국내 특허 문헌 내 화학 용어 자동 추출을 위한 알고리즘 연구

이하영O, 김홍기*, 박진우*
한국특허정보원 R&D센터O
한국특허정보원 R&D센터*
e-mail: {youngsllO, khkcap*, znu808*}@kipi.or.kr

A study on the Algorithm for automated extraction for chemical term in Korean patents

Hayoung LeeO, Hongki Kim*, Jinwoo Park*
R&D Center, Korea Institute of Patent InformationO
R&D Center, Korea Institute of Patent Information*

● 요 약 ●

본 논문에서는 열 및 전기특성 플라스틱 복합수지와 한글에 특화된 인공지능 기술을 개발하기 위한 조성/물성 정보 복합수지 지식베이스를 구축하고자 국내 특허 문헌에서 화학 용어를 추출하고자 한다. 이를 위해 전문용어가 많이 쓰인 특허 문헌의 특수성을 고려하여 UIMA(Unstructured Information Management Architecture) 규칙 기반의 라이브러리를 사용해 한국어 화학 용어 코퍼스를 구축하고 이를 기반으로 딥러닝 알고리즘 중 하나인 Bidirectional LSTM-CRF를 기반으로 특허 문헌에서 화학 용어를 자동으로 추출하는 알고리즘을 연구하고자 한다.

키워드: 인공지능(Artificial intelligence), 특허(patent), Bidirection LSTM-CRF, UIMA

I. Introduction

최근 인공지능에 대한 관심이 높아지면서 다양한 분야에 이를 적용하기 위한 노력이 늘어나고 있다. 그중 화학 분야에서 인공지능을 적용하는 사례는 해외에서는 점점 늘어나는 추세지만 우리나라에서는 한글의 특수성으로 인하여 연구가 진행되기 힘든 상황이다. 따라서 본 논문에서는 열 및 전기특성 플라스틱 복합수지와 한글에 특화된 인공지능 기술을 개발하기 위한 조성/물성 정보 복합수지 지식베이스를 구축하고자 국내 특허 문헌에서 화학 용어를 추출 하려 한다. 이를 위해 전문용어가 많이 쓰인 특허 문헌의 특수성을 고려해 화학 용어에 대한 규칙을 정의하려고 한다. 또한 정의된 규칙을 활용하여 Apache에서 개발한 라이브러리인 UIMA(Unstructured Information Management Architecture)를 사용해 한국어 화학 용어 코퍼스를 구축하고 이를 기반으로 딥러닝 알고리즘 중 하나인 Bidirectional LSTM-CRF를 기반으로 특허 문헌에서 화학 용어를 자동으로 추출하는 알고리즘을 연구하고자 한다.

II. Preliminaries

1. Related works

1.1 UIMA(Unstructured Information Management Architecture)

UIMA는 Apache에서 개발한 라이브러리로 지식정보서비스 개발에 필요한 비정형데이터의 태깅과 메타데이터 생성을 위한 프레임워크이다. Java, C++ 프레임워크를 지원하며, C/C++, Perl, Python, TCL에서 사용할 수 있다. UIMA는 annotator 구성요소를 구성하고 파이프라인을 실행하며 사용자 정의에 의해 구조화되지 않은 정보를 구조화하여 보여준다. UIMA는 분석 엔진들(Analysis Engines)을 기본 아키텍처로 하고 있으며 문서를 분석하고 문서에 대한 속성을 추측하여 기록한다. 또한 분석 결과는 문서의 메타데이터 정보로 사용된다. 분석된 정보들은 공통 분석 구조(Common Analysis Structure)를 통해 보여주며 데이터 구조를 object, property, value로 구조화한다. [1] 따라서 본 논문에서는 화합물명명법을 기준으로 화학 용어에 대한 규칙을 정의하여 UIMA를 통해 한국어 화학 용어 코퍼스를 구축하고자 한다.

1.2 Bidirectional LSTM-CRF 네트워크

Bidirectional LSTM 네트워크는 과거와 미래의 데이터에 접근을 가능하게 하기 위한 알고리즘이다. 기존의 LSTM의 은닉계층은 과거에서 온 정보만을 가지고 있고 미래에 대한 정보는 가지고 있지 않았다. 이런 문제를 해결하기 위해 Bidirectional LSTM(Dyer et al., 2015)이 사용된다. Bidirectional LSTM은 과거와 미래의 정보를 각각 저장하기 위해 두 개의 분리된 은닉 계층을 각 시퀀스 앞뒤로 제시한다. [2] Fig·1에서 볼 수 있듯이 Bidirectional LSTM-CRF 네트워크는 LSTM-CRF 네트워크의 발전된 형태로 앞서 언급한

LSTM 네트워크의 문제점을 해결한 개체명 인식 모델이다. [3] 본 논문에서는 UIMA를 통해 만들어진 코퍼스를 학습데이터로 활용하여 Bidirectional LSTM-CRF 네트워크로 특허 문헌 내의 화학 용어를 추출할 것이다.

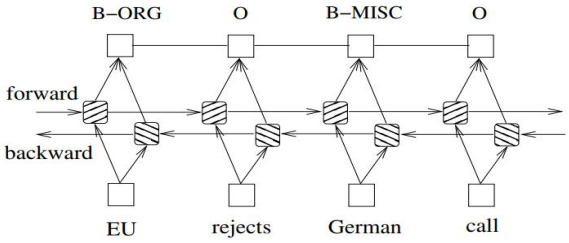

Fig. 1. A Bidirectional LSTM-CRF model

III. The Proposed Scheme

1. 특허 문헌 내 화학 용어 추출 프로세스

본 논문의 대상이 되는 화학 용어는 특허 IPC 분류 C08K에 해당하는 특허 문헌 중 열 및 전기 특성 플라스틱 수지와 관련된 280건의 특허를 기준으로 한다. 대상 건에서 화학 용어를 자동으로 추출하기 위해 화합물명명법을 기준으로 화학 용어를 분석해 규칙을 정의하여 UIMA를 사용하여 한국어 화학 용어 코퍼스를 구축한 후 Bidirectional LSTM-CRF를 이용하여 특허 문서 내에서 화학 용어를 자동으로 추출하고자 한다.

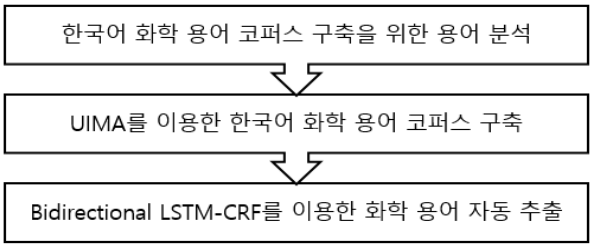

Fig. 2. A process of automated extraction for chemical substances.

2. 한국어 화학 용어 코퍼스 구축

2.1 UIMA 정규표현식 정의를 위한 화학 용어 분석

UIMA는 정규표현식을 기반으로 사용자가 정의한 패턴을 문서에 적용하여 추출해내는 라이브러리다. 따라서 이에 필요한 정규표현식을 작성하기 위해 matweb에서 제공하는 레진, 필러 분류의 화학 용어를 한국어로 번역한 후 화합물명명법을 기준으로 화학 용어를 접두·접미어, 전체가 한 단어로 활용될 수 있는 용어의 세 부분으로 나누어 정규표현식을 정의하였다.

2.2 UIMA를 이용한 한국어 화학 용어 코퍼스 구축

정의한 정규표현식은 Fig.3과 같이 접두어, 접미어, 단어의 각각의 분석 엔진을 구축하는 데 사용되었다. 접두어 엔진은 총 46개의 정규표현식이 정의되어 있으며 접미어 엔진은 총 30개의 정규표현식이 정의되어 있다. 마지막으로 단어 엔진에는 2385개의 정규표현식이 정의되어 있으며 이렇게 구축된 엔진들은 화학 용어라는 하나의 엔진으로 합쳐진다. UIMA 라이브러리는 이렇게 완성된 엔진을 사용해 접두어-단어-접미어를 하나로 묶어 단어로 인식해 낸다.

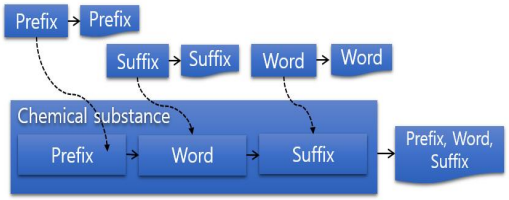

Fig. 3. A structure of analysis engine

하지만 띄어쓰기로 구분된 하나의 용어를 인식하기 위한 정규표현식과 숫자, 특수기호가 들어간 정규표현식을 정의하는 데 어려움이 있어 완벽한 화학 용어 추출은 어렵다. 따라서 본 논문에서는 UIMA를 이용해 직접적으로 화학 용어를 추출하는 것이 아니라 화학 용어 자동 추출 학습데이터로 사용할 수 있는 코퍼스 구축을 위해 사용한다.

3. Bidirectional LSTM-CRF를 이용한 특허 문헌 내 화학 용어 추출

3.1 학습데이터 구축

우선 대상 특허 문헌 280건에 대하여 형태소 분석을 진행하였다. 형태소 분석을 결과로 각 문서에 대한 BIO(Begin Inside Outside) 학습데이터를 구축하였다.

또한 학습데이터에 따른 모델 성능을 비교하기 위해 대상 특허 문헌 280건 전체를 띄어쓰기 단위로 잘라낸 후 UIMA에서 추출된 용어를 기준으로 잘못 추출된 용어, 추출되지 않은 용어, 연결어 돼야 되는 용어 등을 확인하여 BIO 학습데이터를 구축하였다.

Sentence	Word	POS	Tag
Sentence: 47959	Indian	JJ	B-gpe
Sentence: 47959	forces	NNS	O
Sentence: 47959	said	VBD	O

Fig. 4. An example of general BIO Tagging

3.2 실험 및 결과

3.2.1 형태소 분석한 결과로 구축한 학습데이터 기반 실험

본 연구에서는 형태소 분석을 기반으로 구축된 학습데이터를 활용하여 Bidirectional LSTM-CRF 네트워크를 이용한 모델 학습을 시행하였다. 활성 함수는 ReLU를 사용하였고 최적화를 위해 RMSProp함수를 사용하였으며 가장 좋은 성능을 보이는 하이퍼파라미터를 찾기 위해 파인 튜닝을 진행하였다. Test size, Dropout, Epoch 등 총 3가지의 하이퍼파라미터를 대상으로 실험을 진행하였다.

Test size가 30%일 때 실험 결과는 Table 1과 같으며 Epoch가 100, Dropout이 0.2일 때 F1-score 52.5%로 가장 높은 성능을 보였다. Table 2는 Test size가 25%일 때 실험 결과이며 Epoch가 75, Dropout이 0.2일 때 F1-score가 54.8%로 가장 높은 성능을 보였다.

Table 1. experiment results(test size: 30%)

Epoch	Dropout	F1-score (%)
50	0.1	42.3
50	0.15	45.5
50	0.2	50.0
75	0.1	51.7
75	0.15	50.5
75	0.2	52.1
100	0.1	50.2
100	0.15	51.0
100	0.2	52.5
150	0.1	48.4
150	0.15	49.7
150	0.2	48.9

Table 2. experiment results(test size: 25%)

Epoch	Dropout	F1-score (%)
50	0.1	50.2
50	0.15	44.0
50	0.2	54.3
75	0.1	49.3
75	0.15	52.2
75	0.2	54.8
100	0.1	51.4
100	0.15	51.8
100	0.2	52.6
150	0.1	52.1
150	0.15	52.1
150	0.2	51.0

하지만 전체적으로 약 42%~55% 사이의 성능을 보여 실제 예측에서는 좋은 결과를 나타내지 못했다. 이와 같은 결과는 형태소 분석을 했을 때 한 단어 임에도 불구하고 여러 개의 형태소로 쪼개지는 현상이 발생해 화학 용어를 제대로 추출할 수 없었기 때문으로 사료된다. 따라서 새로운 기준으로 학습데이터를 구축하여 실험을 진행하였다.

3.2.2 UIMA로 추출된 화학 용어를 활용해 구축한 학습데이터 기반 실험

UIMA를 사용해 추출된 데이터를 활용하여 특허 문헌에 친화적인 학습데이터를 구축한 후 Bidirectional LSTM-CRF 모델학습을 실시하였다. 모든 조건은 형태소 분석한 결과로 구축한 학습데이터를 사용했을 때와 같으며 파인 튜닝을 위한 실험을 진행하였다. Table 3은 Test size가 30%일 때의 실험 결과이며 Epoch가 150, Dropout이 0.15일 때 F1-score가 78.1%로 가장 높은 성능을 보였다. Table 4는 Test size가 25%일 때의 실험결과로 Epoch가 100, Dropout이 0.15일 때 F1-score가 80%로 가장 높은 성능을 보였다.

Table 3. experiment results(test size: 30%)

Epoch	Dropout	F1-score (%)
50	0.1	75.5
50	0.15	76.5
50	0.2	74.3
75	0.1	76.2
75	0.15	78.7
75	0.2	74.3
100	0.1	74.9
100	0.15	77.1
100	0.2	75.8
150	0.1	74.9
150	0.15	78.1
150	0.2	74.2

Table 4. experiment results(test size: 25%)

Epoch	Dropout	F1-score (%)
50	0.1	78.1
50	0.15	77.5
50	0.2	77.9
75	0.1	78.1
75	0.15	73.5
75	0.2	79.6
100	0.1	76.4
100	0.15	80.0
100	0.2	74.5
150	0.1	77.8
150	0.15	77.4
150	0.2	77.7

본 연구에서 실험한 결과를 토대로 특허 문헌에서 화학 용어를 추출할 때 UIMA를 이용한 화학 용어 코퍼스를 사용하여 Bidirectional LSTM-CRF 네트워크를 사용하는 것이 효과적일 것이라 사료된다.

IV. Conclusions

본 논문에서는 특허 문서 내의 한국어 화학 용어 자동 추출을 위해서 UIMA 라이브러리와 Bidirectional LSTM-CRF 네트워크를 사용한 알고리즘을 구현하고 성능 개선을 위해 학습데이터에 대한 연구와 하이퍼파라미터에 대한 실험을 진행하였다. 결론적으로 형태소 분석을 결과로 구축한 학습데이터를 이용한 모델 학습보다 UIMA를 이용해 추출된 화학 용어를 기반으로 만들어진 학습데이터를 사용한 모델 학습의 결과가 F1-score 기준으로 약 30% 정도 좋은 결과를 보여줬다. 또한 파인 튜닝을 통해 Test size 25%, Epoch 100, Dropout 0.15일 때 가장 좋은 F1-score가 나타난다는 것을 알 수 있다.

향후 연구에서는 본 논문의 알고리즘을 기반으로 실제 특허 문헌에서 자동으로 발명의 상세한 설명 부분에 나타난 조성, 물성, 부가정보, 도표 등을 추출하여 실험 정보에 대한 지식베이스를 구축할 예정이다.

ACKNOWLEDGEMENT

이 논문은 2019년도 정부(산업통상자원부)의 재원으로 한국산업기술평가관리원의 지원을 받아 수행된 연구임
(No.20000391, 열 및 전기특성 플라스틱 복합수지의 빅데이터 구축과 인공지능 기술을 활용한 정확도 90% 이상의 조성/물성 예측 및 용도 추천 플랫폼 개발)

REFERENCES

[1] Apache UIMA™ Development Community, “UIMA Overview & SDK Setup”, Version 3.0.2

[2] Xuezhe Ma and Eduard Hovy, "End-to end Sequence Labeling via Bi-directional LSTM-CNNs-CRF", arXiv:1603.01354v5 [cs.LG] 29 May 2016.

[3] Zhiheng Huang, Wei Xu, Kai Yu, “Bidirectional LSTM-CRF Models for Sequence Tagging”, arXiv:1508.01991v1 [cs.CL] 0 Aug 2015.

[4] Abhinav Walia, “Annotated Corpus for Named Entity Recognition-Feature Engineered Corpus annotated with IOB and POS tags version 4”