전력 도메인 특화 정보 추출을 위한 개체명 인식기 개발

하유이, 황호진, 송종협
한국전력 데이터사이언스연구소

Named Entity Recognition in Electric Power Field for Extracting Domain-Specific Information

Yui Ha, Hojin Hwang, JongHyup Song
Data Science Lab., Korea Electric Power Corporation

Abstract - 도메인 특화 정보 추출 기술은 챗봇 등 고도화된 AI 서비스 개발 및 대규모 학습용 DB 구축 등을 위한 필수 기반 기술이다. 그러나 전력 분야의 정보 추출을 위한 개체명 유형 정의 및 해당 기술 개발에 대응하는 학습 데이터는 미비한 실정이다. 이 연구는 9개의 전력 도메인 개체명 유형을 구축하고 이를 분류할 수 있는 KoBERT 기반 개체명 인식 모델을 제안한다. 이를 통해 일반 기술을 통해 추출할 수 없는 전력 분야 특화 정보(예: 설비명, 배전 선로명 등)를 인식하고, 각종 AI 개발을 위한 전력 분야 특화 DB 구축 등에 기여하는 것을 목표로 한다.

1. 서 론

최근 인공지능 기반 도메인 맞춤형 챗봇, 정보검색, 질의응답, 지식 관리 시스템 등의 개발 및 각종 산업 분야로의 도입이 활발하게 이뤄지고 있다. 해당 기술 개발을 위해서는 각 도메인에 특화된 정보를 포함하는 대규모 학습용 DB 구축이 선행되어야 한다. 한국전력 사내 각종 보고서, 지침서, 공문 및 교재 등의 방대한 문서 데이터는 각종 설비, 자재, 공사명세, 업체/인력 등의 정보를 포함하고 있으며, 이러한 전력 분야의 비정형 텍스트 데이터는 빠르게 증가하는 추세이다. 그러나 문서, 엑셀 등으로 구성된 비정형 포맷의 텍스트 데이터를 도메인 전용 모델 개발을 위한 대규모 학습 DB로 구축하기 위해서는 도메인 특화 정보 추출 기술 개발이 필요하다.

개체명 인식(Named Entity Recognition, NER) 기술은 문장 내 고유한 의미를 지니는 단어 또는 어절을 사전에 정의한 유형으로 인식하여 분류 및 추출하는 자연어처리 기법이다. 입력된 문자열의 각 단어에 해당하는 유형 태그를 출력하는 다중클래스(Multiclass) 분류 작업의 성격을 지닌다. 기관명, 지역, 시간, 날짜, 인물 등의 일반적인 의미를 가지는 개체명을 인식하는 기술이 대중화 되어 있으며, Bio, IT 등의 전문 분야의 용어를 분류하기 위한 도메인 특화 개체명 인식 기술도 개발되고 있다. 개체명 인식은 엔터티 링킹(Entity linking)과 더불어 텍스트 정보 추출을 목적으로 하는 대표 기술로 발전하여, 정보 검색 및 요약, 추천 시스템, 질의응답 시스템, 지식베이스 구축 등의 기반 기술로 활용되고 있다. 그러나, 일반적인 개체명 인식 모델의 경우, 소셜미디어, 웹 문서 및 뉴스 등과 같이 보편적인 어휘가 다수 포함된 데이터를 바탕으로 개체명 유형을 구축하고 모델을 학습함에 따라 전력 분야의 설비명, 설비용량, 유관기관 등의 특화 용어를 인식하지 못하며, LA, 전주 등의 설비명을 지역명으로 오인식하는 한계점을 지닌다.

이 연구에서는 전력 분야 도메인 특화 정보 추출을 위한 개체명 인식기 개발을 목표로 한다. 이를 위해, 사내 각종 시스템에 수록된 문서 데이터를 수집하고, 전력 분야 개체명 유형 및 학습용 태깅 데이터를 구축하였다. 이후, 전력 일반, 설비, 조직 인사, 방법론, 기관, 설비용량 등 9개의 전력 분야 개체명 유형 인식을 위한 딥러닝 기반의 개체명 인식 모델을 학습 및 검증한다. 이를 통해 일반적인 텍스트 기술 적용 시, 전력 도메인 용어 인식 불가 및 동음이의어 등에 의한 오인식 문제를 해결할 수 있게 되었다. 한편, 개발된 기술은 향후 챗봇 등의 고성능 대화형 AI 서비스 개발을 위한 필수 기반 기술로 제공될 수 있으며, 대규모 비정형 텍스트 데이터의 DB화를 보다 효율적이고 시스템적으로 접근할 수 있는 기반을 마련하는 데 활용될 수 있을 것으로 예상한다.

2. 본 론

2.1 연구 방법

2.1.1 데이터 수집

전력 분야 특화 개체명 인식모델 개발을 위해 한국전력공사 사내외에서 운영되는 각종 사이트에 등록된 문서 및 텍스트 로그 약 15,000건을 수집하였다. 수집된 데이터는 교재, 현장 교육 매뉴얼, 규정, 각종 공법, 기술 표준, 월간 리포트, 배전 고장정보 분석요약 등을 포함한다.

2.1.2 데이터 전처리

수집한 데이터 중 문서 데이터는 출처와 작성 목적에 따라 문서의 구성 형태, 길이, 표 또는 그림 등의 포함 요소가 모두 상이했으나, 엑셀로 추출된 로그 데이터의 경우 평균 102자로 작성된 단일 문장 또는 짧은 단락으로 구성되어 있었다. 문서 데이터 전처리를 위해 다양한 형식의 문서를 pdf 파일로 변경하였으며, 표와 그림을 제외한 모든 텍스트를 추출하였다. 또한, 마침표를 기반으로 문장 단위로 텍스트를 분리하였으며, 마침표가 없이 끝나는 문장일 경우, 마침표로 끝나는 다음 줄의 글과 병합 후 분리하였다. 15글자 이하의 문장은 데이터에서 제외했으며, 결측치, 중복 데이터, 특수문자, 2개 이상의 공백 등을 식별 및 제거하였다. 이를 통해 최종 평균 120자의 글자로 구성된 문장 115,427건을 구축하였다.

2.1.3 전력 분야 개체명 구축

전력 분야의 정보 추출 문제는 명확하게 정의되지 않고 있으며, 개체명 인식 모델 개발에 대응하는 학습 데이터도 미비한 실정이다. 도메인 내 주요 사용 어휘 탐색 및 개체명 유형 구축 선행 필요에 따라, 사내 배전, 송변전, 사무 분야 전문 인력과 함께 전력 분야 문헌 분석 및 사전 실험 결과를 검토하여 도메인 전용 개체명 유형을 다음 표와 같이 정립하였다.

<표 1> 전력 도메인 개체명 유형

	개체명유형	정의	예시
1	전력일반 (TRM)	전력분야 일반용어 및 사내 시스템 관련 등	비상전력, 배전계통, AMI, NDIS, DAS
2	전력설비 (EQP)	배전, 송변전 설비 관련 용어 및 수식어	부싱, 변압기, 인하선, 배선용 인장클램프
3	조직인사 (CVL)	한국전력 사내 직급 및 직무 관련 용어	검침원, 계통운영자, 선로순시원, 처장, 실장
4	방법론	복구, 시공, 진단,	공사, 작업,

	(MTD)	탐지 등 방법론 관련 용어	시공법, 고장점 탐지, 펄스 주입
5	기관 (ORG)	한전 내 부서, 유관기관, 협회, 학회 등 조직 관련 용어	전력시장처, 한국전력거래소, 벽산파워, 1차 사업소
6	지역 (LOC)	배전선로, 계통 등 위치 나타내는 용어 및 전국 시도군 등의 지역명	배전선로, 원성간44R2, 영월SS, 연당DL, 서울시, 경기
7	문서 (DOC)	시공절차서, 관리대장, 보고서 등 사내 문서	해저케이블 긴급복구 절차서, 표준시공 절차서
8	설비용량 (KV)	설비 및 전압 용량 관련 용어	kVA, kV
9	고장 (PRB)	고장 원인, 결과 등 관련 용어	정전, 단락, 순시사고, 훼손, 파손, 결상, 차량충돌, 조류접촉, 까치

2.1.4 개체명 인식 학습용 데이터 구축

BIO 태깅은 각 어휘가 어떠한 종류의 개체명인지를 나타내는 유형 라벨(예: 기관 'ORG', 지역 'LOC' 등)과 함께 개체명의 시작에 'B-(begin)', 중간에는 'I-(inside)', 개체명 외 토큰에는 'O(outside)'를 붙여 여러 개의 어휘를 하나의 개체명으로 묶기 위해 사용되는 가장 보편적인 태깅 시스템이다.

매뉴얼 레이블링 작업을 통해 9개의 전력 분야의 개체명 항목 및 숫자, 시간, 날짜 개체명에 대한 약 10만 개의 BIO 태그 데이터를 구축하였으며, 해당 어휘를 포함하는 데이터는 총 39,193개의 문장으로 구성되었다.

<표 2> 전력 개체명 인식 모델 학습용 데이터 예시

Idx	text	tar	Idx	text	tar
0	데드	EQP_B	0	원성간44R2	LOC_B
1	앤드	EQP_I	1	고압고객	-
2	클램프	EQP_I	2	일광전자재	ORG_B
3	에	-	3	950kW	KV_B
4	조류	PRB_B	4	VCB	EQP_B
5	접촉	PRB_I	5	불량	PRB_B

2.1.5 전력 분야 개체명 인식 모델

최근 문장 내 단어의 의미 분산을 벡터 형태로 표현할 수 있는 단어 임베딩을 기반으로 희소성 문제를 완화하고, 단어를 딥러닝 모형의 입력으로 변환할 수 있는 기법이 마련되었다[1]. Bidirectional LSTM-CRFs[2], KoBERT[3] 등의 딥러닝 모델이 개체명 인식 분야의 좋은 성과를 나타내는 것으로 알려져 있다.

전력 분야 9개의 개체명을 태깅한 약 39,000문장의 BIO 태깅 데이터를 기반으로 두 가지 모델에 대해 학습 및 성능 비교 등의 실험을 진행하여 최종 알고리즘을 선정하였다. 모델 학습을 위해 데이터 셋을 8:2의 비율로 훈련(31,354건), 검증 데이터(7,839건)으로 분리하였다. Bidirectional LSTM-CRFs의 은닉층 및 임베딩 벡터는 128차원으로 설계하였고, 확률적 경사하강법(Stochastic gradient descent)을 통해 최적의 파라미터를 선정하였다(Learning rate=0.01, Gradient clipping=5.0, Dropout=0.3, Epochs=80). 한국어 사전학습 모델인 KoBERT는 훈련 데이터를 통해 미세조정(fine tuning)을 수행하였다(Batch size=32, random_state=2018, Learning rate=0.03, test_size=0.1, Dropout=0.5, Epochs=3).

2.2 연구 결과

2.2.1 전력분야 개체명 인식 성능 평가

전력 분야 개체명 인식 모델 학습을 위해 구축한 데이터를 대

<표3> 전력 분야 개체명 인식 성능 평가

	Precision	Recall	F1-Score
Bidirectional LSTM-CRFs	0.85	0.82	0.83
KoBERT	**0.92**	**0.90**	**0.91**

상으로 KoBERT 및 Bi-directional LSTM CRFs 모델을 학습한 결과, KoBERT 모델이 개체명 유형 인식에서 f1-score 91%로 더 높은 성능을 나타내는 것을 확인하였다.

<그림 1> 전력분야 개체명 인식 결과

해당 모델에 실제 문장을 입력하여 개체명 인식 결과를 검토하였다[그림1]. '폴리머 현수애자', 'COS 1차리드', 'RA' 등과 같은 배전 설비를 하나의 용어로 인식할 수 있게 되었으며, '송림간207R8호', '홍천SS', '성수 DL'등 배전 선로를 나타내는 특수한 용어도 위치 유형으로 인식할 수 있는 것을 확인했다.

3. 결 론

개발한 모델은 f1-score 91%의 성능으로 총 9개 전력 분야 개체명 유형을 인식함에 따라, 기존 텍스트 전처리 라이브러리를 통해 인식할 수 없던 각종 전력 설비, 기관명, 설비 용량, 전략 분야 일반 어휘 등을 추출할 수 있게 되었다. 나아가 LA, 전주 등 전력 도메인에서 설비로 일컬어지는 어휘이지만 일반적으로는 지역명을 지칭하는 동음이의어에 의한 문제 또한 해소되었으며, 애자, 완철, 변대주 등과 같이 일반 개체명 인식 모델에서는 인물 유형으로 주로 분류되었던 전력 설비명도 올바르게 인식된다. 그러나, TRM_I, MTD_I, ORG_I, CVL_I 유형은 상대적으로 낮은 인식률을 나타내고 있으며, '배전 해저 케이블 긴급복구 절차서' 등과 같이 혼용 어절에 대한 인식률 또한 낮아 향후 고도화 과정이 필요할 것으로 보인다.

개발 기술은 향후 챗봇 등의 대화형 AI 서비스 개발의 기반 기술로 제공될 수 있으며, 비정형 포맷의 전력 분야 텍스트 데이터 처리 및 개체 유형에 따른 DB 구축 등에 활용될 수 있다. 또한, 분류 기준에 의한 텍스트 정보는 향후 AI 학습을 위한 학습용 DB 구축에 활용될 수 있으며, 이를 문장 형식 구조에서의 키워드 추출 및 의도 해석 등과 같은 고도화 서비스로의 확장이 가능할 것으로 예상된다.

[참 고 문 헌]

[1] 손대능, 이동주, 이용훈, 정유진, & 강인호., "딥러닝 모형 기반 한국어 개체명 연결", 한국어정보학회 학술대회, 90-95, 2016
[2] Yu, H., & Ko, Y., "Expansion of word representation for named entity recognition based on bidirectional lstm crfs", Journal of KIISE, 44(3), 306-313, 2017
[3] 이녕우, 최호진, "KoBERT 기반 한국어 시나리오 개체명 인식 모델 성능과 한국어 조사의 관계 분석", 한국정보과학회 학술발표논문집, 615-617, 2021