Journal of KIIT. Vol. 19, No. 1, pp. 23-32, Jan. 31, 2021. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2021.19.1.23

특허 기술 개체명 식별을 위한 단어 임베딩 기반 Bi-LSTM CRF 연구

김세빈*, 김장원**

A Bi-LSTM CRF Model with Word Embedding for Patent Technology Named Entity Recognition

Sebin Kim*, Jangwon Gim**

이 성과는 2018년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No.NRP-2018RICB600862).

요 약

최근 4차 산업혁명의 시대에서 다양한 신기술들이 등장함에 따라 가치 있는 신기술들에 대한 권리를 주장하기 위한 목적으로 특허 출원 및 발명 건수가 급증하고 있다. 따라서 선행 특허의 기술 검색 및 핵심 기술 추출 등의 연구들이 수행되고 있다. 특히 특허 문헌에 등장하는 기술 개체명은 특허 기술에 대한 분석 정확도 및 검색 재현율 성능 향상에 중요한 역할을 한다. 본 논문에서는 이러한 기술 개체명 식별을 위해 단어 임베딩 기반 특허 기술 개체명 식별 방법을 제안한다. 다양한 도메인 데이터를 대상으로 기술 개체명 식별 실험 결과 제안 방법이 가장 우수함을 확인하였다. 따라서 제안 모델을 통해 다양한 특허 기술 분야에 포함된 기술 개체명 식별 및 기계 학습용 한국어 사전 자동 구축을 위한 기반 연구로 활용 가능한 것으로 기대된다.

Abstract

With the emergence of various new technologies in the recent era of the 4th industrial revolution, the number of patent applications and inventions is rapidly increasing to claim the right to valuable new technologies. Therefore, studies such as technology search and core technology extraction of prior patents are being conducted. In particular, technology entity names appearing in patent documents play an important role in improving analysis accuracy and retrieval reproducibility performance for patent technologies. This paper proposes a patent technology entity name identification method based on word embedding for such technology entity name identification. It was confirmed that the proposed method was the best as a result of an experiment for identifying the technical entity name for various domain data. Therefore, it is expected that the proposed model can be used as a basis for the identification of technology entity names included in various patent technology fields and the automatic construction of a Korean dictionary for machine learning.

Keywords
named entitiy recognition, patent technologies, word embedding, Bi-LSTM CRF

* 군산대학교 소프트웨어융합공학과 학부생
- ORCID: https://orcid.org/0000-0001-8794-7747
** 군산대학교 소프트웨어융합공학과 교수(교신저자)
- ORCID: https://orcid.org/0000-0002-4480-7944

· Received: Oct. 16, 2020, Revised: Dec. 02, 2020, Accepted: Dec. 05, 2020
· Corresponding Author: Jangwon Gim
Dept. of Software Convergence Engineering, Kunsan National University, 558, Daehak-ro, Gunsan, Jeollabuk-do, Korea.
Tel.:+82-63-469-8916, Email: jwgim@kunsan.ac.kr

Ⅰ. 서 론

최근 다양한 기술들이 급속도로 발전하고 기술에 대한 소유권이 중요해지며 특허 출원 및 등록 건수가 급증하고 있다. 특허 문헌에는 해당 특허의 핵심이 되는 핵심 기술이나 해당 분야에서 최근 주목받는 기술 및 시스템 등에 대한 내용이 존재한다. 따라서 특허나 과학기술 문헌에서 핵심 기술을 추출하여 신기술이나 유망한 기술의 트렌드를 분석하기 위해 개체명 및 핵심어 추출의 중요성이 대두되었다[1].

개체명(Named entity)이란 문장에서 인명, 기관명, 지명, 날짜, 시간 등 고유한 의미를 가지는 단어로, 웹 뉴스, 과학기술 문헌 등에서 단어의 의미를 인식하거나 비정형 텍스트 분석에서 널리 사용될 수 있다[2][3]. 개체명을 식별하는 것을 개체명 식별(Named entity recognition)이라 하며 문서에서 개체명을 추출하여 미리 정의된 범주로 분류하는 작업을 말한다. 개체명은 정보 검색이나 질의응답, 관계 추출 등을 위해 사용되기 때문에 자연어 처리 분야에서 개체명 식별 연구가 진행되고 있다.

개체명 식별을 위한 전통적인 방법으로 사전 기반 방식이 있다. 사전 기반 개체명 식별 방법은 자원이 충분할 경우 유용한 방법이지만 다른 도메인이나 언어에 적용할 수 없고, 자원의 생성 및 유지에 시간과 비용이 많이 든다는 단점이 있다. 따라서 사전 기반 방식을 개선하기 위해 기계학습 방법과 딥러닝 기반의 방식이 제안되었다. 기존 개체명 식별을 위한 기계학습 방법은 조건에 따라 상황을 고려한 확률 값을 이용한 CRF(Conditional Random Field) 모델이 개체명 식별 성능에 좋기 때문에 CRF가 주로 사용된다[4][5]. 또한 딥러닝 기법을 활용한 개체명 식별 방법이 등장하며 최근에는 LSTM(Long Short-Term Memory) CRF[6][7] 혹은 Bi-LSTM CRF (Bidirectional LSTM CRF)모델을 이용한 방법이 우수한 성능을 보이고 있다[8]-[13].

개체명 식별에 대한 연구는 다양한 언어로 진행되고 있다. 영어 개체명 식별의 경우 대문자나 호칭 기호 등의 특징으로 개체명 식별에 높은 성능을 보인다. 그렇지만 한국어의 경우, 어순이 중요시 되는 영어와 다르게 어근에 접사가 붙어 의미와 문법적 기능이 부여되는 특징이 있다[14]. 예를 들어, 교착어의 경우, 원형이 '잡다'일 경우, 파생될 수 있는 단어들은 '잡히다', '잡히시다', '잡았다', '잡혔다' 등 다양하다. 그러므로 한국어 개체명 식별이 영어 개체명 식별보다 어렵기 때문에 한국어 개체명 식별의 정확도가 영어 개체명 식별보다 낮다. 또한 기존의 한국어 개체명 식별 모델에서는 한국어 개체명 식별을 위한 기본적인 자질이 풍부하지 않기 때문에, 기존 개체명 식별 방법에 형태소 분석, 전처리 등 다양한 기법을 사용하거나 다양한 자질을 추가하여 성능 향상을 목적으로 할 필요가 있다. 그 결과, 한국어 개체명 식별의 성능 향상을 위한 자질 추가 연구가 진행되었으며, 임베딩 자질을 추가하여 기존 개체명 식별 모델들보다 성능이 향상됨을 보였다[15]-[17].

개체명 식별에 대한 연구는 다양한 분야에서 진행되고 있다. [7]은 식품 도메인에서 식품 분야 전문 용어에 대한 개체명 식별 연구를 진행하였으며 [16]는 TV 도메인, 스포츠 도메인 등에서 사용되는 개체명 식별 연구를 수행하였다. [4][5]는 특허 분야에서 사용되는 용어에 대한 개체명 식별을 위해 기계학습 기반인 CRF 모델을 사용하였다. 그렇지만 영문 특허 데이터를 대상으로 실험이 진행되었기 때문에 한국어 특허 발명 문헌에서의 개체명 식별을 위한 추가적인 연구가 필요하다. 또한, 기존 개체명 식별에 사용된 CRF 모델뿐만 아니라 다양한 개체명 식별 모델들을 비교평가하고 한국어 특징을 반영한 특허 개체명 식별을 위한 연구가 필요하다.

기계 학습 기반 기술 개체명 식별을 위해서 한국어 개체명 사전이 필요하다. 그렇지만 기존 개체명 사전은 SNS, 웹 문서 및 영화 리뷰 등과 같이 일반적인 어휘가 다수 포함된 텍스트로부터 구축되었다. 그러므로 특허 문헌에서 새로운 기술을 표현하기 위해 등장한 신조어, 전문 용어 및 복합 어휘 형태의 기술명 식별을 위해 이러한 특허 문헌의 특징을 반영한 특허 기술 개체명 사전이 필요하다.

본 논문의 구성은 다음과 같다. 2장에서 기존 개체명 식별 방법을 소개하고 3장에서 제안 방법을 소개한다. 4장에서 실험 및 비교 평가 결과를 보이고, 마지막으로 5장에서는 결론 및 향후 연구를 설명한다.

II. 관련 연구

2.1 CRF

CRF는 입력 벡터에 대해 레이블 값을 출력하는 과정을 가지는 시퀀스 레이블링(Sequence labeling) 기법으로, 품사 판별이나 개체명 식별에서 자주 사용하는 기계학습 모델이다. CRF는 레이블의 인접성에 대한 정보를 바탕으로 레이블을 추측하는 특징을 가지고 있기 때문에 개체명 식별 분야에서 널리 사용되고 있다.

CRF 모델을 이용한 개체명 식별 연구로 [4]는 특허 문헌에서의 개체명 식별을 위해 CRF 모델에 우수 자질을 추가한 실험을 진행하였다. [5]는 특허 문헌에서 개체명 식별 연구로 수작업으로 진행한 개체명 식별의 성능과 CRF 모델을 이용한 개체명 식별의 성능을 비교하였다. 그 결과 수작업으로 진행한 개체명 식별과 CRF 모델을 사용한 개체명 식별의 성능이 크게 차이가 나지 않기 때문에 CRF를 이용한 개체명 식별 시스템이 수작업 태깅을 대신하여 실용적으로 활용될 수 있음을 보였다.

2.2 LSTM CRF

LSTM은 순차적인 데이터를 입력받아 결과 값을 도출하는데 사용하는 딥러닝 모델로, 짧은 시퀀스에 대해서는 결과가 좋지만, 긴 시퀀스에 대해서는 성능이 저조한 RNN의 단점인 장기 의존성 문제를 해결한 모델이다. LSTM CRF 모델은 문맥에 흐름을 살펴볼 수 있는 LSTM의 특징과 개체명을 식별하는 규칙을 해석할 수 있는 특징을 가진 CRF가 반영되어 개체명 태깅 문제에 뛰어난 성능을 보인다.

LSTM CRF 모델을 이용한 개체명 식별 연구에서 [6]는 문자 단위 LSTM CRF의 성능 평가를 위해 한국어 및 영어 개체명 식별 데이터를 사용한 실험을 진행하였으며, 실험 결과 기존 CRF 모델보다 높은 성능을 보였다. [7]는 LSTM CRF 모델을 식품 분야에 적용하여 식품 분야의 전문 용어에 대한 개체명 식별 진행한 연구로, 확장 단어 표상을 추가하여 식품 분야에서의 전문용어 개체명 식별에 높은 성능을 보였다.

2.3 Bi-LSTM CRF

기존 LSTM 모델은 입력을 시간 순서대로 받기 때문에 결과물이 직전 패턴을 기반으로 수렴하는 경향을 보이는 한계가 있다. 이 단점을 해결하기 위해 LSTM 계층에 입력 문장의 단어 순서를 반대로 처리하는 LSTM 계층을 추가한 Bi-LSTM이 제안되었다. 이러한 특징을 가지는 Bi-LSTM은 주변 정보를 동시에 학습할 수 있기 때문에, 균형 있게 학습할 수 있다.

Bi-LSTM과 CRF를 결합한 Bi-LSTM CRF 모델은 기존 LSTM CRF 모델의 단점을 해결함으로써 최근 다양한 개체명 식별 연구에서 주로 사용된다. [9]에서는 시퀀스 태깅을 위해 Bi-LSTM CRF 모델을 제안하여 POS 태깅 및 개체명 식별에서 높은 성능을 나타내는 것을 보였다. [13]은 한국어 개체명 식별을 위해 Bi-LSTM CRF 모델을 사용하고, 입력 단어의 표상 확장을 위한 사전 학습된 벡터들을 사용하여 기존 개체명 식별 모델보다 높은 성능을 보였다.

2.4 Word Embedding

단어 임베딩은 텍스트를 구성하는 하나의 단어를 고유의 벡터로 표현하는 방법으로, 유사한 단어일수록 가까운 거리에 위치하여 각 단어에 해당하는 벡터 값을 찾는 기법이다.

Word2Vec은 단어 임베딩 방법 중 하나로 단어 간의 유사성을 구해 말뭉치에 존재하는 단어 간의 관계를 구하는 방법이다. Word2Vec의 학습방법은 CBOW와 Skip-Gram 두 가지 방식이 존재한다. CBOW는 주변 단어를 이용하여 중심 단어를 예측하는 방법이고, Skip-Gram은 한 단어를 기준으로 주변 단어를 예측하는 방법이다. Skip-Gram은 기준 단어에서 앞, 뒤에 대한 문맥 정보를 알 수 있기 때문에 CBOW 모델보다 좋은 성능을 보인다[15].

개체명 식별을 위해 단어 임베딩 기법을 적용한 [16]은 개체명 식별의 성능 향상을 위해 CRF 모델에 Word2Vec 자질을 사용하고 있으며 우수한 성능을 보이고 있음을 보였다. [17]은 일반 도메인에서의 한국어 개체명 식별을 위해 Bi-LSTM CRF 모델

에 각 개체명 태그에 따른 한국어의 특징(LC-국, 동, 구 등으로 끝남 / TI-오전, 오후 등을 포함 등)을 담은 원 핫 벡터를 사용하였다. 또한 '하다' 접사의 동사형이 보존된 단어 임베딩 벡터를 사용하여 성능 향상을 보였다. 그렇지만 일반 도메인에서만 진행한 실험이라는 한계가 있으며, 개체명의 경계가 틀린 경우 개체명 사전을 활용한 실험에서 성능이 낮아지는 결과를 보였다.

따라서 본 논문에서는 한국어 및 특허 문헌의 기술 개체명 식별을 위해 단어 임베딩 기반 개체명 식별 방법을 제안한다. 다양한 한국어 데이터를 사용하여 기존 개체명 식별 모델들(CRF, LSTM CRF, Bi-LSTM CRF)의 성능을 비교하고, 이 실험을 통해 가장 좋은 성능을 보인 Bi-LSTM CRF 모델에 단어 임베딩 자질을 추가하여 한국어 및 특허 기술 개체명 식별의 성능을 높이는 방법을 제안한다. 실험에 사용하는 단어 임베딩 기법은 Word2Vec 중 Skip-Gram 방식을 사용하여 단어 벡터를 생성하고, 이를 Bi-LSTM CRF 모델의 입력으로 사용하여 개체명 식별을 수행한다.

III. 제안 방법

3.1 제안 모델 아키텍처

제안 모델의 전체 아키텍처는 그림 1에서 보인다. 입력 단계에서 한국어 데이터를 단어 단위로 입력하여 그림 1의 개체명 식별 모델에 전달한다. 단어 임베딩 기반 개체명 식별 모델에서는 개체명 사전을 참조하며 한국어 특허 기술 개체명 사전과 국립국어원 개체명 사전[18], NLP Challenge 개체명 사전[19]을 사용한다. 단어 임베딩 기반 개체명 식별 모델은 입력을 임베딩 벡터로 받으므로 개체명 사전을 참조하여 단어 임베딩을 수행하며, 이를 위해 Word2Vec 모델로 입력 단어를 벡터화한다. 벡터화된 단어는 Bi-LSTM의 입력으로 사용되며 입력으로 들어온 값은 정방향 LSTM과 역방향 LSTM을 거쳐 마지막 상태 값을 합친 값으로 출력된다. 출력된 값은 CRF의 입력으로 사용되며 CRF에서는 Bi-LSTM 모델에서 출력된 값이 입력이 되어 각 단어들의 개체명 태그를 예측한다. 그 결과 CRF에서는 개체명 태그가 부착된 단어인 개체명 후보 집합을 출력한다. 마지막으로 출력된 개체명 후보 집합을 정답 개체명과 비교한다. 이를 위해 개체명 사전을 참조하며, 최종적으로 식별된 특허 기술 개체명을 출력한다.

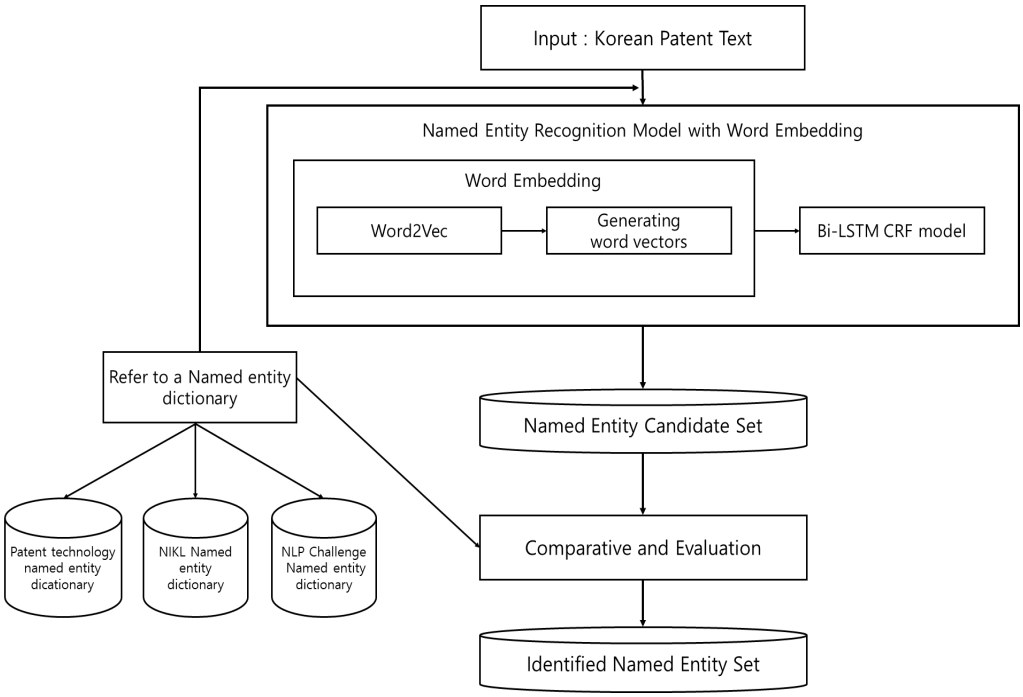

그림 1. 제안 모델의 아키텍처
Fig. 1. Architecture of proposed model

3.2 제안 모델의 구조

그림 2의 (a)부분은 전체 제안 모델 중 LSTM 셀을 표현한 것으로 RNN의 장기 의존성 문제를 해결하기 위해 사용된다. LSTM 노드는 1개의 메모리 셀과 3개의 게이트를 사용하며 메모리 셀은 과거 정보를 기억하는 기능으로 이전에 있던 상태를 기억하여 다음 상태로 전달한다. 게이트의 종류는 forget, input, output 게이트 3가지가 있다. LSTM 셀의 구조의 전체적인 수식은 식 (1)에서 (6)과 같다.

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (1)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (2)$$

$$\widetilde{C}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_C) \quad (3)$$

$$C_t = f_t * C_{t-1} + i_t * \widetilde{C}_t \quad (4)$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (5)$$

$$h_t = o_t * \tanh(C_t) \quad (6)$$

식 (1)은 forget 게이트, 식 (2), (3), (4)는 input 게이트, 식 (5), (6)은 output 게이트에 대한 수식을 나타낸다. f, i, C, o, h는 각각 forget 게이트, input 게이트, 메모리 셀, output 게이트, hidden state를 의미한다. x_t는 입력 값을 나타내며, h_{t-1}은 이전의 상태 값을 나타낸다. forget 게이트는 메모리 셀로부터 어떤 정보를 제거할지 결정하는 부분으로, 식 (1)은 이전 상태 값과 입력 값을 받아 시그모이드를 취해준 값으로 결정된다. 예를 들어 시그모이드 값이 0.8일 경우, 80%의 이전 값만 기억을 유지하고 20% 값은 삭제한다. input 게이트에서는 새로운 정보 중 어떤 정보를 메모리 셀에 저장할 것인지 결정하는 부분으로, 식 (2)는 입력 값을 메모리 셀에 포함시킬지 여부를 결정하는 부분이며, 식 (3)은 메모리 셀에 값을 얼마나 추가할지를 결정하는 부분이다.

식 (4)는 메모리 셀을 업데이트 하는 부분으로, f_t값을 곱해 메모리 셀에서 값을 삭제하고, $i_t * \widetilde{C}$ 값을 더해 메모리 셀에 값을 업데이트 한다. output 게이트는 메모리 셀에서 어떤 값을 출력할지 결정하는 부분으로, 상태 값과 메모리 셀의 값이 계산되어 출력된다. 식 (5)는 현재 메모리 셀 값을 출력에 반영함으로써 구하고, 식 (6)은 메모리 셀에서 얼마나 출력할지를 결정하는 부분이다.

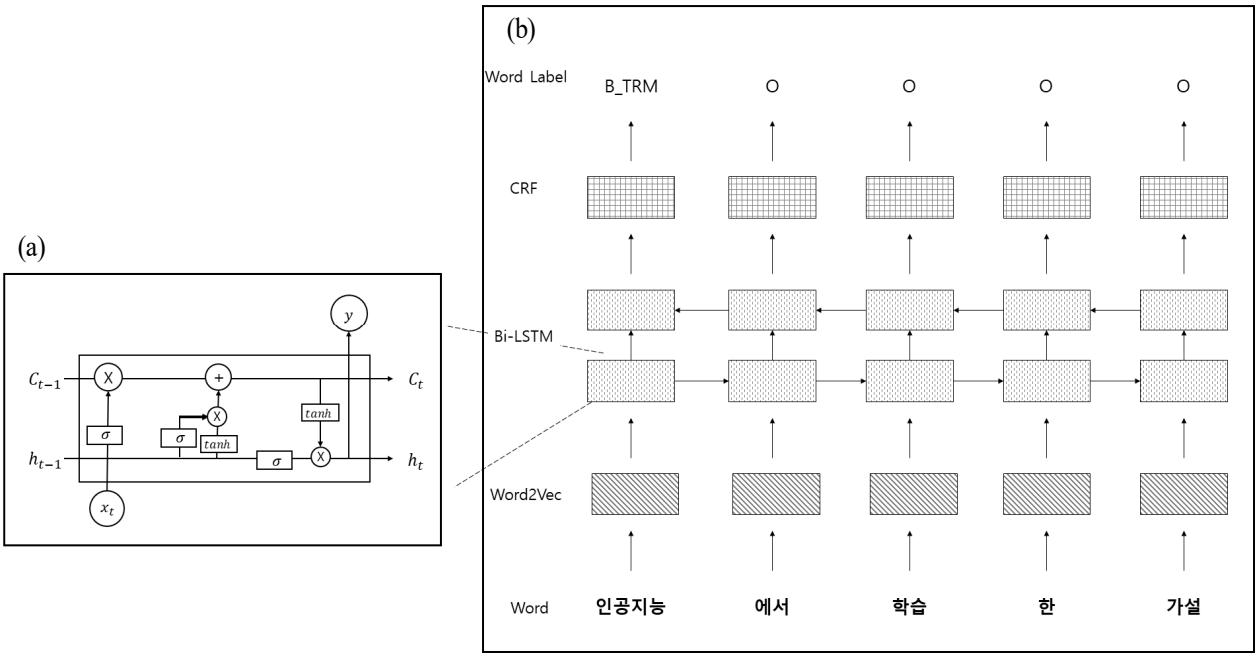

그림 2. 제안하는 단어 임베딩 기반 Bi-LSTM CRF 모델의 구조
Fig. 2. Structure of the proposed Bi-LSTM CRF model with word embedding

그림 2의 (b)는 앞서 기술한 LSTM (a)와 단어 임베딩이 포함된 Bi-LSTM CRF 모델의 구조를 보인다. 입력 단계에서 단어가 입력이 되면 Word2Vec를 통해 벡터화 단계를 거치고, 출력 벡터 값을 이용하여 Bi-LSTM의 입력 값으로 전달한다.

그림 3은 Word2Vec을 사용한 입력 단어의 벡터화 단계에 대한 알고리즘이다. LSTM 기반의 모델은 각 단어의 임베딩 벡터를 입력으로 받기 때문에 그림 3의 단계를 거쳐 입력 단어를 벡터로 변환한다. 먼저 제안 모델에서 임베딩 벡터를 사용하기 위해 학습 데이터를 사용한 Word2Vec 모델을 생성하였다. 그리고 생성한 Word2Vec 모델을 사용하여 해당 단어가 Word2Vec 모델에 있을 경우 임베딩 값을 반환하고 그렇지 않다면 'None'을 반환하여 최종적으로 L에 벡터화 된 단어를 저장한다.

LSTM 모델을 2개의 층으로 쌓은 Bi-LSTM 층은 Word2Vec를 통해 벡터화된 값을 입력으로 한다. Bi-LSTM 층은 단어 벡터를 양방향으로 받아 단어 간의 의존성을 고려한다. Bi-LSTM 모델을 거쳐 출력된 값은 CRF 층의 입력 값으로 전달되어 입력된 정보를 BIO 태그로 출력한다.

출력된 태그는 입력 단어의 개체명 유무를 판별하고, 개체명일 경우 정의된 개체명 태그 중 어떤 태그에 해당하는지 판별한다.

Algorithm : Word vectorization algorithm using Word2Vec

```
D : word dictionary (index : word)
M : Trained Word2Vec vector set
L : Array to store vectorized words
(zero array)

def get_vector(word):
    if word in M:
        return M[word]
    else:
        return None

for i, word in D:
    t = get_vector(word)
    if t is not None:
        L[i] = t
```

그림 3. Word2Vec 모델을 사용한 단어 벡터화 알고리즘
Fig. 3. Word Vectorization algorithm using Word2Vec model

Ⅳ. 실 험

4.1 실험 데이터

4.1.1 특허 기술 개체명 사전

특허 문헌에서의 기술 개체명 추출을 위한 실험에 필요한 한국어 특허 개체명 사전이 존재하지 않기 때문에, 실험을 위해 특허 데이터를 수집하고 개체명 사전을 구축하였다. 구축 방법은 그림 4와 같다. 한국특허정보원(KIPRIS)에서 제공되는 특허 중 4차 산업혁명과 관련된 7개 분야(사물인터넷, 빅데이터, 클라우드 컴퓨터, 인공지능, 3D 프린팅, 지능형 로봇, 지율주행 자동차) 특허들을 수집한다. 실험에 사용된 데이터는 특허에서 기술 용어 빈도수가 가장 높은 청구항을 대상으로 진행하였으며 문장 단위로 구분하여 개체명 태깅을 수행하였다.

전처리 작업은 특허 문헌에 적합한 MeCab 형태소 분석기를 사용하였다[20][21]. 그리고 자연어처리를 전공한 15명의 연구원들이 각 단어에 대해 BIO 태깅을 수작업으로 사전을 구축하였으며 TTA 정보통신용어사전과 IT 위키 및 백과사전을 이용하여 개체명 태깅된 기술 용어를 추가 검증하였다.

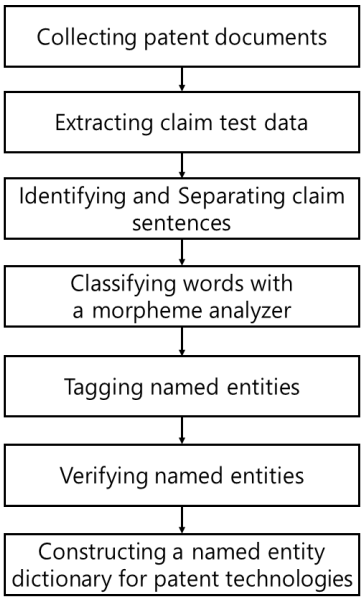

그림 4. 특허 기술 개체명 사전 구축 절차
Fig. 4. Patent technology named entity dictionary construction procedure

특허 문헌에서 추출하는 개체명 유형은 ETRI에서 정의한 총 15개의 개체명 종류에서 FLD(학문 분야), AFW(대상물), ORG(기관), MAT(물질), TRM(기술) 총 5개로 지정한다.

4.1.2 국립국어원 데이터

국립국어원 데이터는 국립국어원 언어정보나눔터에서 제공하는 개체명 식별용 말뭉치 데이터로, 2016년 국어 정보 처리 시스템 경진 대회에서 배포한 개체명 식별 말뭉치이다. 정의된 개체명 태그의 종류는 PS(사람), OG(기관), TI(시간), LC(장소), DT(날짜) 총 5가지 종류로 구분된다.

4.1.3 NLP Challenge 데이터

NLP Challenge 데이터는 네이버와 창원대가 2018년 개최한 한국어 자연어처리 기술대회에서 제공한 개체명 식별 말뭉치 데이터이다. 정의된 개체명 태그 종류는 PER(사람), FLD(학문 분야 및 이론), ORG(기관 및 단체), LOC(지역 명칭), AFW(인공물 및 대상물) 등 총 14개로 구분된다. 실험에 사용된 특허 데이터, 국립국어원 데이터, NLP Challenge 데이터 정보는 표 1과 같다.

표 1. 실험 데이터 정보
Table 1. Experimental data information

Data / Count	Patent data	NIKL data	NLP challenge data
Total number of sentences	9,891	4,056	90,000
Total number of words	598,651	138,383	1,063,571
Total number of unique words	5,416	15,004	331,196
Total number of words tagged with BIO	42,255	11,732	286,278
Number of unique named entity	10,351	6,181	176,138

4.2 실험 방법

기존 개체명 식별에서 주로 사용되는 세 가지 모델 (CRF, LSTM CRF, Bi-LSTM CRF)의 성능을 비교 평가하기 위해 3.3절에 기술된 세 가지의 한국어 데이터를 사용하여 실험을 진행한다. 그리고 제안 모델인 단어 임베딩 자질을 추가한 Bi-LSTM CRF 모델의 성능을 검증하기 위해 세 가지 데이터를 사용하고, Word2Vec 기법 중 Skip-Gram 방식을 이용한 단어 벡터를 사용하여 실험을 진행한다. 기존 개체명 식별 모델과 제안 모델의 실험을 위해 데이터셋을 8:2 비율로 학습데이터와 실험데이터로 분류하고 K-Fold를 적용하여 각 실험데이터(10회)에 대한 산술평균을 도출한다. 그리고 각 모델의 정확도를 측정하기 위해 정밀도(Precision), 재현율(Recall), F1-score를 사용한다.

4.3 실험 평가

표 1에서 정의한 다종 데이터를 기반으로 실험한 결과를 표 2에서 보인다. CRF 모델은 특허 데이터에서 개체명 식별 정확도가 가장 높으며 국립국어원 데이터에서 개체명 식별 정확도가 가장 낮다. 그 이유는 국립국어원 데이터의 양이 적고, 태깅된 기술 개체명 수가 적다. 또한 국립국어원 데이터에서 태깅된 개체명 태그의 형태에 대한 이슈가 존재한다. 예를 들어 '데이비드', '베컴'이 사람 이름 개체명일 경우 특허 데이터나 NLP Challenge 데이터는 각각 'B-PS', 'I-PS'로 구분하지만 국립국어원 데이터의 경우 'B-PS', 'I'로 구분된다. 따라서 'I'가 사람 이름 개체명으로부터 추출된 것인지 구별할 수 없다.

개체명이 연속된 어휘의 집합이라는 특징을 반영할 수 있는 LSTM CRF 모델의 경우 특허 데이터와 국립국어원 데이터에서 기존 CRF 모델의 결과보다 F1-score가 각각 9.93%p, NLP Challenge 데이터에서 기존 CRF 모델의 결과보다 정밀도와 재현율이 증가하여 F1-score가 각각 9.93%p, 10.19%p가 향상됨을 보인다. 그리고 NLP Challenge 데이터의 경우 4.21%p가 향상됨을 보인다.

Bi-LSTM CRF 모델은 세 가지 데이터 모두에서 기존 LSTM CRF 모델보다 개체명 식별 성능이 높음을 보인다. Bi-LSTM CRF 모델은 LSTM CRF 모델보다 특허 데이터의 경우 F1-score가 4.7%p 증가하였으며 국립국어원 데이터의 경우 2.23%p 증가하였다. 그리고 NLP Challenge 데이터의 경우 F1-score가 기존 LSTM CRF보다 2.69%p 증가함을 보인다.

그러므로 기존에 연구된 개체명 식별 방법에 대한 비교 결과 Bi-LSTM CRF 모델이 다양한 도메인에 데이터에 적용이 가능하며, 그 성능 또한 가장 높음을 알 수 있다.

표 2(우측 음영)는 단어 임베딩 기법을 적용한 Bi-LSTM CRF 결과를 나타낸다. 제안 방법의 실험결과 특허 데이터의 경우 0.16%p, 국립국어원 데이터의 경우 7.72%p, NLP Challenge 데이터의 경우 0.95%p 증가하였다. 특히 세 가지 데이터 중 특허 데이터에 대한 F1-score가 81.08%로 가장 높다. 실험결과를 통해 단어 임베딩 자질을 적용한 Bi-LSTM CRF 모델의 성능이 가장 뛰어남을 알 수 있다. 그러므로 실험에 사용된 다양한 데이터에서 제안 모델의 성능이 기존 개체명 식별 분야에서 SOTA 성능을 보이는 Bi-LSTM CRF 모델보다 우수함을 알 수 있다. 따라서 제안 방법은 일반적인 어휘가 포함된 데이터 뿐 아니라 다양한 분야의 전문적인 용어들이 출현하는 특허 문헌을 대상으로 한 개체명 식별에서도 그 성능이 우수하므로 기술 개체명을 추출에 다양하게 적용할 수 있음을 보인다.

Ⅴ. 결론 및 향후연구

최근 4차 산업혁명 시대에 다양한 기술들이 등장하며 기술 권리를 주장하기 위한 특허 출원 및 발명 건수가 증가하고 있다. 그러므로 선행 특허 검색 및 핵심 기술 도출을 위해 특허 문헌에서의 기술 개체명은 검색 성능 향상을 위해 중요한 역할을 하게 되었다.

이 논문에서는 한국어 특허 문헌에서의 기술 개체명 식별 모델을 제안한다. 기존 특허 도메인에서의 개체명 식별 연구는 영문 특허로만 진행되었다는 한계점이 있으며, 한국어 특허 문헌에서의 개체명 식별을 위한 개체명 사전이 존재하지 않는다. 그러므로 한국어 특허 문헌의 개체명 식별을 위해 특허 기술 개체명 사전을 구축 및 검증하였다. 또한, 기존에 개체명 식별에서 기계학습으로 주로 사용되는 CRF 모델뿐만 아니라 딥러닝 기반 개체명 식별 모델 (LSTM CRF, Bi-LSTM CRF)들의 성능을 비교하였다. 이를 위해 특허 데이터뿐만 아니라 국립국어원 데이터, NLP Challenge 데이터로 실험을 진행하여 Bi-LSTM CRF 모델의 성능이 가장 우수함을 알 수 있었다.

따라서 본 논문에서는 기존 개체명 식별 모델인 Bi-LSTM CRF보다 정밀도, 재현율을 높이기 위해 단어 임베딩 기법을 Bi-LSTM 모델의 추가 자질로 적용하여 기술 개체명 식별을 수행하였다. 제안 방법인 단어 임베딩 기반 Bi-LSTM CRF모델의 실험결과 제안 방법의 성능이 가장 뛰어남을 확인하였으며, 특히 특허 데이터에서의 개체명 식별의 정확도가 가장 높음을 보였다.

따라서 제안 방법을 다양한 한국어 도메인(영화 리뷰, 감성분류 등)에서 적용할 수 있을 뿐만 아니라 대용량의 특허 데이터를 대상으로도 기술 개체명을 식별할 수 있음을 확인하였다. 그러므로 특허, 논문 및 보고서와 같은 과학기술문헌으로부터의 개체명 자동식별 연구에 활용될 수 있음을 보인다.

향후 연구로 특허 기술 개체명 식별 성능 향상을 위해 특허 청구항의 계층적인 명세 구조를 고려한 자질 추가 연구가 필요하다. 이를 통해 특허 기술 개체명 식별의 성능을 높일 것으로 기대하며, 4차 산업분야 이외의 기존 학문분야에서 발명된 기술개체명 식별 및 대용량 특허 개체명 사전 구축에 활용할 수 있을 것으로 기대된다.

표 2. 제안 모델을 포함한 데이터 별 개체명 식별 모델의 정확도
Table 2. Accuracy of named entity identification model by data including the proposed model

Model / Data	CRF			LSTM CRF			Bi-LSTM CRF			Bi-LSTM CRF using W2V		
	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
Patent data	66.16%	66.78%	66.45%	78.01%	74.81%	76.38%	79.85%	82.02%	80.92%	79.94%	82.25%	**81.08%**
NIKL data	63.66%	42.1%	50.69%	62.8%	59.07%	60.88%	65.59%	60.11%	62.73%	67.9%	62.5%	**65.05%**
NLP challenge data	85.38%	50.89%	63.77%	92.83%	53.63%	67.98%	91.62%	57.51%	70.67%	92.31%	58.51%	**71.62%**

References

[1] Y. S. Lim, D. J. Seo, and Y. C. Jung, "Fine-tuning BERT Models for Keyphrase Extraction in Scientific Articles", JAITC, Vol. 10, No. 1, pp. 45-56, Jul. 2020.

[2] ETRI EXOBrain Korean language analysis Toolkit v3.0, http://tech.btp.or.kr/download.php?idx=4066. [accessed: Dec. 01. 2020]

[3] J. Y. Chang, "An Experimental Evaluation of Box office Revenue Prediction through Social Bigdata Analysis and Machine Learning", Journal of JIIBC, Vol. 17, No. 3, pp. 167-173, Jun. 2017.

[4] T. S. Lee, H. W. Chun, and S. S. Kang, "Recognizing Named Entities in Patent Documents by using Conditional Random Fields", Proceedings of KIISE Conference, pp. 612-613, Jun. 2014.

[5] T. S. Lee, S. M. Shin, and S. S. Kang, "Named Entity Recognition for Patent Documents Based on Conditional Random Fields", KIPS Transactions on Software and Data Engineering, Vol. 5, No. 9, pp. 419-424, May. 2016.

[6] S. H. Na and J. W. Min, "Character-Based LSTM CRFs for Named Entity Recognition", Proceedings of KIISE Conference, pp. 792-731, Jun. 2016.

[7] J. W. Min, H. J. Oh, and S. H. Na, "Character-Based LSTM CRFs for Food Domain Named Entity Recognition", Proceedings of KIISE Conference, pp. 500-502, Dec. 2016.

[8] D. J. Park and C. W. Ahn, "Named Entity Recognition using Bidirectional LSTM-CRF Combining Named Entity Ratio Dictionary", Proceedings of KIISE Conference, pp. 721-723, Jun. 2019.

[9] Z. Huang, W. Xu, and K. Yu, "Bidirectional LSTM-CRF Models for Sequence Tagging", arXiv:1508.01991, 2015.

[10] Y. J. Jang, T. H. Min, and J. S. Lee, "Using Stacked Bi-LSTM-CRF Ensemble Model", Proceedings of KIISE Conference, pp. 2049-2051, Jun. 2018.

[11] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural Architectures for Named Entity Recognition", arXiv preprint arXiv:1603.01360, 2016.

[12] J. P. C. Chiu and E. Nichols, "Named Entity Recognition with Bidirectional LSTM-CNNs", arXiv:1511.08308, 2015.

[13] H. Y. Yu and Y. J. Ko, "Expansion of Word Representation for Named Entity Recognition Based on Bidirectional LSTM CRFs", Journal of KIISE, Vol. 44, No. 3, pp. 306-313, Mar. 2017.

[14] What makes Korean natural language processing moredifficult, https://kh-kim.gitbook.io/natural-language-processing-with-pytorch/00-cover/04-korean-is-hell. [accessed: Dec. 01. 2020]

[15] Word2Vec : CBOW, Skip-gram, https://vhxpffltm.tistory.com/94. [accessed: Dec. 01. 2020]

[16] Y. S. Choi and J. W. Cha, "Korean Named Entity Recognition and Classification using Word Embedding Features", Journal of KIISE, Vol. 43, No. 6, pp. 678-685, Jun. 2016.

[17] S. H. Nam, Y. G. Choi, and K. S. 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.

[18] National Institute of Korean Language data, https://ithub.korean.go.kr/user/total/referenceView.do?boardSeq=5&articleSeq=118&boardGb=T&isInsUpd=&boardType=CORPUS. [accessed: Dec. 01. 2020]

[19] NLP Challenge data, https://github.com/naver/nlpchallenge/tree/master/missions/ner/data/train [accessed: Dec. 01. 2020]

[20] Y. J. Lee, S. B. Kim, H. S. Hong, and J. W. Gim, "Comparison and Evaluation of Morphological Analyzers for Patent Documents", Proceedings of KIIT Conference, pp. 264-265, Jun. 2019.

[21] Y. J. Lee, S. B. Kim, and J. W. Gim, "A study on technology name recognition method based on word embedding", Proceedings of KIPS Conference, pp. 750-781, Nov. 2019.

저자소개

김 세 빈 (Sebin Kim)

2021년 1월 현재 : 군산대학교
소프트웨어융합공학과(학부생)
관심분야 : 자연어 처리, 텍스트 마이닝, 딥러닝

김 장 원 (Jangwon Gim)

2005년 2월 : 상명대학교
컴퓨터소프트웨어공학과(공학사)
2008년 2월 : 고려대학교
컴퓨터학과(이학석사)
2012년 8월 : 고려대학교 컴퓨터·
전파·통신공학과(공학박사)
2013년 3월 ~ 2017년 3월 :
한국과학기술정보연구원 선임연구원
2017년 4월 ~ 현재 : 국립군산대학교
소프트웨어융합공학과 조교수
관심분야 : 텍스트 마이닝, 실시간 빅데이터 분석,
지식그래프, 메타데이터