J Korean Soc Phys Med, 2023; 18(1): 77-86
https://doi.org/10.13066/kspm.2023.18.1.77

Online ISSN: 2287-7215
Print ISSN: 1975-311X

Research Article

Open Access

자연어 알고리즘을 활용한 한국표준건강분류(KCF) 코드 검색

최년식 · 송주민†
신라대학교 물리치료학과

Korean Standard Classification of Functioning, Disability and Health (KCF) Code Linking on Natural Language with Extract Algorithm

Nyeon-Sik Choi, PhD · Ju-Min Song, PT, PhD†
Department of Mechanical Convergence Engineering, Silla University

Received: January 30 2023 / Revised: January 31 2023 / Accepted: February 9 2023

| Abstract |

PURPOSE: This study developed an experimental algorithm, which is similar or identical to semantic linking for KCF codes, even if it converted existing semantic code linking methods to morphological code extraction methods. The purpose of this study was to verify the applicability of the system.

METHODS: An experimental algorithm was developed as a morphological extraction method using code-specific words in the KCF code descriptions. The algorithm was designed in five stages that extracted KCF code using natural language paragraphs. For verification, 80 clinical natural language experimental cases were defined. Data acquisition for the study was conducted with the deliberation and approval of the bioethics committee of the relevant institution. Each case was linked by experts and was extracted through the System. The linking accuracy index model was used to compare the KCF code linking by experts with those extracted from the system.

RESULTS: The accuracy was checked using the linking accuracy index model for each case. The analysis was divided into five sections using the accuracy range. The section with less than 25% was compared; the first experimental accuracy was 61.24%. In the second, the accuracy was 42.50%. The accuracy was improved to 30.59% in the section by only a weight adjustment. The accuracy can be improved by adjusting several independent variables applied to the system.

CONCLUSION: This paper suggested and verified a way to easily extract and utilize KCF codes even if they are not experts. KCF requires the system for utilization, and additional study will be needed.

Key Words: Code extraction algorithm, Code search, KCF, Linking, Natural language

†Corresponding Author : Ju-Min Song
jmsong@silla.ac.kr, http://orcid.org/0000-0001-8469-3550

Ⅰ. 서 론

2001년 세계 보건 기구(WHO)에서는 국제표준분류 중의 하나로 건강과 건강관련 상태를 표현하기 위한 통일되고 표준화된 언어를 제공하기 위해 국제표준기능건강장애분류 (International Classification of Function, Disability and Health, ICF)를 발표하였다[1].

ICF는 과거 단순히 의학적이나 생물학적인 측면에서만 장애를 바라보던 것에서 사회적, 물리적 환경 요인이 결합된 새로운 장애개념을 도입하고, 건강과 관련된 광범위한 정보를 구분하는 형태로 구성되었다. ICF가 표준화된 코드 체계로 제공됨에 따라 건강 및 건강관련, 다양한 기준과 분야 간에 원활한 의사소통이 가능하게 되었다. 한국에서는 ICF를 한국어로 번역하고 문화에 적용시킨 한국표준건강분류 (Korean Standard Classification of Functioning, Disability and Health, KCF)가 발표되기에 이르렀다[2].

ICF는 인구조사와 같은 통계도구, 삶의 질과 인간개발의 통합 또는 환경요인의 측정과 같은 연구도구, 임상실습 적용 장애와 재활, 재활과 중재효과 평가 등으로 활용할 수 있는 임상도구, 퇴역군인 재사회화, 사회보장제도 및 정책 설계와 같은 사회정책도구 또한 교육과정계발과 같은 교육도구로써 사용될 수 있다. 이러한 의미에서 ICF와 이를 기반으로 하는 KCF는 장애가 있는 사람뿐만 아니라 모든 사람을 위해 사용될 수 있는 보편성을 가진 분류체계라 할 수 있다[3-8].

2001년부터 ICF 개념을 다양한 영역에서 소개하는 것을 시작으로 ICF 개념을 사용하는 방법에 대한 연구가 발표되었다[9]. 국가별로 다양한 분야에서 다양한 용도로 사용되기도 한다. 특히 ICF의 개념과 시스템 설명, 임상 애플리케이션, 기존 도구와의 연결 및 데이터 수집 분석에 대한 연구가 임상 영역에서 활발히 수행되고 있으며 ICF의 사용은 증가 추세이다[10].

한국어로 번역된 KCF에 IT 기술을 적용하여 KCF의 자연어 기반 코딩을 데이터베이스를 구축하고, 이를 기반으로 정해진 데이터 영역내에서 상호 작용을 활용하는 데이터 마이닝 기법과 인공지능을 활용하는 기술은 쉽고 편리하게 KCF를 활용할 수 있는 프로그램을 구현할 수 있는 기술로 사료된다. 데이터 마이닝은 크고 복잡한 데이터 세트에서 구조와 패턴을 발견하는 기술로 모델 구축과 패턴 감지가 있다[11]. 모델 구축은 통계 모델링과 유사하며 패턴 감지는 알고리즘을 사용하므로 대규모 데이터베이스에서 정보와 지식을 추출해 내는 것은 많은 연구자들에 의해 데이터베이스 시스템과 기계 학습의 핵심 연구 주제로 인식되어 왔다[12]. 대규모 검색 시스템에서 상대적 가중치를 부여해서 중요도를 측정하는 기술인 PageRank 알고리즘은 연구를 통해 확인된 데이터베이스와 구조를 IT 시스템과 결합하는 방안으로 KCF 코딩의 중요도 계산에 적용될 수 있다. 인공지능은 인간의 지식 체계 하에서 행동하는 것과 유사하게 기계가 인간의 행동을 모방하도록 프로그래밍 되어 작동하는 것을 기반으로 한다[13]. 컴퓨터 과학을 통한 알고리즘 접근, 탐색적 추론, 자연어 처리, 마이크로월드 등 기술 발전이 가속화되고 있다[14-16]. 본 연구에서는 다루는 자연어란 우리가 일상 생활에서 사용하는 언어를 말하며, 자연어 처리(Natural Language Processing, NLP)란 이러한 자연어의 의미를 분석하여 컴퓨터가 처리할 수 있도록 하는 것이다. 자연어 처리에 대한 연구와 텍스트 처리, 구문 분석 및 의미 분석에 대한 연구가 수행되었다[17,18].

KCF가 추구하는 다양한 기준과 분야 간의 원활한 의사소통을 위해서는 KCF를 이해하고 전문적으로 다룰 줄 알아야만 정확한 코드의 추출이 가능하다. 이런 관점에서 각 분야별 전문가들은 KCF의 사용에 한계를 느끼고 있으며, 누구나 쉽게 코드를 검색해서 도구에 접근할 수 있는 방법이 필요하다. IT 기술은 과거에 보지 못했던 다양한 애플리케이션이 새로운 형태의 산업군을 형성하고 있다[19]. 이에 IT기술을 활용하는 방안을 제시한다.

본 연구에서 언급하고자 하는 가장 큰 핵심은 KCF 코드 추출 방법론의 변화에 있다. 기존의 의미적 코드 추출 방법을 KCF 코드체계와 코드의 설명 등에 나타나 있는 단어 형태를 이용하는 형태적 코드 추출 방법으로 변환해서도 그 의미적 추출 방법과 유사 또는 동일한 코드 추출이 시스템을 통해 이루어지게 하는 것으로 자연어 처리를 통한 코딩과 추출(Linking)을 통해 달성

할 수 있다[20]. KCF 활용 측면에서 일상언어 및 자연어에 의한 코딩과 추출을 지원하는 실험 시스템을 구축하고 이를 검증하는 것이 목적이다.

Ⅱ. 연구방법

1. KCF 코드별 단어 분석과 알고리즘

본 연구 목적 달성을 위한 형태적 추출 방법을 정의하기 위해서는 KCF의 코드 별 정의 및 설명에 대한 단어들의 형태를 분석해야 한다. 단어 단위로 형태를 나열하여 단어의 검색 빈도가 어느 정도인지에 따라 코드 적합성이 나타날 수 있다.

KCF 코드를 시스템을 통해 추출하고 이를 검증하기 위하여 KCF 코드 추출 알고리즘이 포함된 실험적 시스템을 구축하였다. 알고리즘을 통한 KCF 코드 추출 예제로 “나는 각 일들의 개념과 특성을 잘 이해하므로 계획수립을 통해 실행 가능한 일들을 구체화시키고 이를 실행하기 위해 행동한다”라는 일상용어에 대하여 시스템 내부에서는 단어를 분석하고 노출 빈도에 따라 코드를 결정하게 된다.

Fig. 1은 KCF 코드에 대하여 그 의미를 판단할 수 있는 단어들을 나열한 것이다. 형태적 추출 방법은 단어별로 검색이 이루어지지만, 해당 단어가 포함하고 있는 의미들이 복합적으로 작용하는 것이므로 각 코드별 정의와 설명을 구성하는 단어는 의미가 있다. b164는 20가지의 단어로 그 의미를 파악해서 코드를 구분할 수 있다. b1640의 경우는 b164가 가지고 있는 단어 20가지와 고유한 9개의 단어로 그 의미가 표시되었다. KCF 코드 별 단어화에 대한 의미는 일상어를 통해 자주 언급되어지는 단어에 대하여 KCF 코드 별 단어가 얼마나 많이 노출되는가에 따라 일상어의 의미가 확률적으로 해당 코드일 가능성이 높다는 것이다. 알고리즘에 의해 b164는 3개의 단어가 노출되었다고 판단되었고, b1640는 6개의 단어가 노출되었다고 판단되었으므로 b1640가 더 적합하다는 것을 시스템은 알려 주고 있다.

또 다른 예로 통증 설문지 중 “당신은 복합적 사고를 할 수 있고, 침대에서 일어나거나 이를 닦거나 옷을 입는 등의 생활을 할 때 통증이 얼마나 개인 관리에 지장을 줍니까?”라는 일상적 언어의 표현에 대하여 Cieza 등[21]은 b280(통증감각), d5(자기관리) 외에 추가적으로 d4100(눕기), d5201(치아 관리), d540(몸단장)와 연결이 된다고 하였다. 이를 시스템을 통해 확인한 바로 b280(통증)외에 s320(입구조), d299(상세불명의 일반적 과제와 요구), d540(몸단장) 코드가 추출되었다. 여기서 가장 핵심인 b280(통증)이 추출된 점과 자연어에서 비롯된 입의 구조, 일반적 과제 및 요구 및 옷 입고 벗기의 내용이 예로 살펴본 내용과 의미적 유사점이 있다는 것을 파악할 수 있다. 의미적 코드 추출 방법을 통해서 제시된 KCF코드가 형태적 코드 추출 방법을 통해서도 가능하다는 것을 확인할 수 있다.

자연어 문장을 통해 KCF 코드를 추출해 내는 알고리즘은 총 5단계로 나뉘어 설계되었다. 가장 먼저 자연어

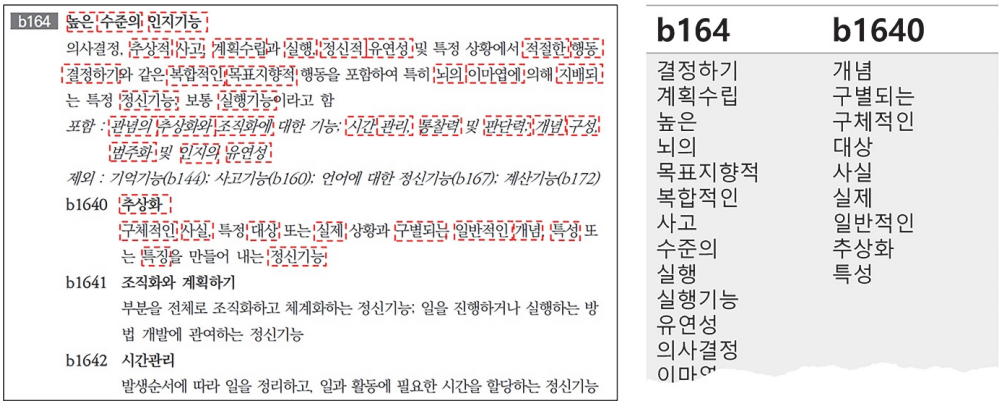


Fig. 1. Wordization for KCF code using the morphological extraction method.

문장을 단어화(Word graph)하는 작업을 진행한다. 문장의 단어화를 위해서는 임상 용어와 한글 사전의 용어 데이터베이스를 통해 단어를 추출하게 된다. 다음으로 대표단어를 추출하여 대표단어들 만으로 문장화(Sentence graph)를 진행한다. 대표단어는 대부분 명사와 동사로 구성되어 있는데, 형용사로 분류된 단어 중 명사화 가능한 경우는 명사화 시켜 처리하고 순수 형용사의 경우는 실험적 시스템에서는 참조하지 않았다. 대표단어는 시스템의 학습에 의해 대표단어 데이터베이스를 꾸준히 작성한다. 다음으로 대표단어 별 중요도(Weight)를 부여하게 되는데 의학용어는 일반용어보다 중요도가 높게 설정되어 있다. 대표단어와 KCF 코드에 대한 연결을 시스템에서 진행하고 이를 정량화된 수치로 변환한다. 이때 PageRank기법이 사용된다[22]. PageRank 기법에 의해 노출이 많은 순서로 나열하여 형태적 코드 추출 방법이 완료된다.

시스템의 구축에 필요한 KCF용어 데이터베이스(1,587레코드), 의학용어 데이터베이스(1,414레코드), 임상자연어 데이터베이스(73,308레코드), 대표단어 데이터베이스(1,787레코드)가 구축되고 활용되었다.

2. 데이터의 수집과 연구 모델

연구를 위한 데이터 수집은 임상 자연어 수집 기관을 모집하고, 교육한 뒤 관련 기관의 생명윤리위원회의 심의 및 승인(승인번호 1041449-202006-HR-009)을 득한 후 진행되었다. 임상 자연어 수집을 위해 참여한 연구 대상자는 언제라도 원치 않을 경우 연구 참여를 중단할 수 있었으며, 연구 대상자용 설명문을 제공하였고, 동의서를 작성한 후 연구에 참여하였다.

참여 기관은 총 17개로 서울경기 8개 기관, 부산경남 2개, 대구경북 6개, 강원 1개 기관이 임상 자연어 수집에 참여했고, 영역별로는 물리치료 7개, 작업치료 3개, 언어치료 7개로 구성되었다.

임상 자연어 수집에 참여한 대상자는 145명으로 남자 75명, 여자 70명이며, 환자 당사자가 대상인 경우는 101명이고, 보호자가 대상자인 경우는 44명이다. 대상자의 질환은 뇌졸중, 척수손상 등의 신경계 환자가 47명이었고 허리통증, 어깨통증 골절 등의 근골격계 환자가 43명이었다. 또한 유창성장애, 청각장애, 지적장애 등 언어치료 대상자가 55명 참여하였다.

임상 자연어 수집 방법은 상담을 녹음하여 145개의 음성파일로 수집하였다. 전문가와 연구대상자의 중앙에 녹음기를 놓고 녹음을 진행하였고, 질문이 완전히 끝난 후에 대답하고 대답이 완전히 끝난 후에 질문을 이어 나가도록 하였다. 이로 인해 환자의 음성과 치료사나 평가자의 목소리가 모두 녹음되었다. 수집된 음성파일은 텍스트 파일로 변환되어 임상 자연어 샘플로 활용되었다. 음성파일을 텍스트로 변환하기 위해 음성텍스트 변환 프로그램(다글로)을 이용하였다. 임상 자연어 샘플은 5분에서 20분의 음성파일로 텍스트로 변환 후 확인한 결과 샘플은 500개에서 2,000개의 단어가 사용되었다.

형태적 코드 추출 방법이 얼마나 의미적 코드 추출 방법과 일치할 수 있는지를 파악하기 위해 145개의 임상 자연어 샘플을 10가지 상황에 대하여 단락별로 쪼개어 3,689개의 실험 데이터를 확보하여 연구를 진행했다. 10가지 상황은 오른쪽 발목 골절(Right ankle fracture), 어깨 충돌 증후군(Shoulder impingement syndrome), 척추 측만 증(Scoliosis), 타격(Stroke), 회전근개 파열(Rotator Cuff Tear), 골반 통증(Pelvic pain), 외상성 뇌손상(Traumatic brain injury), 파킨슨병(Parkinson's disease),

Table 1. Sample of CLT (Clinical Natural Language Voice Text)

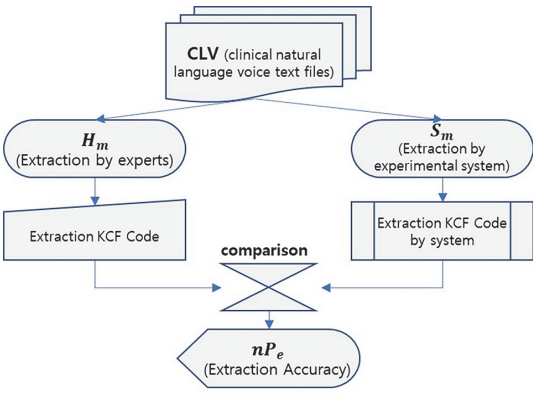
CLV No.	Contents	Count of H_m
CLV_01	Right ankle fracture	138
CLV_02	Shoulder impingement syndrome	461
CLV_03	Scoliosis	142
CLV_04	Stroke	553
CLV_05	Rotator Cuff Tear	92
CLV_06	Pelvic pain	121
CLV_07	Traumatic brain injury	876
CLV_08	Parkinson's disease	922
CLV_09	Transient synovitis of the hip	230
CLV_10	Forward head posture	154

일과성 고관절 활액막염(Transient synovitis of the hip), 일자목 증후군(Forward head posture)이다.

Table 1은 10가지 상황에 대한 내용과 단락의 수를 나타낸 것이다. CLV_01은 138개의 단락을 대상으로 하며, CLV_02는 461개의 단락을 대상으로 하고, CLV_10의 경우는 154개의 단락을 대상으로 하고 있음을 나타내고 있다.

실험 데이터는 자연어의 형태로 구성되어 있고, 이에 대해서 해당 전문가 2명이 합의하에 추출된 KCF 코드와 KCF 코드 추출 알고리즘을 활용한 KCF 코드에 대한 정확도를 파악함으로써 시스템의 활용 가능성에 대하여 파악했다. 전문가에 의해 제시된 KCF 코드와 시스템에서 추출된 코드를 비교하기 위하여 Fig. 2와 같은 정확도 지표 모델을 제시하였다.

정확도 지표 모델은 전문가에 의해 제시된 KCF 코드의 개수(Extraction by experts: *Hm*)와 시스템을 이용해서 추출된 KCF 코드의 개수(Extraction by system: *Sm*)을 비교하여 산술적으로 유사성을 계산하는 방식이다. 특히 *Hm*의 경우는 각 단락별로 전문가가 제시한 KCF 코드의 개수로 해당 단락의 기준이 된다. 예를 들어 전문가가 해당 단락의 KCF 코드를 3개 제시하였다면 *Hm*은 3이 된다. 시스템에서 추출한 KCF 코드 중에서 전문가가 제시한 코드와 같은 것이 2개 존재하면 *Sm*은 2가 되게 되며, *Sm*은 아무리 크더라도 *Hm*보다는 클 수 없다. 본 정확도 지표에서는 전문가에 의해 제시된 KCF 코드를 표준으로 볼 수 있다고 가정하고 시스템에서 추출된 KCF코드를 참조해서 상호 비교하여 전문가가 제시한 일부 코드에 대한 오류 수정도 진행되었다. 3,689개의 실험 데이터에 대한 정확도 분석을 통해 시스템을 통한 KCF 코드 추출 가능성을 확인할 수 있다.



$$nP_e = \left(\frac{1S_m \cap 1H_m}{1H_m} + \frac{2S_m \cap 2H_m}{2H_m} + \cdots \frac{nS_m \cap nH_m}{nH_m}\right)/n$$

$$CLV\_nP_e = \sum \frac{nS_m \cap nH_m}{n \times nH_m}$$

Fig. 2. Index of the linking accuracy.

Ⅲ. 연구결과

시스템을 이용한 KCF 코드 추출의 정확도를 확인해 보았다. 입력된 실험 데이터인 자연어 단락 자체가 인터뷰를 통해 이루어진 내용이라 의미가 없는 입력 데이터도 포함되어 있으며, 의미 파악이 쉽지 않은 단락도 존재하였다.

Table 2는 실험 데이터 중에서 정확도가 높은 일부 샘플을 나열한 것이다. 단락의 내용이 일상 용어로 표현되더라도 의미적 유사성을 높게 형성한 것을 확인할 수 있다. 또한 Table 3의 1차 실험에서는 100%의 일치성을 가진 샘플이 213개(5.77%)를 나타냈으며, 75%이상의 정확도를 가진 샘플은 272개(7.37%), 50%이상의 정확도는 418개(11.33%), 25%이상의 정확도를 가진 샘플이 527개(14.29%)로 나타났다. 그러나 25% 미만의 정확도를 가진 샘플이 2,259개(61.24%)로 나타난 것을 확인할 수 있다. 정확도 향상을 위해 전문(의학) 용어에 대한 가중치를 조절한 뒤 시스템을 통해 정확도를 재 산정한 2차 실험 결과 100%의 일치성을 가진 샘플이 323개(8.76%), 75%이상의 정확도는 416개(11.28%), 50%이상은 599개(16.24%), 25%이상은 783개(21.23%), 25% 미만의 정확도를 가진 샘플이 1,568개(42.50%)로 나타났다. 가중치에 의해 정확도가 다소 향상된 것을 확인할 수 있었고, 여전히 시스템에 의한 KCF 코드 추출의 내용에 대한 정확도가 상당히 미흡한 것으로 파악되었으나, 시스템을 이용한 형태적 코드 추출 방법에 대한 가능성을 확인할 수 있다.

Table 2. Accuracy by example

Sample No.	Experts (*Hm*)	System (*Sm*)	Accuracy (%)
CLV_01_P036	s720(어깨 부위의 구조)	s5400(소장) s720(어깨 부위의 구조) s230(눈 주위의 구조)	100
CLV_01_P046	s720(어깨 부위의 구조)	s720(어깨 부위의 구조) s7200(어깨 부위의 뼈) s7202(어깨 부위의 근육) s7203(어깨부위의 인대와 근막	100
CLV_02_P023	s75001(고관절)	s75001(고관절)	100
CLV_02_P069	s750(다리의 구조)	s7501(아래다리의 구조) s75010(아래다리의 뼈) s75012(아래다리의 근육) s75019(상세불명의 아래다리...)	100
CLV_05_P004	d760(가족관계) e310(직계가족)	d760(가족관계) d7602(형제자매 관계) d7608(기타 명시된 가족 관계) e310(직계 가족)	75
CLV_05_P007	d7701(배우자 관계) e310(직계 가족)	d7701(배우자 관계) d7702(성적인 관계) e310(직계 가족)	75
CLV_08_P003	b280(통증) s750(다리의 구조) s7502(발목과 발의 구조)	s230(눈 주위의 구조) s7202(어깨 부위의 근육) s75012(아래다리의 근육) s75022(발목과 발의 근육) s8301(발톱)	67
CLV_09_P001	d410(기본자세 변경) d4103(앉기) e5350(통신서비스)	d4151(쪼그려 앉은 자세 유지) d4104(서기) d4153(앉은 자세 유지) d4101(쪼그려 앉기) d4103(앉기)	67
CLV_05_P063	d430(물건을 나르기) d850(유급 고용)	d4309(상세불명의 들어...) d4308(기타 명시된 들어...) d4300(들어 올리기) b3401(음역의 산출) s560(간의 구조)	50
CLV_06_P003	b280(통증) s7501(아래다리의 구조)	s7501(아래다리의 구조) s75010(아래다리의 뼈) s75012(아래다리의 근육) s75013(아래다리의 인대와...) s75019(상세불명의 아래다리...)	50

Table 3. Distribution by the accuracy range

Accuracy Range		$nP_e = 100$	$100 > nP_e \geq 75$	$75 > nP_e \geq 50$	$50 > nP_e \geq 25$	$25 > nP_e$
1st Experiment	Count	213	272	418	527	2259
	Distribution	5.77%	7.37%	11.33%	14.29%	61.24%
2nd Experiment	Count	323	416	599	783	1568
	Distribution	8.76%	11.28%	16.24%	21.23%	42.50%

Ⅳ. 고 찰

결과에서 확인한 바와 같이 실험을 통한 현상은 크게 3가지의 문제로 파악할 수 있다.

첫째로 입력 데이터에 대한 문제로 일상적인 인터뷰를 통해 추출된 일상 용어 데이터를 가공 없이 그대로 입력 데이터로 활용함에 따른 중복적 단어의 활용, 전문 용어에 대한 잘못된 사용 및 의미를 내포하지 못하는 대화 내용 등이 있으며, 둘째로 시스템 적용 가능성에 대한 실험적 시스템 구축으로 유사, 동의어에 대한 샘플 데이터 베이스의 한정적 적용에 따라 일상용어의 유사, 동의어 기반의 검색에 대한 정확도가 현저히 떨어지는 결과를 초래하였다. 이에 대하여는 우리말샘사전 및 기존의 사전의 유사, 동의어를 데이터베이스화하여 적용하게 되면 정확성을 상당히 높일 수 있을 것으로 확인되었다. 세번째로 Table 3에서와 같이 전문용어에 대한 중요도 가중치를 조절하여 적용함으로써 정확도 향상에 크게 기여할 수 있다는 것을 확인하였다.

1. 입력 정보에 대한 고찰

본 연구의 실험 데이터는 상담을 통해 얻어진 음성파일을 텍스트로 변환하여 실험에 적용되었다. 일반인이 활용하더라도 시스템에서 요구 코드를 추출하기 위한 실험으로 상담의 데이터를 여과 없이 직접 활용하는 방법을 택했다. 이러한 데이터는 자연어에 최대한 가까운 데이터임에는 분명하나, 대화의 과정에서 자연스럽게 의미를 파악할 수 없는 내용도 포함되는 것이다. 이러한 근본적 이유로 인해 정확도 25% 미만의 데이터가 다수 발생하였다.

실험 데이터의 문단을 분석한 결과 각 문단에 사용된 단어는 1개에서 281개까지 사용되어, 그 차이를 보였다. 즉 짧은 문단과 긴 문단이 존재하게 된다. 이러한 문단의 길이 차이에 의해 짧은 문단에서 핵심 단어가 노출되면 그 단어의 영향력이 커지고, 긴 문단에서는 핵심단어의 영향력이 적어지는 효과가 발생되었다. 예를 들면 1개의 단어로 형성된 문단인 “발등이?”라는 샘플은 해당 내용만으로 s7502(발목과 발의 구조) 코드가 바로 연계될 수 있지만, 133개의 단어를 포함하고 있는 “그러면 지금 아까 전에 화장실을 …<중략>… 약 아니야 그거는 음료수”로 되어 있는 샘플에서는 전문가들이 e115(개인의 일상생활용 제품과 기술), e310(직계가족), d460(다른 장소로의 이동), b525(배변기능), b620(배뇨기능), b630(비뇨기능과 관련된 감각), e110(개인 소비용 제품 또는 물건), e355(보건 전문가) 코드가 제시되었다. 단어의 수가 많을 경우 전반적으로 코드를 통해 의미를 파악하기 어렵다.

Cieza 등[21]의 연구를 살펴보면 연구의 샘플이 되는 데이터가 의료 기록으로 한정되었을 경우 자연어 처리를 통해 ICF를 활용하는 효과성이 입증되기도 하였다. 실제 KCF 사용의 경우 문진표나, 의료 기록 등 입력 데이터가 정제되어 정확도가 높아질 수 있겠지만, 다양한 분야의 다양한 사용자를 대상으로 사용의 폭을 넓히기 위해서는 데이터 입력에 대한 방식도 개선이 되어야 할 필요성이 있다.

2. 실험 결과에 대한 고찰

연구 방법에 따라 실험이 진행되었고 그 결과도 도출이 되었다. 실험 데이터에 대한 전문가의 코딩 결과를 기준으로 시스템에서 추출한 코드를 비교하여 정확도를 확정지었으나, 1차 실험에 대한 결과, 정확도 100%가 5.77%, 정확도 100%미만 75%이상이 7.37%, 정확도 75%미만 50%이상이 11.33%, 정확도 50%미만 25%이상이 14.29%,

정확도 25% 미만이 61.24%로 나타났다. 정확도 100%가 나온 샘플이 있는가 하면 0%의 정확도가 나온 샘플도 존재하는 등 상당히 낮은 정확도를 나타내고 있다.

다양한 원인 중 가장 중요한 원인은 사용된 시스템의 데이터베이스가 실험을 위해 간략히 구축된 데이터베이스라는 것이다. 시스템에 관계되는 데이터베이스의 경우 KCF용어 데이터베이스(1,587레코드), 의학용어 데이터베이스(1,414레코드), 임상자연어 데이터베이스(73,308레코드) 대표단어 데이터베이스(1,787레코드)이다. 이는 KCF 코드의 추출이 시스템으로 가능한지를 확인하기 위한 최소 조건의 데이터베이스 구축이다.

분석을 통해 확인한 바로는 임상자연어 데이터와 대표단어 데이터의 경우 사용량이 늘어날수록 데이터 베이스의 내용은 학습에 의해 기하급수적으로 증가할 것이고, 추가적으로 국어사전에 반영되어 있는 434,216개의 단어에 대한 동의어 및 유사단어에 대한 데이터베이스가 구축되어야 할 필요성이 있다. 결과적으로 데이터베이스를 추가적으로 구축하여 정확도 향상이 가능하게 된다.

2차 실험의 경우에는 알고리즘에 의학용어 및 임상자연어에 대한 가중치를 조절하고 진행하였다. 그 결과 정확도 100%가 8.76%, 정확도 100%미만 75%이상이 11.28%, 정확도 75%미만 50%이상이 16.24%, 정확도 50%미만 25%이상이 21.23%, 정확도 25% 미만이 42.50%로 나타났다. 이는 정확도 구간별로 100%구간에서 51.82%, 100%~75%구간에서 53.05%, 75%~50%구간에서 43.34%, 50%~25%구간에서 48.57%, 25%~0%구간에서 30.60%의 정확도 향상이 나타난 것을 알 수 있다. 이러한 결과를 바탕으로 정확도 향상을 위해 추가적인 가중치 연구가 필요하다.

3. 정확도 향상에 기여하는 변수

실험을 통해 정확도 향상을 조절하는 인자를 확인할 수 있었다. 그 첫번째가 입력 데이터에 대한 정도 향상, 두번째로 관련 데이터 베이스에 대한 확대 개발, 세번째가 알고리즘 내의 각종 가중치에 대한 연구 및 적용을 제시할 수 있다.

입력 데이터의 정도 향상에 대해서는 시스템 설계에 있어서 사용자의 편의 및 활용성을 고려하고 정보 수집이 용이하도록 UI(User Interface)를 개발할 필요가 있다. 휴대가 간편한 스마트폰을 통해 질의 응답을 유도하는 방식이나, 로봇을 통해 질문하고 정보를 추출하는 방식 등도 고려해 볼 수 있다.

데이터 베이스에 대한 확대 개발은 지속적인 추가 연구를 통해 개발이 진행될 필요가 있다. 실제 KCF용어 관련 데이터 베이스는 거의 일정하게 유지될 것이며, 의학용어 데이터 베이스의 경우는 의학용어에 대한 추가적인 데이터를 수집하여 반영하면 가능하다. 임상자연어나 대표단어의 경우 알고리즘에 의해 지속적으로 추가 및 수정이 이루어지므로 사용량이 늘어나면 그 역할이 더 늘어날 것이다. 단 추가적으로 우리말 사전에 나타나 있는 동의어나 유사어의 경우는 데이터베이스를 n:n의 경우로 분석하고 적용해야 한다. 이러한 경우 검색에 대한 사례가 기하급수적으로 증가하여 추출에 상당한 시간이 소요될 수 있어 합리적인 데이터베이스 구성이 필요하다.

알고리즘에 적용되는 가중치의 경우 전문영역(KCF용어, 의학, 임상자연어 등)의 데이터베이스에 관여되는 가중치와 국어사전 데이터베이스의 유사단어 및 동의어에 대한 가중치를 이야기한다. 본 연구의 실험을 통해 확인한 바에 의하면, 의학용어에 대한 가중치만을 간단히 조절함으로 인해 상당한 정확도의 향상을 이룰 수 있었다. 그 만큼 알고리즘에 관여하는 가중치는 중요한 요소인 것으로 판단된다. 이러한 가중치에 대한 연구는 향후 추가적으로 수행되어야 할 필요가 있다.

4. AI 활용 방안

본 연구에서 활용되는 알고리즘에는 학습과 관련되는 논리시스템이 포함되어 있으나 인공지능의 수준은 아니며 데이터 베이스를 이용한 마이닝에 학습기능을 이용해서 데이터를 추가 또는 수정하는 정도이다. 정확도 향상에 기여하는 변수에 있어서 국어사전의 동의어나 유사단어의 검색 등은 분명한 정확도 향상에 기여하겠지만, 시스템에 가해지는 부하는 상당할 것으로 예상된다. 그러므로 이러한 시스템의 경우 빅데이터나, 인공지능을 활용하면 정확도 향상에 상당한 기여를 할 수 있을 것으로 판단된다. 자연어 검색을 통한 다양한

기법들은 이미 인공지능으로 구현되어 있으므로 이를 활용할 필요가 있다.

연구에서 제시한 방법은 자연어 단락을 단어 형태적 노출의 빈도에 따른 PageRank 기법을 활용하여 의미를 부여하는 방식으로 접근하였다. 최근 유사 연구 사례[20]에서는 자연어의 의미 단어를 기반으로 적합한 ICF 코드 후보군을 제시하고 이에 따른 유사 코드를 상호 비교하면서 가장 적합한 코드를 찾아내는 방법의 연구도 있으나, 제한적인 범위 내에서 입력 데이터가 다루어지므로 다양한 활용에 한계가 존재하지만, 본 연구의 경우 범위의 제한이 없이 적용 가능하다는 차이가 있다.

KCF의 활용성 증대를 위해서는 그 범위를 확대시킬 필요성이 있고, 자연어 기반의 코드 추출이 완벽히 이루어진 이후에 각 코드 별로 건강상태, 기능상태, 삶의 질 등을 확인할 수 있는 추가적인 평가값(Qualifier)에 대한 활용도 가능하다. 코드 추출에 대한 정확도를 향상시키기 위한 기본 방향으로 기존의 데이터베이스를 활용함과 동시에 최근 4차산업혁명의 중심이라고 할 수 있는 이기종 간의 데이터활용에 대한 방안도 연구할 필요가 있으며, 인공지능(AI)기반의 검색기능에 대한 추가적인 연구를 통해 코드 추출의 정확도를 향상시킬 수 있다.

결과적으로 KCF에 시스템 적용 방안을 추가함으로 인해 KCF가 추구하는 건강상태, 기능상태, 웰빙, 삶의 질 등의 확인과 활용에 있어서 각 분야 전문가의 전유물이 아니라 누구나 쉽게 코드를 검색하고 도구에 접근하여 활용할 수 있는 기반이 될 수 있으며, 시스템에 적용되는 몇가지의 독립변수(각 데이터베이스별 가중치, 유사 및 동의어 데이터베이스의 확대 등)를 조절함으로써 KCF 코드 추출의 정확도를 향상시킬 수 있음을 발견하였다. 현재까지의 KCF 사용에 대한 한계를 뛰어 넘을 수 있는 방안의 제시가 될 수 있다는 점에서 본 연구의 의의가 크다 하겠다.

Ⅴ. 결 론

KCF가 프로토콜 개념으로 이용되면 향후 다양한 산업에 활용될 수 있다. 본 연구는 각 분야의 활용성을 극대화하기 위해 자연어를 입력 데이터로 받아 시스템을 통해 KCF 코드를 추출하는 실험적 시스템을 구축하고 이에 대한 정확도를 확인했다. 임상 자연어를 통해 KCF 코드를 추출함으로써 전문가의 전유물이 아닌 누구나 쉽게 코드를 검색하고 활용할 수 있는 방안을 제시하였다.

아직 초기 단계의 시스템을 통한 코드 추출 정확도는 미흡한 실정이나, 시스템을 통해 KCF 코드를 추출해 낼 수 있는 기반 연구를 수행하고 향후 방향을 제시한 것은 큰 의미가 있다.

References

[1] Organization WH. Towards a common language for functioning, disability, and health: ICF. The international classification of functioning, disability and health. 2002.

[2] Bioelectrical Impedance Analysis at Popliteal Regions of Human Body using BIMS. Sensors. 2016;25(1):1-7.

[3] Cieza A, Brockow T, Ewert T, et al. Linking health-status measurements to the international classification of functioning, disability and health. J Rehabil Med. 2002;34(5):205-10.

[4] Stucki G, Cieza A, Ewert T, et al. Application of the international classification of functioning, disability and health (icf) in clinical practice. Disabil Rehabi. 2002; 24(5):281-2.

[5] Üstün TB, Chatterji S, Bickenbach J, et al. The International Classification of Functioning, Disability and Health: a new tool for understanding disability and health. Disab Rehab. 2003;25(11-12):565-71.

[6] Imrie R. Demystifying disability: a review of the International Classification of Functioning, Disability and Health. Sociol Health Illn. 2004;26(3):287-305.

[7] McDougall J, Wright V, Rosenbaum P. The ICF model of functioning and disability: incorporating quality of life and human development. Develop neurorehab. 2010;13(3):204-11.

[8] Fox MH, Krahn GL, Sinclair LB, et al. Using the

international classification of functioning, disability and health to expand understanding of paralysis in the United States through improved surveillance. Disab Health J. 2015;8(3):457-63.

[9] Stucki G. International Classification of Functioning, Disability, and Health (ICF): a promising framework and classification for rehabilitation medicine. Am J Phys Med Rehabiln. 2005;84(10):733-40.

[10] Jelsma J. Use of the International Classification of Functioning, Disability and Health: a literature survey. J Rehabil Med. 2009;41(1):1-12.

[11] Hand DJ, Adams NM. Data mining. Wiley StatsRef: Statistics Reference Online. 2014:1-7.

[12] Chen M-S, Han J, Yu PS. Data mining: an overview from a database perspective Trans Knowl Data Eng. 1996;8(6):866-83.

[13] Minsky M. A framework for representing knowledge. de Gruyter. Berlin, Boston. 2019.

[14] Manning C, Schutze H. Foundations of statistical natural language processing. MIT press. 1999.

[15] Hoyles C, Noss R, Adamson R. Rethinking the microworld idea. J. Educ Comput Res 2002;27(1):29-53.

[16] Mitchell R, Michalski J, Carbonell T. An artificial intelligence approach. Springer. 2013.

[17] Indurkhya N, Damerau FJ. Handbook of natural language processing. CRC Press. 2010.

[18] Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. Journal of the American Medical Informatics Association. 2011;18(5): 544-51.

[19] Thamhain HJ. Management of technology: Managing effectively in technology-intensive organizations. John Wiley & Sons. 2005.

[20] Newman-Griffis D, Maldonado JC, Ho PS, et al. Linking Free Text Documentation of Functioning and Disability to the ICF With Natural Language Processing. Front Rehabil Sci. 2021;2.

[21] Cieza A, Geyh S, Chatterji S, et al. ICF linking rules: an update based on lessons learned. J Rehabil Med. 2005;37(4):212-8.

[22] Haveliwala TH. Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. Trans Knowl Data Eng. 2003;15(4):784-96.