovsky et al., 2006)

연구 내용 - Dialogue Breakdown

System

"오늘은 축구하기 좋은 날씨네요."

User

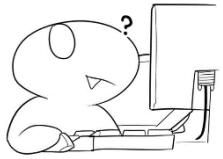
"맞아, 그런데 곧 비가 온대."

"일기예보에서 저녁에 비가 온다고 했어요."

"나는 비가 싫어서 저녁에 집에 있을 거야."

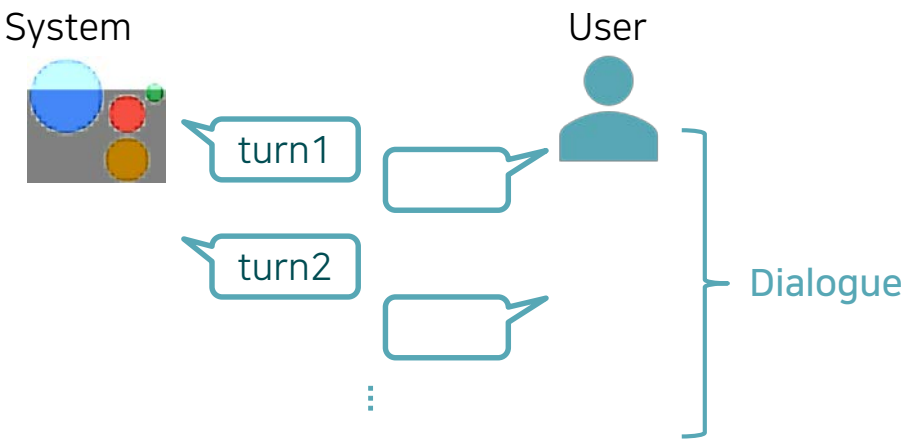
"그러면 저녁에 나가서 축구하면 되겠네요."

"..."



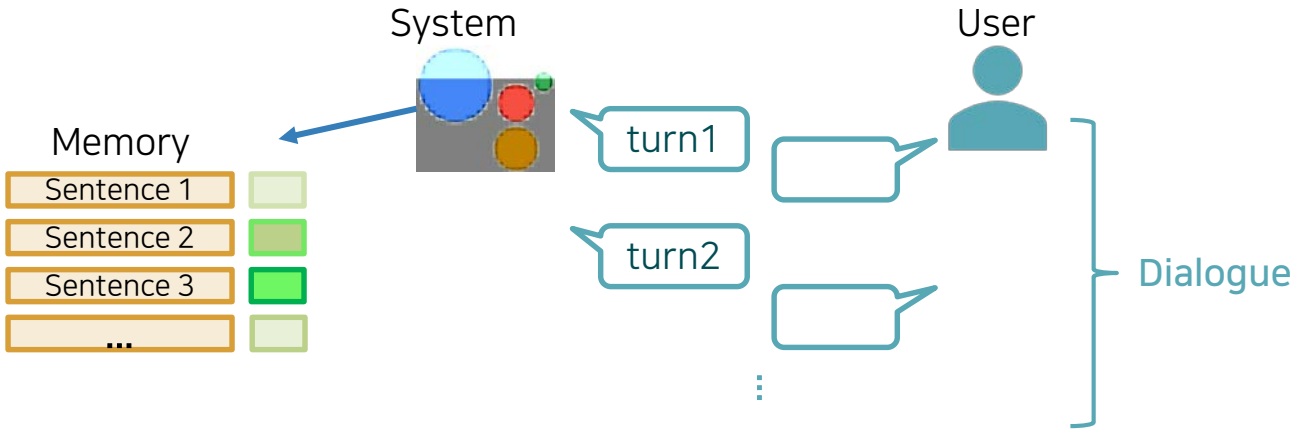
연구 내용 - Motivation (1/2)

- 대화 시스템내에 Breakdown이 발생하는 이유
  - 사람과 시스템이 발화를 여러 번 주고 받는 multi-turn으로 인해 시스템이 사용자의 의도를 잘못 이해
  - 시스템이 대화의 모든 흐름을 파악하기 어려움



연구 내용 - 연구 내용 – Motivation (2/2)

- Breakdown detection을 위한 기존 방법들은 hand-crafted feature들을 이용하거나 단순한 구조의 end-to-end방식을 사용함
  - 대화 시스템 고유의 특성을 반영하지 못함
  - 시스템이 이전 발화들을 기억하게 하면서 그 중 중요한 발화들에 중점을 둬보자

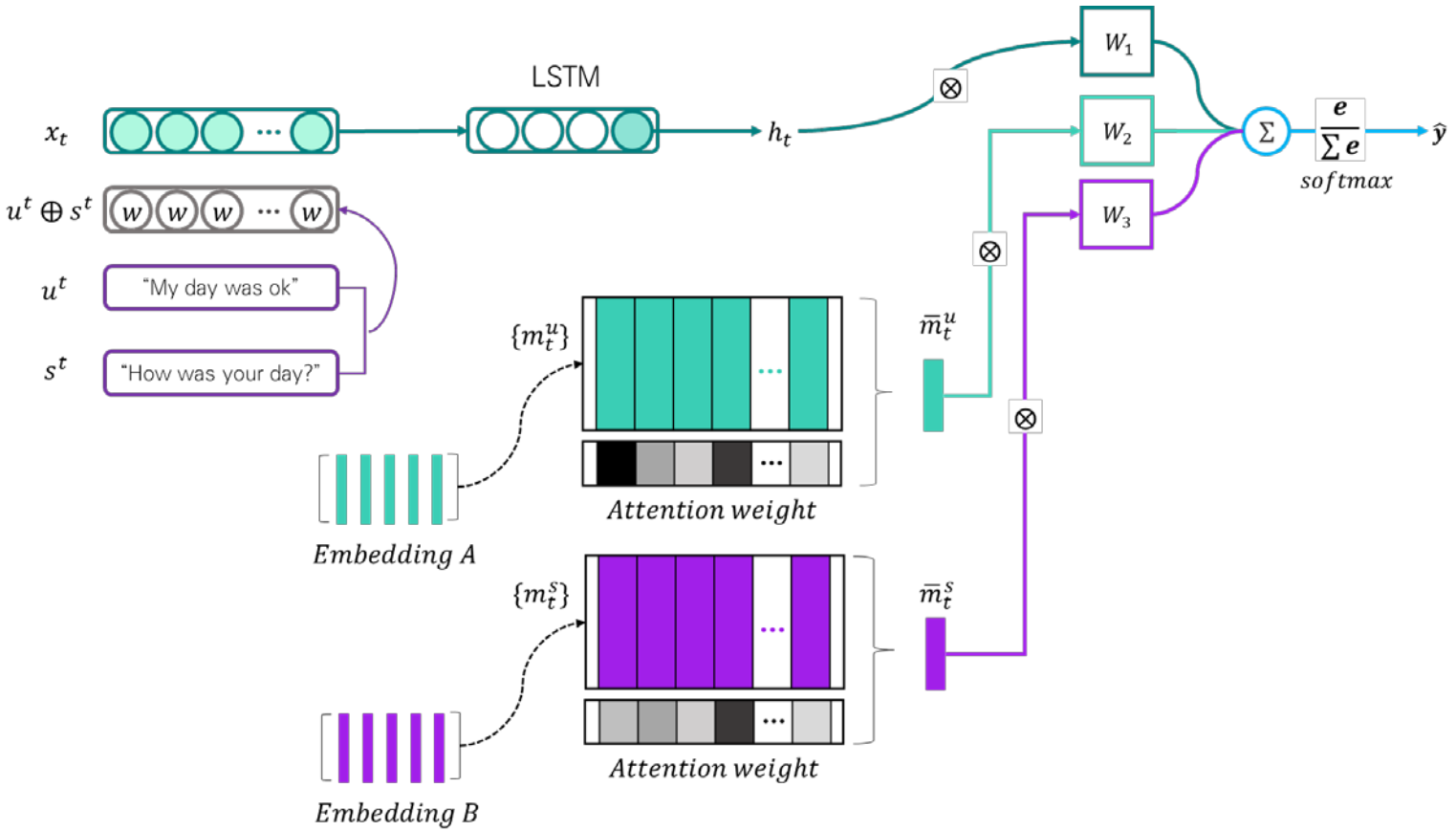


연구 내용 - Proposed model

- Temporal Utterance Encoding model
  - End-to-End
  - Memory Attention (memory network based)

- 제안한 모델은 LSTM을 이용하여 대화내에 사용자와 시스템의 발화를 인코딩하고 시스템 발화에 대해 memory network기반의 attention을 적용함

연구 내용 - Proposed model



s^t: *system utterance at time step t*
u^t: *user utterance at time step t*
$\{m_t^s\}$: *system memory vectors*
$\{m_t^u\}$: *user memory vectors*

연구 내용 - Dataset

- DSTC6(Dialog system technology challenges)
- DBDC3 dataset for training
  - IRIS-100, TKTK-100, CIC-115, YI-100 (English)
  - Dialogue 415개
  - 20 또는 21개의 발화로 구성된 dialogue
  - 각 system utterance에 대해 사람이 세 개의 label (Non breakdown, possibly breakdown, Breakdown) 중 태깅하여 voting
- Evaluation dataset
  - IRIS-50, TKTK-50, CIC-50, YI-50

연구 내용 - Evaluation

- Baseline – CRF(Conditional random field)
- Evaluation metrics
  - Classification-related: Accuracy, precision, recall, f1
  - Distribution-related: JS divergence, Mean squared error
- 비교 모델
  - 제안한 모델 – TU(no memory net), TU+S(system memory), TU+U(user memory), TU+S+U(sys+user memory)
  - KTH – LSTM, SVM
  - PLECO – Memory network
  - RSL17BD - ExtraTreesRegressor

연구 내용 - Experiment results(1/3)

- Classification-based
  - TU+S+U에서 기존 모델보다 뛰어난 성능을 보임

		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

연구 내용 - Experiment results(2/3)

- Distribution-based
  - 기존 모델의 성능이 더 뛰어났음

		Model								
		Proposed model				CRF Baseline	Majority Baseline	KTH run2	PLECO run1	RSL17BD run2
		TU	TU+S	TU+U	TU+S+U					
JSD	JSD(NB,PB,B)	0.0775	0.0749	0.075	0.0976	0.4409	**0.0393**	0.0481	0.0714	0.0412
	JSD(NB,PB+B)	0.056	0.0483	0.0478	0.0536	0.2687	**0.0237**	0.0267	0.0427	0.0256
	JSD(NB+PB,B)	0.0375	0.0388	0.0367	0.06	0.2985	0.0257	0.0262	0.0535	**0.0225**
MSE	MSE(NB,PB,B)	0.0452	0.0435	0.0432	0.058	0.2185	**0.0224**	0.0281	0.0415	0.0241
	MSE(NB,PB+B)	0.0651	0.0567	0.0561	0.0622	0.2171	**0.0278**	0.0315	0.0455	0.0301
	MSE(NB+PB,B)	0.0413	0.0436	0.0413	0.0707	0.2578	0.0264	0.0286	0.0632	**0.0246**

연구 내용 - Experiment results(3/3)

- Qualitative evaluation
  - Memory attention distribution(TU+S+U)을 이용한 정성적 평가

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

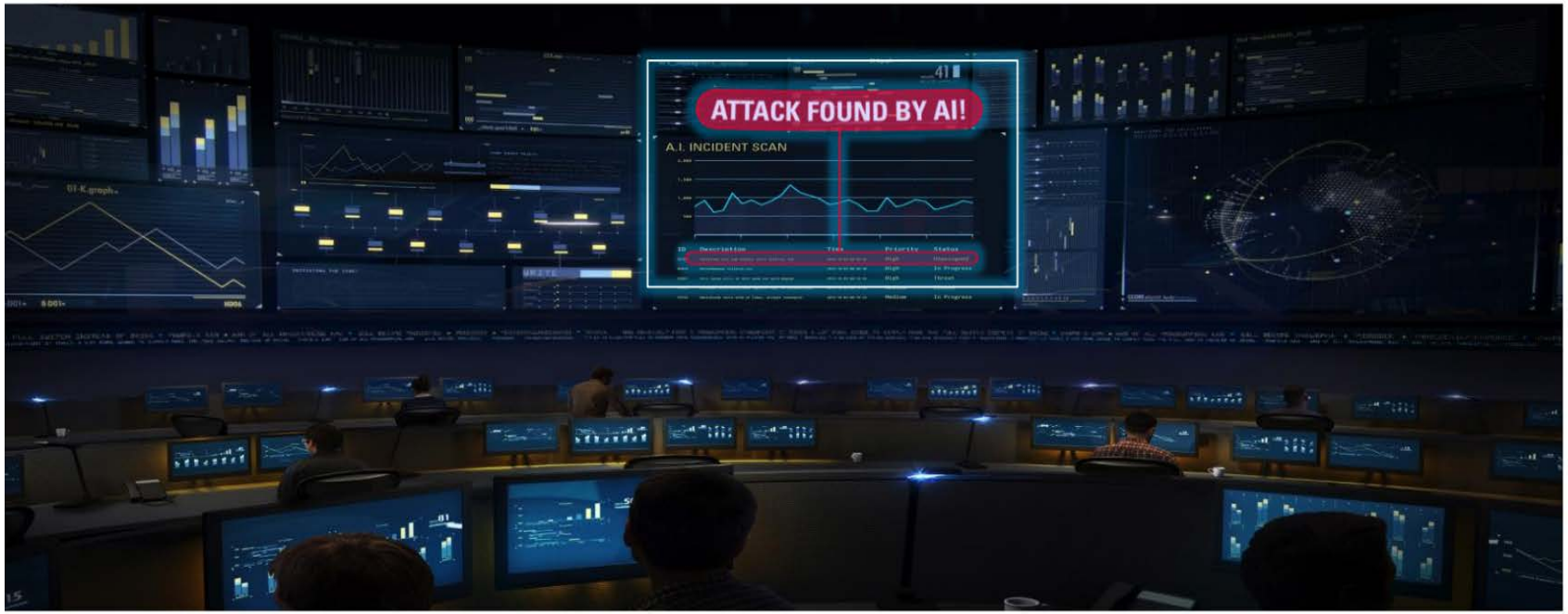
Table 3: Example predictions of user and system memory attention distribution on DBDC3 task.

An Analysis on AI in CyberSecurity

고려대학교 NLP&AI Lab.

Aiyanyo Imatitikua Danielle

연구 내용 - The Ultimate goal



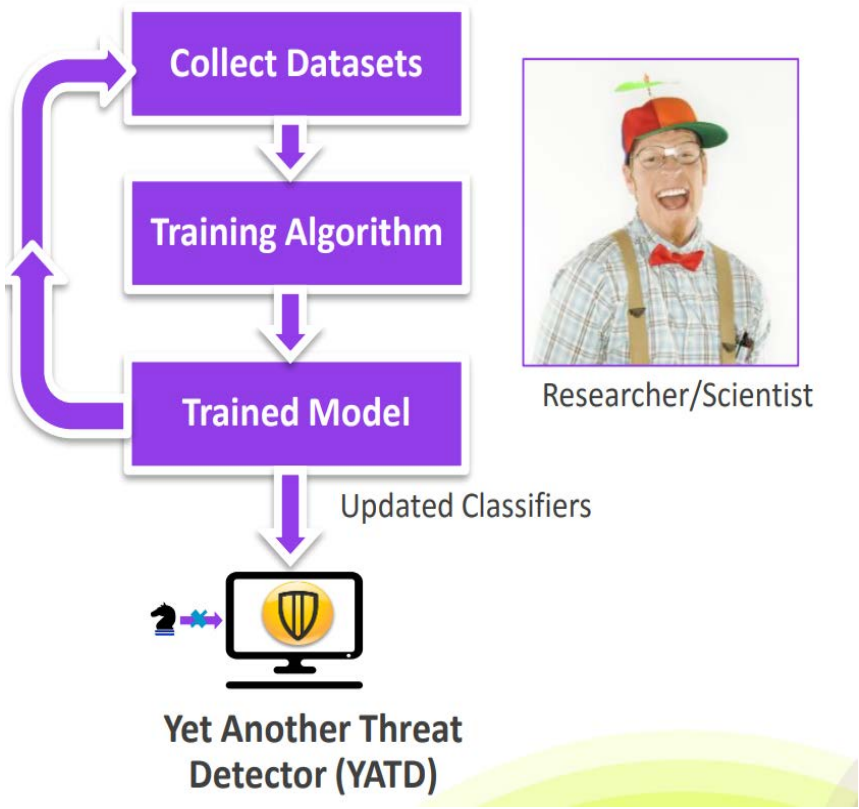
Predictive Protection

연구 내용 – The Ultimate goal

- A system that anticipates attacks and automatically reconfigures for protection.
- Real-time conversation monitoring for
  - Social engineering
  - Cyberbullying
  - Fake news etc
- Threat detection systems that learn to learn

연구 내용 – Powered by Learning

- ❖ Yet Another Threat Detector
  - Data with labels
  - Debate about techniques
  - Rely on data scientists
    - Feature engineering
    - Updates & tweaks
- ❖ Primarily by automation



연구 내용 - The Actors

Abundant Data

- 9 Trillion rows of security data
- 4.5B queries processed daily from 175M endpoint devices
- 2B emails scanned daily
- 1B previously unseen web requests scanned daily
- Outputs from other systems and products

Leading Experts

- Dedicated organization of experts
- World renowned attack investigation team
- Centuries of combined AI experience.

Features

- Static attributes
- Dynamic behaviors
- Reputation
- Relations
- Sequential state

Techniques

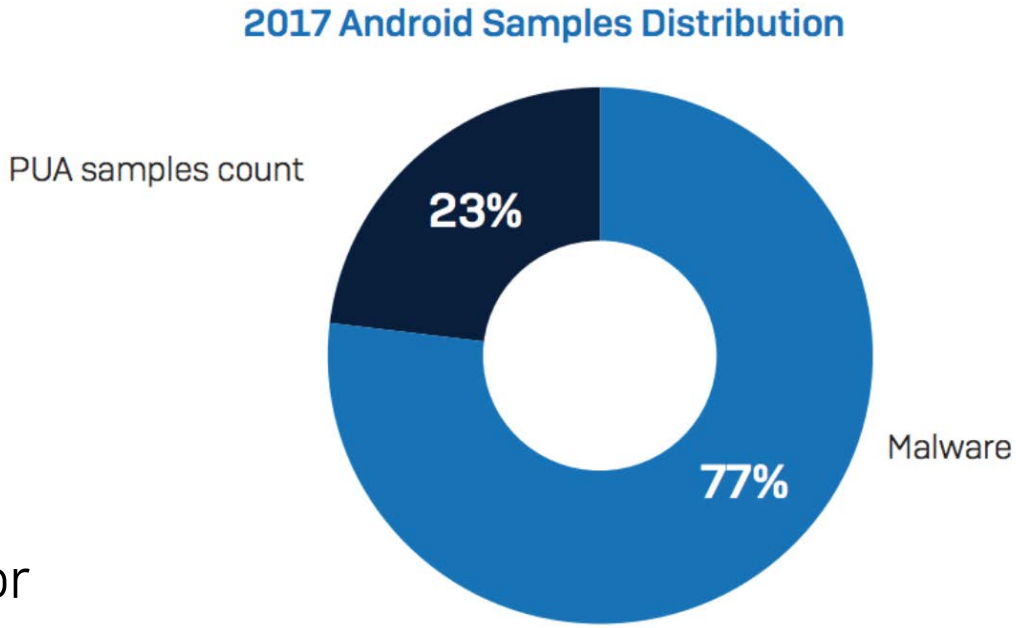
- Ensembling
- Boosting
- Sequential Learning
- Deep Learning
- Automation

연구 내용 - The Tools

- Supervised or Unsupervised Learning
  - In the field of malware classification many of the techniques used to implement supervised or unsupervised Machine Learning offer significant benefits in terms of accurate and fast classification of malware with very low False Positive rates, but it does require a large training base, extensive data expertise and resources.
  - Supervised: Malware classification, Spam identification, MLSec project on firewall data
  - Unsupervised: DNS analytics, Threat intelligence, Tier 1 analyst automation, User and Entity Behavior Analytics.
- Deep Learning

연구 내용 - The Challenges

- Ransomware Evolution
- AI Expansion
- IOT threats
- Blockchain
- Serverless Apps
- Network traffic data
- Real time scenarios
- Adaptive attacking behavior
- Anomaly detection for cloud servers
- Testing deep learning based intrusion detection



연구 내용 - Future Direction

- Reducing training and detection time
- Real world scenarios
- Outlier detection algorithms to take into consideration the evolving nature of threat
- Integrate file specific behavioral detection into endpoint anti-virus systems

보고서 자동 분석 및 Topic 자동 추출 기술 연구

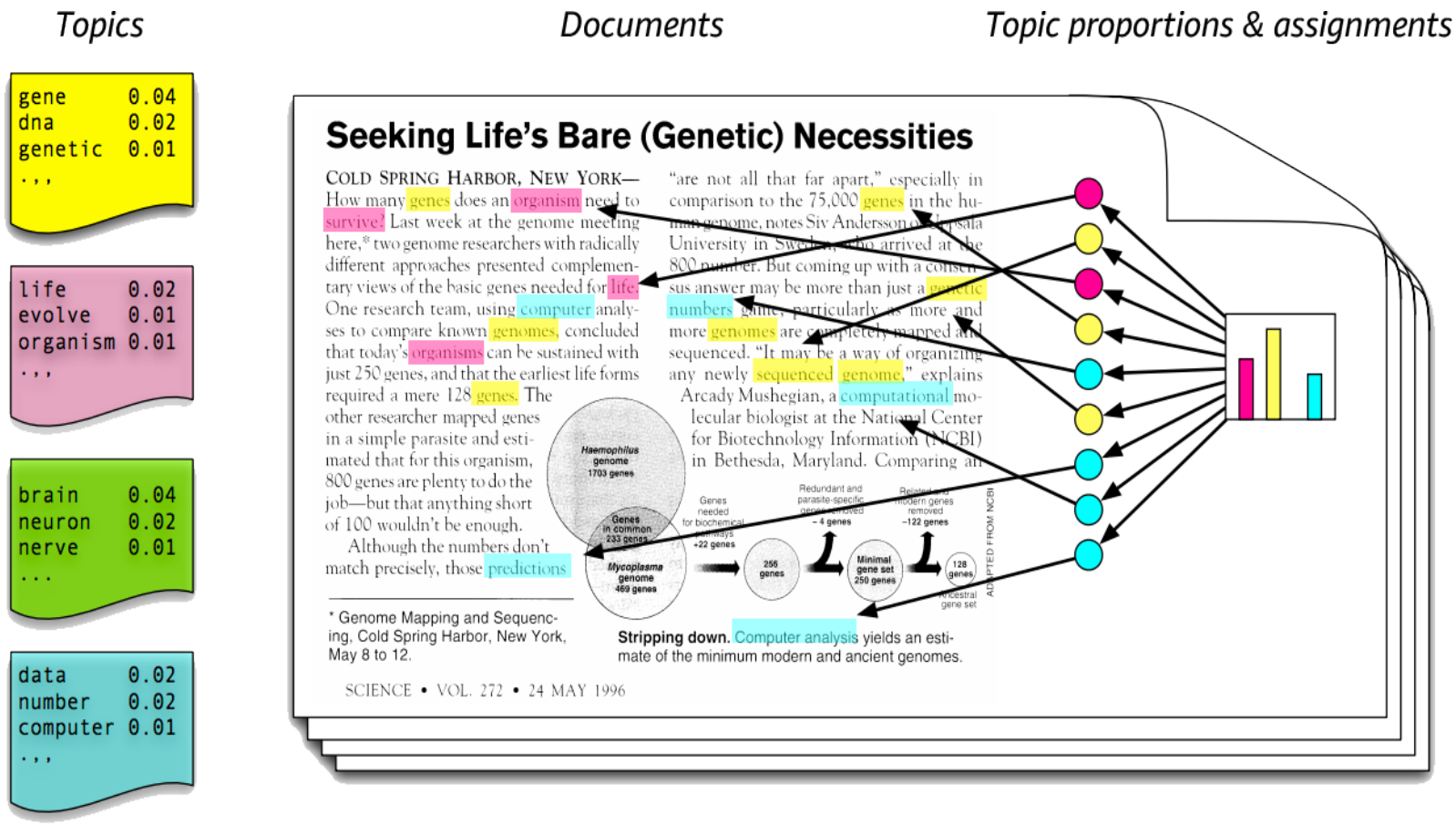
고려대학교 NLP&AI Lab.

허윤아

연구 목표 - 보고서 Topic 자동 추출

- Topic Modeling이란 문서 집합의 추상적인 “주제”를 발견하기 위한 통계적 모델링임
- 정보화 시대가 도래하면서 매일 생성되는 텍스트는 인간이 직접 처리할 수 있는 양을 크게 넘어서는데, 토픽 모델은 자동적으로 비정형 텍스트의 집합을 이해하기 쉽도록 조직하고 정리하는 데에 쓰일 수 있음
- 실제로 문헌 내에 어떤 주제가 들어있고, 주제 간의 비중이 어떤지는 문헌 집합 내의 단어 통계를 수학적으로 분석함으로써 알아낼 수 있음
- 또한, 토픽 모델은 원래 개발된 목적인 텍스트 마이닝 분야 이외에도 유전자 정보, 이미지, 네트워크와 같은 자료에서 유의미한 구조를 발견하는 데에도 유용하게 사용되고 있음

연구 내용 - 보고서 Topic 자동 추출



연구 내용 - Latent Dirichlet Allocation(LDA)

- LDA 란?
  - LDA(Latent Dirichlet Allocation) 주어진 문서에 대해서 각 문서는 어떤 주제들이 존재하는지 확률적으로 추출하는 Topic Modeling 기법 중 하나임

[LDA 모델 예시]

- Document 1: I had a peanut butter sandwich for breakfast.
- Document 2: I like to eat almonds, peanuts and walnuts.
- Document 3: My neighbor got a little dog yesterday.
- Document 4: Cats and dogs are mortal enemies.
- Document 5: You mustn't feed peanuts to your dog.

LDA 모델을 통해 문서에서 나타나는 여러 주제를 발견함

[LDA 결과]

- Topic 1: 30% peanuts, 15% almonds, 10% breakfast… (-> food 라고 해석)
- Topic 2: 20% dogs, 10% cats, 5% peanuts… (->동물이라고 해석)

연구 내용 - Latent Dirichlet Allocation(LDA)

- LDA 란?
  - LDA(Latent Dirichlet Allocation) 주어진 문서에 대해서 각 문서는 어떤 주제들이 존재하는지 확률적으로 추출하는 Topic Modeling 기법 중 하나임

[LDA 모델 예시]

- Document 1: I had a peanut butter sandwich for breakfast.
- Document 2: I like to eat almonds, peanuts and walnuts.
- Document 3: My neighbor got a little dog yesterday.
- Document 4: Cats and dogs are mortal enemies.
- Document 5: You mustn't feed peanuts to your dog.

[Topic 결과]

- Document 1 and 2 : 100% Topic 1
- Document 3 and 4 :100% Topic 2
- Document 5 : 70% Topic 1, 30% Topic 2

LDA 모델을 통해 문서에서 나타나는 여러 주제를 발견함

[LDA 결과]

- Topic 1: 30% peanuts, 15% almonds, 10% breakfast… (-> food 라고 해석)
- Topic 2: 20% dogs, 10% cats, 5% peanuts… (->동물이라고 해석)

연구 내용 - Topic Modeling

❖ Procedure

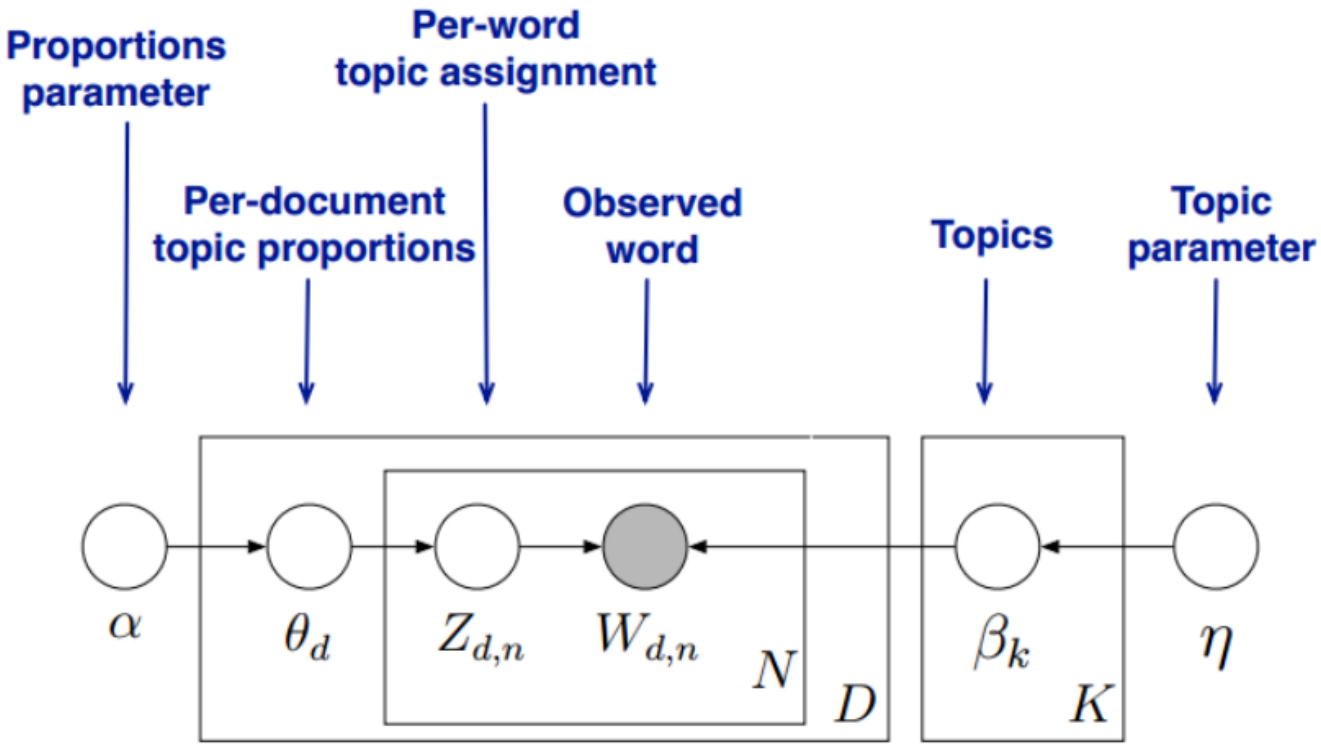
- 전체 문서에 대한 전처리와 Topic 이 될 수 있는 단어 생성을 위해 전처리를 진행함
  - am / is / are / of / a / the / but / ...와 같은 단어는 " 주제"에 대한 정보가 없기 때문에 제거함
- 각 문서 내에서 중요하고 효과적인 단어를 구성하여 단어 집합을 생성함
- 실제 문서(d) 에서 각 단어 w를 통해 연산함
  - p(topic t | document d): 문서에서 topic(word)이 나온 비율(확률)

	W1	W2	W3	Wn
D1	0	2	1	3
D2	1	4	0	0
D3	0	2	3	1
Dn	1	1	3	0

연구 내용 - Topic Modeling

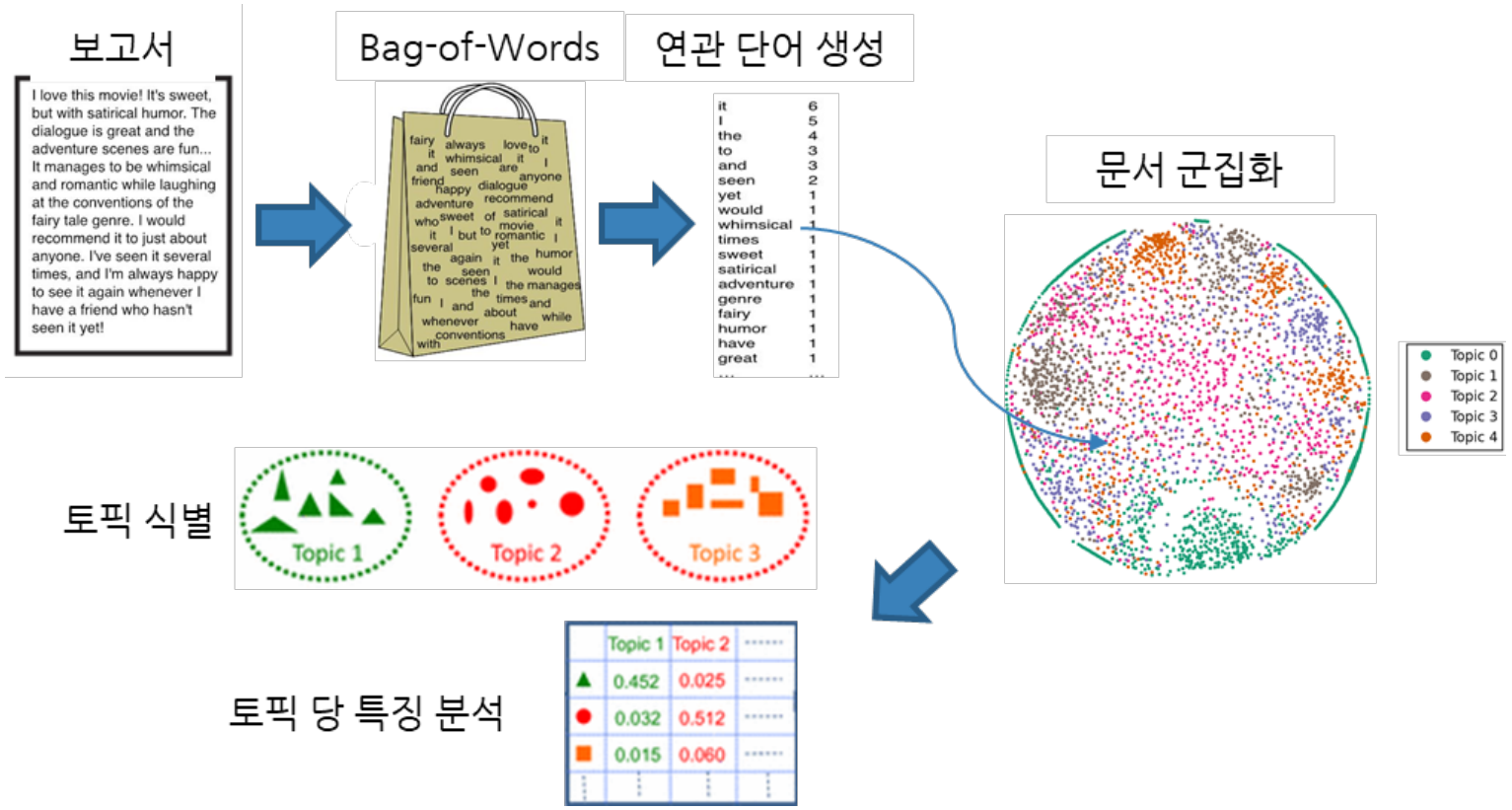
❖ Topic Model Definition

- 목표 : 전체 문서에서 Topic을 추측 또는 추정하는 것



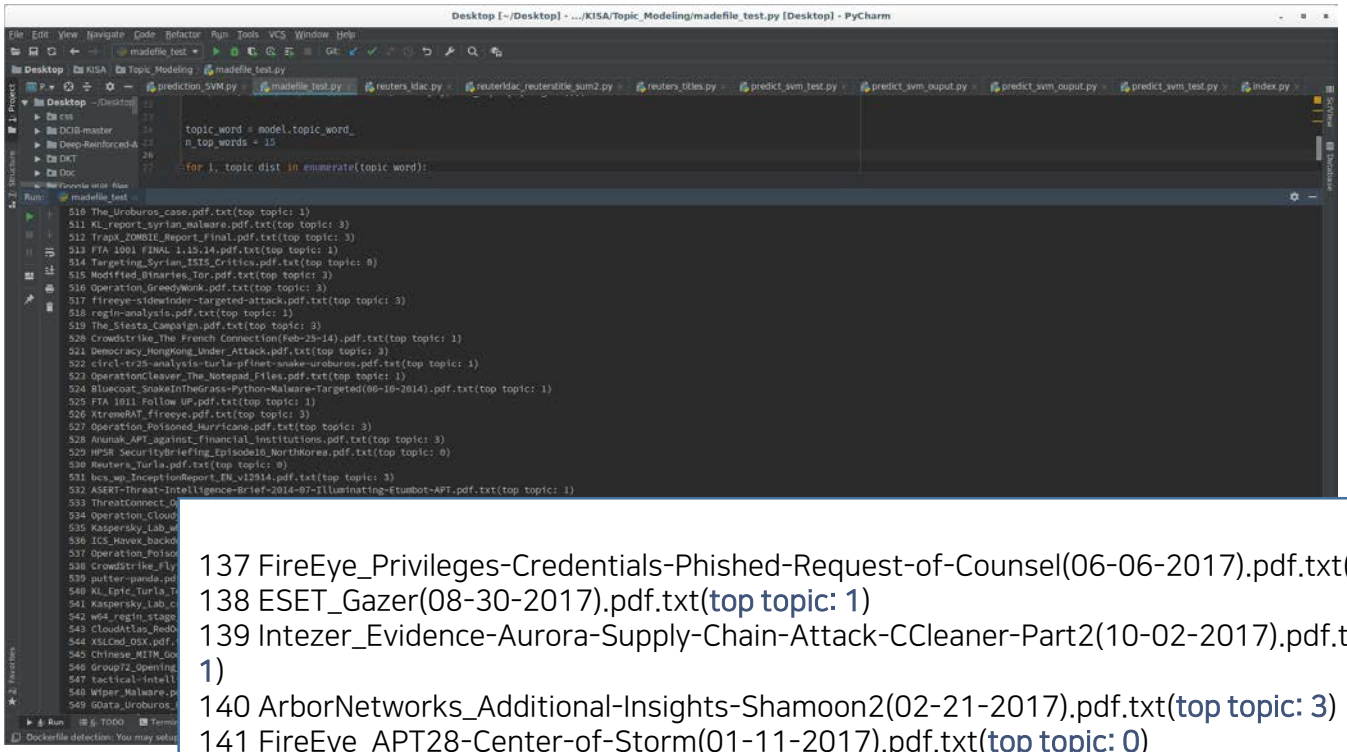
연구 내용 - Topic Modeling

❖ 문서 토픽 자동 추출 시스템 구조도



연구 결과 - Topic Modeling

❖ 각 인텔리전스 보고서 마다 토픽 자동 추출



137 FireEye_Privileges-Credentials-Phished-Request-of-Counsel(06-06-2017).pdf.txt(top topic: 4)
138 ESET_Gazer(08-30-2017).pdf.txt(top topic: 1)
139 Intezer_Evidence-Aurora-Supply-Chain-Attack-CCleaner-Part2(10-02-2017).pdf.txt(top topic: 1)
140 ArborNetworks_Additional-Insights-Shamoon2(02-21-2017).pdf.txt(top topic: 3)
141 FireEye_APT28-Center-of-Storm(01-11-2017).pdf.txt(top topic: 0)

Character-Level Feature Extraction with Densely Connected Networks

고려대학교 NLP&AI Lab.

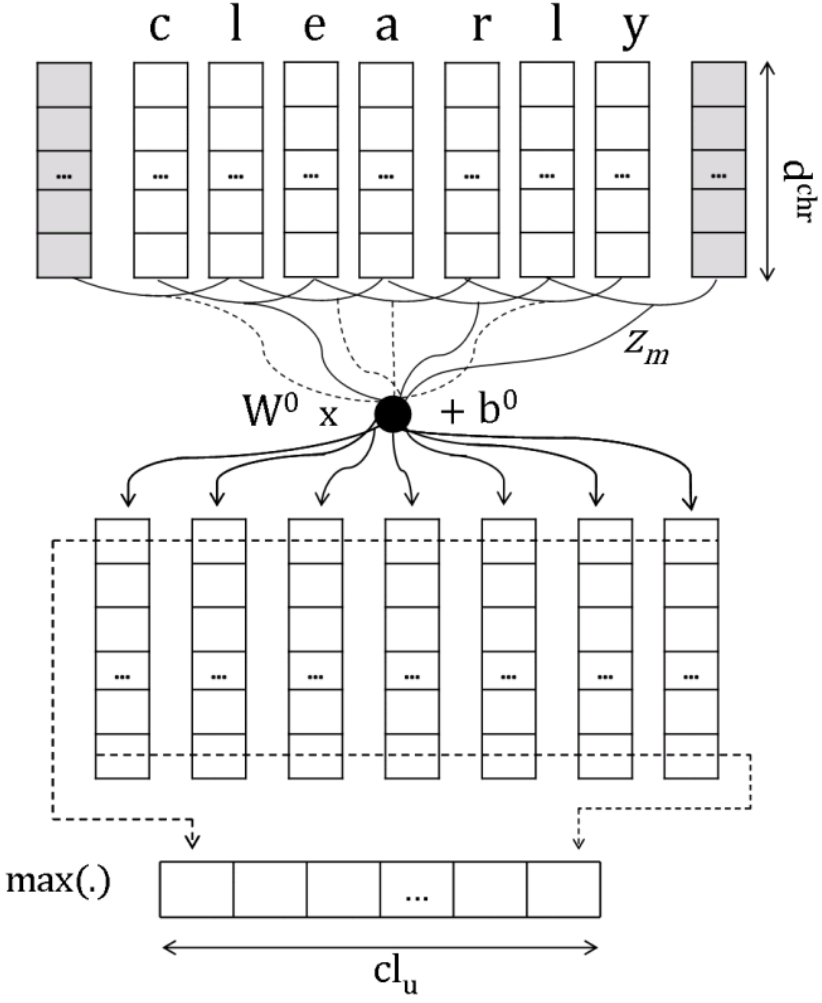
이찬희

Introduction | Background

- 효과적인 문자 단위 feature를 만들어내는 것은 많은 자연어처리 task에서 매우 중요함
  - Named Entity Recognition (NER), Part-of-Speech (POS) Tagging, Slot tagging 등
- State-of-the-art method들은 어떤 형태로든 문자 단위 feature를 활용함
  - Hand-crafted feature
  - Convolutional Neural Network (CNN) 활용
  - Recurrent Neural Network (RNN) 활용
- 이 중 CNN이나 RNN과 같은 Neural Network을 활용하는 방법은 전문가 지식이나 시간 투자의 필요성이 낮으면서도 탁월한 성능을 보여 각광 받음

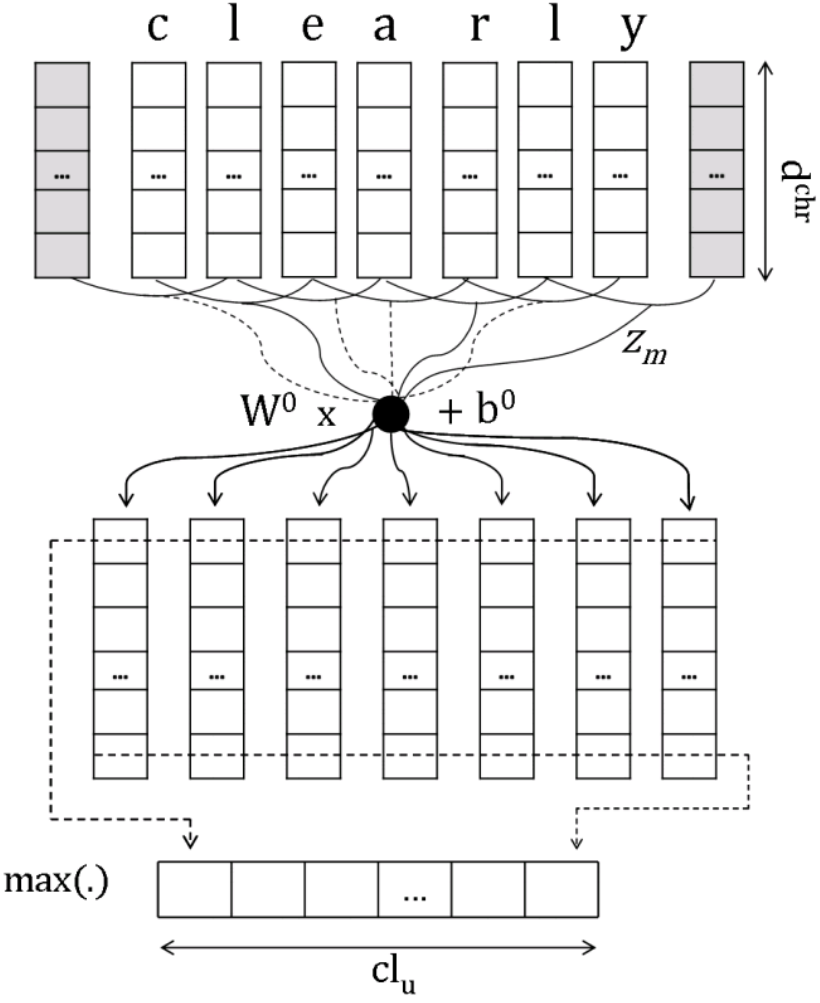
Introduction | CNN

- Convolutional Neural Network (CNN)을 이용하여 단어를 구성하는 문자로부터 feature vector 생성
  - Word embedding과 같은 방법으로 각 문자를 character embedding로 변환
  - 단어를 구성하는 문자들의 character embedding에 convolution 적용 후, pooling을 사용하여 단어 벡터 생성
- 단점 - 단어 내 n-gram의 순서와 단어 벡터가 독립적이므로 위치 정보가 손실됨



Introduction | RNN

- Recurrent Neural Network (RNN)을 이용하여 단어를 구성하는 문자로부터 feature vector 생성
  - 단어를 구성하는 문자들의 character embedding에 RNN을 적용하여 단어 벡터 생성
  - Word embedding을 보조하는 역할로 사용, 혹은 word embedding 없이 character RNN만을 사용
- 단점 – 문자를 순차적으로 처리해야 하므로 계산 속도가 느림



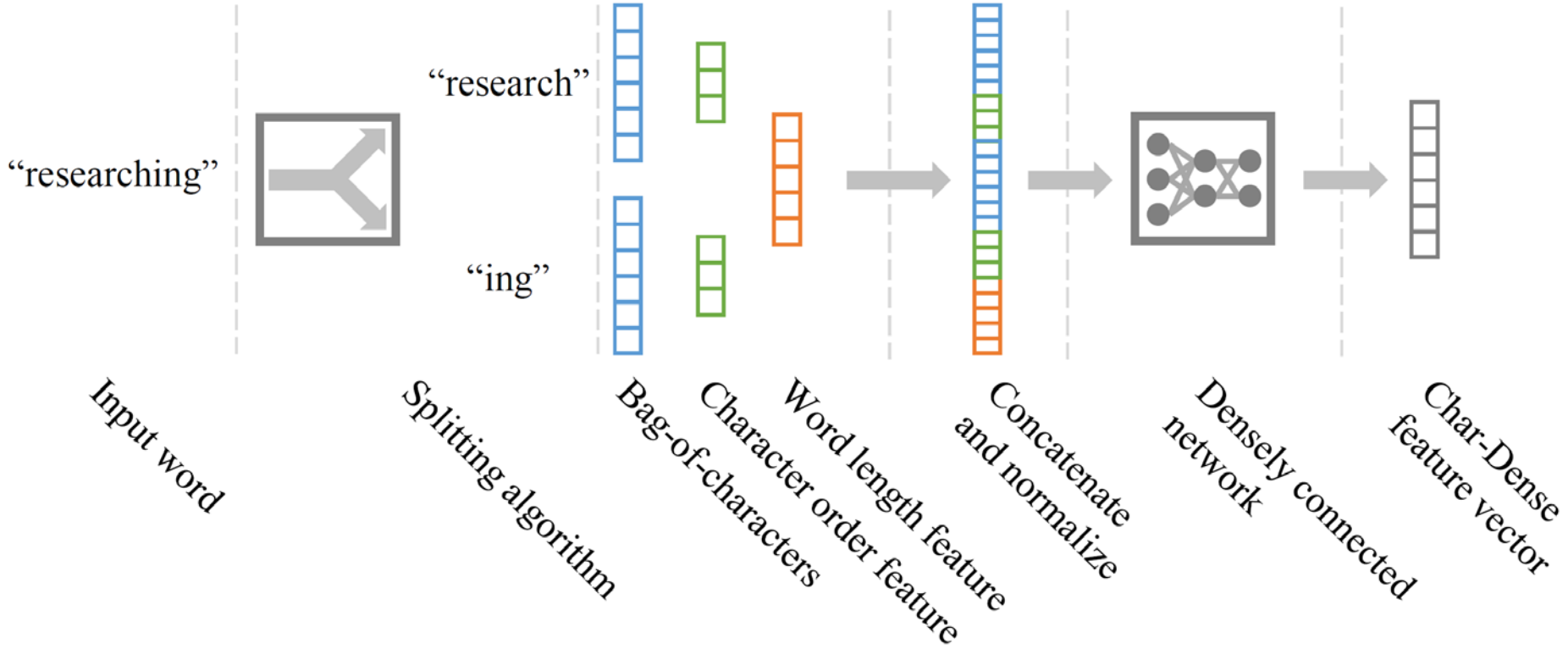
Introduction | Contribution

- Densely connected network (i.e. feed-forward NN, fully-connected NN, multi-layer perceptron)을 이용하여 효과적인 문자 단위 feature를 만들어내는 방법을 제안함
- 제안된 방법의 장점
- Hand-crafted feature나 data 전처리가 필요하지 않음
- RNN/CNN 대비 빠른 계산 속도를 보여주면서 성능도 우수함
- 특정 언어나 task에 국한되지 않음 – 알파벳권에만 존재하는 대소문자 정보나, NER에 많이 활용되는 gazetteer (개체명 사전) 불필요
- Contribution
- 효과적인 동시에 효율적인 문자 단위 feature 생성 방법을 제안함
- 다수의 실험을 통해 제안된 방법이 CNN 및 RNN보다 우수함을 정량적으로 증명함
- 가장 많이 연구된 NLP task들인 NER, POS, Slot tagging에서 state-of-the-art (POS, Slot) 혹은 그에 근접한 (NER) 성능을 보임

Proposed Method | Overview

- 제안된 방법은 Bag-of-Characters (BOC)를 바탕으로 함
- BOC는 문자의 순서 정보가 손실되기 때문에 word collision 현상이 발생함
  - Word collision – 서로 다른 단어가 동일한 vector 표현을 갖는 현상
- 이를 극복하기 위해 단어를 분할하고, 추가적인 feature를 추가
  - 이 때, feature들은 모두 특정 언어/task에 국한되지 않고 자동적으로 추출 가능
- 문자 BOC, 문자 순서 feature, 단어 길이 feature를 concatenate 하여 sparse vector를 생성함
  - 이 sparse vector는 단어마다 유일하고 변하지 않으므로 속도 향상을 위해 캐싱
- Sparse vector를 하나의 hidden layer가 있는 densely connected network의 입력으로 사용해서 dense feature vector를 생성

Proposed Method | Overview



Proposed Method | Splitting Words

- 각 단어는 k개의 조각으로 분할됨
  - Word collision을 줄이기 위함
- 학습 말뭉치의 n-gram 통계를 활용, 단어가 고빈도 n-gram들로 분할되도록 함
- Byte-pair encoding (BPE) 방법과 유사하지만, BPE는 임의의 개수로 단어를 분할하는 반면 제안된 방법은 단어가 k개의 조각으로 분할됨을 보장함
  - Concatenation과 Densely Connected Network을 이용하기 때문에 일정한 개수로 분할하는 것이 필수
- 각 단어 조각은 BOC를 이용하여 vector로 변환

Proposed Method | Splitting Words

Algorithm 1: Splitting word into k pieces

Input : word $w = (c_1, c_2, \cdots, c_n)$, n-gram statistics C_{ng}, number of pieces k
Output: $S = (s_1, \cdots, s_k)$ where $s_1 + s_2 + \cdots + s_k = w$

$S \leftarrow w$
while $|S| > k$ **do**
  $m = \underset{i}{\operatorname{argmax}}\, C_{ng}(c_i + c_{i+1})$
  $S \leftarrow (\cdots, s_{m-1}, s_m + s_{m+1}, s_{m+2}, \cdots)$
end
while $|S| < k$ **do**
  Append empty string to S
end
return S

Proposed Method | Character Order Feature

- BOC로 인해 문자 순서 정보가 손실되므로, 이를 보충할 feature를 추가함
- 디지털로 표현 가능한 모든 문자는 character-set (e.g. UTF-16)으로부터 숫자로 변환될 수 있으며, 이에 따라 문자 간 순서 비교가 가능함
- 길이 n의 문자열 T={c_1,c_2,⋯, c_n}로부터 아래 식에 따라 $C_asc\ (T)$와 $C_des\ (T)$를 계산함
- 각 단어 조각은 $C_asc\ (T)$>$C_des\ (T)$, $C_asc\ (T)$=$C_des\ (T)$, $C_asc\ (T)$<$C_des\ (T)$ 중 하나로 분류되며, 이 정보를 3차원 one-hot vector로 표현함

$$F_{asc}(T,k) = \begin{cases} 1, \text{if } c_k < c_{k+1} \\ 0, \text{otherwise} \end{cases}, \quad F_{des}(T,k) = \begin{cases} 1, \text{if } c_k > c_{k+1} \\ 0, \text{otherwise} \end{cases}$$

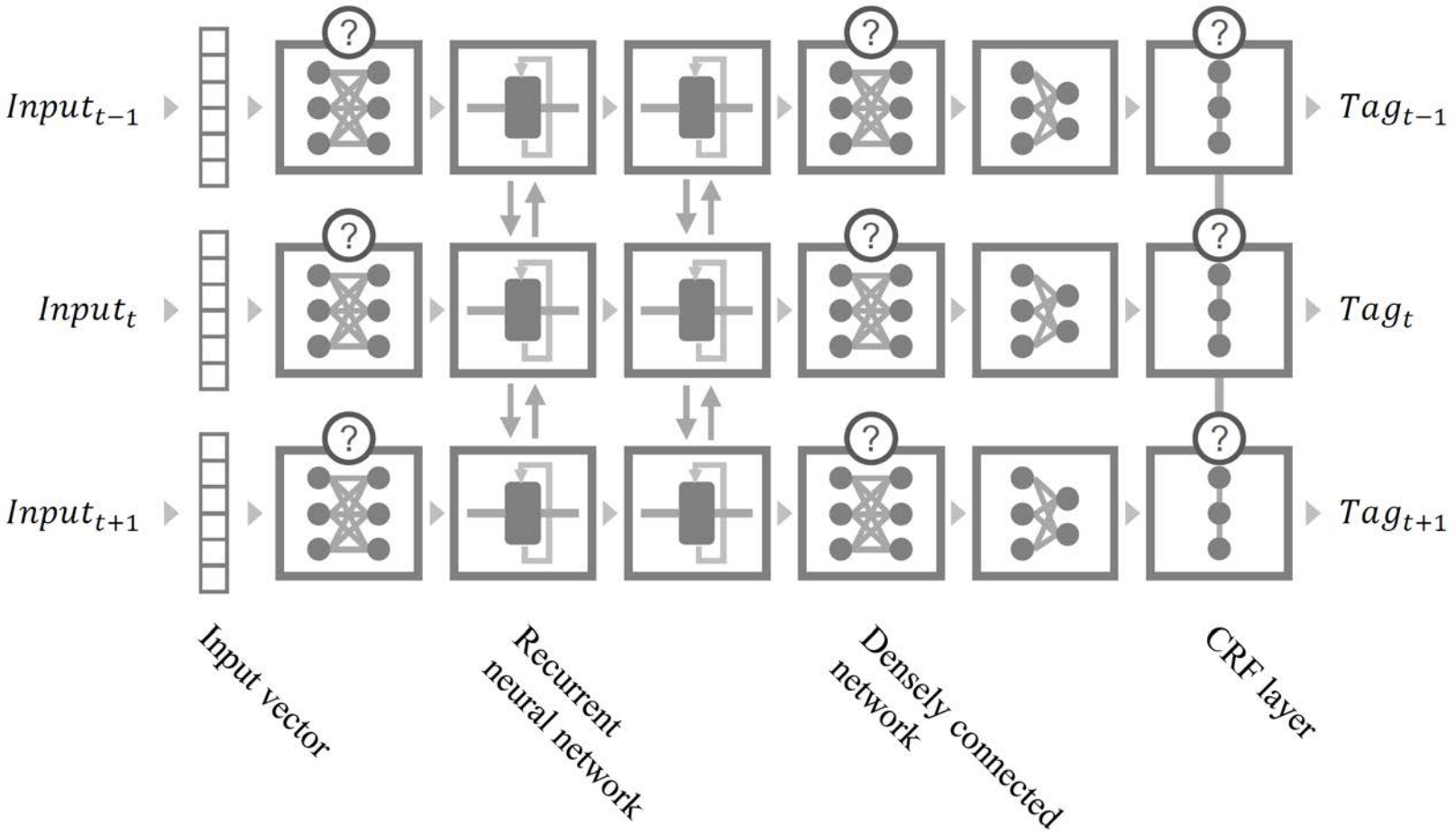
$$C_{asc}(T) = \sum_{k=1}^{n-1} F_{asc}(T,k), \quad C_{des}(T) = \sum_{k=1}^{n-1} F_{des}(T,k)$$

Proposed Method | Word Length Feature

- Word collision의 추가적인 감소를 위해, 각 단어를 구성하는 문자의 수를 feature로 활용
- 0~20의 숫자를 one-hot encoding을 이용하여 vector로 표현하며, 20글자를 초과하는 단어는 20글자로 취급

Experimental Setup | Model

- Sequence tagging은 Bidirectional-LSTM-CRF 모델을 사용함



Experimental Setup | Evaluation

- 제안된 모델을 3개의 task를 통해 정량적으로 검증함
- Slot tagging
  - Airline Travel Information System (ATIS) dataset 사용
  - 84종의 slot label (BIO 적용으로 127종의 class label)
- Part-of-Speech Tagging
  - Wall Street Journal (WSJ) portion of the Penn TreeBank (PTB) dataset 사용
  - 45종의 POS label
- Named Entity Recognition
  - CoNLL-2003 dataset 사용
  - 4종의 entity label (BIO 적용으로 9종의 class label)

Experimental Setup | Baseline Models

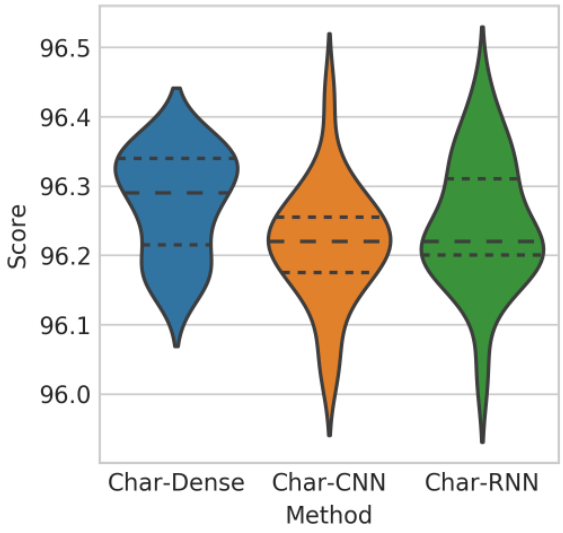
- CNN/RNN을 이용하여 문자 단위 feature를 생성하는 모델을 baseline으로 사용
- 문자 단위 feature 생성 방법 외의 모든 모델 구조는 동일하게 적용
- CRF layer는 모델 구조의 변화에 대해 robust하게 만드는 영향이 있으므로, 모델 간 차이를 확인하기 위해 baseline 실험에서는 CRF 대신 softmax layer 사용

Results and Discussion | Baseline Models

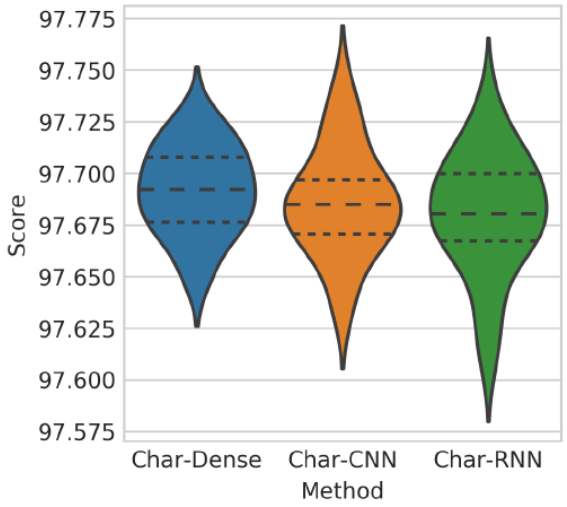
Method	Task		
	Slot	POS	NER
Char-CNN	96.22 (SD 0.08)	97.68 (SD 0.03)	89.08 (SD 0.20)
Char-RNN	96.25 (SD 0.09)	97.68 (SD 0.03)	**90.15** (SD 0.14)
Char-Dense (Ours)	**96.28** (SD **0.07**)	**97.69** (SD **0.02**)	90.10 (SD **0.13**)

- Baseline model들과의 성능 비교. SD=standard deviation

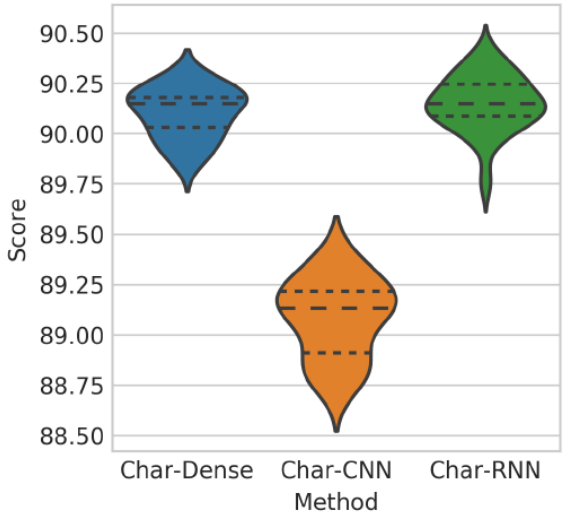
Results and Discussion | Baseline Models



(a) Slot



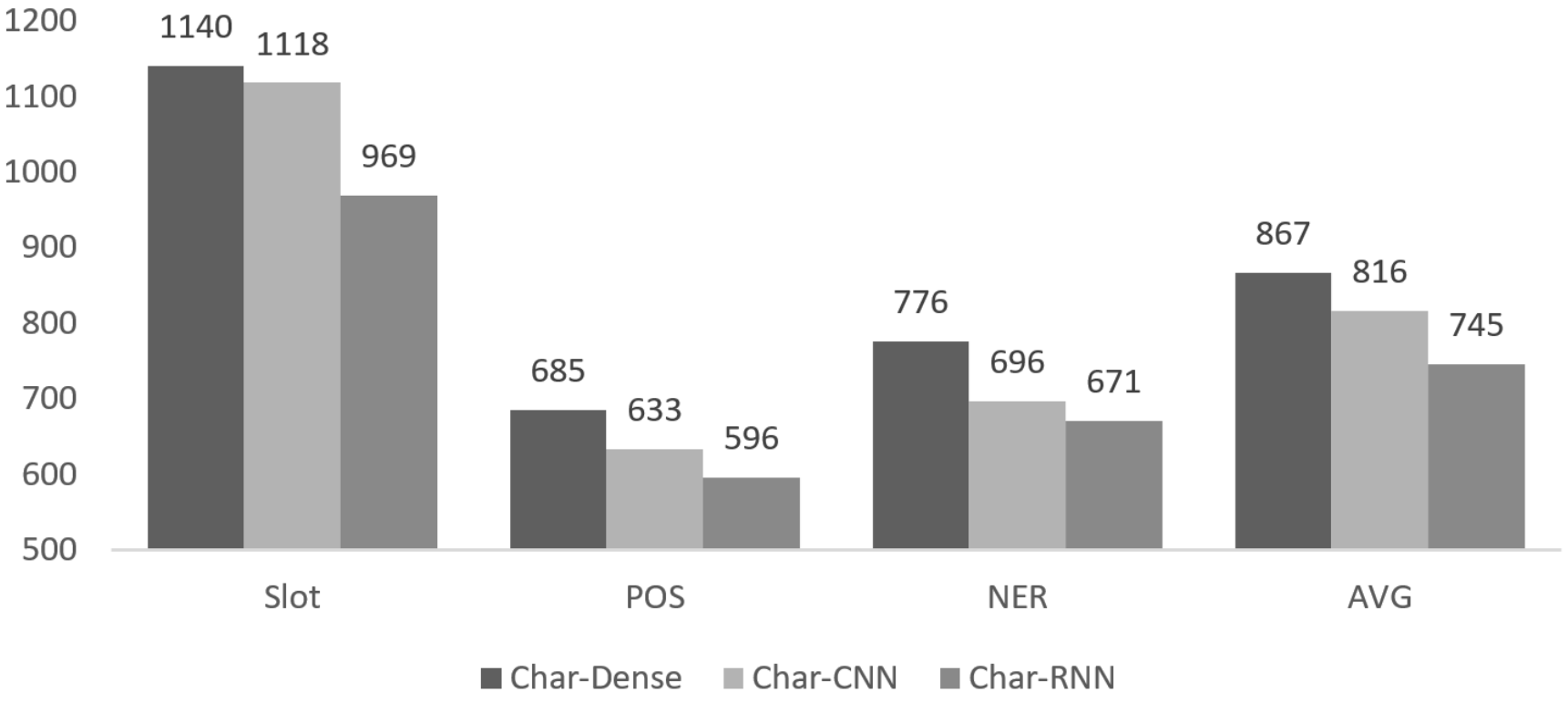
(b) POS



(c) NER

- 성능 분포를 나타낸 violin plot. 점선은 quartile을 나타냄

Results and Discussion | Tagging Speed



- 모델별 tagging 속도 비교. (초당 처리 문장 수)
- 제안된 모델에서 RNN 대비 16.32%, CNN 대비 6.29% 속도 향상이 나타남

Results and Discussion | Comparison

Slot		**POS**		**NER**	
Approach	F1	Approach	Acc.	Approach	F1
Mesnil et al. (2015)	94.73	Toutanova et al. (2003)	97.24	Ando and Zhang (2005)	89.31
Yao et al. (2014)	94.85	Manning (2011)	97.32	Collobert et al. (2011)	89.59
Liu and Lane (2015)	94.89	Shen et al. (2007)	97.33	Huang et al. (2015)	90.10
Yao et al. (2014)	95.08	Sun (2014)	97.36	Chiu and Nichols (2015)	90.77
Peng and Yao (2015)	95.25	Moore (2015)	97.36	Ratinov and Roth (2009)	90.80
Vu et al. (2016)	95.56	Hajič et al. (2009)	97.44	Lin and Wu (2009)	90.90
Vu (2016)	95.61	Søgaard (2011)	97.50	Passos et al. (2014)	90.90
Kurata et al. (2016)	95.66	Tsuboi (2014)	97.51	Lample et al. (2016)	90.94
Zhu and Yu (2017)	95.79	Huang et al. (2015)	97.55	Luo et al. (2015)	91.20
Zhai et al. (2017)	95.86	Choi (2016)	97.64	Ma and Hovy (2016)	**91.21**
Char-Dense w/o CRF (Ours)	96.36	Char-Dense w/o CRF (Ours)	**97.73**	Char-Dense w/o CRF (Ours)	90.28
Char-Dense w/ CRF (Ours)	**96.62**	Char-Dense w/ CRF (Ours)	97.65	Char-Dense w/ CRF (Ours)	91.13

Conclusion

- 효과적인 동시에 효율적인 문자 단위 feature 생성 방법을 제안하였음
- 제안된 방법은 parameter initialization의 영향을 CNN/RNN에 비하여 적게 받음
- 제안된 방법은 특정 언어나 task에 국한되지 않으며 generalizability가 우수함

- Future Work
  - 철자 오류에 대한 robustness 실험
  - Machine translation이나 automatic text summarization과 같은 추가적인 자연어처리 task에 적용

Structured “Image – Text” Mapping for Bidirectional “Image – Text” Transitions

고려대학교 NLP&AI Lab.

김규경

연구 소개 - 문제점 및 목표 소개

- Image-Text processing의 문제점:
  - 이미지 → 텍스트의 경우:
    - ▶ 비(非)구조적인 접근으로 인하여 오류 수정 및 학습에 개입하기 힘듬
  - 텍스트 → Object의 경우:
    - ▶ 텍스트가 표현하는 대상(이미지)의 속성을 고려하지 않아 당연한 논리적 결과를 인지 못함

- 목표:
  - 구조적인 이미지 – 텍스트 Embedded Map을 구축하여 구조적인 이미지 분석과 대상 속성을 고려하는 Semantic Text Embedding을 가능케 하는 모델의 구축

연구 내용 - 문제점 및 목표 소개

- 이미지 → 텍스트의 경우:구조적인 접근의 부재

Okapi

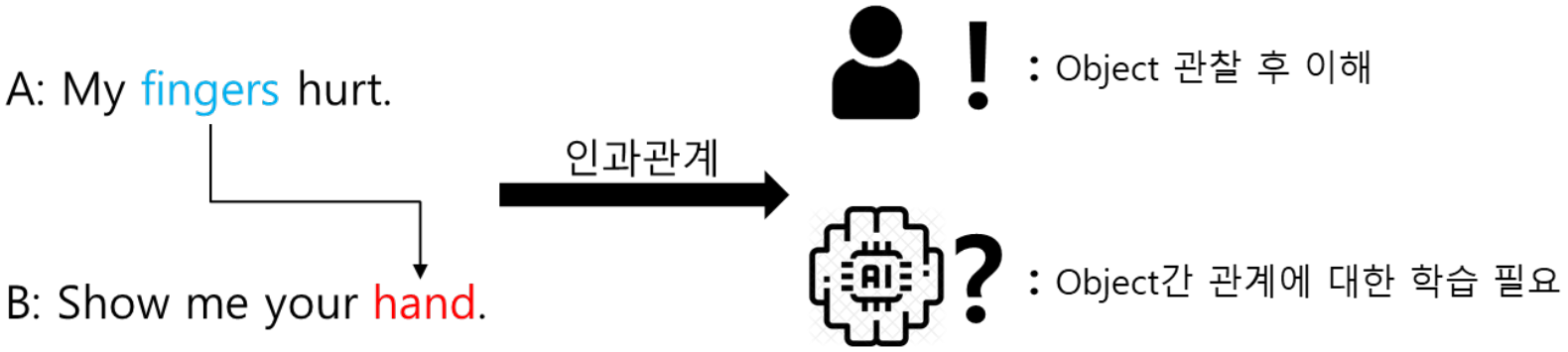
Deer

Zebra

- 각각 Deer, Zebra, Okapi를 학습하여도 3 동물의 구조적인 공통점과 차이점을 분석 못함.
- 이미지내의 수많은 Feature들을 분석하지 않고 한번에 처리하면서 생기는 결과
- 해당 연구에서는 이러한 Feature들의 구조적인 분석을 거치는 Recognition을 시도함

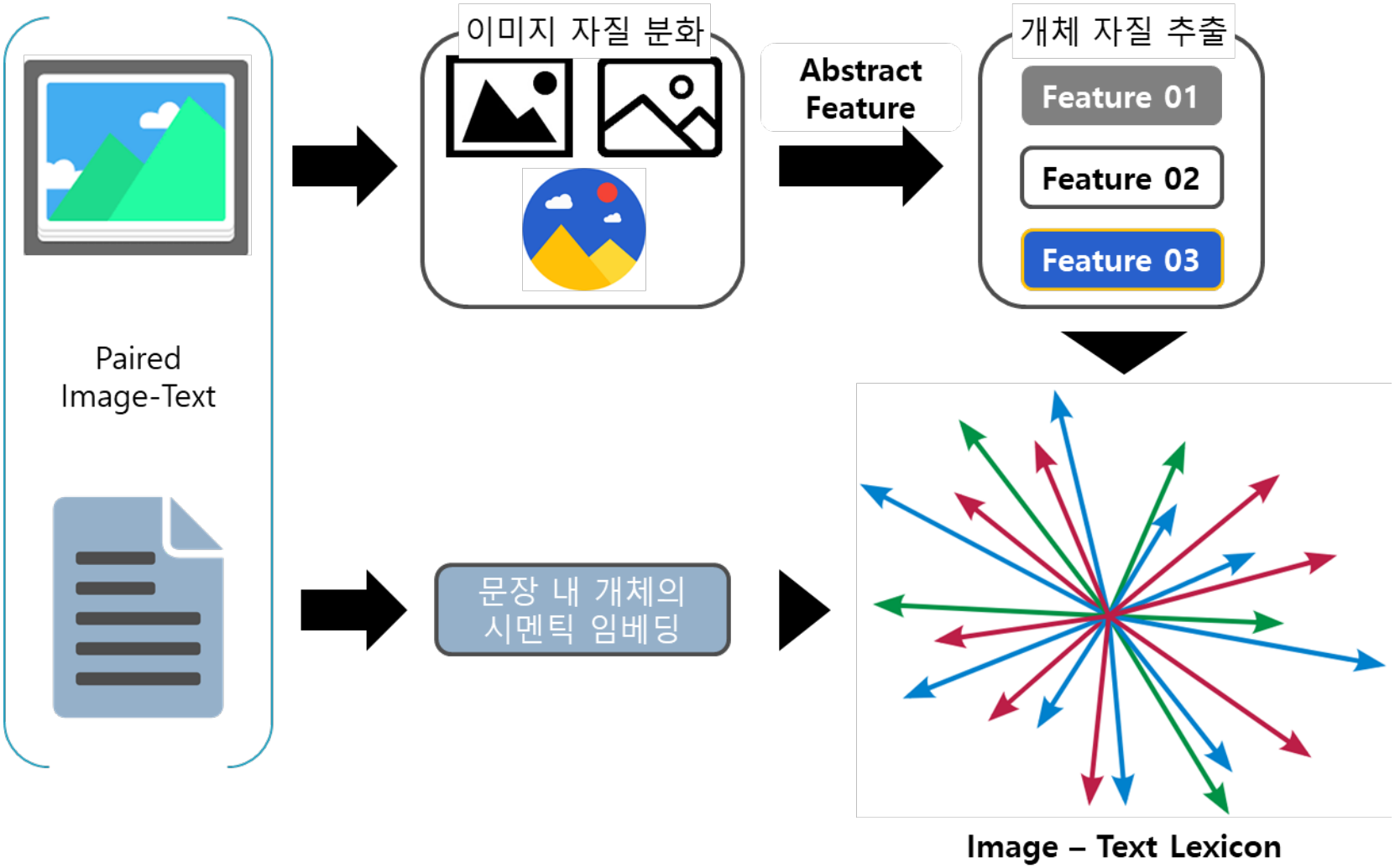
연구 내용 - 문제점 및 목표 소개

- 텍스트의 경우: 언급하는 대상 속성에 대한 이해 결여

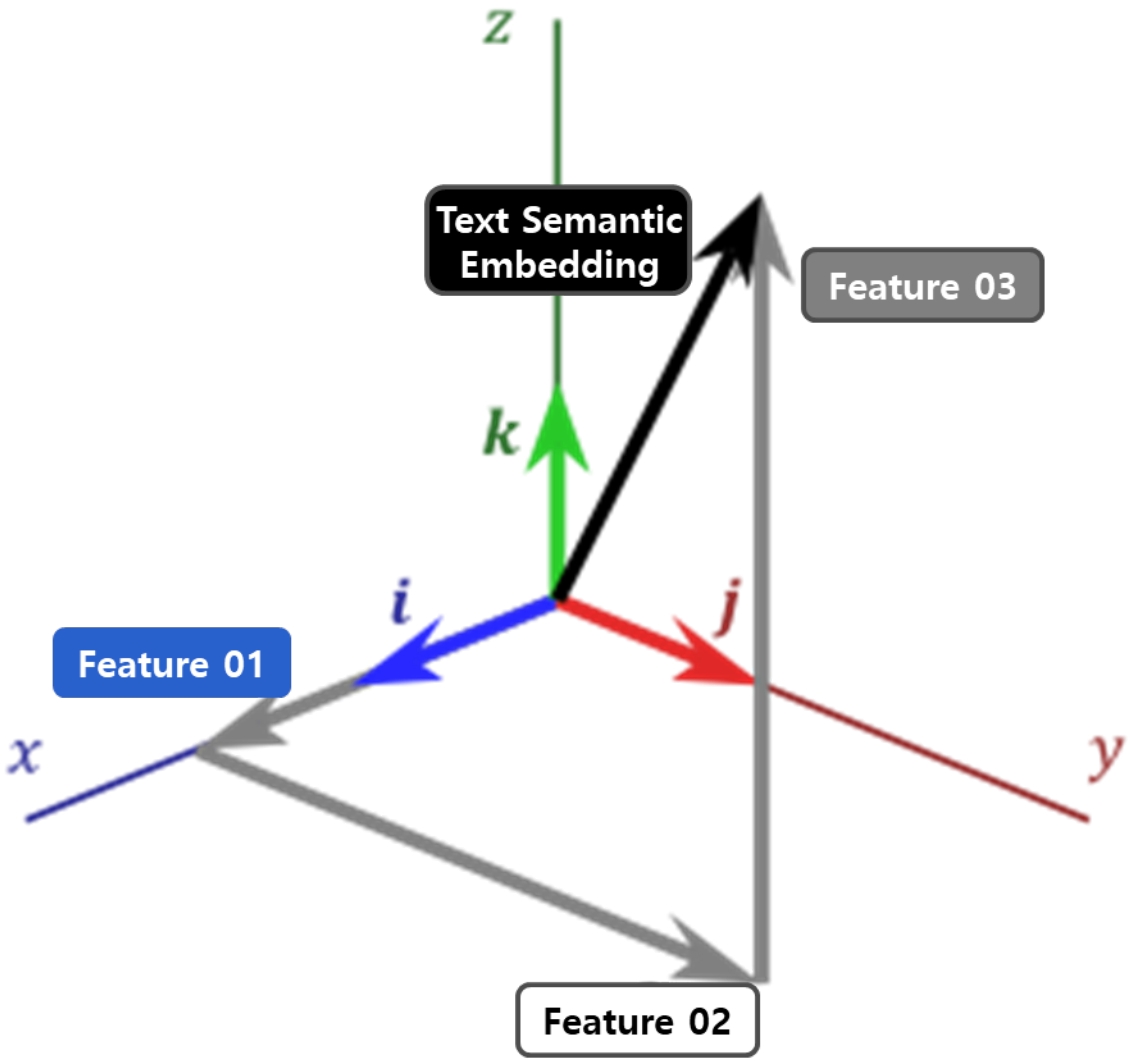


- Finger와 Hand가 가지는 두 개체간 공통된 속성을 텍스트 데이터만으로는 알 수 없음
- 이는 이미지 개체로서는 너무나 당연하나, 텍스트에는 해당 속성이 반영이 안돼서 생기는 문제
- 해당 연구에서는 이미지 Feature들을 통해 텍스트에 Object가 가지는 속성을 반영할 수 있게 함

연구 내용 - 모델 개요

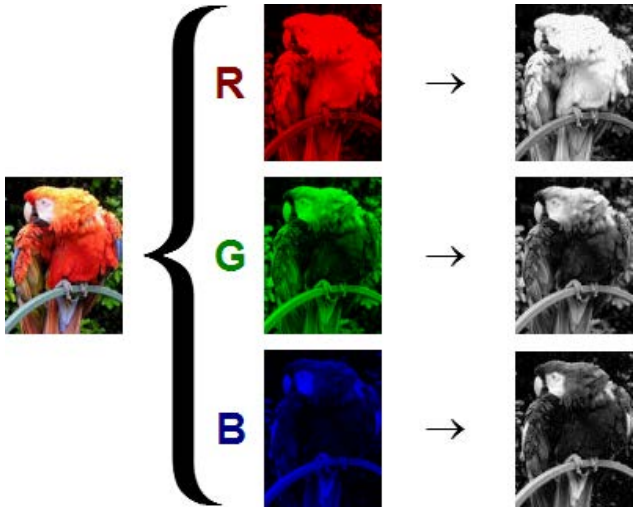

이미지 자질 분화
Abstract Feature
개체 자질 추출
Feature 01
Feature 02
Feature 03
Paired Image-Text
문장 내 개체의 시멘틱 임베딩
Image – Text Lexicon

연구 내용 - Output


z
Text Semantic Embedding
Feature 03
k
i
j
Feature 01
x
y
Feature 02

연구 내용 - 모델 설명

- 이미지 자질 분화



RGB channel seperation

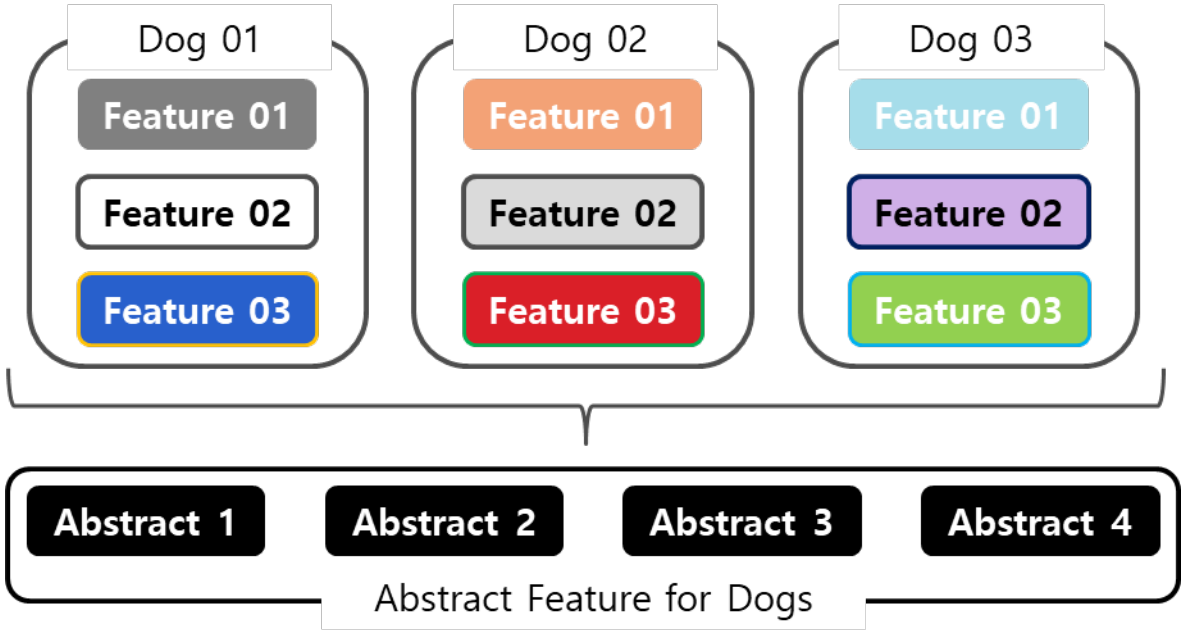
SURF

Edge Detection

- 이미지 내의 각 Feature의 특징을 극대화 및 다양화하기 위해 3가지 방법으로 이미지를 처리함
- 처리된 이미지에서 뽑아낸 자질들은 해당 Object의 Feature는 본래의 것보다 단순해짐

연구 내용 - 모델 설명

- 이미지 자질 추출

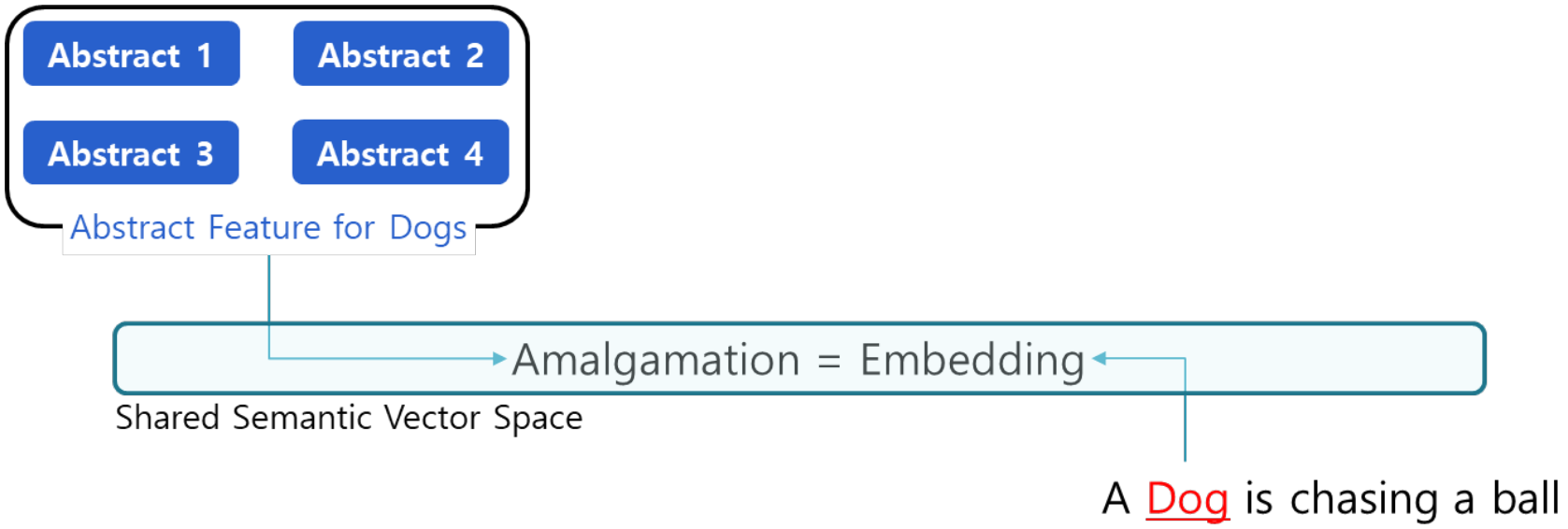


- Train a classifier that classifies features based on which object it belongs to.
- Based by this trained classifier, construct “Abstract Features” for each class(object)

연구 내용 - 모델 설명

- Aligning with Semantic Embeddings
  - The embeddings are based from BERT embeddings
  - As each paired sentences contains similar contexts, so does the entities within it.
  - Based by this we align the amalgamation of the abstract feature to become like this.

→ In other words, the amalgamation is used like a sentences used in BERT



모델 실험 - 모델 실험 설계

- Experiment: Visual Question & Answering
  - A set of images paired with questions which requires visual understanding to answer
  - A fair test for testing both Text-to-Image and Image-to-Text

  → The model itself will be a pre-existing model, but trained with our vector space.

결론

- Contribution:
  - Constructing a true bidirectional Text-Image semantic space that benefits both Natural Language Processing and Image Processing

    → Binding Text to physical information via Image Features

    → Structural Image Feature Analysis/Processing via Textual information

- Futureworks:
  - As this is only a construction of a shared semantic space…

    → Generating Images via Text description

    → Describing Image features via Text

    ex) finding common features between 2 images and point out the different ones

Utilizing Dual Memory Network with Domain Knowledge for Goal-Oriented Dialogues

고려대학교 NLP&AI Lab.

이동엽

연구 내용 - 이중 메모리와 도메인 지식을 이용한 목적 지향적 대화 시스템

- 기존 대화 시스템을 구성하는 방식으로는 rule-based와 hand-craft 자질들을 이용한 방법들이 존재하였음
- 최근 딥러닝 모델들을(LSTM, Seq2Seq…) 대화 시스템에 적용하고 있지만, 많은 양의 학습 데이터를 필요로 한다는 한계점이 존재
- 도메인 지식을 대화 시스템에 활용하는 연구가 존재하지만, 해당 연구는 대화의 전체 이력을 고려할 수 없고, hand-craft 자질을 필요로함
- 본 연구에서는 도메인 지식을 활용하여 시스템 응답 템플릿을 구성하고, 이중 메모리를 통해 대화의 전체 이력을 고려할 수 있는 대화 시스템 모델을 제안함

연구 내용 - 시스템 응답 생성시 도메인 지식 반영

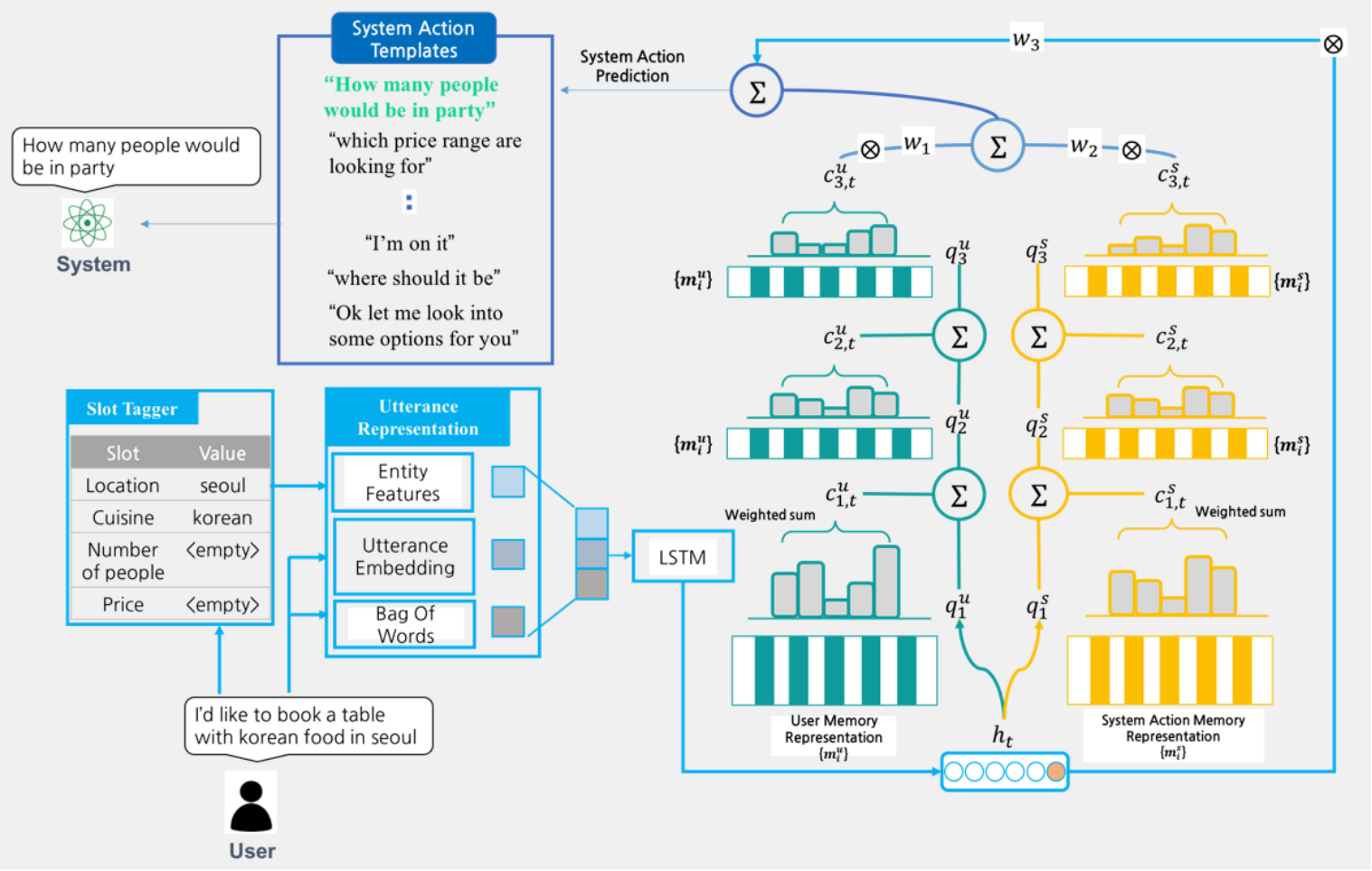
System Action Templates
hello what can I help you with today
any preference on a type of cuisine
api_call <price> <number of people> <cuisine> <location>
great let me do the reservation
here it is <info_address>
here it is <info_phone>
how many people would be in your party
I'm on it
is there anything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

- 도메인 지식을 반영하여 시스템 응답 액션 템플릿 구성
- 적은 양의 데이터로, 대화 모델의 학습이 가능

연구 내용 - 이중 메모리를 이용한 대화 이력 추적

- 사용자 발화와 시스템 응답 발화 표현(rerpesentation)을 구성하고 이를 각각 사용자 발화 메모리, 시스템 응답 메모리로 구성
- 주어진 사용자 발화에 대한 응답 생성시, 이중 메모리에서 어떤 발화에 각각 집중할 지 고려하여 시스템 응답 생성

연구 내용 - 제안 모델 구조도



제안 모델 구조도

연구 내용 - 실험 데이터

Dataset	Task 5	Task 6
Training dialogs	1000	1618
Validation dialogs	1000	500
Test dialogs	1000	1117
Slot label size	4	3
System action templates size	16	58
Vocab size	3747	1229

Table 4: Statistics of the bAbI task datasets.

- End-to-End 방식의 목적 지향적 대화 시스템 모델을 위한 평가로 많이 사용되는 bAbI 데이터를 사용
- 해당 데이터는 식당 예약에 관련한 도메인 데이터임

연구 내용 - Evaulation Matrix

- Evaulation Matrix
  - Per-response accuracy = $\frac{(Number\ of\ Correct\ Utterance)}{(Total\ Number\ of\ Utterance)}$
  - Per-dialogue accuracy = $\frac{(Number\ of\ dialogs(every\ response\ is\ corret))}{(Total\ Number\ of\ Dialogue)}$

연구 내용 - 제안 모델 성능

Apporach			Task-5(OOV) Per-response	Task-5(OOV) Per-dialogue	Task-6 Per-response	Task-6 Per-dialogue
(Bordes et al., 2016)			77.7	0.0	41.1	0.0
(Perez and Liu, 2016)			79.4	0.0	48.7	1.4
(Seo et al., 2016)			96.0	-	51.1	-
(Eric and Manning, 2017)			-	-	48.0	1.5
(Williams et al., 2017)			**100**	**100**	53.1	1.9
Proposed Model	BoW		99.9	98.6	54.9	1.0
	UE		99.7	97.3	55.1	0.9
	BoW+UE		**100**	99.3	55.1	1.1
	BoW+UE+EF(RB)		**100**	**100**	55.8	1.3
	BoW+UE+EF(ST)		**100**	**100**	55.8	1.4
	1-hop	UM	**100**	**100**	55.5	1.8
		SM	**100**	**100**	55.2	1.5
		UM+SM	**100**	**100**	**55.9**	2.0
	2-hop	UM+SM	**100**	**100**	55.7	2.1
	3-hop	UM+SM	**100**	**100**	55.7	**2.3**

Table 6: Performance results of per-response and per-dialog accuracy on bAbI tasks. Key: BoW = bag-of-words representation; UE = utterance embedding using a word2vec; EF(RB) = entity features representation with rule-based entity names; EF(ST) = entity features representation with slot tagging model (section 3.1); UM = user memory representation; SM = system memory representation. Experiments on memory hop(1 to 3) were based on utterance representation utilizing BoW + UE + EF(ST).

제안 모델 성능

Adversarial Learning for Natural Language Understanding

고려대학교 NLP&AI Lab.

이동엽

연구 내용 - 기존 연구 내용

- 자연어 이해 모델 구성을 위한 방법으로는 CRF, LSTM과 같이 순차열 라벨링을 위한 모델이 많이 이용되고 있음
- 하지만 이러한 모델을 학습하기 위해서는 많은 양의 라벨링 된 데이터를 필요로 함
- 또한 새로운 도메인으로 모델을 확장하고자 할 때, 새로운 도메인에 대한 데이터를 다시 라벨링 해야함 (non-scalable)

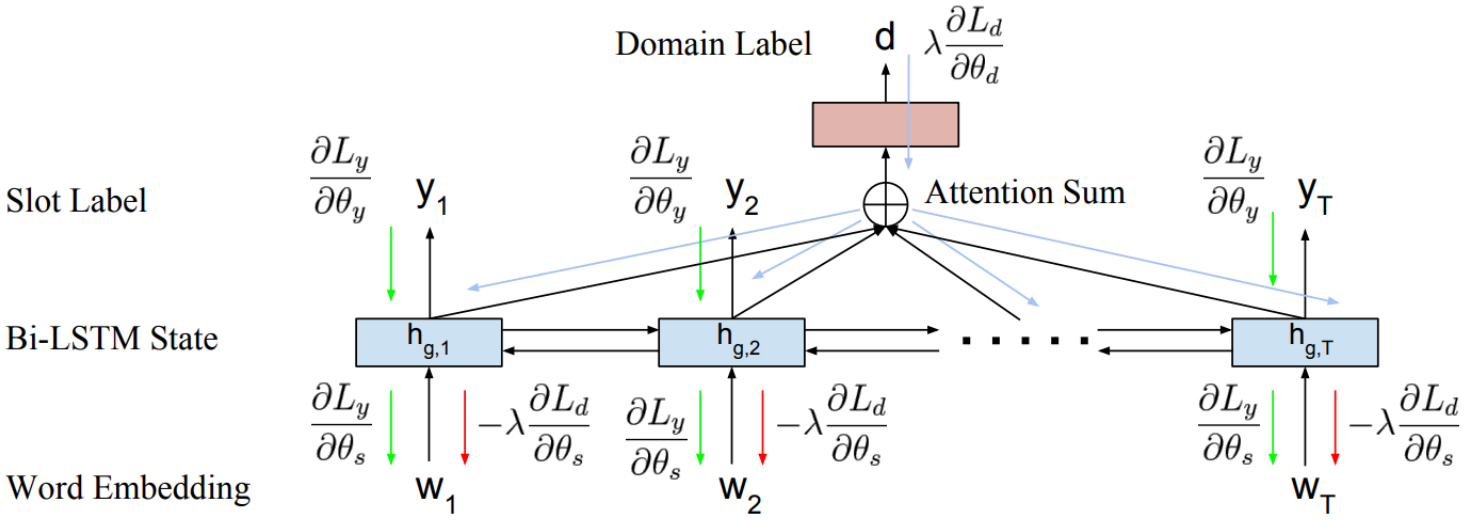
연구 내용 - 적대 학습을 이용한 자연어 이해

- Motivations
  - 새로운 도메인(target)으로 시스템을 확장하고자 할 때, 관련있는 기존 도메인(source)의 데이터를 활용하여 대상 도메인(target)의 모델 성능을 향상
  - 적대 학습을 이용하여 두 도메인간의 공유 될 수 있는 자질(shared features)을 학습하고, 이를 대상 도메인을 위한 모델 학습에 이용

- Benefits
  - 새로운 도메인을 위한 모델 학습 시, 보다 적은 양의 데이터로 모델의 학습이 가능
  - 또한 적은양의 데이터만으로 모델 성능의 향상 가능

연구 내용 - Learning Shared Features

- Negative Gradient 방법을 이용하여 Domain Classification Loss 학습을 방해



Multi-Domain Adversarial Learning for Slot Filling in Spoken Language Understanding [NIPS 2017]

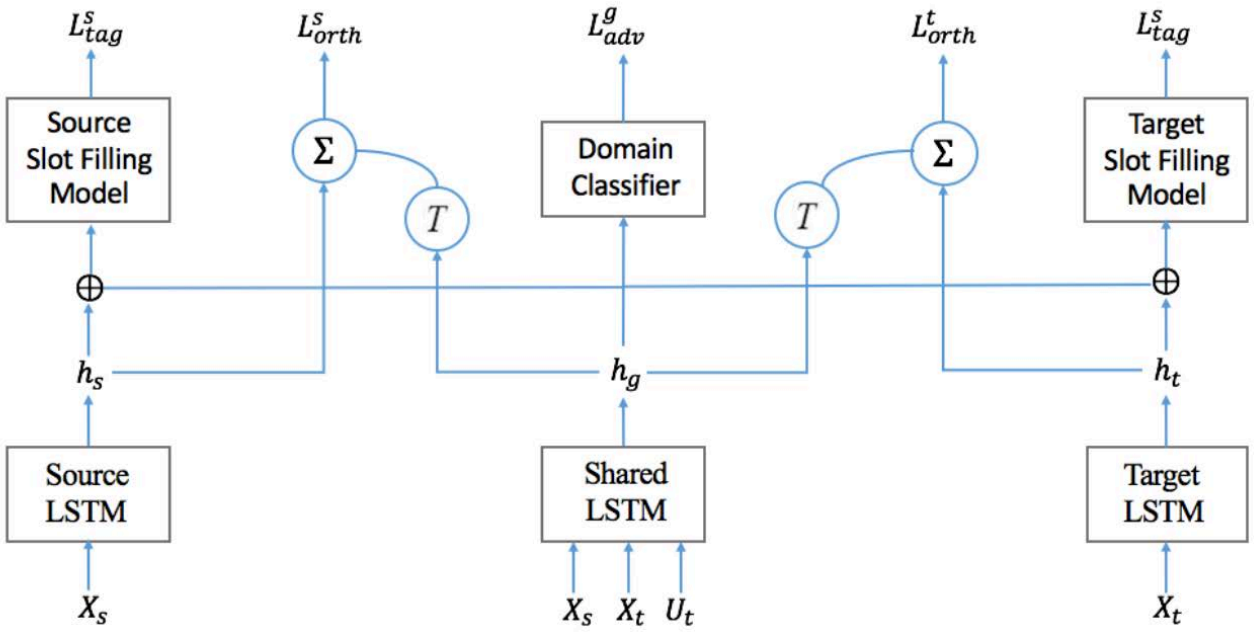
연구 내용 - Learning Shared Features

- Random Prediction
  - 도메인 라벨값을 랜덤하게 할당하고, 발화로부터 정확하지 않은 도메인 라벨을 예측하도록 함으로써 공유 자질 학습

    Adversarial Adaptation of Synthetic or Stale Data [ACL 2017]

- 두 가지의 데이터 전이 (data shift) 문제를 풀기 위해 제안됨
  - Synthetic data => Live user data (deployment shift)
  - Stale data => Current data (temporal shift)

연구 내용 - 제안 모델 구조도



Input representation:

X_s: Labeled source domain data

X_t: Labeled target domain data

U_t: Unlabeled target domain data

Parameters:

h_s: Source domain-specific features

h_g: Domain-general features

h_t: Target domain-specific features

연구 내용 - Adversarial Domain Classification Loss

- Adversarial Domain Claasification Loss
  - 도메인 공유 자질을 학습하기 위해 적대 학습 분류기를 학습
  - 적대 학습 방식으로 domain random prediction 방법을 이용
  - h_g로부터 도메인 $d\hat{}i$ 예측시 Source 도메인 50%, Target 도메인 50% 확률로 랜덤하게 예측 => 데이터로부터 해당 데이터의 도메인을 잘 구분하지 못하도록 학습이 진행되기 때문에 도메인 공유 자질이 학습됨
  - $\lambda L_{adv}^{g} = -\sum logp(d^{i}|h_{g})$, where λ is coefficient

연구 내용 - Joint Optimization

Input representation:

X_s: Labeled source domain data

X_t: Labeled target domain data

U_t: Unlabeled target domain data

Parameters:

h_s: Source domain-specific features

h_g: Domain-general features

h_t: Target domain-specific features

Each epoch

1) Train source slot filling model with X_s based on h_s and h_g
2) Train domain classifier with X_s, X_t, U_t based on h_g
3) Train target slot filling model with X_t based on h_t and h_g

End of epochs

[알고리즘 1] Joint Optimization을 통한 슬롯 채우기 모델 학습 과정

연구 내용 - 실험 데이터

- 실험 데이터

Datasets	Slot	Train Label/Unlabel	Test	Description
ATIS	127	4978/0	893	Flight reservation
MIT Rest.	17	766/6894	1521	Restaurant reservation

[표 1] 실험 데이터 통계 정보

- Slot Filling 모델 학습을 위해 ATIS(항공 예약)와 MIT Rest.(식당 예약) 데이터를 이용
- 일반적인 도메인 적응 실험 환경에 따라, 라벨링된 Source 도메인 데이터는 충분하고 Target 도메인 데이터는 10%만 라벨링 되어있다고 가정
- Source 도메인 : ATIS, Target 도메인 : MIT Rest.

연구 내용 - 실험 결과

- 실험 결과
  - Slot Filling 모델의 성능 평가를 위해 f1-score metric을 사용
  - TGT는 적대 학습을 적용하지 않은 모델, ADV는 적대 학습을 적용한 모델
  - 실험 결과 λ=0.1일때, 67.12의 f1-score로 가장 높은 성능을 나타냄

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

[표 2] 적대 학습 방법을 적용한 슬롯 채우기 모델 성능

개체명 인식(Named Entity Recognition) 시스템 연구 및 개발

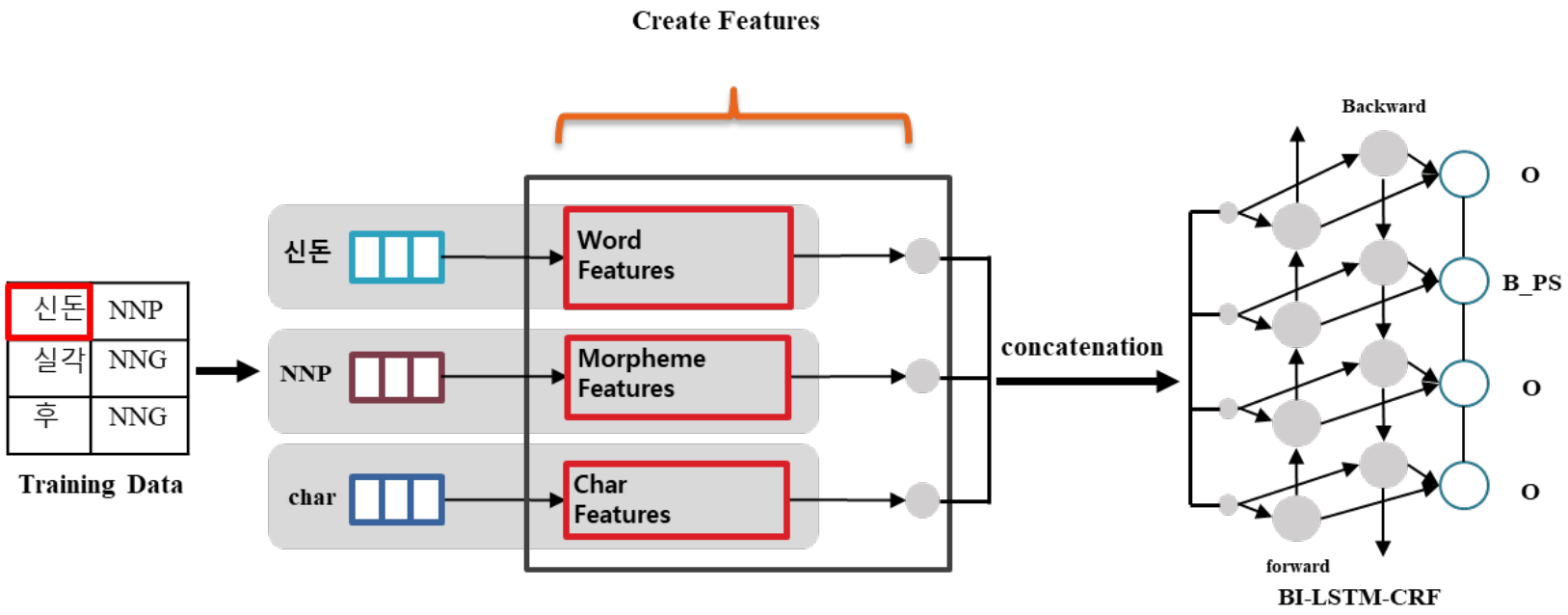
고려대학교 NLP&AI Lab.

김경민

연구 소개 - 개체명 인식

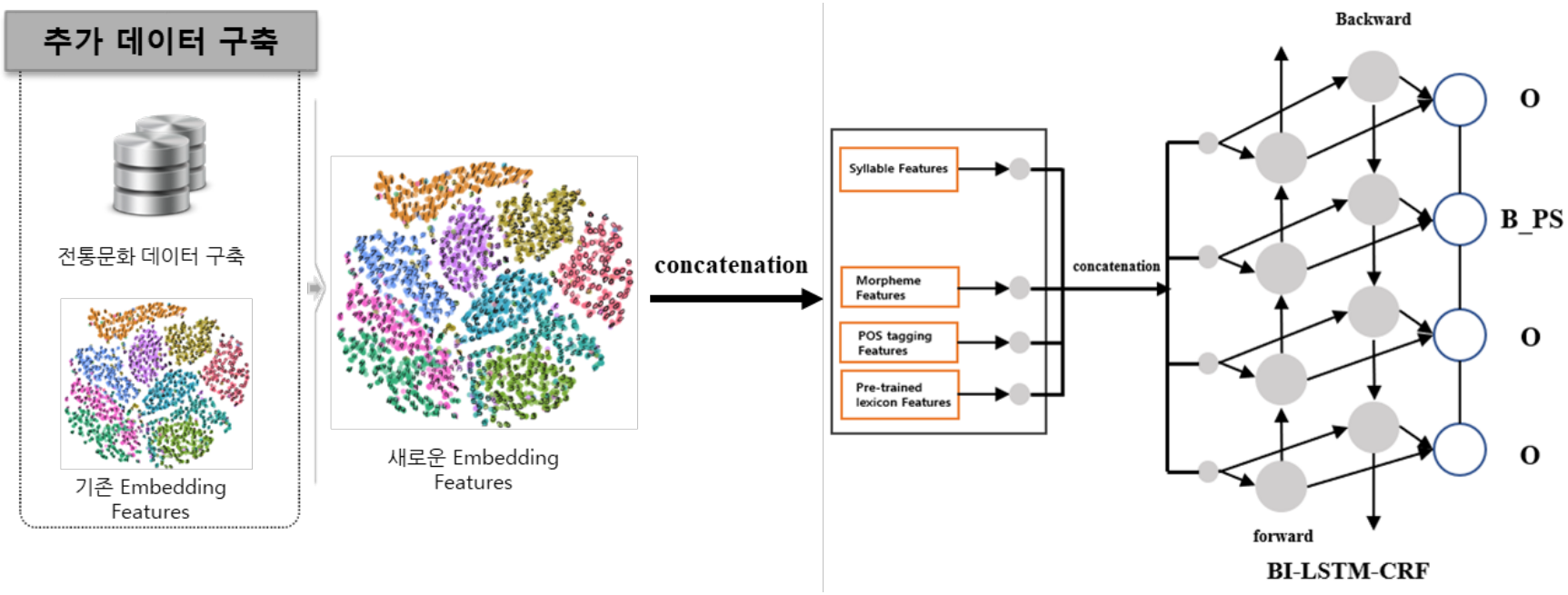
- 목표
  - 기존 영어 혹은 한글 기반 개체명 인식시스템 base model을 기반으로 word, char, morpheme 자질을 활용한 BI-LSTM-CRF 모델 구축
  - 새로운 데이터를 구축하여 기존 입력 데이터에 추가한 모델 학습

연구 소개 - Bi-LSTM-CRF 기반의 개체명 인식 모델



- Word, Morpheme, Char 단위의 각각의 자질을 활용하여 모델 성능을 평가함

연구 소개 - 구축 데이터 기반으로 기존 모델에 추가 학습



- 기존 embedding feature에 구축한 데이터를 추가 입력으로 학습하여 새로운 Embedding Feature를 생성
- 생성된 것을 기존 Bi-LSTM-CNN-CRF 모델의 입력으로 넣어 성능 평가

연구 소개 - 구축 데이터 기반으로 기존 모델에 추가 학습

```
input> 1597년 정유재란이 일어나자 원균은 삼도수군통제사가 되어 일본 수군과 대전했으나 조선 수군은 다대포·칠천량에서 대패하여 제해권을 상실하였고 원균과 이억기 등의 장수는 전사하였다.

output> <1597년:DT> <정유재란:DT>이 일어나자 원균은 삼도수군통제사가 되어 <일본:LC> 수군과 대전했으나 <조선:DT> 수군은 <다대포·칠천량:LC>에서 대패하여 제해권을 상실하였고 원균과 이억기 등의 장수는 전사하였다.

1597년    DT

정유재란    DT

일본    LC

조선    DT

다대포·칠천량    LC
```

- 기존 개체명 시스템 결과

"개체명 인식 결과": "<1420년:DT>에 다시 집현전이 설립되었을 때는 10명에서 <1423년:DT> 16인이었다가 <1435년:DT>에는 32인까지 증가되었다. 기능 또한 확대되어 유교주의적 의례, 제도, 문화의 정리 사업인 고제 연구와 편찬 사업이 시작되었다. <조선 문종:OG>이 즉위하면서 집현전이 대간 차출의 본거가 되어 호간고론적인 집단으로 변하였다."}

- 새로운 개체명 시스템 결과

개체명 인식 결과:
"<1420년:DT>에 다시 <집현전:OG>이 설립되었을 때는 10명에서 <1423년:DT> 16인이었다가 <1435년:DT>에는 32인까지 증가되었다. 기능 또한 확대되어 유교주의적 의례, 제도, 문화의 정리 사업인 고제 연구와 편찬 사업이 시작되었다. <조선:LC> 문종이 즉위하면서 <집현전:OG>이 대간 차출의 본거가 되어 호간고론적인 집단으로 변하였다."

  - 기존에 학습된 모델로 인식 하기 어려웠던 전통문화 텍스트 인식

Enhanced Sequential Representation Augmented with Utterance-level Attention for Response Selection

고려대학교 NLP&AI Lab.

황태선

연구 내용 - Response Selection

- 두 사용자의 Multi-turn 대화에서 마지막에 올 발화 내용을 정답 후보들 중 정답을 예측하는 연구
- Human-computer 대화 시스템은 자연스럽고 일관된 대화를 이끌어 나가는 것을 목표로 함. 특히 검색 기반 대화 시스템에서는 대화 코퍼스 내에서 가장 적합한 응답 검색을 바탕으로 응답
- 정답 이전의 대화를 입력으로 하여 전체 대화가 자연스럽게 이어질 수 있는 정답을 선택
- 대화 내에서 단어, 문장과 같은 중요한 정보와 정답의 관련성을 어떻게 확인할 것인지, 또한 대화 내 발화 간의 관계를 어떻게 모델링 할 지 등에 대한 연구가 필요함

연구 내용 - Multi-turn 대화 내 발화 예측

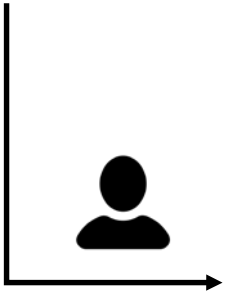
My skype doesn't start, anyone knows why?

Skype is having an issue; https://twitter.com/#!/Skype

So I can only wait?

At this time, that's what I suggest

Thanks, better to know that its in progress than missing out. Maybe skype has this problem with Ubuntu because MS bought skype



Response Selection

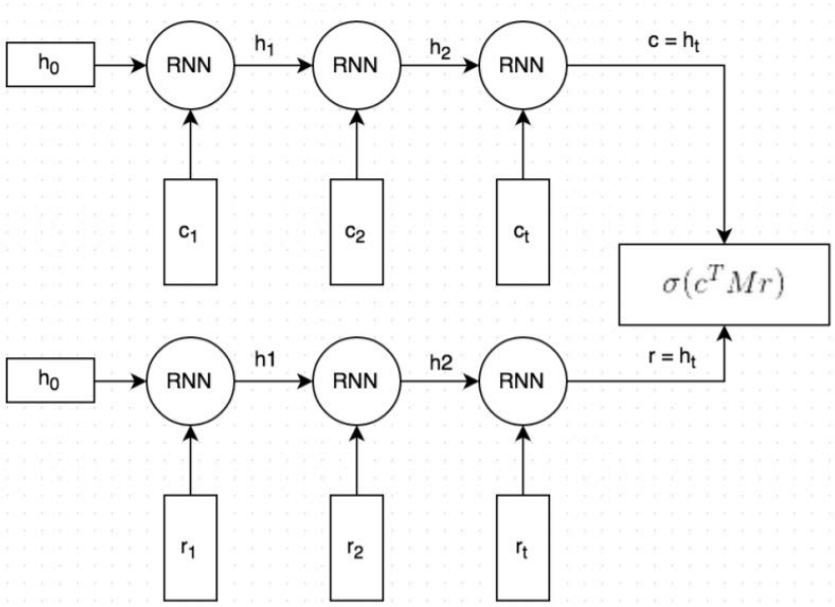
1. have you tried right shift?
2. there are several apps for data backup, after that do a 64 bit clean install
3. **Don't be silly, it's a cross-platform issue**

⋮

99. It is cool, one sec
100. thanks for the link

연구 내용 - Dual Encoder w/LSTM [Lowe et al, 2015]

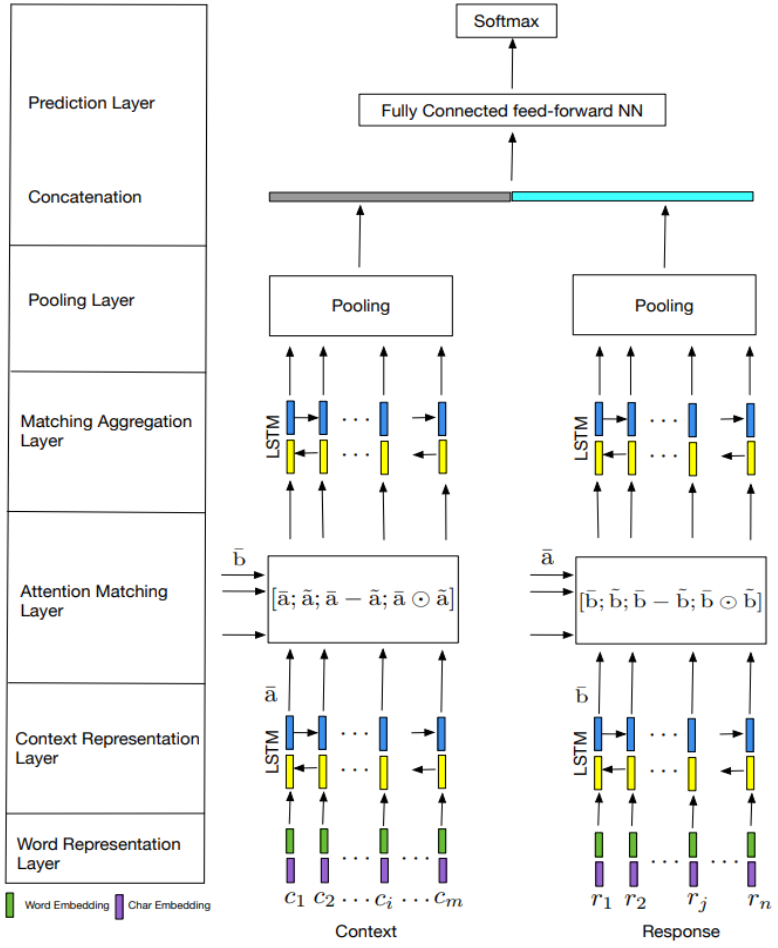
- 두 사용자간 대화 문장(Context)와 각 정답 문장(Response)의 단어들이 RNN 또는 LSTM의 입력으로 들어감
- 양방향 LSTM을 통한 모델의 경우 두 문장의 단어 간 관련성을 반영하지 않기 때문에 문장 간 유사도를 판단하기 어려움
- LSTM을 통한 word-sequence 모델은 대화 발화 turn별로 고려하지 않기 때문에 대화가 길어질 경우 대화의 중간 맥락의 정보 손실이 발생할 수 있음



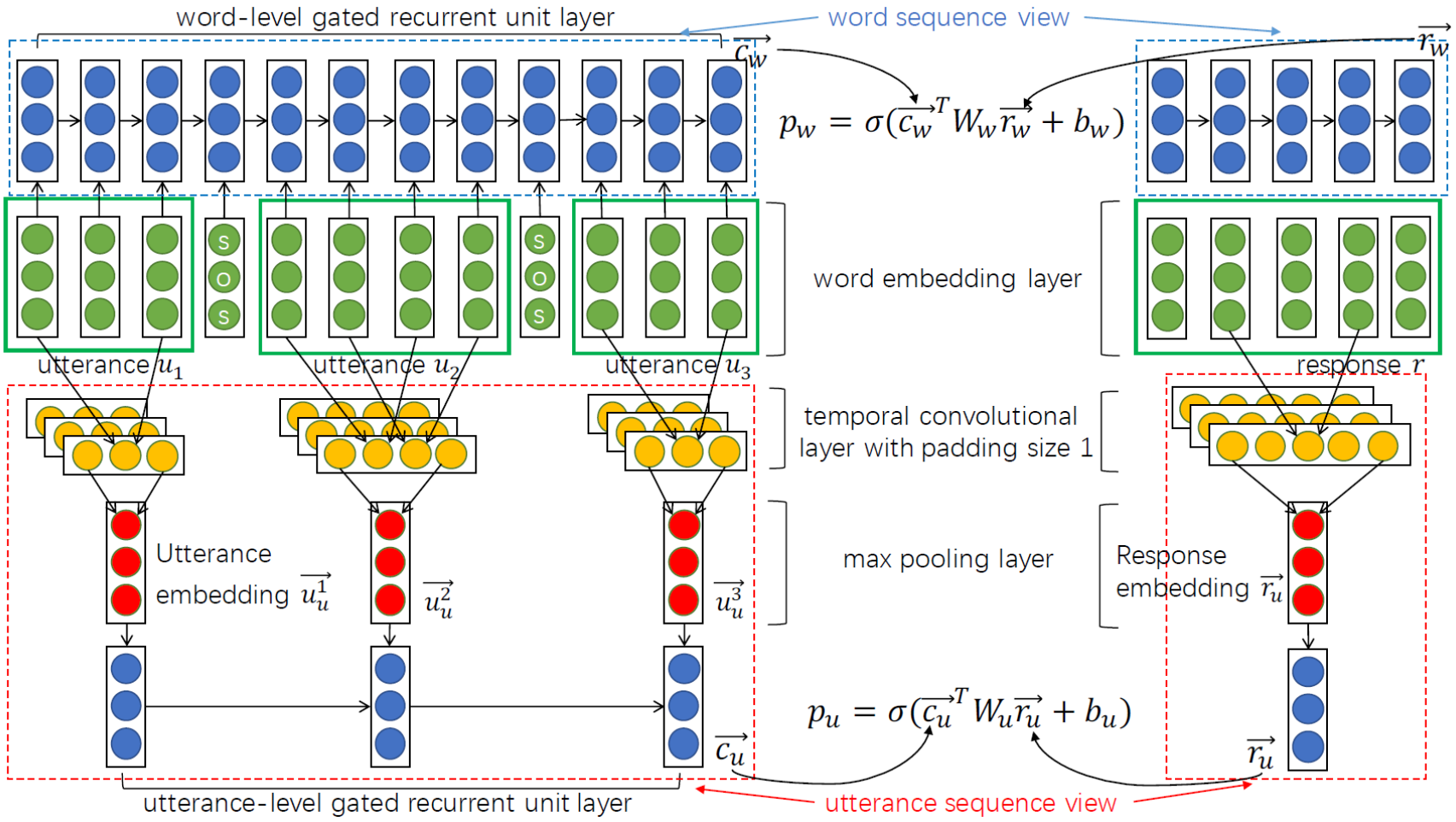
Lowe et al. The Ubuntu Dialogue Corpus: A Large Dataset for Research in unstructured Multi-Turn Dialogue Systems. In Proceedings of SIGDIAL, 2015.

연구 내용 - ESIM [Dong and Huang, 2018]

- NLI(Natural Language Inference) task 에서 뛰어난 성능을 보인 ESIM 모델을 Response Selection에서 사용함으로써 대화 및 응답 발화의 관련 정보들을 반영
- ESIM 기반의 모델의 경우, 두 문장의 관련 단어 들의 정보를 반영하지만 대화(Dialog Context)가 Utterance(발화) 단위가 아닌 전체 하나의 문장 단위로 들어가게 됨 [대화 내 발화의 구분은 __eot__(end of turn) 태그를 사용함]
- 본 모델에서도 대화 발화 turn별로 고려하지 않고 대화 전체를 하나의 입력으로 받기 때문에 대화 또는 각 발화들의 길이가 길어질 경우 정보 손실이 발생할 수 있음

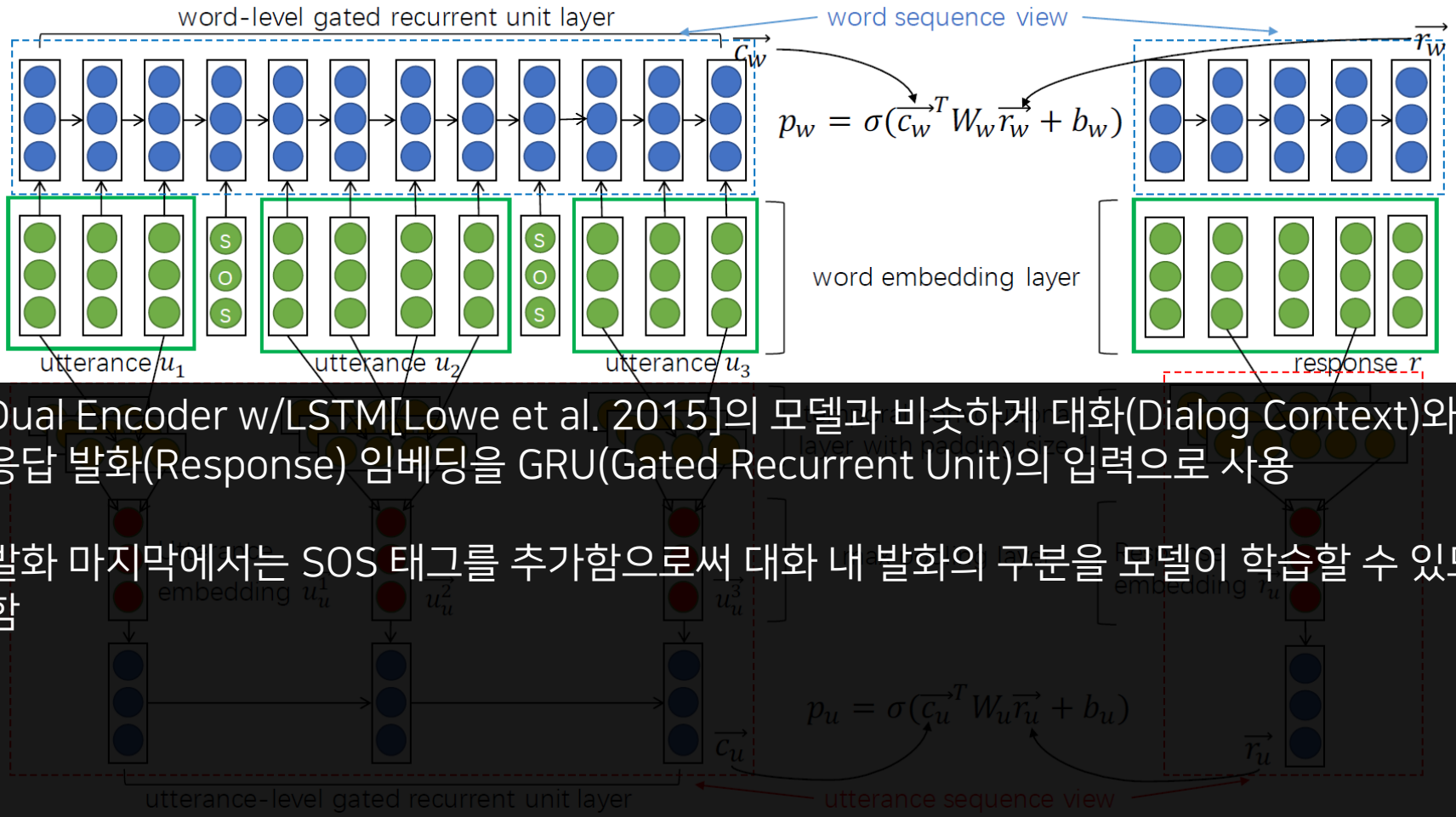


연구 내용 – Multi-view [Zhou et al, 2016]



Zhou et al. Multi-view Response Selection for Human-Computer Conversation. In Proceedings of EMNLP 2016.

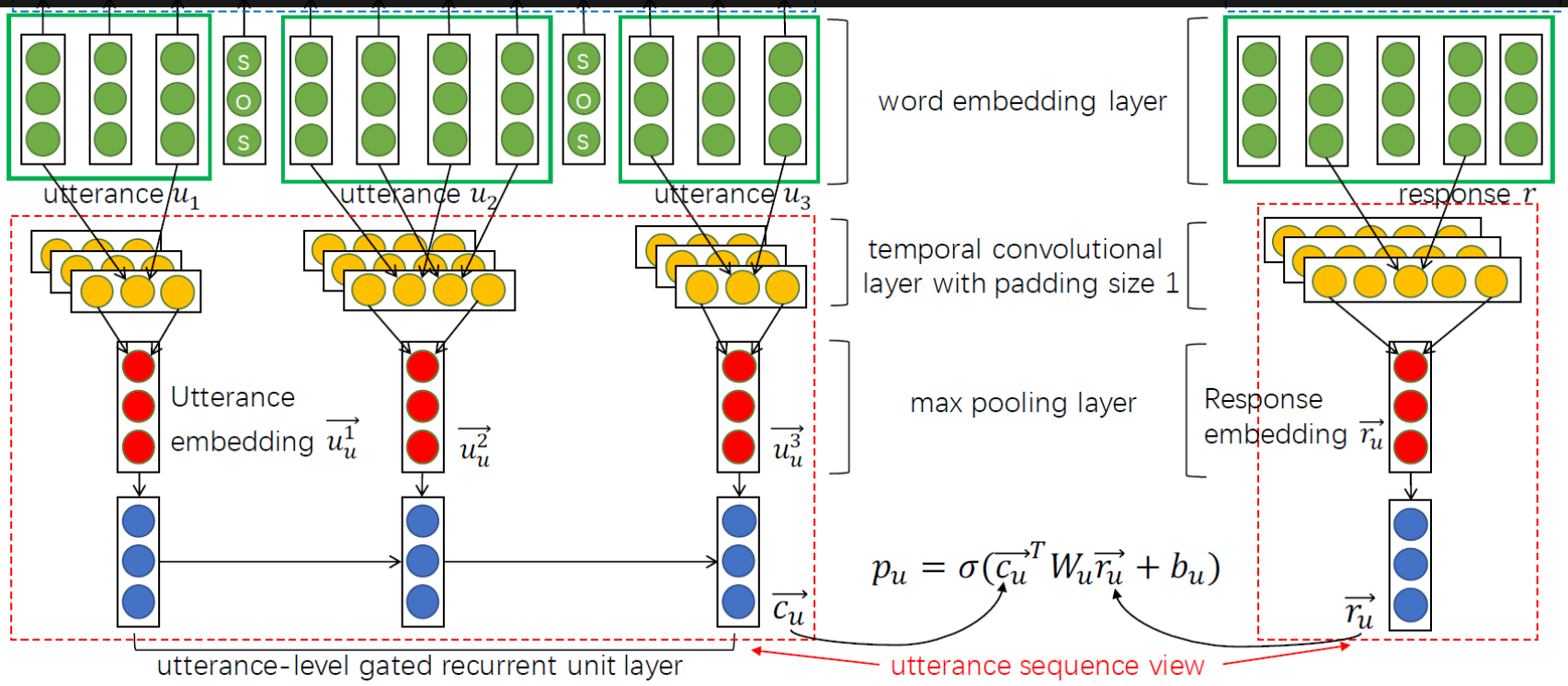
연구 내용 – Multi-view [Zhou et al, 2016]



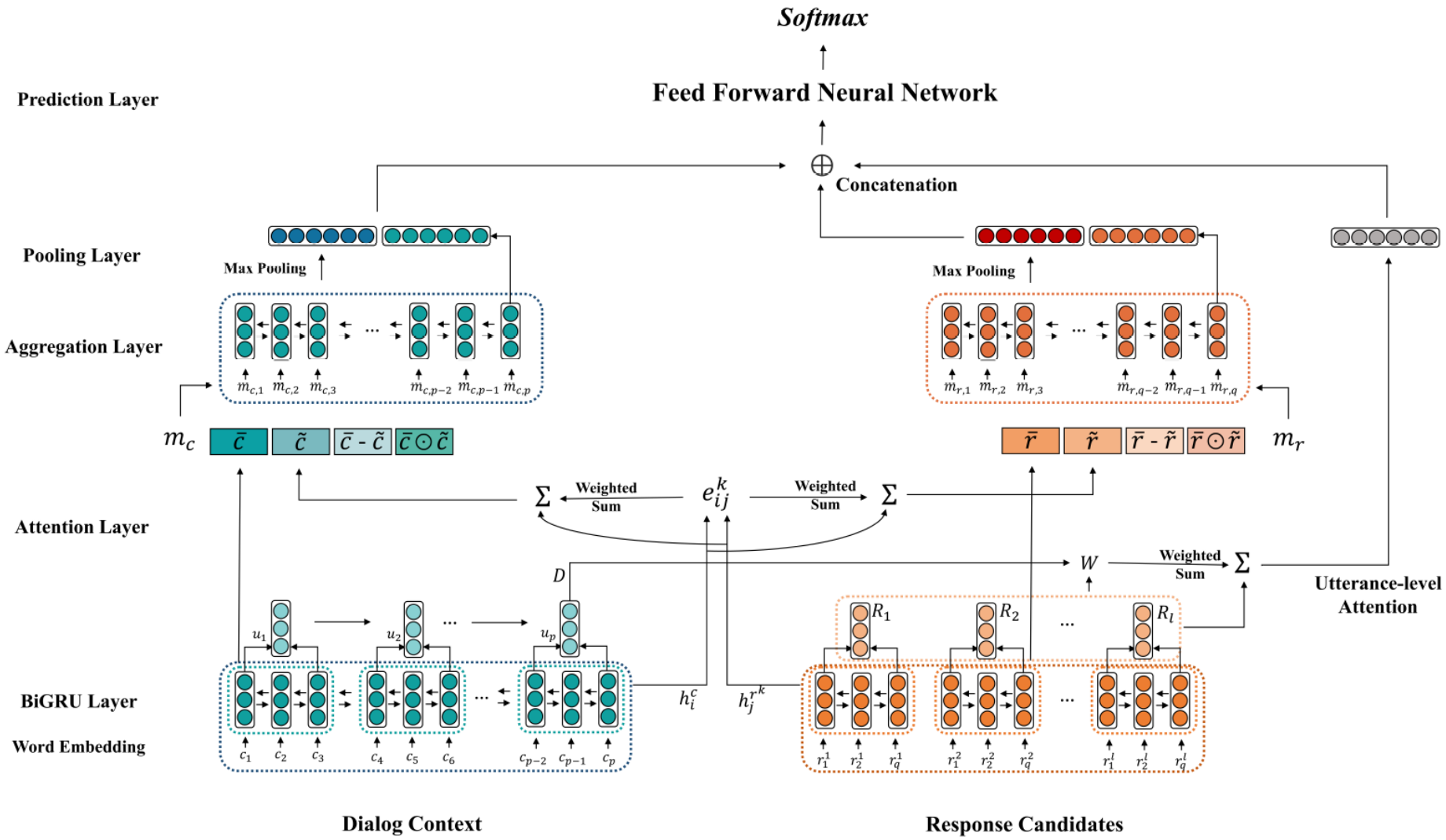
- Dual Encoder w/LSTM[Lowe et al. 2015]의 모델과 비슷하게 대화(Dialog Context)와 응답 발화(Response) 임베딩을 GRU(Gated Recurrent Unit)의 입력으로 사용
- 발화 마지막에서는 SOS 태그를 추가함으로써 대화 내 발화의 구분을 모델이 학습할 수 있도록 함

연구 내용 – Multi-view [Zhou et al, 2016]

- 대화 내 발화(Utterance) 단위로 학습을 진행하기 위해서 CNN을 사용하였음. 각 발화의 Convolved feature들을 max pooling하여 GRU의 입력으로 사용. Multi-view 모델에서는 대화의 단어, 발화 단위 모두를 학습할 수 있게끔 모델을 설계
- 하지만 대화와 응답 발화의 상관 관계를 모델 학습에 충분히 반영할 수 없는 한계가 존재



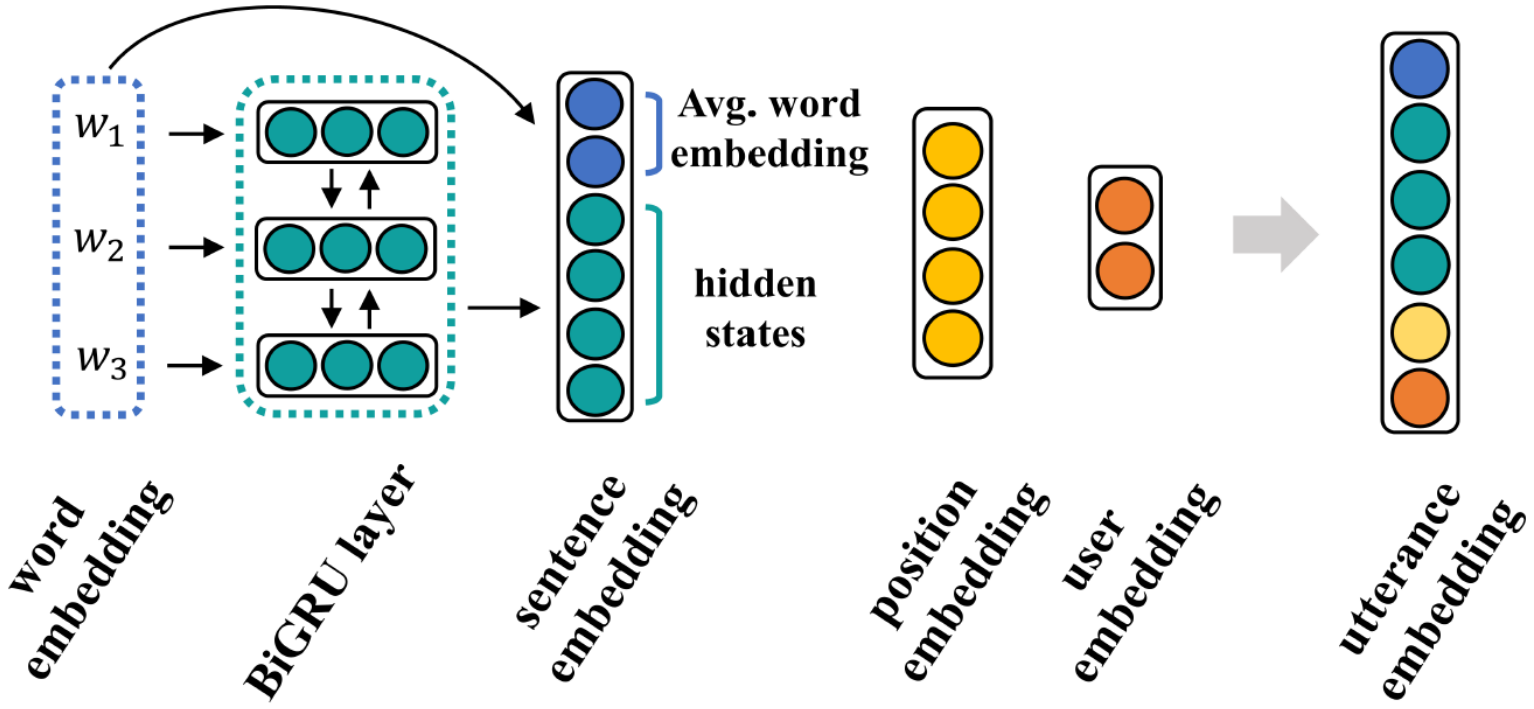
연구 내용 - ESIM w/Utterance-level Attention



연구 내용 - Utterance Representation

- Sentence Embedding
  - 발화 내 단어 임베딩의 평균(Avg. word embeddings)과 bidirectional GRU의 hidden states를 결합하여 이를 발화 문장 임베딩으로 사용
- Position Embedding
  - 대화 내에서 각 발화(utterance)들의 위치 정보는 마지막 발화를 예측하는 데에 중요한 정보로 쓰일 수 있음. 발화의 position 임베딩은 Vaswani et al. (2017)에서 제안했던 방식과 동일하게 생성함
  - 기존 연구에서 position 임베딩을 입력에 더한 것과 달리 본 연구에서는 문장 임베딩과 결합하여 사용함
- User Embedding
  - 두 사용자의 multi-turn 대화 내에서는 각자의 특징 및 대화 목적이 존재함. 각 사용자들을 one-hot encoding으로 벡터화하여 utterance feature로 사용(e.g. advisor:[1,0], student[0,1])

연구 내용 - Utterance Representation



연구 내용 – Experiments

- Datasets
  - Dialog System Technology Challenges 7(DSTC7)
    - Ubuntu IRC (Internet Relay Chat) Dialog Corpus
- Ubuntu 시스템 내에서 발생하는 문제에 대해 질문 및 답변하는 대화 코퍼스
  - Advising (Flex) Dataset
    - 학생이 다음 학기에 어떤 수업을 들을 지 Course Adviser에게 조언을 구하는 대화 데이터
- Task Description
  - Task 1 : 총 100개의 응답 발화 후보 중 1개의 정답을 고르는 것
  - Task 4 : Task 1가 동일하지만, 정답이 없는 경우도 있음(NONE)

연구 내용 - Experiments

Task	Dataset	Ubuntu			Advising		
		#dialogs	#tokens	#Avg. turns per dialog	#dialogs	#tokens	#Avg. turns per dialog
Task 1	Training	100000	221112	5.490	100000	5302	6.985
	Develop.	5000	117853	5.436	500	2433	7.234
	Test1	1000	49788	5.588	500	2494	7.232
	Test2	-	-	-	500	2332	7.166
Task4	Training	100000	220999	5.446	100000	5304	6.985
	Develop.	5000	118229	5.431	500	2415	7.234
	Test1	1000	49772	5.588	500	2466	7.232
	Test2	-	-	-	500	2316	7.166

연구 내용 - Negative Sampling

- 각 대화(Dialog Context)마다 1개의 positive 응답과 99개의 negative 응답들을 포함하는 100개의 정답 후보들이 존재. 하지만 학습 시 매번 100개 중 하나의 적합한 정답을 예측하는 것은 오랜 학습 시간이 소요됨

- 따라서, 정답 후보 내의 negative 응답들을 랜덤하게 샘플링을 진행하여 학습에는 축소된 정답 후보들을 사용함. Negative 샘플의 수의 경우, 10에서부터 30까지 5개 단위로 실험을 진행한 결과 20개를 사용하였을 때 가장 좋은 성능을 보여줌

- 전체 100개를 사용하였을 때보다 축소된 negative 샘플을 사용하였을 때 약 2~3% 좋은 성능을 보여주었으며, 매 epoch 마다 다른 negative 샘플을 사용하였을 때 5~6% 좋은 성능을 보여줌

연구 내용 - Evaluation

- Baselines
  - Dual Encoder w/LSTM(Lowe et al. 2015)
    - 단어 단위에서 대화와 정답 후보들의 sequence 정보만을 사용
  - ESIM(Dong and Huang 2018)
    - Context-aware, response-aware Attention을 사용함으로써 대화와 정답 후보들과의 상관 관계를 반영하여 학습을 진행
- 제안된 모델
  - ESIM + SE
    - Sentence embedding(문장 임베딩)을 생성하여 단어 단위 ESIM 모델과 함께 사용
  - ESIM + SE + PE + UE
    - Utterance embedding(발화 임베딩)을 sentence, position, user에 대한 세 가지 feature로 구성하여 단어 단위 ESIM 모델과 함께 사용

연구 내용 - Experiments

Task 1	Ubuntu						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	0.320	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	0.294	0.430	0.834	0.204
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	**0.505**	0.106	0.160	**0.306**	**0.460**	0.858	0.213

(a) Model performance for task 1

Task 4	Ubuntu						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)	0.215	0.328	0.500	0.622	0.925	0.351	0.162	0.200	0.262	0.346	0.736	0.228
(Dong and Huang 2018)	0.378	0.507	0.634	0.717	0.931	0.500	0.156	0.242	0.400	0.546	0.888	0.278
ESIM + SE (GloVe)	0.386	0.512	0.636	0.722	0.939	0.507	0.170	**0.278**	0.406	0.536	0.894	0.293
ESIM + SE + PE + UE (GloVe)	0.394	0.515	0.651	0.728	**0.941**	0.512	**0.190**	0.272	**0.424**	**0.578**	**0.908**	**0.311**
ESIM + SE + PE + UE (ELMo)	**0.417**	**0.542**	**0.662**	**0.735**	0.938	**0.533**	0.186	0.266	0.386	0.498	0.854	0.291

(b) Model performance for task 4

연구 내용 - Results

Task	Dataset	Submitted Results				Improved Results			
		R@1	R@10	R@50	MRR	R@1	R@10	R@50	MRR
Task 1	Ubuntu	0.471	0.700	0.926	0.5570	-	-	-	-
	Advising case1	0.122	0.474	0.868	0.2337	0.290	0.614	0.898	0.4019
	Advising case2	0.064	0.398	0.800	0.1611	0.106	0.436	0.826	0.2071
Task4	Ubuntu	0.487	0.772	0.936	0.5932	-	-	-	-
	Advising case1	0.224	0.552	0.896	0.3276	0.340	0.782	0.974	0.4926
	Advising case2	0.178	0.470	0.856	0.2703	0.196	0.544	0.898	0.3154

- Batch Shuffling on Advising Dataset
  - 기존 실험에서 Advising 데이터에 대해 학습이 잘 진행되지 않았음
  - Response Selection 데이터 구축 시 하나의 대화에서 turn 별로 데이터 셋을 구축하여 대화 데이터 분포가 일정하지 않아 모델 학습이 진행되지 않는 문제가 발생
  - 총 100,000개의 데이터 셋에 대한 Batch Shuffling을 진행하여 위의 문제를 해결하였음

Answer Generation from Span Prediction in MRC Tasks

고려대학교 NLP&AI Lab.

양기수

연구 내용 - Span Prediction in MRC Tasks

Passage Sentence

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity.

Question

What causes precipitation to fall?

Answer Candidate

gravity

Example of Span Prediction in MRC Task

연구 내용 - Answer Generation Model

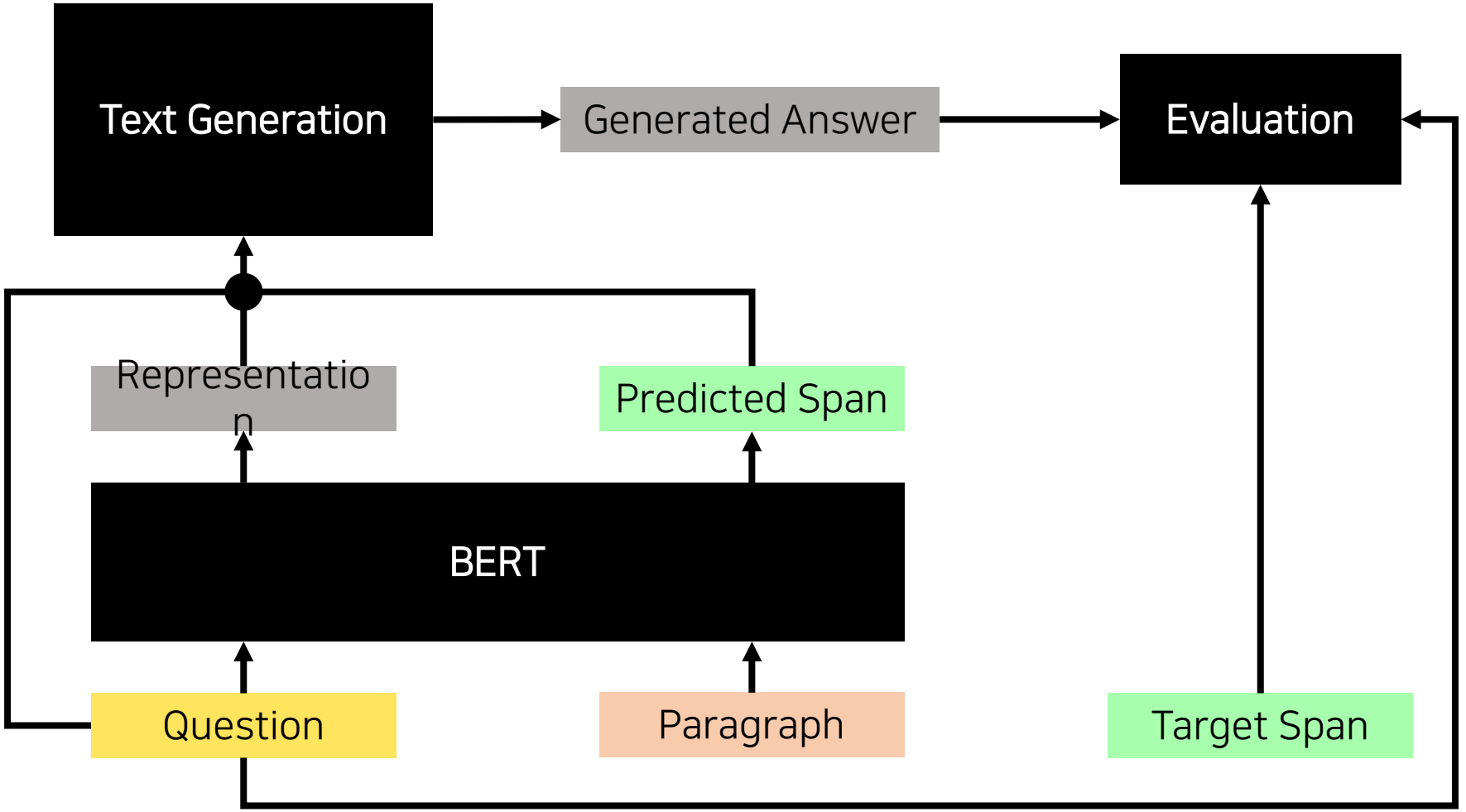


Figure of Answer Generation Model

연구 내용 - Related Works

BERT

1. Inputs: [CLS] + Question + [SEP] + Paragraph
2. Encoding the Inputs through Pre-trained BERT Model
3. Multiple Fine-tuners
   - First of all, train the span prediction fine-tuner, then freeze its variables.
   - Second, train the answer generation fine-tuner
4. Related Experiments
   - Text Classification with **Bidirectional Encoder Representations from Transformers(BERT)** in a Korean Language Classification Task

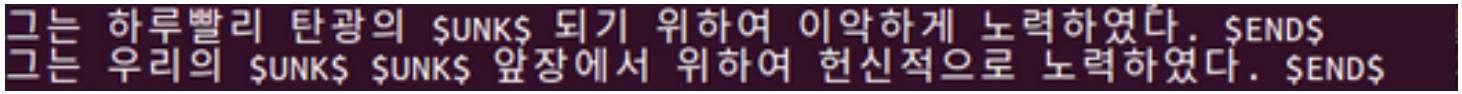
s	합성재를 쓰는 대부분의 화가가 건강상의 문제나 불편을 호소하기 때문이다.
n	온 세계의 자주화를 실현하기 위한 투쟁의 지침을 마련해주신 그이께 가장 숭고한 경의를 드린다.
n	버럭더미가 있던 곳에 성토를 하고 꾸린 야외휴식터는 그야말로 한폭의 그림을 방불케 하였다.
s	김민하, 하준호, 김대우의 경쟁 구도는 이미 깨진 지 오래다.
n	모두가 이런 생각으로 그이를 우러르는데 위대한 장군님의 우렁우렁하신 음성이 장내를 울리였다.
n	도, 시, 군당위원회를 비롯하여 모든 당조직들이 철도수송문제를 푸는데 큰 관심을 돌려야 하겠습니다.
s	서울은 인천 유나이티드에서 이석현을 영입했다고 2일 공식 발표했다.
s	누구나 서피스2와 함께 요리를 만들 수 있는 것이다.

Korean Language Distinction Dataset

연구 내용 - Related Works

Text Generation

1. Decoder
   - Train to generate answers from hidden representation
2. GAN with RL e.g. SeqGAN(2017), LeakGAN(2017), GAN with IRL(2018)
   - Being actively researched, but there is no placeholder for the hidden representation
3. Related Experiments
   - Style Transfer in Korean Text using **Sequence-to-Sequence(S2S)**
   - Image **Generative Adversarial Networks(GAN)** with Densely Connected Layers
   - "Auto Thinking" Game Development based on **Reinforcement Learning(RL)**



Example of Text Generation using Seq2Seq

연구 내용 - Related Works

Evaluation

1. BLEU score → when comparing **with an original Question**
   - Having a lack when a target text is too short and when similar yet different words are generated.
2. 14 Metrics in『The Evaluation of Sentence Similarity Measures (2018)』→ **with a predicted span**

Sentence Similarity Measures
$sim_{jaccard}$
$sim_{overlap}$
$sim_{overlap,IDF}$
$sim_{overlap,phrase}$
$sim_{TF,vector}$
$sim_{TFIDF,vector}$
$sim_{TFIDF,nov}$
$sim_{identity}$
sim_{ssv}
sim_{sem}
$sim_{simsem,IDF}$
sim_{wo}
sim_{ssv+wo}
sim_{sem+wo}

14 Sentence Similarity Measures

Metrics	MLE	SeqGAN	RankGAN	LeakGAN	IRL	Ground Truth
$BLEU_F$-2	0.798	0.821	0.850*	**0.914**	0.829	0.836
$BLEU_F$-3	0.631	0.632	0.672*	**0.816**	0.662	0.672
$BLEU_F$-4	0.498	0.511	0.557*	**0.699**	0.586	0.598
$BLEU_F$-5	0.434	0.439	0.544*	**0.632**	0.542	0.557
$BLEU_B$-2	0.801	0.682	-	0.790	**0.868**	0.869
$BLEU_B$-3	0.622	0.542	-	0.605	**0.718**	0.710
$BLEU_B$-4	0.551	0.513	-	0.549	**0.660**	0.649
$BLEU_B$-5	0.508	0.469	-	0.506	**0.609**	0.601
$BLEU_{HA}$-2	0.799	0.745	-	0.847	**0.848**	0.852
$BLEU_{HA}$-3	0.626	0.584	-	**0.695**	0.689	0.690
$BLEU_{HA}$-4	0.523	0.512	-	0.615	**0.621**	0.622
$BLEU_{HA}$-5	0.468	0.454	-	0.562	**0.574**	0.578

Results on COCO image caption dataset

연구 내용 - Expected Contribution

1. To derive complete sentence through answer generation instead of span prediction
2. To propose multiple fine-tuners of BERT
3. To improve generative models by feeding sentence representation to GAN
4. To generate meaningful text by combining to MRC tasks
5. Answer Data Augmentation (QG is also possible with this model)

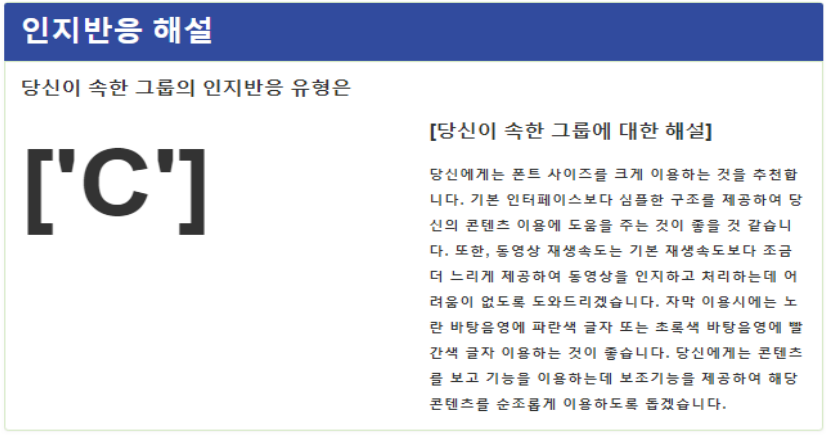
스마트 시니어세대의 문화향유를 위한 인지반응 맞춤형 UI/UX 기술 개발

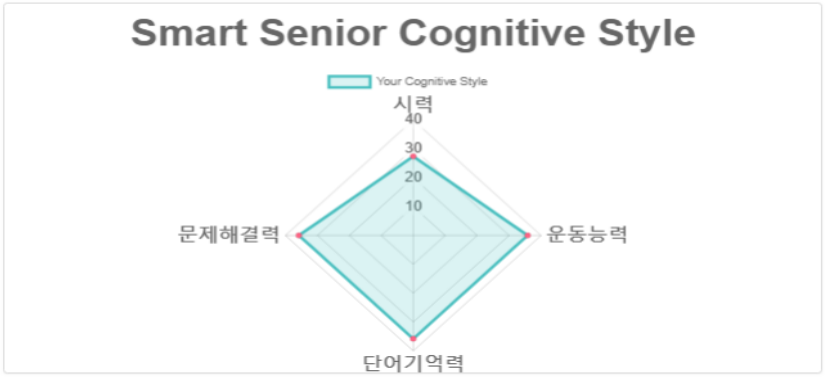
고려대학교 NLP&AI Lab.

- 인지반응 분석 시스템 개발
- 스마트 시니어 맞춤형 프로파일 관리시스템 고도화
- 스마트 시니어 맞춤형 추천 시스템 개발
- 표준안 및 기술 보고서 작성

연구 내용 - 인지반응 분석 시스템 개발

- 인지반응 9종 검사에 따른 시각화 결과 도출

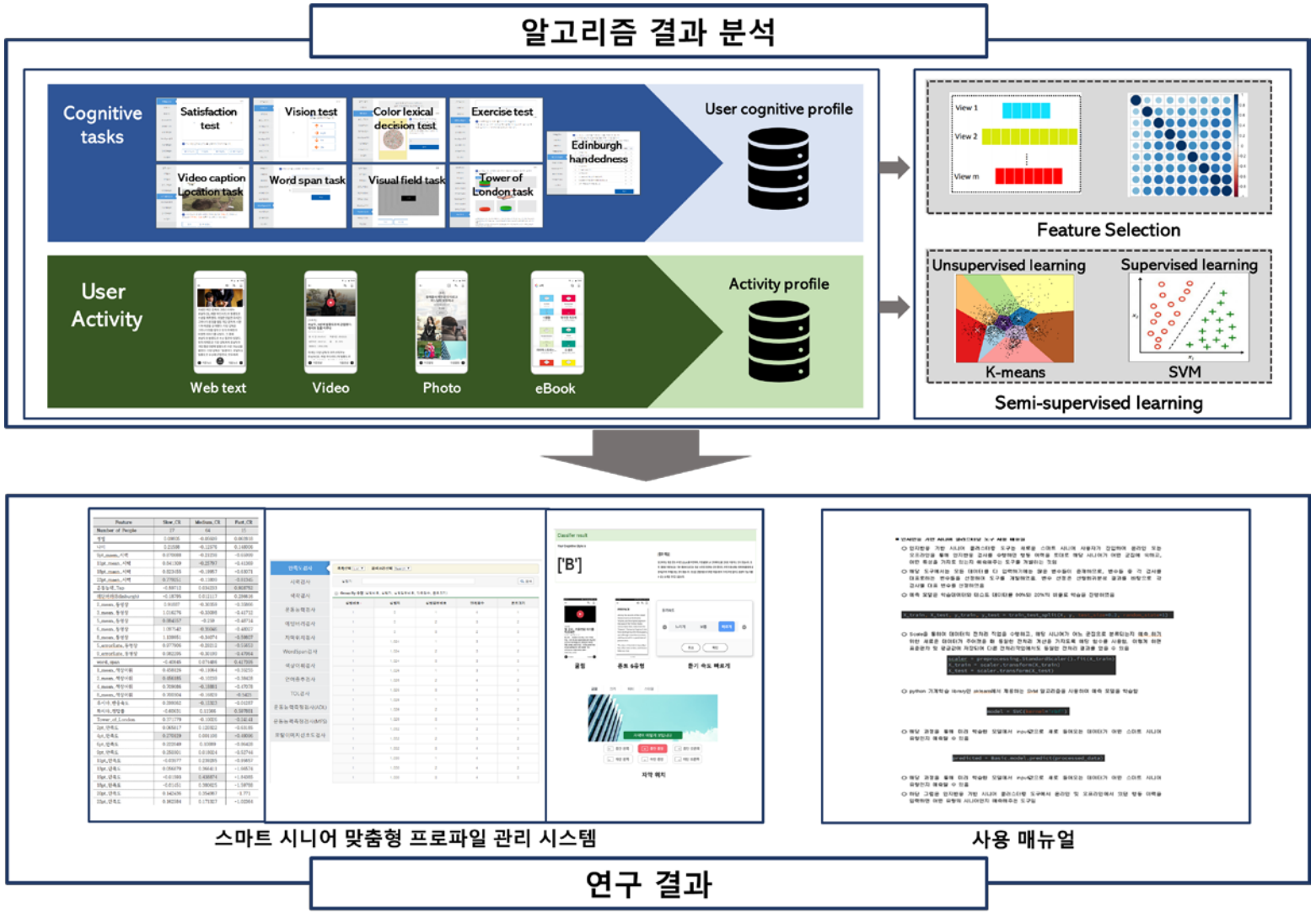




연구 내용 - 스마트 시니어 맞춤형 프로파일 관리시스템 고도화

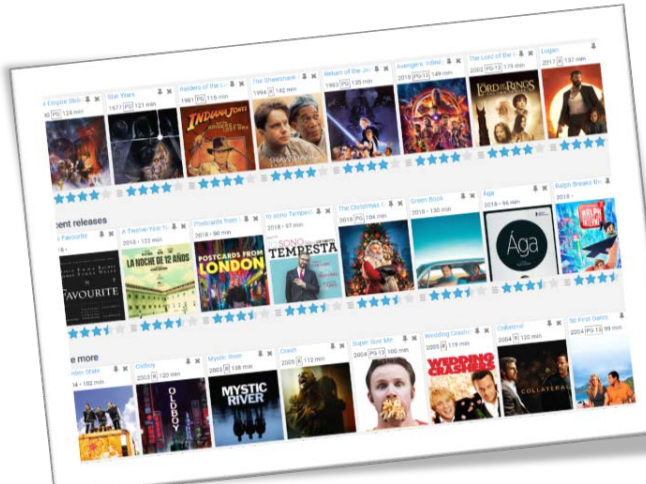
- 스마트 시니어 맞춤형 프로파일 관리시스템은 통합적인 프로파일 관리를 지원하고, 프로파일 분석을 통해 발생한 데이터는 시니어 맞춤형 디지털 콘텐츠 추천, 시니어 맞춤형 콘텐츠 생성 및 최적화를 위해 개발
- Feature selection을 통해 약 180여개의 인지반응 검사지에서 의미 있는 검사지만 추출하여 검사 시간을 단축하고, 집중도를 향상시킴
- 또한 인지반응 검사지 데이터와 온라인 활동 기반 데이터를 기반으로 classification을 통해 새로운 스마트 시니어가 인지반응 검사를 했을 때 정확한 맞춤형 시스템을 제공할 수 있도록 개발함

연구 내용 - 스마트 시니어 맞춤형 프로파일 관리시스템 고도화



연구 내용 - 스마트 시니어 맞춤형 추천 시스템 개발

- Zhang et al. (2017)이 SIGIR에서 발표한 Contractive Auto-encode를 이용한 Collaborative Filtering 모델인 AutoSVD++ 모델을 사용하여 추천 시스템을 개발하였음
- 사용자가 이미 관람한 영화에 대해 1~5 점 사이의 평점을 내리면 추천 모델이 상위 K개의 영화를 사용자에게 추천함. 본 실험에서 사용한 Movielens-latest-small 데이터 셋에 대하여 RMSE 0.887의 성능을 확인
- 데이터 셋에 존재하는 9205개에 해당하는 모든 영화에 대한 포스터 및 줄거리 크롤링을 진행하였으며, 해당 포스터 및 줄거리는 사용자의 인지 능력에 따라 다른 UI/UX로 사용자에게 제공됨



연구 내용 - 표준안 및 기술 보고서 작성

- 1건의 표준이 표준의결 되었으며 기술 보고서 4건 표준 공고됨

구분	제목	현재상태
표준	온라인 콘텐츠 사용자 프로파일 구조	운영위원회 조정
기술 보고서	고령자용 디지털 콘텐츠 서비스 개발을 위한 3D 디지털 콘텐츠 안정성	표준공고
기술 보고서	고령자를 위한 디지털 콘텐츠 서비스 개발을 위한 그래픽 사용자 인터페이스 구성요소 현황	표준공고
기술 보고서	고령자를 위한 동영상 자막 설계 지침	표준공고(예정)
기술 보고서	고령자를 위한 인지반응 측정 방법 및 적용사례	표준공고(예정)

사용자 중심의 지능형 패션 검색 및 맞춤형 코디네이션 제품 개발

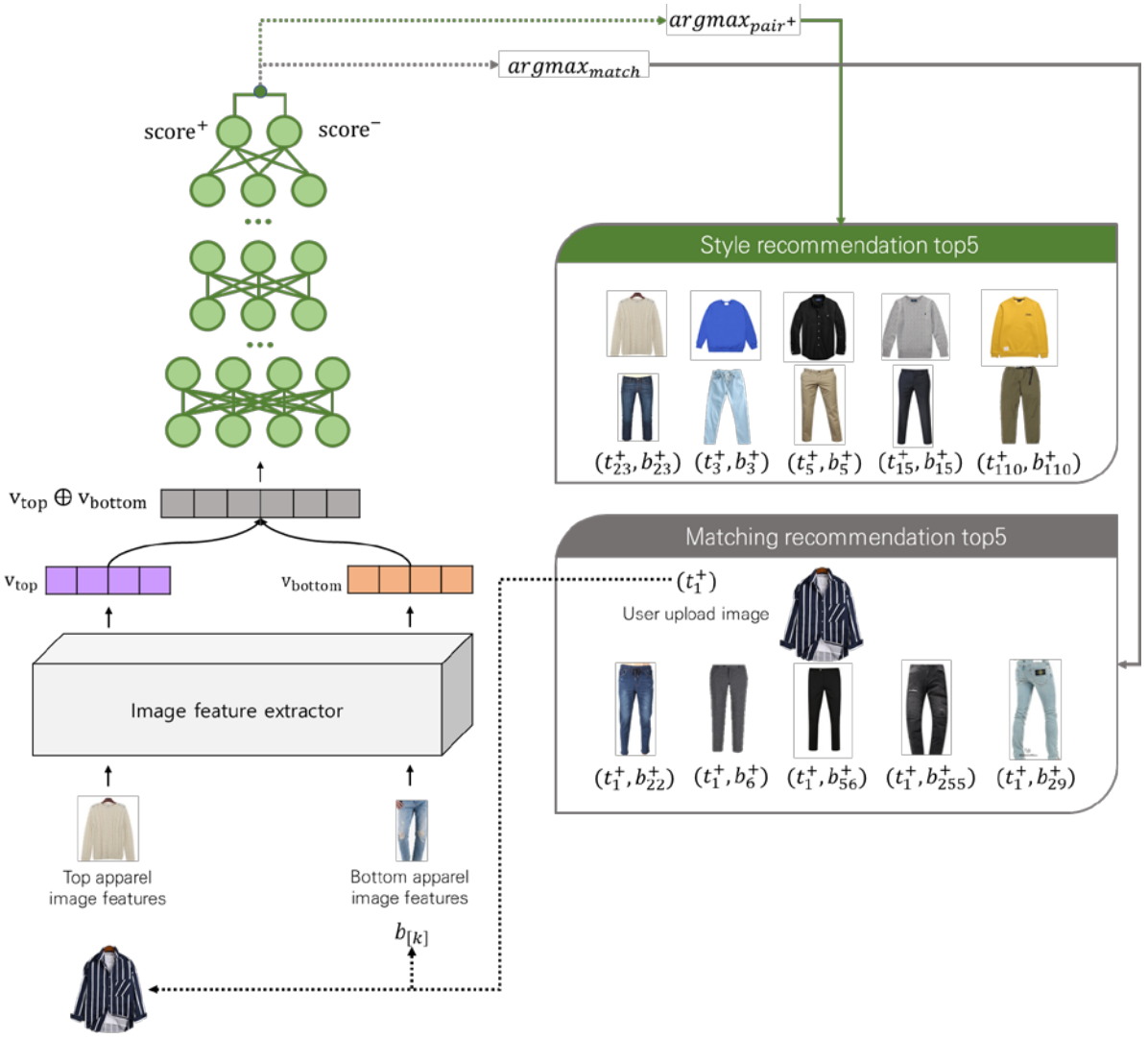
고려대학교 NLP&AI Lab.

- 벡터기반 사용자 선호 패션 프로파일링· 코디 및 매칭 기술 개발
- 드로잉 자질 추출 기술
- 딥러닝 기반의 패션 상품 유사 이미지 검색 모델

연구 내용 - 벡터기반 사용자 선호 패션 프로파일링· 코디 및 매칭 기술 개발

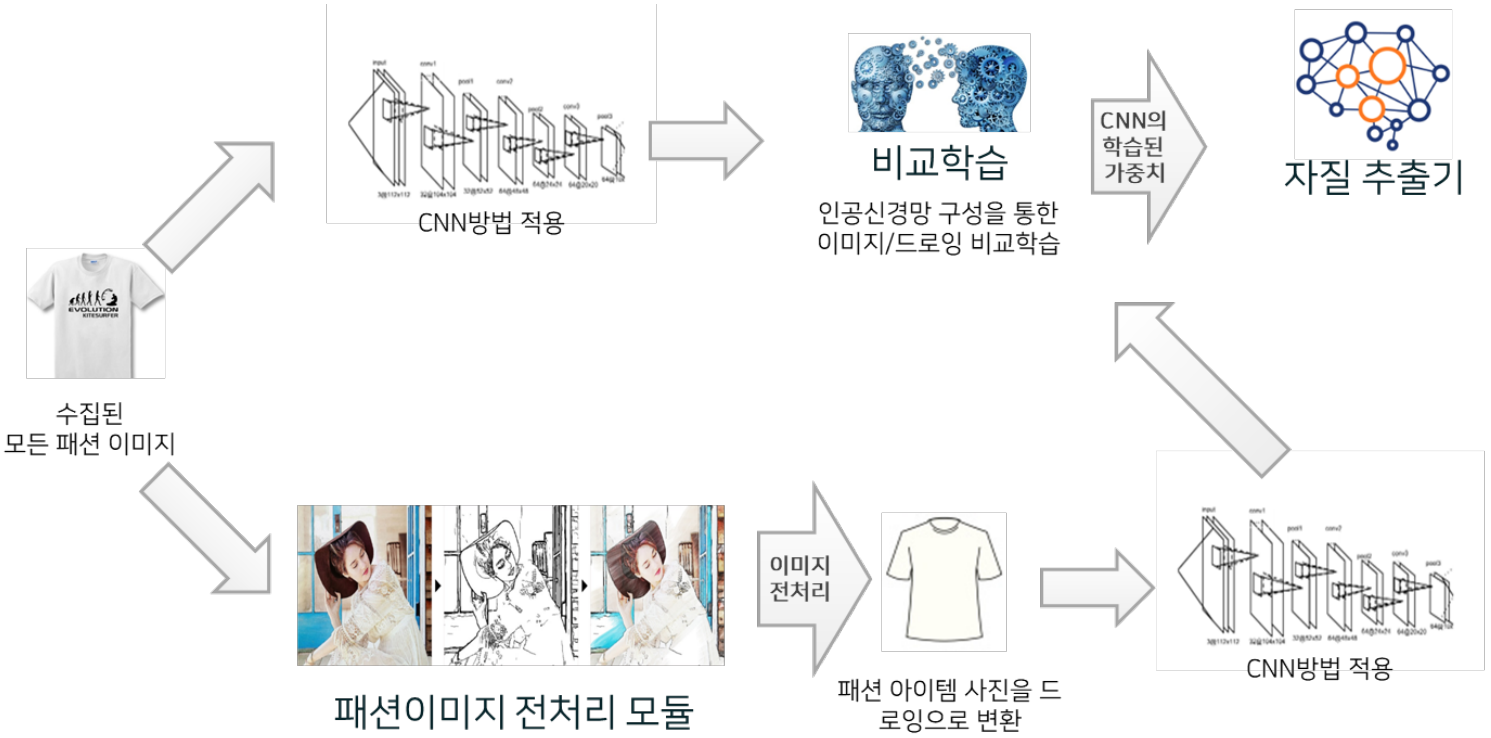
❖ **패션 추천 기술**

- 코디 추천
  - 상의 및 하의
  - 계절감
- 매칭 추천
  - 어울리는 하의 추천
  - 어울리는 상의 추천
  - 계절감



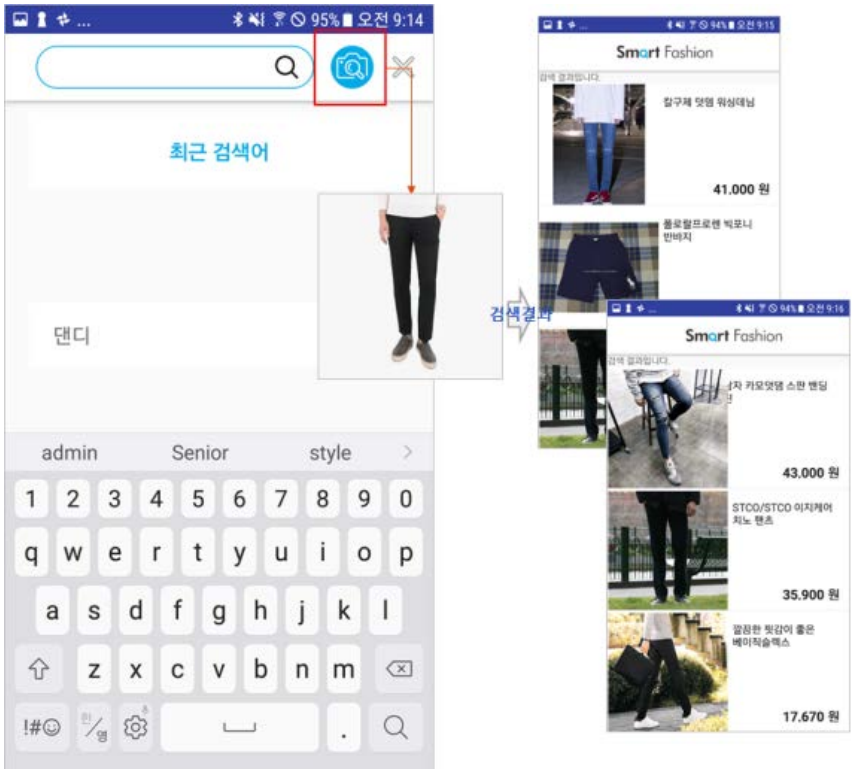
연구 내용 - 드로잉 자질 추출 기술

- CNN & FFNN 방법을 활용하여 드로잉 이미지 자질 추출모듈 개발
  - 패션 아이템 사진들을 이미지 전처리 과정을 통해 드로잉 이미지로 변환시켜 학습 데이터로 사용



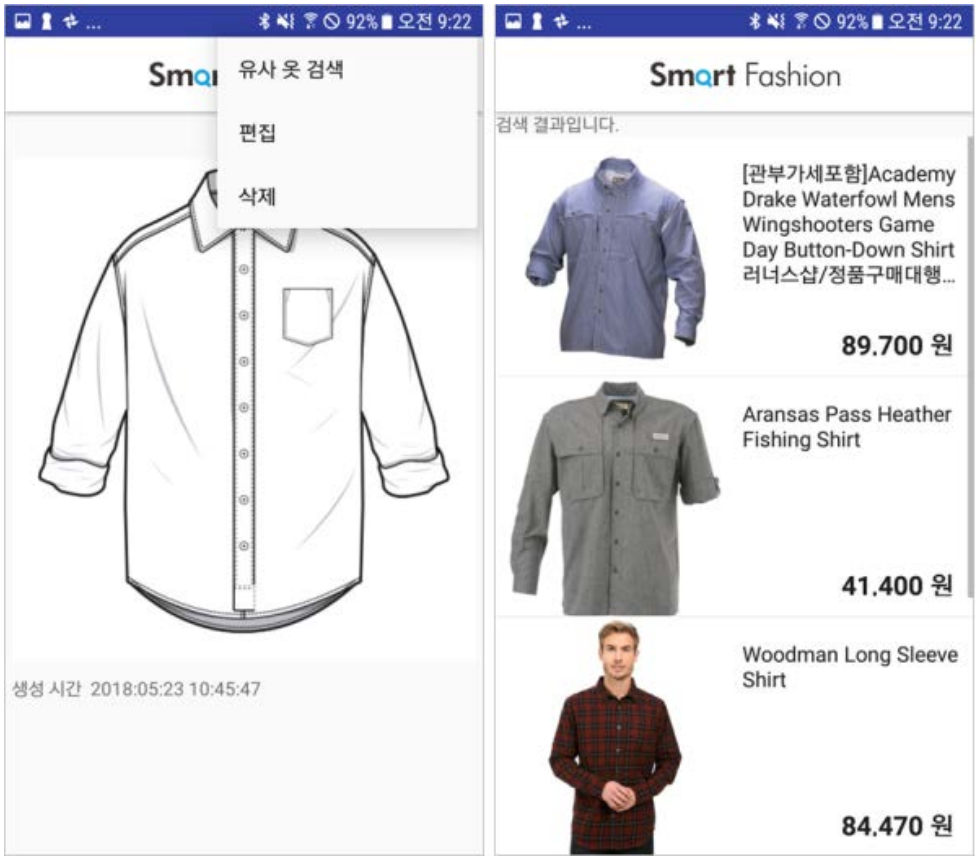
연구 내용 - 딥러닝 기반의 패션 상품 유사 이미지 검색 모델

- CNN 모델을 이용하여 패션 상품 이미지의 자질을 추출한 후 이를 기반으로한 유사 이미지 검색 모델 개발
- 이미지의 유사도는 색, 모양, 패턴 등을 고려하여 결정될 수 있음



연구 내용 - 딥러닝 기반의 패션 상품 유사 이미지 검색 모델

- 패션 상품 드로잉으로부터 업샘플링된 패션 상품 이미지를 이용한 유사 이미지 검색 모델 개발



전통문화 융복합 지원을 위한 지능형 검색 플랫폼 구축

고려대학교 NLP&AI Lab.

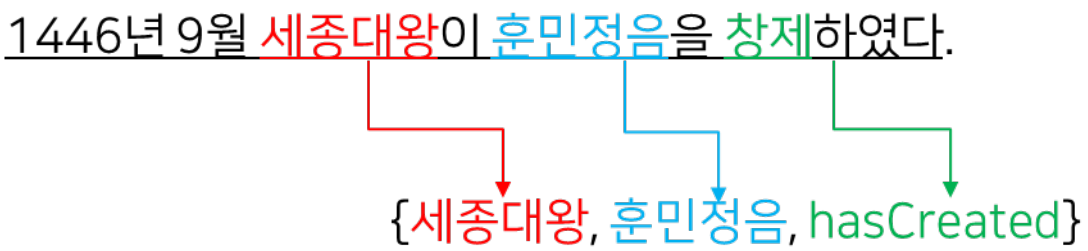
- Meta Learning based Global Relation Extraction
- 고유명사 개체인식 기술 개발 추가 연구 – 데이터 구축

Meta Learning based Global Relation Extraction

고려대학교 NLP&AI Lab.

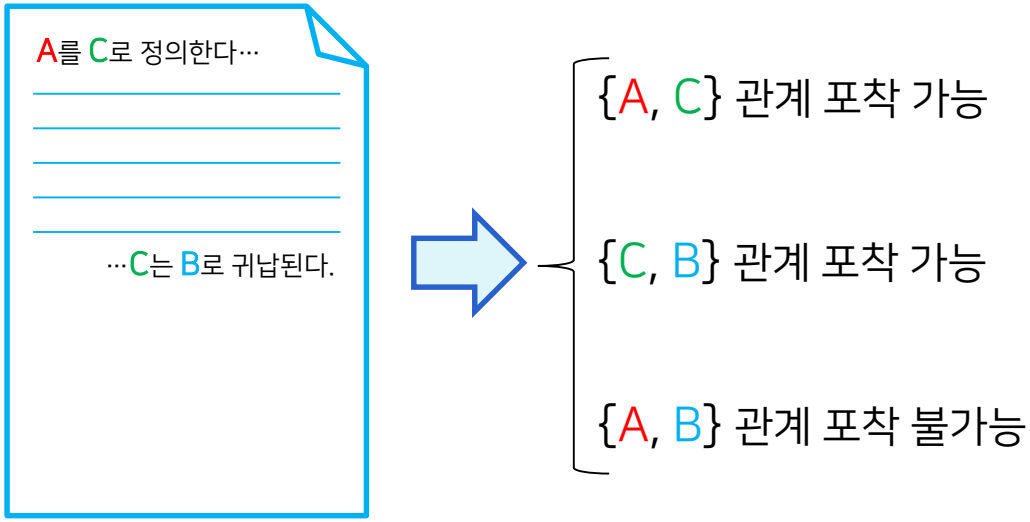
연구 내용 - 시멘틱 분석 기법

- 시멘틱 분석 기법에 사용될 기술 선정:
  - 문선 내 개체간 시멘틱 분석 기법으로 Relation Extraction을 선택함
- 관계 추출(Relation Extraction)이란?
  - 텍스트 내의 개체간 상호작용들을 고정된 시멘틱 관계 유형으로써 표현하는 것
- 텍스트 내 개체 간 관계 :
  - 텍스트의 의미는 텍스트 내에 개체를 정의하고, 각 개체간의 상호작용을 통해 부여된다.
  - 이러한 상호작용들은 다양한 형태의 텍스트로 표현되며, 그 방식과 표기는 일정하지 않다.



연구 내용 - 시멘틱 분석 기법

- 문제점
  - 장문의 텍스트를 처리할 경우, 여러 문장에 걸쳐 표현되는 문서 내 개체 간의 관계를 놓친다.
  - 이는 주어 및 목적어의 생략이 잦은 한국어의 경우 이 문제점이 더 크게 두드러진다.

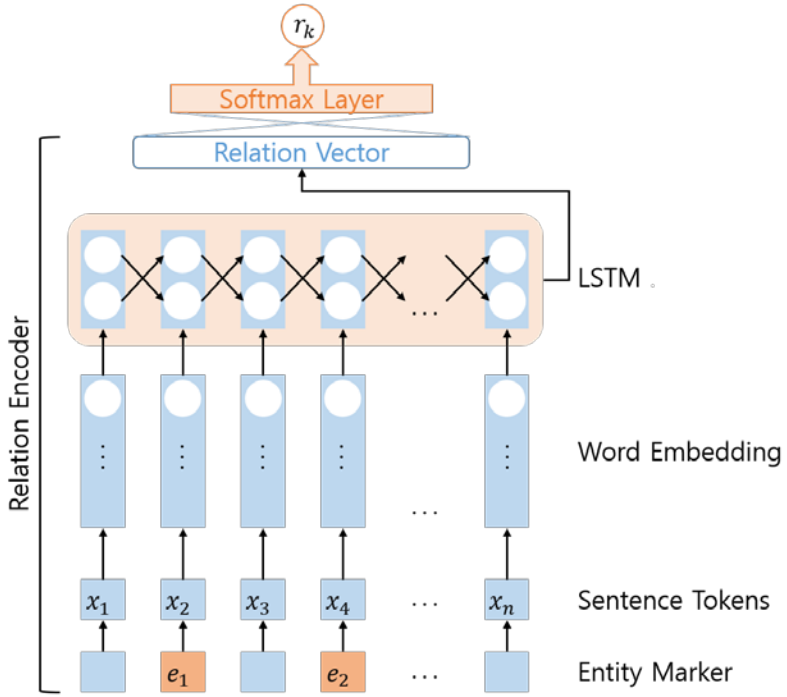


- 위 문제를 해결하기 위하여 외부 메모리를 사용하는 RE 모델을 제시한다.

모델 설명 - Initial Relation Extraction

- 관계 분석 이전 전처리
  - 개체명 인식기를 통한 문서 내 개체 파악함

- 단편적 관계 분석
  - 문서 내에 존재하는 관계를 문장 단위로 분석 및 추출
  - 문장 내에 관계를 정의할 개체 1, 2 (e_1,e_2)을 표시함
  - 문장 x 내의 토큰 x_1, x_2, ⋯,x_n들을 벡터에 임베딩
  - 토큰 임베딩들은 LSTM 신경망을 거쳐 Relation 벡터로 변환됨
  - Relation 벡터를 통하여 (e_1,e_2)간 관계 여부 및 관계 분류를 판별함

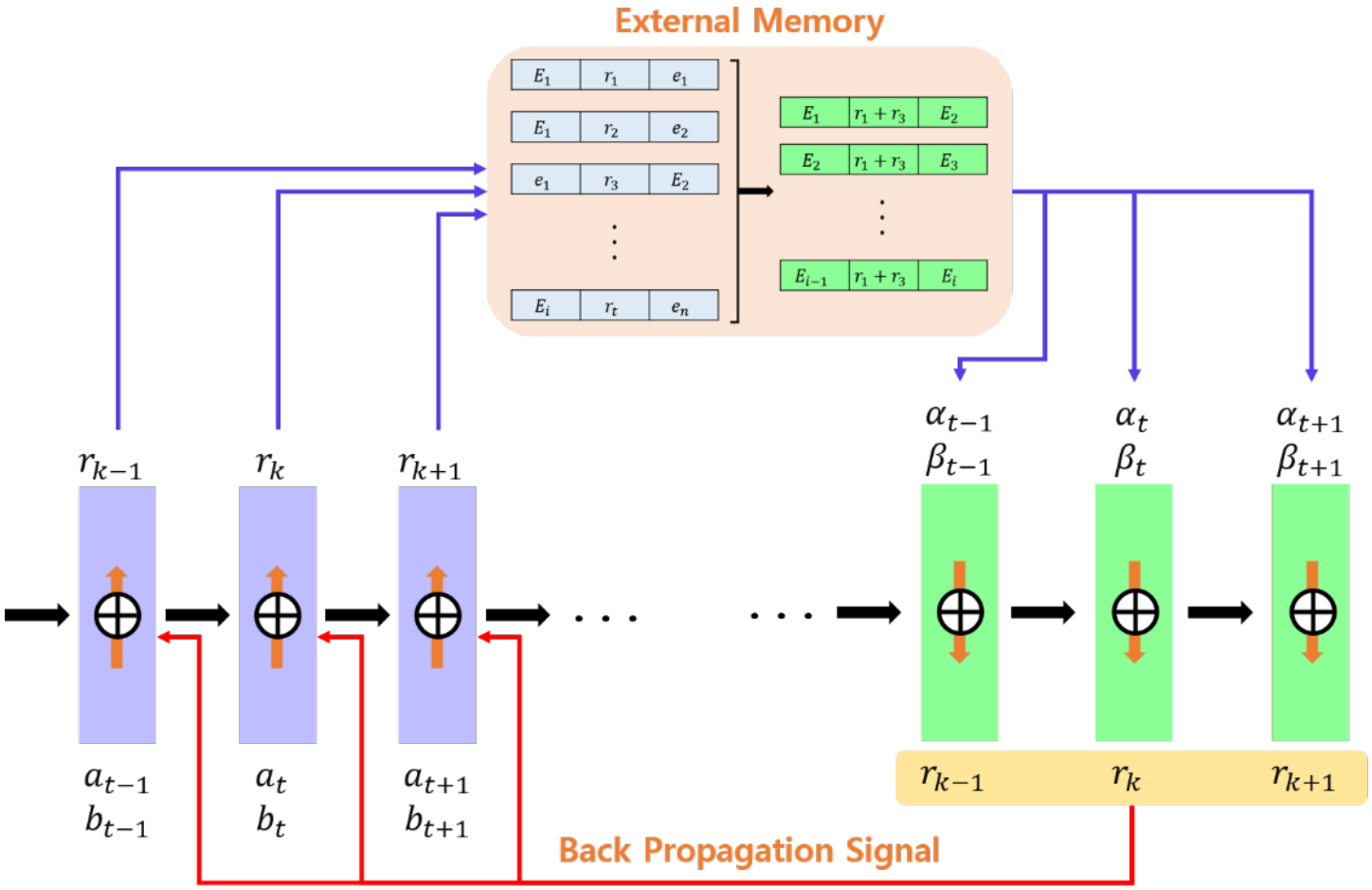


모델 설명 - Augmented External Memory

- 문제점
  - 단편적 관계 추출 모델에서 얻어낸 결과들은 그러나 어디까지나 언급 수준의 관계들
  - 여러 문장에 걸쳐져 있는 관계를 가진 개체들을 분석하지 못함

- 외부 메모리 사용
  - 1차적으로 외부 메모리에 텍스트에서 얻어낸 언급 수준의 관계들의 출력벡터를 차례대로 저장하며 문장을 건너뛰는 관계 후보들을 검색함
  - 각 개체 사이에 공유되는 중간 개체들이 있는 관계들을 모두 통합하여 새로운 예측 벡터를 생성
  - 생성된 예측 벡터는 훈련된 메모리 증강 신경망을 통해 개체들 간의 전역 관계들을 정의함

모델 설명 - Augmented External Memory



모델 평가 - Experiments & Results

❖ 평가방법

- 전역적 관계 추출의 정확도를 평가하기 위하여 타 모델들과 함께 문서 단위의 평가 데이터로부터 관계 추출을 실행함
- 단편적 관계 추출과 전역적 관계 추출의 Precision, Recall 그리고 F1 Score을 비교함

	[A]	[B]	[C]	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

결론

- 모델의 목적
  - 문서 내의 여러 문장을 통해 관계가 형성되는 개체들의 관계도 추출하려고 하는 RE 모델의 개발
  - 문장 내에 주어, 목적어의 생략이 잦은 한글 텍스트에 효과적인 RE 모델의 개발
- Contribution
  - 장문의 텍스트에 대한 심도 깊은 분석을 제시하는 것
  - 자연어로 이루어진 문서에서는 단 하나의 문장 내에서 성립되는 관계보다는 여러 문장에 걸쳐 표현되어지는 관계가 해당 문서 내의 주제를 더 잘 반영함을 반영하는 모델 제시
- Future Improvements:
  - 대량의 데이터를 2가지 형태로 요구하므로 이에 대한 완화
  - 오랜 훈련 시간 단축

고유명사 개체인식 기술 개발 추가 연구 - 데이터 구축

고려대학교 NLP&AI Lab.

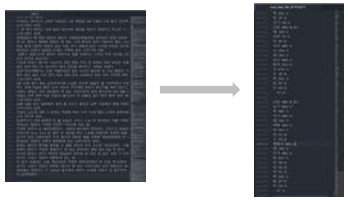
연구 내용 - 전통문화 개체명 인식 시스템 연구

- 개체명 인식기(Named Entity Recognition) 시스템이란?
  - 문장 내에서 고유한 의미를 갖는 명사를 추출하여 인명(PS), 지명(LC), 시간(TI) 등으로 고유명사를 인식하는 시스템

- 2차년도 연구 목표
  - 1차년도에서 부족했던 학습 데이터 추가 구축
  - 기존 모델에 추가 학습하여 자질 보강

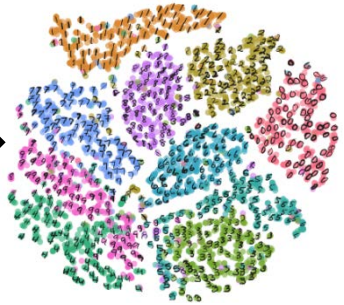
연구 내용 - 전통문화 개체명 인식 시스템 연구

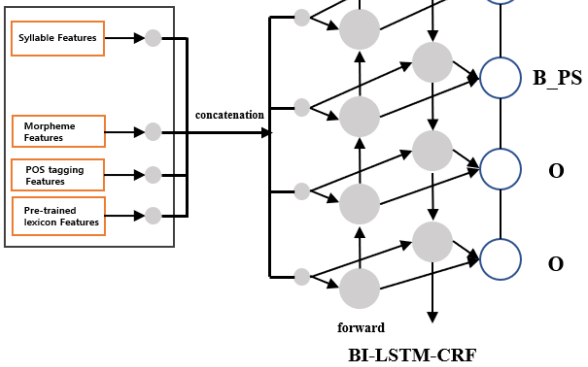
추가 데이터 구축



전통문화 데이터 구축

Contextual Word Representation





입력 문장

형태소 단위 개체명 분석 결과

개체명 인식 결과

태그 결과

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:
{"조선": "B_LC", "세종": "B_PS", "은": "O", "의": "O", "제": "O", "4": "O", "대": "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", "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:
"<조선 세종:PS>은 <조선:LC>의 제4대 군주이며 언어학자이다. 그의 업적에 대한 존경의 의미를 담은 명칭인 <세종:PS>대왕으로 자주 일컬어진다. 성은 이, 휘는 도, 본관은 <전주:LC>, 자는 원정, 아명은 막동이다."

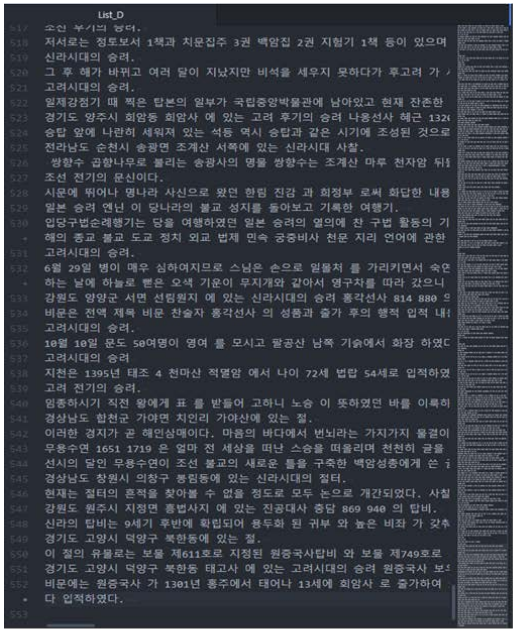
태그 결과:
('조선' : 'LC')
('세종' : 'PS')
('전주' : 'LC')

연구 내용 - 전통문화 개체명 인식 시스템 연구

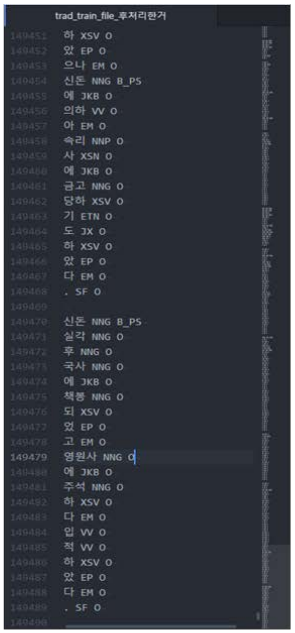
- 한국학 중앙연구원 사이트(http://dh.aks.ac.kr) 데이터 [정의] 및 [내용] 크롤링

연구 내용 - 전통문화 개체명 인식 시스템 연구

- 고려대학교 형태소 분석기를 사용하여 크롤링 데이터 문장 형태소 분석 및 POS tagging
- 지난 5월, 5명이 한 달간 15만여 건의 태깅 작업 진행
- 이후 추가 데이터 training을 위한 문장 전처리 작업 진행



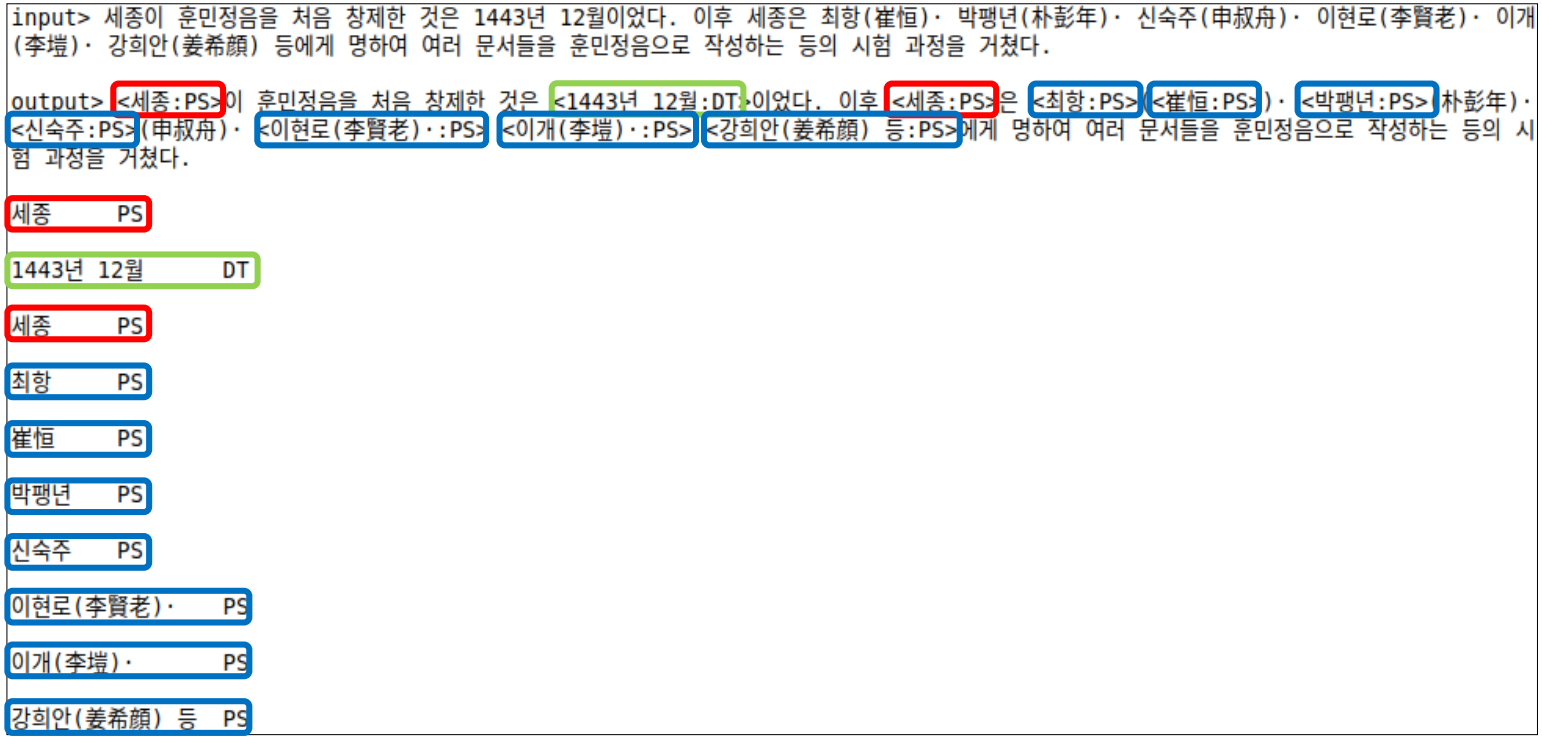
데이터 4702 문장



태깅 데이터 15만 개

연구 내용 - 전통문화 개체명 인식 시스템 연구

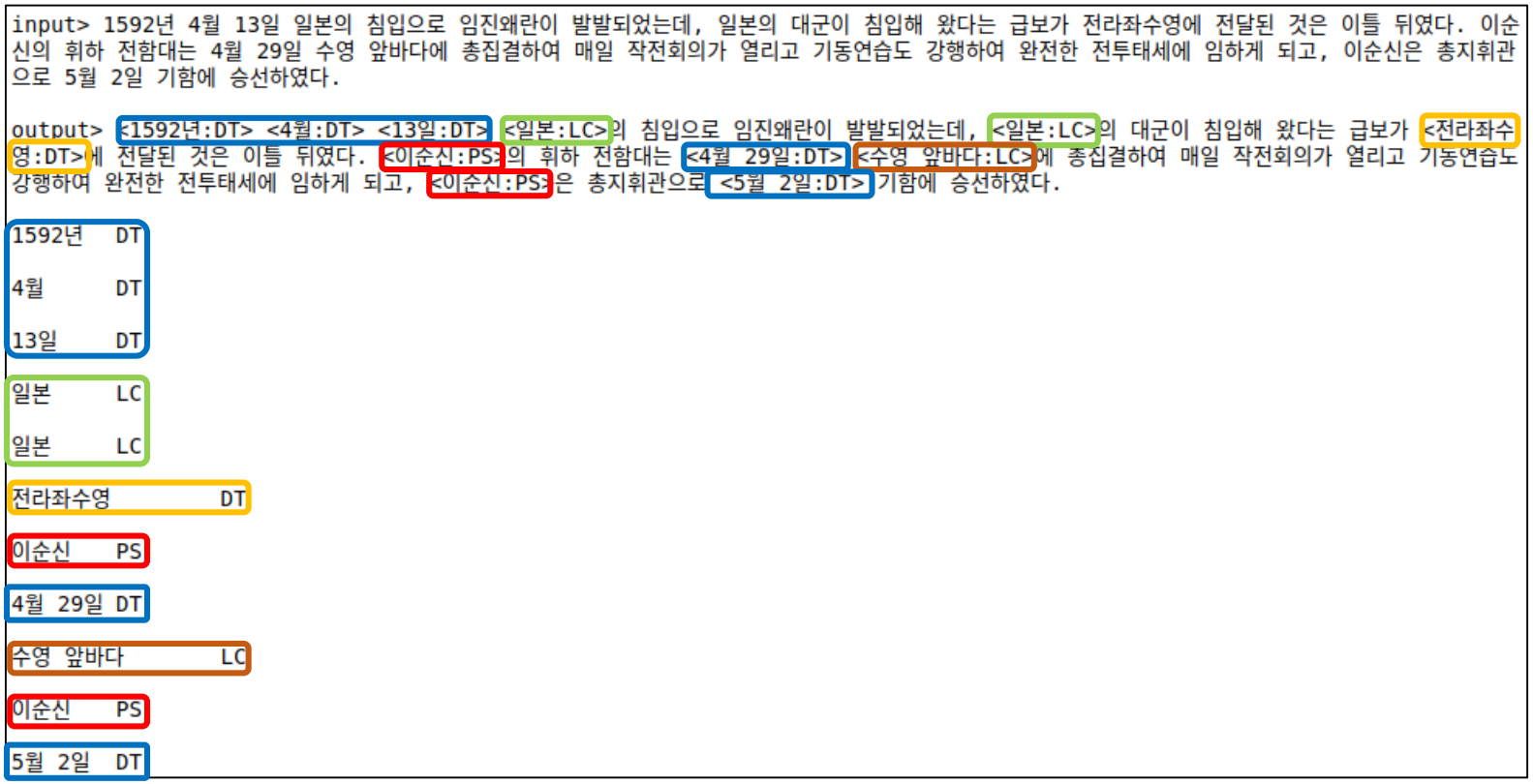
- 기존 4258 문장 + 4702문장 학습 및 결과



세종(PS), 년도(DT), 그 외 인물(PS) 인식

연구 내용 - 전통문화 개체명 인식 시스템 연구

- 기존 4258 문장 + 4702문장 학습 및 결과



장소(LC), 년도(DT), 그 외 인물(PS) 인식

머신러닝을 이용한 비정형 위협정도 자동추출 방안 연구

고려대학교 NLP&AI Lab.

- 비정형 위협정보 자동 인식 및 추출
- 머신러닝을 이용한 문서 자동 요약
- 유사 문서 검색 및 시각화
- Keyword Extraction
- 위협정보 개체명 인식 데이터 구축
- OCR 프로그램 조사 및 성능 비교

연구 목표

선행 기술 연구

국내·외 비정형 빅데이터 분석 사례 조사

기반 기술 연구

딥러닝 기술을 이용한 비정형 위협정보 자동 인식 및 추출 기술 연구

머신러닝 기반 보고서 자동 분석 및 키워드 추출 기술 연구

머신러닝을 이용한 문서 자동 요약 기술 연구

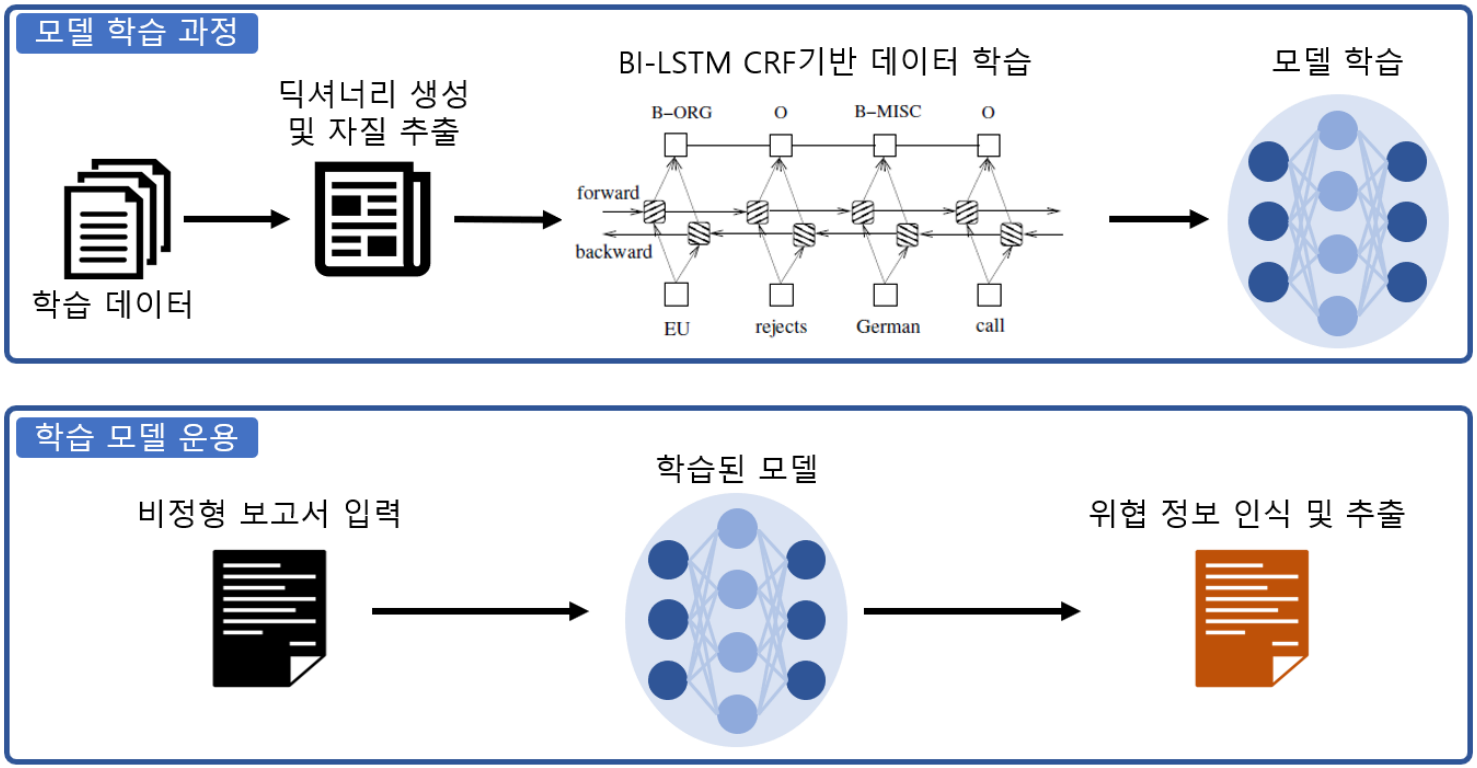
응용 기술 연구

딥러닝을 이용한 문서 토픽 자동 추출 기술 연구

문서 임베딩 벡터 기반 유사 문서 검색 및 시각화 기술 연구

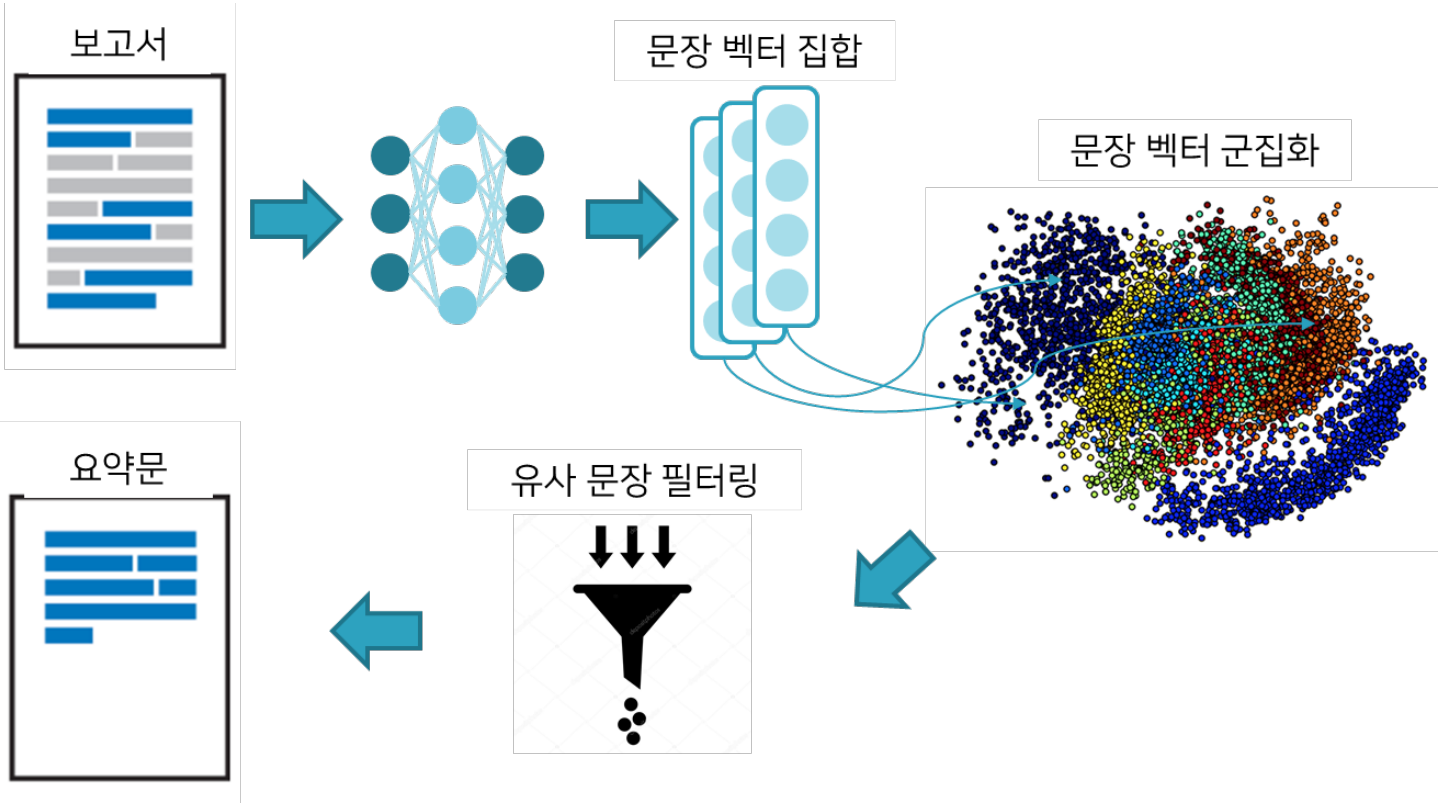
연구 내용 - 비정형 위협정보 자동 인식 및 추출

- BI-LSTM-CRF 기반 모델을 통해 데이터 학습
- 비정형 보고서가 입력될 경우, 학습된 모델을 사용하여 위협정보로 판단되는 개체명 인식 및 추출



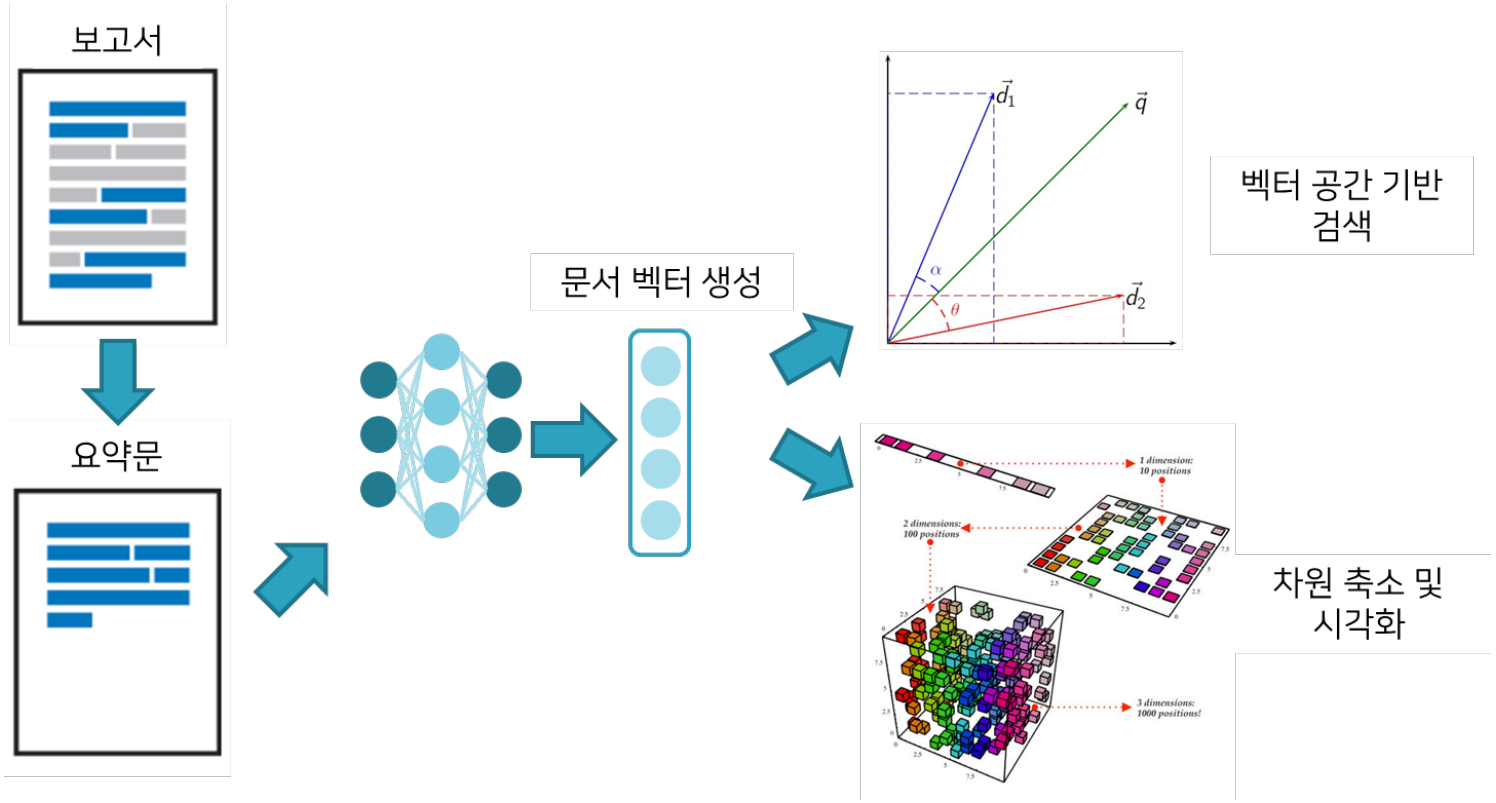
연구 내용 - 머신러닝을 이용한 문서 자동 요약

- 보고서로부터 요약문을 자동적으로 생성할 수 있는 기술 연구
- 보고서 검색 시 요약문을 보고서와 함께 효과적으로 제시할 수 있는 방법 연구



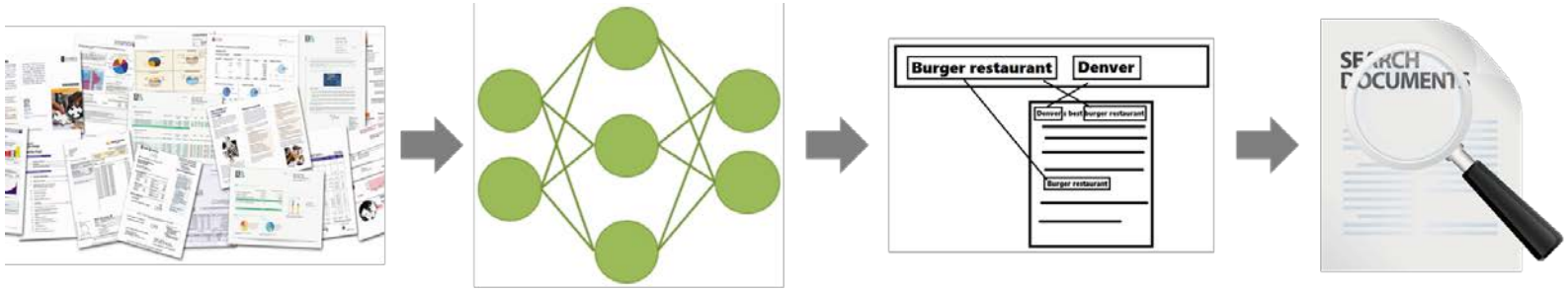
연구 내용 - 유사 문서 검색 및 시각화

- 벡터 공간 상에 문서를 임베딩하고, 이 벡터 공간의 차원을 축소하는 기술 연구
- 차원 축소된 벡터 공간을 효과적으로 시각화하고, 이를 이용하여 유사 문서를 검색하는 기술 연구



연구 목표 - Keyword Extraction

- 인텔리전스 보고서가 증가함에 따라 사용자가 원하고자 하는 문서를 짧은 시간 내에 판단하여 찾는 것은 쉽지 않음
- 사용자가 효율적으로 원하는 문서를 찾을 수 있도록 키워드 추출 알고리즘을 이용함



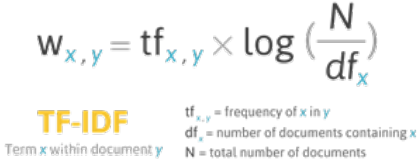
인텔리전스 보고서 | 키워드 추출 알고리즘 | 키워드 추출 | 키워드 검색 용이

연구 내용 - Keyword Extraction

- TF-IDF 활용
  - TF : 한 문서 내에 자주 출현하는 단어는 그 문서를 대표할 수 있다는 명제를 통해 문헌 내에 단어 빈도(Term Frequency)를 계산
  - IDF : 여러 문서에 동시에 출현하는 단어는 범용적인 확률이 높다는 전체 아래 역 문서 빈도수를 계산

보고서 집합

토큰화
및 전처리



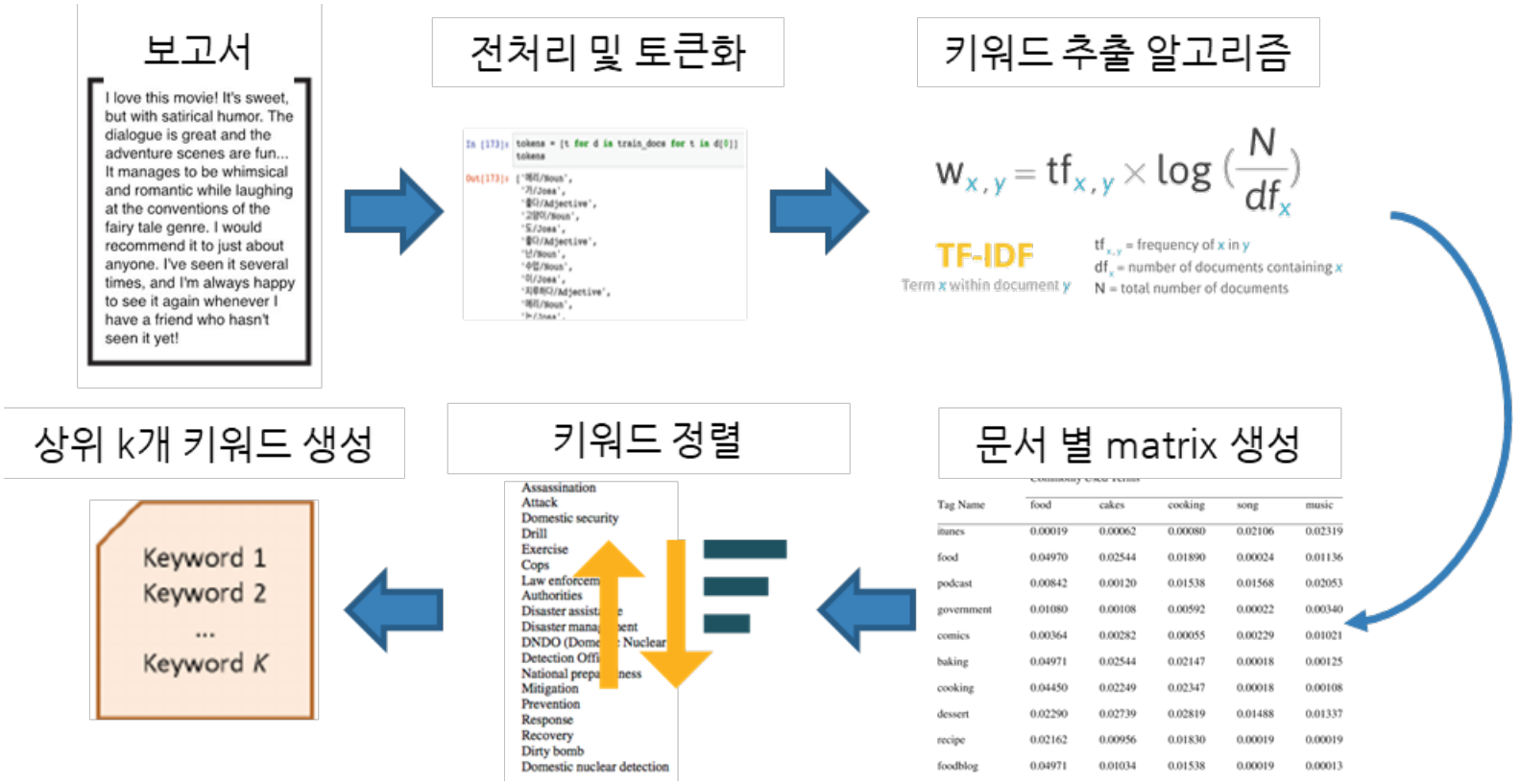
TF-IDF 모델

Terms	Scores
Term 1	0.5897
Term 2	0.6985
Term 3	0.2146
...	...
Term N	0.7891

후보 키워드
생성

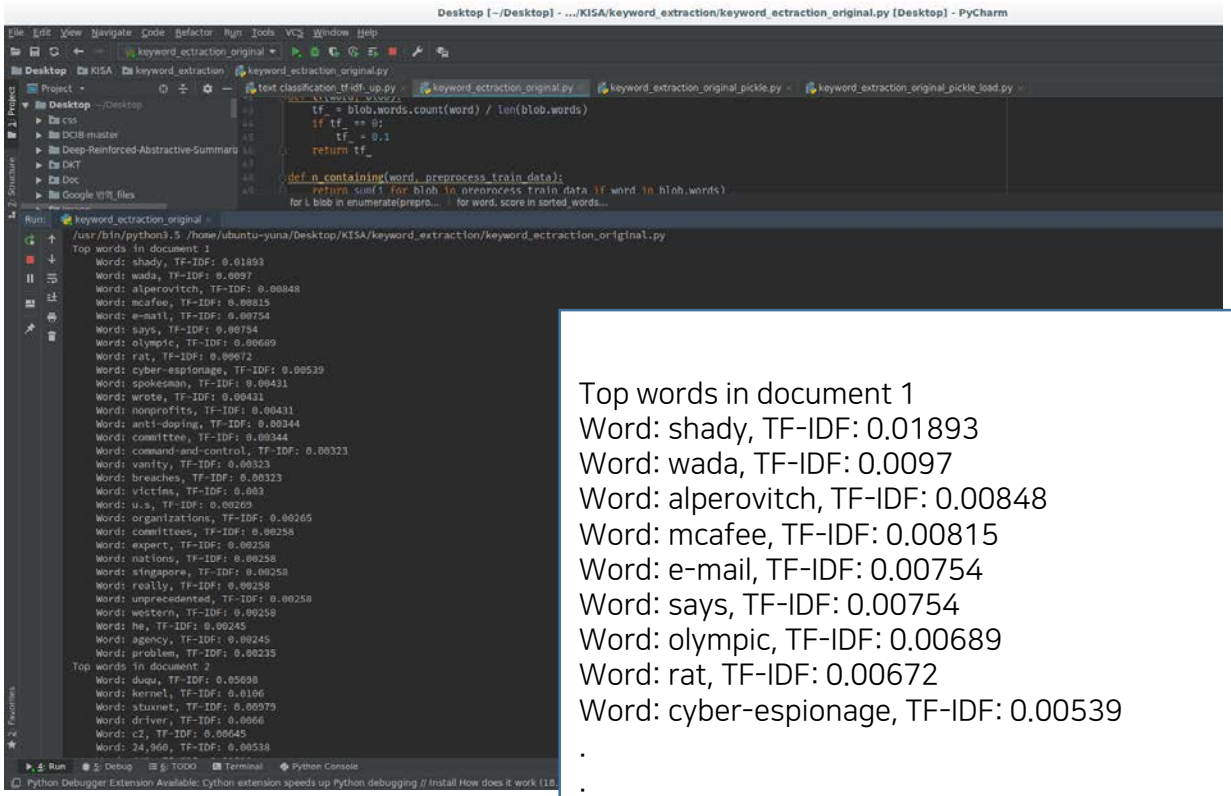
연구 내용 - Keyword Extraction

- 키워드 추출 시스템 구조도



연구 결과 - Keyword Extraction

- 각 인텔리전스 보고서 마다 keyword 추출



Top words in document 1
Word: shady, TF-IDF: 0.01893
Word: wada, TF-IDF: 0.0097
Word: alperovitch, TF-IDF: 0.00848
Word: mcafee, TF-IDF: 0.00815
Word: e-mail, TF-IDF: 0.00754
Word: says, TF-IDF: 0.00754
Word: olympic, TF-IDF: 0.00689
Word: rat, TF-IDF: 0.00672
Word: cyber-espionage, TF-IDF: 0.00539
.
.
.

연구 결과 - Keyword Extraction

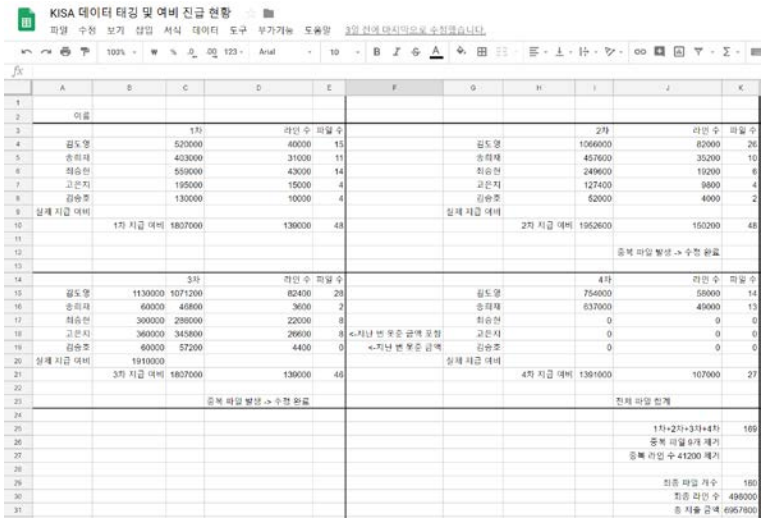
- 각 인텔리전스 보고서 keyword 추출 성능
  - 수작업으로 추출한 키워드와 결과와 비교 시 80%의 정확도를 보임

수작업을 통한 키워드 추출	본 모델을 이용한 키워드 추출
MSPs	Word: apt10
APT10	Word: plugx
Cloud hopper	Word: msp
Threat actor	Word: organisations
Plug X	Word: msps
PWC	Word: chches
BAE	Word: china-based
China based	Word: pwc
Organisation	Word: japanese
Dynmaic-dns	Word: bae

수작업을 통한 키워드 추출	본 모델을 이용한 키워드 추출
Utc+8	Word: dynamic-dns
Chches	Word: utc+8
poison	Word: annex
Japen	Word: hopper
Annex	Word: quasar
Poision ivy	Word: ivy
Attack	Word: 2016
IVY	Word: poison
C2 domain	Word: actor
FYP	Word: fyp

연구 결과 - 위협정보 개체명 인식 데이터 구축

- 데이터 구축
  - 국내외 각 보안 분야 기업에서 작성한 인텔리전스 리포트임
  - 데이터 구축을 위한 보안학과 재학 중인 5명의 연구원들이 참여
  - 가장 라인 빈도수 높은 파일(1500~5000어절)을 선정하여 배포
  - 데이터 구축을 위해 가이드 라인 생성
  - 지난 4달간 160개 파일로부터 약 50만 어절 생성



연구 결과 - 위협정보 개체명 인식 데이터 구축

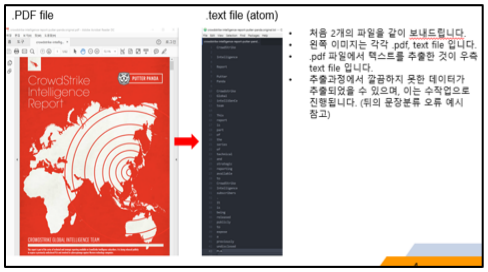
- 가이드라인 생성

• 프로젝트 설명

- 본 과제는 KISA에서 진행하는 위협 정보 자동추출 프로젝트입니다.
- 해당 과제에서 필요한 Datasets 구성을 위해 사전 태깅 작업을 진행 하려고 합니다.
- 작업 환경은 상관 없으나 저희가 진행하는 OS가 linux이므로, 윈도우에서 작업 시 Editor는 Atom에서 진행 해주시길 바랍니다.

- 배포하는 PDF 파일은 각 보안 분야의 회사에서 기록한 Intelligence Report 입니다.
- Text 파일은 각 report를 크롤링 및 사전작업을 통하여 정리한 text문서입니다.
- 현재 Text 파일은 한 줄에 한 단어 씩 있으며 문장은 공백 줄로 분리되어 있습니다.
- 그러나 간혹 전처리로 처리되지 못한 text들이 있을 수 있습니다.
- 이것은 사람이 직접 수정해야 합니다.

• 배포 파일 설명



- 처음 2개의 파일을 같이 보내드립니다.
- 왼쪽 이미지는 각각 .pdf, text file 입니다.
- .pdf 파일에서 텍스트를 추출한 것이 우측 text file 입니다.
- 추출과정에서 깔끔하지 못한 데이터가 추출되었을 수 있으며, 이는 수작업으로 진행됩니다. (뒤의 문장분류 오류 예시 참고)

• BIO 태깅

주의사항

태깅 시, BIO tagging scheme 방식을 사용합니다.

예시 :
New York의 경우 앞의 'New'와 뒤의 'York'이 합쳐서 하나의 장소를 나타내는 단어 이므로, 두 개의 단어 의미를 이어주기 위해서

구분자 간격 = 탭

New B_LC
York I_LC (영어 i 대문자)

으로 태깅 합니다.

BIO 태깅은 단어 뒤에 붙는 '조사'에 사용하는 것이 아닌, 의미를 가진 명사가 '띄어쓰기'로 구분되어 있는 텍스트의 경우 그 의미를 이어주기 위해 사용하는 것입니다.

• Class 정의

대분류	소분류	태깅명
Hash	Hash	hash
악성코드	가상통화채굴	malware.mining
	백도어(원격제어)	malware.backdoor
	DDoS 공격	malware.ddos
	드랍퍼(다운로더, 런처)	malware.drop
	랜섬웨어	malware.ransom
	정보유출	malware.infosteal
	기타	malware.unknown
	정상(악성코드 아님)	malware.normal
IP	C&C, 정보유출지	ip.cncsvr
	공격시도(scan, brute force, ping, sql injection 시도 등 다양한 공격시도)	ip.attack
	유포지	ip.distribute
	DDoS공격 수행(봇넷)	ip.ddos
	기타	ip.unknown
	정상	ip.normal
도메인/URL	C&C, 정보유출지	url.cncsvr
	경유지	url.codevia
	유포지	url.distribute
	기타	url.unknown
	정상	url.normal

Class 정의

- 태깅 클래스는 **20 가지**로 정의 되어있습니다. (태깅명 부분)
- 이는 태깅 시 20가지의 다양한 클래스로 태깅이 진행되야 함을 의미합니다.
- 태깅 이외의 단어는 **아무 태깅도 하지 않습니다.**
- 구분자는 탭으로 구분합니다.(뒷장 참고)

6

대분류	소분류	태깅명
Hash	Hash	hash
악성코드	가상통화채굴	malware.mining
	백도어(원격제어)	malware.backdoor
	DDoS 공격	malware.ddos
	드랍퍼(다운로더, 런처)	malware.drop
	랜섬웨어	malware.ransom
	정보유출	malware.infosteal
	기타	malware.unknown
	정상(악성코드 아님)	malware.normal
IP	C&C, 정보유출지	ip.cncsvr
	공격시도(scan, brute force, ping, sql injection 시도 등 다양한 공격시도)	ip.attack
	유포지	ip.distribute
	DDoS공격 수행(봇넷)	ip.ddos
	기타	ip.unknown
	정상	ip.normal
도메인/URL	C&C, 정보유출지	url.cncsvr
	경유지	url.codevia
	유포지	url.distribute
	기타	url.unknown
	정상	url.normal

대분류 : Hash (태깅명 : hash)

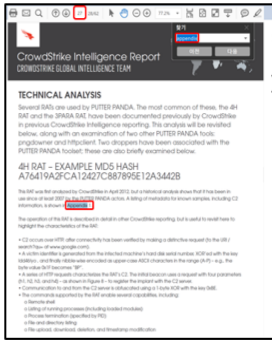
- Text에서 Hash로 보이는 것은 hash로 태깅 합니다.
- 예시

11

• Table, appendix 등 주요사항

중요!

- 태깅 시, Table이나 appendix가 언급되면,
- Table과 appendix 자체는 hash값이 아니지만,
- 해당 Table이나 appendix의 내용을 보고 hash로 정리된 것일 경우, hash로 태깅 해야 합니다. (다음 장 예시 참고)



첨부된 pdf 27페이지 참고.

- pdf 27페이지에 'Appendix 1' 언급
- 실제 Appendix 1은 해당 pdf 50페이지에 정리 되어 있습니다.

19

Appendix 50페이지에 정리된 HASH 값

- pdf 27페이지에 'Appendix 1' 언급
- 실제 Appendix 1은 해당 pdf 50페이지에 정리 되어 있습니다.

20

연구 결과 - 위협정보 개체명 인식 데이터 구축

- 오류수정 도구 생성
  - 태그 작업은 사람에 의해 발생된 부분으로 실수가 발생할 수 있음
  - 따라서 오류를 방지하고자 태그확인이 가능한 반자동 도구를 생성하여 배포
  - 오류 파악 도구에서, 도구가 자동적으로 말뭉치를 수정하는 경우는 모호성이 존재하지 않는 경우에 한함
  - 이는 모호성이 있는 상황에서 잘못된 오류 수정으로 인해 말뭉치가 손상되는 것을 방지하기 위함임
  - 모호성이 존재할 가능성이 있는 경우, 도구는 오류를 수정하는 대신 오류가 발생한 파일과 발생 위치를 사용자에게 제시하여 직접적인 수정을 유도함

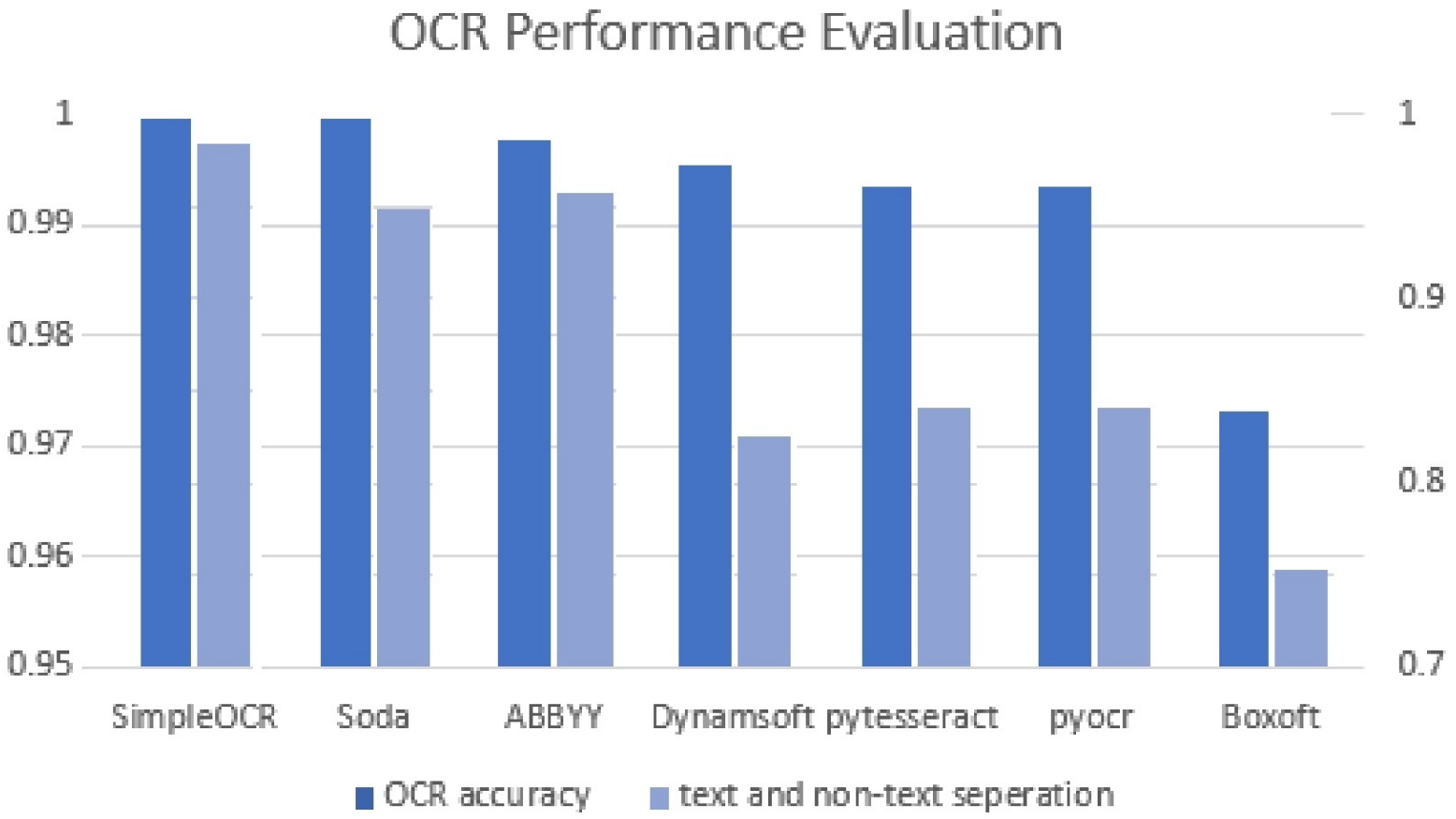
연구 결과 - 위협정보 개체명 인식 데이터 구축

- 오류수정 도구 생성
  - 태그 작업은 사람에 의해 발생된 부분으로 실수가 발생할 수 있음
  - 따라서 오류를 방지하고자 태그확인이 가능한 반자동 도구를 생성하여 배포
  - 오류 파악 도구에서, 도구가 자동적으로 말뭉치를 수정하는 경우는 모호성이 존재하지 않는 경우에 한함
  - 이는 모호성이 있는 상황에서 잘못된 오류 수정으로 인해 말뭉치가 손상되는 것을 방지하기 위함임
  - 모호성이 존재할 가능성이 있는 경우, 도구는 오류를 수정하는 대신 오류가 발생한 파일과 발생 위치를 사용자에게 제시하여 직접적인 수정을 유도함

연구 내용 - OCR 프로그램 조사 및 성능 비교

- 다양한 형태로 제공되는 OCR 프로그램을 7종으로 추려 2가지 성능 평가를 진행함
- 주어진 문자 이미지를 해당 문자로 정확히 변환하는지, 문자를 포함한 이미지와 문자가 아닌 이미지를 구분할 수 있는지를 측정해 정확도로 표현함
- 하지만 각 프로그램마다 줄 바꿈과 문장 순서 등 출력 형태가 달라 정답 데이터와 예측 데이터의 문자들을 일대일로 비교하기 어려움
- 모든 예측 데이터를 정답 데이터에 맞춰 일일이 병렬 데이터를 구축하는 작업은 확장성이 낮음
- 이에 따라 대안으로서, 정답 데이터에 등장한 문자의 출현 횟수를 사전(dictionary)에 저장한 뒤, 예측 데이터마다 사전을 구축해 정답 데이터 사전과 값 차이의 절대값을 모두 더해 오차를 구함

연구 내용 - OCR 프로그램 조사 및 성능 비교



<그림 4> OCR 성능 비교 결과

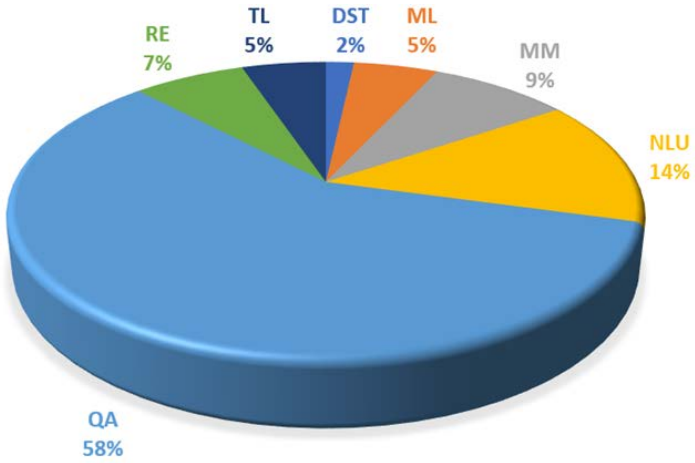
기계학습을 활용한
자연어 질의/응답 인공지능 모델 개발

고려대학교 NLP&AI Lab.

- 기술 동향 보고서
- 기술 개발 전략 보고서
- 자동차 법률 응답 대화 시스템

연구 내용 - 기술 동향 보고서 및 기술 개발 전략 보고서

- 과제 배경
  - 4차 산업혁명의 핵심 기술인 AI 기술의 선도적 확보를 위해 차세대 업무 지원이 가능한 AI DSS (Decision Support System) 서비스 모델 및 질의응답 모델 개발 연구
  - 최신 NLP관련 Top Conference 및 학회지에서 발표된 논문들을 분석함
  - Machine Learning, Question & Answering, Natural Language Understanding, MultiModal Approach, Relation Extraction, Transfer Learning,
    Dialogue State Tracking 분야를 다룸



연구 내용 - 기술 동향 보고서 및 기술 개발 전략 보고서

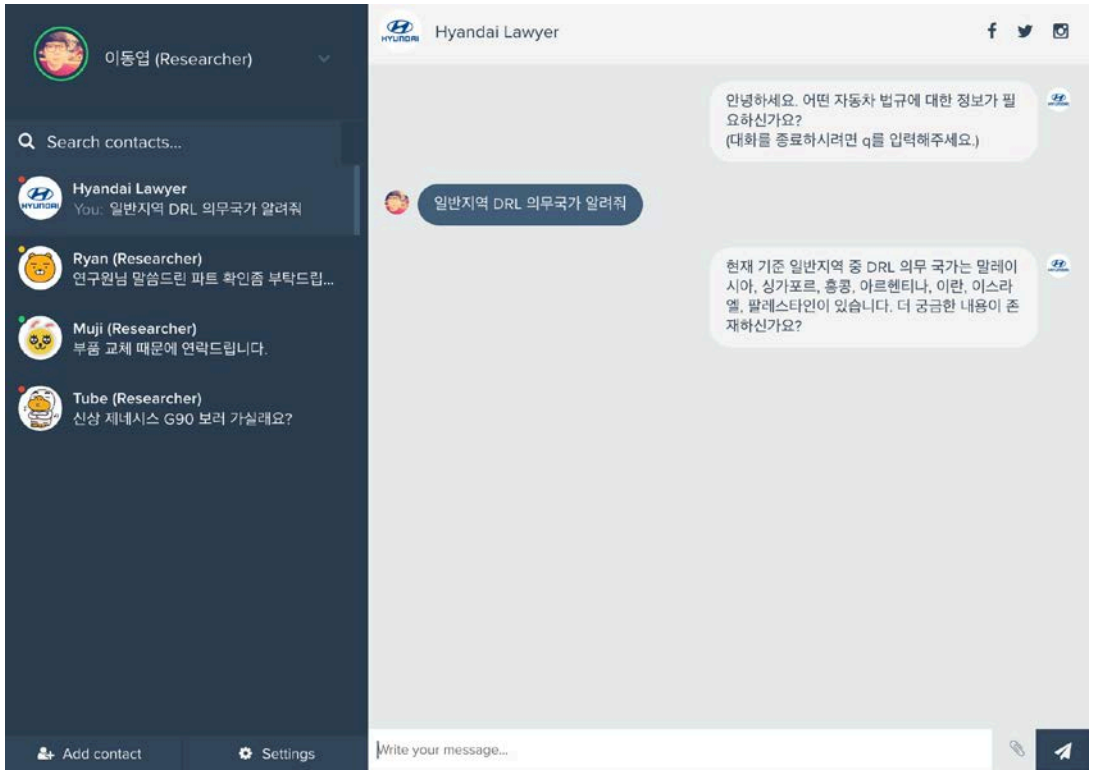
- 분석 및 논문 요약
  - Know What You Don't Know: Unanswerable Questions for SQuAD
  - Large-Scale QA-SRL Parsing
  - Learning to Ask Good Questions: Ranking Clarification Questions using Neural Expected Value of Perfect Information
  - Multi-Relational Question Answering from Narratives: Machine Reading and Reasoning in SimulatedWorlds
  - Question Answering on Knowledge Bases and Text using Universal Schema and Memory Networks
  - Attention Is All You Need
  - Learning to Paraphrase for Question Answering
  - Latent Space Embedding for Retrieval in Question-Answer Archives
  - The NarrativeQA Reading Comprehension Challenge
  - Adversarial Adaptation of Synthetic or Stale Data

연구 내용 - 기술 동향 보고서 및 기술 개발 전략 보고서

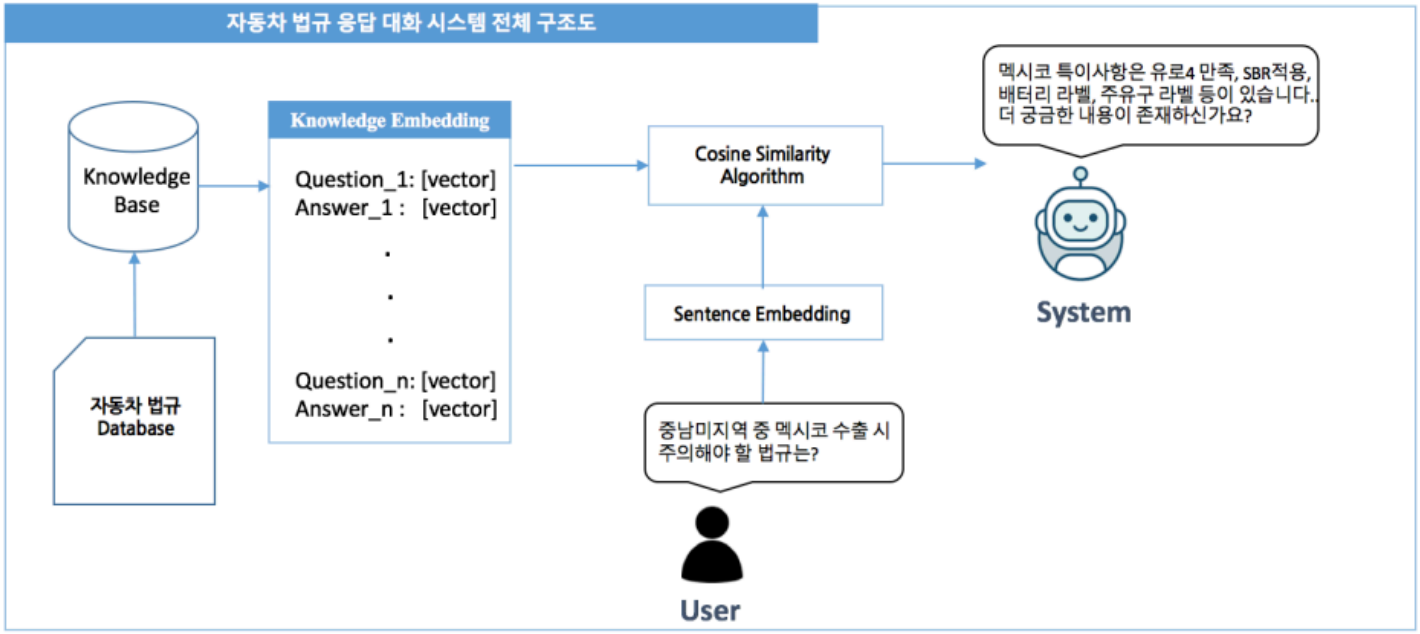
- 분석 및 논문 요약
  - New Transfer Learning Techniques for Disparate Label Sets
  - ONENET: JOINT DOMAIN, INTENT, SLOT PREDICTION FOR SPOKEN LANGUAGE UNDERSTANDING
  - ReadingWikipedia to Answer Open-Domain Questions
  - Never-Ending Learning
  - Visual Relation Extraction via Multi-modal Translation Embedding Based Model
  - Training Classifiers with Natural Language Explanations
  - Zero-Shot Transfer Learning for Event Extraction
  - Cross-Domain Sentiment Classification with Target Domain Specific Information
  - Question-Answer Driven Semantic Role Labeling
  - Zero-Shot Recognition via Semantic Embeddings and Knowledge Graphs
  - Semantic Parsing of Technical Support Questions
  - Adaptive Document Retrieval for Deep Question Answering

연구 내용 - 자동차 법률 응답 대화 시스템

- 자동차에 관련한 법률 지식을 대화 시스템을 통해 연구자들에게 제공해주는 시스템
- 본 시스템을 통해, 연구자들은 자동차 법률 지식에 쉽게 접근할 수 있음



연구 내용 - 자동차 법률 응답 대화 시스템



- 자동차에 관한 법규들을 지식 기반(knowledge base)의 형태로 저장
- 연구자로부터 질문(query)이 들어왔을 때, 구축된 지식을 활용하여 연구자에게 관련된 법규 조항들을 알려줌
- 질문과 구축된 법률 지식 기반과의 유사도를 측정하고, 해당 질문과 가장 유사도가 높은 법률 지식에 대한 답을 전달

연구실적

고려대학교 NLP&AI Lab.

- 논문
- 특허
- 소프트웨어등록
- 시제품

연구 실적 - 논문

❖ 해외 학술지

1. Seolhwa Lee, Dongyub Lee, Danial Hoosyar, Jaechoon Jo, Heuiseok Lim, “Integrating Breakdown Detection into Dialogue Systems to Improve Knowledge Management: Encoding Temporal Temporal Utterances with Memory Attention”, the journal of information technology and management(ITEM), 2018. (under review)
2. Danial Hooshyar, Seolhwa Lee, Yeongwook Yang, Jaechoon Jo, Heuiseok Lim, ”SmartSenior: automatic content personalization through semi-supervised learning”, Wireless Personal Communications, 2018. (SCIE)
3. Danial Hooshyar, Seolhwa Lee, Yeongwook Yang, Jaechoon Jo, Heuiseok Lim, ”Long-term effects of adaptive customization support on elderly people”, Cognition Technology and Work, 2018. (SSCI/SCIE)

연구 실적 - 논문

❖ 해외 학술대회

1. Aram So, Natasia Wulansari, HeuiSeok Lim, " A study on building corpus for restaurant reservation system " , The 2nd International Conference on Interdisciplinary research on Computer Science, Psychology, and Education (ICICPE), pp. 126-128, 2018.
2. Seolhwa Lee, Heuiseok Lim, "A Study on Vector-Based User-Preferred Fashion Matching", The 2st International conference on Interdisciplinary research on Computer Science, Psychology, and Education (ICICPE' 2018), 2018.
3. YunA Hur, Heuiseok Lim, "Empirical Evaluation of Traditional Culture Document Classification Systems with TF-IDF/Bag-of-Words Features", The 2nd International Conference on Interdisciplinary research on Computer science, Psychology, and Education(ICICPE), 2018.
4. Chanhee Lee, Youngbum Kim, Dongyub Lee, Heuiseok Lim. "Character-Level Feature Extraction with Densely Connected Networks", International Conference on Computational Linguistics (COLING 2018), 2018.08.
5. Andrew Matteson, Chanhee Lee, Young-Bum Kim, HeuiSeok Lim. "Rich Character-Level Information for Korean Morphological Analysis and Part-of-Speech Tagging", International Conference on Computational Linguistics (COLING 2018), 2018.08.

연구 실적 - 논문

❖ 해외 학술대회

6. Chanhee Lee, Heuiseok Lim. “Sketch to Image Upsampling using Generative Adversarial Networks for Query by Image Content“, International Conference on Interdisciplinary research on Computer Science, Psychology, and Education (ICICPE), 2018.01.
7. Kuekyeng Kim, Hyesung Ji, Heuiseok Lim, “Comparisons of multiple approaches in improving Image Caption Generation Models”, ICICPE 2018, 2018.02
8. Gyeongmin Kim, Heuiseok Lim, “A study of Dataset and User interface for a Restaurant Reservation Dialogue System", International Conference on Interdisciplinary research on Computer Science, Psychology, and Education(ICICPE), 2018.
9. Taesun Whang, Dongyub Lee, Chanhee Lee, Heuiseok Lim, “Enhanced Sequential Representation Augmented with Utterance-level Attention for Response Selection“, AAAI’2019 Workshop on Dialog System Technology Challenges (DSTC7), 2019. 01.

연구 실적 - 논문

❖ 국내 학술지

1. 이설화, 이찬희, 조재춘, 임희석, " 암묵적 사용자 프로파일링을 통한 딥러닝기반 지능형 선호 패션 추천", 한국융합학회논문지 2018.
2. 김규경, 김경민, 조재춘, 임희석, “전통 문화 데이터를 이용한 메타 러닝 기반 전역 관계 추출”, 한국융합학회논문지, 2018.11, vol.9, no.11, pp. 23-28 (6 pages)
3. 김경민, 김규경, 조재춘, 임희석, " 한국 전통문화 말뭉치구축 및 Bi-LSTM-CNN-CRF 기반의 개체명 인식 모델 개발", 한국융합학회, 2016년 추계학술발표대회 논문집, 제 23권, 제 2호, pp. 327-328, 2018.

연구 실적 - 논문

❖ 국내 학술대회

1. 소아람, 이설화, 허윤아, 황태선, 임희석, “맞춤형 UX 설계를 위한 시니어와 주니어 사용자의 생체능력 비교 분석 연구”, 한국컴퓨터교육학회 하계 학술발표논문지 제22권 제2호, 2018.
2. 소아람, 박기남, 임희석, “식당예약 및 추천을 위한 한국어 대화 코퍼스 구축 연구”, 제30회 한글 및 한국어정보처리 학술대회 논문집, pp. 630-632, 2018.
3. 이설화, 소아람, 허윤아, 황태선, 임희석, "고령 사용자를 위한 UI/UX 컴포넌트 매핑 및 시각화 방법", 컴퓨터교육학회 2018 하계학술발표논문집, 제22권, 제2호, pp.85~87, 2018
4. 이설화, 조재춘, 이찬희, 이동엽, 임희석, "사용자 프로파일링을 통한 지능형 선호 패션 추천", 2018 한국패션비지니스학회, pp.72~73, 2018.
5. 허윤아, 황태선, 소아람, 이설화, 임희석, "고령 사용자 맞춤형 UI/UX 제공을 위한 인지 및 온라인 활동 기반 프로파일링", 컴퓨터교육학회, 하계학술 발표논문집, 제 22권, 제 2호, 2018
6. 허윤아, 김경민, 이찬희, 임희석, “딥러닝 기술을 이용한 비정형 보안 위협정보 자동 탐지 및 추출 기술 연구", 한글 및 한국어 정보처리 학술대회정보처리학회, 2018.
7. 이찬희, 조재춘, 이동엽, 이설화, 임희석. “스케치를 이용한 남성 의류 검색“, 한국패션비즈니스학회 2018 춘계학술대회 논문집, pp.79-80, 2018.04.
8. 이찬희, 이동엽, 허윤아, 양기수, 임희석. “음절 단위 및 자모 단위의 Byte Pair Encoding 비교 연구“, 제30회 한글 및 한국어 정보처리 학술대회 논문집, 2018.

연구 실적 - 논문

❖ 국내 학술대회

9. 김규경, 김경민, 조재춘, 임희석 "문서 내 전역 관계 추출: 생략된 개체의 고려",
제30회 한글 및 한국어 정보처리 학술대회, 2018

10. 이동엽, 황태선, 이찬희, 임희석, "적대 학습을 이용한 자연어 이해", 한글 및 한국어정보처리 학술대회,
2018

11. 이동엽, 이찬희, 이설화, 조재춘, 임희석 “패션 상품 이미지를 이용한 패션 상품 카테고리 자동 분류기 개발”
한국패션비즈니스학회 2018 춘계학술대회 논문집, pp.131-132, 2018.04.

12. 김경민, 허윤아, 김규경, 임희석, " 정보보안 분야의 위협정보 개체명 인식 시스템 개발을 위한 데이터셋 구축", 한글 및 한국어 정보처리학회, 2018.

13. 황태선, 이동엽, 임희석, “Attention 기반의 대화 발화 예측 모델",
제 30회 한글 및 한국어 정보처리 학술 대회, pp 40-43, 2018.

14. 황태선, 이설화, 허윤아, 소아람, 임희석, “시니어 인지 반응 측정 데이터 기반 뉴스 형태 추천 시스템 설계”,
한국 컴퓨터 교육학회 하계 학술대회, 제 22권 2호. pp. 143-145. 2018.

15. 양기수, 이동엽, 이찬희, 임희석, "오토인코더와 적대 네트워크를 활용한 한국어 문체 변환", 한국정보과학회,
제30회 한글 및 한국어 정보처리 학술대회 논문집, 2018

연구 실적 - 특허

❖ 특허(국내)

1. 스마트 시니어 인지반응 기반의 모델링 방법 및 장치, 출원 완료 및 등록 진행 중, 이설화
2. 사용자 프로파일링을 이용한 사용자의 선호 패션 추천 방법 ,출원 완료, 이설화
3. 스마트 시니어 온라인 활동 기반 프로파일을 이용한 맞춤형 콘텐츠, 출원 완료, 허윤아
4. 한글 문서 내 전역적 관계 추출기, 2018.12.03 출원, 김규경, 김경민, 조재춘
5. 딥 러닝을 이용하여 이미지를 검색하는 단말 장치 및 방법, 이동엽, 임희석
6. 스마트 시니어를 위한 맞춤형 콘텐츠 추천 장치 및 방법, 출원 완료, 황태선

❖ 특허(국외)

1. 스마트 시니어 인지반응 기반의 모델링 방법 및 장치, 출원 완료, 이설화

❖ 기술이전

1. 한국어 개체명 인식 시스템, 이동엽, 임희석

연구 실적 – 시제품 제작, 강연, 수상

❖ 시제품 제작

1. 인지반응 분석 시스템 플랫폼 제작, http://nlplab.iptime.org:4321/seol2/
2. 전통문화 관계 추출기 개발 DEMO: http://nlplab.iptime.org:32277/
3. 한국어 식당 예약 시스템, http://nlplab.iptime.org:8886/
4. 현대 자동차 자동차 법률 상담 시스템, http://nlplab.iptime.org:8894/
5. 딥러닝 기반의 패션 상품 유사 이미지 검색 모델, http://nlplab.iptime.org:8854/
6. 전통문화 개체명 인식기 개발 DEMO, http://nlplab.iptime.org:32280/ner_demo/index.html

❖ 강연

1. 현대자동차, “Bilingual Word Embedding”, 2018. 이설화
2. KT, “Bilingual Word Embedding”, 2018. 이설화
3. 현대자동차 Sequence Tagging (Named Entity Recognition) 튜토리얼
4. KT End-to-End Sentence Classification 튜토리얼

연구 실적 – 시제품 제작, 강연, 수상

❖ 수상

1. 국어정보처리 경진대회 (의존 구문 파서), 차다은, 이동엽, 임희석

❖ 국내포럼

1. Samsung AI Forum 2018, Character-Level Feature Extraction with Densely Connected Networks.