구텐베르그 프로젝트 텍스트 데이터를 활용한 시각화 및 용례 검색

김동성, 신연수, 이지안, 유지민○
이화여대 인문대학
dsk202@ewha.ac.kr, dorn21@ewha.ac.kr, 1501299@ewhain.net, kekekele2007@ewhain.net

Text Visualization and Concordance Search Using Gutenberg Project Text Data

Dongsung Kim,Yeonsu Shin, Jian Lee, Jimin Yu○
College of Art & Science, Ewha Womans University

요 약

본 연구는 거시적 빅데이터 인문학과 미시적 언어 텍스트 검색 시스템을 구축하고, 이를 통해서 언어를 통한 문화의 역동적 변화를 시간적 순서에 따라 살펴보고자 한다. 연구의 최종적인 목표는 문화도 생물체처럼 변화하는 존재라 여기고 그 구성요소들을 연구한다는 뜻인 '문화체학(文化體學; Culturomics)'과 같은 '인문학 + 정보과학 + 사회과학' 등등의 다학문간의 융합적 연구에 있다. 이 시스템을 통해서 인류 역사의 기록인 텍스트 빅데이터를 통한 인문학적 성찰을 시각화하고 있다. 이러한 구글의 업적은 인문학과 정보기술의 융합을 통해서 인문학 자체의 지평을 넓히고, 사회과학을 변형시키고, 산업과 상아탑 사이의 관계를 재조정하는데 있다[1].

주제어: 구글 엔그램, 텍스트 시각화, 용례 검색, 구텐베르그 프로젝트

1. 서론

현재 빅데이터를 통한 언어 데이터의 시각화 그리고 문화의 역동적 변화에 대한 연구인 문화체학(文化體學; Culturomics)에 대한 접근은 현재 인문학과 융합된 연구들의 시대적 흐름이다. 특히 인문학 연구 관점에 볼 때 텍스트를 통한 문화 자체의 동향 및 추세(trend)를 포괄적으로 이해하는 것은 필수불가결한 요소이다.

본 연구는 거시적 빅데이터 인문학과 미시적 언어 텍스트 검색 시스템을 구축하고, 이를 통해서 언어를 통한 문화의 역동적 변화를 시간적 순서에 따라 살펴보고자 한다. 연구의 최종적인 목표는 문화도 생물체처럼 변화하는 존재라 여기고 그 구성요소들을 연구한다는 뜻인 '문화체학' 과 같은 '인문학 + 정보과학 + 사회과학' 등등의 다학문간의 융합적 연구에 있다.

2. 관련 연구

정보과학 분야에서는 인문학의 추상적 내용들은 정량화하는 여러 연구가 진행되고 있다. 그 중에서 특히 구글은 전 세계 많은 서적 텍스트를 디지털화하고 이를 기반으로 Ngram에 기반을 두고 텍스트 정보를 검색하는 Google Ngram Viewer[1] 서비스를 구축했다. 이 시스템을 통해서 인류 역사의 기록인 텍스트 빅데이터를 통한 인문학적 성찰을 시각화하고 있다. 이러한 구글의 업적은 인문학과 정보기술의 융합을 통해서 인문학 자체의 지평을 넓히고, 사회과학을 변형시키고, 산업과 상아탑 사이의 관계를 재조정하는데 있다[1]. [2]는 Google Ngram Viewer 서비스를 통한 텍스트 출현빈도에 기반을 두고 문화체학을 설명했다.

대용량 텍스트 용례 검색을 위한 시스템으로 [3], [4]는 IMS Corpus Workbench (이하 CWB)를 개발했다. CWB는 천만에서 20억 단어로 구성된 텍스트를 다양한 복잡한 조건의 검색조건에 맞춰서 검색할 수 있는 시스템이다.

이러한 시스템적 연구를 통해서 문화체학의 연구는 질적으로 국어를 대상으로 관련 시스템을 구축하고 이를 통해서 핵심어 연구[5], 특정 트렌드 연구[6], 개념어 연구[7] 등의 양적 연구를 통한 언어 내부적 표현의 변화등의 연구들로 발전하고 있다.

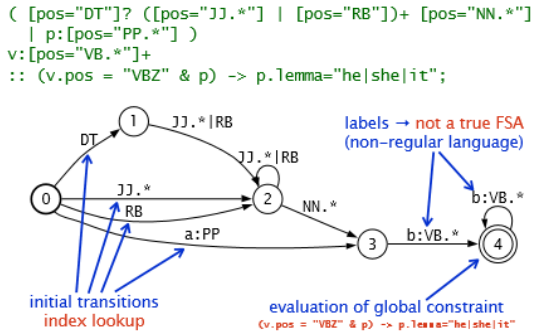
3. 시스템 구성도

CWB는 특정 구성 방식으로 텍스트 데이터를 구성하고 이를 바이너리(binary) 형식으로 변환하는 컴파일(compile) 과정을 거친다. 이러한 작업은 단어를 특정 인덱스로 변환해서 이를 활용한 효과적인 검색 방식으로 전환하기 위함이다. 특히 허프만 인코딩(Huffman Encoding)과 같이 효율적 검색 알고리즘도 구현되어 있다. 또한 인덱스를 바이너리 형식으로 전환해서 기계 친화적인 구조로 바꾸어 놓는다.

1) https://books.google.com/ngrams

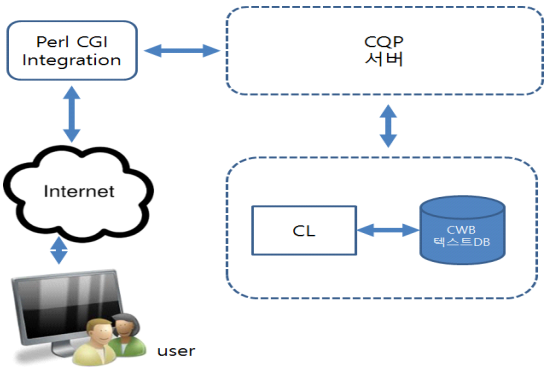
CWB의 장점은 크게 두 가지 인데, 하나는 대용량 텍스트 데이터의 효율적 처리이고, 다른 하나는 복잡한 검색 구조가 활용 가능하다는 것이다. CWB가 처리 가능한 데이터는 천만에서 20억 단어로 이루어진 대용량 텍스트이다. 또한 그림 1과 같이 정규 언어(Regular Language)가 아닌 복잡한 유한 상태 오토마타(Finite State Automata)도 처리가 가능하다.

그림 1 검색 FSA예 [4]



구축된 시스템 개요는 그림 2와 같다.

그림 2 시스템 개요도



전체 시스템은 CQP 서버에 의해서 구동되는데, CWB 형식으로 만들어진 데이터는 CL 모듈에 의해서 검색이 가능한다. 전체적으로 웹 인터페이스는 Perl 프로그래밍 언어인 Perl CGI로 구현된다.

4. 처리 과정

현재 시험 제작 원형으로 기 구축된 내용은 웹상에서 저작권이 없거나 무료인 서적들을 전자정보로 구축한 Gutenberg Project에서 전체 5만여 권 중 약 1%에 해당하는 500여권 소설 텍스트를 대상으로 용례 검색 및 연도별 시계열 그래프를 구성했다. 텍스트는 대략 5천만 어절로 구성되었으며, 텍스트에 대한 색인은 문장, 단어, 품사, 기본형이 가능하게 했다. 텍스트는 원시 코퍼스로 이를 자연어처리 시스템을 활용해서 품사, 문장 분리, 기본형 추출을 했다.[2)]

자료의 시대별 분포는 표 1과 같다.

표 1 연대별 소설 분포

연대	%
1950년대 이전	61
1950~1960	23
1970~1990	1
1990~2010	9
2010~2017	1

Gutenberg Project가 저작권이 무료이거나 없는 텍스트의 경우를 대상으로 하기 때문에 1950년대 이전 데이터가 주류를 이루고 있으며, 특히 1920년대 데이터가 가장 많다.[3)] 그러나 전체 텍스트를 대상으로 하면 더 다양한 연대의 텍스트가 수집될 것이다. 표 2와 같은 다양한 소설 장르들이 기 구축된 자료에 포함되어 있다.

표 2 소설 세부 장르별 구성

세부 장르	%	세부 장르	%
모험	2	학교	8
범죄	2	유머	9
탐정	7	영화	5
판타지	6	미스터리	0.4
고딕풍	1	과학	46
공포	5	서부극	8.8

그림 3과 같이 문장 구분을 <s>...</s> 태그로 구분하고 각 열은 텍스트에서 사용된 단어 자체, 단어의 품사, 단어의 기본형으로 구성했다. 각각은 검색은 단어, 품사, 기본형을 모두 조합해서 가능하게 했다.

그림 3 코퍼스 작성의 예

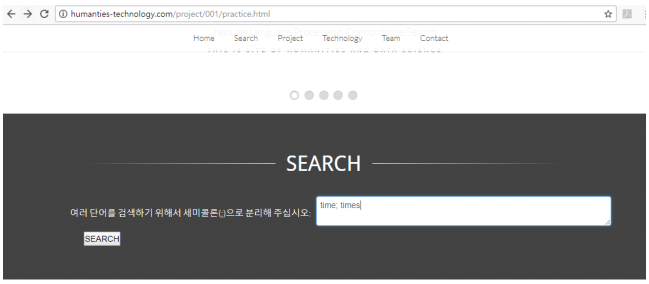
```
#     word        pos     lemma

0     A           DET     a
1     fine        ADJ     fine
2     example     NN      example
3     .           PUN     .

4     Very        ADV     very
5     fine        ADJ     fine
6     examples    NN      example
7     !           PUN     !
```

2) 사용된 자연언어 처리 시스템은 Stanford POS Tagger, NLTK sentence splitter, WordNet 등등이다.
3) 저작권 문제 등으로 인해서 1920년대 데이터에 집중되어 있다. 더 많은 데이터를 수집하면 이 문제는 해결되리라 생각된다.

5. 데모 서비스

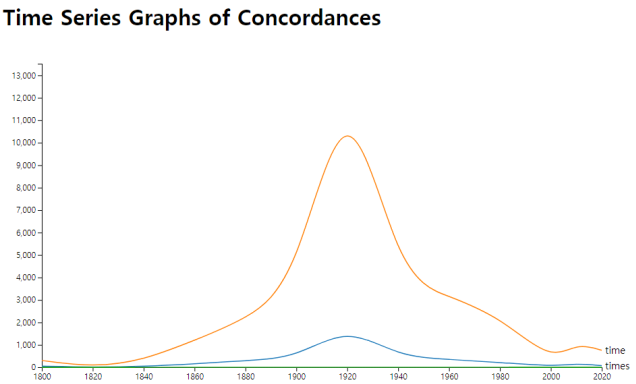
현재 그림 4와 같이 기 구축되어서 데모 웹사이트를 통해서 서비스고 있다.[4] 여러 개의 검색어를 ";(세미콜론)"으로 구분해서 입력할 수 있다. 그림 4는 "time, times"와 같이 두 개의 다른 단어를 입력한 것이다.

그림 4 두 개 검색어 입력



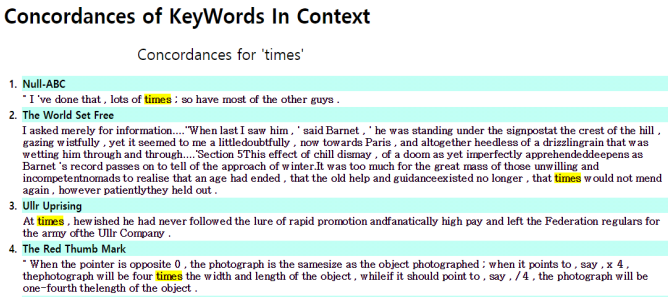
결과물은 텍스트 시각화와 용례 검색 화면에서 나탄다. 텍스트 시각화를 위해서 d3.js가 활용되었다.

그림 5 텍스트 시각화



현재 수집된 데이터는 1920년대에 집중되어 있기 때문에 1920년대에 많은 양의 데이터가 나타난다. 향후 영문 구텐베르그 프로젝트 텍스트 데이터가 활용되면 다양한 결과가 발견될 것이다. 그림 5에서 나타난 결과는 time이 times보다 10배 이상의 더 많은 용례가 발견되는 것을 보여준다.

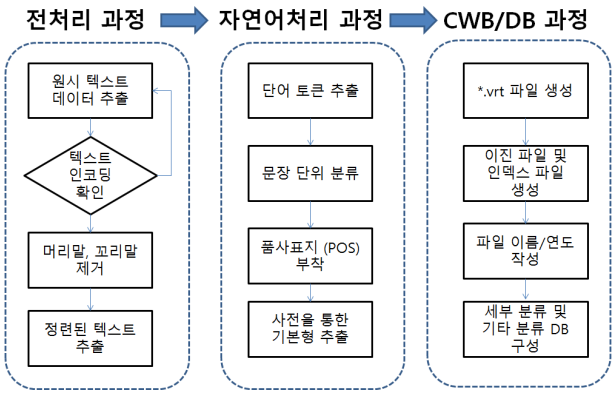
그림 6 times 용례 검색 결과



현재 데모 웹 서비스는 전체 용례 검색 결과를 보여주는 대신에 20개만 나타낸다. Php, MySQL 기반 웹 서비스가 CWB에서 개발되어 있기 때문에, 이를 적용하면 용례 결과 전체를 보여줄 것이다. 여러 용례 검색 결과 중 문장 단위 검색 결과만 보여준다. 문장 단위를 선택한 이유는 여러 검색 결과 중 가장 적절하다는 연구자들의 판단에서이다. 이 부분도 웹 서비스를 적용하면 더 다양한 검색 내용을 보여 줄 수도 있다.

Gutenberg Project에서 수집된 파일들을 정련하는 과정은 다음과 같다. 우선 텍스트 전처리(preprocess) 과정, 자연어처리 과정을 거치고, CWB 및 기타 파일 이름 및 연도, 세부 분류 DB를 구성하는 과정을 거쳐야 한다. 전처리 과정에서는 텍스트 인코딩이 무엇인지 확인하고 사용이 가능한지를 확인 작업해야 한다. 그리고 텍스트 내부의 머리말 및 꼬리말을 제거해야 한다. 파일의 메타데이터 및 법적 내용이므로 텍스트 자체와 연관성이 없는 것을 제거하기 위함이다. 다음으로는 자연처리 과정으로 단어 토큰을 분류하고 문장 단위로 텍스트가 구성되어 있지 않기 때문에 문장 단위로 텍스트를 분류한다. 이를 기반으로 품사표지를 부착하고 사전을 활용해서 기본형을 추출한다. 그림 7은 전체 작업 공정도이다.

그림 7 작업 공정도



4) http://humanties-technology.com/project/001/practice.html

6. 결론

이 연구는 언어 자료 구축과 연관되어서 구글 엔그램 뷰어와 유사한 구조인 텍스트 시각화를 보여준다. 구글 엔그램은 텍스트 용례 검색과 같이 미시적 텍스트 처리가 없는 반면에 본 시스템은 용례 검색도 가능하게 했다. 이를 위해서 CWB를 사용해서 전체 시스템을 구축했다.

향후 한국어 텍스트도 처리를 하기 위해서 노력하고 있으며, 한국어 인코딩 문제를 해결하고 있다. 또한 군집어 분석, 핵심어 추출과 같은 2단계 텍스트 분석 작업도 연구 중에 있다.

참고문헌

[1] 에이든·미셸 (2015) 빅데이터 인문학: 진격의 서막, 김재중 번역, 사계절.

[2] 문상호 (2015) 엔그램 뷰어를 이용한 인문학 빅데이터 사례 연구, Asia-pacific Journal of Multimedia Services Convergent with Art, Humanities, and Socialogy, 5(6), 57-65.

[3] Evert, S. and A. Hardie (2011) Twenty-first century corpus workbench. Proceedings of the Corpus Linguistics. Univ. of Birmingham, UK.

[4] Evert, S. (2008) Inside the IMS corpus workbench. Presentation at IULA, Univ. of Pompeu Fabra, Barcelona, Spain.

[5] 최재웅·김일환·홍정하·이도길 (2015) 핵심어로 본 시대상의 변화, 새국어생활, 26(4), 36-76.

[6] 조수곤·조재희·김성범 (2015) 텍스트마이닝을 활용한 대통령 취임 연설문의 트렌드 연구, 41(5), 435-460.

[7] 김영회·윤상길·최운호 (2011) 대한매일신보 국문논설의 언론 관련 개념 분석, 한국언론학보, 55(2), 77-102.