한국학 제48권 제1호 통권 178호 77~106쪽
doi: 10.25024/ksq.48.1.202503.77

연구논문

근대 한국학 텍스트의 개체명 주석 연구

1920~1930년대 신문 기사를 중심으로

강범일
연세대학교 언어정보연구원 연구교수, 전산언어학 전공
kangbeomil@gmail.com

Ⅰ. 머리말

이 연구에서는 연세대학교 근대한국연구소에서 '근대 한국학의 지적 기반 성찰과 21세기 한국학의 전망'이라는 어젠다 실현을 위해 2021년부터 2024년까지 수행한 근대 한국학 텍스트의 개체명 분석 과정을 정리하고 방법론적 쟁점을 논의한다.[1] 구체적으로 데이터 수집, 선별, 주석 작업의 각 단계를 상세히 소개하고, 그 과정에서 직면했던 문제점들과 그에 대응한 경험을 공유할 것이다. 이를 통해 관련 연구를 계획 중인 후속 연구 기관 및 연구자들에게 실질적인 제언과 고려 사항을 제시함으로써, 역사 텍스트(historical text)를 기반으로 한 한국학 연구의 질적 향상 및 지속 가능한 발전을 위한 실천적 토대를 마련하고자 한다.

개체명(named entity)은 텍스트 내에서 특정한 개체를 지칭하는 단위로, 인물, 장소, 기관, 사건 등의 정보를 포함한다. 이러한 개체명 분석 관련 연구들은 대체로 개체명을 기계적으로 탐지하고 분류하는, 자동화된 개체명 인식(named entity recognition) 시스템 개발 및 성능 향상에 초점을 두고 있다. 이는 개체명 인식이 형태소 분석과 더불어 자연어 처리의 핵심적인 전처리 단계 중 하나로 간주되고 있기 때문이다. 그러나 이 과업에서 수행한 개체명 주석은 개체명 인식 시스템 개발이 아닌, 역사 텍스트의 내용 분석을 위한 개체명 데이터베이스 구축을 목표로 했으므로, 이 연구는 자동화된 주석 방

※ 이 논문은 2017년 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2017S1A6A3A01079581).

1 근대한국학연구소에서 수행한 개체명 분석은 그간 학계에 '근대 한국학 지식 DB 구축'이라는 명칭으로 소개되어 왔다. 그러나 'DB'라는 용어는 실제로는 개체명 주석 결과물을 가리키는 것으로, 이 작업의 본질인 '개체명 주석'을 명확히 드러내지 못한다는 한계가 있어 이 연구에서는 '개체명 분석'이라는 용어를 채택했다. '개체명'은 디지털 인문학, 자연어 처리, 언어학 등 다양한 학문 분야에서 통용되는 보편적인 용어라는 점 또한 고려했다.

식이 아니라 전문가의 면밀한 검토를 통해 개체명을 수동으로 식별하고 분류한 과정을 중점적으로 논의할 것이다.

개체명은 디지털 인문학 연구 중 특히 역사 텍스트를 기반으로 하는 분야에서 중요한 역할을 한다. 역사 텍스트에서 개체명은 역사적 사실의 기본 구성 요소로서, 특정 사건이나 현상의 주체와 대상, 시공간적 배경을 명확히 하는 역할을 하며 이들 간의 관계는 역사적 사실의 구조를 형성한다. 따라서 이러한 개체명은 단순한 명칭을 넘어 당대의 사회적, 문화적, 정치적 맥락을 함축하고 있는 의미 단위라고 할 수 있을 것이다. 특히, 방대한 양의 정보를 담고 있는 역사 텍스트에서 이러한 개체명은 더욱 두드러진 가치를 지닌다. 텍스트에 존재하는 수많은 개체명들은 그 자체로 풍부한 정보 자원이 되며, 이를 효과적으로 활용할 경우 다양한 방식의 탐색과 정보 검색이 가능해진다. 예를 들어, 특정 인물과 관련된 사건들을 시간 순서대로 추적하거나, 특정 지역과 관련된 사회·문화적 현상을 분석하는 등의 연구가 가능해진다. 이와 같이 개체명이 대규모 역사 텍스트의 복잡한 정보 구조를 요약하고 이를 기반으로 다양한 탐색과 정보 검색을 가능하게 한다는 점에서, 이를 식별하고 분류하는 작업은 역사 텍스트의 활용 가능성을 확장하는 과정이라고 할 수 있을 것이다.

Ⅱ. 개체명 주석 과정

한국학 텍스트의 개체명 주석 과정은 크게 한국학을 주제로 하는 텍스트를 선별하는 단계, 선별된 텍스트에서 개체명을 식별하고 분류하는 주석(annotation) 단계로 구분해 볼 수 있다.

1. 한국학 텍스트의 선별

개체명 주석을 위한 첫 번째 단계는 한국학을 주제로 하는 텍스트를 선별하는 것이다. 이를 위해 선별의 기반이 되는 자료로서 1920년부터 1940년까지 발행된《조선일보》와《동아일보》기사를 수집했다. 1920년대 자료의 선별은 전적으로 수작업 검토를 통해 이루어졌다. 말 그대로 해당 시기에 발행된 전수 기사를 일일이 검토하여 한국학 기사 여부를 판별한 것이다.[2] 이때 고려된 한국학 관련 기사의 판별 기준은 앞선 연구에서 다음과 같이 제시된 바 있다.[3]

> '조선 역사 문화' 관련 기사라 함은 조선(한국)의 역사, 어문, 사상(철학), 민속, 고적, 음악, 미술 등을 다룬 기사를 의미하는 것으로 그에 관한 사건, 운동 등에 관한 기사도 포함한다. 즉, 소재와 대상으로서 '조선 역사 문화'에 관련된 기사를 선별 한 것이다. 따라서 당대 국내외의 사건이나 정세를 전하거나 그에 관해 논평하는 기사는 대부분 포함되지 않으며, 일반적인 의미의 '역사', '언어', '문학', '철학', '민속' 등을 다루거나 타국, 타민족의 역사 문화에 관련된 기사 또한 포함되지 않는다. 그리고 '조선 역사 문화'를 소재로 했더라도 시(詩), 소설, 희곡 등 문예 작품은 제외했으며, '조선 역사' 관련 기사는 1920년대 발행된 신문 기사가 분석 대상이므로 임의적으로 1910년 강제 병합을 기준으로 삼아 그 이전 시기를 다루는 기사로 한정했다.

---

2 작업자들이 가장 선호한 방식은 엑셀 파일의 각 시트에 연도별 발행 기사 전체가 저장된 형식의 자료를 검토하여 한국학 기사 여부를 각 행에 코딩하는 방식이었다.

3 홍정완, 「신문으로 읽는 1920년대 식민지 조선의 '조선 역사·문화': 『동아일보』, 『조선일보』의 '조선 역사·문화' 관련 텍스트 계량 분석을 중심으로」, 『동방학지』 198(2022), 1~37쪽.

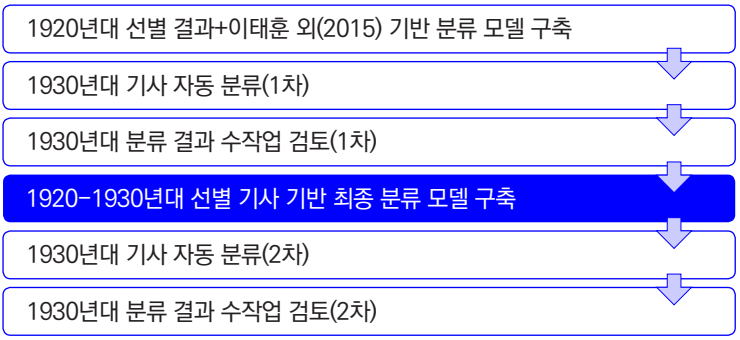

그림1-1930년대 한국학 관련 기사 선별 과정

이 과정을 통해 1920년대 한국학 관련 기사를 선별한 결과, 약 1만 2,000건의 기사가 최종적으로 추출되었다. 이는 전체 기사 수인 약 83만 건의 1.4%에 불과한 것으로, 수작업 기반의 기사 선별 과정이 상당한 시간과 노력을 요구함을 보여준다. 이에 따라 1930년대 자료에 대해서는 선별의 효율성을 높이기 위해 기계 학습 기반의 자동 분류 기법을 부분적으로 도입했다. 일차적으로 자동 분류를 통해 한국학 관련 기사 후보군을 추출한 뒤, 추출된 후보군에 대해 수작업 검토를 수행하는 단계적 접근 방식을 적용했다(그림1).

자동 분류를 시도할 수 있었던 이유는 기본적으로 1920년대 선별 작업을 통해 기계 학습을 위한 훈련 데이터가 마련되었기 때문이다. 여기에 더해 이태훈 외(2015)에서 목록화한 1930년대 《조선일보》 및 《동아일보》의 조선 역사·문화 관련 기사를 찾아 추가 훈련 데이터로 활용할 수 있었다.[4] 이러한 자료들을 기반으로 분류 모델을 생성하고 이를 이용해 1930년대 기사를 자동으로 분류했다. 이후 그 결과를 다시 수작업으로 검토하고 오류를 보정하는

4 이태훈·정용서·채관식, 『일제하 '조선 역사·문화' 관련 기사 목록 1』(서울: 선인, 2015).

표1-한국학 관련 기사의 규모 (단위: 건, 개)

구분	기사 수	문자 수	어절 수(환산)
1920년대 기사	11,789	8,652,517	2,575,154
1930년대 기사	27,649	20,886,921	6,216,346
합계	39,438	29,539,438	8,791,500

과정을 거쳐 모델의 성능을 개선했다.[5] 개선된 모델을 다시 1930년대 기사에 적용하고 재차 수작업 검토를 수행함으로써, 최종적인 1930년대 한국학 관련 기사 목록을 확보할 수 있었다. 〈표1〉은 이와 같은 과정을 거쳐 구축한, 개체명 주석의 대상이 되는 한국학 관련 기사의 규모를 보여 준다.

1930년대 한국학 관련 기사의 수는 2만 7,649건으로, 1920년대의 1만 1,789건에 비해 2배 이상 증가했음을 알 수 있다. 이는 1920년대에 비해 1930년대의 전체 신문 기사 발행량이 증가한 것과 관련이 있다. 자료의 규모를 더 객관적으로 가늠해 보기 위해 한국어 텍스트의 기본적인 계량 단위인 어절 수를 산출했다.[6] 그러나 1920~1930년대 신문 기사의 경우 현대의 띄어쓰기 규범과는 상당한 차이가 있고, 이 과업에서는 전처리 단계에서 별도의 띄어쓰기 교정을 수행하지 않았으므로, 단순히 띄어쓰기를 기준으로는 텍스트의 실제 규모를 파악하기 어렵다.[7] 이러한 문제를 보정하기 위해 현대 신문

---

5 자동 분류 모델 구축과 관련된 자세한 내용은 강범일, 「20세기 전반기 한국어 텍스트의 자동 분류 연구: 1920-1939년 한국학 관련 신문 기사를 중심으로」, 『인문과학연구논총』 44-3(2023), 221~242쪽을 참고.

6 이때의 어절은 국어학에서 정의하는 단위가 아닌, 단순히 텍스트 내 공백으로 구분되는 물리적 단위를 의미한다. 어절 수 정보는 다른 말뭉치와의 규모 비교를 가능하게 하며, 연구의 분석 결과가 지니는 타당성과 일반화 가능성을 객관적으로 평가할 수 있는 근거를 제공한다는 측면에서 중요한 정보이다.

7 극단적인 예로, 띄어쓰기가 전혀 없는 기사는 어절 수가 1로 계산되는 문제가 발생한다.

```
▼<xml>
  ▼<header>
     <id>19300102000209207002</id>
     <date>19300102</date>
     <publisher>동아일보</publisher>
     <section>생활/문화</section>
     <genre>칼럼/논단</genre>
     <title>馬의 學的考察 (上)</title>
   </header>
   <body> 馬의 學的考察 (上) 東京 金 益 鎬 原始人類의 文化過程에잇서서가장 重大한 問題는"어떠케하면그들의 軟弱한 手足의힘을 幇助할 手段을 講究할까"함에잇섯슬것이다 그러고이 目的을 貫徹키 爲하야는 動物을 家畜化함이아니면 到底히 野蠻狀態를벗어나 文化의 領域에 得達키 不能하얏슬것이다 이 點에 關하야 馬의 養育利用이 人類에게 莫大한 影響을주니만치 戰時나 平時를 勿論하고 重大한 關係를가지고잇섯다 萬一馬를 利用치안햇든들 人類의 文化는 決코 今日의 狀態에 發展되지못하얏슬것은 明瞭한 事實인 同時에 이 動物이아니드면 旣成의 文化도 維持키어려웟든것은 古代의 歷史的記錄이 雄辯으로 證明하고잇다 그런데 馬의 發生地에 對하야는 各說이 紛紛하다 歐洲或은 北米를 原産地라고 主唱하는 學者들도잇슴으로 亞細亞一源說이 一時그 勢力을일케되엇스나 吾人은 地質時代의 馬의 祖先에 對한 問題는 姑置하고 적어도 歷史時代또는 遺物遺跡으로알수잇는 時代에잇서서는 中央亞細亞"이란"(Iran)高原地方이主되는 原産地임은"헤ㅣㄴ"(Hehn)氏의 學說인 同時에 一般學界의 定評이다뿐만아니라"콜럼버스"發見當時의 米大陸에는 馬가업섯든것도 周知의 事實이다 또한 語源의 方面으로보아도 馬가各地에 傳播되기 前에"아리안"(Aryan)人種만은먼저알앗든것갓다 이 問題에 關하야좀더 具體的으로 考察컨대 于先東方에잇서서 支那의 唐代와 六朝時代에키가놉고다리가늘은 馬를 表現한 藝術品어만히생겻다 大宛傳에 依컨대 漢武帝가 大 宛國의 汗血馬를 貪내어 貳師將軍李廣利를 시켜서 大宛國을 征服한 後로 汗血馬가만히 支那에 輸入되엇스며 藝術品에도 此種의 馬가"모ㅣ텔"이되엇다 또 記異錄을보건대 唐代天寶年間에 大宛國에서 紅▨騮紫▨騮 靑▨騮 黃▨騮 丁香▨騮 桃花▨騮이란 六匹의 汗血馬를 獻上하얏다 그리고 白樂天의 ▨雪中馬上妓란 詩中에도 花襜宜乘▨騮駒란 句가잇다 또한 唐의 召陵六驟가운대 赤馬를 什伐이라하얏다 이제"이란 地方의 馬에 對한 原語를 考察컨대 大宛國에서는 Aspah,安息國에서는Aspa,波斯에서는Asp或은Asb이라고함을보아도 ▨騮或什伐이란이름은 馬의 輸入에딸하서그 名詞까지 傳來되엇다고 推想된다 또한 昭陵六驟은할티(馬口)를 三結하얏는데 이 結束의 數에딸하 三花馬或은 五花馬라고하얏다 白樂天의 詩中에도"馬口剪三花"란 句가잇스며 宋의 郭君虛見聞誌와 唐의 韓幹의 貴戚關馬圖, 張萱의 虢國夫人乘馬出行圖에도 三花馬가잇다 이것은 支那固有의 風이아니엇고 唐代부터 流行되엇는데"이란"地方에서는 支那의 漢代頃에벌서 大流行이엇든것은"에니새이"(R.pyenisei)河上流의 壁畫와 波斯"삿산"朝의"코스러스"二世의 彫刻物等을보아도 疑心할 餘地가업다 그러고 新羅古墳에서 發堀되어 現在京城及慶州博物館에 所藏된 馬의 裝飾品인 杏葉에잇서서도 朝鮮의 三國時代,日本의 奈良朝時代,支那南北朝時代에 大流行되엇다 唐代王勃의 詩에도"杏葉裝金轡"란 句가잇다 이것도 亦是龜玆國의 壁畫와"삿산 朝"의 遺物을보아"이란"地方에서 傳來하얏다고밋는다 </body>
   <category1>스트레이트/담론/삭제/재확인</category1>
   <category2>고적/기행/문학/미술/민속/신간/어학/역사/음악/철학/총론/한의학/기타
   </category2>
</xml>
```


그림2-한국학 관련 기사의 XML 형식 구조

기사의 띄어쓰기 양상을 기준으로 어절 수를 환산하는 방법을 적용했다.[8] 〈표 1〉에 제시된 환산된 어절 수는 이러한 과정을 거쳐 산출된 결과이다. 이를 통해 개체명 주석 대상 텍스트는 약 879만 어절에 달하는 규모임을 알 수 있다.

8 구체적으로 2023년 발행된《조선일보》기사 전체를 분석하여 평균 어절 길이(3.36문자)를 산출하고, 이를 기준으로 1920~1930년대 기사의 어절 수를 환산했다.

이렇게 구축된 데이터는 〈그림2〉와 같이 XML 형식으로 구조화했다.

2. 개체명 주석

1) 주석 대상

개체명이란 텍스트 내에서 특정 대상이나 개념을 지칭하는 고유한 명칭을 의미한다. 이는 인물, 장소, 조직, 날짜와 같은 텍스트의 주요 정보를 나타내는 언어 단위로, 디지털 인문학 연구에서는 대규모 역사 텍스트의 복잡한 정보 구조를 요약하고 이를 통해 다양한 탐색과 정보 검색을 가능하게 하는 역할을 한다.

최종 선별된 한국학 관련 기사에서 식별하고 분류한 개체명은 인물, 저자, 기관, 레퍼런스, 레퍼런스 저자이다.[9] 먼저 '인물' 유형은 기사에서 언급되는 역사적 인물이나 당대의 실존 인물을, '저자' 유형은 해당 기사를 작성한 저자를 지칭한다. '기관' 유형은 정부 기관, 교육 기관, 단체 등 조직의 명칭을 포함한다. '레퍼런스' 유형은 텍스트 내에서 인용되거나 언급되는 문헌, 서적, 논문 등의 저작물을 가리키며, '레퍼런스 저자' 유형은 이러한 레퍼런스의 저자를 의미한다. 구체적인 주석 지침은 다음과 같다.[10]

---

9 개체명은 일반적으로 인물(Person), 장소(Location), 기관(Organization)의 세 가지 기본 유형(PLO)을 기반으로 정의되며 연구의 목적과 대상 텍스트의 특성에 따라 확장될 수 있다.

10 이 외에도 근대한국학연구소의 개체명 주석 작업 관련해서 다음의 자료들을 참고할 수 있다. 연세대 근대한국학연구소 인문한국플러스(HK+)사업단(편), 『디지털 인문학과 근대한국학』(서울: 소명출판, 2020); 홍정완, 「1920~30년대 식민지 조선의 종합잡지에 나타난 '조선 역사': 텍스트 계량 분석을 중심으로」, 『동방학지』 202(2023), 1~39쪽; 홍정완·정유경·심희찬, 『근대한국학 데이터베이스 자료집 1』(서울: 소명출판, 2020).

① **인물:** 특정 인물을 지시하는 모든 단어(호, 별명, 외국인 등을 모두 포함). 신화나 설화, 종교상 인물 포함. 문수보살, 보현보살, 천사들 이름도 포함. '연호'(광무, 융희, 명치, 대정, 소화 등)나 '연호+帝'(강희제, 영락제, 옹정제, 건륭제 등)는 인물명으로 태깅. 동일 기사 내에 한 인물이 다양한 칭호로 표시될 때('씨', '옹', '선생', '박사', '公', 姓 등을 모두 포함), 이를 모두 태깅하고 note에 본명을 기입. 기사의 저자가 자신을 지칭하는 표현은 이것이 역사적 행위의 주체일 때만 태깅. 본명이 나오지 않더라도 누구인지 특정할 수 있는 경우("조선을 개국한 왕" 등)에는 태깅하고, note에 본명을 기입. '기씨조선'이나 '세종3년' 같이 특정인을 지칭한다고 볼 수 없는 항목이라도 '기씨'와 '세종'을 〈인물〉로 태깅

② **저자:** 해당 기사를 작성한 저자

③ **기관/조직:** 특정 단체, 기관, 조직 태깅. 사색당파 등 파벌, 주전파·주화파, 척사파, 개화당·수구당, 훈구·사림 등 포함, 친일·친미·친중(청)·친러파, 동학당, 동학군, 만민공동회 태깅. 하위 조직을 포함하는 명칭의 경우('총독부 학무국' 등) 모두 하나로 이어서 태깅

④ **레퍼런스:** 기사 중에 등장하는 서적 및 텍스트

⑤ **레퍼런스 저자:** 레퍼런스와 더불어 저자가 언급될 경우 태깅. '氏', '金氏', '某' 등 특정할 수 없는 표기인 경우에도 모두 태깅

㉠ 서명 및 레퍼런스 저자명이 불명확한 경우에는 다음 방식에 따른다.

- 서명이나 텍스트명 없이 인용구만 있는 경우
  - 서명 혹은 텍스트명을 조사하여 '노트(Notes)'란에 기입
  - 서명 혹은 텍스트명을 확인할 수 없는 경우에는 인용 구절 일부를 〈레퍼런스〉 태깅
- '아무개운', '아무개왈' 등으로 표기된 경우
  - '아무개'를 〈레퍼런스저자〉, '아무개운' 혹은 '아무개왈'을 〈레퍼런스〉 태깅
- '아무개가 …시에서 말하길' 등, 〈레퍼런스〉 태깅의 대상이 분리되어 있는

경우

- '아무개'와 '시'를 'Add Frag.' 기능을 활용하여 '아무개시'로 〈레퍼런스〉 태깅

㉡ 신간에 관한 기사 중 잡지와 게재 글들이 소개된 경우에는 다음 방식에 따른다.

- 잡지 제목 〈레퍼런스〉 태깅
- 각론은 〈기사 선별 기준〉에 부합하는 것이 확실한 경우 〈레퍼런스〉 태깅

개체명 외에도 기사의 종류(스트레이트/담론), 잘못 선별된 기사의 삭제 또는 재검토 여부, 기사의 주제에 대한 텍스트 단위 정보를 주석했다. 〈그림2〉의 xml 태그 중 category1, category2가 이에 해당한다.

2) 주석 도구

개체명 주석 작업을 위해 오픈소스 텍스트 주석 도구인 BRAT[11]을 연구소 내부 서버에 설치해 사용했다. BRAT은 웹 기반 시스템으로, 기사 텍스트 파일(XML)과 주석 대상 태그 정보를 담은 파일을 지정된 폴더에 업로드하면 작업자들이 별도의 소프트웨어 설치 없이 웹 브라우저를 통해 어디에서나 주석 작업을 수행할 수 있다.[12] 구체적으로는 웹 브라우저를 통해 서버에 접속하여 로그인한 후, 작업 대상 텍스트를 선택하는 것으로 시작되며, 이후 텍스트 내에서 주석할 내용을 마우스로 드래그하여 선택하면, 〈그림3〉과 같이 사전에 정의된 개체명 목록이 나타나고, 작업자는 이 중 하나의 유형을 선택하여 주석한다. 〈그림4〉는 이렇게 주석한 결과를 예시한 것이다.

---

11 BRAT, https://brat.nlplab.org.

12 일반적인 PC 수준의 서버 환경에서도 20명 내외의 사용자를 대상으로 장기간 안정적인 운영이 가능했다. 이는 BRAT이 단순하고 가벼운 시스템임을 확인해 주는 것으로, 이러한 특징은 통상적인 인문학 연구 기관이 가지는 서버 운영 부담을 완화해 주는 요인이 될 수 있다.

그림3-사전 정의된 개체명 선택 메뉴

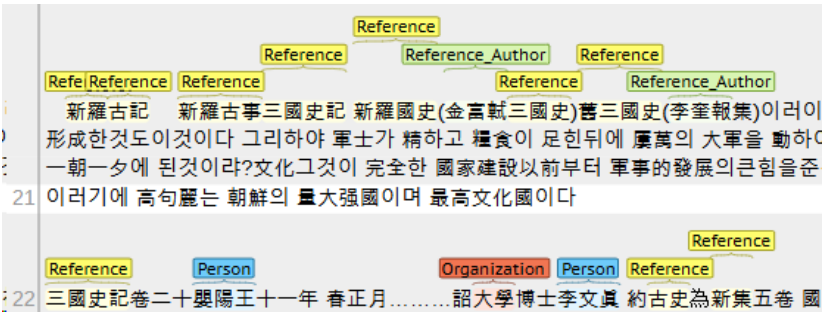

그림-4 BRAT을 이용한 개체명 주석

3) 주석 결과

웹 브라우저상에서 수행된 주석 결과는 BRAT이 설치된 서버의 지정된 디렉토리에 별도의 파일로 저장된다. 이때 원문 텍스트 파일은 수정되지 않으며, 주석 정보만을 담은 새로운 파일이 생성되는데, 이 파일은 원문 텍스트 파일과 동일한 파일명에 .ann이라는 확장자를 가진다. 주석 파일은 탭으로 구분된 값(TSV: Tab Separated Value) 형식을 따르며, 각 행은 다음과 같은

구조를 가진다.

T1	Person 156 159	韓圭卨
T2	Organization 1114 1117	一進會
T3	Person 1125 1127	伊藤
T4	Person 1234 1236	光武
T5	Person 1246 1249	長谷川
#1	AnnotatorNotes T3	이토 히로부미
#2	AnnotatorNotes T5	하세가와 요시미치

첫 번째 컬럼은 주석의 고유 식별자(ID)이며, 두 번째 컬럼은 개체명 유형(person, organization 등)과 해당 개체명이 원문 텍스트에서 출현하는 위치 정보를 나타낸다. BRAT이 Python으로 개발되었기 때문에, 이 위치 정보는 Python의 문자열 슬라이싱 규칙을 따른다. 예를 들어, T1의 '156 159'는 원문 텍스트에서 인덱스 156부터 159 이전까지의 범위를 의미한다. 마지막 컬럼은 실제 개체명 텍스트에 해당한다.

작업자는 개체명 주석 외에, 부가 정보를 작성할 수도 있는데 이러한 정보는 '#' 기호로 시작하는 별도의 행으로 기록된다. 앞의 예에서 #1은 ID가 'T3'인 개체명 '伊藤'에 대해 주석자가 '이토 히로부미'라는 추가 정보를 기록했음을 나타낸다. 이러한 주석자 노트는 동일한 대상을 가리키는 다양한 이표기에 대한 표준형을 기록하거나, 개체명에 대한 부가 설명을 기록하는 데 사용되었다.

이와 같이 구조화된 형식으로 저장된 주석 파일은 간단한 프로그래밍을 통해 쉽게 파싱할 수 있으며, 이를 통해 텍스트에 출현하는 개체명의 빈도 및 분포 등 다양한 통계적 분석이 가능해진다.

4) 반자동 주석

1930년대 신문 기사의 개체명 주석은 반자동 방식으로 수행되었다. 이는 수동 주석의 효율성을 높이기 위해 자동 주석을 선행한 후, 그 결과를 수동으로 교정하는 방식이다. 여기서 자동 주석이란 1920년대 자료의 주석 결과를 활용하여 1930년대 텍스트의 개체명을 자동으로 식별하는 과정을 의미한다.

구체적인 수행 절차는 다음과 같다. 우선 1920년대 주석 결과에서 개체명과 그 출현 빈도를 추출하여 목록화했다. 이 목록에서 오주석(false annotation) 가능성이 높은 개체명은 제외했는데, 기본적으로 출현 빈도가 낮거나 단음절인 개체명을 배제했으며, 목록을 수작업으로 검토하여 중의성을 가지는 개체명 또한 제외했다.

다음으로, 정제된 개체명 목록을 바탕으로 1930년대 텍스트에 대한 자동 주석을 수행했다. 이 과정은 단순 문자열 매칭(string matching) 방식으로 이루어졌다. 예를 들어, 개체명 목록에 '伊藤'가 포함되어 있다면, 텍스트 내에서 이 문자열이 출현하는 모든 위치에 자동으로 개체명을 주석하는 것이다. 주석된 각 개체명의 텍스트 내 위치 정보를 계산하여 BRAT의 주석 파일(.ann)에 인위적으로 저장함으로써 BRAT상에서 자동 주석 결과를 시각적으로 확인할 수 있도록 했다.

마지막으로, 생성된 자동 주석 결과를 검토하고 교정하는 수동 작업이 수행되었다. 이 과정에서 잘못 주석된 개체명을 삭제하고, 자동 주석 과정에서 식별되지 않은 개체명을 추가하는 작업이 이루어졌다.[13] 이러한 반자동 주석 방식은 완전 수동 주석에 비해 작업 효율성을 향상시킬 수 있다는 장점이 있다. 물론 자동 주석의 오류를 제거하는 데 완전 수동 주석보다 더 많은 시간

13 자동 주석은 작업 과정에서 여러 차례 수행되었다. 특정 시점까지 새롭게 축적된 개체명 정보를 검토하여 활용 가능한 것들을 선별한 후, 이를 바탕으로 미작업 텍스트에 대한 추가 자동 주석을 수행하는 방식을 반복했다.

이 소요되는 사례도 일부 있었으나, 반자동 주석에 대한 작업자들의 전반적인 만족도는 높게 나타났다.

5) 주석 결과

이러한 과정을 통해 약 4만 개의 한국학 관련 기사에서 식별된 개체명 정보는 〈표2〉와 같다. 1920년대와 1930년대 기사에서 주석된 개체명의 수는 총 10만 332개였으며, 이 중 중복을 제외한 고유 개체명은 3만 1,632개로 나타났다. 특히 1930년대 기사에서 식별된 고유 개체명이 2만 1,706개로, 1920년대의 1만 2,394개에 비해 약 1.8배 많았다. 이는 분석 대상 기사의 규모 차이에서 기인한 결과로 보인다.[14]

공개된 개체명 주석 역사 코퍼스(NE-annotated Historical Corpora)에 대해 조사한 최근의 한 연구에서는 개체명이 주석된 역사 코퍼스의 규모를 소규모(1만 개 미만), 중규모(1~3만 개), 대규모(3~10만 개), 초대규모(10만 개 이상)로 제시한 바 있다.[15] 이 기준에 따르면 이 연구의 자료는 고유 개체명 수가 3만 개를 상회하므로 대규모 코퍼스에 해당한다.

고유 개체명의 유형별 분포를 살펴보면, 인물이 49%로 가장 높은 비중을 차지했으며, 레퍼런스(23%), 기관(18%), 레퍼런스 저자(7%), 저자(3%) 순으로 나타났다. 이는 전체 주석된 개체명의 절반 가까이가 인명이었음을 보여

---

14 중복을 포함된 개체명 수는 약 4.4배의 차이를 보이는데 이는 주석 방식의 변화에서 비롯된 것으로 판단된다. 1920년대 자료는 텍스트 내에서 한 번 주석된 개체명은 이후에 출현하더라도 중복 주석을 하지 않았으나 1930년대 자료에 적용된 반자동 주석에서는 자동 주석 단계 중복을 허용해 주석하고 잘못된 주석이 아니라면 검토 단계에서도 이를 삭제하는 작업을 하지 않았다. 따라서 이 결과만을 가지고 1930년대 기사에서 동일 개체명이 더 빈번하게 재등장했다고 판단할 수는 없다.

15 M. Ehrmann, A. Hamdi, E. L. Pontes, M. Romanello, and A. Doucet, "Named entity recognition and classification in historical documents: A survey," *ACM Computing Surveys*, Vol. 56, No. 2(2024), pp. 1~47.

표2-주석된 개체명 통계 (단위: 개)

구분	개체명 수(tokens)	고유 개체명 수(types)
1920년대 기사	18,656	12,394
1930년대 기사	81,676	21,706
합계	100,332	31,632

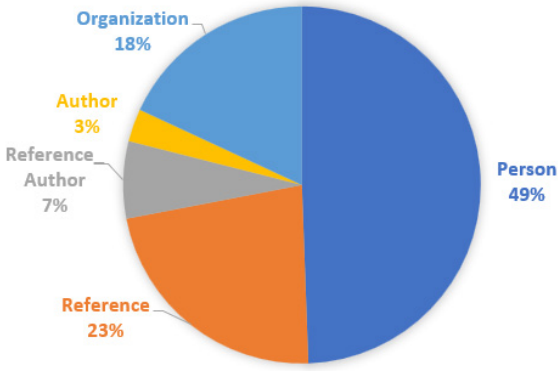

그림5-고유 개체명의 유형별 분포

준다. 또한 레퍼런스와 레퍼런스 저자를 합산하면 30%로, 인용 관련 개체명이 전체의 3분의 1을 차지하는 것으로 확인되었다. 이러한 분포는 한국학을 다룬 기사의 특성이 잘 반영된 결과로, 인물을 중심으로 한 서술과 함께 다수의 참고문헌이 활용되었음을 보여준다.

Ⅲ. 방법론적 쟁점

이 연구에서 소개하는 개체명 주석은 다수의 작업자가 방대한 양의 텍스트를 장기간에 걸쳐 처리해야 하는 작업이다. 따라서 주석 지침의 수립과 주

석 도구의 선정 등 사전 준비 단계에서의 세심한 의사결정이 작업의 일관성과 효율성뿐만 아니라 주석 결과의 활용 범위까지 결정하게 된다. 또한 한자와 한글이 혼용된 텍스트라는 자료의 특성상, 동일 개체가 한자와 한글로 다르게 표기되는 경우나, 이체자 및 이형자 등과 같은 이표기 문제를 처리하기 위한 방안도 사전에 고려되어야 한다. 이 장에서는 개체명 주석 작업에서 직면했던 주요 쟁점들을 검토하고, 이를 해결하기 위해 기울인 노력들에 대해 구체적으로 논의하고자 한다.

1. 주석 도구의 선정

수동 주석(manual annotation)은 대부분 자연어 처리 모델 개발을 위한 학습 데이터 구축을 목표로 수행된다. 최근 사전 학습된 대규모 언어 모델의 발전으로 인해, 별도의 주석 데이터 없이도 다양한 자연어 처리 과제를 수행할 수 있는 제로샷 러닝(zero-shot learning)이 가능해졌다. 그러나 이 연구의 대상인 근대 한국학 텍스트와 같은 특수한 도메인의 경우, 기존 모델이 학습한 일반적인 언어 지식만으로는 정확한 처리가 어렵기 때문에 여전히 수동 주석을 필요로 한다. 이러한 수동 주석 작업에서는 적절한 주석 도구의 선택이 핵심적인 고려 사항이 되는데, 이는 도구의 사용성과 완성도가 주석 작업의 속도와 품질에 직접적인 영향을 미치기 때문이다.

이 과업의 개체명 주석 작업에서 BRAT을 채택한 것은 오픈소스 주석 도구들 중 설치와 운영이 용이하다는 실용적 측면을 고려했기 때문이다. 또한 앞서 언급한 바와 같이 주석 결과가 단순히 텍스트 파일로 생성되어 관리되기 때문에 파일 편집을 통해서 주석 결과를 쉽게 수정할 수 있다는 것은 큰 장점에 해당한다. 그러나 BRAT을 장기간 사용하는 과정에서 여러 한계 또한 확인되었다.

BRAT의 주요 한계는 크게 세 가지 측면으로 구분해 볼 수 있다. 첫째, 2012년 이후 업데이트가 중단되어 기능이 제한적이며 유지보수가 이루어지지 않는다는 문제가 있다. 둘째, 프로젝트 관리 기능의 측면에서 작업 배분, 수정 이력 관리 등 협업을 위한 기본적인 기능이 결여되어 있다. 셋째, 보안 및 안정성 측면에서 사용자 권한 관리 체계가 미비하고, 간헐적으로 발생하는 시스템 오류 중 원인을 파악하기 어려운 경우가 있다.

최근 개발된 주석 도구들은 이러한 결여된 기능들이 기본적으로 포함되어 있을 뿐만 아니라, 주석 작업을 돕기 위한 기계학습 기반의 주석 추천 기능을 제공하기도 한다. 그러나 이러한 자동화 기능이 한국어에도 효과적으로 적용될 수 있는지에 대해서는 면밀한 검토가 필요하다. 한국어는 교착어적 특성으로 인해 주석 대상이 어절 내부의 부분 문자열인 경우가 빈번한데, 일부 최신 도구 중에는 부분 문자열에 대한 주석 자체를 지원하지 않는 경우도 있기 때문이다. 또한 최근의 주석 도구들은 단순한 개체명 태깅을 넘어서, 개체 식별자(entity ID)를 활용한 고도화된 기능들을 제공한다. 이러한 기능을 통해 동일한 개체가 서로 다른 표현으로 언급되는 경우(예: '신채호' - '단재' - '신씨')나, 반대로 동일한 표현이 서로 다른 개체를 지칭하는 경우의 중의성을 효과적으로 해소할 수 있으나 BRAT은 개체 식별자 관련 기능을 지원하지 않는다.

주석 도구를 선정하는 일은 기술적 완성도, 작업 관리 기능의 효율성, 목표 언어의 특성 등의 다양한 요소를 종합적으로 고려해야 하는 중요한 의사결정이다. 현재 사용 가능한 주석 도구는 매우 다양하며, 각 도구마다 지원하는 기능의 범위도 상이하여 이를 모두 검토하여 최적의 도구를 선정하는 것이 쉬운 일은 아니지만 장기간의 작업 환경을 결정하는 일인 만큼 신중한 검

토와 선택이 필수적이다.[16]

2. 한자의 정규화

1920~1930년대 한국학 관련 기사는 한자와 한글이 혼용된 형태를 보인다. 〈그림7〉에서 확인할 수 있듯이, 언어 변이가 극심했던 1920년대 초반을 제외하면 전체 텍스트에서 한자가 차지하는 비율은 대체로 35%에서 45% 사이를 유지하고 있다.[17] 이러한 한자가 포함된 텍스트는 전산적 처리 시 고려해야 할 특수한 문제들이 존재한다.

첫째, 모양은 같지만 음가가 다른 한자의 문제이다. 예를 들어 '樂'은 문맥에 따라 '낙(樂, U+F914)'과 '락(樂, U+F95C)'으로 서로 다른 음가를 가지며, 이는 서로 다른 유니코드 코드 포인트에 배정되어 있다. 즉, 코드 값이 달라 컴퓨터는 이 둘을 다른 글자로 인식하게 된다. 둘째, 이체자 및 이형자의 문제로, 동일한 의미와 음가를 가지지만 정자(正字)의 획수를 간략화한 약자(略字)나 고자(古字)가 혼용되는 경우이다.

이러한 요인은 주로 정보 검색 성능에 영향을 주는데, 이 연구의 개체명 주석 과정에서는 자동 주석 단계에 직접적인 영향을 미쳤다. 자동 주석이 문자열의 형태 매칭에 기반하기 때문에, 동일한 음과 뜻, 형태의 한자라도 서로 다른 코드 포인트가 할당되어 있다면 별개의 문자로 인식된다. 이러한 문제를 해결하기 위해 이 연구에서는 공개된 한자 자원을 활용하여 문자 변이를

---

16 M. Neves and J. Ševa, "An extensive review of tools for manual annotation of documents," *Briefings in Bioinformatics*, Vol. 20, No. 1(2021), pp. 146~163에서는 78개의 도구가 검토된 바 있다.

17 보다 상세한 문자 사용 양상에 대해서는 다음 논문을 참고할 수 있다. 강범일, 「한국어 통시 신문 말뭉치의 구축과 활용」, 『언어사실과관점』 54(2021), 7~33쪽.

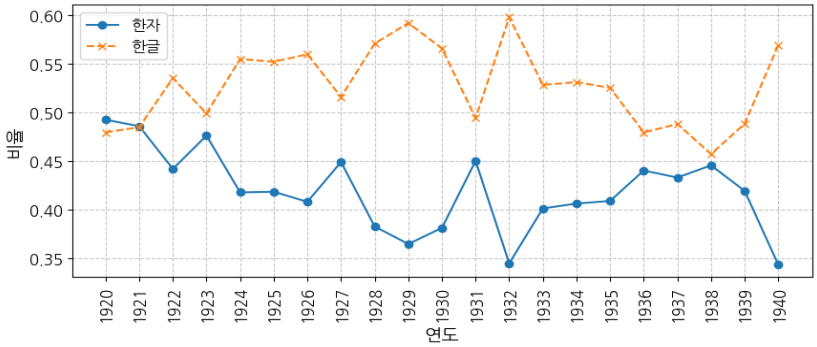

그림7-한국학 관련 기사의 연도별 한글·한자 비율

17 檄文과 阜陷城報가 第壹報됨을 따라 官土豪貴虐에 울분衆衆學旗幟아래 湖南列邑을 論全國
鼓舞하엿스며 天幸福龍운四歲와 年으로 萬義兵蜂起것으로보와 當時民衆업마 狂的이엿슴을
18 이가 各地에 起한衆은모 學胞機關인"包"容되엇나 形勢는"村村設旗相應"(大東紀年)라
東學陣營 韌性과範圍廣汎함을작할수가잇다러 한衆으로말암 展開學軍 續엽떠하엿든
意味로쓴것임)
19 A.東學軍 土戰捷
20 民衆 亂이지낼수록더욱더 大되여 年三四月頃에 全國 然함에르럿스 珍準에 率學黨
金開南(泰仁古縣內사람으로 時에 原에 起包하 州 間에두 珍準 西에서 應하얏다)泰
21 이에이르 羅監司金文鉉 까 觀할수 슴으로 珍準 殺 的으로 客二人 草商으 裝시
目的을이두지못하얏다고한다리하야드듸 업 하 負 하기로되 後統將李在 利川郡
四月七日 阜德面(今德川面)黃土峴에 出되엇다리하 學軍 襲아래 軍 數 傷者 이

그림8-한자 코드 문제로 인한 BRAT의 오류 예시

표준화하는 작업을 수행했다. 다만 원문의 형태를 보존하는 것도 중요하므로, 원문을 직접 수정하는 대신 표준화된 형태를 별도의 층(layer)으로 구성하여 원문과의 대응 관계를 유지했다.

한자 코드의 문제는 수동 주석 과정에도 영향을 미쳤다. 특히 유니코드 CJK Extension B 영역의 한자가 포함된 텍스트의 경우, 〈그림8〉과 같이 BRAT상에서 텍스트의 일부 문자들이 중첩되어 표시되는 현상이 발생하여 주석 단위를 정확히 선택할 수 없는 문제가 발생했다. 이 문제를 해결하기 위

해 텍스트를 서버에 탑재하기 전, 해당 유니코드 영역의 한자들을 미리 추출하여 대표 한자로 변환하는 전처리 작업을 수행했다. 전체 한국학 기사에서 문제가 되는 한자는 총 32종으로 확인되었으며, 이들이 출현하는 모든 기사에서 해당 한자들을 대표 한자로 일괄 변환한 후 업로드하는 방식을 적용했다.[18] 이는 해당 시기 데이터의 특성으로 인한 문제라기보다는 라틴 문자 기반의 언어를 중심으로 설계된 주석 도구와의 호환성 문제라고 할 수 있을 것이다. 따라서 특히 한자가 포함된 텍스트의 주석 작업을 계획할 때는 한자의 정규화 문제와 더불어 주석 도구와의 호환성 문제를 면밀히 검토하여 적절한 텍스트 전처리 방안을 수립해야 할 것이다.[19]

3. 이표기의 정규화

앞서 논의한 한자 코드의 변이 문제와는 별개로, 또 다른 유형의 이표기 형태가 존재한다. 이는 단순한 문자 코드의 차이가 아닌 표기 방식의 차이에서 기인하는 것으로, 특히 근대 시기에 외래어를 비롯한 단어들의 표기법이 통일되지 않은 문제, 약칭이 혼용된 문제 등에서 비롯된다. 이러한 이표기가 처리되지 않을 경우, 동일 개체의 출현 빈도가 여러 표기 형태에 분산되어 정확한 통계 분석이 어려워지는 문제가 발생한다. 이 과업에서는 이러한 문제

---

18 일부 한자의 경우 적절한 대체 문자를 찾을 수 없어, 해당 한자를 구성하는 부수의 조합으로 치환했다. 예를 들어, 纚는 '糸+彌', 晷는 '日+咎', 炤는 '火+言'과 같이 변환했다.

19 동시기 텍스트의 또 다른 특징으로 옛한글 표기의 문제를 들 수 있다. 옛한글은 과거 유니코드 PUA(Private Use Area) 영역의 코드를 사용해 인코딩되는 경우가 많았는데, 이는 시스템 간 호환성 문제를 야기할 수 있다. 이러한 문제를 해결하기 위해 현재는 유니코드 첫가끝(Hangul Jamo) 코드로의 변환이 표준적인 처리 방식으로 권장된다. 다만 이 연구의 분석 대상인《조선일보》와《동아일보》기사의 경우, 디지타이징 과정에서 옛한글이 현대 한글로 일괄 변환되었기 때문에 이 문제를 고려할 필요가 없었다.

표3-이표기 정규화 테이블의 일부

변이형	표준형	변이형	표준형
加里波地	가리발디	개르하르트·하움트만	게르하르트 하우프트만
加里波	가리발디	倭將淸正	加藤淸正
깐듸	간디	淸正	加藤淸正
칸듸	간디	京都大學	京都帝國大學
게테	괴테	경도제국대학	京都帝國大學
께데	괴테	嘉聖王朔	嘉聖王

를 해결하기 위해 총 2,005개의 이표기에 대한 정규화 테이블을 구축했는데 〈표3〉은 그 일부를 보여준다.

'加里波地/加里波/加尼巴'(가리발디), '깐듸/칸듸'(간디) 등은 동일한 외래어를 서로 다른 한자나 한글로 음차한 경우, '淸正/倭將淸正(加藤淸正)'은 동일 인물의 다양한 지칭 방식이 혼용된 경우이다. '京都帝國大學(경도제국대학)'과 같이 동일 단어의 한자 표기와 한글 표기가 공존하는 경우도 둘 중 하나의 표기로 통일할 필요가 있다.

이러한 이표기 정보 역시 기본 텍스트와 별도의 층으로 구성하여 원본의 형태를 보존하면서도 정규화된 형태에 접근할 수 있도록 했다. 정규화 테이블의 구축은 두 단계로 진행되었다. 우선 BRAT의 주석자 노트 기능[20]을 통해 작업자들이 입력한 정규화 형태 정보를 수집했고, 이후 전체 개체명 목록을 검토하여 추가적인 이표기 관계를 보완했다.

이러한 정규화 작업은 주석 작업 이전에 완성된 형태로 작업자들에게 제공되어 모두가 통일된 표준형을 사용하는 것이 이상적이나 이는 현실적으로 불가능한 일이다. 따라서 주석 작업 과정에서 정기적인 검토 회의를 통해 새

20 Ⅱ-2-3) 참고.

롭게 발견되는 이표기들을 공유하고, 표준형 선정에 대한 기준을 협의하여 작업자들 간의 일관성을 확보하는 것이 중요하다.

4. 주석의 방식

한국학 기사의 개체명 주석 작업에서는 효율성을 고려하여 동일 개체명이 한 텍스트 내에서 반복 출현할 경우 최초 출현 시에만 주석을 부착하는 방식을 채택했다. 이는 대규모 텍스트를 수작업으로 처리해야 하는 현실적 제약을 고려하는 동시에, 문헌 빈도(document frequency)만으로도 의미 있는 분석이 가능할 것이라는 판단에 기반한 결정이었다.

그러나 이러한 주석 방식은 결과적으로 데이터의 활용도를 감소시키는 요인이 되었다. 〈표2〉에서 확인된 바와 같이 식별된 개체명의 유형이 매우 다양하고 〈그림9〉[21]와 같이 저빈도 개체명의 비중이 높아 문헌 빈도 중심의 분석만으로 유의미한 패턴을 포착하기 어려운 경우가 존재했다. 이를 다르게 말하면 단어 빈도 정보의 부재로 인한 제약이라고 할 수 있을 것이다. 이로 인해 TF-IDF(Term Frequency-Inverse Document Frequency)와 같은 표준적인 텍스트 분석 지표나 산포도(dispersion) 지표의 정확한 산출이 어려워졌고, 이는 개체명의 중요도나 특징을 정량적으로 평가하는 데 한계로 작용했다.

이러한 주석 방식은 텍스트 내 개체명의 분포와 맥락 분석에도 영향을 미쳤다. 특정 개체가 어떤 문서들에 등장했는지는 파악할 수 있으나, 각 문서

---

21 개체명의 문헌 빈도는 전형적인 롱테일(long-tail) 분포를 보여준다. 소수의 고빈도 개체명이 분포의 머리(head) 부분을 차지하고, 다수의 저빈도 개체명이 긴 꼬리(tail) 부분을 형성한다. 전체 개체명 중 약 90%가 5회 이하의 빈도를 보이는 것으로 나타나, 저빈도 개체명의 비중이 매우 높음을 알 수 있다.

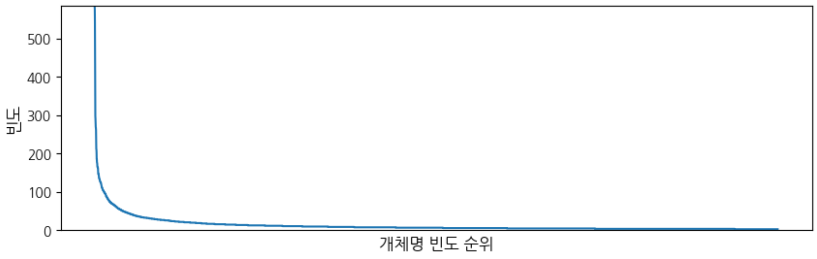

그림9-개체명의 빈도 순위에 따른 빈도 분포

내에서의 출현 빈도나 출현 위치에 따른 중요도 변화는 추적할 수 없게 되었다. 이는 개체명 간의 연관성 분석에도 영향을 미쳐, 문장이나 문단 단위의 정밀한 공기 관계(co-occurrence) 분석이 불가능해졌다. 결과적으로 개체명 간의 강한 연관성이나 개체명이 언급되는 특정 맥락과 같은 미시적 패턴을 포착하기 어려워졌다.

이러한 한계는 형태소 분석기를 활용하거나 최초 주석된 개체명을 텍스트 내에서 자동 검색하여 개체명 빈도를 파악하는 방식으로 일정 부분 보완할 수 있다. 그러나 이러한 자동화된 접근은 필연적으로 오류를 포함하게 되므로, 수작업으로 구축한 주석 데이터의 정확성과 신뢰성을 온전히 활용하기 어렵다는 문제가 있다.

이러한 문제를 통해 주석 방식을 결정할 때 작업의 효율성뿐만 아니라 결과물의 활용 가능성까지 종합적으로 고려하는 것이 중요함을 알 수 있다. 특히 장기적인 관점에서 데이터의 재사용성과 확장성을 확보하기 위해서는, 작업 초기 단계에서 예상되는 분석 방법론과 활용 목적을 면밀히 검토할 필요가 있다.

5. 자료의 분석과 해석

개체명 주석 데이터는 일차적으로 빈도 분석이나 공기어 분석과 같이 빈도를 기반으로 하는 분석에 활용될 수 있다. 그러나 이러한 계량적 분석 결과만으로는 단어가 실제로 사용된 맥락과 의미를 정확히 파악하기 어렵다는 한계가 있다. 같은 단어라도 문맥에 따라 그 의미가 크게 달라질 수 있기 때문이다. 이러한 한계를 보완하기 위해서는 코퍼스 언어학의 방법론인 용례(concordance) 분석이 유용하게 활용될 수 있다. 이 분석은 〈그림10〉과 같이 검색어를 중심으로 그 전후 맥락을 검토하는 방식으로 수행되며, 이를 통해 텍스트를 처음부터 끝까지 순차적으로 읽는 선형적 읽기 방식으로는 포착하기 어려운 단어 사용의 패턴과 의미적 특성을 효과적으로 파악할 수 있게 해 준다는 장점이 있다.[22]

이러한 분석을 지원하기 위해 코퍼스 언어학 연구를 위해 설계된 CQP-Web을 서버에 설치해 활용했다. CQP-Web은 대규모 텍스트 데이터의 효율적 탐색과 분석을 지원하며, 정규 표현식 및 주석 정보를 활용한 상세 검색 기능을 제공한다.[23] 앞서 살펴본 〈그림9〉는 실제로 '京城府_Oraganization'와 같이 개체명 주석 정보를 함께 입력하여 검색한 결과이며, [word="加里波" & tag="Person" & standardForm="가리발디"]와 같이 복합 조건을 설정해 더 정교한 검색을 할 수도 있다. 이 외에도 단어 빈도 테이블 생성, 키워드 분석, 연어 분석 등 코퍼스 언어학의 주요 분석 방법론[24]을 지원하며, 전체 자료뿐

---

22 T. McEnery and A. Hardie, *Corpus Linguistics: Method, Theory and Practice* (Cambridge: Cambridge University Press, 2012).

23 A. Hardie, "CQPweb: Combining power, flexibility and usability in a corpus analysis tool," *International Journal of Corpus Linguistics*, Vol. 17, No. 3(2012), pp. 380~409.

24 키워드 분석(keyword analysis)은 특정 코퍼스에서 통계적으로 유의미하게 빈번히 출

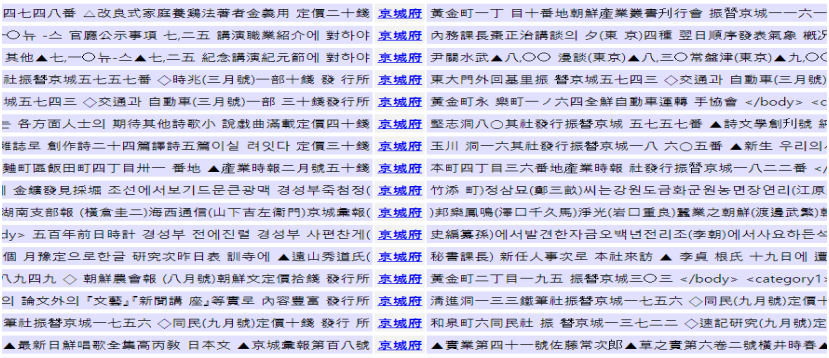

그림10-CQP-Web의 용례 검색 결과

만 아니라 텍스트의 시기나 주제와 같은 메타 정보를 기준으로 하위 코퍼스를 구성하여 분석할 수도 있다. 또한 CQP-Web은 웹 기반의 오픈소스이면서 사용자 권한 관리를 지원하고, 관계형 데이터베이스를 기반으로 빠르고 강력한 검색을 지원하므로, 여러 연구자가 웹 브라우저를 통해 동시 접속하여 활용할 수 있다는 장점이 있다.[25]

현하는 단어를 추출하는 방법론으로 비교 대상이 되는 참조 코퍼스(reference corpus)를 설정하여 분석한다. 예컨대 1920년대와 1930년대 한국학 기사들을 비교하여 1920년대에 특징적으로 나타나는 단어 또는 개체명을 파악할 수 있다. 연어(collocation) 분석은 공기어 분석과 유사하나 언어학에서는 공기 범위를 분석 단어의 앞뒤 n개의 문맥으로 좁게 설정하는 경향이 있다. 단순히 공기 빈도뿐만 아니라 다양한 연관성 척도(association measure)를 지원한다.

25 CQP-Web과는 별개로, 저작권 허용 범위 내의 자료를 대상으로 주석 결과에 대한 검색을 지원하는 외부 공개용 검색 시스템이 현재 개발 중이다.

Ⅳ. 맺음말

이 연구에서는 연세대학교 근대한국학연구소에서 수행한 근대 시기 한국학 관련 기사의 개체명 주석 과정과 결과를 소개하고, 그 과정에서 제기된 주요 방법론적 쟁점들에 대해 논의했다. 구체적으로는 주석 도구의 선정, 한자 및 이표기의 정규화, 주석의 방식, 자료의 분석과 해석과 관련된 문제를 검토하고 문제 해결을 위해 기울였던 노력을 공유하고자 했다.

대부분의 문제들의 근본적인 원인은 근대 시기 한국어에 대한 자연어 처리 기술의 부재에 있다. 해당 시기 한국어 자료에서 의미 있는 언어 단위를 높은 정확도로 추출하는 기술이 존재했다면 모든 단계를 더 수월하게 처리할 수 있었을 것이다. 그동안 이러한 기술 개발에 대한 노력이 없었던 것은 아니었지만, 현대 한국어를 대상으로 하는 기술의 수준과는 상당한 격차가 있는 것이 현실이며, 이는 동 시기의 자료를 대상으로 하는 디지털 인문학 연구의 주요 장애 요인이 되고 있다. 이러한 맥락에서 이 연구에서 소개한 수작업 중심의 주석 작업은 중요한 의미를 지닌다. 비록 근대 시기의 지식 데이터베이스 구축을 목적으로 시작된 작업이지만 주석 결과를 비롯해 주석 과정의 각 단계에서 생성된 언어 자원은 근대 시기 한국어 처리 기술 개발에 활용 가능한 중요하고 희소한 자원임이 분명하다. 궁극적으로는 이러한 언어 자원을 토대로 단순 사전 기반 주석(dictionary-based annotation) 방식을 넘어서는, 문맥 기반의 정교한 자동 개체명 주석 기술을 개발하고, 이를 학계에 공개하여 연구자들이 자유롭게 활용할 수 있게 되어야 할 것이다. 이러한 방향으로 후속 연구를 이어 갈 계획이다.

여러 한계에도 불구하고 이 과업은 대규모의 근대 시기 한국어 기사를 대상으로 한 주석 작업이라는 점에서 선구적인 의의를 지닌다. 이 논문에서 소개하고 논의한 내용이 부족하나마 해당 시기 데이터를 연구 재료로 삼는 후

속 연구 기관 및 연구자들에게 실용적인 제안과 고려 사항으로 활용되기를 바란다.

참고문헌

1. 1차 자료

《동아일보》.《조선일보》.

2. 논저

강범일,「한국어 통시 신문 말뭉치의 구축과 활용」,『언어사실과관점』54, 2021, 7~33쪽.

강범일,「20세기 전반기 한국어 텍스트의 자동 분류 연구: 1920-1939년 한국학 관련 신문 기사를 중심으로」,『인문과학연구논총』44-3, 2023, 221~242쪽.

연세대 근대한국학연구소 인문한국플러스(HK+)사업단(편),『디지털 인문학과 근대한국학』, 서울: 소명출판, 2020.

이정현,「유니코드 한자 검색의 문제점 및 개선방안」,『정보화정책』19-3, 2012, 50~63쪽.

이태훈·정용서·채관식,『일제하 '조선 역사·문화' 관련 기사 목록 1』, 서울: 선인, 2015.

홍정완,「신문으로 읽는 1920년대 식민지 조선의 '조선 역사·문화': 『동아일보』,『조선일보』의 '조선 역사·문화' 관련 텍스트 계량 분석을 중심으로」,『동방학지』198, 2022, 1~37쪽.

홍정완,「1920~30년대 식민지 조선의 종합잡지에 나타난 '조선 역사': 텍스트 계량 분석을 중심으로」,『동방학지』202, 2023, 1~39쪽.

홍정완·정유경·심희찬,『근대한국학 데이터베이스 자료집 1』, 서울: 소명출판, 2020.

Ehrmann, M., Hamdi, A., Pontes, E. L., Romanello, M., and Doucet, A., "Named entity recognition and classification in historical documents: A survey," *ACM Computing Surveys*, Vol. 56, No. 2, 2024, pp. 1~47.

Hardie, A., "CQPweb: Combining power, flexibility and usability in a corpus analysis tool," *International Journal of Corpus Linguistics*, Vol. 17, No. 3, 2012, pp. 380~409.

McEnery, T., and Hardie, A., *Corpus Linguistics: Method, Theory and Practice*, Cambridge: Cambridge University Press, 2012.

Neves, M., and Ševa, J., "An extensive review of tools for manual annotation of documents," *Briefings in Bioinformatics,* Vol. 20, No. 1, 2021, pp. 146~163.

3. 기타

BRAT, https://brat.nlplab.org

국문초록

이 연구에서는 연세대학교 근대한국학연구소에서 수행한 1920-1930년대《조선일보》·《동아일보》 기사의 개체명 주석 과정을 소개하고 방법론적 쟁점을 논의했다. 구체적으로는 발행된 전수 기사로부터 한국학 텍스트를 선별하고, 선별된 텍스트에서 개체명을 식별하고 분류하는 과정과 그 결과를 살펴보았다. 또한 주석 도구의 선정, 한자 및 이표기의 정규화, 주석의 방식, 자료의 분석과 해석 등 주석 과정에서 직면했던 문제를 검토하고 문제 해결을 위해 기울였던 노력을 공유하고자 했다. 이 연구에서 소개한 과업은 근대 한국학 기사를 대상으로 한 최초의 대규모 개체명 주석 작업이라는 점에서 의의를 지닌다. 소개된 주석 과정과 논의된 쟁점들이 해당 시기 데이터를 연구하는 후속 연구자들에게 실질적인 참고가 되기를 기대한다.

투고일 2025. 2. 7.
심사일 2025. 2. 27.
게재 확정일 2025. 3. 4.
주제어(keywords) 개체명(named entity), 언어 주석(linguistic annotation), 한국학(Korean studies), 디지털 인문학(digital humanities), 텍스트 분석(text analysis)

Abstract

Named Entity Annotation in Modern Korean Studies Texts: Newspaper Articles from the 1920s–1930s

Kang, Beomil

This study presents the named entity annotation process conducted on Chosun Ilbo and Donga Ilbo articles from the 1920s–1930s by the Institute for the Study of Korean Modernity at Yonsei University, highlighting key methodological considerations. It examines the selection of Korean studies texts from the full corpus of published articles, the identification and classification of named entities, and the outcomes of this process. Additionally, the study reviews challenges encountered during annotation—including annotation tool selection, normalization of Chinese characters and variant spellings, annotation methods, and data interpretation—and details efforts to address these issues. As the first large-scale named entity annotation project on early 20th-century Korean newspaper articles, this study provides valuable insights for future researchers working with historical Korean texts.