반자동구축된 개체명 주석코퍼스 DecoNAC과 KoBERT를 이용한 개체명인식 플랫폼 DecoNERO

김신우○, 황창회, 윤정우, 이성현, 최수원 & 남지순
한국외국어대학교 DICORA연구센터/언어인지과학과
siinwoo0306@gmail.com, hch8357@naver.com, skyjw1211@gmail.com,
olsunghyun0416@gmail.com, soown0607@gmail.com, jeesun.nam@gmail.com

A Named Entity Recognition Platform Based on Semi-Automatically Built NE-annotated Corpora and KoBERT

Shin-Woo Kim○, Chang-Hoe Hwang, Jeong-Woo Yoon,
Seong-Hyeon Lee, Soo-Won Choi & Jee-Sun Nam
DICORA/Department of LCS, Hanguk University of Foreign Studies

요 약

본 연구에서는 한국어 전자사전 DECO(Dictionnaire Electronique du COréen)와 다단어(Multi-Word Expressions: MWE) 개체명을 부분 패턴으로 기술하는 부분문법그래프(Local-Grammar Graph: LGG) 프레임에 기반하여 반자동으로 개체명주석 코퍼스 DecoNAC을 구축한 후, 이를 개체명 분석에 활용하고 또한 기계학습에 필요한 도메인별 학습 데이터로 활용하는 DecoNERO 개체명인식 플랫폼을 소개하는 데에 목적을 두었다. 최근 들어 좋은 성과를 보이는 것으로 보고되고 있는 기계학습 방법론들은 다양한 도메인을 기반으로한 대규모의 학습데이터를 필요로 한다. 본 연구에서는 정교하게 설계된 개체명 사전과 다단어 개체명 시퀀스에 대한 언어자원을 바탕으로 하는 반자동으로 학습데이터를 생성하는 방법론을 제안하였다. 본 연구에서 제안된 개체명주석 코퍼스 DecoNAC 기반 접근법의 성능을 실험하기 위해 온라인 뉴스 기사 텍스트를 바탕으로 실험을 진행하였다. 이 실험에서 DecoNAC을 적용한 경우, KoBERT 모델만으로 개체명을 인식한 결과에 비해 약 7.49%의 성능향상을 기대할 수 있음을 확인하였다.

주제어: 개체명 주석코퍼스 DecoNAC, DECO 전자사전, LGG 프레임, 개체명 인식 플랫폼 DecoNERO

1. 서론

본 연구는 한국어 전자사전 DECO(Dictionnaire Electronique du COréen) 시스템[1]과 두 단어 이상으로 구성된 다단어(Multi-Word Expressions: MWE) 개체명을 부분패턴으로 기술하는 부분문법그래프(Local-Grammar Graph: LGG) 프레임[2][3]에 기반하여 반자동으로 개체명 주석코퍼스(Deco Named entity Annotated Corpus: DecoNAC)를 구축한 후, 이를 개체명 분석에 활용하고 또한 기계학습에 필요한 도메인별 학습 데이터로 활용하는 개체명인식 플랫폼(Deco Named Entity Recognizer & Organizer: DecoNERO)을 소개하는 데에 목적을 두었다.

최근 들어 기계학습 기반 접근법이 좋은 성과를 보이는 것으로 보고되고 있으나[4][5][6], 학습 데이터 도메인의 의미속성에 직접적인 영향을 받으며 대규모의 데이터가 필요로 한다는 한계가 문제가 된다. 이러한 한계를 극복하기 위해서 본 연구에서는 정교하게 설계된 개체명 사전과 MWE 개체명 시퀀스에 대한 언어자원을 바탕으로 반자동으로 학습데이터를 생성하는 방법론을 제안하였다. DecoNAC은 DecoSAC 감성주석코퍼스와 함께 다양한 도메인에 대한 대규모 학습데이터를 반자동으로 구축하는 한국외대 DICORA 연구센터의 DECO-LGG 프로젝트의 일환으로 진행되었다.

일반적으로 개체명(named entity)이란, 실세계에 존재하는 특정 대상들에 대한 명칭으로 일반적으로 고유명사가 이에 해당된다. 그러나 고유명사뿐만 아니라 일반명사 또한 개체명 부류에 포함될 수 있으며, 실제 응용분야의 궁극적인 목적에 따라 개체명의 유형도 다양하게 분류될 수 있다.

개체명 어휘는 정보 추출에 필요한 핵심어로 사용되며, 기하급수적으로 정보량이 증가하는 현대 사회의 특성에 따라 더욱 다양한 유형으로 실현될 수 있다. 예를 들어, '대한민국'이라는 고유명사가 의미적으로 확장됨에 따라 '갓한민국', '헬조선' 등의 변이형으로 나타날 수 있고, 고유명사 '메가박스'는 영화관 체인 조직을 나타내는 추상적인 의미와 동시에 구체적 공간 장소를 지칭하는 의미로 사용될 수 있다. 이렇게 확장되거나 중의적 의미를 갖는 개체명 어휘들에 대한 실제 분류를 위해서는 개체명에 대한 상세 분류 기준을 제시하는 연구만으로는 충분하지 않다. 이러한 분류 기준에 따라 실제 어휘들이 구체적으로 분류되는 작업이 반드시 수반되어야 하기 때문이다. DECO 전자사전은 이러한 개체명 부류에 대한 상세 분류 체계와 함께, 각 카테고리에 해당하는 어휘 표제어들을 분류하여 이를 텍스트 처리 시에 사용할 수 있는 형식으로 내장하고 있다는 점에서 그 차별성을 가진다. 이를 통해 실제 코퍼스에서 실현되는 개체명 어절들의 모든 표면형을 인식하고 이들의 의미분류 유형을 자동으로 마크업하는 작업을 가능하게 한다.

그런데 실제 텍스트에서 개체명은 단일어뿐만 아니라 MWE 형태로도 실현된다. 이렇게 나타나는 개체명을 인식해 내기 위해 본 연구에서 제안하는 LGG 프레임은

DECO 사전에 등재된 표제어의 개체명과 의미분류 태그를 바탕으로 부분적인 패턴문법을 비순환 방향성 그래프(directed acyclic graph) 형식으로 표상한다. 이를 통해 MWE로 실현되는 개체명 인식뿐만 아니라, 도메인별로 상이한 분류로 기술되어야 하는 단일어 개체명 표현들도 유연하고 효율적으로 기술하는 것이 가능하게 된다.

DECO 사전과 LGG는 부트스트랩 방식으로 지속적으로 보완되고 확장된다. DECO사전은 활용후치사 트랜스듀서 사전을 통해 모든 표면형 어절을 인식할 수 있도록 컴파일되어 있어 UNTIEX 플랫폼[7]에서 곧바로 코퍼스 처리에 적용될 수 있다. 이를 통해 본 연구에서는 반자동으로 개체명을 주석한 DecoNAC을 생성할 수 있다. 앞서 지적한 바와 같이 이렇게 생성된 주석코퍼스는 두 가지 방향에서 의의가 있는데, 첫째는 체계적으로 구축된 언어자원을 기반으로 생성된 데이터로 개체명 분석을 수행하는 것이다. 둘째는 DECO 사전과 LGG을 기반으로 특정 도메인에 얽매이지 않고 대용량의 DecoNAC을 구성하는 것이 가능하므로, 기계학습에 필요한 학습데이터를 반자동으로 생성하여 공급하는 것이 가능하다는 점이다.

본 연구에서는 DECO 사전과 LGG를 활용한 반자동 개체명 주석코퍼스 DecoNAC과, 이를 기반으로 개체명 분석을 수행하고 동시에 이를 기계학습을 위한 학습데이터로 사용하는 DecoNERO 플랫폼을 소개한다. 끝으로 온라인 기사에 대한 개체명 인식 실험을 통해 DecoNAC 기반 접근법의 성능을 실험하였다.

2. 관련 연구

최근 다양한 유형의 기계학습 기반의 개체명 인식 방법론이 제안된 가운데, 이를 위해서는 우선 개체명이 태깅된 코퍼스와 개체명 태그가 실제 부여된 개체명 리스트 및 관련 분류체계 등이 필수적으로 요구된다. 개체명을 태깅된 학습용 주석코퍼스을 구축하기 위해서는 우선 개체명에 대한 분류 체계가 설정되어야 한다.

TTA 개체명 분류체계[8]를 보면 15가지의 대분류와 146가지의 세부 분류가 제시되어 있다. 표 1은 15가지 대분류 체계를 보인다.

표 1. TTA 분류체계

대분류	태그	의미	하위분류
Person	PS	인명/별칭	1
Study Field	FD	학문 분야	6
Theory	TR	이론, 법칙, 원리	6
Artifacts	AF	인공물	13
Organization	OG	기관/단체 명칭	15
Location	LC	지역/장소, 지형/지리명칭	14
Civilization	CV	문명/문화	19
Date	DT	날짜	8
Time	TI	시간	5
Quantity	QT	수량	18
Event	EV	특정 사전/사고 명칭	5
Animal	AM	동물	9
Plant	PT	식물	7
Material	MT	물질	4
Term	TM	기타 개체명 용어	16

이 분류체계는 현재 한국어 개체명 인식에 많이 사용되고 있는 체계로서, PS 태그를 제외하고는 하위 태그가 최소 4개에서 최대 19개까지 존재한다. 세분류가 많은 이유는 새로운 범주가 필요할 때 대분류를 추가하기보다 하위분류에 새로운 범주를 생성하는 방식으로 추가하는 것이 효과적이기 때문으로 보인다.

하지만 TTA 개체명 태그셋이 실제 어휘에 대한 분류체계로 적용이 될 때는 주로 대분류인 15가지를 기반으로 약간의 추가 혹은 축소 방식으로 적용되는 경우가 많다. 실제로 총 분류체계가 146가지로 세세히 분류되어 제안되어 있으나, 해당 분류체계에 맞는 표제어 리스트나 사전 같은 데이터는 공개된 바가 없어 그 적용에 있어 어려움이 있다.

영어 개체명분류 체계는 MUC(Message Understanding Conferences)에서 사용된 주석 체계가 가장 잘 알려져 있다. MUC-6에서는 'SGML text markup data'로 평가가 이루어졌으며, 해당 마크업 태그 정보는 이름(고유명사), 지명, 기관명으로 구성된 ENAMEX 태그와 날짜와 시간으로 구성된 TIMEX 태그, 돈과 수치(퍼센트)로 구성된 NUMEX 태그로 구성된다[9].

이와 같이 연구소나 학회에서 제시한 개체명 기준 이외에도 해외에서는 기업에서 활용하는 개체명 인식 시스템에 대한 태그셋도 존재한다. 그림 1의 워싱턴대학교 외에서 만든 AllenNLP의 NER Tagger는 8가지 태그로 분류되며, 해당 태그는 장소(Loc), 인명(Per), 조직(Org), 기타 범주(Misc), 정치(GPE), 날짜(Date), 시간(Time) 그리고 법(Law)으로 구성된다[10].

이외에도 6가지 분류의 태그로 개체명을 인식해주는 Dandelion의 NER Tagger[11], 14개의 범주와 23개의 세범주가 존재하는 Microsoft 사에서 제공하는 개체명 인식 시스템[12], 8개의 개체명 태그가 사용된 IBM 사의 개체명 인식 시스템[13] 등이 존재한다.

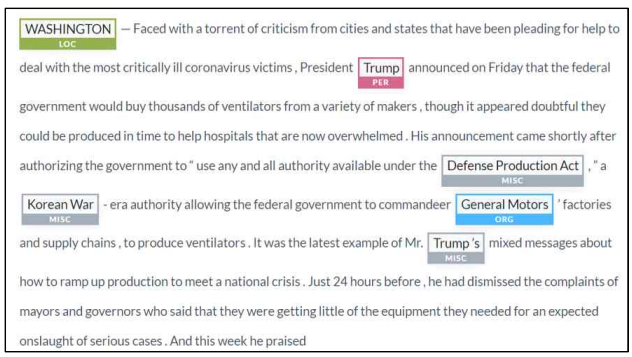

그림 1. AllenNLP의 NER Tagger

본 연구에서는 DECO사전에서 사용하는 EntLEX 개체명 분류체계에 기반하여 11가지의 대분류와 30가지의 세분류를 적용하였고, 그리고 DECO-SemOnto 의미분류 체계에 기반한 98가지의 의미분류 표제어 정보를 적용하였다. 여기서 단일어 유형의 개체명은 이처럼 사전에 직접 수록되는 반면, MWE 유형의 개체명은 LGG 프레임을 이용하여 패턴문법 유형으로 구조화되었다.

다음 3장에서는 이처럼 구축된 DECO 전자사전과 LGG 패턴문법을 이용해 반자동으로 DecoNAC을 구축하는 과정에 대해 소개한다. 본 연구의 4장에서 소개하는 개체명 인식 플랫폼 DecoNERO는 DecoNAC과 KoBERT 기계학습 접근법을 결합한 방식으로 적용하여 개체명을 인식하는 플랫폼으로, 여기서는 실제 언어자원을 구축하는 연구자가 그 적용결과를 시각적으로 확인하여 부트스트랩 방

식으로 DecoNAC 생성을 위한 사전과 문법을 보완 확장할 수 있도록 하였다는 점에서 차별성을 갖는다.

3. 개체명 주석코퍼스 DecoNAC 구축과정

DecoNAC은 한국어 표제어에 대한 형태, 통사, 의미, 극성 정보 및 개체명 정보를 제공하는 기계가독형 전자사전 DECO를 기반으로 생성된다. 사전에 기반한 개체명 패턴 인식을 위해, DECO 사전의 개체명 태그를 바탕으로 도메인별 코퍼스에서 나타나는 단일어 혹은 MWE 개체명 패턴을 LGG로 구축한다. DECO 사전과 LGG 언어자원을 기반으로, 이와 연동되는 Unitex 플랫폼을 통해 개체명 태그가 반자동으로 주석된 DecoNAC을 생성한다.

위와 같은 방법으로 생성된 DecoNAC은 한국어 개체명 인식 플랫폼 DecoNERO를 통해 개체명 분석에 사용된다. 분석 결과는 DecoNERO 플랫폼에서 시각화된 형태로 확인된다. 또한, 반자동으로 대량 생산된 새로운 DecoNAC을 플랫폼 안에 내장되어있는 KoBERT 모델의 추가학습 데이터로도 활용할 수 있다. 이와 같은 진행 과정을 그림으로 나타내면 그림 2과 같다.

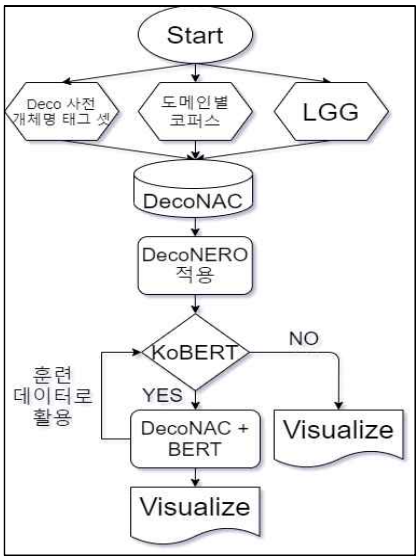

그림 2. DecoNAC 구축과정과 DecoNERO 처리 흐름도

3.1 개체명 주석코퍼스 DecoNAC 반자동구축

개체명 주석코퍼스를 구축하기 위해 DecoSOC 크롤러[14]를 사용하여 온라인상의 원하는 도메인의 코퍼스를 수집한다. 이렇게 수집된 코퍼스는 Unitex 플랫폼에서 DECO 전자사전과 LGG 패턴문법을 기반으로 자동 처리되어 다음과 같은 XML 형식의 주석코퍼스로 생성된다. 그림 3은 단일어 유형에 대한 개체명 분류 정보가 내장된 DECO 사전과 MWE 유형의 개체명 인식을 기술하고 있는 LGG 패턴문법을 기반으로 Unitex 플랫폼에서 자동 구축된 DecoNAC의 일부 예를 보인다.

```
<XXEV>아시안 게임 폐막식</XXEV> 때 당시.
<XXPE>이지은</XXPE> <XXHU>기자</XXHU>.
<XXPE>조명균</XXPE> <XXHU>통일부 장관</XXHU>의 <XXOR>여자
아이스하키 남북 단일팀</XXOR>을 꾸리는 방안.
<XXLO>대구 경북</XXLO>만 바로서면 <XXOR>대한민국</XXOR>이 변한다.
<XXOR>한겨레</XXOR>는 <XXLO>북한</XXLO> 언론인가?
```

그림 3. DecoNAC의 일부

3.2 DECO사전과 개체명관련 태그셋 DecoTagset

DecoNAC이 구축되기 위해 필요한 언어자원 중 하나는 단일어 개체명 인식에 필요한 DECO 전자사전이다. 표 2는 DECO 사전에서 사용하는 EntLEX 개체명분류 체계의 대분류 유형을 보인다. <인물성/공간성/시간성/사물성>으로 분류되고 다시 하위분류되어 전체 11가지의 분류 카테고리를 나타낸다. 이 분류에서는 숫자를 중심으로 구성되는 시퀀스(예: 3천 달러) 유형은 일반적으로 MWE 구성을 보이므로, DECO 사전의 단일어 표제어 대신 LGG에서 <XXNU>의 태그를 통해 주석되는 형식을 취한다.

표 2. Deco사전 개체명 태그셋

개체분류	중분류	태그
인물성	특정개인	XXPE
	일반개인	XXHU
	집단조직	XXOR
공간성	자연공간	XXGE
	인공공간	XXLO
시간성	사건표현	XXEV
	시간표현	XXTI
사물성	특정구체	XXPR
	이동구체	XXTH
	고정구체	XXCO
	개념추상	XXCR

현재 표 2에 제시된 단일어 개체명 분류 태그의 각 예시를 들어보면 다음과 같다.

(1)ㄱ. **트럼프(XXPE)** 대통령의 이번 발언.
ㄴ. 새로 부임한 **국회의장(XXHU).**
ㄷ. **한국노총(XXOR)**은 문 대통령에게 꽃다발과 자체 제작한 벽시계를 전달했다.
ㄹ. **도봉산(XXGE)**을 찾는 방문객들이 늘어났다.
ㅁ. **성남시(XXLO)**, 공공시설 장애물 조사
ㅂ. 2012년 **대통령선거(XXEV)**에 개입했다.
ㅅ. **오늘(XXTI)**(27일) 경남 밀양 세종병원 화재 참사 현장을 방문했다.
ㅇ. 새로 나온 **갤럭시폰(XXPR).**
ㅈ. **린스(XXTH)**가 다 떨어졌어요.
ㅊ. **남대문(XXCO)**이 재건되었다.
ㅋ. **라이언킹(XXCR)**이 방영되었다.

이 분류에 대해서 다시 30가지 하위분류가 수행된다. 예를 들어 '특정개인(XXPE)'에는 한국인(XQKN)인지 외국인(XQFR)인지 등이 다시 분류되고 '집단조직(XXOR)'에는 '은행(XQBN)'인지 '학교(XQSC)'인지 등이 하위분류된다. '개념추상(XXCR)'의 경우에도 '캐릭터(XQCH)'인지 '게임(XQGM)'인지, '영화명(XQMV)'인지 등이 다시 하위분류된다.

개체명 세분류와 함께 DECO 사전의 SemOnto 의미분류 체계가 함께 사용되었다. DECO 사전에 수록된 표제어들은 모두 98가지의 의미분류 태그로 분류되는데, 이들은 크게 서술형과 비서술형으로 나눌 수 있다. 표 3은 비서술형 의미분류의 예의 일부를 보인다.

표 3. 비서술형 의미분류 태그셋 일부

태그	의미	태그	의미
QHUM	인물	QLAN	문자
QANM	동물	QHIS	기념일
QPLT	식물	QCRR	화폐
QSUS	광물	QVEH	교통
QCLL	집단	QINS	제도
QNTR	자연	QART	예술
QREG	지리	QDIS	발견
QCLT	의생활	QCRA	발명
QFOO	식생활	QPRD	상품
QBUI	주생활	QPES	인간속성

비서술형 의미분류는 다시 구체물 관련 분류와 추상적 개념, 속성 관련 분류로 하위분류된다. 이러한 의미분류 체계를 활용하면 현재 DECO 사전에 분류되어있는 개체명 대분류의 상세 유형을 분류해 인식하는 작업이 가능해진다. 그 결과 전체 139개의 분류 태그를 바탕으로 개체명 인식을 위한 분류정보로 활용할 수 있게 된다.

3.3 DECO 전자사전 표제어의 개체명 분류정보

DECO사전의 개체명 분류의 특징은, 이러한 분류체계가 개념적 하향식으로 '체계' 자체만 설계되어있는 것이 아니라, 어휘적 상향식(bottom-up)으로 실제 표제어 '어휘' 각각에 대한 분류 작업이 전체 수행되어 있다는 점이다. 현재 375,380개의 표제어를 내장하고 있는 DECO 명사 사전(ver5.2)의 개체명 분류 대상 어휘는 208,130여 개에 이른다.

DECO 사전에는 개체명 태그뿐 아니라 개체명 인식에 직접적인 연관성을 갖는 의미 태그와 개체명 세분류, 그리고 감성분석을 위한 극성(polarity) 태그, 활용·품사·형태·통사 분류 등 여러 유형의 태그가 부여되어 있다. 사전의 초기 버전은 CSV 형식의 테이블로 저장된 후, Unitex 플랫폼에서 코퍼스 처리를 위한 사전 형식으로 컴파일되기 이전에, 일련의 사전 형식의 변환이 수행되어야 한다. DecoLEXO[15]는 이를 위해 별도 구현된 플랫폼으로 Unitex와 호환되는 형식으로 구분자를 자동 변환하고 활용후치사 트랜스듀서와 결합 가능한 어간변이형 사전을 자동 생성하는 기능을 수행한다.

3.4 LGG를 통한 다단어 개체명 패턴의 기술

3.2에서 기술한대로 DECO 사전의 태그가 코퍼스에 부착되면, LGG 그래프를 통해 개체명 인식에 필요한 태그들을 기저로 개체명 주석이 가능하다. 이때, 코퍼스에서 실제 개체명은 단일어뿐 아니라 MWE 형태로도 실현된다. 예를 들어 다음을 보자.

(2)ㄱ. 여야 단체
ㄴ. 중앙 정부
ㄷ. 해군 작전사령부

위의 예는 실제 코퍼스에서 관찰되는 MWE 개체명들이다. (2ㄱ)의 '단체'와 (2ㄴ)의 '중앙' 같은 경우, 각각의 개별 단일어는 사전에서 개체명으로 분류되기 어려운 일반명사 부류이다. 그러나 코퍼스에서 이들은 공기한 성분과 함께 하나의 개체명으로 인식되는 것이 바람직하다. (2ㄷ)의 경우도 각 어휘 성분이 하나의 개체명을 이루는 MWE 개체명으로 인식되는 것이 필요하다. 이러한 일련의 부분적 패턴들은 LGG(Local-Grammar Graph) 프레임을 통해 효율적으로 기술하는 것이 가능하다. LGG 프레임은 방향성 그래프 형식으로 구성되어, 문장보다 작은 단위의 언어현상을 부분적으로 표상하는 것을 가능하게 한다. 그림 3은 정치 기사 코퍼스에 MWE로 실현된 개체명을 인식하고 주석하기 위한 LGG의 예를 보인다. 이때, 산발적으로 과생성을 유발할 수 있는 몇몇 개체명 세분류 태그는 제외하고 그래프를 구축하였다.

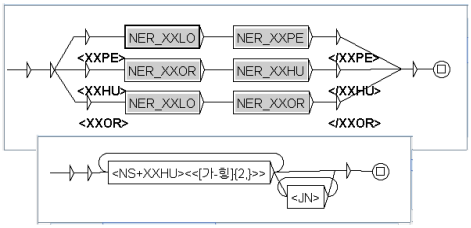

그림 4. MWE 개체명 인식을 위한 그래프

위의 그래프에 기술된 회색 박스들은 각 카테고리별 개체명 어휘를 표상하는 서브그래프들을 호출하는 그래프들이다. 이 그래프는 Unitex 플랫폼에서 유한 상태 트랜스듀서(finite-state transducer)로 컴파일되어 MWE 개체명을 인식하고 주석하는 데에 사용된다.

4. DecoNERO 플랫폼

본 연구에서 소개하는 DecoNERO 플랫폼은 앞서 논의된 DECO 사전과 LGG 패턴문법을 통해 반자동으로 생성된 DecoNAC을 토대로 개체명 인식의 결과를 시각화해주는 동시에 이를 플랫폼에 내장된 KoBERT 기반 기계학습기(trainer)의 학습용 데이터로써 활용할 수 있도록 구현된 프로그램이다. DecoNERO의 세부 기능은 크게 5가지로 분류되며, 이에 대한 상세는 다음과 같다.

4.1 KoBERT를 이용한 하이브리드 방식의 개체명 인식

단일어 개체명을 인식하기 위한 DECO 사전의 태그 정보와 MWE 개체명을 포착하기 위한 LGG를 바탕으로 생성된 DecoNAC에서 처리되지 못한 새로운 유형의 개체명들이 실현될 수 있다. KoBERT와 같은 기계학습 기반 방법론은 이러한 재현율(recall)의 문제를 효과적으로 보완할 수 있다. DecoNERO는 DecoNAC에서 인식한 개체명을 배제하고 남은 문서에 대해서 기계학습 기반의 개체명 인식을 수행하는 하이브리드 방식을 사용함으로써 재현율을 보완한다. 본 연구에서는 KoBERT 모델에 토큰 단위 개체명 분류를 수행할 지도적 모델을 전이 학습(transfer learning)을 통해 구성하고, 문장 내 토큰에 대한 선형분류를 통해 개체명을 처리하는 모듈을 오픈 소스로 공개된 [16]의 알고리즘에 기반하여 구현하였다. 이때, 학습에 사용된 개체명 태그를 본 연구의 DECO 개체명 태그로 대응시키는 과정이 진행되었다. 이 과정에서 KoBERT를 통해 주석된 개체명 태그는 'XX-B' 방식으로 차별화되어, DecoNAC을 통해 자동 주석된 개체명과 구분될 수 있게 하였다.

4.2 DecoNAC 코퍼스를 이용한 KoBERT의 추가학습

DecoNERO에서 사용한 KoBERT 모델의 학습데이터는 'Naver NER Challenge 2018'에서 제공한 데이터셋이다 [17]. 본 연구에서는 이를 그림 5와 같이 DECO 사전의 개체명 태그로 변환하여 모델을 학습시켰다.

```
롯데, 연첩 행렬 멈춰 XXOR-B O O O
이곳 플랑큰달동물원에서는 지방에서  O XXLO-B O
오초아 전성시대이다 . XXPE-B O O
```

그림 5. KoBERT 모델 학습데이터 일부

그림 5는 실제로 KoBERT 모델을 학습시킬 때 사용되었던 데이터의 일부를 보인다. 학습된 모델은 학습데이터의 도메인에 의존적인 한계를 보이기 때문에, 새로운 도메인에서의 성능 향상을 위해서는 그 도메인에 대한 학습 코퍼스를 제공하는 것이 중요하다. 본 연구에서 제안하는 DecoNAC의 반자동 구축 과정과 이를 기계학습에 사용할 수 있도록 하는 DecoNERO 플랫폼은 이런 점에서 중요한 의의를 갖는다.

4.3 개체명 인식 결과 시각화

개체명 인식 결과는 그림 6과 같은 방식으로 시각화되어 제시된다. 사용자는 시각화하고자 하는 태그를 테이블 형식으로 선택하고 편집할 수 있어 특정 유형의 개체명 범주만을 시각화하는 것이 가능하다. 이러한 기능은 부트스트랩 방식으로 학습데이터를 확장할 때 사용자가 효율적이고 직관적으로 데이터를 관리할 수 있다는 장점이 있다.

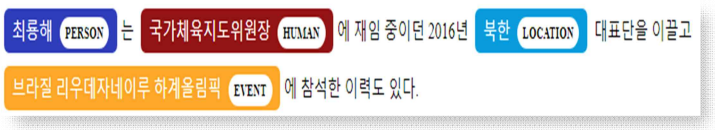

그림 6. DecoNERO 시각화 화면

4.4 개체명 분석 결과에 대한 통계 정보

DecoNAC과 KoBERT를 기반으로 개체명이 인식되면, 그 결과에 대한 통계 결과도 획득할 수 있다. 그림 7은 현재 DecoNERO에서 제공하는, 시각화된 개체명 통계 결과의 예를 보인다. 각 개체명 유형별로 실제 출현한 개수와 그 비중이 바(bar) 형태로 실현된 것을 볼 수 있다. 각 바 상단에 적혀있는 숫자는 개체명 태그의 빈도순으로 내림차순 나열되어 사용자가 빈도수에 따른 비교를 수월하게 진행할 수 있도록 구성하였다.

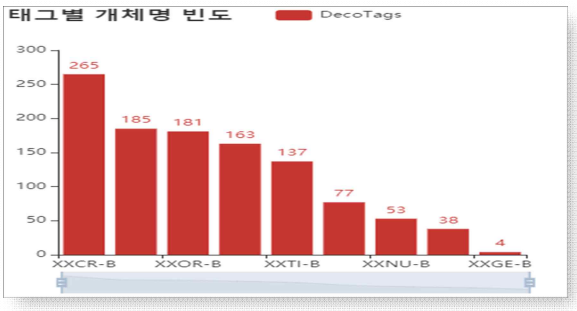

그림 7. 태그 종류별 개체명 빈도 그래프

그림 8의 좌측 하단의 파이(pie) 그래프는 코퍼스 내에서 주석된 개체명 태그 분류의 비율을 보여주는 그래프이다. 그래프 하단의 개체명 태그 라벨을 선택해서 특정 태그의 비중을 한눈에 쉽게 알아볼 수 있으며 마우스 커서를 그래프 주변에 위치하면 해당 태그의 빈도수와 정보가 출현하는 유동적 그래프 형식을 취하고 있다. 반면 그림 8의 우측 상단의 물결 그래프는 사용자가 호출한 코퍼스의 전체 어절당 개체명의 비율을 나타낸다. 해당 그래프를 통해 사용자는 개체명이 주석된 어휘의 비중을 한 눈에 확인할 수 있다.

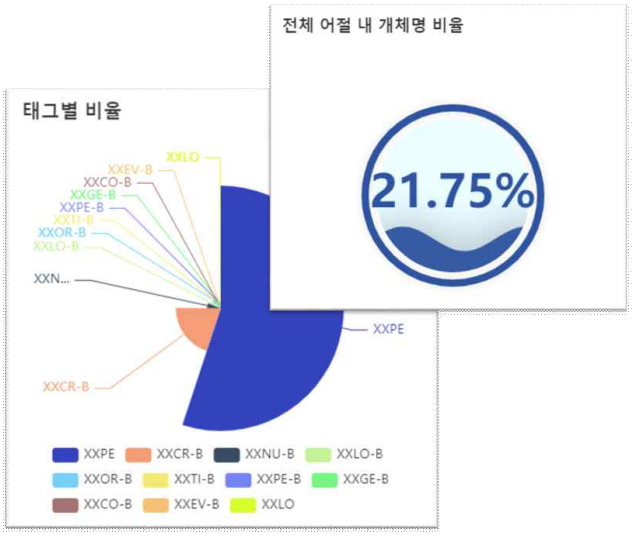

그림 8. 개체명 분포 파이 그래프와 물결 그래프

이상과 같이 제공되는 3가지 통계 그래프를 바탕으로 개체명 인식 결과에서 획득되는 중요한 통계적 자료를 사용자가 직접 확인할 수 있다.

4.5 개체명 표제어 분석 테이블

DecoNERO에서는 인식된 개체명 어휘의 목록을 표 4와 같이 각 카테고리 태그와 함께 제공한다. 1열은 개체명이 주석된 어휘들을 보여주며, 2열은 해당 어휘에 부착된 태그를 보여준다. 3열은 각 어휘가 코퍼스에 나타난 빈도수를 나타낸다.

표 4. 개체명 어휘표

표제어	태그	빈도
⋮		
올림픽	XXEV-B	7
판문점	XXLO-B	7
한반도	XXLO-B	7
홍준표	XXPE-B	7
대통령	XXCR-B	6
북한	XXLO-B	6
오후	XXTI-B	6
트럼프	XXPE-B	6
⋮		

이와 같은 자료는 그 텍스트에 실현된 개체명의 유형을 한눈에 확인할 수 있도록 한다. 특히 KoBERT를 통해 획득된 개체명(즉, XX-B 태그류)의 경우, DECO사전에 누락된 형태들로 추정되므로, 추후 이들을 추가하고 확장하

는 데에 중요한 언어자원으로 활용할 수 있다.

5. 실험 및 성능 평가

5.1 실험 데이터

본 연구에서 소개한 DECO-LGG 언어자원에 기반한 하이브리드 방식의 접근법의 성능을 평가하기 위하여 이를 KoBERT 기반 접근법과 비교하는 실험을 진행하였다. 여기 사용된 코퍼스는 KoBERT 모델의 평가데이터로 선정하였다. 해당 코퍼스는 KoBERT 모델의 학습데이터와 동일한 뉴스 기사 도메인으로, 총 100,109어절 규모로 구성되어있다.

5.2 실험 과정

실험 코퍼스를 DECO 사전과 LGG를 기반으로 개체명 태그를 주석해 DecoNAC으로 변환한다. 주석된 코퍼스에 KoBERT 모델을 이용해 하이브리드 방식으로 결과 파일을 생성한다. 해당 결과물을 정답 데이터셋과 비교하여 성능을 측정하였다. 한편 주석되지 않은 코퍼스를 바탕으로 KoBERT 모델을 사용하여 개체명 인식을 진행한 데이터와 이를 비교하여 다음과 같은 결과를 획득하였다.

5.3 실험 결과

표 5는 DecoNAC 기반 접근법의 성능을 KoBERT 모델만을 사용했을 때의 성능과 이 두 가지를 하이브리드하게 적용하여 추출했을 때의 성능과 비교한 결과를 보인다. DecoNAC을 기반으로 하는 경우, 89.65%의 높은 정밀도를 보이지만 상대적으로 낮은 재현율(67.88%)을 나타내었다. 반면 KoBERT 모델만 사용한 경우 DecoNAC보다 정밀도는 1.66% 낮게 나왔지만, 재현율은 77.9%로 10.02% 더 높게 나타났다.

DecoNAC에 KoBERT 모델을 적용한 하이브리드 방법론은 정확율이 KoBERT 모델이나 DecoNAC에 기반한 경우에 비해 86.98%로 다소 감소하였으나, 재현율은 93.52%로 높게 상승한 것을 볼 수 있었다.

전체적으로 DecoNAC 접근법이 77.26%의 조화평균을 보인 반면, KoBERT 방식은 82.64%를 나타냈고, 하이브리드 방식은 90.13%로 현재 3가지 접근법 중 가장 좋은 성능을 보이는 것을 확인할 수 있었다.

표 5. 성능 실험 결과

	Precision	Recall	F1 score
DecoNAC	89.65 %	67.88 %	77.26 %
KoBERT	87.99 %	77.9 %	82.64 %
DecoNAC+KoBERT	86.98 %	93.52 %	90.13 %

6. 결론

본 연구에서는 현재 제공된 KoBERT의 학습데이터 도메인에 한정하여 실험을 진행하였다. 그러나 해당 도메인의 학습데이터가 제공되지 않는 KoBERT 접근법의 성능은 더 많은 한계를 보일 것으로 예상된다. 이러한 한계를 극복하기 위해서는 본 실험에서 제안한 DecoNAC 기반 접근법을 더 다양한 도메인에 적용하여 이를 통해 개체명 인식 성능을 향상시키는 것이 필요하다.

또한, 본 연구에서 제안하는 반자동 개체명 주석코퍼스 생성 방법론은 체계적인 언어 자원을 바탕으로 생성되어 향후 다양한 기계학습 알고리즘의 성능을 향상하는 데에 필요한 학습데이터를 제공하는 데에 중요한 의의를 가질 것으로 기대된다.

참고문헌

[1] 남지순. 코퍼스 분석을 위한 한국어 전자사전 구축방법론. 역락출판사 (2018)

[2] 남지순. 프랑스어 언어 자원 구축을 위한 부분문법 (Grammaire locale) 방법론의 소개. 한국프랑스논문집, 49, 67-94 (2005)

[3] Gross, M.. The Construction of Local Grammars. Finite-State Language Processing, The MIT Press (1997)

[4] 박광현, 나승훈, 신종훈, 김영길.KoBERT를 이용한 한국어 자연어처리: 개체명 인식, 감성분석, 의존 파싱, 의미역 결정. 한국정보과학회 학술발표논문집, 584-586 (2019)

[5] 양성민, 정옥란. DeNERT: DQN과 KoBERT를 활용한 개체명 인식 모델. 한국컴퓨터정보학회 논문지 25, 29-35 (2020)

[6] 유소엽, 정옥란. KoBERT와 지식 그래프를 이용한 한국어 문맥 정보 추출 시스템. 인터넷정보학회논문지, 123-131 (2020)

[7] Paumier, S.. Unitex Users' Manual. France (2003)

[8] TTA, 개체명 태그 세트 및 태깅 말뭉치. TTAK.KO-10.0852 (2015)

[9] MUC-6, https://cs.nyu.edu/grishman/muc6.html (1996)

[10] The Allen Institute for Artificial Intelligence. Named Entity Recognition. https://demo.allennlp.org/named-entity-recognition (2018)

[11] SpazioDati. Dandelion, Entity Extraction. https://dandelion.eu/semantic-text/entity-extraction-demo/ (2017)

[12] Microsoft. Supported entity categories in the Text Analytics API v3. https://docs.microsoft.com/en-us/azure/cognitive-services/text-analytics/named-entity-types?tabs=general (2020)

[13] IBM. MAX-Named-Entity-Tagger: Locate and tag named entities in text. https://github.com/IBM/MAX-Named-Entity-Tagger (2019)

[14] 황창회 & 남지순. DecoSOC. DICORA-TR-2019-02, HUFS (2019)

[15] 김신우, 이성현, 황창회 & 남지순. DecoLEXO, DICORA-TR-2020-01, HUFS (2020)

[16] Park, J.. KoBERT-NER. GitHub. https://github.com/monologg/KoBERT-NER.git (2020)

[17] 네이버, 창원대학교, Data.ly. Naver NLP Challenge 2018. GitHub. https://github.com/naver/nlp-challenge/blob/master/missions/ner/data/train/train_data (2018)