고전언어에서의 어휘 의미망 구축을 위한 전문용어 추출 자동화 방안

[1] 백영윤, [2*]박용범

Automated Approaches for Extracting Specialized Terminology in Building Semantic Networks for Classical Languages

[1]Young Yun Baek, [2*]Young Bom Park

요 약

지식이나 정보를 찾는 경우 아날로그적인 인쇄된 책이나 출판물 등등의 종이로 기록된 매체보다는 디지털적으로 구현되는 웹을 이용하는 방법이 증가하고 있다. 이러한 현상은 고전적인 종이 사전 보다 디지털 사전을 사용하는 것이 더 효과적이고 시간을 절약할 수 있다는 인식이 증가되고 있다. 따라서 이러한 어휘를 구성하는 어휘 의미망 구축은 언어학계와 전산언어학, 자연어 처리 전공자들에게 있어서 중요한 문제로 떠오르고 있다. 이를 위해 언어학자들은 어휘의 의미와 개념을 구조화하여 분류할 수 있는 방법을 찾기 위해 수많은 연구가 진행되었다. 이러한 연구에서 어휘 의미망을 구성하기 위한 전문용어는 일반어와 같이 중요한 요소이다. 하지만 이러한 과정에서 전문용어를 찾고 축적하는 과정에서 여전히 종이로 된 사전 문서나 디지털로 된 방대한 자료를 사람이 직접 확인하고 그 중에서 전문용어를 추출하고 정리하는 과정을 수작업으로 거치고 있다. 본 논문에서는 이러한 인적 작업의 오류를 보완하기 위해서 디지털로 된 자료에서 사용자가 원하는 전문용어를 추출할 수 있는 자동화된 프로그램을 제안한다.

Abstract

The trend of seeking knowledge or information has been increasingly shifting towards the digital implementation on the web rather than relying on analog printed media such as books or publications. This shift is driven by the perception that using digital resources, particularly digital dictionaries, is more effective and time-saving compared to traditional paper dictionaries. Consequently, the construction of a semantic network for vocabulary has emerged as a significant issue for linguists, computational linguists, and natural language processing specialists. To address this, linguists have conducted numerous studies to find methods for structuring and classifying the meanings and concepts of vocabulary. In these studies, specialized terminology for constructing vocabulary semantic networks is as crucial as common language. However, in the process of finding and accumulating specialized terminology, there is still a manual step where individuals directly verify and extract specialized terms from paper documents or vast digital datasets. In this paper, we propose an automated program to extract the specialized terms that users desire from digital materials, aiming to compensate for errors in human-operated tasks and streamline the process.

Keywords: *Digital dictionary, Vocabulary semantic network, Natural language processing, Specialized terminology, Automation*

[1] *단국대학교 컴퓨터과학과 박사과정 (youngyun.baek@dankook.ac.kr)*
[2*] *단국대학교 컴퓨터과학과 교수 (ybpark@dankook.ac.kr)*

Received: Dec. 21, 2023, Revised: Jan. 12, 2024, Accepted: Jan. 15, 2024

I. 서론

최근의 기술의 발전과 인터넷의 발달로 인해 점차 다양한 국가와 문화의 교류가 다양해지고 이러한 문화의 교류도 다양해지고 있다. 따라서 최근에는 지식이나 정보를 찾는 경우 아날로그적인 인쇄된 책이나 출판물 등등의 종이로 기록된 매체보다는 디지털적으로 구현되는 웹을 이용하는 방법이 증가하고 있다. 모르는 문장이나 모르는 지식을 찾을 때에도 종이로 된 사전보다는 인터넷 검색 시스템을 통한 검색을 통해 디지털화된 사전을 활용하는 경우가 많아지고 있다. 이러한 현상은 고전적인 종이 사전 보다 디지털 사전을 사용하는 것이 더 효과적이고 시간을 절약할 수 있다는 인식이 증가되고 있다. 디지털 사전을 사용하여 정보에 접근하는 경우, 찾고자 하는 정보에 대한 텍스트 자료만 아니라 해당 정보에 관련되는 이미지, 동영상, 기타 관련 텍스트 내용을 제공하는 데이터 등 관련된 정보들이 종이 사전에 비해 다양하고 풍부하게 제공되기 때문이다[1][2][3].

따라서 이러한 어휘를 구성하는 어휘 의미망 구축은 언어학계와 전산언어학, 자연어 처리 전공자들에게 있어서 중요한 문제로 떠오르고 있다. 어휘 의미망 구축이 중요하게 된 이유는 어휘 의미망이 컴퓨터를 통한 자연어 처리의 기초 자원으로 사용되면서 중요해지기 시작하였다. 특히 정보 검색, 기계어 번역 등의 작업에서 어휘 의미망 자료는 중요한 지식 베이스로서 사용되고 있다. 이러한 어휘 의미망 자료를 통해서 컴퓨터는 자연어 처리를 위한 학습을 할 수 있고 이를 통해서 사용자에게 더 다양한 검색을 할 수 있게 해주었다. 기계어 번역에서 어휘 의미망은 단어 의미를 파악하는 과정에서 중의성을 해결하는 중요한 요소로서 사용된다. 이를 위해 언어학자들은 어휘의 의미와 개념을 구조화하여 분류 할 수 있는 방법을 찾기 위해 수 많은 연구가 진행되었다[4][5].

이러한 연구에서 어휘 의미망을 구성하기 위한 전문용어는 일반어와 같이 중요한 요소이다. 이를 반증하듯 전문어 사전이 아닌 일반 사전에서 각 분야의 전문용어들이 표제어로 등록되고 있다. 고전적인 종이 사전만 아니라 웹에 존재하는 디지털 사전 또한 마찬가지다. 하지만 이러한 전문용어를 어떠한 방법으로 구성할 것인가 또한 중요한 문제이다. 또한 이런 전문용어들은 시대가 지속됨에 따라서 점점 새로운 의미의 단어도 추가되고 필요에 따라서 색다른 전문용어들이 추가됨으로써 대량의 데이터가 축적되게 된다. 따라서 이러한 데이터들을 사용자가 얼마나 찾기 쉽게 검색 시스템을 구성하고, 관리하고, 보수하는가에 달려 있다[6][7].

하지만 이러한 과정에서 전문용어를 찾고 축적하는 과정에서 여전히 종이로 된 사전 문서나 디지털로 된 방대한 자료를 사람이 직접 확인하고 그 중에서 전문용어를 추출하고 정리하는 과정을 수작업으로 거치고 있다. 이러한 과정은 인적 작업이므로 작업을 하는 중에 인적 오류가 발생하여서 다른 의미의 데이터로 저장을 할 수 있고 또한 데이터를 수집하고 정리하는 시간과 비용이 발생할 수 있다. 이는 데이터의 수가 많으면 많을수록 비례하여 증가하기 때문에 그만큼의 손해가 발생하게 된다.

본 논문에서는 이러한 인적 작업의 오류를 보완하기 위해서 디지털로 된 자료에서 사용자가 원하는 전문용어를 추출할 수 있는 자동화된 프로그램을 제안한다. 제안하는 프로그램을 통해서 사용자는 분석하고자 하는 디지털 자료를 입력하고 프로그램이 분류한 분류 체계를 통한 전문용어 선택을 통해서 자동으로 전문용어를 저장하고 관리할 수 있다.

II. 관련 연구

J. H. Koo et al [1] 에서는 특정 정보 요소를 지정하고 해당하는 정보 요소를 문장에서 추출하여서 XML 로 구성하여서 사용자에게 제공하는 방법으로 구성하였다. 해당 XML 은 관계 구조로 구성되어서 각각의 요소들이 관계되는 요소들에 대해 쉽게 파악할 수 있는 장점이 있으나 이를 구성하기 위해서는 수동으로 각 문장에서 해당 요소를 찾아서 XML 에 입력해야 하는 단점이 있다.

B. K. Kang [4] 에서는 언어 별로 제공되는 사전 자료를 활용하여서 자동으로 어휘 의미망을 구성하는 작업을 제안하였다. 자동으로 어휘 의미망을 구성할 수 있다는 장점이 있지만 단어에

해당하는 어휘만 분석을 할 수 있다는 단점이 있으며 사전에 존재하지 않는 단어에 대해서는 분석을 할 수 없다는 단점이 있다.

Y. S. Bang [6] 에서는 전문용어를 선정하는 방법에 대한 방법론은 제안하였으며 이를 통해 선택된 전문용어들이 다양한 사용자들이 사용할 수 있는 전문용어집으로 사용할 수 있도록 구성하였다. 다양한 사용자들이 구성된 전문용어를 통해서 사용할 수 있다는 장점이 있으나 자동화를 지원하지 않기 때문에 수작업으로 전문용어를 추출해야하는 단점이 있다.

E. J. Kwon [7] 에서는 역사 관점에서의 전문용어 추출을 위해서 전문용어를 추출하는 방법을 제안하였다. 이를 통해서 역사 관점에서의 전문용어를 구성하였지만 해당 전문용어들은 역사 용어로만 사용할 수 있다는 단점이 있다.

Y. J. Yoo [8] 에서는 시소러스 분류 체계를 보완하기 위해서 전문용어 추출을 통해서 분류체계를 보안하는 방법을 제안하였다. 이를 통해서 용어 분류에 있어서 유연성과 특징성을 가지는 분류 체계를 구성하는 장점이 있으나 모든 작업을 수작업으로 하는 단점이 있다. 위와 같은 연구들의 특징을 정리한 내용은 표 1 과 같다.

Table 1. Comparison of Characteristics in Related Research
표 1. 관련 연구 특징 비교

Name	Usability	Versatility	Automation
J. H. Koo et al [1]	√	√	
B. K. Kang [4]		√	√
Y. S. Bang [6]	√	√	
E. J. Kwon [7]	√		
Y. J. Yoo [8]	√		

III. 전문용어 추출 자동화 방안

본 논문에서 제안하는 전문용어 추출 자동화 프로그램은 그림 1 과 같이 구성된다. 사용자는 고전언어를 선택하고 분석하고자 하는 전문용어를 선택한다. 전문용어 추출기는 입력된 고전언어와 전문용어를 확인하고 룰 베이스 기반의 점수를 부여하여 최종적인 전문용어를 선택한다. 전문용어 추출기는 최종적인 결과를 엑셀 파일로 사용자에게 제공하여서 쉽게 추출된 전문용어를 확인 할 수 있도록 하였다.

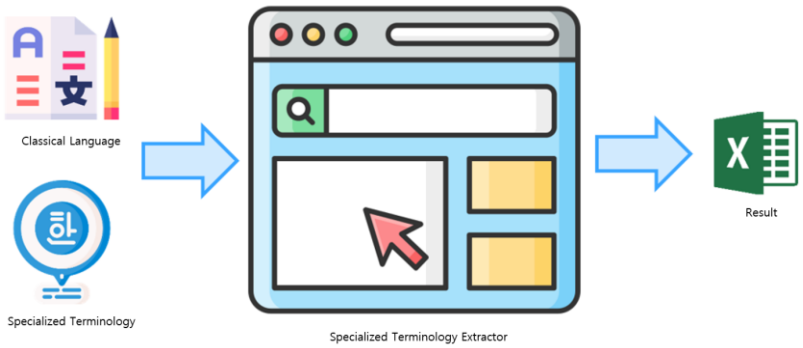

Figure 1. Automated Specialized Terminology Extraction Program Overview
그림 1. 전문용어 추출 자동화 프로그램 요약

사용자는 해당 화면에서 분석하고자 하는 디지털로 된 자료와 사용자가 전문용어로 추출하기를 원하는 사전 데이터를 각각 Analysis File, Index File 로 입력한다.

사용자가 파일을 입력하고 분석 버튼을 선택하면 시스템은 디지털로 된 자료와 사용자의 사전 데이터를 비교하여서 전문용어로 선정될 수 있는 단어들을 분류한다. 분류가 완료되면 하나의 문장을 기준으로 하여 해당하는 문장과 단어들을 표시해준다. 해당 결과는 그림 2 와 같이 구성된다.

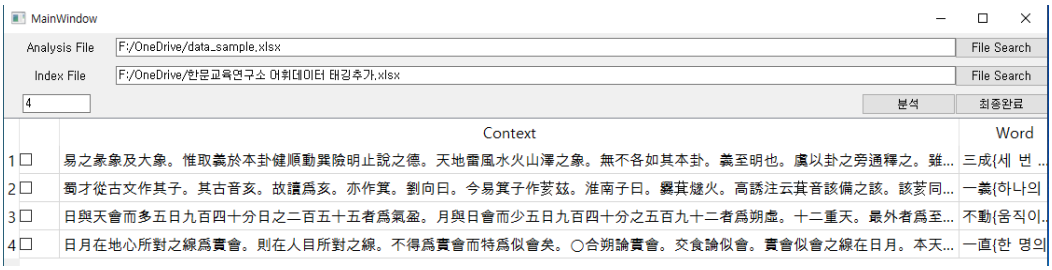

Figure 2. Automated Specialized Terminology Extraction Program
그림 2. 입력 데이터 분석 결과

사용자는 분류된 결과를 확인하고 문장에서 전문용어를 추출하기 위해서 분석을 원하는 문장을 선택한다. 시스템은 사용자가 문장을 선택하면 해당하는 문장에서 전문용어를 추출할 수 있는 화면을 제공한다. 해당 화면은 그림 3 과 같다.

사용자는 해당 화면에서 사용자가 선택한 문장과 방점을 기준으로 분리한 전체 문장을 볼 수 있다. 또한 방점을 기준으로 한 문장씩 분석을 할 수 있도록 현재 분석되고 있는 문장인 확인 문장을 확인할 수 있고 해당 확인 문장에서 전문용어로 선택될 수 있는 후보들인 출현 단어들이 표시되게 된다. 출현 단어의 경우 사용자의 사전 데이터의 단어들 중 룰 베이스 기반의 점수 배분을 통해서 점수가 높은 순서대로 표시된다. 룰 베이스 기준은 다음과 같다.

- 확인 문장에서 도출될 수 있는 단어 중 사전 데이터에 단어가 있다면 점수를 얻는다.
- 추출되는 단어 중 더 긴 단어가 높은 점수를 받는다.
- 추출되는 단어가 동일한 길이라면 사전 데이터에서 더 먼저 기록된 단어가 높은 점수를 받는다.
- 추출되는 단어가 동일한 단어라면 필터 단어의 사전 순으로 빠른 단어가 높은 점수를 받는다.
- 만약, 단어 필터가 지정되면 필터에 포함되지 않는 단어는 점수를 모두 잃는다.

위의 룰 베이스를 기준으로 점수를 배분하고 배분된 점수에 따라서 출현 단어의 순서가 지정된다. 전문용어로써 추출하고 싶은 단어가 있다면 출현 단어에서 선택 단어로 단어를 드래그 앤 드롭하면 된다.

예를 들어 “易之彖象及大象”라는 분석 문장이 존재할 때, 해당하는 전문용어는 “易之{자(字).}”, “易之{당(唐) 이이간(李夷簡)의 자.}”, “易之{원(元) 내현(迺賢)의 자.}”, “大象{북주(北周) 정제(靜帝:宇文衍)의 연호(579 ~ 580).}”가 존재하게 된다. 해당되는 전문용어는 위에서 작성된 룰 베이스 점수에 기반하여서 각 단어들 별로 점수를 얻게 되고 이를 기반으로 하여서 UI 에 표시되게 된다.

단어 선택이 완료되면 다음 문장 버튼을 통해서 다음 문장으로 전환할 수 있다. 만약 선택 단어를 잘못 선택하였다면 단어 리셋 버튼을 통해서 다시 단어 선택을 수행할 수 있으며, 다음 문장으로 넘어간 상태에서 이전문장으로 전환하고 싶으면 이전 문장 버튼을 선택하면 된다.

위와 같은 과정을 반복하여서 수행하여 전체 문장을 모두 분석을 수행하였다면 최종 확인 버튼을 선택하여서 그동안 분석한 전문용어 단어들을 저장한다. 저장이 완료되면 그림 2 의 화면에서 다시 분석하고자 하는 문장을 반복해서 선택하고 위와 같은 과정을 반복해서 수행한다. 모든 문장에 대한 분석이 완료되면 그림 2 의 최종 완료 버튼을 선택하여 작업을 종료한다. 최종 완료 버튼을 선택하면 시스템은 사용자가 그동안 분석한 단어를 정리하여 파일로 구성하여 사용자에게 제공한다. 최종 정리된 파일은 그림 4 와 같다.

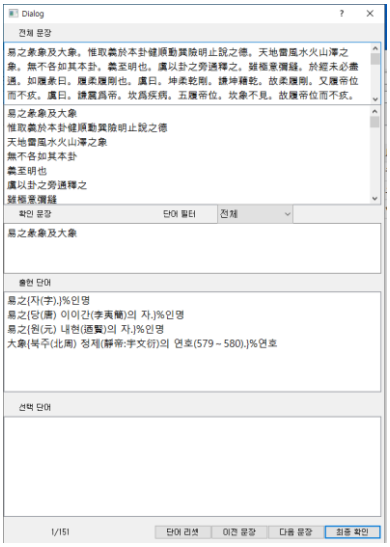

Figure 3. Specialized Terminology Extraction Interface
그림 3. 전문용어 추출 화면

	A	B	C	D	E	F	G	H
1	ID	TITLE	TEXT	RESULT	WORD			
	02_0001	周易疏義攷	易之象象及大象。惟取義於本卦健順動巽險明止說之德。天地雷風水火山澤之象。無不各如其本卦。義至明也。虞以卦之旁通釋之。雖極意彌縫。於經未必盡通。如履彖曰。履柔履剛也。虞曰。坤柔乾剛。謙坤雜乾。故柔履剛。又履帝位而不疚。虞曰。謙震爲帝。坎爲疾病。五履帝位。坎象不見。故履帝位而不疚。此謂履與謙通。謙上體有坤。互體有震	[易之{자(字	易之{자(字)%인명大象{%연			

Figure 4. Final Result
그림 4. 최종 결과

IV. 결론

디지털로 된 자료에서 사용자가 지정한 사전 데이터를 바탕으로 전문용어를 추출할 수 있는 자동화된 시스템을 구성하였다. 이 시스템을 통해서 사용자가 지정된 사전 데이터를 기반으로 전문용어를 추출하기 위한 필터 단어를 구성하고 디지털로 된 자료에 적용하였다. 필터 단어를 통해서 시스템은 사용자에게 전문용어로 선정할 수 있는 단어 후보를 제공하고 이를 사용자가 최종 선택하여서 전문용어를 선택하는 과정을 거치게 된다. 최종적인 선택이 완료되면 시스템은 그 동안 사용자가 선택한 전문용어를 정리하여 파일로서 제공하여 사용자가 분류한 전문용어를 확인할 수 있도록 하였다. 논문에서 제시한 시스템을 통해서 사용자는 시스템이 제안하는 단어를 선택함으로써 자동으로 디지털로 된 자료에서 사용자가 원하는 전문용어들을 추출할 수 있게 된다.

V. 감사의 글

이 논문은 2021 년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임(NRF-2021S1A5C2A02086984)

VI. 참고문헌

[1] J. H. Koo, Y. S. Kim, “A Study on the Development of an XML Data Model for Digitalized Dictionary of Terms : - The Case of Korean-Japanese Dictionary of Diplomatic Terms in the Joseon Dynasty –,” The Journal of Korean Studies, 2014, pp. 7-31.
[2] H. O. Son, D. M. Kim, M. H. Cha, W. J. Kim. “A Study on the Characteristics of Headwords in Chinese Dictionaries for a Development of a Classical Chinese Lexical Semantic Network,” Journal of the Oriental Studies , Dong Yang Hak, 2022, pp. 37-56
[3] E. H. Bae, Y. B. Park, C. Heo “Research on the Necessity and Prerequisite Problems which Chinese Classical written language works in Machine translation,” Journal of the The Association Of Korean Literature In Chinese, 2019, pp. 39-54
[4] B. K. Kang, “A Study on the Construction of Korean-Chinese-Japanese-English Multi-Lingual WordNet,” JOURNAL OF CHINESE LANGUAGE AND LITERATURE, 2007, pp. 107-132
[5] K. B. Choi “The Ontology of "Mulmyeonggo" and Its Significance in Lexicography,” Journal of the The Society Of Korean Semantics, 2005, pp. 21-42
[6] Y. S. Bang, “On the Selection and Processing of Specialized Terms in the 'Open Dictionary': Focusing on Economic Terminology,” Journal of Korealex, 2013, pp. 69-80.
[7] E. J. Kwon, “A Construction of Historical Terminology for Terminology in Gaebanghyeong-Hangugeo-Jisik-Daesajeon,” Journal of Korealex, 2012, pp. 31-51.
[8] Y. J. Yoo “A Study on Classification System of Korean Literatures Thesaurus,” Journal of the Korean Society for Library and Information Science, 2006, pp. 415-434

저자소개

백영윤(*Young Yun Baek*)

2014 년 3 월 단국대학교 대학원 컴퓨터과학 석사
2016 년 3 월 ~ 현재 단국대학교 대학원 컴퓨터과학 박사과정

관심분야 : 인공지능, NLP, ChatGPT

박용범(*Young Bom Park*)

1987 년 3 월 Polytechnic Institute 대학교 전자계산학과(공학석사)
1991 년 3 월 Polytechnic Institute 대학교 전자계산학과 지능형 SE(공학박사)
1993 년 3 월~현재 단국대학교 교수

관심분야 : Information Architecture, Narrative Contents & Technology, Intelligent Software Engineering, 패턴인식, 의미 정보 추출