어문청정 빅데이터 분석: 위문기거 일례

스노우버거 다니엘 아론 · 이충호*

한밭대학교

A Big Data Analysis of Yumentingzheng: Weiwenqiju as an Example

Aaron Daniel Snowberger · Choong Ho Lee*

Hanbat National University

E-mail : aaron.snowberger@gmail.com / chlee@hanbat.ac.kr

요 약

청나라 황제가 신하들과 정사를 논한 내용을 기록한 중국의 어문청정은, 한국의 조선실록과 같은 중요한 문서이다. 본 논문은 만주글자로 쓰여진 어문청정을 빅데이터 분석하기 위한 방법과 그 단계를 기술한다. 만주글자로 씌여진 문서의 빅데이터 분석에는 사전에 해결해야 할 많은 문제가 있으며 이에 대한 연구가 선행되어야 한다. 본 논문에서는 앞으로 이루어질 사전 연구를 통하여 만주 글자로 씌여진 텍스트가 라틴문자로 전사된 단계에서, R언어를 이용하여 빅데이터 분석을 하는 방법을 제안하였다. 제안된 방법에서는 어문청정을 전사하는 방식은 압카이 방식을 채택하였고, 위문기거 부분의 텍스트를 이용하여 빅데이터 분석 결과를 제시하였다.

ABSTRACT

Yumentingzheng, which records the contents of the Qing dynasty's discussions with his subjects, is an important document like the Annals of Joseon in Korea. This paper describes the method and steps for big data analysis of Yumentingzheng written in Manchu alphabet. In big data analysis of documents written in Manchu characters, there are many problems that need to be solved in advance, and research on these should be preceded. In this paper, a method of big data analysis using the R language was proposed in the stage where the text written in Manchurian characters was transliterated into Latin characters through a preliminary study to be conducted in the future. In the proposed method, Apkai method was adopted for the transliteration of Wumentingzheng, and the results of big data analysis were presented using the text of Weiwenqiju.

키워드

Yumentingzheng, Big Data Analysis, Manchurian characters, R language, Manchu script

Ⅰ. 서 론

청나라의 어문청정(御門聽政)[1]은 한국의 조선실록에 비견할 만한 중요한 문서이다. 이것은 임금이 드나드는 어문에서 청나라 황제가 대신들과 정사를 논한 대화내용을 기록한 것이다.

청대 300여년간의 통치 중에 중국과 대만에는 만주글자[2]로 기록된 수백만권의 기록이 남아 있고, 일부는 한국, 일본에도 남아 있어 연구되고 있다. 그런데 이것을 연구하기에는 만주글자를 독해할 수 있는 인원수가 많지 않아서 사람이 일일이 독해하기에는 많은 비용과 노력이 필요하다.

이 문제를 공학적 입장에서 해결하기 위한 연구가 있어 왔다. 이와 관련된 최근의 연구는 딥러닝 기술을 적용하여 텍스트 추출과 단어 분류에 관한 [3]과 같은 연구가 있다. 하지만, 필자가 아는 한 아직까지 완벽한 글자단위 인식이 이루어지지 않는 것으로 알고 있다. 이것은 만주글자로 저장된 문자들은 대부분이 필기체로 기록된 경우가 많고, 초기 만주글자와 후기 만주글자가 다른 점이 있는 등 해결해야할 많은 문제가 있기 때문이다.

본 연구는 만주글자 자동 인식에 대한 선행 연

* corresponding author

구가 다 이루어져 있다는 가정 하에, 만주글자로 기록된 어문청정을 빅데이터 분석을 한 것이다. 이미 묄렌도르프 방식으로 전사되어 있는 것을, 압카이방식으로 변환한 다음 R언어를 이용하여 단어의 빈도수를 이용하여 빅데이터 분석을 한 것이다.

II. R언어를 이용한 어문청정 빅데이터분석

본 연구에는 R언어로 빅데이터를 분석하였다[4]. 어문청정의 텍스트를 전사한 내용은 [1]에서 그대로 가져왔다. 이것을 그대로 빅데이터 분석하기에는 용이하지 않다. 왜냐하면 이 문헌에서 전사한 방식은 묄뢴도르프 표기법으로 표기되어 알파벳 특수기호 때문에 빅데이터 분석이 용이하지 않기 때문이다. 그래서 전사방식을 특수기호가 없는 압카이(Abkai) 표기방식[5]을 채용하여 다시 바꾸었다. 이것은 한글편집기에 넣고 다음 표와 같이 일률적으로 바꾸면 된다.

묄렌도르프	압카이
š	x
c	q
ū	v

그림 1. 묄렌도르프 표기방식과 압카이 표기방식

다음으로 잘못 전사된 단어를 찾아내어 올바르게 수정하였다. 이것은 만주글자를 독해할 수 있는 능력이 필요하므로 실제로는 자동인식이 가능하도록 선행연구가 이루어져야 할 부분이다.

III. 실험결과

실험은 참고문헌 [1] 의 위문기거(慰問起居) 부분을 가지고 실험하였다. 그림 2에 만주글자로 씌여진 텍스트 중에 첫 페이지를 보였으며 그림 3은 묄렌도르프 방식으로 전사된 것이다. 이것을 압카이방식으로 전사하기 위하여 다음과 같이 하였다.

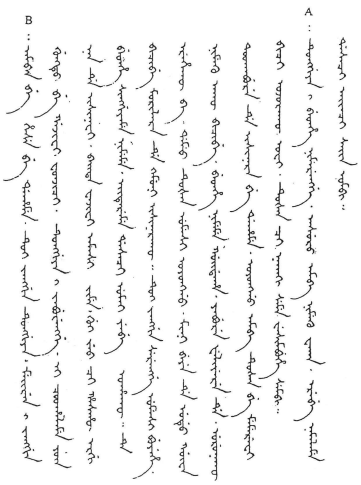

그림 2. 만주글자로 기록된 위문기기

B : amban be hese be dahame, tui janggin tunggiya, meiren i janggin gungtu be gajifi fonjici, tunggiya i jaburengge, i coohalame yūn nan de isinafi, baita wajifi .bpr gui jeo ci hashū ergi bethe singgiyame nimeme, eitereme dasaci umai yebe ohakū. te bicibe morilara de kemuni isinarakū, tui janggin serengge umesi buyehe, erehekū ba, damu tušan umesi oyonggo, mini beye de udu gūwa nimeku akū bicibe, bethe nimeme goidaha, yabure feliyere de, urunakū tookabure de isinara be dahame, oyonggo amba tušan be memerefi bici ojorakū ofi, tušan ci nakaki seme wesimbuhe sembi.

A : tunggiya i bethe nimerengge, asuru amba nimeku waka, erebe namalame dsaaci ainci sain ombi.

그림 3. 묄렌도르프 표기 방식으로 전사된 문서

먼저, 텍스트를 수기로 입력한 다음 워드프로세서를 이용하여 그림 1에 나타난 대응되는 알파벳을 일괄적으로 변환한다. 그 다음에 전사 시 두 군데의 오자가 있는 부분 '.bpr'과 'dsaaci'를 수정한 것은 그림 3과 같다.

B: amban be hese be dahame, tui janggin tunggiya,
meiren i janggin gungtu be gajifi fonjiqi, tunggiya
i jaburengge, i qoolhalame yvn nan de isinafi,
baita wajifi amasi jimen gui jeo qi hashv ergi bethe
singgiyame nimeme, eiterme dasaqi umai yebe
ohakv. te biqibe morilara de kemuni isinarakv,
tui janggin serengge umesi buyehe, erehekv ba,
damu tuxan umesi oyonggo, mini beye de udu gvwa
nimeku akv biqibe, bethe nimeme goidaha,
yabure feliyere de, urundakv tookabure de isinara
be dahame, oyonggo amba tuxan be memerefi
biqi ojorakv ofi, tuxan qi nakaki seme wesimbuhe sembi.
A: tunggiya i bethe nimerengge, asuru amba nimeku
waka, erebe namalame dasaqi ainqi sain ombi.

그림 4. 압카이방식으로 전사한 결과

같은 방식으로 위문기거 전체 텍스트를 전사한 .txt파일을 가지고 R 언어로 분석하여 워드클라우드로 표현한 것은 그림 5와 같다.

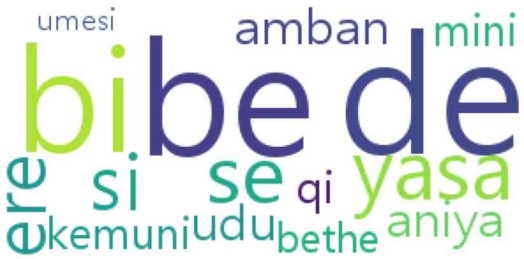

그림 5. 위문기거의 텍스트를 빅데이터 분석한 워드클라우드

그림 5에서 큰 글씨로 나타난 단어들 중에 의미가 없는 단어들을 골라내고 의미 있는 중요한 단어들로 워드클라우드를 나타내야 한다. 그림 5에서 큰 글씨로 나타난 7개 단어들의 의미는 다음과 같다.

표 1. 워드클라우드에 나타난 만주어 단어들의 의미

단어	의미
be	~을,~를, ~로써, ~로 하여금
de	~에, ~에서, ~에로, 에 대하여
amban	대신(大臣)
kemui	늘, 언제나
aniya	해, 년
yasa	눈(目)
udu	~라 할지라도

표 1에서 알 수 있는 것은 아직 만주어 전자사전이 없어서 명사를 추출할 수 없기 때문에 조사에 해당하는 단어 'be', 'de'가 가장 크게 나타나 있음을 알 수 있다. 향후 만주어 전자사전 패키지가 제공된다면 그것에 맞는 워드클라우드가 만들어 질 수 있음을 알 수 있다.

Ⅳ. 결론 및 향후 연구 계획

본 논문은 만주글자로 씌여진 문헌의 빅데이터 분석 방법을 제시하였다. 현재까지 만주어 사전 패키지가 R언어로 제공되지 않으므로 현재 상태에서 단어의 빈도수로 워드클라우드를 만들었다. 만주글자를 로마자화 하는 방법은 특수기호가 없는 압카이방식으로 변환하였다. 실제로 어문청정의 위문기거 부분을 가지고 실험을 하여 이 방법의 유효성을 보였다.

References

[1] Zhuang Jifa, Yumen tingzheng, *Wenshizhe Press*, 2000.

[2] Manchu alphabet. [Internet] Available: https://en.wikipedia.org/wiki/Manchu_alphabet

[3] Diandian Zhang, Yan Liu, Zhuowei Wang, and Depei Wang, "OCR with the Deep CNN Model for Ligature Script-Based Languages like Manchu," Hindawi Scientific Programming, vol. 2021, Article ID 5520338, https://doi.org/10.1155/2021/5520338

[4] Jang Yongsik, Kang Higu, Learning to code in R language, Saengneung Press, 2018.

[5] Manchu Language. [Internet] Available: https://namu.wiki/w/%EB%A7%8C%EC%A3%BC%EC%96%B4