東洋學 第84輯(2021年 7月) 檀國大學校 東洋學硏究院

인공지능 시대의 사전과 데이터

-한국역사인물사전(데이터)을 위한 탐색-***

김 바 로***

국문초록

본 논문은 인공지능 시대 사전의 미래는 인간을 위해서 컴퓨터가 이해할 수 있는 방식으로 여러 가지 사항을 모아 일정한 순서로 배열하고 그 각각에 해설을 붙인 내용인 '기계가독형 데이터'에 있다고 판단했다. 사전학의 데이터학으로의 전환 가능성을 타진하기 위해서, 인공지능 통합 플랫폼인 AIHUB에 공개된 레이블드 데이터, 병렬코퍼스(말뭉치) 데이터, 스쿼드 데이터의 구조와 내용의 장단점을 살펴보았다. 데이터 설계와 전문 도메인 영역의 데이터 구축에서는 충분히 기존 사전학의 연구 성과를 계승할 수 있다고 판단했다. 그리고 딥러닝을 위한 데이터의 한계를 타파하기 위한 또 다른 인공지능 데이터인 시맨틱 데이터의 구조와 내용에 대해서 탐색했고, 기존 사전학의 연구 성과가 온톨로지 설계와 시맨틱 데이터 구축에서 충분히 활용될 수 있다고 판단했다.

이를 토대로 실제적인 사례 연구를 위해서, 한국역사인물사전을 주제로 딥러닝을 위한 데이터와 온톨로지 설계를 통한 시맨틱 데이터를 각각 실험적으로 구축했다. 우선 딥러닝의 객체명식별(NER) 방법을 활용하여, 『조선왕조실록』 데이터를 토대로 하는 객체명식별(NER) 학습 데이터를 구축하고 실험했다. 딥러닝을 통한 객체명식별(NER)의 객체 식별이 객체의 속성만을 식별할 뿐, 객체의 고유성을 인지하지 못한다는 문제를 확인했다. 이에 객체의 시간, 공간, 인물 등과의 관계를 통한 다양한 고유성을 내재하고 있는 지식으로서의 시맨틱 데이터를 구축하기 위한 실험적인 온톨로지 설계를 통해서 시맨틱 데이터의 가능성을 살펴보았다.

결론적으로 전통적인 사전학에서 기계가독형 데이터에 대한 개념과 이해가 더해질 수 있다면 딥러닝을 위한 데이터와 온톨로지를 통한 시맨틱 데이터 모두에서 충분히 경쟁력이 있다고 본다.

[주제어] 디지털인문학, 사전, 데이터, 인공지능, 역사인물

목 차

Ⅰ. 서 론
Ⅱ. 인공지능 시대의 데이터
Ⅲ. 인공지능을 위한 한국역사인물 데이터
Ⅳ. 결 론

* 이 논문은 2020년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임. (NRF-2020S1A5B5A16082267)
** 이 논문은 2020년 단국대학교 동양학연구원 제11회 사전학 학술대회에서 "인공지능을 위한 한국역사인물사전(데이터)의 개념과 설계"라는 제목으로 발표한 원고를 수정·보완한 것임.
*** 한국학중앙연구원 디지털인문학연구소 학술연구교수 / ddokbaro@gmail.com

Ⅰ. 서 론

전통적인 종이 사전은 2000년대에 들어와서 점점 일반 사용자들에게서 멀어져서, 1990년대부터 2010년대 중반까지 매년 300종 내외의 사전만이 출간되었다. 특히 2019년에는 127종 출간이라는 최하의 수치를 보이며, 이는 사전 판매 부수의 급격한 감소로 인한 결과로 보인다.[1] 이와 반대로 온라인 사전의 편찬양과 사용자는 지속적인 증가 추세이다.[2] 또한, 디지털 기술은 사전 편찬 방법에도 직간접적으로 영향을 주고 있다. 사전 편찬은 전문가에 의한 비교적 주관적인 서술에서 데이터[3]를 바탕으로 하는 비교적 객관적인 서술로 변화해 가고 있다. 90년대의 언어 사전 편찬 방법은 전문가에 의한 표제어 선정과 내용 서술이 기본이었다. 그런데 2000년대 후반부터는 세종21 사업의 대규모의 말뭉치(코퍼스) 자료를 토대로 디지털 기술을 적용하여 표제어 선정과 내용의 서술을 하는 비교적 객관적인 방법으로 변화했다. 또한, 설령 직접 데이터를 토대로 하지 않더라도, 디지털을 기반으로 운용되는 사전 편찬 관리기가 도입되어 사전 항목의 선정, 집필 의뢰, 집필, 검수, 출판의 일련의 과정이 디지털상에서 이루어지고 있으며, 기존의 종이 사전으로 출판된 사전도 디지털로 전환되고 있다.[4] 그리고 디지털의 영향으로 사전 편찬 주체도 전문가에서 대중으로 변화하고 있다. 2010년대부터 우리말샘[5]이나 네이버 사전[6] 사용자 참여 기능을 통해서, 전문가에 의한 사전 편찬에서 일반 이용자가 직접 사전 편찬에 참여하는 형태로 변화해 가고 있다. 마지막으로 사전의 이용 방법에서도 디지털 기술을 활용한 자동완성, 필기 입력, 단어 발음 등의 다양한 새로운 사용자의 사전 접근 방식도 선보이고 있다.[7]

종이 사전의 몰락 속에서 이제 시대의 화두는 데이터가 되었다. 민간 시장뿐만이 아니라 한국 정부에서도 지속해서 데이터에 힘을 집중하고 있다. 2013년 공공데이터법[8]을 통해서 공공기관이 생성 또는 취득하여 관

---

1) 도원영・차준경・신상현・최호철, 「국내 출간 사전의 동향 분석: 연도별 주제별 통계 분석을 중심으로」, 『한민족문화연구』 70, 한민족문화학회, 2020, 264쪽.
2) 김종환, 「2016 네이버 어학사전의 현황과 미래」, 『한국사전학회 학술대회 발표논문집』, 2016: 8, 한국사전학회, 2016, 1~8쪽.
3) 본 논문에서 사용하는 데이터의 개념은 DIKW(Data, Information, Knowledge, Wisdom)에서의 데이터(data) 개념이 아니다. 다시 말해서, 국립국어원 『표준국어대사전』의 "데이터 01 이론을 세우는 데 기초가 되는 사실. 또는 바탕이 되는 자료"가 아니다. DIKW를 배제한 이유는 언어 말뭉치(코퍼스)도 단순한 사실이나 값인 '데이터'가 아니라, 언어학적 의미를 내재한 '정보'로도 볼 수 있기 때문이다. 『온라인 표준국어대사전』도 컴퓨터에서 작동하는 데이터이지만, DIKW의 기준으로는 최소한 '정보'라고 보아야 할 것이다. 이는 컴퓨터를 기반으로 하는 데이터는 그 데이터의 구조와 내용에 따라서, DIKW에서 정의하는 '데이터'와 '정보', 심지어는 '지식'도 될 수 있다는 의미이다. 따라서 본 논문에서 사용하는 데이터의 개념은 국립국어원 『표준국어대사전』의 "데이터 03 컴퓨터가 처리할 수 있는 문자, 숫자, 소리, 그림 따위의 형태로 된 정보"이다.
4) 한국학중앙연구원의 『한국민족문화대백과사전』은 개정 증보 사업을 통해서 2001년에 『디지털 한국민족문화대백과사전 EncyKorea』(CD-ROM 6장, DVD 1장)이 발매되었고, 엠파스, 네이트, 네이버 지식백과, 다음 백과 및 자체적으로 온라인 서비스를 제공하고 있다(https://encykorea.aks.ac.kr/Main/Intro). 단국대학교의 『한한대사전』도 2010년부터 본격적으로 디지털화 작업을 수행하고 있다.
5) 국립국어원, 「우리말샘」, 2016, https://opendict.korean.go.kr/ (검색일 : 2021.02.17.).
6) NAVER Corp., 「NAVER 사전」, 1999, https://dict.naver.com/ (검색일 : 2021.02.17.).
7) 김종환, 앞의 논문, 2016, 1~8쪽.
8) 국가법령정보센터, 「공공데이터의 제공 및 이용 활성화에 관한 법률」, http://www.law.go.kr/법령/공공데이터의제공및이용활성화에관한법률/(11956,20130730) (검색일 : 2021.02.17.).

리하는 데이터를 이용자가 자유롭게 활용할 수 있도록 개방하고 제공했다. 2020년에는 "데이터3법"[9]을 통해서 개인정보를 가명 처리를 통해서 활용할 수 있도록 하고, 한국판 뉴딜의 핵심 산업으로서 데이터 댐을 제시했다. 데이터 댐은 데이터 수집, 가공, 거래, 활용 기반을 강화하여 데이터 경제를 가속화하고 5세대 이동통신(5G) 전국망을 통한 전(全)산업 5세대 이동 통신(5G), 인공지능(AI) 융합 확산을 핵심 내용으로 하고 있다.[10] 특히 정부에서는 인공지능을 위한 데이터를 구축하는 일을 디지털 뉴딜 사업의 토대이자 시작으로 보고 막대한 예산을 배분하고 있다.[11]

이미 우리에게 와 있는 데이터 시대에 사전의 미래에 대해서 논의하기 위해서는 사전의 개념부터 재검토해볼 필요가 있다. 사전(事典)[12] 혹은 사전(辭典)[13]은 "여러 가지 사항을 모아 일정한 순서로 배열하고 그 각각에 해설을 붙인 책"으로 정의할 수 있다. 전통적인 사전의 정의는 종이 형태의 '책'을 필요조건으로 제시했지만, "최근에는 콤팩트디스크 따위와 같이 종이가 아닌 저장 매체에 내용을 담아서 만들기도 한다."라는 문장을 통해 매체의 변화를 언급하였다. 다시 말해서, 사전의 정의에서 '책'이라는 종이 매체는 필수적이지 않고, 어디까지나 특정 대상에 대한 해설이 중요한 것으로 생각할 수 있다. 그런데 기존 사전의 정의에서 누락되어 있는 중요한 요소가 사전의 사용자이다. 지금까지 사전의 사용자는 '인간'이었다. 그런데 디지털 시대에도 최종 사용자는 여전히 '인간'이지만, 그 중간 매개자로 '컴퓨터'가 등장했고, 이에 따라 '기계가독형 데이터' 개념이 필요하게 되었다.

기계가독형 데이터(Machine Readable Data)는 인간가독형 데이터(Human Readable Data)와 대비되는 개념으로 인간이 아닌 컴퓨터가 쉽게 처리할 수 있는 포맷으로 된 데이터를 의미한다. 인간가독형 데이터(Human Readable Data)의 핵심은 인간이다. 인간이 편하게 읽기 위한 데이터가 바로 인간가독형 데이터다. 대표적인 인간가독형 데이터로는 PDF, HWP, DOC 형태의 우리가 자주 사용하는 컴퓨터 파일들이 있다. 이들도 컴퓨터에서 작동을 하는 데이터이지만, 인간을 위한 데이터이기에 컴퓨터에게는 제한적인 정보밖에 제공하지 못한다.[14] 그래서 등장한 것이 컴퓨터의 거대한 저장 공간과 빠른 처리 속도를 활용하기 위해서 컴퓨터가 이해하는 정보를 구축하는 개념인 기계가독형 데이터(Machine Readable Data)이며[15], 현재 정부

9) 문화체육관광부 정책위키, 「데이터 3법」, 2020, http://www.korea.kr/special/policyCurationView.do?newsId=148867915 (검색일 : 2021.02.17.).

10) 기획재정부, 「한국판 뉴딜」, http://www.moef.go.kr/mp/nd/newDeal.do (검색일 : 2021.02.17.).

11) NIA 한국정보화진흥원, AIHUB, 「인공지능(AI) 학습용 데이터 구축 2차」, 2020, http://aihub.or.kr/node/6227 (검색일 : 2021.02.17.).

12) 국립국어원, 『표준국어대사전』, 2008, 「사전12(事典)」, "여러 가지 사항을 모아 일정한 순서로 배열하고 그 각각에 해설을 붙인 책. 최근에는 콤팩트디스크 따위와 같이 종이가 아닌 저장 매체에 내용을 담아서 만들기도 한다.", https://stdict.korean.go.kr/search/searchView.do?word_no=438466&searchKeywordTo=3 (검색일 : 2021.02.17.).

13) 국립국어원, 『표준국어대사전』, 2008, 「사전22(辭典)」, "어떤 범위 안에서 쓰이는 낱말을 모아서 일정한 순서로 배열하여 싣고 그 각각의 발음, 의미, 어원, 용법 따위를 해설한 책. 최근에는 콤팩트디스크 따위와 같이 종이가 아닌 저장 매체에 내용을 담아서 만들기도 한다. ≒말광, 사림, 사서, 어전.", https://stdict.korean.go.kr/search/searchView.do?word_no=438470&searchKeywordTo=3 (검색일 : 2021.02.17.).

14) 예를 들어서, PDF나 HWP에서 '이순신'을 검색하면 '이순신'이 검색되어 나온다. 그런데 '충무공'을 검색하면 '이순신'이 검색되어 나오지는 않는다. 인간들은 '이순신'과 '충무공'을 자연스럽게 연결해서 사고하지만, 컴퓨터에게 '이순신'과 '충무공'은 단순히 다른 문자조합일 뿐이다. 또한 '이순신'과 '충무공'을 연결하지 못하는 것과 같은 원리로, '이순신이 한산도에서 승리하였다.'와 같은 문장도 역시 이해를 하지 못한다.

의 데이터댐에서 요구하는 데이터가 바로 기계가독형 데이터이다.[16)]

그런데 우리가 암묵적으로 생각하는 사전의 정의인 “(인간을 위해서) 여러 가지 사항을 모아 일정한 순서로 배열하고 그 각각에 해설을 붙인 내용”을 “(인간을 위해서 컴퓨터가 이해할 수 있는 방식으로) 여러 가지 사항을 모아 일정한 순서로 배열하고 그 각각에 해설을 붙인 내용”으로 바꿀 수 있다면, 다시 말해서 사전학의 사용자에 ‘컴퓨터’를 추가할 수 있다면, 사전학의 선행 연구 성과들은 데이터 영역에서 충분히 경쟁력이 있다고 생각한다[17)]. 실제 데이터를 다루는 영역에서는 “Garbage-In Garbage-Out[18)]”이라는 불문율이 존재한다. 아무리 뛰어난 알고리즘이라도 문제가 있는 데이터를 사용하면 그 결과도 신뢰할 수 없다는 말이다. 비록 정보의 기록 형식이 종이에서 데이터로 변경되었고, 중간 매체로 컴퓨터가 등장하였지만, 정보의 수집, 항목 선정, 항목 집필, 편찬의 과정과 이를 위한 방법론을 통해서 ‘좋은 정보’를 만들어야 하는 점은 절대 변화하지 않았다는 것이다. 특히 데이터 라벨링 기술의 특성상 시간이 지날수록 전문 영역의 전문 지식에 대한 전문가의 판단 능력이 필요하다는 점[19)]에서 오랜 시간 동안 정보를 다루어왔던 사전학 연구의 성과의 필요성과 활용 가능성이 충분히 존재한다.

이에, 정부의 인공지능 통합 플랫폼인 AIHUB[20)]에 공개된 인공지능 데이터의 구조와 내용을 살펴보고, 이를 토대로 인공지능을 위한 한국역사인물사전의 가능성 탐색을 통해서, 데이터를 중심으로 하는 사전의 미래를 제시해 보고자 한다.

Ⅱ. 인공지능 시대의 데이터

인공지능은 인간의 학습능력과 추론능력, 지각능력, 자연언어의 이해능력 등을 컴퓨터 프로그램으로 실현한 기술[21)]이다. 인공지능은 크게 인간의 지식을 컴퓨터가 이해 가능한 기호로 표기해서 처리하는 심볼릭(Symbolic)과 지식의 표현이 없거나 최소화해서 처리하는 서브심볼릭(Subsymbolic)으로 구분할 수 있다.

---

15) 김바로, 「인공지능 시대를 위한 인문데이터의 역사와 과제」, 『인공지능인문학연구』 3, 중앙대학교 인문콘텐츠연구소, 2019, 77~97쪽.

16) 인문사회 분야에 대한 한국연구재단의 토대연구지원사업에서도 HWP나 DOC가 아닌 비교적 기계가독형에 가까운 Excel, Access, RDB(Relational database), XML(Extensible Markup Language) 등의 형태로 데이터 제출을 요구하고 있으며, 이에 따라서 참여연구원 중 컴퓨터 및 문헌정보학 관련 전공자가 최소 1인 이상 참여하도록 필수화하였다.

17) 비록 사전학에서도 21세기 세종계획의 성과물 중에 하나인 표준화된 전자사전을 구축하는 등 기계가독형데이터를 구축하기 위한 노력이 있었으나, 당시의 기계가독형데이터는 현재의 시맨틱데이터 개념처럼 발전하지 않았으며, 기계가독형데이터 구축 자체가 사업의 중심 개념이 아니었던 아쉬움이 있다.

18) Corrales, David Camilo, Agapito Ledezma, and Juan Carlos Corrales, “A conceptual framework for data quality in knowledge discovery tasks (FDQ-KDT): A Proposal.” *Journal of Computers* 10: 6, 2015, pp.396~405.

19) 성수영, 「21세기판 인형 눈붙이기……한국판 뉴딜 일자리 체험해보니」, 『한경닷컴』, 2020년 07월 25일자, https://www.hankyung.com/economy/article/202007245362i (검색일 : 2021.02.17.).

20) NIA 한국정보화진흥원, 「AI HUB」, 2019, http://www.aihub.or.kr/ (검색일 : 2021.02.17.).

21) DOOSAN, 『doopedia』, 2000, 「인공지능」, https://www.doopedia.co.kr/doopedia/master/master.do?_method=view&MAS_IDX=101013000752124 (검색일 : 2021.02.17.).

2010년 이전에는 심볼릭 계열의 소위 '전문가시스템(expert system)'이 인공지능 영역의 대세였다. 그런데 심볼릭 계열은 어렵고 복잡한 인간 지식을 컴퓨터에게 이해시키기 위해서 해당 지식 체계를 이해하는 다수의 전문 인력이 장기간 투입되어야 하는 한계가 있었다. 이에 최근에는 방대한 데이터를 토대로 어렵고 복잡한 인간 지식 체계 자체를 유추하는 서브심볼릭 계열의 딥러닝 기술이 급격하게 부상했다.

1. 딥러닝을 위한 데이터

(1) 레이블드 데이터

딥러닝 기술을 위한 기본적인 데이터 형태는 레이블드 데이터(labeled data)이다.[22] 레이블드 데이터는 대상과 대상에 대한 의미를 표시한 라벨로 구성한다. 예를 들어서, 전형적인 레이블드 데이터인 AIHUB의 "한국어 감정 정보가 포함된 단발성 대화 데이터셋[23]"은 텍스트를 통한 인공지능 감정 식별을 위해서, 한국어 텍스트에 대해서 7개의 감정(기쁨, 슬픔, 놀람, 분노, 공포, 혐오, 중립)을 라벨링한 총 38,594건의 데이터이다.

〈표 1〉 "한국어 감정 정보가 포함된 단발성 대화 데이터셋" 샘플

Sentences	Emotion
언니 동생으로 부르는게 맞는 일인가요..??	공포
의사양반 내가고자라니!	놀람
서해 수자원도 한류라서 그동안 불법조업 방치했나?	분노
그저 보고싶어죽겟어요	슬픔
소량은 먹어도 인체에 해롭지 않고요.	중립
우리모두 준혁이성과 LG선수들이 선전할수 있게 기를 넣어줍시다.	행복
세금을 쳐 내라 양아치들아	혐오

"한국어 감정 정보를 포함된 단발성 대화 데이터셋"에서는 한국어 문장을 대상으로 하고, 그에 대한 의미로 7개의 감정을 선택해서 사람에 의해서 감정을 판별한 라벨이 있다. 심볼릭에서는 개별 문장의 특정 형태소, 특정 형태소의 조합, 특정 형태소의 위치 등의 다양한 감정 판별을 위한 전문가들의 판단 관련 정보를

22) 엄밀하게 이야기해서, 레이블드 데이터는 머신러닝 기술 중에서 지도학습 영역을 위한 데이터에 필요하다. 머신러닝 기술에는 비지도학습과 강화학습의 영역도 존재하며, 이를 위한 데이터의 구조는 레이블드 데이터와는 다르다. 하지만 지도학습의 영역이 현재 가장 보편적이고 상용화된 영역인 만큼 충분히 딥러닝 기술 혹은 머신러닝 기술을 위한 대표적인 데이터로 부를 수 있다고 본다.

23) ㈜아크릴, KETI 지능정보 플래스십 R&D,「한국어 감정 정보가 포함된 단발성 대화 데이터셋」, 2017, http://aihub.or.kr/keti_data_board/language_intelligence (검색일 : 2021.02.17.).

알려주어야 했지만, 서브심볼릭인 딥러닝에서는 알고리즘이 많은 문장-감정 데이터 세트를 토대로 자동으로 감정 판별의 요소들을 추론한다.

따라서, 레이블드 데이터에서는 문장-감정 데이터 세트의 연결 정확도, 문장 데이터의 오·탈자나 대표성 문제가 여전히 중요하긴 하지만, 극단적으로 설령 데이터에 오류가 있더라도 압도적으로 많은 데이터양이 있다면 데이터의 오류로 인한 알고리즘의 오류를 최소화할 수 있다. 이에 따라서 대다수의 레이블드 데이터에서는 비교적 데이터양을 중시하고 있다. 0부터 9까지의 10개 숫자의 손글씨 이미지를 판독하기 위한 레이블드 데이터인 MNIST 데이터셋[24]은 총 7만 개의 데이터로 구성되어 있다. 1개의 숫자를 식별하기 위해서 7천 개 정도의 이미지 데이터가 제공되고 있다. 또한, IMDB 해외 영화 평가 사이트의 리뷰를 대상으로 하고, 그에 대한 긍·부정 평가를 라벨 데이터로 사용한 긍·부정 데이터[25]도 총 5만 개의 데이터로 구성되어 있다. 긍정과 부정이라는 2개의 선택지에 관한 판단을 위해서 긍·부정 결과로 각각 2.5만 개의 데이터를 제공하고 있다.

하지만 레이블드 데이터에서도 여전히 데이터의 질도 중요하기에 사전 편찬의 집필과 검수 방법론을 통한 데이터 정확도 확보가 필요하다. 무엇보다 MNIST 데이터의 숫자나 IMDB의 긍·부정처럼 단순한 데이터가 아닌 감정에 대한 레이블드 데이터에서는 사전의 항목 선정처럼 감정 선정에 대한 연구, 다시 말해서 데이터 설계에 관한 연구가 필요하다. 현재 "한국어 감정 정보가 포함된 단발성 대화 데이터셋"은 로버트 플루치크(Robert Plutchik)의 8감정 분류를 기반으로 신뢰, 기대를 제외하고 중립을 넣은 것으로 보인다. 그런데 감정의 선정에서 "한국어 감정 정보가 포함된 단발성 대화 데이터셋"에서 차용한 방식이 절대적인 정답이 아니며, 기존 연구를 바탕으로 한 다양한 감정 분류에 대한 연구가 필요하다는 점이 중요하다.[26] 이는 전문가에 의한 데이터 설계와 이를 위한 연구가 여전히 중요하다는 점을 의미한다.

(2) 병렬코퍼스(말뭉치)

병렬코퍼스(말뭉치)는 동일한 의미가 있는 두 언어의 문장을 병렬로 구성한 레이블드 데이터의 변형이라고 할 수 있다. 레이블드 데이터가 대상과 대상에 대한 라벨로 구성되어 있는 반면, 병렬코퍼스는 양쪽 언어의 텍스트 모두가 대상이자 라벨이라는 점에서 다르다. 예를 들어서, 전형적인 병렬코퍼스인 AIHUB의 "한국어-영어 번역(병렬) 말뭉치 AI데이터"[27]는 기계번역을 위해서 6개 영역(뉴스, 정부 웹 사이트/저널, 법률, 한국문화, 구어체, 대화체)에서 한국어와 영어의 대응 번역 문장으로 구성한 160만 건의 데이터이다.

---

24) LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P., "Gradient-based learning applied to document recognition", *Proceedings of the IEEE* 86: 11, 1998, pp. 2278~2324.

25) Maas, A., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C., "Learning word vectors for sentiment analysis", *In Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies,* 2011, pp. 142~150.

26) 이유미·김바로·박지영, 「한국어 감정 디지털 온톨로지 구축에 관한 연구」, 『한국어 의미학』 68, 한국어의미학회, 2020, 140~142쪽.

27) ㈜솔트룩스파트너스, ㈜에버트란, ㈜플리토, NIA 한국정보화진흥원, 「한국어-영어 번역(병렬) 말뭉치 AI데이터」, 2019, http://aihub.or.kr/aidata/87 (검색일 : 2021.02.17.).

〈표 2〉 "한국어 – 영어 번역(병렬) 말뭉치 AI데이터" 샘플

원문	번역문
'Bible Coloring'은 성경의 아름다운 이야기를 체험 할 수 있는 컬러링 앱입니다.	Bible Coloring' is a coloring application that allows you to experience beautiful stories in the Bible.
씨티은행에서 일하세요?	Do you work at a City bank?
푸리토의 베스트셀러는 해외에서 입소문만으로 4차 완판을 기록하였다.	PURITO's bestseller, which recorded 4th rough -cuts by words of mouth from abroad.
11장에서는 예수님이 이번엔 나사로를 무덤에서 불러내어 죽은 자 가운데서 살리셨습니다.	In Chapter 11 Jesus called Lazarus from the tomb and raised him from the dead.
6.5, 7, 8 사이즈가 몇 개나 더 재입고 될지 제게 알려주시면 감사하겠습니다.	I would feel grateful to know how many stocks will be secured of size 6.5, 7, and 8.

병렬코퍼스는 이미 오래전부터 언어학, 번역학 연구의 핵심 데이터로 사용됐다. 심볼릭에서는 언어 간의 차이에 대해서 형태소, 형태소의 구성, 문화적 차이 등의 다양한 번역에서의 전문가들의 판단 관련 정보를 알려주어야 했지만, 서브심볼릭인 딥러닝에서는 알고리즘이 병렬코퍼스를 바탕으로 자동으로 번역 판별 요소들을 자동으로 추론한다. 병렬코퍼스의 데이터 설계에서는 사전 항목 추출과 같은 설계는 필요하지 않다. 다만 병렬코퍼스의 질적인 문제인 번역 품질의 향상을 위해서 기존 번역 기업의 데이터를 토대로 번역 내용을 수정, 보완하는 데 집중하고 있다.

현 단계에서 병렬코퍼스는 데이터양을 증대하는 데 집중하고 있다. 그러나 일상대화, 뉴스, 법률문서, 학술대회 등 다양한 언어 환경에서의 사용 어휘의 차이점을 반영하기 위해서는 대상 언어 도메인(Domain)에 대한 판별과 응용이 필요하다. 그런데 개별적인 도메인에 대한 학습에 또다시 방대한 데이터를 구축하기는 쉽지 않다. 이에 따라서, 인공지능 영역에서는 도메인 구별 없이 최대한 많은 양의 텍스트를 바탕으로 구축한 기본 모델을 구현하고, 기본 모델을 토대로 비교적 적은 데이터로 개별 도메인의 특수성을 학습시키는 전이학습(Transfer Learning)에 관한 연구를 수행하고 있다. 기본 모델의 구현에는 사전학이 개입할 여지는 많지 않지만, 반대로 개별 도메인의 특수성을 반영하는 데이터의 구축에는 개별 도메인의 전문가들이 반드시 필요하다. 그리고 이러한 특정 분야에 대한 전문 지식의 필요성과 수요는 종이 사전 시장에서도 드러난다.[28]

(3) 스쿼드

스쿼드(SQuAD, Standford Question Answering Dataset) 데이터는 인공지능이 스스로 문서를 분석하고 질문에 대한 최상의 답을 찾아내는 기계독해(Machine Reading Comprehension, MRC)를 위해서 만들어진 데이터이다. 예를 들어서, AIHUB의 "기계독해"[29]는 딥러닝을 통한 정답 자동 추론을 위해서, 한국어로 구성된 본문과 본문에서 도출되는 질문과 답, 그리고 질문 유형으로 구성한 45만 건의 데이터이다.

28) 도원영, 앞의 논문, 2020, 262쪽.
29) ㈜마인즈랩, NIA 한국정보화진흥원, 「기계독해」, 2019, http://aihub.or.kr/aidata/86 (검색일 : 2021.02.17.).

〈표 3〉 "기계독해" 샘플

본문	질문	답변	종류
한국청소년단체협의회와 여성가족부는 22일부터 28일까지 서울과 충북 괴산에서 '국제청소년포럼'을 연다고 21일 밝혔다. 한국 미국 캐나다 호주 등 전 세계 32개국 75여명의 대학생, 청소년들이 모여 전 세계적 현안문제에 대한 대안과 해결책을 모색하는 자리다. 이번 포럼의 주제는 '청소년과 뉴미디어'다. 스마트폰 SNS 태블릿PC 등 새로운 커뮤니케이션 매체인 '뉴미디어'에 대한 성찰과 문제점에 대해 토론한다. 기조강연을 시작으로 국가별 주제관련사례발표, 그룹 토론 및 전체총회, '청소년선언문' 작성 및 채택 등 다양한 프로그램을 운영한다. 개회식은 22일 서울 방화동에 있는 국제청소년센터 국제회의장에서 한다. 전 세계 32개국 대학생·청소년참가자와 전국의 청소년기관단체장과 청소년지도자 여성가족부주한외교사절 등 100여명이 참석할 예정이다. 23일에는 유엔미래포럼 박영숙 대표가 '뉴미디어의 균형 있는 발전을 위한 청소년의 역할'에 대해 기조강연을 한다. 뉴미디어의 올바른 활용방안과 청소년문화의 형성에 대해 설명할 계획이다. 27일 폐회식에서는 '청소년선언문'을 채택한다. 선언문에는 전 세계적으로 뉴미디어의 바람직한 발전을 촉구하며 각국 청년들이 함께 실천할 수 있는 내용 등이 담길 예정이다. 한국청소년단체협의회는 포럼이 끝난 뒤 UN 등 국제기구와 참가자 각국 정부 등 국제사회에 선언문을 전달할 예정이다.	서울과 충북 괴산에서 '국제청소년포럼'을 여는 곳은?	한국청소년단체협의회와 여성가족부	work_who
	국제 청소년포럼'이 열리는 때는?	22일부터 28일	work_when
	이번 포럼의 주제는?	청소년과 뉴미디어'	work_what
	포럼은 어떻게 진행되는가?	기조강연을 시작으로 국가별 주제관련사례발표, 그룹 토론 및 전체총회, '청소년선언문' 작성 및 채택 등 다양한 프로그램을 운영한다.	work_how
헤럴드POP=고승아 기자그룹 구구단 샐리가 리듬체조에서 실수했다.15일 방송된 MBC '설특집 2018 아이돌스타 육상 볼링 양궁 리듬체조 에어로빅 선수권대회(이하 아육대)'에서는 구구단 샐리가 리듬체조 종목에 출전한 모습이 그려졌다.이날 구구단 샐리는 리듬체조 종목에 처음 출전해 훌라후프로 무대를 꾸몄다. 그러나 연기 도중 훌라후프를 받지 못하는 등 실수를 해 안타까움을 자아냈다.응원을 하던 구구단 멤버들도 열심히 연습했던 샐리를 생각해 눈물을 흘렸다. popnews@heraldcorp.com-Copyrights ⓒ 헤럴드POP & heraldpop.com, 무단 전재 및 재배포 금지 -	아육대에서 리듬체조에 출전한 구구단의 멤버는?	샐리	work_who

스쿼드 데이터는 정보를 재구성하는 사전 편찬보다는 문제 출제에 더욱 가깝다고 할 수 있다. 따라서 스쿼드 데이터의 구축에서는 다양한 영역의 다양한 유형의 질문과 답변을 정확하게 구성하는 것이 중요하다. 스쿼드 데이터도 현재는 일반 상식 수준의 데이터를 중심으로 구축되고 있지만, 추후 전문 영역으로의 확장이 필요할 것으로 보인다. 그리고 결국 병렬코퍼스와 동일하게 특정 도메인 영역 전문가들이 필요하다.

2. 시맨틱 데이터

딥러닝이 인공지능의 전부인 것처럼 여겨지고 있지만, 딥러닝은 블랙박스 문제,[30] 컴퓨팅 파워의 한계[31] 등으로 인하여, 현재의 딥러닝 기술이 그 가능성과 비교하면 과도한 관심을 받고 있기에 오히려 또다시 인공

지능의 겨울이 올 것을 염려하는 의견도 엄연히 존재하고 있다. 이에, 서브심볼릭의 딥러닝과 심볼릭의 시맨틱 데이터를 융합하는 앙상블 기술이 대안으로 제시되고 있다.[32]

시맨틱 데이터란 컴퓨터가 이해할 수 있는 '의미'를 내포하고 있는 데이터로서 시맨틱 웹의 개념을 토대로 생성된 데이터를 의미한다. 시맨틱 웹(Semantic Web)은 "의미론적인 웹"이라는 뜻으로, 현재의 인터넷과 같은 분산환경에서 리소스(웹 문서, 각종 화일, 서비스 등)에 대한 정보와 자원 사이의 관계 – 의미 정보(Semanteme)를 기계(컴퓨터)가 처리할 수 있는 온톨로지 형태로 표현하고, 이를 자동화된 기계(컴퓨터)가 처리하도록 하는 프레임워크이자 기술이다.[33] 시맨틱 데이터는 기본적으로 "주어+동사+목적어"의 구조로 되어 있다.

http://lod.koreanhistory.or.kr/resources/인명_이순신
http://www.ddokbaro.com/ontologies/참가했다.
http://encykorea.aks.ac.kr/Contents/Item/E0061676(한산도대첩)[34]

위의 데이터는 "(국사편찬위원회에서 관리하는) 이순신은 (한국학중앙연구원 한국민족문화대백과사전에서 관리하는) 한산도대첩에 (김바로가 정의하는) "참가했다"로 풀어쓸 수 있다.[35] 일반적인 서술에 비해서 무의미한 정보가 증대했다고 생각할 수도 있지만, 인간가독형 데이터가 아닌 기계가독형 시맨틱 데이터를 채택함으로써 컴퓨터는 국사편찬위원회에서 관리하는 '이순신'의 또 다른 정보인 출생시간(birthData), 출생지(birthPlace), 봉호(bonghoName), 사망시간(deathData), 사망지(deathPlace), 본적(familyPlace)은 물론이고, 이순신과 관련된 작품(hasArtworkInfo),[36] 이순신의 관직정보(hasCarrerInfo),[37] 이순신과 관련된

30) 구본권, 「'인공지능 세번째 겨울, 피할 길은 "설명 가능성"에 있다'」, 『한겨레』, 2019년 04월 08일자, http://www.hani.co.kr/arti/science/future/889159.html (검색일 : 2021.02.17.).

31) 이재구, 「MIT, '딥러닝, 컴퓨팅 파워 한계 접근'……'AI겨울' 해결책은?」, 『AI타임스』, 2020년 07월 17일자, http://www.aitimes.com/news/articleView.html?idxno=130732 (검색일 : 2021.02.17.).

32) 최광민, 「솔트룩스, 학습과 상식 추론 결합한 '앙상블 AI', '설명가능한 AI' 상용화로 '3세대 AI'로 진화」, 2020년 04월 13일자, 『인공지능신문』, http://www.aitimes.kr/news/articleView.html?idxno=16034 (검색일 : 2021.02.17.).

33) 위키미디어 재단, 『위키백과』, 「시맨틱 웹」, https://ko.wikipedia.org/wiki/시맨틱_웹 (검색일 : 2021.02.17.).

34) 한국학중앙연구원 한국민족문화대백과사전은 전통적인 사전을 XML(Extensible Markup Language) 형식으로 디지털화했을 뿐, 시맨틱 데이터로 구성되어 있지 않다. 다만 본 논문에서는 가능성에 대한 예시로서 한국민족문화대백과사전은 시맨틱 데이터로 가정했다.

35) 이에 대한 기술적인 세부 내용은 다음을 참고한다. 김바로, 앞의 논문, 2019, 91~93쪽; 김현, 임영상, 김바로, 『디지털 인문학 입문』, HUEBOOKS : 한국외국어대학교 지식출판원, 2016, 163~183쪽.

36) hlod : 난중일기(亂中日記), hlod : 이충무공전서(李忠武公全書), hlod : 난중일기(亂中日記), hlod : 이순신행장급사적(李舜臣行狀及事蹟), hlod : 이충무공전서(李忠武公全書)

37) hlod : 권지훈련원봉사(權知訓練院奉事), hlod : 함경도동구비보권관(咸鏡道童仇非堡權管), hlod : 훈련원봉사(訓練院奉事), hlod : 충청병사군관(忠淸兵使軍官), hlod : 발포수군만호(鉢浦水軍萬戶), hlod : 훈련원봉사(訓練院奉事), hlod : 건원보권관(乾原堡權管), hlod : 훈련원참군(訓練院參軍), hlod : 사복시주부(司僕寺主簿), hlod : 조산만호(造山萬戶), hlod : 겸 녹도둔전사의(鹿島屯田事宜), hlod : 정읍현감(井邑縣監), hlod : 고사리진병마첨절제사(高沙里鎭兵馬僉節制使), hlod : 만포진수군첨절제사(滿浦鎭水軍僉節制使), hlod : 진도군수(珍島郡守), hlod : 가리포첨절제사(加里浦僉節制使), hlod : 전라좌도수군절도사(全羅左道水軍節度使), hlod : 전라도좌수사(全羅道左水使), hlod : 평안도우후(平安道虞侯), hlod : 전라좌도수군절도사(全羅左道水軍節度使), hlod : 겸 삼도수군통제사(三道水軍統制使), hlod : 동구비보권관(童仇非堡權管), hlod : 선전관(宣

사건(hasRelatedEventInfo)[38]을 식별할 수 있게 된다. 그리고 이런 식별된 정보를 통해 컴퓨터는 '명량해전'에 참가한 인물들의 정보를 대답할 수 있으며, 이순신의 아버지 형제들의 정보를 추론할 수 있게 된다.

이러한 시맨틱 데이터는 딥러닝이 강조되는 현재에도 인공지능에서 광범위하게 활용되고 있다. 예를 들어서, AI 스피커에서 사용자의 질문에 대답하는 과정에서 딥러닝의 활용 영역은 생각보다 많지 않다. 거칠게 말해서, 딥러닝은 사용자의 음성을 텍스트로 변환하는 STT(Speech-to-Text)와 도출된 결과값을 사용자에게 음성으로 전달하기 위한 TTS(Text-to-Speech)에서 핵심으로 사용한다. 그런데 정작 가장 중요하다고 볼 수 있는 사용자 질문에 관한 결과 도출에서는 시맨틱 데이터를 토대로 하는 추론을 활용한다. 다시 말해서, AI 스피커에 이순신의 사망 장소를 질의했을 때, 그에 대한 대답을 추론하는 부분은 기본적으로 시맨틱 데이터를 바탕으로 하고 있다.

사전학에서 주목할 만한 점은 시맨틱 데이터가 정보의 고유성과 확장성을 모두 만족하고 있다는 점이다. 시맨틱 데이터에서는 정보의 출처를 밝히는 동시에 고유식별값을 부여하기 위해서 URL(Uniform Resource Locator)을 URI(Uniform Resource Identifier, 통합 자원 식별자)로 사용한다. 예를 들어서, 국사편찬위원회에서 관리하고 있다는 점을 알려주기 위해서 국사편찬위원회의 URL인 "http://lod.koreanhistory.or.kr"을 사용했다. 동일한 원리로 한국학중앙연구원 한국민족문화대백과사전에서 관리하고 있다는 점을 알려주기 위해서, 한국학중앙연구원 한국민족문화대백과사전의 URL인 "http://encykorea.aks.ac.kr/[39]"를 사용했다. 그리고 '동사'라고 볼 수 있는 "참가했다"도 개념 정의자를 명확하게 하기 위해서 "http://www.ddokbaro.com/ontologies/"를 사용했다. URI를 통해서 개별 정보들의 고유성을 명확하게 해주는 한편, 개별 정보들을 상호 연결하여 정보의 확장이 이루어진다. 비록 국사편찬위원회에서는 "한산도 해전"[40]에 대한 정보를 제공하고 있지만, 한산도 대첩의 발생일(hasDate), 관련인물(hasRelatedPerson) 정도의 정보만 제공하고 있다. 그에 반해서 한국민족문화대백과사전에서는 한산도 대첩에 대한 발생장소, 연구문헌 등 국사편찬위원회에서 제공하지 않는 정보를 제공하고 있다. 따라서 이 둘을 연결함으로써 이순신에 대한 정보가 더 풍부해진다. 반대로 한국민족문화대백과사전은 비록 한산도대첩에 대한 전반적인 지식을 제공하고 있지만, 이순신, 원균과 같은 개별 인물에 대한 상세 정보는 기술하지 못하고 있다. 그런데 국사편찬위원회의 이순신 데이터와 연결하여 이순신 개인에 대한 상세 정보를 획득할 수 있다. 시맨틱 데이터는 이처럼 국사편찬위원회의 이순신과 한국민족문화대백과사전의 한산도대첩을 상호 연결하여, 이순신과 한산도 대첩에 대한 정보를 확장할 수 있다.

시맨틱 데이터는 딥러닝에 대비하여 정보의 정확도와 확장성이 보장되기에, 시맨틱 데이터를 기반으로 하

傳官), hlod : 전라좌도수군절제사(全羅左道水軍節制使), hlod : 삼도통제사(三道統制使), hlod : 삼도통제사(三道統制使), hlod : 동구비보권관(童仇非堡權管), hlod : 훈련원봉사(訓鍊院奉事), hlod : 발포수군만호(鉢浦水軍萬戶), hlod : 건원보권관(乾原堡權管), hlod : 건원보권관(乾原堡權管), hlod : 선전관(宣傳官), hlod : 정읍현감(井邑縣監), hlod : 고사리진병마첨절제사(高沙里鎭兵馬僉節制使), hlod : 만포진수군첨절제사(滿浦鎭水軍僉節制使), hlod : 전라좌도수군절제사(全羅左道水軍節制使)

38) hlod : 옥포대첩, hlod : 당항포대첩, hlod : 한산대첩, hlod : 명량해전, hlod : 사건정보_50326, hlod : 부산포해전

39) 엄밀하게 한국민족문화대백과사전의 '한산도대첩'은 시맨틱 데이터로 구성되어 있지 않으며, 고유URL도 http://encykorea.aks.ac.kr/Contents/Item/E0061676이지만, 시맨틱 데이터의 개념에 대한 이해를 돕기 위해서 예시로 사용했다.

40) 국사편찬위원회, 「한국사 LOD」, 「사건정보_50323」, 2014, http://lod.koreanhistory.or.kr/resources/사건_한산도_해전 (검색일 : 2021.02.17.).

는 정보 공유 방식인 LOD(Linked Open Data)를 다양한 기관에서 이미 구축하고 서비스하고 있다. 서울시의 데이터를 토대로 한 서울시 LOD[41]나 국사편찬위원회의 데이터를 토대로 하는 한국사 LOD[42]가 존재하고 있다. 또한 정부의 데이터댐 프로젝트에서도 시맨틱 데이터 구축을 진행하고 있다. 한국어 위키백과 데이터를 토대로 구축한 총 75만 건의 RDF 데이터인 "일반상식[43]"이나, 법령, 판례와 여기서 도출되는 용어 사전 등을 토대로 구축한 "법률[44]" 등이 있다.

그런데 '이순신'과 '한산도대첩'과 같이 비교적 간단한 연결은 일반인도 수행 가능하지만, 보다 깊은 지식에 대한 정보의 연결은 해당 분야의 전문가만이 수행 가능하다. 문제는 현재 시맨틱 데이터의 구축은 해당 분야의 전문가가 아닌 정보처리 관련 기술자에만 의존하고 있다는 점이다.

예를 들어서, 국사편찬위원회의 한국사 LOD에서는 대부분의 연결을 '관련'으로만 처리하고 있다. '관련 인물(hasRelatedPerson)'을 통해서 인물과 관련한 인물을 연결은 하고 있지만, 구체적인 관련성에 대해서는 기술하고 있지 않다. 단순히 이순신이 선조, 류성룡, 원균과 관련이 있다고만 기술할 뿐이며, 구체적으로 군주-신하 관계, 정치적 후원 관계, 정치적 대립 관계 등의 구체적인 연결에 대해서는 서술하지 못하고 있다. 결국 전문적인 지식 정보를 시맨틱 데이터로 구축하기 위해서는 시맨틱 데이터의 설계, 구축, 검수의 전 과정에서 해당 분야 전문가의 참가가 필수불가결하다.

Ⅲ. 인공지능을 위한 한국역사인물 데이터

1. 딥러닝을 통한 한국역사인물 식별

사료에서 역사인물을 자동으로 식별하기 위해서, 비정형 텍스트 데이터에서 개체명을 식별하는 딥러닝(머신러닝) 기술의 하나인 NER(Named Entity Recognition, 개체명식별)을 사용했다. NER을 구현하기 위해서는 컴퓨터에게 개체명을 식별하는 학습용 데이터를 제공할 필요가 있다. NER을 위한 학습 데이터의 최소한의 스키마는 문장고유ID, 형태소, 개체태그이다. 개체태그는 O, B, I로 구성하는데, O는 개체가 아닌 모든 대상을 의미하고, B는 개체의 시작, I는 개체의 지속을 의미한다. 개체의 종류는 학습의 필요성에 따라서 설정한다.

본 논문에서는 XML 기반의 『조선왕조실록』의 데이터[45]를 토대로 NER 학습 데이터를 구축하였으며,[46]

41) 서울특별시, 「서울 열린데이터 광장 LOD 라이브 서비스」, http://lod.seoul.go.kr/ (검색일 : 2021.02.17.).
42) 국사편찬위원회, 「한국사 LOD」, 2014, http://lod.koreanhistory.or.kr/ (검색일 : 2021.02.17.).
43) (주)마인즈랩, NIA 한국정보화진흥원, 「일반상식」, 2017, http://aihub.or.kr/aidata/84 (검색일 : 2021.02.17.).
44) 와이즈넛, 코아인포메이션, NIA 한국정보화진흥원, 「법률」, 2017, http://aihub.or.kr/aidata/029 (검색일 : 2021.02.17.).
45) 국사편찬위원회의 『조선왕조실록』 데이터는 「공공데이터법」에 의거하여, 공공데이터포털을 통해서 정보공개 신청을 해서 2016년 5월 19일에 확보한 데이터이다.
46) 본 논문은 NER를 본격적으로 도입하는 논문이 아닌 딥러닝의 장단점을 간략하게 보여주기 위한 논문이기에, 국사편찬위원

『조선왕조실록』 데이터에서 제공하는 객체명인 인명, 연호, 지명, 서명을 각각 PER, DAT, LOC, POH로 설정하였다.[47]

〈표 4〉 조선왕조실록의 데이터를 토대로 한 NER 학습 데이터 샘플

Sid	Word	Tag
2	穆	B-PER
2	祖	I-PER
2	聞	O
2	之	O
2	,	O
2	遂	O
2	徙	O
2	居	O
2	江	B-LOC
2	陵	I-LOC
2	道	I-LOC
2	三	B-LOC
2	陟	I-LOC
2	縣	I-LOC

SId는 『조선왕조실록』 개별 항목 기사의 고유값이고, Word는 1자 단위로 구분된 문자이다. 마지막으로 Tag는 개체명을 식별한 데이터로, 穆은 인명의 시작이기에 B-PER, 祖는 인명의 연속이기에 I-PER, 聞은 객체가 아니기에 O이다. 江陵道도 같은 원리로 江은 지명의 시작이기에 B-LOC, 陵과 道는 지명의 연속이기에 I-LOC를 입력하였다. 이렇게 구성된 학습 데이터와 다국어 BERT[48] 모델[49]을 토대로 NER 학습을 수행했

회 『조선왕조실록』의 태조 총서부터 태종 10년 3월 9일 2번째 기사(http://sillok.history.go.kr/id/wca_11003009_002), 총 9,999건의 데이터만을 토대로 진행하였다. 향후 본격적으로 『조선왕조실록』과 『한국문집총간』 등의 최대한 많은 한국 역사 데이터를 토대로 NER을 적용하는 연구를 진행할 예정이다.

47) 『조선왕조실록』의 데이터를 토대로 한 전체 NER 학습 데이터: https://osf.io/h2npb/ 혹은 https://drive.google.com/file/d/1MU-u3Uy055lZ8EBW-zYTpSJtiZTmsMrd/

48) BERT(Bidirectional Encoder Representations from Transformers, 버트)는 위키피디아, 뉴스 등의 방대한 텍스트를 토대로 사전 훈련을 수행한 범용 언어 모델이다. 특정한 언어 처리를 하기 이전에 언어에 대한 기본 모형을 만들고, 이를 토대로 개별 언어 처리에 활용함으로써 기존보다 언어 처리 성능을 비약적으로 향상시킨 모델이다. 본 논문의 핵심은 알고리즘이 아닌 데이터이기에 구체적인 내용을 생략한다.

49) 본 발표에서 사용한 BERT 모델은 “bert-base-multilingual-cased”이다. bert-base-multilingual-cased는 104개의 언어를 토대로 구축한 BERT 모델이다. Google Research, 「bert」, https://github.com/google-research/bert (검색일 : 2021.02.17.).

다.[50][51] 이를 『승정원일기』에서 랜덤으로 추출한 항목[52]에 대입하면 다음과 같이 자동으로 인물을 식별하는 것을 볼 수 있다.[53]

〈표 5〉 조선왕조실록 기반 NER 모델을 적용한 승정원일기 결과 샘플

Predict	Word
O	掌
O	令
B-PER	金
I-PER	載
I-PER	顯
O	啓
O	曰
O	,
O	請
B-PER	柳
I-PER	炳
I-PER	然
O	削
O	奪
O	官
O	爵
O	事
O	。

또한 인명, 연호, 지명, 서명 이외에 관직명, 물품명, 한자의 표제자음, 뜻갈래 등과 같은 사항을 식별하고 싶다면, 그에 해당하는 NER 학습 데이터를 구성해야 한다. 그리고 이를 위해 필요한 것은 더욱 강력한 딥러닝 알고리즘이 아니라, 관직명, 물품명 혹은 한자의 표제자음, 뜻갈래 등 식별을 원하는 대상에 대한 학습 데이터이며, 이 과정에서 전문가가 반드시 필요하다는 점이 중요하다. 본 연구에서 활용한 조선왕조실록 데이

50) Multi_NERwithBert_조선왕조실록 전체 코드: https://osf.io/7bvge/ 혹은 https://drive.google.com/file/d/1zGx_IuMaluvXAEy8hKL9luczWsCvFk88/ (검색일 : 2021.02.17.).

51) 학습 결과 Validation Accuracy는 0.9917179931830413이 나왔다. 99.17% 이상의 정확도를 보인다고 할 수 있다.

52) 국사편찬위원회, 『승정원일기』, 2015, 277책 (탈초본 14책) 숙종 6년 7월 1일 무자 4/9 기사, 「掌令金載顯啓曰, 請柳炳然削奪官爵事。」, http://sjw.history.go.kr/id/SJW-D06070010-00400 (검색일 : 2021.02.17.).

53) 본 NER 학습에서 사용한 모든 데이터와 알고리즘은 다음 URL을 통해서 제공한다. https://doi.org/10.17605/OSF.IO/5FBHP

터도 전문가에 의해 인명, 연호, 지명, 서명이 식별된 것이며, 이는 딥러닝 기술로 인하여 전문가의 일이 점차 없어지는 것처럼 여겨지지만, 실제로는 대체하기 어려운 전문가의 역할이 더더욱 중요해지고 있다는 의미이다.

그런데 딥러닝을 통한 인명의 '식별'은 어디까지나 해당 단어의 인명 여부를 판별하는 것이며, 인문학에서 말하는 인물의 정체성에 대해서 파악하는 '식별'과는 다르다. 인문학의 인명 식별을 위해서는 인물에 대한 모든 배경 정보를 파악할 필요가 있다. 그런데 인물의 생몰년, 생몰지, 혈연, 학연, 관직 등의 다양한 데이터는 딥러닝용 데이터 형태로는 편찬이 어렵거나 불가능하기에, 시맨틱 데이터의 형태가 필요하다.

2. 온톨로지를 통한 한국역사인물 식별

인물에 대한 시맨틱 데이터를 구축하기 위한 온톨로지 설계는 일반적으로 대상 자원을 클래스(Class)로 범주화하고, 각각의 클래스에 속하는 개체(individuals)들이 공통의 속성(attribute)을 갖도록 하고, 그 개체들이 다른 개체들과 맺는 관계(relation)를 명시적으로 기술한다.[54] 인물, 그리고 이와 관련된 주요 요소인 공간, 사건, 기관 등의 클래스(Class)를 선정한다. 클래스는 기존 사전 편찬에서의 분류 체계 구축과 거의 유사하며, 개별 항목들에서 서로 유사한 성격을 묶어서 범주로 정하는 것이다.

〈표 6〉 한국역사인물 클래스 설계 샘플

Class Name	한글명칭	subClassOf	설명
System	제도		(인사) 제도에 관한 정보
Term	용어		용어(직품, 직위 등)에 관한 정보
Event	사건		역사적 변화(사건)에 관한 정보
Actor	행위자		모든 행위자에 관한 정보
Group	조직	Actor	조직(관청 등)에 관한 정보
Person	인물	Actor	인물에 관한 정보

그 뒤에 개별 클래스의 속성을 선정한다. 속성(attribute, data property, metadata)은 기존 사전 편찬의 메타데이터 설계와 거의 동일하다. 다만 시맨틱 데이터에서는 기본적으로 해당 클래스에만 귀속되는 내용을 속성으로 선정하며, 다른 클래스에 포함될 수 있는 항목은 관계(relation, object property)로 정의한다. 또한 기존 사전 편찬에서 종이 지면의 한계상 소략되었던 내용도 비교적 덜 중요하다고 판단한 정보도 모두 편찬이 가능하며, 오히려 최대한 많은 정보를 편찬하는 것을 권장한다. 예를 들어서, 인물에 대한 속성(attribute, data property, metadata)은 이칭과 생몰년(生歿年)이 대표적이다. 생몰년은 현대에 사용하는

54) 이에 대한 상세 내용은 다음 문헌을 참조하라. 김현, 앞의 책, 2016, 163~183쪽.

서력(그레고리력)으로 기술할 수도 있지만, 필요에 따라서는 태음태양력, 시헌력(時憲曆), 수시력(授時曆) 등의 태음태양력이나 연호(年號)와 같은 방식으로도 기술할 수 있다. 한국역사인물에 대한 이칭 정보는 단순하게는 dcterms에서 정의한 alternative를 사용할 수 있다. 하지만 실제 당시 상황과 기존 인문학 연구 성과를 반영하기 위해서는 호(號), 자(字), 화명(化名) 등으로 세분화할 수 있다. 또한 더욱 더 정밀한 데이터 구축을 위해서, 호(號)를 더욱 세분화하여 봉호(封號), 작호(綽號), 당호(堂號), 아호(雅號), 재호(斋號), 능호(陵號), 묘호(廟號), 시호(諡號), 휘호(徽號)로 세분화 할 수 있다.

〈표 7〉 인명 이칭 정보에 관한 Data Property 설계 예시

Data Property Name	명칭	Domain	Range	subPropertyOf	비고
dct:alternative	대체명	owl:Thing	rdfs:Literal		모든 이칭 정보
hasAliasName	자(字)	Perosn	rdfs:Literal	dct:alternative	
hasNickname	호(號)	Perosn	rdfs:Literal	dct:alternative	
hasEpithetByEmperor	존호(尊號)	Perosn	rdfs:Literal	hasNickname	
hasTitleOfHonor	휘호徽號	Perosn	rdfs:Literal	hasNickname	
hasPosthumousEpithet	시호諡號	Perosn	rdfs:Literal	hasNickname	
hasPosthumousTitleOfEmperor	묘호廟號	Perosn	rdfs:Literal	hasNickname	
hasTombName	능호陵號	Perosn	rdfs:Literal	hasNickname	
hasTitleByEmperor	봉호(封號)	Perosn	rdfs:Literal	hasNickname	
hasHouseName	재호斋號	Perosn	rdfs:Literal	hasNickname	= 斋号, 斋名, 室名, 庵号
hasAncestralShrine	당호堂號	Perosn	rdfs:Literal	hasNickname	
hasLiteraryName	아호雅號	Perosn	rdfs:Literal	hasNickname	
hasSobriquet	작호(綽號)	Perosn	rdfs:Literal	hasNickname	=诨号, 绰名, 浑号, 别号, 混号, 诨名, 外号, 混名, 昵称
hasAssumedName	가명(化名)	Perosn	rdfs:Literal	dct:alternative	특정 목적으로 사용한 명칭
hasAdultName	관명(冠名)	Perosn	rdfs:Literal	dct:alternative	
hasChildhoodName	아명(兒名)	Perosn	rdfs:Literal	dct:alternative	
hasPedigreeName	보명(譜名)	Perosn	rdfs:Literal	dct:alternative	
hasStudentName	훈명(訓名)	Perosn	rdfs:Literal	dct:alternative	= 學名
hasPenName	필명(筆名)	Perosn	rdfs:Literal	hasAssumedName	

Data Property Name	명칭	Domain	Range	subPropertyOf	비고
hasBaptismalName	세례명(洗禮名)	Perosn	rdfs:Literal	hasAssumedName	
hasBuddhistName	법명(法名)	Perosn	rdfs:Literal	hasAssumedName	= 戒名, 法讳
hasDharmaNickname	법호(法号)	Perosn	rdfs:Literal	hasAssumedName	
hasTaoistMonasticName	도명(道号)	Perosn	rdfs:Literal	hasAssumedName	
hasGisaengName	기명(妓名)	Perosn	rdfs:Literal	hasAssumedName	

마지막으로 시맨틱 데이터의 꽃이라고 볼 수 있는 클래스와 클래스를 연결하는 관계(relation, object property)는 인물-인물, 인물-공간, 인물-사건으로 구분할 수 있다. 인물-인물은 다시 혈연관계, 혼인관계, 사회관계로 구분할 수 있으며, 사회관계는 다시 정치, 교육, 군사, 의료, 예술 등으로 세분화할 수 있다. 인물-공간은 생몰지(生歿地), 부임지(赴任地) 등의 정보를 연결하기 위해서 사용할 수 있으며, 인물-사건은 인물과 역사적 사건의 연결로 "참여하다", "주도하다" 등의 정보를 연결하기 위해서 사용할 수 있다.

〈표 8〉 인명 정보에 관한 Object Property 설계 예시

Object Property Name	명칭	Domain	Range	subPropertyOf	inverseOf	비고
isLineageKinOf	친족	Person	Person			
hasFather	아버지	Person	Person	isLineageKinOf	isFatherOf	
isFatherOf	아버지(피동)	Person	Person	isLineageKinOf	hasFather	
hasMother	어머니	Person	Person	isLineageKinOf	isMotherOf	
isMotherOf	어머니(피동)	Person	Person	isLineageKinOf	hasMother	
hasSon	아들	Person	Person	isLineageKinOf	isSonOf	
hasAdoptedHeir	양자	Person	Person	hasSon	isAdoptedHeirOf	
hasDaughter	딸	Person	Person	isLineageKinOf	isDaughterOf	
isAffinalKinOf	인척	Person	Person			
hasHusband	남편	Person	Person	isAffinalKinOf	isHusbandOf	
hasWife	부인	Person	Person	isAffinalKinOf	isWifeOf	
hasConcubine	첩	Person	Person	hasWife	isConcubineOf	
hasFatherInLaw	장인	Person	Person	isAffinalKinOf	isFatherInLawOf	
hasMotherInLaw	장모	Person	Person	isAffinalKinOf	isMotherInLawOf	
hasSonInLaw	사위	Person	Person	isAffinalKinOf	isSonInLawOf	
hasDaughterInLaw	며느리	Person	Person	isAffinalKinOf	isDaughterInLawOf	
hasSocialRelations	사회관계	Actor	Actor			

Object Property Name	명칭	Domain	Range	subPropertyOf	inverseOf	비고
foaf:knows	지인	Actor	Actor	hasSocialRelations		
hasMaster	스승	Person	Person	hasSocialRelations	isMasterOf	
hasDisciple	제자	Person	Person	hasSocialRelations	isDiscipleOf	
isFellowOf	동문/동류	Person	Person	hasSocialRelations		
wasOrdainedBy	성직임명자	Person	Person	hasSocialRelations		계(戒)
foaf:member	구성원	Group	Person	hasSocialRelations		
participatesIn	참여	Actor	Event			
administrates	주관	Actor	Event			
performed	주도	Actor	Event			
hasNation	국적	Actor	Place			
hasBitrhPlace	출생지	Actor	Place			
isEnshrinedIn	배향	Person	Place			

이러한 관계에 대한 설계는 기존의 인간을 위한 사전 내용 서술을 기계(컴퓨터)가 이해하는 데이터로 전환하기 위한 설계이다. 관계 설계 자체는 기존의 사전 서술 내용을 추상화하여 종합하는 것이라 할 수 있으며, 이를 바탕으로 구축되는 시맨틱 데이터의 내용 자체도 해당 분야 전문가들에게는 크게 어렵지 않은 내용이라고도 할 수 있다. 다만 온톨로지 설계와 이를 바탕으로 한 시맨틱 데이터 구축을 통해서, 개개인의 연구자로서는 접근하기 힘들었던 방대한 데이터에 접근할 수 있게 된다. 예를 들어서, 전통적인 방법으로 역사인물에 대한 모든 혈연, 혼인, 사회 관계에 대한 탐색을 위해서는 사료, 논문, 사전 등의 자료와 그 자료를 탐색하는 시간을 필요로 하고, 설령 충분한 시간이 투자되더라도 다양한 정보를 누락하는 경우가 많다. 하지만 상호 공유되는 시맨틱 데이터에서는 역사 인물에 대한 탐색이 한 줄의 명령어로 처리 가능하다.[55]

사실 한국에는 이미 다양한 역사인물 데이터가 존재한다. 하지만 시맨틱 데이터의 형태로 이루어져 있지 않기에 각각의 역사인물데이터는 타 기관 간의 연결은 물론이고, 동일 기관 내의 데이터 간에도 연결이 되어 있지 않다.[56] 만약 입사기록 중심의 한국역대인물종합정보시스템 유관 데이터,[57] 조선왕조실록 기록에 토대

55) 구체적인 실례로, 인물 A의 제자의 친구를 찾거나, 인물 B의 장인의 동문을 탐색할 수 있다. 또한 C 문헌의 작가와 그 서발문을 작성한 사람들 사이의 혈연 관계 혹은 사회적 관계도 살펴 볼 수 있다.

56) 물론 시맨틱 데이터 형식이 아니더라도 API(Application Programming Interface) 등의 기술을 통해서 서로 다른 데이터를 상호 연계하는 것이 불가능한 것은 아니다. 하지만 시맨틱 데이터 형식으로 LOD 서비스를 하는 것에 대비하여 연결에 다양한 제약이 존재한다. 특히 데이터 형태가 아닌 다른 방식의 데이터 제공은 저작권 보장에 취약한 면이 분명 존재한다. 시맨틱 데이터는 지속적인 데이터 갱신과 유지, 그리고 다른 데이터와의 연계를 통해서 저작권이 보장된다.

57) 한국학중앙연구원, 「한국역대인물종합정보시스템」, 2005, http://people.aks.ac.kr/ (검색일 : 2021.02.17.).; 공공데이터포털, 한국학중앙연구원, 「한국학중앙연구원_한국역대인명정보」, 2020, https://www.data.go.kr/data/15052748/fileData.do (검색일 : 2021.02.17.).; 공공데이터포털, 한국학중앙연구원, 「한국학중앙연구원_조선조 관인 정보」, 2020, https://www.data.go.kr/data/15052754/fileData.do (검색일 : 2021.02.17.).; 공공데이터포털, 한국학중앙연구원, 「한국학중앙연구원_

를 둔 조선왕조실록 부가 인물 자료 유관 데이터,[58] 문헌 창작자에 기반을 둔 한국문집총간 유관 데이터,[59] 족보자료를 바탕을 둔 한국족보자료시스템[60] 등의 역사인물 정보들이 상호 연결이 된다면 그 효과는 자명하다. 따라서, 각각의 기관이 시맨틱 데이터를 구축하고, 이를 상호 연결할 필요가 있다. 그런데 상호 연결을 하기 위해서는 서로 다른 기관의 인물들에 대한 동일인 식별 작업뿐만이 아니라, 다양한 관점에서의 역사인물 정보를 유의미하게 포함하기 위해서 온톨로지 설계부터 해당 분야 전문가의 직접적인 참여가 요구된다.

Ⅳ. 결 론

본 논문은 인공지능 시대 사전의 미래가 인간을 위해서 컴퓨터가 이해할 수 있는 방식으로 여러 가지 사항을 모아 일정한 순서로 배열하고 그 각각에 해설을 붙인 내용인 '기계가독형 데이터'에 있다고 판단했다. 이에, 사전학의 데이터학으로의 전환 가능성을 타진하기 위해서, 인공지능 통합 플랫폼인 AIHUB에 공개된 레이블드 데이터, 병렬코퍼스(말뭉치) 데이터, 스퀴드 데이터의 구조와 내용의 장단점을 살펴보았다. 비록 데이터의 질보다는 양을 추구하며, 데이터 설계가 복잡하지는 않지만, 데이터 설계와 전문 도메인 영역의 데이터 구축에서는 충분히 기존 사전학의 연구 성과를 계승할 수 있다고 판단했다. 그리고 딥러닝을 위한 데이터의 한계를 타파하기 위한 또 다른 인공지능 데이터인 시맨틱 데이터의 구조와 내용에 대해서 탐색했고, 기존 사전학의 연구 성과가 온톨로지 설계와 시맨틱 데이터 구축에서 충분히 활용 될 수 있다고 판단했다.

이를 토대로 실제적인 사례 연구를 위해서, 한국역사인물사전을 주제로 딥러닝을 위한 데이터와 온톨로지 설계를 통한 시맨틱 데이터를 각각 실험적으로 구축했다. 우선 딥러닝의 객체명식별(NER) 방법을 활용하여, 『조선왕조실록』 데이터를 토대로 하는 객체명식별(NER) 학습 데이터를 구축하고 실험했다. 비록 『승정원일기』에 대한 객체명식별 실험 결과는 좋았지만, 딥러닝을 통한 객체명식별(NER)의 객체 식별이 객체의 속성

조선조문과급제자」, 2020, https://www.data.go.kr/data/15052752/fileData.do (검색일 : 2021.02.17.).; 공공데이터포털, 한국학중앙연구원, 「한국학중앙연구원_조선조사마시급제자」, 2020, https://www.data.go.kr/data/15052749/fileData.do (검색일 : 2021.02.17.).; 공공데이터포털, 한국학중앙연구원, 「한국학중앙연구원_조선조 무과급제자 장보」, 2020, https://www.data.go.kr/data/15052755/fileData.do (검색일 : 2021.02.17.).; 공공데이터포털, 한국학중앙연구원, 「한국학중앙연구원_조선조잡과급제자」, 2020, https://www.data.go.kr/data/15052750/fileData.do (검색일 : 2021.02.17.).; 공공데이터포털, 한국학중앙연구원, 「한국학중앙연구원_조선조 음관정보」, 2020, https://www.data.go.kr/data/15052753/fileData.do (검색일 : 2021.02.17.).

58) 국사편찬위원회, 「조선왕조실록 부가 인물 자료」, 2006, http://sillok.history.go.kr/manInfo/nameList.do (검색일 : 2021.02.17.).; 공공데이터포털, 국사편찬위원회, 「교육부 국사편찬위원회_조선왕조실록 정보_부가정보 인물 데이터」, https://www.data.go.kr/data/15053645/fileData.do (검색일 : 2021.02.17.).; 공공데이터포털, 국사편찬위원회, 「교육부 국사편찬위원회_조선왕조실록 정보_실록원문_20200903」, https://www.data.go.kr/data/15053647/fileData.do (검색일 : 2021.02.17.).; 공공데이터포털, 국사편찬위원회, 「교육부 국사편찬위원회_조선왕조실록 정보_고순종실록 원문」, https://www.data.go.kr/data/15053646/fileData.do (검색일 : 2021.02.17.).

59) 한국고전번역원, 「한국고전종합DB」, 2001, https://db.itkc.or.kr/ (검색일 : 2021.02.17.).; 공공데이터포털, 한국고전번역원, 「한국문집총간」, 2020, https://www.data.go.kr/data/3074298/fileData.do (검색일 : 2021.02.17.).; 공공데이터포털, 한국고전번역원, 「고전원문」, 2020, https://www.data.go.kr/data/15022432/fileData.do (검색일 : 2021.02.17.).

60) 성균관대학교, 「한국족보자료시스템」, 2006, http://jokbo.skku.edu/ (검색일 : 2021.02.17.).

만을 식별할 뿐, 객체의 고유성을 인지하지 못한다는 문제를 확인했다. 이에 객체의 시간, 공간, 인물 등과의 관계를 통한 다양한 고유성을 내재하고 있는 지식으로서의 시맨틱 데이터를 구축하기 위한 실험적인 온톨로지 설계를 통해서 시맨틱 데이터의 가능성을 살펴보았다.

결론적으로 전통적인 사전학에서 기계가독형 데이터에 대한 개념과 이해가 더해질 수 있다면 딥러닝을 위한 데이터와 온톨로지를 통한 시맨틱데이터 모두에서 충분히 경쟁력이 있다고 생각한다. 다만, 이를 위해서는 전통적인 사전학의 틀을 벗어나서 인공지능 시대, 데이터에 대한 이해 능력을 의미하는 데이터 리터러시(Data Literacy) 능력의 향상이 필요하다.

다만, 본 연구는 인공지능 시대의 사전학의 가능성 탐색을 우선시 하였기에, 한국역사인물 사전을 주제로 수행하는 객체명식별(NER)과 온톨로지 설계에 대한 초안을 제시한 수준에 불과하다. 따라서, 추후『조선왕조실록』,『승정원일기』,『한국문집총간』 등의 다수의 빅데이터를 토대로 최적화된 객체명식별(NER) 연구를 진행할 예정이다. 또한 한국학중앙연구원, 국사편찬위원회, 한국고전번역원, 성균관대학교 등에 산재해 있는 역사인물데이터를 토대로 온톨로지 설계, 시맨틱 데이터 구축, LOD(Linked Open Data) 시스템 구축 및 연계를 진행할 예정이다.

〈참고문헌〉

1. 단행본

김현 · 임영상 · 김바로,『디지털 인문학 입문』, HUEBOOKS : 한국외국어대학교 지식출판원, 2016.

2. 논문

김바로,「인공지능 시대를 위한 인문데이터의 역사와 과제」,『인공지능인문학연구』3, 2019.

김종환,「2016 네이버 어학사전의 현황과 미래」,『한국사전학회 학술대회 발표논문집』, 2016: 8, 2016.

도원영 · 차준경 · 신상현 · 최호철,「국내 출간 사전의 동향 분석: 연도별 주제별 통계 분석을 중심으로」,『한민족문화연구』70, 2020.

이유미 · 김바로 · 박지영,「한국어 감정 디지털 온톨로지 구축에 관한 연구」,『한국어 의미학』68, 2020.

Corrales, David Camilo, Agapito Ledezma, and Juan Carlos Corrales, "A conceptual framework for data quality in knowledge discovery tasks (FDQ-KDT): A Proposal." *Journal of Computers* 10: 6, 2015.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P., "Gradient-based learning applied to document recognition", *Proceedings of the IEEE* 86: 11, 1998.

Maas, A., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C., "Learning word vectors for sentiment analysis", *In Proceedings of the 49th annual meeting of the association for*

computational linguistics: Human language technologies, 2011.

3. 법률, 신문, 공보물

구본권, 「'인공지능 세번째 겨울, 피할 길은 "설명 가능성"에 있다'」, 『한겨레』, 2019년 04월 08일자, http://www.hani.co.kr/arti/science/future/889159.html (검색일 : 2021.02.17.).

국가법령정보센터, 「공공데이터의 제공 및 이용 활성화에 관한 법률」, http://www.law.go.kr/법령/공공데이터의제공및이용활성화에관한법률/(11956,20130730) (검색일 : 2021.02.17.).

기획재정부, 「한국판 뉴딜」, http://www.moef.go.kr/mp/nd/newDeal.do (검색일 : 2021.02.17.).

문화체육관광부 정책위키, 「데이터 3법」, 2020, http://www.korea.kr/special/policyCurationView.do?newsId=148867915 (검색일 : 2021.02.17.).

성수영, 「21세기판 인형 눈붙이기……한국판 뉴딜 일자리 체험해보니」, 『한경닷컴』, 2020년 07월 25일자, https://www.hankyung.com/economy/article/202007245362i (검색일 : 2021.02.17.).

이재구, 「MIT, '딥러닝, 컴퓨팅 파워 한계 접근'……'AI겨울' 해결책은?」, 『AI타임스』, 2020년 07월 17일자, http://www.aitimes.com/news/articleView.html?idxno=130732 (검색일 : 2021.02.17.).

최광민, 「솔트룩스, 학습과 상식 추론 결합한 '앙상블 AI', '설명가능한 AI' 상용화로 '3세대 AI'로 진화」, 2020년 04월 13일자, 『인공지능신문』, http://www.aitimes.kr/news/articleView.html?idxno=16034 (검색일 : 2021.02.17.).

NIA 한국정보화진흥원, AIHUB, 「인공지능(AI) 학습용 데이터 구축 2차」, 2020, http://aihub.or.kr/node/6227 (검색일 : 2021.02.17.).

4. 웹서비스

국립국어원, 「우리말샘」, 2016, https://opendict.korean.go.kr/ (검색일 : 2021.02.17.).

국립국어원, 『표준국어대사전』, 2008, https://stdict.korean.go.kr/ (검색일 : 2021.02.17.).

국사편찬위원회, 『승정원일기』, 2015, http://sjw.history.go.kr/ (검색일 : 2021.02.17.).

국사편찬위원회, 『조선왕조실록』, 2006, http://sillok.history.go.kr/ (검색일 : 2021.02.17.).

국사편찬위원회, 「한국사 LOD」, 2014, http://lod.koreanhistory.or.kr/ (검색일 : 2021.02.17.).

서울특별시, 「서울 열린데이터 광장 LOD 라이브 서비스」, http://lod.seoul.go.kr/ (검색일 : 2021.02.17.).

성균관대학교, 「한국족보자료시스템」, 2006, http://jokbo.skku.edu/ (검색일 : 2021.02.17.).

위키미디어 재단, 『위키백과』, 2001, https://ko.wikipedia.org/ (검색일 : 2021.02.17.).

한국고전번역원, 「한국고전종합DB」, 2001, https://db.itkc.or.kr/ (검색일 : 2021.02.17.).

한국학중앙연구원, 「한국역대인물종합정보시스템」, 2005, http://people.aks.ac.kr/ (검색일 : 2021.02.17.).

DOOSAN, 『doopedia』, 2000, https://www.doopedia.co.kr/ (검색일 : 2021.02.17.).

Google Research, 「bert」, 2018, https://github.com/google-research/bert (검색일 : 2021.02.17.).

NAVER Corp., 「NAVER 사전」, 1999, https://dict.naver.com/ (검색일 : 2021.02.17.).
NIA 한국정보화진흥원, 「AI HUB」, 2019, http://www.aihub.or.kr/ (검색일 : 2021.02.17.).

5. DATA

공공데이터포털, 국사편찬위원회, 「교육부 국사편찬위원회_조선왕조실록 정보_고순종실록 원문」, https://www.data.go.kr/data/15053646/fileData.do (검색일 : 2021.02.17.).
공공데이터포털, 국사편찬위원회, 「교육부 국사편찬위원회_조선왕조실록 정보_부가정보 인물 데이터」, https://www.data.go.kr/data/15053645/fileData.do (검색일 : 2021.02.17.).
공공데이터포털, 국사편찬위원회, 「교육부 국사편찬위원회_조선왕조실록 정보_실록원문_20200903」, https://www.data.go.kr/data/15053647/fileData.do (검색일 : 2021.02.17.).
공공데이터포털, 한국고전번역원, 「고전원문」, 2020, https://www.data.go.kr/data/15022432/fileData.do (검색일 : 2021.02.17.).
공공데이터포털, 한국고전번역원, 「한국문집총간」, 2020, https://www.data.go.kr/data/3074298/fileData.do (검색일 : 2021.02.17.).
공공데이터포털, 한국학중앙연구원, 「한국학중앙연구원_조선조문과급제자」, 2020, https://www.data.go.kr/data/15052752/fileData.do (검색일 : 2021.02.17.).
공공데이터포털, 한국학중앙연구원, 「한국학중앙연구원_조선조사마시급제자」, 2020, https://www.data.go.kr/data/15052749/fileData.do (검색일 : 2021.02.17.).
공공데이터포털, 한국학중앙연구원, 「한국학중앙연구원_조선조잡과급제자」, 2020, https://www.data.go.kr/data/15052750/fileData.do (검색일 : 2021.02.17.).
공공데이터포털, 한국학중앙연구원, 「한국학중앙연구원_조선조 관인 정보」, 2020, https://www.data.go.kr/data/15052754/fileData.do (검색일 : 2021.02.17.).
공공데이터포털, 한국학중앙연구원, 「한국학중앙연구원_조선조 무과급제자 장보」, 2020, https://www.data.go.kr/data/15052755/fileData.do (검색일 : 2021.02.17.).
공공데이터포털, 한국학중앙연구원, 「한국학중앙연구원_조선조 음관정보」, 2020, https://www.data.go.kr/data/15052753/fileData.do (검색일 : 2021.02.17.).
공공데이터포털, 한국학중앙연구원, 「한국학중앙연구원_한국역대인명정보」, 2020, https://www.data.go.kr/data/15052748/fileData.do (검색일 : 2021.02.17.).
와이즈넛, 코아인포메이션, NIA 한국정보화진흥원, 「법률」, 2017, http://aihub.or.kr/aidata/029 (검색일 : 2021.02.17.).
㈜마인즈랩, NIA 한국정보화진흥원, 「기계독해」, 2019, http://aihub.or.kr/aidata/86 (검색일 : 2021.02.17.).
㈜마인즈랩, NIA 한국정보화진흥원, 「일반상식」, 2017, http://aihub.or.kr/aidata/84

(검색일 : 2021.02.17.).
㈜솔트룩스파트너스, ㈜에버트란, ㈜플리토, NIA 한국정보화진흥원, 2019, 「한국어 – 영어 번역(병렬) 말뭉치 AI데이터」, http://aihub.or.kr/aidata/87 (검색일 : 2021.02.17.).
㈜아크릴, KETI 지능정보 플래스십 R&D, 「한국어 감정 정보가 포함된 단발성 대화 데이터셋」, 2017, http://aihub.or.kr/keti_data_board/language_intelligence (검색일 : 2021.02.17.).

7. 논문 데이터 및 코드

논문 사용 데이터 및 코드 종합 정리:
https://doi.org/10.17605/OSF.IO/5FBHP

『조선왕조실록』의 데이터를 토대로 한 전체 NER 학습 데이터:
(1) https://osf.io/h2npb/
(2) https://drive.google.com/file/d/1MU-u3Uy055lZ8EBW-zYTpSJtiZTmsMrd/

『조선왕조실록』 데이터를 토대로 한 NER 학습 코드:
(1) https://osf.io/7bvge/
(2) https://drive.google.com/file/d/1zGx_IuMaluvXAEy8hKL9luczWsCvFk88/

* 이 논문은 2021년 4월 22일에 투고되어,
2021년 6월 11일에 심사위원을 확정하고,
2021년 6월 30일까지 심사하고,
2021년 7월 7일에 게재가 확정되었음.

Abstract

Dictionaries and Data in the Artificial Intelligence Era
– Exploring for the Korean Historical Dictionary (data) –

Kim, Ba-ro*

This paper concludes that the future of the artificial intelligence era dictionary is in the 'machine readable data', which is a collection of various things in a way that computers can understand for human beings and arranged in a certain order and commentary on each of them. In order to explore the possibility of transition to the data of dictionary studies, researcher examined the advantages and disadvantages of labeled data, parallel corpus data, and squad data released on AIHUB, an artificial intelligence integration platform. In data design and data construction in the specialized domain area, it was decided that it could inherit the research results of existing dictionary studies. And researcher searched the structure and contents of semantic data, another artificial intelligence data to overcome the limitations of data for deep learning, and decided that the results of existing dictionaries could be fully utilized in ontology design and semantic data construction.

Based on this, for practical case study, data for deep learning and semantic data through ontology design were experimentally constructed on the theme of Korean historical character dictionary. First, researcher constructed and experimented with the learning data of object name identification (NER) based on the annals data of the Joseon Dynasty using the object name identification (NER) method of deep learning. The problem that object identification of object name identification (NER) through deep learning only identifies the attribute of the object and does not recognize the uniqueness of the object is confirmed. Therefore, the possibility of semantic data was examined through experimental ontology design to build semantic data as knowledge that has various uniqueness through the relationship with the object's time, space, and characters. In conclusion, if the concept and understanding of machine-readable data can be added in traditional dictionary studies, it is considered to be competitive enough in both deep learning data and semantic data through ontology.

[Keywords] Digital Humanities, Dictionary, Data, Artificial Intelligence, Historical Figures

* Research Professor, Center for Digital Humanities, The Academy of Korean Studies