특허문서의 한국어 화합물 개체명 인식

신진섭○, 김경민, 김성찬†, 이문용◇
한국과학기술정보연구원†, 한국과학기술원◇
{js.shin, kkmkorea, sckim}@kisti.re.kr† , munyi@kaist.ac.kr◇

Korean Chemical Named Entity Recognition in Patent Documents

Jinseop Shin○, Kyung-min Kim, Seongchan Kim† , Mun Yong Yi◇
KISTI† , KAIST◇

요 약

화합물 관련 한국어 문서는 화합물 정보를 추출하여 그 용도를 발견할 수 있는 중요한 문서임에도 불구하고 자연어 처리를 위한 말뭉치의 구축이 되지 않아서 활용이 어려웠다. 이 연구에서는 최초로 한국 특허 문서에서 한국어 화합물 개체명 인식(Chemical Named Entity Recognition, CNER)을 위한 말뭉치를 구축하였다. 또한 구축된 CNER 말뭉치를 기본 모델인 Bi-LSTM과 KorBERT 사전학습 모델을 미세 조정하여 개체명 인식을 수행하였다. 한국어 CNER F1 성능은 Bi-LSTM 기반 모델이 83.71%, KoCNER 말뭉치를 활용하는 자연어 처리 기술들은 한국어 논문에 대한 화합물 개체명 인식으로 그 외연을 확대하고, 한국어로 작성된 화합물 관련 문서에서 화합물 명칭뿐만 아니라 물성, 반응 등의 개체를 추출하고 관계를 규명하는데 활용 될 수 있을 것이다.

주제어: 한국어 화합물 개체명 인식, 특허 문서, LSTM, BERT

1. 서론

과학기술 분야에서 자연어 처리(Natural Language Processing)를 통해 문서에 포함된 화합물 명칭을 인식하고 그 용도를 밝혀내는 작업은 문서 기반의 발견(Literature Based Discovery) 연구를 수행하는데 있어서 필수적이며 단계이다. 그러나 한국어에 대한 충분한 CNER 코퍼스가 없어서 한국어로 작성된 문서에서 화합물에 대한 정보를 추출하고 화합물의 용도를 규명하는 연구는 본격적으로 수행되지 못하는 상황이다.

매년 의약품, 고분자, 소재 등 상당수의 새로운 화합물과 이들의 용도가 특허문서를 통해 공개되고 있다. 세계 5대 특허청(The five largest intellectual property offices; IP5)에는 미국, 유럽, 일본, 중국 및 한국의 특허청들이 속하며, 이들 특허청들은 세계 특허 출원의 80%를 처리한다[1]. 그리고 출원인이 각국의 특허청에 특허 출원을 할 때는 속지주의 원칙에 따라서 개별 국가의 언어로 작성된 출원서를 제출해야 한다. 미국특허청(USPTO)은 2001년부터 출원된 특허 출원서와 공보를 XML 형태로 공개하면서 문서에 포함된 테이블, 수식, 화합물 구조, 유전자 염기 서열 등의 데이터를 기계가독성이 높은 형태로 가공하여 포함시키고 있다[2].

그러나 한국 특허청(KIPO)에는 매년 약 20만 건의 특허가 한국어로 출원되고 있으나, KIPO는 수식, 화합물 구조 등의 데이터에 대한 정리와 공개에 대해서는 지원하지 못하고 있다. 한국어로 작성된 논문과 특허 등의 문서에 화합물 정보가 수록되고 있지만, 화합물 자연어 처리(Natural Language Processing)는 대부분 영어 문서를 다루는데 집중되어 있다. 그래서 영어가 아닌 한국 특허 문서에서 자연어 처리를 통해 문서에 포함된 화합물 정보를 인식하고 조직화 하는 연구와 기술 개발이 필요한 시점이다.

Lowe와 Sayle(2015)이 개발한 LeadMine 소프트웨어는 체계적인 화합물 명명법(systemic chemical nomenclature)을 기술하기 위한 규칙을 부호화하고, 화합물의 일반 명칭(trivial name)을 위한 사전을 함께 사용하여 화합물 명칭을 인식하였다. 규칙과 사전을 이용하는 이러한 방식은 정교한 규칙을 만들기가 어렵지만, 기계학습 방법에 비해 상대적으로 결과를 쉽게 해석할 수 있는 장점이 있다[3].

새로운 화합물이 논문이나 제품 카탈로그에 지속적으로 출판되거나 특허로 출원되고 있다. 그리고 동일한 화합물에 대한 다수의 명칭들이 존재하기 때문에 완전히 사전만을 기반으로 하는 화합물 인식에는 한계가 있다. 그래서 최근의 화합물 개체명 인식에는 Conditional Random Field (CRF), Long Short-Term Memory (LSTM) 등의 머신러닝과 딥러닝 방법들이 적용되고 있다[4].

인공신경망을 적용하여 기계번역의 성능을 향상시키기 위한 표현학습 방법으로 Recurrent Neural Network (RNN), Sequence to Sequence 등이 적용되었으며, 특히 Google은 RNN을 전혀 사용하지 않고 Attention만을 이용하는 Transformer 모델을 발표하였다. 이 모델은 Bilingual Evaluation Understudy Score (BLEU) 평가에서 영어-독일어 번역 테스크 (WMT 2014 Englishto-German translation task)에서 28.4점, 영어-프랑스어 번역 테스트 (WMT 2014 English-to-French translation task)에서 41.0점으로 최고 성능을 개신하였다[5]. Google은 Transformer의 Encoder만을 이용하는 BERT라는 언어 모델을 2018년에 공개하였으며, 이 모델은 질의응답 (SQuAD) 등의 테스크에서 당시 최고 성능(SOTA)을 달성하였다. BERT는 질의응답, 다음 문장 예측 등의 자연어 처리 분야에서 인간 보다 더 나은 정확도를 보이는 기계학습 모델이며, 구글이 2018년 말에 발표하

다. BERT는 전이학습 방법을 적용하고 있어서, 사전 학습(Pre-training)된 모델을 활용하여 미세 조정(Fine Tuning)을 통해 문서 분류, 개체명 인식 등의 문제를 기존의 방법들보다 높은 정확도로 해결 할 수 있다[6].

최근 들어 인공신경망을 기반으로 하는 자연어 처리 기술의 급격한 발전은 화학정보학 분야에서도 영어로 작성된 문서에 적용되어 그 성능이 크게 개선되었다. 영어 문서는 사람에 의해 표지된 화합물 개체명 인식(Chemical Named Entity Recognition; CNER) 말뭉치가 있어서 최신 기술들을 적용하여 성능을 향상 시킬 수 있었다[7]. 그러나 지금까지 한국어로 작성된 화합물 관련 문서에 기계학습 방법을 적용하지 못하는 가장 큰 이유는 화합물이 표지된 말뭉치가 구축되지 않아서이다.

이 연구에서는 화학분야 전공자들이 다단계 검토(multi-level review)를 통해 한국 특허 문서에서 화합물 NER 말뭉치를 구축하고, 양방향 LSTM(Bi-LSTM)과 한국전자통신연구원(ETRI)에서 공개한 한국어 언어모델(KorBERT)을 이용하는 전이학습을 통해 한국어 화합물 개체명 인식을 수행하였다[8].

2. 한국어 화합물 개체명 인식 말뭉치 구축

한국어 화합물 개체명 인식 말뭉치는 화학분야 한국 특허의 실시 예와 청구항(claim) 98,740문단에서 363,471개의 화합물 명칭을 태깅한 텍스트이다. 작업자들은 특허 문서의 본문 텍스트를 보면서, 화합물 명칭 부분을 마우스로 드래그하여 색을 칠하는 방식으로 태깅하였다. 말뭉치 구축은 온라인 작업도구에서 다단계 검토(multi-level review) 절차를 도입하여 데이터의 품질을 높였다. 즉, 5명의 관련분야 전공자들이 1차로 화합물 명칭 표지를 수행하고, 학학 전공자가 2차와 3차에 걸쳐서 검사 및 수정을 하였다.

말뭉치의 문단 안에서 화합물 명칭을 <FULL_NAME>과 </FULL_NAME> 태그로 감싸서 표시하였다. 화합물이 표지된 문장의 예시는 표 1과 같다. 말뭉치에 포함된 문장의 평균 길이는 121개 음소이며 가장 긴 문장은 965개의 음소로 구성되어 있다(그림 1).

화합물 개체명 인식을 위해서 말뭉치를 BIO 태그를 적용하였다. B 태그와 I 태그는 각각 화합물명 시작 음소와 화합물명 중간 음소를 나타내며, O 태그는 화합물명이 아닌 음소를 의미한다. 한국어 화합물 개체명 인식 말뭉치의 BIO 태그별 음소 분포는 표 2에 나타내었다. 한국어 화합물 개체명 인식을 위한 음소별 BIO 태그 예시는 표 3에 수록하였다.

표 2 화합물 명칭 태깅 문장 예시

플라스크에 <FULL_NAME>인산</FULL_NAME>을 투입하였다. 온도계를 붙인 플라스크에 <FULL_NAME>에리트리톨</FULL_NAME>과 <FULL_NAME>인산</FULL_NAME>을 투입하였다.

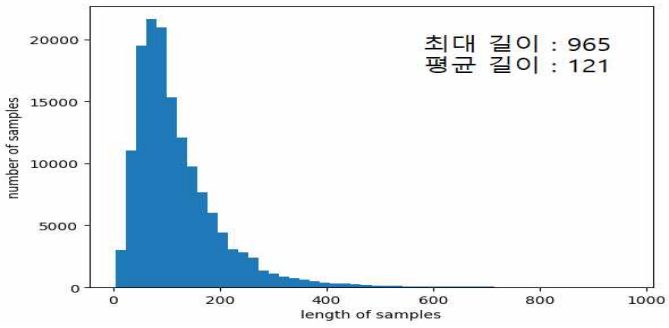

그림 1 한국어 화합물 개체명 인식 말뭉치의 문장 길이 분포

일반적으로 자연어 처리에서 토크나이징을 위한 구분자로 사용되는 공백, 쉼표, 작은따옴표, 대괄호, 소괄호 등의 문자들은 화합물 명칭에서는 중요한 의미를 가지는 구성 요소이다. 그래서 일반적인 토크나이징 방법을 적용했을 때 무시되는 이들 문자들을 보존하기 위해서 음소 단위로 토크나이징을 수행하였다. BIO 태그별 음소 분포는 표-3과 같다. 표-4에서처럼 음소별 BIO 태그가 적용된 결과 파일은 문장번호, 음소, BIO 태그로 구성되어 있다.

표 3 한국어 화합물 개체명 인식 말뭉치의 BIO 태그별 음소 분포

태그	의미	음소 수
B	화합물명 시작 음소	318,726
I	화합물명 중간 음소	4,297,180
O	비-화합물명 음소	13,267,580

3. 화합물 개체명 인식 실험

3.1 Bi-LSTM 모델 학습

화합물 개체명 인식을 위해서 일반적으로 개체명 인식에 많이 사용되는 Bi-LSTM모델을 구현하였다. 화합물 명칭에 존재하는 괄호, 쉼표 등의 특수문자를 고려하기 위해서 화합물 명칭이 표지된 문단을 음소 단위의 개체명 인식을 위한 코퍼스로 변환하였다(표 3). 학습은 NVIDIA TITAN Xp GPU에서 약 10 시간이 소요되었다.

3.2 BERT 모델 미세 조정

ETRI에서 공개한 한국어 BERT 언어모델 KorBERT를 사전 학습 모델로 활용하였다. 한국어 화합물 개체명 인식 말뭉치의 98,740문단에서 363,471개의 화합물명칭을 태깅한 텍스트를 이용해 NER 미세 조정 학습을 진행하였다. 학습, 검증 및 테스트를 위해 데이터를 60%, 20%, 20%의 비율로 분할하였다. 11,700개의 화합물 부분단어를 BERT의 기존 사전에 저빈도 단어 대신 추가 하였다. Korean Sentence Separator를 사용하여 문단의 문장을 분리하였다. []()- 등의 분리 문자를 추가 하여 토크나이저를 성능을 개선하였다. 학습은 NVIDIA GTX2080TI GPU에서 약 72 시간이 소요되었다.

표 4 한국어 화합물 개체명 인식을 위한 음소별 BIO 태그 예시

문장번호	음소	BIO 태그
1	플	O
1	라	O
1	스	O
1	크	O
1	에	O
1		O
1	인	B
1	산	I
1	을	I
1		O
1	투	O
1	입	O
1	하	O
1	였	O
1	다	O
1	.	O

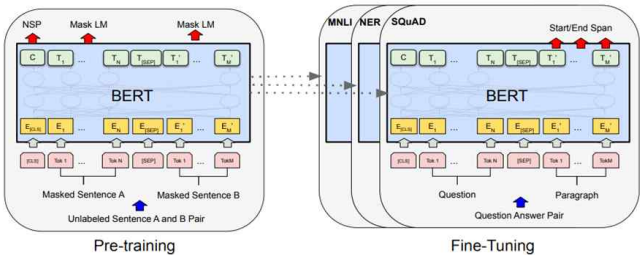

그림 2 BERT 모델의 사전학습과 미세조정[6]

4. 실험 결과 및 평가

한국 특허문헌에서 표지한 한국어 화합물 개체명 인식 말뭉치를 활용해 Bi-LSTM 모델과 KorBERT 모델을 이용하여 한국어 화합물 명칭에 대한 개체명 인식을 수행한 결과 F1 값은 Bi-LSTM이 83.71, KorBERT 모델이 86.05를 기록하였다(표 4). KorBERT를 NER을 위해 미세 조정한 모델이 2.34 높은 성능을 보였다.

표 5 모델별 화합물 개체명 인식 성능 비교

모델	Precision	Recall	F1
Bi-LSTM	79.00	89.00	83.71
KorBERT	84.25	87.94	86.05

5. 결론 및 향후 연구

이 연구에서는 최초로 한국 특허에서 화합물 개체명 인식(Chemical Named Entity Recognition)을 위한 말뭉치를 구축하였다. 구축된 CNER 말뭉치를 기본 모델인 Bi-LSTM과 KorBERT 사전학습 모델을 미세 조정하여 개체명 인식을 수행하였다.

향후에는 한국어 화합물 자연어 처리에 대한 연구 결과를 한국어 논문에 대한 화합물 개체명 인식으로 외연을 확대하고, 한국어로 작성된 화합물 관련 문서에서 화합물 명칭뿐만 아니라 물성, 반응 등의 개체를 추출하여 연결하고 관계를 규명하는데 활용 될 수 있을 것이다(그림 3).

그림 3 화합물 개체명 인식 및 관련 데이터 연결 예시

사사

본 연구는 문화체육관광부 및 한국콘텐츠진흥원의 2021년도 저작권보호 및 이용활성화 기술개발(R&D) 사업으로 수행되었음 (과제명: 학술자료 이미지(표, 도표 등)에 대한 저작권 검증 기술 개발, 과제번호: CR202104001, 기여율: 50%)

참고문헌

[1] Five Intellectual Property Offices, 2020, https://www.fiveipoffices.org/
[2] USPTO, Bulk Data Storage System (BDSS) Version 1.1.0, https://bulkdata.uspto.gov/
[3] Lowe and Sayle, LeadMine: a grammar and dictionary driven approach to entity recognition, Journal of Cheminformatics, 2015, 7(Suppl 1):S5, http://www.jcheminf.com/content/7/S1/S5
[4] Corbett and Boyle, Chemlistem: chemical named entity recognition using recurrent neural networks, Journal of Cheminformatics, 2018, https://doi.org/10.1186/s13321-018-0313-8
[5] A. Vaswani, N. Shazeer et al., Attention Is All You Need, NIPS Proceedings, 2017, arXiv:1706.03762 [cs.CL]
[6] J. Devlin, M.W. Chang et al., BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Proceedings of NAACL-HLT, 2019.
[7] Krallinger et al. The CHEMDNER corpus of chemicals and drugs and its annotation principles, Journal of Cheminformatics, 2015, http://www.jcheminf.com/content/7/S1/S2
[8] 공공 인공지능 오픈 API·DATA 서비스 포털, ETRI, 2019.12. http://aiopen.etri.re.kr