특집논문 (Special Paper)
방송공학회논문지 제24권 제1호, 2019년 1월 (JBE Vol. 24, No. 1, January 2019)
https://doi.org/10.5909/JBE.2019.24.1.58
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

효율적 대화 정보 예측을 위한 개체명 인식 연구

고 명 현[a], 김 학 동[a], 임 헌 영[a], 이 유 림[b], 지 민 규[c], 김 원 일[d]‡

A Study on Named Entity Recognition for Effective Dialogue Information Prediction

Myunghyun Go[a], Hakdong Kim[a], Heonyeong Lim[a], Yurim Lee[b], Minkyu Jee[c], and Wonil Kim[d]‡

요 약

대화 문장 내 고유명사와 같은 개체명에 대한 인식 연구는 효율적 대화 정보 예측을 위한 가장 기본적이며 중요한 연구 분야이다. 목적 지향 대화 시스템에서 가장 주요한 부분은 대화 내 객체가 어떤 속성을 가지고 있느냐 하는 것을 인지하는 것이다. 개체명 인식 모델은 대화 문장에 대하여 전처리, 단어 임베딩, 예측 단계를 통해 개체명 인식을 진행한다. 본 연구는 효율적인 대화 정보 예측을 위해 전처리 단계에서 사용자 정의 사전을 이용하고 단어 임베딩 단계에서 최적의 파라미터를 발견하는 것을 목표로 한다. 그리고 설계한 개체명 인식 모델을 실험하기 위해 생활 화학제품 분야를 선택하고 관련 도메인 내 목적 지향 대화 시스템에서 적용 할 수 있는 개체명 인식 모델을 구축하였다.

Abstract

Recognition of named entity such as proper nouns in conversation sentences is the most fundamental and important field of study for efficient conversational information prediction. The most important part of a task-oriented dialogue system is to recognize what attributes an object in a conversation has. The named entity recognition model carries out recognition of the named entity through the preprocessing, word embedding, and prediction steps for the dialogue sentence. This study aims at using user - defined dictionary in preprocessing stage and finding optimal parameters at word embedding stage for efficient dialogue information prediction. In order to test the designed object name recognition model, we selected the field of daily chemical products and constructed the named entity recognition model that can be applied in the task-oriented dialogue system in the related domain.

Keyword : Task-Oriented Dialogue System, Word Embedding, NER(Named Entity Recognition), Bi-LSTM

a) 세종대학교 디지털콘텐츠학과(Department of Digital Contents, Sejong University)
b) 세종대학교 인공지능언어공학과(Department of Artificial Intelligence and Linguistic Engineering, Sejong University)
c) 세종대학교 소프트웨어융합학과(Department of Software Convergence, Sejong University)
d) 세종대학교 소프트웨어학과(Department of Software, Sejong University)
‡ Corresponding Author : 김원일(Wonil Kim)
E-mail: wikim@sejong.ac.kr
Tel: +82-3408-3795
ORCID: https://orcid.org/0000-0002-1489-8427

※ This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(2015R1D1A1A01060693)
· Manuscript received November 15, 2018; Revised December 31, 2018; Accepted December 31, 2018.

Ⅰ. 서 론

개체명은 인명, 지명, 기관명 등과 같은 고유명사나 명사구를 의미하며 개체명 인식이란 문장 내 어절을 기본 단위로 분석하여, 문장에 출현한 모든 개체명의 경계를 인식하고, 해당 개체명이 어떤 태그로 분류될 수 있는지 자동으로 인식하는 것을 말한다. 개체명 인식 기술은 자연어 이해를 위한 어휘 수준의 이해 과정에 속하며 목적 지향 대화시스템, 정보 검색 시스템에서 중요성이 크다.

목적 지향 대화시스템에서 개체명 인식은 각 도메인 특성에 따라 방대한 고유명사로 인해 어려움이 있고, 각 고유명사에 대한 개체명 경계를 판단하기 어렵다. 개체명 인식 모델은 대화 문장에 대하여 전처리, 단어 임베딩, 예측 단계를 통해 개체명 인식을 진행한다. 본 연구에서는 개체명 인식 단계 중 전처리 부분에서 발생할 수 있는 문제점에 대하여 분석하고 이를 해결하기 위한 개체명 인식 모델을 제안한다. 그리고 설계한 개체명 인식 모델을 실험하기 위해 생활 화학제품 분야를 선택하고 관련 분야의 목적 지향 대화 시스템에서 적용 할 수 있는 개체명 인식 모델을 구축하였다.

본 논문의 구조는 다음과 같다. 제 2장에서 이전 연구를, 제 3장에서 본 논문에서 제안하는 생활 화학제품 관련 새로운 개체명과 효율적 대화 정보 예측을 위한 개체명 인식 모델을 소개한다. 제 4장에서는 제안 모델을 이용한 실험을 진행하고 결과를 기술한다. 제 5장에서는 본 연구의 결론 및 향후 연구 방향에 대해 논의한다.

Ⅱ. 이전 연구

1. 목적 지향 대화 시스템

대화 시스템은 목표에 따라 목적 지향 대화 시스템(Task-Oriented Dialogue System)과 재미를 위한 대화 시스템(ChatBot)으로 나눌 수 있다. 목적 지향 대화 시스템은 호텔이나 식당에서 예약, 질의응답과 같은 특별한 목적에 해당하는 업무를 수행하기 위한 시스템으로 가상 비서 대화 시스템으로 말할 수 있다. 재미를 위한 대화 시스템은 사람과 수다를 떠는 시스템으로 잡담 대화 시스템으로 말할 수 있다.

목적 지향 대화 시스템은 기본적으로 자연어 이해, 대화 관리, 자연어 생성의 과정을 포함하며 자연어 이해 부분에서 질문의 도메인과 의도를 파악하고 대화 관리 부분에서 대화 맥락을 고려한 질문의 답변을 검색하고 답변 발화를 결정한다. 자연어 생성 모듈에서는 답변 검색을 통해 획득한 지식을 제공하기 위한 답변을 생성한다[1].

2. 개체명 인식

개체명 인식에 대한 연구는 1995년, MUC-6(Message Understanding Conference)에서 처음 촉발되었다. 초기에는 연구자들이 서로 다른 입출력 방식을 사용하여 연구 진척에 어려움이 있었지만, 향후 BIO 태깅(Tagging)을 통한 통일된 입출력 방식을 이용하여 체계적으로 개체명 인식에 대한 연구가 되고 있다. BIO 태깅이란 특정 문장에 대하여 개체명의 시작을 B(Beginning)문자로 표시하고 개체명 중간에 있는 문자는 I(Inside), 개체명이 아닌 문자의 경우에는 O(Outside)로 표시하는 태깅 방법이다. 개체명 인식 연구에서 개체명 태그 세트는 제한 사항 없이 각 연구자의 연구 목적에 따라 설정하고 사용한다[2].

영어권의 경우 대문자, 소문자의 사용 정보가 개체명 인식에 큰 도움이 되므로 개체명 인식 기술의 연구 성과가 높다. 하지만 한국어는 언어적 특질이 영어와 달라 영어 개체명 인식에서 쓰인 기술을 그대로 적용하기에는 어려움이 있다. 최근에는 딥 러닝을 적용해 한국어의 특질을 분석한 한글 개체명 인식 연구도 활발히 진행되고 있다[3][4].

3. 자연어 처리를 위한 신경망 모델

최근 자연어 처리를 위한 신경망 모델로 가장 많이 사용되는 모델은 장단기 기억 구조(Long Short-Term Memory)[5] 모델이다. 장단기 기억 구조 모델은 순환 인공 신경망(Recurrent Neural Network)[6] 모델의 문제점을 해결하기 위하여 고안되었다. LSTM은 기존 RNN이 이전 입력에 대한 결과를 다음 단계에 전이하는 것과 달리 입력 게이트, 망각 게이트, 출력 게이트를 이용하여 불필요한 정보는 지

우고 중요한 정보들만 전달할 수 있도록 만든 구조이다. 이로 인해 LSTM은 RNN과 달리 긴 시퀀스의 입력 값을 처리하는데 좋은 성능을 보인다.

개체명 인식 기술에서 최근 가장 많이 사용되는 모델은 양방향 장단기 기억 구조(Bidirectional-LSTM) 모델이다[7][8]. Bi-LSTM 모델은 기존 단방향으로 상태를 전이 시키는 LSTM 모델에 대하여 반대 방향으로도 상태를 전이시키는 모델로 확장한 것이다.

그림 1과 같이 여러 개의 단어로 구성된 문장과 같은 시퀀스 데이터에 대하여 LSTM 모델은 각 입력으로 단어를 입력 받는다. 각 타임 스텝마다 단어를 입력 받아 메모리 셀에 있는 게이트에서 연산을 통해 결과를 전이하고 출력한다. Bi-LSTM의 경우 전위 방향(Forward)에 대하여 시간 t일 경우 W2의 단어를 입력 받고 이전의 W0, W1에 의한 상태를 전이 받는다. 후위 방향(Backward)에 대하여 시간 t일 경우 마찬가지로 W2의 단어를 입력 받고 이전의 W4, W3에 의한 상태를 전이 받는다. 이처럼 Bi-LSTM 셀은 각 시간에 대하여 앞뒤로 이전 상태를 고려한 출력 결과를 생성하므로 개체명 인식과 같은 순차 라벨링 문제를 해결하는데 특화된 모델이다.

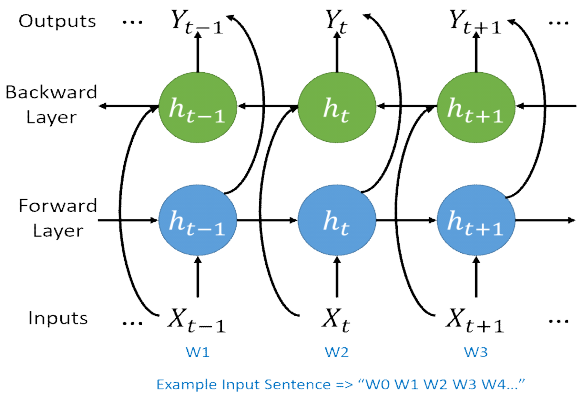


그림 1. Bi-LSTM 모델 개요
Fig. 1. Introduction of Bi-LSTM Model

Ⅲ. 제안 방법

1. 생활 화학제품 관련 개체명 정의

국내 개체명 인식과 관련한 표준으로 한국정보통신기술협회(TTA : Telecommunications Technology Association)가 발표한 개체명 태그 세트 및 태깅 말뭉치[9]가 있다. 기존 표준에는 개체명 대분류로 인물, 학문분야, 이론, 인공물, 기관, 지역, 문명, 날짜, 시간, 수량, 이벤트, 동물, 식물, 물질, 용어로 모두 15개로 구성되어 있다.

생활 화학제품 위해 정보 제공 대화 시스템 설계를 위하여 사용자가 질의한 문장에 대하여 제품 제조사, 제품 품명, 제품 세부 품목, 제품 제형 분류, 성분, 유입 경로 등의 정보를 추출할 필요가 있다. 따라서 본 연구에서는 생활 화학제품과 관련한 단어에 대한 개체명 인식을 위해 한국정보통신기술협회 표준[9]을 참고하여 아래 표 1과 같이 새로운 개체명 태그 세트를 정의하였다. 제품 세부 품목의 경우 위해우려 제품 품목과 동일하며 세정제, 합성세제, 표백제, 섬유유연제 등이 있다.

표 1. TTA 표준 태그 세트와 제안 태그 세트 비교
Table 1. Comparison Between TTA Standard and Proposal Tag Set

TTA Standard Tag Set		Proposal Tag Set	
meaning	Tag	meaning	Tag
PERSON	PS	PERSON	PS
LOCATION	LC	LOCATION	LC
ORGANIZATION	OG	ORGANIZATION	OG
ARTIFACTS	AF	DATE	DT
DATE	DT	TIME	TI
TIME	TI	BRAND	BR
CIVILIZATION	CV	MODEL NAME	MN
ANIMAL	AM	PRODUCT TYPE	PT
PLANT	PT	SHAPE TYPE	ST
QUANTITY	QT	QUANTITY	QT
STUDY_FIELD	FD	MATERIAL	MT
THEORY	TR	INFLOW ROUTE	IR
EVENT	EV		
MATERIAL	MT		
TERM	TM		

2. 개체명 인식

2.1 전처리(Preprocessing)

개체명 인식은 앞서 말한 것과 같이 문장 내 어절을 기본 단위로 분석한다. 개체명 인식은 입력 받은 문장에 대하여

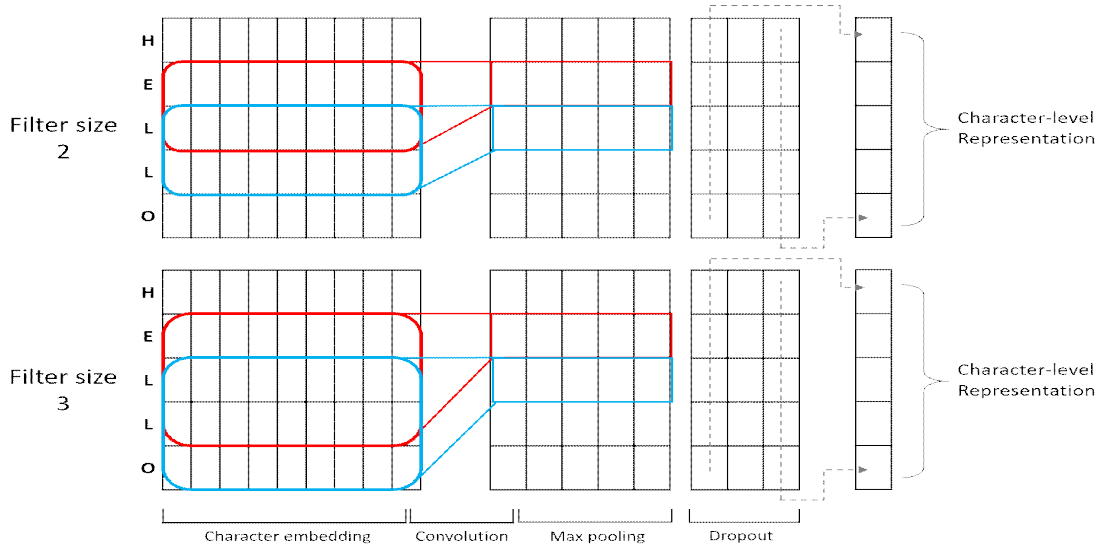


그림 2. 문자 단위 합성 곱 신경망(Char-CNN)
Fig. 2. Character Level Convolutional Neural Network

형태소 분석을 통해 형태소 단위로 분리(Tokenizing)하고 각 형태소 별로 개체명을 분류한다. 문장을 형태소 별로 분리하는 전처리 과정에서 개체명과 직접적으로 관련 있는 고유 명사가 분리 되는 문제가 발생 할 수 있다. 예를 들어, 섬유유연제와 같은 단어가 형태소 분석 결과 ('섬유', 'NNP'), ('유연', 'NNP'), ('제', 'XSN')로 분리 되고 다른 예로 트리하이드록시스테아린과 같은 고유 화학 물질에 대한 형태소 분석 결과 하나의 고유 명사 단위로 결정하지 못한다. 본 연구에서는 도메인과 관련한 사용자 사전을 구성하였고 전처리 과정의 형태소 분석 시 이용하였다.

2.2 단어 임베딩(Word Embedding)

본 연구에서 단어 임베딩을 위하여 형태소 사전, 출현 단어 사전, 개체명 태그 사전을 이용하여 원-핫(One-Hot) 인코딩을 진행하였고 단어를 벡터 차원으로 표현하는 분산 표상 모델인 Word2Vec 모델[10]과 Glove 모델[11]을 이용하였다. 그리고 문자 단위 합성 곱 신경망(Char-CNN : Character Level Convolutional Neural Network)[12]을 이용하여 단어 임베딩을 진행하였다.

문자 단위 합성 곱 신경망은 아래 그림 2와 같다. 입력 문장의 문자 시퀀스에 대하여 각 문자를 표현하는 차원을 추가하여 2차원 행렬로 표현하고 2D 합성 곱 필터와 합성 곱 연산을 진행한다. 필터의 크기를 변경하여 합성 곱 과정을 반복하고 이후 각 필터 크기에 대한 합성 곱 결과물에 맥스 풀링(Max-Pooling)과 드롭아웃(Dropout)을 거쳐 문자 단위 임베딩 벡터를 생성하였다.

위 과정을 통해 형태소, 출현 단어, 개체명 태그 사전 및 분산 표상 모델을 이용하여 단어 수준의 특징을 표현하고 문자 단위 합성 곱 신경망을 이용하여 문자 수준의 특징을 표현하였다.

2.3 인공 신경망 모델(Neural Network Model)

본 연구에서 입력 시퀀스 데이터에 대하여 각 타임 스텝 별 정보를 추출 할 수 있는 LSTM 모델을 확장한 Bi-LSTM을 이용하여 문장 수준의 특징을 표현하였다. 입력 문장의 각 단어 시퀀스 정보를 아래 그림 3과 같이 Bi-LSTM 모델에 입력한다. Bi-LSTM 모델을 통해 시퀀스 데이터에 대하여 각 타임 스텝 별 특징 정보를 추출하고 마지막 단계에서 모든 특징 정보를 종합하여 전체 문장에 대한 특징 벡터를 생성한다. 이를 통해 문장의 구문적 정보의 의미적 정보를 추출 할 수 있다. 다음으로 기본 인공 신경망 형태의 완전 연결 층(Fully Connected Layer)을 통해 문장 벡터의 크기를 줄인다. 그리고 드롭아웃 층을 거쳐 CRF 모델[13]을 통해 각 단어의 개체명 태그에 대한 확률을 계산하여 분류한다.

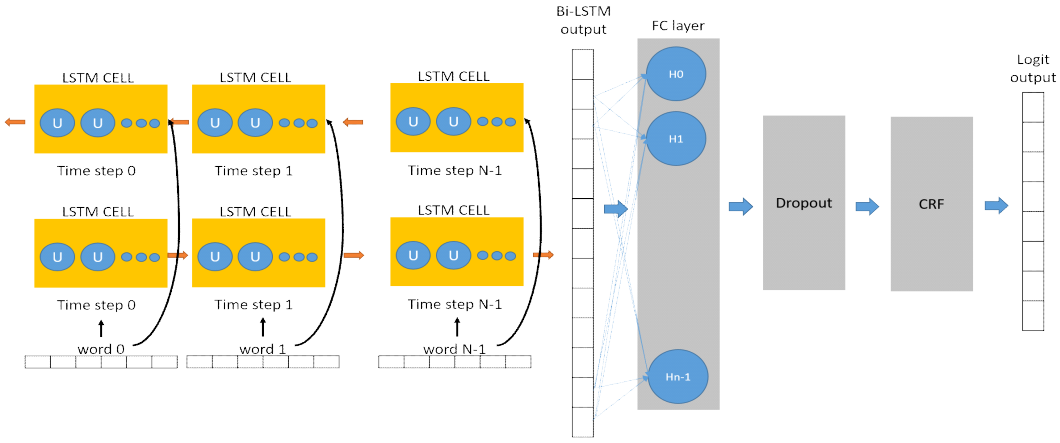


그림 3. Bi-LSTM 인공 신경망 모델
Fig. 3. Bi-LSTM Neural Network Model

Ⅳ. 실험 및 결과 논의

본 연구에서 실험에 사용하는 개체명 인식 모델의 구조는 아래 그림 4와 같다. 실험을 위해 전처리 과정에서 파이썬 한글 형태소 분석 라이브러리 KonNLPy의 Komoran 패키를 이용하고[14] 단어 임베딩 단계에서 파이썬 오픈소스 라이브러리 gensim의 word2vec 모델[15]과 스탠포드 대학교의 Glove 모델[16]을 이용하였다. 마지막으로 신경망을 구성하기 위해 구글(Google)에서 배포한 딥 러닝 라이브러리 텐서플로(Tensorflow)를 이용하였다[17].

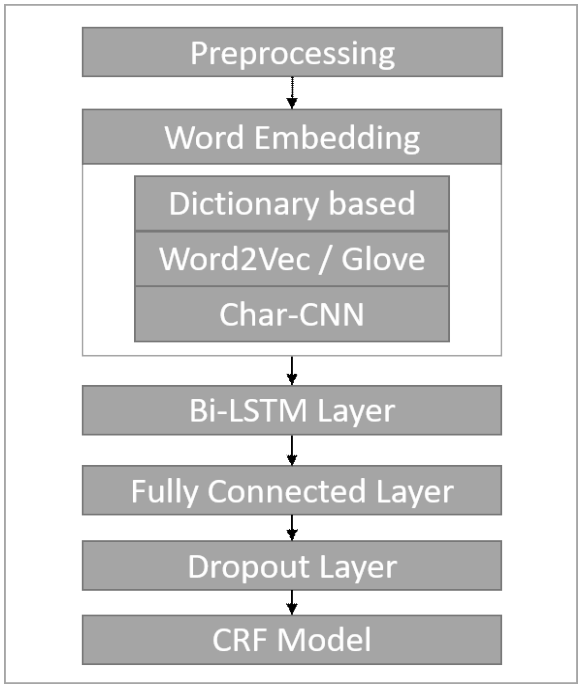


그림 4. 개체명 인식 모델 구조도
Fig. 4. Named Entity Recognition Model Architecture

1. 실험 데이터

본 연구에서 실험을 위해 생활 화학제품과 관련한 1500개 문장을 수집하고 이용하였다. 그리고 사전에 정의한 생활 화학제품 관련 개체명 기준을 적용하여 BIO 태깅을 진행하고 훈련용 데이터로 사용하였다. 전처리 과정에서 형태소 분석 시 사용하는 사용자 사전은 훈련 데이터 내 존재하는 생활 화학제품 관련 개체명과 생활 화학제품의 제조사, 제품명 등을 이용하여 전체 937개의 단어 사전을 구성하였다. 단어 임베딩을 위해 Word2Vec 모델과 Glove 모델을 사전에 학습하여 사용하였다. 단어 임베딩 모델의 사전 학습 데이터는 생활 화학제품과 관련 있는 뉴스 기사 40,000개를 이용하였다. Word2Vec 모델은 파라미터 기본 설정에서 윈도우 크기를 6으로 설정하고 Skip-gram 방법으로 20회의 반복 횟수를 통해 훈련하였다. Glove 모델도 파라미터 기본 설정에서 윈도우 크기와 훈련 반복 횟수만 Word2Vec 모델과 동일하게 설정하여 훈련하였다. 훈련 결과 Word2Vec 모델은 50749 단어 사전, Glove 모델은 125706개의 단어 사전을 구축하였다.

표 2. 제안 모델 실험 결과(정확도/F1점수)
Table 2. Proposed Model Experiment Result(Accuracy/F1 score)

			Filter A size 2,3		Filter B size 2,3,4,5		Filter C size 2,4,6,8	
User dictionary	Embedding model	Word dimension	Character dimension					
			100	150	100	150	100	150
use	Word2Vec	100	95.961 / 81.864	95.132 / 78.762	95.536 / 80.249	95.541 / 80.350	95.718 / 80.880	95.528 / 80.297
		150	96.427 / 83.431	95.154 / 79.512	95.548 / 80.541	95.397 / 79.747	95.761 / 80.317	95.504 / 80.528
	Glove	100	95.045 / 76.904	96.098 / 82.034	95.387 / 80.341	95.268 / 80.214	95.509 / 80.050	95.251 / 80.192
		150	95.467 / 80.150	95.169 / 79.369	95.476 / 80.744	95.31 / 80.019	95.667 / 80.763	95.395 / 80.400
not use	Word2Vec	100	93.083 / 65.513	92.16 / 61.414	92.383 / 61.300	92.61 / 61.518	92.346 / 63.154	92.995 / 61.124
		150	93.051 / 63.358	91.782 / 58.732	92.888 / 60.797	92.616 / 61.130	92.955 / 63.169	92.655 / 62.648
	Glove	100	92.561 / 63.424	92.888 / 63.599	92.598 / 63.566	92.675 / 62.827	92.282 / 60.491	92.437 / 61.862
		150	92.818 / 63.578	92.589 / 60.729	92.999 / 63.331	92.608 / 61.964	92.599 / 63.071	91.844 / 59.497

2. 실험

실험의 비교 데이터를 구성하기 위해 형태소 분석 시 사용자 사전을 사용한 데이터 세트와 사용하지 않은 데이터 세트를 준비하였다. 그리고 각 데이터에 대하여 두 가지 단어 임베딩 방법을 적용하고, 문자 단위 합성 곱 신경망의 서로 다른 필터 사이즈와 문자 표현 차원 크기를 다르게 적용하여 비교 실험하였다. 신경망을 구성하는 파라미터는 다음과 같다. 모델 훈련을 위해 배치 크기는 20으로 설정하고 훈련 반복 횟수를 20회 하였다. 문자 단위 합성 곱 신경망의 필터 수는 각 필터 사이즈 별로 128개, Bi-LSTM 신경망의 은닉 노드 수는 600개로 설정하였다.

위와 같이 여러 환경에서 생활 화학제품과 관련한 문장 데이터를 이용하여 개체명 인식 모델을 구성하고 훈련하였다. 실험 결과는 아래 표2와 같고 각 항목에서 '/' 기준 좌측 값은 정확도, 우측 값은 F1 점수이다.

3. 결과 논의

실험 결과 가장 눈에 띄는 것은 전처리 과정에서 형태소 분석 시 사용자 사전의 사용 유무이다. 표 3과 같이 사용자 사전을 이용하여 형태소 분석을 했을 경우 정확도와 F1 점수가 사용자 사전을 사용 한 경우와 사용자 사전을 사용하

표 3. 실험 결과 요약(사용자 사전-임베딩 모델)
Table 3. Summary of Experiment Result(user dictionary - embedding model)

User dictionary	Embedding model	Accuracy average	F1 score average
use	Word2Vec	95.601	80.540
	Glove	95.420	80.098
	total	95.510	80.319
not use	Word2Vec	92.627	61.988
	Glove	92.575	62.328
	total	92.601	62.158

표 4. 실험 결과 요약(임베딩 모델-필터 모양)
Table 4. Summary of Experiment Result(embedding model - filter shape)

Embedding model	Filter shape	Accuracy average	F1 score average
Word2Vec	Filter A	94.094	71.573
	Filter B	94.065	70.704
	Filter C	94.183	71.515
	total	94.114	71.264
Glove	Filter A	94.079	71.223
	Filter B	94.04	71.626
	Filter C	93.873	70.791
	total	93.998	71.213

지 않은 경우 차이가 큰 것을 알 수 있다. 또한 Word2Vec 모델과 Glove 모델의 실험 결과가 유사한 것을 알 수 있고 표 4를 통해 본 실험에서 여러 필터 사이즈에 대한 성능 차이는 적은 것을 알 수 있다.

사용자 사전을 이용한 경우와 이용하지 않은 경우의 성능을 분석 한 결과 다음과 같다. 전처리 과정에서 사용자 사전을 이용하지 않은 경우 개체명 인식 과정에서 형태소로 분리된 고유 명사에 대한 의미 파악과 개체명의 경계 인식이 어렵다. 또한 입력 시퀀스의 길이가 더 길어 모델의 복잡성을 증가시켰다. 이는 개체명 인식 과정에서 사용자 사전을 이용하여 분석하는 방법보다 좋은 결과를 얻을 수 없다. 따라서 기존의 방법인 훈련 데이터에 형태소 분석 후 개체명 인식 신경망 모델을 학습하는 방법보다 제안 방법인 형태소 분석 시 개체명과 관련한 고유명사에 대한 전처리 이후 신경망 모델을 학습하는 것이 성능이 좋다는 것을 알 수 있다.

V. 결 론

본 연구는 목적 지향 대화 시스템에서 자연어 이해를 위한 과정 중 개체명 인식 방법에 대해 기술하고 생활 화학제품 관련 도메인에 적용하였다. 효율적인 대화 정보 예측을 위해 제안한 개체명 인식 모델은 전처리 과정에서 사용자 사전, 도메인 관련 기 훈련 단어 임베딩 모델, 문자 단위 합성 곱 신경망을 이용하여 대화 객체의 속성을 표현하였다. 그리고 시계열 데이터 처리에서 유용하게 사용되는 Bi-LSTM 모델을 이용하여 문장을 표현하였고, 문장을 구성하는 순차적인 단어에 대하여 개체명 태그를 예측하였다. 제안 모델에 대해 실험을 진행하기 위해 생활 화학제품과 관련한 도메인을 이용하였다. 실험 결과 사용자 사전을 사용한 경우와 사용하지 않은 경우 성능의 차이가 뚜렷하고, 문자 단위 합성 곱 신경망의 필터 사이즈 및 수에 대해서는 큰 성능 차이를 보이지 않았다. 따라서 효율적인 대화 정보 예측을 위한 개체명 인식을 위해서는 특정 도메인에서 사용자 사전을 이용하여 전처리를 진행하는 것이 효율적이고, 문자 단위 합성 곱 신경망의 필터 수를 적게 하여 모델 복잡성을 줄이는 것이 좋다는 것을 알 수 있었다. 향후에는 문장을 표현하는 Bi-LSTM 신경망 모델과 CRF 예측 모델 부분에 대한 연구를 통해 좀 더 효율적인 대화 정보 예측을 위한 개체명 인식 모델을 기대 할 수 있다.

참 고 문 헌 (References)

[1] J. Huang, O. Kwon, K. Lee and Y. Kim, “A Chatter Bot for a Task-Oriented Dialogue System”, KIPS Transactions on Software and Data Engineering, Vol.6, No.11 pp.499-506, Nov 2017

[2] D. Nadeau and S. Sekine, “A survey of named entity recognition and classification”, Lingvisticae Investigationes, Vol.30, No.1, pp.3-26, Jan 2007

[3] S. Na and J. Min, “Character-Based LSTM CRFs for Named Entity Recognition”, Proceedings of KISS Korea Computer Congress, pp.729-731, Jun 2016

[4] S. Nam, Y. Hahm and K. Choi, “Application of Word Vector with Korean Specific Feature to Bi-LSTM model for Named Entity Recognition”, Human & Cognitive Language Technology(HCLT 2017), Oct 2017

[5] S. Hochreiter and J. Schmidhuber, “LONG SHORT-TERM MEMORY”, Neural Computation Archive, Vol.9, No.8, pp.1735-1780 , Nov 1997

[6] JL. Elman, “Finding Structure in Time”, Cognitive Science , Vol.14, No.2, pp.179-211, Mar 1990

[7] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami and C. Dyer, “Neural Architectures for Named Entity Recognition”, Proceedings of NAACL-HLT 2016, pp.260-270, Jun 2016

[8] X. Ma and E. Hovy, “End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF”, arXiv preprint arXiv:1603.01354, 2016

[9] TTA, “Tag Set and Tagged Corpus for Named Entity Recognition”, TTAK.KO-10.0852, 2015

[10] T. Mikolov, I. Sutskever, K. Chen, GS. Corrado, and J. Dean, “Distributed Representations of Words and Phrases and their Compositionality”, In Advances in Neural Information Processing Systems, pp.3111 - 3119, 2013

[11] J. Pennington, R. Socher, and C. Manning, “GloVe: Global Vectors for Word Representation”, In Proceedings of EMNLP-2014, pp.1532 - 1543, Oct 2014

[12] X. Zhang, J. Zhao and Y. LeCun, “Character-level Convolutional Networks for Text Classification”, Advances in Neural Information Processing Systems 28 (NIPS 2015), Vol.1, pp.649-657, 2015

[13] C. Sutton and A. McCallum, “An Introduction to Conditional Random Fields for Relational Learning”, Foundations and Trends® in Machine Learning, Vol.2, 2006

[14] E. Park and S. Cho, “KoNLPy: Korean natural language processing in Python”, The 26th Annual Conference on Human & Cognitive Language Technology, pp.133-136, Oct 2014

[15] Gensim Topic Modelling for Humans, https://radimrehurek.com/gensim (accessed Jul. 1, 2018).

[16] GloVe: Global Vectors for Word Representation, https://nlp.stanford.edu/projects/glove/ (accessed Jun. 1, 2018).

[17] Tensorflow, https://www.tensorflow.org (accessed Jun. 1, 2018).

저 자 소 개

고 명 현

- 2016년 : 세종대학교 디지털콘텐츠학과 학사
- 2016년 ~ 현재 : 세종대학교 디지털콘텐츠학과 석사과정
- ORCID : https://orcid.org/0000-0002-6036-4717
- 주관심분야 : 텍스트 마이닝, 기계학습, 딥러닝

김 학 동

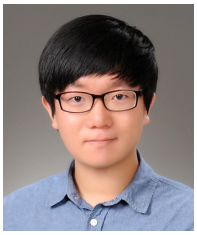
- 2016년 : 경성대학교 컴퓨터공학과 학사
- 2017년 ~ 현재 : 세종대학교 디지털콘텐츠학과 석,박사통합과정
- ORCID : https://orcid.org/0000-0003-3816-1224
- 주관심분야 : 머신러닝, 딥러닝, 자연어처리

임 헌 영

- 2017년 : 세종대학교 디지털콘텐츠학과 학사
- 2017년 ~ 현재 : 세종대학교 디지털콘텐츠학과 석사과정
- ORCID : https://orcid.org/0000-0002-8547-6248
- 주관심분야 : 컴퓨터 비전, 기계학습, 딥러닝

이 유 림

- 2018년 : 세종대학교 디지털콘텐츠학과 학사
- 2018년 ~ 현재 : 세종대학교 인공지능언어공학과 석사과정
- ORCID : https://orcid.org/0000-0001-8309-090X
- 주관심분야 : 텍스트 마이닝, 자연어 처리, 딥러닝



지 민 규

- 2018년 : 세종대학교 천문우주학과 학사
- 2018년 ~ 현재 : 세종대학교 소프트웨어융합학과 석사과정
- ORCID : https://orcid.org/0000-0002-3089-1452
- 주관심분야 : 텍스트 마이닝, 기계학습, 딥러닝

저 자 소 개

김 원 일

- 1981년 12월 ~ 1985년 7월 : ㈜대한항공 전산실 재무 시스템 개발원
- 1982년 : 한양대학교 공과대학 금속공학 학사
- 1987년 : 미국 일리노이주 서던일리노이대학교 컴퓨터 과학 학사
- 1990년 : 미국 일리노이주 서던일리노이대학교 컴퓨터 과학 석사
- 1994년 : 미국 인디아나주 인디애나 대학교 대학원 컴퓨터 과학 전공
- 2000년 : 미국 뉴욕주 시러큐스 대학교 대학원 컴퓨터 & 정보과학 공학 박사
- 2000년 1월 ~ 2001년 3월 : 미국 펜실바니아주 외인시 소재 Bhasha, INC Technical Staff (연구원)
- 2002년 3월 ~ 2003년 8월 : 아주대학교 정보통신전문대학원 BK 교수
- 2003년 9월 ~ 2017년 2월 : 세종대학교 전자정보공학대학 교수
- 2017년 3월 ~ 현재 : 세종대학교 소프트웨어융합대학 교수
- ORCID : https://orcid.org/0000-0002-1489-8427
- 주관심분야 : 인공지능, 지능형 시스템, 딥러닝 등