KONG-DB: 웹 상의 어휘 사전을 활용한 한국 소설 지명 DB, 검색 및 시각화 시스템*

KONG-DB: Korean Novel Geo-name DB & Search and Visualization System Using Dictionary from the Web

박성희 (Sung Hee Park)**

초 록

본 연구의 목적은 1) 소설 속 지명 데이터베이스(DB)를 구축하고, 2) 확장 가능한 지명 DB를 위해 자동으로 지명을 추출하여 데이터베이스를 갱신하며, 3) 데이터베이스 내의 소설지명과 용례를 검색하고 시각화하는 파일럿시스템을 구현하는 데 있다. 특히, 학습자료(training)에 해당하는 말뭉치(corpus)를 확보하기 어려운, 소설지명과 같이 현재 잘 쓰이지 않는 개체명을 자동으로 추출하는 것은 매우 어려운 문제이다. 효과적인 지명 정보 추출용 학습자료 말뭉치 확보 문제를 해결하기 위해 본 논문에서는 이미 수작업으로 구축된 웹 지식(어휘사전)을 활용하여 학습에 필요한 충분한 양의 학습말뭉치를 확보하는 방안을 적용하였다. 이렇게 확보된 학습용 코퍼스와 학습된 자동추출 모듈을 가지고, 새로운 지명 용례를 찾아 추가하는 지명 데이터베이스 확장 도구를 만들었으며, 소설지명을 지도 위에 시각화하는 시스템을 설계하였다. 또한, 시범시스템을 구현함으로써 실험적으로 그 타당성을 입증하였다. 끝으로, 현재 시스템의 보완점을 제시하였다.

ABSTRACT

This study aimed to design a semi-automatic web-based pilot system 1) to build a Korean novel geo-name, 2) to update the database using automatic geo-name extraction for a scalable database, and 3) to retrieve/visualize the usage of an old geo-name on the map. In particular, the problem of extracting novel geo-names, which are currently obsolete, is difficult to solve because obtaining a corpus used for training dataset is burden. To build a corpus for training data, an admin tool, HTML crawler and parser in Python, crawled geo-names and usages from a vocabulary dictionary for Korean New Novel enough to train a named entity tagger for extracting even novel geo-names not shown up in a training corpus. By means of a training corpus and an automatic extraction tool, the geo-name database was made scalable. In addition, the system can visualize the geo-name on the map. The work of study also designed, implemented the prototype and empirically verified the validity of the pilot system. Lastly, items to be improved have also been addressed.

키워드: 소설지명추출, 개체명 인식, 웹지식베이스, 조건부랜덤필드, 텍스트마이닝, 기계학습, 소설지명 및 용례정보 시각화

novel geo-name extraction, named entity recognition, Web knowledge base, conditional random fields, text mining, machine learning, novel geo-name map visualization

* 이 논문 또는 저서는 2015년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2015S1A5A8017833),

** 한남대학교 문헌정보학과 조교수(sunghee@hnu.kr)

■ 논문접수일자 : 2016년 8월 23일 ■ 최초심사일자 : 2016년 9월 5일 ■ 게재확정일자 : 2016년 9월 13일

■ 정보관리학회지, 33(3), 321-343, 2016. 〔http://dx.doi.org/10.3743/KOSIM.2016.33.3.321〕

1. 서 론

1.1 연구배경 및 목적

최근 빅데이터 분석에 대한 관심이 높아지면서, 통계자료, 수치자료와 같이 정형화된 자료뿐만 아니라 인문학(문학, 사학, 철학 등) 분야를 포함한 텍스트 기반 빅데이터 자료를 분석하려는 시도가 있어 왔다. 구글의 북스 프로젝트[1]가 그 대표적인 예다. 이는 기존의 장서를 디지털화하고 그것들을 정보처리/관리기술을 이용하여 빅데이터를 생성한 것이다. 이 프로젝트의 결과물을 가지고 엔그램뷰어[2]를 통하여 특정단어의 연도별 빈도수를 분석할 수 있게 되었다(문상호, 2015).

단순히 키워드의 출현 빈도수 분석에서 발전하여, 좀 더 의미 있는 빅데이터 분석을 위해서는, 비정형의 속성을 지닌 데이터로부터, 관심 있는 속성을 지닌 키워드들을 선택적으로 추출하여 정형 자료화하는 것이 필수적이다. 이러한 문제는 흔히 정보추출(Information Extraction) 혹은 개체명 인식(Named Entity Recognition)이라고 알려져 있다. 비정형 텍스트 기반 인문 자료로부터 정형 자료화를 해야 하는 개체명 중 지명 정보는 특별한 의미를 갖는다. 인문 자료에서 사람의 장소에 대한 지각과 인지는 일차적으로 지명을 통해 이루어지기 때문이다(장노현, 2008). 지금까지 이러한 지명을 데이터베이스로 만드는 것에 대한 시도들이 문화, 지리, 역사 분야에서 있어 왔다(장문현, 2015).

일단 구축된 지명 정보는 문학 연구뿐만 아니라 지역학 및 문화콘텐츠 산업 등의 다양한 분야에 활용될 수 있기 때문에, 정보로서 체계적이고 효율적인 관리가 필요하다. 지명 데이터베이스 구축을 통하여 얻는 이점은 연구자의 임의성이나 식견에 좌우되지 않고 객관적인 장소성과 관련한 자료 획득이 가능하다는 것이다(장노현, 2008).

그렇다면, 왜 이런 문학작품 속 지명을 추출하여 시각화하는 것이 중요한 일일까? 문학연구자들이 작품지명을 통해 특정지역(예를 들어, 대전)과 관련된 문학작품 속에서의 이미지를 파악해서 신작품을 작성하고자 할 때 작가들의 그 지역에 대한 인식을 손쉽게 파악할 수 있다. 하지만, 한국 신소설에는 현재 잘 쓰이지 않는 옛날 지명과 현대 지명이 공존해서 나타난다. 지금은 잘 쓰이지 않는 옛날 지명에 대한 정보추출은 학습자료(training)에 해당하는 말뭉치(Corpus)를 확보하는 일이 어려운데 이것은 자동으로 지명 추출을 어렵게 만드는 중요한 요인 중 하나다.

위에서 서술한 것과 같은 배경 하에, 본 연구는 소설 작품으로부터 지명 데이터베이스를 구축하고, 데이터베이스에서 그 지명이 어느 작품에서 나왔는지 어떤 용례에서 쓰였는지 검색할 수 있으며, 지도 위에 그 지역을 시각화하는 파일럿 시스템을 설계하고 구현함으로써 그 실현 가능성을 확인하는 데 목적이 있다. 또한 지명 데이터베이스의 확장을 손쉽게 하기 위하여 학습자료 말뭉치를 확보하기 어려운 신소설 내의 소설 지명을 자동추출 문제를 해결하는 방법을 제시하고 그 방법의 실현가능성을 확인하고자 한다.

이러한 목적을 달성하기 위하여 다음과 같은 연구질문을 가지고 연구를 수행하였다.

---

1) Google Books, Retrieved from https://books.google.com/
2) Google Books - Ngram Viewer, Retrieved from https://books.google.com/ngrams

- RQ 1. (기초 지명DB구축) 지명DB를 구축할 수 있는 효과적인 방법은 무엇인가?
- RQ 2. (학습자료 수집의 어려움) 지명DB를 확장 가능하도록, 소설 속 지명정보를 자동으로 추출할 때, 학습자료를 수집의 어려움을 어떻게 극복할 것인가?
- RQ 3. (검색지명 시각화) 질의로 들어온 지명 키워드를 어떻게 지도 위에 시각화할 수 있을까?

1.2 연구의 내용 및 방법

첫 번째 연구질문의 연구내용/방법은 수동구축의 시간적, 비용적 제약을 극복하기 위하여, 웹 상의 어휘사전을 활용하여 기초 지명DB를 구축하는 것이다. 이를 위해 다음과 같은 순서로 진행하기로 하였다.

1) 이미 구축되어 있는 한국신소설 어휘사전으로부터의 지명에 관한 어휘 및 용례를 추출한다.
2) 추출된 소설 속 지명정보의 현대 지명을 수동으로 매핑한다.

두 번째 연구질문에 대한 연구내용은 확장 가능한 DB갱신문제로, 소설 지명 정보를 자동으로 추출・갱신하는 방법에 관한 것이다. 자동 추출에 의한 소설 속 지명정보 추출은 학습데이터 말뭉치를 확보하기 어려운 문제점을 야기하였다. 이를 해결하기 위한 방법은 다음과 같다.

1) 이미 구축되어 있는 한국신소설 어휘사전으로부터의 지명에 관한 어휘 및 용례를 추출한다(RQ 1의 1) 방법과 동일).
2) 용례로부터 텍스트 전처리를 통해 학습데이터를 구축한다.
3) 학습데이터를 이용하여 CRF 기반의 지명 자동 추출 모델을 학습한다.

이와 같이 자동화된 방법으로 지명정보에 대한 학습을 수행한 후,

4) 일반신소설 텍스트를 입력 받았을 때, 자동으로 지명정보를 추출하고,
5) 그 추출된 지명 정보에 대하여 검색한 결과 DB 내에 있는 지명이면 그 지명의 용례 테이블에 현재 예를 추가하고, 그렇지 않고 그 지명이 없다면 지명 테이블과 용례 테이블 두 곳에 함께 추가한다.

마지막 연구질문을 위해서는 키워드에 해당하는 소설 속 지명을 입력으로 받아 지명 DB검색, 소설 지명을 현대지명으로 바꾸면 지명에 대응되는 위경도로 변환하여 반환해 주는 웹서비스를 활용한 Mash-up 방법을 적용한다.

2. 관련연구

2.1 문학지리학 활용연구

문학과 문학작품 속에서 표현된 공간에 대한 연구를 문학지리학이라고 하는데, 지명으로 대표되는 공간과 그 공간의 속성을 접목하여 지도 위에 시각화하는 시도가 여러 분야에서 있었다. 장문현(2015)은 최근에 섬진강 역사문화 유적을 대상으로 지도 상에 감성 문화 지도를 GIS 기법을 활용하여 시각화하였다. 이번 연구에서는 역사 유적분야를 주로 사용하고 고고, 건축, 문학

분야 관련 자료를 보조자료로 활용 감성인자를 추출하고, 시각화하였다. 작품 속 지명 정보를 체계적으로 구축하여 활용하려는 방안에 대한 기초 연구도 이루어졌다. 장노현(2008)은 한국 신소설 속 지명 정보에 대하여 개념적 정의, 한국소설지명 데이터베이스의 개념적 구조를 지명, 작품, 작가 세 테이블로 설계하고 기술 원칙을 제시하였다. 이은숙, 김일림, 정희선(2007)은 서울 종로 사례를 중심으로 시나 소설과 같은 문학 속에 포함된 공간(이하, 문학공간이라고 지칭함)에 대한 데이터 베이스를 구축하는 방안에 대한 연구를 진행하였다. 박철수(2008)는 문학작품 속으로부터 특정 지역(북촌 도시한옥마을)에 대해 묘사된 부분을 수집, 그 지역에 대한 시대적 변화와 독자들의 그 지역에 대한 인식을 분석하였다. 한순미(2013)는 소설 속의 지명 연구에 대하여 다시 강조하고 감성지도에 활용 방안을 제시하였다.

한편, 최진무, 김민주, 최돈곤(2014)은 우리나라 지명DB와 수치지도DB를 연계하기 위하여 지명DB스키마에 지명이력 관리테이블을 추가 수정하였다.

살펴본 바와 같이 대부분의 연구들은 특정지역에 대한 자료들을 수작업으로 수집/분석하여 속성을 부여하고 시각화하는 연구들이었다. 그리고, 문학소설작품 속 지명에 대한 데이터베이스의 개념적 설계가 있지만 실제 데이터베이스 구축의 후속 연구가 필요한 상황이다. 특히, 초기 DB 구축에 많은 비용이 소요되는 인문DB 구축의 특성상, 꾸준한 DB 확장을 위해 확장 가능한 방안이 필요하다. 여기에 본 연구는 이러한 요구사항을 충족하기 위해 자동 개체명 인식 기법을 도입하고자 한다.

2.2 개체명인식기법

장소(Location), 인물(Person), 기관명(Organization)과 같이 명명된 것을 개체명(Named Entity)이라고 하는데, 텍스트로부터 개체명을 인식하는 기법들과 관련하여 최근까지 많은 연구들이 있어 왔다. 개체명 인식 기법은 크게 두 가지로 나눌 수 있다. 하나는 규칙기반 개체명 인식과 다른 하나는 기계학습기반 개체명 인식 기법이다. 규칙기반 개체명 인식은 개체명에서 비슷한 규칙이 변화가 크지 않고 거의 일정하게 나타날 때 적은 비용으로 구현할 수 있는 장점이 있다. 반면, 변화가 많고 그 규칙을 명시적으로 기술하기 어려울 때에는 기계학습기반 정보추출이 유용하다. 다양한 기계학습 기법들, 즉, Hidden Markov Model(HMM)(황이규, 윤보현, 2003), Conditional Random Field(CRF)(Lafferty, McCallum, Pereira, 2001), Maximum Entropy(김성원, 나동렬, 2008), Structured Support Vector Model(SVM) + Pegasus(이창기, 장명길, 2010), Deep Learning(이창기 외, 2014)이 개체명 인식에 적용되었다.

이러한 기계학습기반 개체명 인식에서의 단점은 학습자료를 구축하는 일이다. 이러한 문제해결을 위하여 Park, Ehrich, Fox(2012)는 참고문헌으로부터 메타데이터 자동추출과 관련하여 학습자료를 기존의 외부 지식으로부터 손쉽게 구축하여 기계학습과 규칙기반 개체명 인식을 결합한 하이브리드 방식을 제안하였다. 또한, 이러한 기계학습 기반의 개체명 인식 및 관계식과 같은 의미추출은 다른 분야, 특히, 생물의학분야에서 활발하게 이루어지고 있다(최성필, 2016). 이은령(2009)은 19세기 문헌 국역본에

대해서 ACE[3] 개체명인식 기초 연구를 수행함으로써 고문헌에 대한 개체명 인식 기법을 시도하였다.

앞에서 살펴보았듯이, 개체명 인식 문제에 다양한 기법들이 다양한 분야에 적용되었다. 본 연구는 확장 가능한 지명 DB 갱신 방법으로 웹 상의 외부지식을 활용한 개체명 자동추출 기법을 적용하고자 한다.

3. 외부지식을 이용한 한국 문학 작품 속 소설지명 DB구축 및 갱신

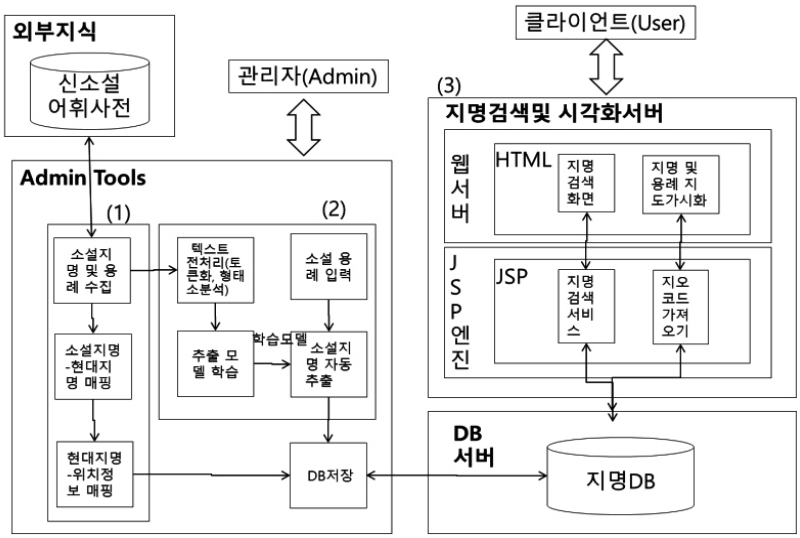
KONG-DB시스템은 크게 세 부분으로 나눌 수 있다. 관리자 도구(Admin Tool), DB 서버, 지명 검색 및 시각화 서버들로 이루어진다.

3.1 시스템 구성도

먼저 관리자도구(Admin Tool)들은 1) 소설지명 및 용례 수집을 통한 초기DB 입력 도구와 2) 신소설 지명 추가 모듈로 구성된다. 이는 모두 소설지명 DB 구축을 위한 백엔드(back-end) 관리자 툴이다. 이러한 관리자 툴을 이용하여 소설지명 DB 구축하는 과정은 다음 절에서 자세하게 설명한다.

DB 서버는 초기 어휘사전으로부터 수집된 지명들을 저장한 데이터베이스를 관리하는 요소이다. 또한, 새로운 소설 작품으로부터 자동으로 추출된 지명들이 저장되어 관리되는 곳이기도 하다.

마지막으로 검색 및 시각화 서버는 사용자들로부터 소설지명에 대한 키워드 질의가 있을 때, 그 지명에 대한 용례와 지도 상의 위치정보를 함께 시각화하는 서브시스템이다. 〈그림 1〉은 전체 시스템 구성도를 보여준다.



〈그림 1〉 시스템 구성도

3) Automatic Content Extraction(ACE) Program. Retrieved from https://www.ldc.upenn.edu/collaborations/past-projects/ace

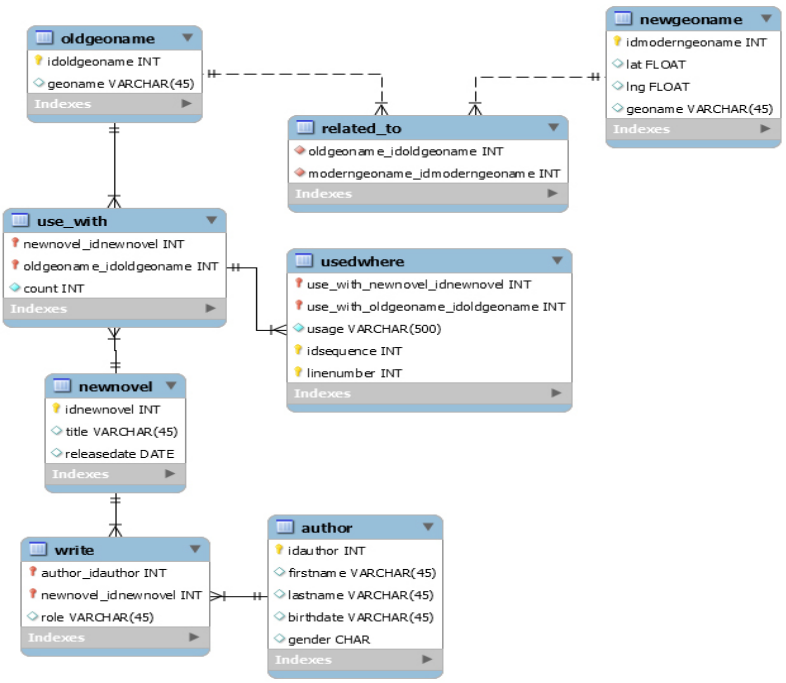
3.2 소설지명 초기DB 구축단계

이 절은 연구질문 1에 대한 해결과정을 설명한다. 즉 소설지명을 초기에 효과적으로 구축하는 과정은 다음과 같다. 먼저 지명데이터베이스를 설계한다(3.2.1). 웹 상의 어휘사전으로부터 지명정보와 용례를 수집한다(3.2.2). 수집된 소설지명으로부터 현대지명을 매핑한다(3.2.3). 현대지명은 Google Geocoding API를 통해 위치정보(위·경도)로 매핑한다(3.2.4). 이렇게 채워진 값들은 관리자도구에 의해 DB에 업로드된다(3.2.5).

3.2.1 지명 데이터베이스 설계

소설지명 데이터베이스는 크게 지명정보, 작품정보, 저자정보로 나눌 수 있다. 지명정보는 소설지명 정보 외에 시각화를 위하여 현대 지명정보와 그 지명의 위치(공간)정보까지 포함해야 한다.

먼저, 지명정보는 oldgeoname, newgeoname, related_to 테이블들로 구성된다. oldgeoname 테이블은 소설 속 지명 정보를 저장한다. 속성은 지명(geoname) 자체만을 담고 있다. 장노현(2008)에서 언급한 *서술정보*는 추후 추가할 예정이다. newgeoname은 현대 지명정보를 저장한다. 속성은 geoname과 위치정보(위, 경도)를 담고 있다. 이 소설지명과 현대 지명 간에는 다대응관계를 형성한다. 이러한 대응정보를 표현한 테이블이 related_to이다. 작품정보는 newnovel 테이블에, 저자정보는 author테이블에 대응된다. 저자와 작품은 write관계에서 다대응관계로 모델링하였다. 〈그림 2〉는 소설 지명 데이터베이스의 E-R 다이어그램을 보여준다. 〈표 1〉은



〈그림 2〉 소설지명 RDB E-R Model

〈표 1〉 지명 데이터 베이스와 테이블 예

oldgeoname	ID	(일괄 부여)
	지명	계야
newgeoname	ID	(일괄 부여)
	지명	개풍군
	위도(lat)	37.954173
	경도(lng)	126.453924
usedwhere	ID	(일괄 부여)
	용례	이곳이 간성 이라 니
	신소설ID	30
	문학속의 지명ID	5
	책속에서의 위	103
newnovel	ID	(일괄 부여)
	제목	강상
	출판일	19xx.xx.xx
	책 정	강상루는...
author	ID	(일괄 부여)
	성	이
	이름	인직
	생일	1862.07.27
	성별	남
use_with	newnovel_idnewnovel	(일괄 부여)
	oldgeoname_idoldgeoname	(일괄 부여)
	count	3(작품 내 어휘출현 빈도수)

설계된 지명 정보 데이터베이스의 E-R모델에 따라 생성된 테이블과 인스턴스 예를 보여준다.

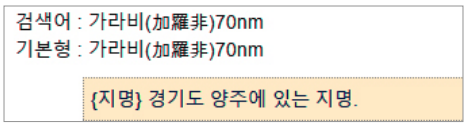
3.2.2 작품 속 지명정보 및 용례수집

한국 신소설 어휘사전 사이트(http://newnovel.aks.ac.kr)는 '신소설 어휘 사전 편찬 연구(Compilation of Word Dictionary of the Korean New Novels)'라는 3개년 연구과제의 결과물로 구축된 어휘사전 사이트이다. 총 59편의 연구 대상 자료로부터 90만 여 개의 어휘를 추출하여, 각 어절에 대한 정규화, 기본형 확정 등의 작업을 완료하였다. 기본형 표제어 4만 여종의 어휘에 대한 뜻풀이를 완료하였다. 그 성과물은 MS Access 데이터베이스의 테이블로 수록되어 있으며, 그 목록은 다음과 같다.[4)]

1) 신소설 목록(총 59종에 대한 서지 정보 제공)
2) 신소설 용례색인(총 90만 어절에 이르는 어휘 용례색인)
3) 신소설 어휘사전(기본형 어휘 4만여 종에 대한 어휘 풀이)

4) 목록정보는 다음 링크의 문서에서 발췌하였음.
http://newnovel.aks.ac.kr/Documents/Introduce_New_Novel.pdf

위의 한국 신소설 어휘사전 사이트에서 한가지 주목할 사실은 구축된 어휘 중 지명에 해당하는 어휘들이 '{지명}'이라는 어휘 풀이로 입력되어 있다는 점이다(〈그림 3〉 참조). 하지만, 한국 신소설 어휘사전 사이트에서 바로 '지명'이라는 단어로 검색하지 않고, 고유명사를 뜻하는 'nm'을 검색한 뒤 웹 페이지를 가져왔다. 그 이유는 한국 신소설 어휘사전에는 '지명'뿐 아니라 장소에 해당하는 '국명', '국가명'을 모두 분리해 내야 했기 때문이다. 예를 들어, 'nm'은 지명뿐만 아니라, 국명, 국가명, 인명, 선박명을 다 포함하고 있어 지명만을 검색했을 때 놓칠 수 있는 국명, 국가명 어휘까지 포함할 수 있었다. 검색한 결과는 〈그림 4〉와 같다.



〈그림 3〉 지명 '가라비'의 어휘 풀이 예

검색결과화면을 보면, 최상단에 검색 창이 위치하고, 아래 왼쪽 패널에 어휘를 포함하고 있는 작품들의 리스트가 위치한다. 오른쪽 패널 상단에는 기본형 링크들이 위치하는데, 이는 실제 nm을 포함한 어휘사전 내 어휘들을 언급빈도에 따라 정렬을 하여 일정 상위 어휘만을 보여준다. 마지막은 오른쪽 패널 하단의 내용으로 용례정보 제시 부분이다. 여기는 각 어휘들이 출현한 용례를 보여준다. 검색된 어휘의 기

〈그림 4〉 한국 신소설 어휘사전 사이트 nm 검색 결과 화면

본형, 활용형과 용례로 구성된다.

여기서, 기본형링크는 'nm'을 포함한 모든 어휘를 보여주지 않고 빈도수가 높은 상위 어휘만 보여주므로 모든 어휘를 수집하는 방법을 취해야 한다. 그 방법은 용례 정보 제시 부분의 기본형에 포함되어 있다. 이 정보를 모두 엑셀에 저장하고 중복 어휘를 제거하여 유일하게 'nm'을 포함한 어휘리스트를 추출한다.

추출한 nm어휘리스트의 모든 어휘에 대해서, HTML파서를 내장한 웹크롤러, Python 라이브러리 *BeautifulSoup*[5]에 검색어로 넣고 어휘사전 사이트에서 검색을 수행한다. 예를 들면, http://newnovel.aks.ac.kr/Search?keyword=가라비(加羅 非)70nm이다. 그 결과 (〈그림 3〉 참조. 보여지는 바와 같이 '지명'이 표시되어 있다) 단어들에 '지명', '국명', '국가명'으로 표시되어 있는 어휘를 가져온다. 가져온 결과는 〈표 2〉에 보여진다.

〈표 2〉 지명, 국명, 국가명 추출 예

가라비(加羅非)70nm
가맹특도70nm(카트만두00nm)
가산(家產)04ng
가평(加平)02nm
간성(杆城)03nm
갈마(葛麻)70nm
갈모봉(_帽峯)70nm
감은돌[黑石]70nm

이 과정에서 기본형, 활용형, 용례, 작품명도 함께 가져와 csv(Comma Separated Version) 파일에 저장한다. 〈표 3〉은 CSV 파일에 저장된 결과 예를 보여준다.

3.2.3 소설지명-현대지명 맵핑

소설지명 - 현대지명 맵핑은 수작업으로 하였다. 〈표 4〉는 소설지명과 현대지명을 매핑한 예를 보여준다.

수작업으로 변환하는 과정에서 발생할 수 있는 문제의 경우는 세 가지로 예상해 볼 수 있었

〈표 3〉 CSV 파일에 저장된 결과 예

기본형	활용형	용례	작품명	작품속 위치
대안동(大安洞)70nm	대안동	(물론 이 소식과 이 사실을 대안동자긔 시집에도 젼얏슬 일)	금국화(하)	85
오복동(五福洞)70nm	오복동은	(오복오일을 류오복니오복불안오복안이라) 오복이가 오 일 동안을 오복동에셔 머무니 오복이 편안치 못데오복동은편안더라	구의산(하)	065:06
서울01㉡nm	셔울로	「가 네를 다리고 셔울로 도로 가셔 어머니 네 정을엿줍고십으나	금강문	111
다방골(茶坊_)70nm	다방골	「고 그길로 나 셩즁을드러셔셔다방골모통이를도라드니 포 그리던 촌을 맛날 터인즉	구마검	059:15
왕십리(往十里)70nm	왕심리를	「고 뒤도 안이 도라보고왕심리를향고가거	구마검	037:26

5) https://pypi.python.org/pypi/beautifulsoup4

다. 첫 번째로 사라진 지명, 두 번째로 현재는 다른 지역과 합쳐져서 확대된 경우(예를 들어, 광역시), 마지막으로 과거에는 넓은 지역을 지칭했으나 현재는 행정구역이 세분화된 경우다.

〈표 4〉 소설지명과 현대지명 매핑 예

소설지명	현대지명
가라비(加羅非)	광적면
가맹특도(카트만두)	Kathmandu
가산(嘉山)	박천군
가평(加平)	가평(加平)
간성(杆城)	간성
갈마(葛麻)	갈마반도

매핑과정에서 있었던 몇 가지 예를 들면, 큰 강의 하류 또는 상류에 붙은 작은 강들은 큰 강들로 표시하였다. 종로구 '~동'과 '~동' 사이에 있던 마을은 종로구로 표시하였다(예: 대묘골(종로구 훈정동, 묘동, 봉익동, 종로3가, 종로4가에 걸쳐있던 마을)은 종로구로 표시).

소설지명을 현대지명으로 맵핑하는 부분은 전문가의 검증이 필수적이며 앞으로 충분히 개선해야 할 필요가 있으므로 향후 연구에서 다시 언급할 것이다.

3.2.4 현대지명에 대한 위치정보(위·경도)로의 변환

Java와 Google GeoCoding API를 사용해서 현대지명의 위치정보(위·경도)를 변환하였다. Google API 중 GeoCoding API는 위치정보를 알고자 하는 지명을 address 인자(parameter)의 값으로 포함하여, URL로 만들어 서버에 요청하면 JSON 파일 형태로 위치정보(위·경도)를 알려준다(아래 URL 참조).

https://maps.googleapis.com/maps/api/geocode/json?address=″지명″&key=xxxxxxxxx

URL을 웹서버에 보내서 되돌려 받은 JSON 파일의 위경도 정보는 〈그림 5〉와 같다.

〈그림 5〉 Google GeoCoding API를 통한 '서울'의 위치정보 반환 예(JSON 포맷)

이러한 작업을 매번 웹브라우저의 주소창에서 수작업으로 하면 시간이 많이 걸리기 때문에 Java에서 지명정보를 담은 텍스트 파일을 불러와 웹서비스를 요청할 URL을 JAVA로 프로그램을 만들어서 자동화하였다.

위의 사진에서 보여지는 JSON파일을 파싱한 뒤, "location"객체의 lat(위도)와 lng(경도)값을 가져와 csv 파일에 저장하였다. 〈표 5〉는 저장된 위치정보변환결과.csv를 한 예를 보여준다.

〈표 5〉 위치정보(위・경도) 결과.csv

소설지명	현대지명	위치정보(위・경도)
가라비(加羅非)	광적면	37.8246038 126.9836808
가맹특도(카트만두)	Kathmandu	27.7172453 85.3239605
가산(嘉山)	박천군	39.7203647 125.5862806
가평(加平)	가평(加平)	37.8253113 127.5165397
간성(杆城)	간성	38.37836430000001 128.4672371
갈마(葛麻)	갈마반도	39.1947222 127.4772222

3.2.5 소설지명-현대지명-위치정보(위・경도) DB 업로드

본 단계는 수집된 지명정보들을 DB에 업로드하는 과정이다. 수집된 지명정보들은 CSV 포맷으로 만들었다. 지금까지 만든 CVS파일을 DB에서 불러오기(import)하였다. 지명데이터베이스 스키마에 대해서는 앞 절(3.2.1)에 상세하게 설명하였다. 〈표 6〉과 〈표 7〉은 각각, 수집되어 지명 DB에 불러오기 위한 소설지명 정보와 현대지명 정보에 대한 CSV 파일 예를 보여준다.

〈표 6〉 소설지명.csv

ID	지명
1	가라비(加羅非)
2	가맹특도(카트만두)
3	가산(嘉山)
4	가평(加平)
5	간성(杆城)
6	갈마(葛麻)

〈표 7〉 현대지명.csv

ID	위도	경도	지명
1	37.8246	126.9837	광적면
2	27.71725	85.32396	Kathmandu
3	39.72036	125.5863	박천군
4	37.82531	127.5165	가평(加平)
5	38.37836	128.4672	간성
6	39.19472	127.4772	갈마반도

이 때 용례의 기본형들을 각각 그에 맞는 id값으로 excel 함수를 이용해 변경시킨 뒤 불러오기(import)하였다. 〈그림 6〉은 id값으로 변경한 후 불러오기(import)한 용례(usedwhere) 테이블 예를 보여준다.

use_with_n	use_with	usage	idsequence	linenumber
53	1	다만 그 모진 명녕으로 라비 단겨오면 실을 일일히 리약이...	1	064:16
23	2	가 특도는 네팔국의 뎨일 큰 도회라	2	135
23	2	즉시 가 특도를 나 다시 셔북을 향 고 가다가 잡함 디방에를...	3	136
38	3	[양] 가산 별령에셔 낫단다	4	034:02
38	3	[양] 시신은 가산 안쥬 박현 셰 고을 사 이라 죽은 모양...	5	034:04
38	3	고마루에 가산 사 돌이 죽엇	6	034:11

〈그림 6〉 id값으로 변경한 후 불러오기(import)한 용례 테이블 예

3.3 소설지명 DB 갱신

지명 데이터베이스를 자동으로 갱신하기 위해서는 지명 정보 자동 추출 기능을 갖추어야 했다. 지명 정보 자동 추출 과정은 크게 두 단계로 이루어진다. 첫 단계는 지명정보 추출기(named entity tagger)를 학습하는 과정이고 두 번째 단계는 학습된 지명정보 추출기를 이용하여 새로운 비정형(unstructured) 소설자료가 입력되었을 때, 자동 추출하는 과정이다. 먼저 학습과정을 살펴보자.

3.3.1 외부지식으로부터의 학습

지명데이터베이스를 자동으로 갱신하기 위하여 지명정보 자동추출기법을 사용하였다. 지명정보 자동갱신을 위하여 MALLET(McCallum, 2002)이라는 기계학습(CRF) 기반의 개체명 인식기를 사용하였다. 기계학습 기반의 개체명 인식기는 학습과정과 디코딩 과정으로 크게 구분되는데, 학습과정에서 학습데이터 확보가 필수적이다. 학습데이터 구축의 두 가지 방법으로 수동구축과 반자동 구축이 있다. 여기에서는 외부지식을 활용한 반자동 학습데이터 구축을 시도하였다.

특별히 주목할 점은 이러한 지명 어휘와 그 지명 어휘가 사용되었던 문장에 대한 용례를 얻을 수 있다면 그 어휘-용례 쌍을 학습코퍼스 구축에 활용할 수 있을 것이라는 점이다. 즉, 초기 지명 DB 구축 시 가져온 지명과 용례를 학습코퍼스 구축에 재사용한다는 접근방법이다. 앞에서도 언급했듯이 한국 신소설 어휘 사전 사이트는 신소설에 나타나는 어휘들의 뜻과 용례를 수작업으로 구축한 사전이다.

MALLET의 학습데이터 포맷은 다음과 같다. 〈그림 7〉에서 보여지듯이 지명단어만 Location으로 태깅하고 나머지 단어는 'Null'을 의미하는 'O'로 표시한다. '/'는 new line을 의미한다.

해야 할 다음 작업은 당연히 용례로부터 위의 MALLET 입력 포맷으로 변환하는 작업이다. 〈그림 8〉은 지명 어휘의 검색 결과 화면을 보여준다. 웹 화면에서 지명 어휘가 빨간색으로 강조되어 있다는 점에 주목하라.

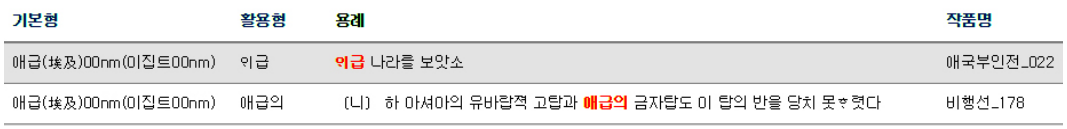
빨간 색으로 처리된 지명만을 찾기 위해서 웹페이지를 가져올 때 태그를 포함한 웹페이지 소스를 가져왔다. 이 페이지의 소스를 보면 *〈font class = "inflectional"〉애급〈/font〉*와 같이 되어있다. 이번에는 태그를 포함해 웹 페이지를 가져온 뒤(〈표 8〉검색 결과, 태그를 포함한 예) 〈tr〉,〈td〉와 같은 태그들은 제거하고 지명에 색깔을 바꿔준 〈/font〉은 체크 표시(c)로 바꾸었다(〈표 9〉용례 체크 표시).

```
긔왕에 O /예지 O /려왓스니 O/일본으로 Location/건너가 O

삼남 Location/등디로 O/도라단인다 O/소문이 O/잇더니 O

셔울로 Location/ 올나와셔 O/ 바로 O/죵로를 Location/지나 O/사동 Location/골목으로 O/들어서셔 O/서울 Location/양반이 O/너를 O/졍실부인으로 O/장가드실 O/터이다 O
```

〈그림 7〉 MALLET 입력 포맷의 예



기본형	활용형	용례	작품명
애급(埃及)00nm(이집트00nm)	이급	**이급** 나라를 보앗소	애국부인전_022
애급(埃及)00nm(이집트00nm)	애급의	〔니〕 하 아셔아의 유바랍젹 고탑과 **애급의** 금자탑도 이 탑의 반을 당치 못ᄒᆞ렷다	비행선_178

〈그림 8〉 검색 결과 사이트 화면

〈표 8〉 검색 결과, 태그를 포함한 예

기본형	활용형	용례	작품명	작품속 위치
〈td〉애급(埃及)00nm(이집트00nm)〈/td〉	〈td〉애급의〈/td〉	〈td〉〈span〉〔니〕 하아셔아의유바랍젹고탑과 〈font class="inflectional"〉애급의〈/font〉금자탑도이탑의반을당치못렷다〈/span〉〈/td〉	〈td〉비행선	178〈/td〉

〈표 9〉 용례 체크표시

기본형	활용형	용례	작품명	작품속 위치
애급(埃及)00nm(이집트00nm)	애급의	〔니〕 하 아셔아의유바랍젹 고탑과 **애급의c** 금자탑도 이 탑의 반을 당치 못렷다	비행선	178

이를 이용하여 모든 지명에 체크 표시를 할 수 있었다. 한 가지 주의할 점은 한 문장에 지명이 두 개가 있는 경우이다. 이 경우, 동일한 문장의 용례가 두 개 있을 수 있다. 예를 들어, "*오복오일을 류오복니오복불안오복안이라) 오복이가 오 일 동안을* ***오복동에셔c*** *머무니 오복이 편안치 못데****오복동은c****편안더라*"이다. 이러한 문장의 경우 한 문장으로 만들어 학습시키는 것이 모델에 충실할 것으로 판단하였다. 이 중복된 문장은 수작업으로 합쳤다.

그 뒤 지명이 체크된 용례들로 MALLET NER을 학습시킨다. 학습의 결과, 지명 자동추출기 모델(하나의 학습모델 파일)이 생성된다. 이 모델을 사용하여, 새로운 소설문장(이미 토큰화된 문장 형태의 입력)이 입력으로 들어왔을 때, 지명을 자동으로 인식하게 된다. 입력과 출력의 형태는 〈표 10〉과 같이 입력 토큰이 지명정보인지 아닌지를 인식해서 출력파일로 저장하여 알려준다. 한 가지 주의할 점은 출력파일에는 입력 자료의 토큰들을 담고 있지 않다는 점이다.

〈표 10〉 중복된 문장을 합친 문장

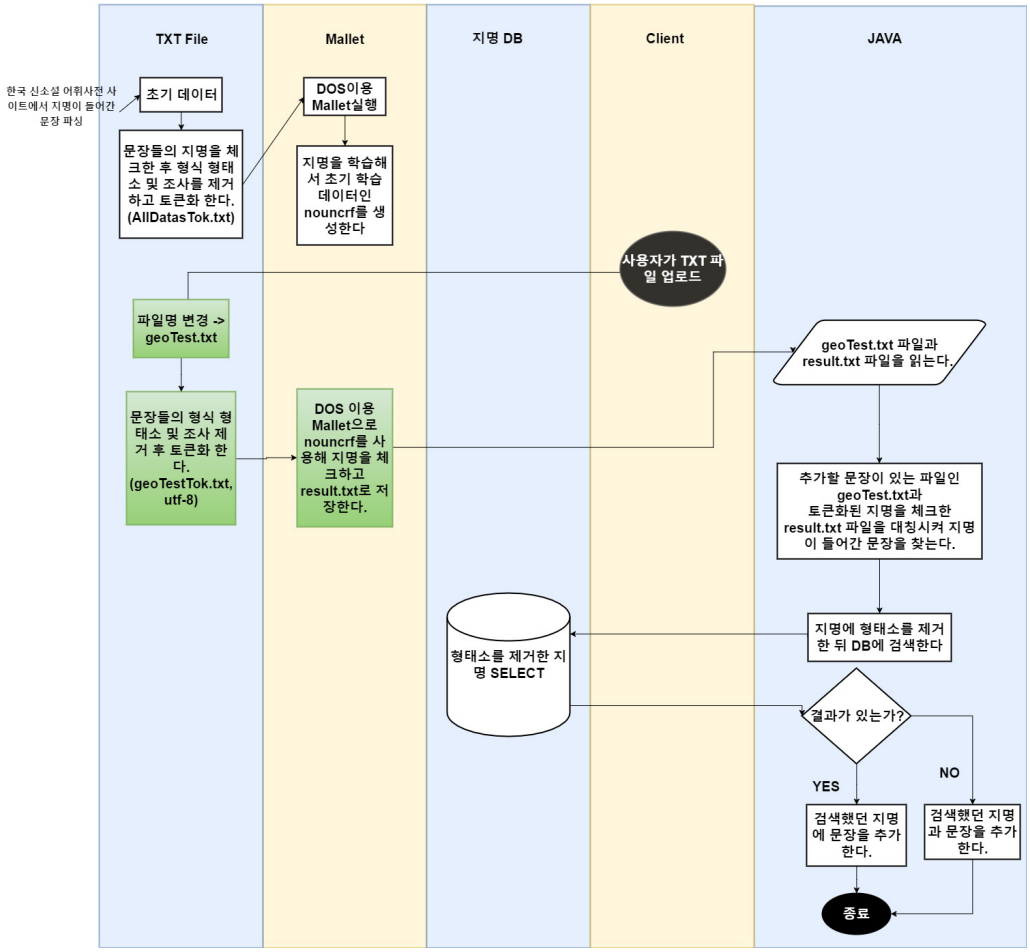
input(geoTestTok.txt)	output(result.txt)
긔왕에	O
예지	O
려왓스니	O
일본으로	Location
건너가	O

출력파일이 나오게 되면 후처리 작업을 통하여 추출된 토큰(지명)만을 가지고 지명 DB 자동 갱신 과정을 수행하게 된다. 이 과정은 새롭게 추출된 소설 지명이DB의 소설지명테이블에 이미 있는지 아니면 없는지 확인하는 단계로 시작한다. 만약 이미 있다면 추가 과정을 마치고, 없다면 새로운 지명이므로 추가 과정을 수

행한다.

상술하면 위의 두 입력과 출력 자료인 geo TestTok.txt파일과 result.txt파일을 읽는다. 검사할 문장이 있는 geoTest.txt파일과 토큰화 되어 있는 지명을 체크한 result.txt파일을 서로 매칭시킨다. 이 때 지명으로 매칭 되어 있는 geoTest.txt문장의 단어에서 형태소 분석을 통하여 형태소를 제거한다. 형태소 분석기는 KAIST, Semantic Web Research Center의 한나눔 분석기[6]를 사용하였다. 그 뒤 이 단어를 DB에서 검색하는데 이 때 DB에 이미 있는 지명이면 DB에 문장만 추가하고 DB에 없는 지명이면 지명과 문장을 추가한다.

이러한 소설지명 정보 자동 갱신 과정은 〈그림 9〉에서 UML의 sequence diagram으로 표현하였다.



〈그림 9〉 신 지명DB 추가 과정의 sequential diagram

6) http://semanticweb.kaist.ac.kr/hannanum/

이 방식으로 구축된 소설 지명 DB 구축 현황은 총 575개이며 용례 문장은 5,823개이다. 다음에 설명할 인식률을 감안하면 자동 DB 갱신은 매우 실현 가능한 것으로 판단되며, DB의 지명 및 용례 엔트리수를 늘리는 방안은 향후 연구에서 다룰 것이다.

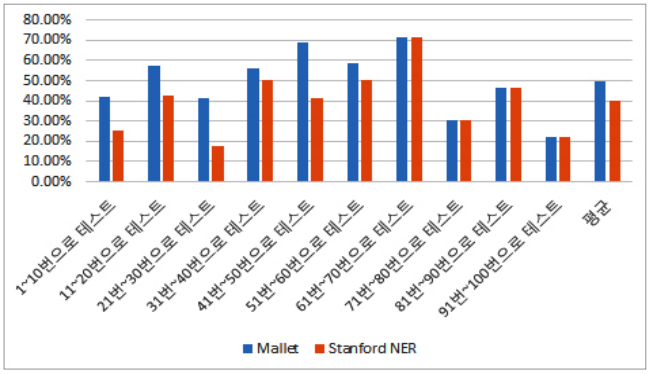
3.3.2 자동추출 실험결과

이 단계에서는 지명정보를 자동으로 추출하기 위하여 기존의 기계학습기반의 개체명인식기의 성능을 테스트하였다. 여기에 사용한 개체명인식기는 Stanford Natural Language Processing Research Group에서 만든 Stanford Named Entity Recognizer(Finkel, Grenager, & Manning, 2006)와 University of Massachussetes, Amherst 컴퓨터과학과에서 만든 MALLET Sequential Labeler이다. 이 두 개체명 인식의 성능 비교는 다음 절에서 다룬다.

1) NER 간 인식률 비교(Stanford NER vs. MALLET)

비교 데이터 셋은 지명정보를 포함한 용례를 임의로 100개를 추출하여 구성하였다. 10-fold validation을 수행하였다. 아래 히스토그램은 두 인식기의 성능을 비교한 결과이다(〈그림 10〉, 〈표 11〉 참조). MALLET의 평균 46.68%와 Stanford NER의 평균이 39.77로 MALLET이 지명 자동 추출에서 성능이 우수한 것으로 나타났다.

통계적으로 유의미한 차이를 알아보기 위해 등분산 t-검정 테스트를 수행한 결과 $p=0.097$로, 두 인식기 간에서 통계적으로 유의미하게 차이가 나지 않는 것으로 나타났다. 하지만, 소수의 test data가 적었던 점(10문장)을 감안하면 좀 더 성능의 차이가 날 수도 있을 것으로 예상되며, 이 또한 향후 연구에서 더 정밀한 평가가 필요하다.



〈그림 10〉 Stanford NER과 MALLET 인식률 비교

〈표 11〉 Stanford NER과 MALLET 인식률 비교

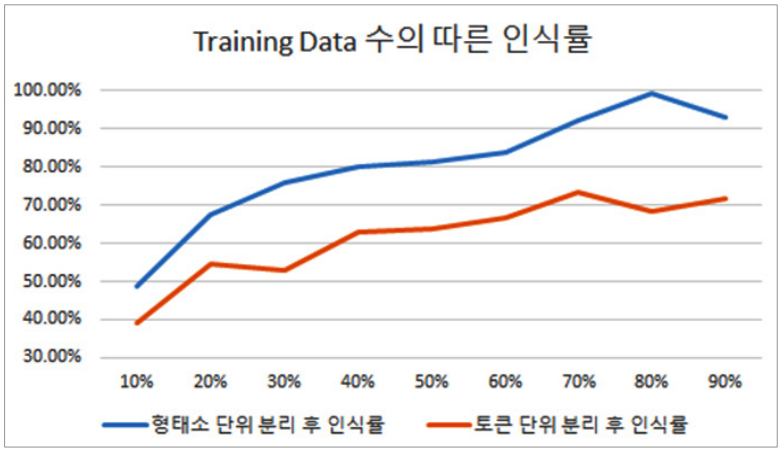
	1~10	11~20	21~30	31~40	41~50	51~60	61~70	71~80	81~90	91~100	평균
MALLET	41.66%	57.14%	41.17%	56.25%	69%	58.33%	71.42%	30.76%	46.66%	22.22%	46.68%
Stanford NER	25.00%	42.85%	17.64%	50.00%	41.17%	50.00%	71.42%	30.76%	46.66%	22.22%	39.77%

2) MALLET의 학습데이터 양에 따른 인식률

앞에서 수행한 성능비교 결과를 바탕으로, 지명 자동 추출기에 MALLET을 사용하기로 하고 MALLET의 성능최적화를 위하여 Training data 수에 따른 성능을 평가하였다. 평가 결과는 〈그림 11〉과 〈표 12〉에 보였다. 단 형태소 분석의 효과성을 검증하기 위하여 형태소 단위 분리 전 인식률과 형태소 단위 분리 후 인식률을 비교하였다. 전체문장(5,800문장)에서 10%씩 training한 뒤 나머지 문장에서 100개의 문장을 랜덤으로 뽑은 후 인식시켜 본 결과이다. Data 수가 많아질수록 인식률이 높아졌다.

두 변인들 간의 상관을 파악하기 위하여, 피어슨 상관계수를 분석한 결과, 형태소 분리 전후 모두 r= 0.92로, 일반적으로 r이 +0.7과 +1.0 사이이면, 강한 양적 선형관계를 나타낸다고 볼 수 있다. 따라서, 충분한 학습자료를 확보하면 성능을 더욱 높일 수 있을 것으로 보여진다.

형태소 분리가 인식률 향상에 영향을 미쳤는지 알아보기 위하여, 등분산 t-검정 테스트를 수행한 결과 p = 7.29325E-06(〈〈0.05)로, 두 인식률 간에 통계적으로 유의미한 차이가 현저한 것으로 나타났다. 최고 인식률은 형태소 분리 전, 후 각각 99.07%와 73.45%로 나타났다.



〈그림 11〉 Training Data 수의 따른 인식률

〈표 12〉 Training Data 수의 따른 인식률

학습데이터양	10%	20%	30%	40%	50%	60%	70%	80%	90%
형태소 단위 분리 후 인식률	48.72%	67.54%	76.03%	80.18%	81.42%	83.76%	92.04%	99.07%	93.16%
토큰 단위 분리 후 인식률	39.32%	54.39%	52.89%	63.06%	63.72%	66.67%	73.45%	68.52%	71.79%

4. 한국 소설 지명 검색 및 지도 시각화 시스템

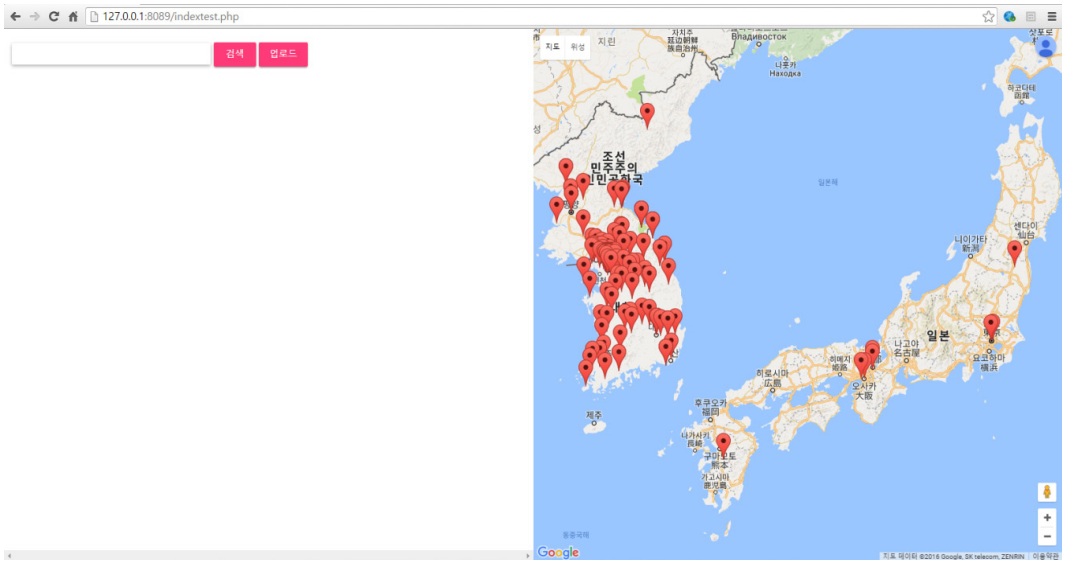
한국 지명 검색 및 시각화 시스템은 지명의 소설 속 용례를 검색하고 지도 위에 시각화하는 시스템이다. 이를 통하여 문학연구자들이나 작가들은 작품 속 지명의 장소성을 용이하게 파악할 수 있다. 먼저, 지도 시각화 과정을 살펴보자.

4.1 한국 소설 지명 지도 시각화

현재 시스템은 웹기반 서비스 시스템[7]으로 인터넷에서 접속이 가능하다. 먼저, 아래 URL로 접속하면, 현재 지명 DB에 포함된 모든 지명정보를 지도 위에 마커로 표시한 화면을 볼 수 있다. 〈그림 12〉는 시스템의 첫화면을 보여준다.

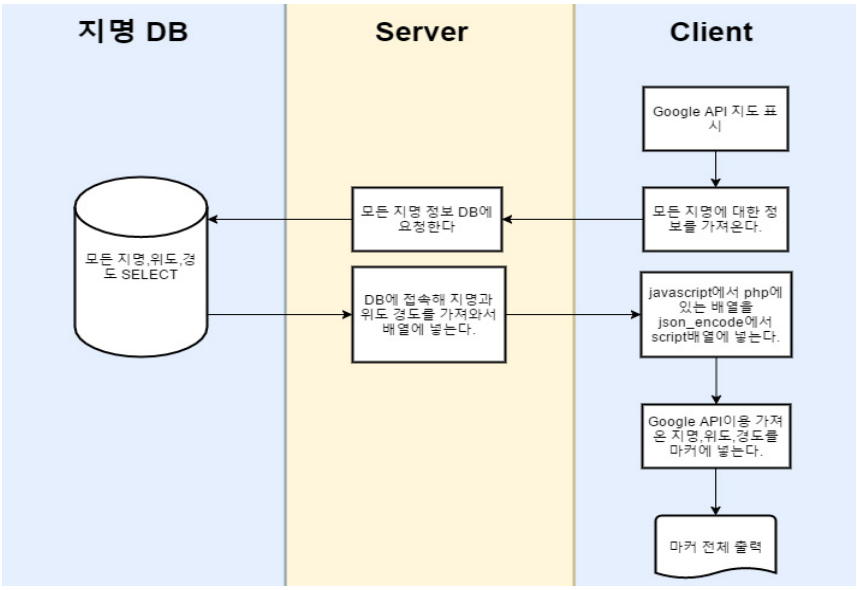
시각화 과정은 크게 세 파트로 나눠볼 수 있다. 1) 지도를 화면에 표시하는 단계, 2) 화면에 표시할 지명정보를 지명 DB로부터 가져오는 단계, 3) DB로부터 가져온 지명정보를 가지고 지도 위에 표시할 마커를 구성하여 지도 위에 표시하는 단계이다.

먼저, 지도를 표시하는 단계에서, 웹사이트 메인에서는 클라이언트에서 Google API 지도를 가져온다. 이 지도에서는 마커를 넣을 수 있는데 마커 정보(지명, 위도, 경도)는 DB에 요청해서 가져온다. 클라이언트는 DB에서 가져온 마커 정보를 지도에 표시한다. 이 과정을 설계하기 위하여 〈그림 13〉과 같은 UML의 sequence 다이어그램을 작성하였다.



〈그림 12〉 지도 시각화(첫화면)

7) http://computelligence.hnu.kr:8089/indextest.php



〈그림 13〉 첫화면의 지도 시각화 과정을 위한 sequential diagram

4.2 한국 소설 지명 및 용례 검색

'검색 기능'은 지명 DB에서 지명, 위도, 경도와 함께, 검색한 지명이 들어간 예문들도 화면에 출력해준다. 〈그림 14〉는 소설 지명을 입력으로 한 검색 결과를 보여준다. 검색 결과는 화면 왼쪽에 검색 지명을 포함한 용례를 보여주고, 오른 쪽에는 검색 키워드에 해당하는 지명의 지도 상의 위치를 마커로 시각화한다. 마커를 클릭하면, 그 지역의 정보가 시각화되고, 왼쪽 용례의 출처 링크를 클릭하면 작품 정보가 표시된다. 〈표 13〉은 검색된 용례가 어떤 포맷으로 시각화되는지를 보여준다. 이 경우, 동일한 지명을 가진 다른 지역(경기도 광주, 전라도 광주)의 용례를 따로 보여주고 있다.

검색한 지명을 쿼리에 넣어 DB에 요청하면 그 지명에 해당하는 위도, 경도, 지명과 그 지명이 들어있는 소설 속 문장(용례)들을 가져와 화면에 시각화한다. 이 과정은 〈그림 15〉에서 보여지듯이 UML의 sequence diagram을 통하여 설계하고 구현하였다.

〈표 13〉 검색 결과

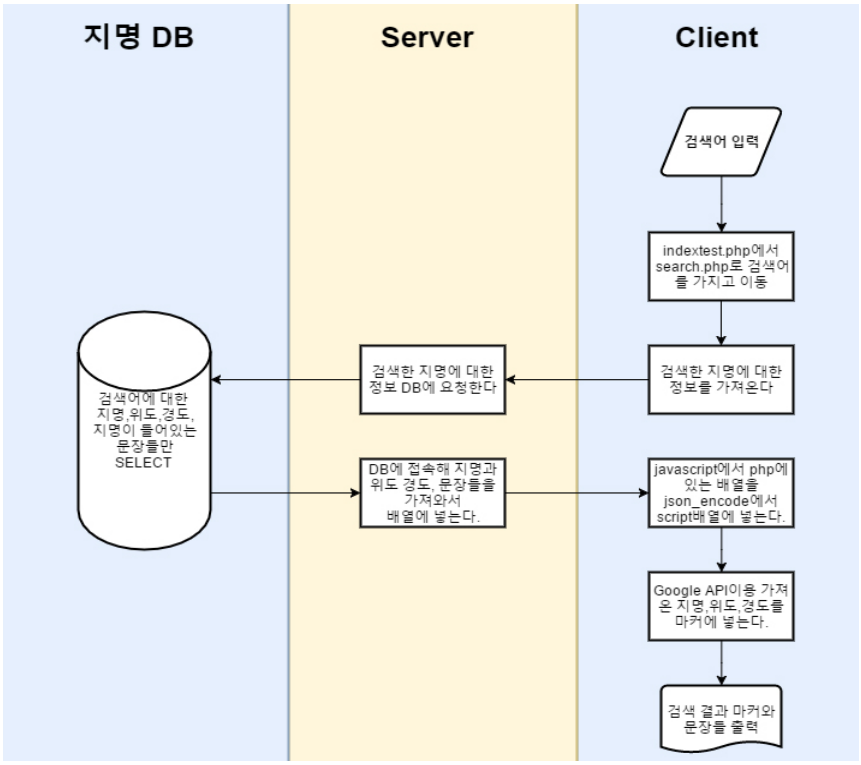
광주 37.4171 127.256 (위도, 경도)

강상촌(책정보)	광명정히 곳은 길로 갈 것 흐면음쥭으로이쳔을 지나 광쥬 길로 갈 터인
귀의성(하)	상노 아 다리고 문 절로 약이 먹으러 다 고 광쥬봉은사로 가니 그거슨왼일인고

광주 35.1595 126.853

비파성	이 길로 긔별야광쥬산셩 우리 외가로 가 잇게

〈그림 14〉 검색 결과 웹 페이지



〈그림 15〉 지명 검색시 sequence diagram

5. 결론 및 향후 연구 방향

본 연구를 통하여, 문학작품 속 지명, 특히, 한국 신소설 작품 속에 나오는 지명들에 대한 DB를 구축하고, 검색, 시각화하는 시스템을 설계 구현하였다. 이는 기존에 이미 수동으로 구축된 인터넷 지식 자원으로부터 텍스트 마이닝을 통하여 DB를 구축하고, 더 나아가 DB를 자동으로 갱신할 수 있는 자동 지명 정보 추출 기능까지 포함하여 구현하였다. 자동 추출에 필요한 개체명 인식기를 성능 비교를 통하여 선택하였으며, 학습데이터의 양에 따라 성능이 개선됨을 확인하였다. 인식률 향상을 위하여 형태소 분석 전후의 인식률을 비교하여 형태소 분리 후의 인식률이 현저히 향상되는 것을 확인하였다. 본 연구를 통하여 구축된 소설 지명 DB 구축 현황은 총 575개이며 용례 문장은 5,823개이다.

본 연구를 통해 개발된 소설지명 DB와 검색 시각화 시스템은 문학연구자들에게 소설 속 지명을 파악하는 데 임의성과 식견을 배제한 객관적 데이터 기반의 문학연구를 돕는 도구가 될 것이다. 또한, 이 연구를 통하여 한 지역이 시대적 흐름을 통해 문학작품 속에서 어떤 방식으로 묘사되어 왔는지 시계열 분석도 가능할 것이다. 데이터 분석가와 문학연구자, 지방자치단체가 함께하는 지역학 연구에 활용된다면 지역의 문화콘텐츠 발굴에도 활용될 수 있을 것이다.

향후 연구로는, 1) 디지털화된 소설 전문이 확보되면 시스템 튜닝과 함께, 소설 지명 엔트리 수를 확대하는 과제와 2) 큰 영역을 지칭하는 지명(상위지명)과 작은 영역을 지칭하는 지명(하위지명)을 구분하는 문제, 3) 현재 수동으로 진행되고 있는 소설 속 지명과 현대 지명을 매핑하는 과정을 자동으로 수행하는 연구와 4) 한국지리원 지명 DB와 통합 연계하는 방안을 모색함으로써 현재 구축된 한국소설지명DB를 양적으로 질적으로 지속적으로 개선해 나갈 필요가 있다.

참 고 문 헌

김성원, 나동렬 (2008). 2단계 최대 엔트로피 모델을 이용한 한국어 개체명 인식. 2008 한국정보과학회 학술 심포지움 논문집, 2(1), 81-86.

문상호 (2015). 엔그램뷰어를 이용한 인문학의 빅데이타 사례 연구. Asia-pacific Journal of Multimedia Services Convergent with Art, Humanities, and Sociology, 5(6), 57-65. http://dx.doi.org/10.14257/AJMAHS.2015.12.10

박철수 (2008). 문학지리학적 관점에서 본 북촌 도시한옥 밀집지역의 물리적 정체성에 관한 연구. 한국주거학회논문집, 19(2), 115-124.

이은령 (2009). 19세기 문헌 국역본의 개체명 인식 및 관계 추출을 위한 기초 연구. 언어학, 53,

141-162.

이은숙, 김일림, 정희선 (2007). 종로 문학공간의 데이터베이스 구축방안. 문화역사지리, 19, 1-14.

이창기, 김준석, 김정희, 김현기 (2014). 딥 러닝을 이용한 개체명 인식. 2014 한국정보과학회 제41회 정기총회 및 동계학술발표회, 423-425.

이창기, 장명길 (2010). Structural SVMs 및 Pegasos 알고리즘을 이용한 한국어 개체명 인식. 인지과학, 21(4), 655-667. http://dx.doi.org/10.19066/cogsci.2010.21.4.009

장노현 (2008). 소설 속 지명정보 활용 방안 기초 연구. 한민족문화연구, 24, 255-283.

장문현 (2015). 공간정보 기반의 감성문화지도 시각화 연구: 섬진강유역 역사문화유적을 대상으로. 국토지리학회지, 49(1), 27-39.

최성필 (2016). 기계 학습을 이용한 바이오 분야 학술 문헌에서의 관계 추출에 대한 실험적 연구. 한국문헌정보학회지, 50(2), 309-336. http://dx.doi.org/10.4275/kslis.2016.50.2.309

최진무, 김민준, 최돈곤 (2014). 지명 활용을 위한 지명 DB 와 수치지도 DB 의 연계 방안 연구. 대한지리학회지, 49(2), 310-319.

한순미 (2013), 소설 속 지명과 감성지도: 지명 연구와 문학 연구의 접점을 기대하며, 지명학, 19, 151-188

황이규, 윤보현 (2003). 한국어 정보처리: HMM에 기반한 한국어 개체명 인식. 정보처리학회논문지 B, b10(2), 229-236. http://dx.doi.org/10.3745/kipstb.2003.10b.2.229

Finkel, J. R., Grenager, T., & Manning, C. (2005). Incorporating non-local information into information extraction systems by Gibbs sampling. Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005), 363-370.

Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. Proceedings of the 18th International Conference on Machine Learning, 282-289.

McCallum, A. K. (2002). MALLET: A machine learning for language toolkit. Retrieved from http://mallet.cs.umass.edu

Park, S. H., Ehrich, R. W., & Fox, E. A. (2012). A hybrid two-stage approach for discipline-independent canonical representation extraction from references. Proceedings of the 12th ACM/IEEE-CS joint conference on digital libraries (JCDL '12). ACM, New York, NY, USA, 285-294. http://dx.doi.org/10.1145/2232817.2232871

Python Beautiful Soup Library (2016. 8. 21). Retrieved from https://pypi.python.org/pypi/beautifulsoup4

Standford Named Entity Recognition. Retrieved from http://nlp.stanford.edu/software/CRF-NER.shtml

• **국문 참고문헌에 대한 영문 표기**
(English translation of references written in Korean)

Choi, Jinmu, Kim, Min Jun & Choi, Don Gon (2014). Linking toponym database with digital map database. Journal of the Korean Geographical Society, 49(2), 310-319.

Choi, Sung-Pil (2016). An experimental study on the relation extraction from biomedical abstracts using machine learning. Journal of the Korean Society for Library and Information Science, 50(2), 309-336. http://dx.doi.org/10.4275/kslis.2016.50.2.309

Han, Soon-mi (2013). A place name and emotional mapping shown in novels - Looking forward to a contact between a place-name study and literary research - Journal of the Place Name Society of Korea, 19, 151-188.

Hwang, Yi-Gyu, & Yun, Bo-Hyun (2003). HMM-based Korean named entity recognition. The KIPS transactions. Part B, b10(2), 229-236.
http://dx.doi.org/10.3745/kipstb.2003.10b.2.229

Jang, Mun Hyun (2015). A study on visualization of an emotional map based on spatial information: Focused on historical and cultural heritage in Seomjin river area. The Geographical Journal of Korea, 49(1), 27-39.

Jang, No Hyun (2008). A basic study on practical use of geographical designation in Korean novel. The Review of Korean Cultural Studies, 24, 255-283.

Kim, Seong-Won, & Ra, Dong-Yul (2008). Korean named entity recognition using two-level maximum entropy model. 2008 Annual Symposium Proceedings of Korean Institute of Information Science and Engineering, 2(1), 81-86.

Lee, Changki, & Jang, Myungil (2010). Named entity recognition with structural SVMs and pegasos algorithm. Korean Journal of Cognitive Science, 21(4), 655-667.
http://dx.doi.org/10.19066/cogsci.2010.21.4.009

Lee, Changki, Kim, Junseok, Kim, Jeonghee, & Kim, Hyunki (2014). Named entity recognition using deep learning. Proceedings of Korean Institute of Information Science and Engineering, 423-425.

Lee, Eunsook, Kim, Il-Rim, & Cheong, Heesun (2007). An inquiry into the database construction of the literary space in Jongno area. Journal of Cultural and Historical Geography, 19, 1-14.

Lee, Eunryoung (2009). Named entity detection and relation extraction in the personal chronology of the 19th century. Journal of the Linguistic Society of Korean, 53, 141-162.

Moon, Sang-Ho (2015). Case study of big data in humanities using N-gram viewer. Asia-pacific

Journal of Multimedia Services Convergent with Art, Humanities, and Sociology, 5(6) December, 57-65 http://dx.doi.org/10.14257/AJMAHS.2015.12.10

Park, Cheol-Soo (2008). Physical identities of Bukchonhanok area viewed from literary geography. Journal of the Korean Housing Association, 19(2), 115-124.