JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 29 No. 9, pp. 197-205, September 2024
https://doi.org/10.9708/jksci.2024.29.09.197

Identifying Missing Concepts in SNOMED CT by Utilizing Attribute Relationships of Sibling Concepts

Wooseok Ryu*

*Professor, Dept. of Health Care Management, Catholic University of Pusan, Busan, Korea

[Abstract]

SNOMED CT is the most widely used comprehensive clinical terminology system worldwide. However, due to the vastness of the terminology and the continuous growth of medical knowledge, the system involve quality issues such as structural inaccuracies and inconsistencies, including missing concepts or relationships and hierarchical errors. In this paper, we propose a method to enhance the consistency of the system by detecting potentially missing concepts by utilizing attributes linked to concepts. The proposed method analyzes the characteristics of the attribute relationships of concepts, extracts sibling concepts that share the same characteristics, and then evaluates whether the parent concepts reflect these characteristics to detect potentially missing concepts. By applying this method to the March 2023 SNOMED CT international release, we identified 564 instances where parent concepts did not reflect the common attributes of their sibling concepts, and a total of 384 potentially missing concepts were detected, including cases involving multiple parent concepts.

▸ **Key words**: SNOMED CT, Inconsistency, Missing Concept, Sibling Concepts, Attribute Relationship

[요 약]

SNOMED CT 용어체계는 전 세계적으로 가장 널리 사용되는 포괄적 임상 용어체계이다. 하지만 용어체계의 방대함 및 의학 지식의 지속적 증가로 인해 용어체계를 구성하는 개념 또는 관계의 누락, 계층 구조의 오류 등 구조적 부정확성 및 불일치 문제가 존재한다. 본 논문에서는 용어체계의 일관성을 높이기 위해 개념에 연결된 속성을 이용하여 잠재적으로 누락된 개념을 탐지하는 방법을 제안한다. 제안하는 방법은 개념이 가지고 있는 속성 관계의 특성을 분석하고 동일한 특성을 공유하는 형제 개념들을 추출한 후 부모 개념들이 해당 특성들을 반영하는지 여부를 평가하여 잠재적으로 누락된 개념을 탐지한다. 제안하는 방법을 2023년 3월 SNOMED CT 국제 배포판에 적용하여 분석한 결과 형제 개념들의 공통 속성을 부모 개념이 반영하지 못하는 경우를 564건 발견하였으며, 다중 부모 개념을 포함하는 경우를 도출하여 총 384개의 잠재적 누락 개념을 탐지하였다.

▸ **주제어**: 의학 용어, 불일치, 개념 누락, 형제 개념, 속성 관계

• First Author: Wooseok Ryu, Corresponding Author: Wooseok Ryu
*Wooseok Ryu (wsryu@cup.ac.kr), Dept. of Health Care Management, Catholic University of Pusan
• Received: 2024. 06. 25, Revised: 2024. 09. 03, Accepted: 2024. 09. 03.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

EMR(Electronic Medical Record) 및 EHR(Electronic Health Record)에서 진단, 처치 등의 진료기록을 작성할 때 사용되는 표준 용어체계는 의학 통계, 의료기관 내외부의 진료기록 교류, 의학 연구 및 진료의 연속성 지원에 필수적인 역할을 한다. 표준 용어체계 중 SNOMED CT는 임상 소견, 질병, 처치, 신체구조, 유기체 등 진료기록에 필요한 방대한 양의 의학용어를 제공하는 대표적인 표준 용어체계이며, 환자 정보의 체계적 기록 및 의료정보의 상호운용성 지원을 위해 전 세계적으로 널리 사용되고 있다. 또한, 국내에서도 상급 종합병원 중심으로 그 활용 범위를 넓혀가고 있다[1].

SNOMED CT는 개념(Concept), 용어(Description) 및 관계(Relationship)로 구성되어 있다. 개념은 고유의 의학적 의미(Clinical Meaning)를 저장하는 단위로 숫자 형태의 식별자인 개념 코드(Concept ID)를 가진다. 개념은 그 의학적 의미를 언어로 표현하기 위해 다수의 용어를 포함하고 있으며, 의미를 구체적으로 표현하기 위해 연관성이 있는 개념들을 다양한 관계를 통해 연결하고 있다.

관계는 계층 관계(is-a Relationship)와 속성 관계(Attribute Relationship, 이하 속성)로 구분된다. 용어체계에 포함된 모든 개념들은 최상위 개념인 “138875005 |SNOMED CT Concept|”에서 계층 관계로 연결되어 있다. 계층 관계는 개념의 의미를 계층적으로 한정하는데 이용되며, 계층 관계에 따라 최상위 개념을 제외한 모든 개념은 하나 또는 그 이상의 부모 개념을 가지는 특징이 있다. 속성 관계는 개념의 특성을 정의하기 위해 다른 개념과 연관 관계를 설정하는 데 사용되며 해당 개념의 의미를 보다 명확하게 정의하는데 활용된다. 그림 1은 SNOMED CT의 예시로서 세균성 폐렴(53084003 |Bacterial pneumonia|)이라는 개념은 계층 관계를 통해 두 부모 개념의 자식 개념으로 정의되어 있다. 그리고 발견 위치(Finding site), 연관 형태(Associated morphology), 병리학적 과정(Pathological process), 원인균(Causative agent)이라는 네 개의 속성을 이용하여 다른 개념들을 연관시키고 있다. 속성을 통해 세균성 폐렴의 발병 위치, 형태, 병리학적 단계, 원인인자를 구체적으로 제시함으로써 세균성 폐렴의 의미를 보다 명확하게 정의한 것을 확인할 수 있다.

SNOMED CT 용어체계는 2023년 3월 국제 배포판(International Release of SNOMED CT) 기준 30만개 이상의 개념들과 약 100만 건의 관계를 포함하고 있으며, 정의되어 있는 속성의 종류도 99종에 달하는 등 방대하고 복잡한 온톨로지 구조를 갖고 있다[2]. 용어체계의 품질 보증(Quality Assurance)은 용어체계의 활용 측면에서 필수적인 요소이며, 특히 온톨로지의 구조적 완전성은 진료기록의 자동화에서 매우 중요한 요소이다[3]. 하지만, 의학 지식이 지속적으로 증가함에 따라 일관적이지 않은 개념의 정의, 개념의 중복 또는 누락, 계층 관계의 오류 등 용어체계의 불일치 및 부정확성으로 인한 온톨로지의 품질 문제는 불가피하게 존재하고 있다[4].

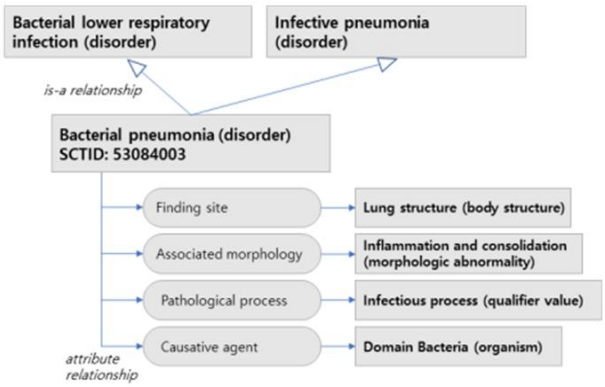


Fig. 1. An example of relationships of a SNOMED CT concept

본 연구에서는 용어체계의 일관성을 높이기 위해 온톨로지 구조에서 개념들 간에 정의되어 있는 관계를 이용하여 잠재적으로 누락된 개념을 탐지하고자 한다. 본 연구의 접근 방법은 SNOMED CT에 정의된 속성 중 계층적 특성을 띄는 속성들에 근거하여 유사한 특성을 보이는 형제 개념들을 찾아내는 것이다. 그 형제 개념들을 이용하여 잠재적으로 누락된 부모 개념을 탐지하는 알고리즘을 제시하고 이를 용어체계에 적용하여 분석 결과를 검증하는 것이 본 연구의 목적이다. 제안하는 방법은 잠재적으로 누락된 개념의 탐지를 위해 개념의 어휘적 특성이나 계층 구조 특성을 이용하는 대신 속성이 가지는 고유의 특성을 이용하여 자동화된 탐지 알고리즘을 제시한 측면에서 기존 연구와 차별성이 있다.

본 논문의 구성은 다음과 같다. 먼저 2장에서는 SNOMED CT 용어체계의 일관성을 높이기 위한 선행 연구들을 분석한다. 3장에서는 속성이 유사한 특성을 가지는 개념들을 추출하고 이를 통해 누락된 부모 개념을 식별하는 알고리즘을 제안하고 4장에서는 제안한 방법을 이용하여 용어체계에서 누락된 부모 개념을 식별한 결과를 제시한다. 마지막으로 5장에서 결론 및 향후연구를 기술한다.

II. Related Work

SNOMED CT의 정확성 및 일관성을 높이기 위한 연구는 계층 관계, 개념, 속성의 누락 여부를 탐지함으로써 잠재적 오류를 찾아내는 연구가 주로 이루어지고 있다. 접근 방법은 크게 두 가지로 구분할 수 있는데 첫 번째는 계층 관계 또는 개념의 용어를 구성하는 규칙을 정의하고 이를 활용하여 잠재적 오류를 탐지하는 방법이다. 연구 [5]에서는 개념의 용어와 부모 개념의 용어를 기반으로 부분 문자열 대체(Substring Replacement)를 통해 새로운 계층 관계를 생성하는 방법을 제안하였으며, 어휘적 특성 집합(Enriched Lexical Feature Set)을 생성하고 이를 다른 개념과 비교하여 잠재적으로 누락된 계층 관계를 탐지하는 연구도 제안되었다[6]. 연구 [7]은 계층 관계에 포함된 부모 개념과 자식 개념 간의 용어 차이 쌍(Term Difference Pair, TDP)을 생성하고 이 TDP의 발생 건수를 분석하여 잠재적 오류에 해당하는 계층 관계를 제안하였다. 계층 구조의 불완전성과 관련한 연구로서 연구 [8]은 계층 구조의 특성을 분석하여 계층 관계가 불완전한 비격자(non-lattice) 서브 그래프를 제안하고 그래프 내 개념들의 어휘적 특성을 분석하여 잠재적으로 누락된 계층 관계를 제안하였다. 이외에도 완전히 정의된 개념(Fully Defined Concept)과 부분적으로 정의된 개념(Partially Defined Concept)들의 어휘적 특성 비교를 통해 부분적으로 정의된 개념에 잠재적으로 누락되어 있는 속성을 탐지하는 방법도 연구되었다[9].

두 번째는 규칙을 정의하는 대신 기계 학습(Machine Learning)을 통해 용어체계를 학습시키는 방법이다. 연구 [10]은 개념의 용어를 word2vec으로 학습하여 어휘적 유사성을 수치화하고 유사한 개념들을 그룹화 한 후 유사 개념들 간 속성의 불일치를 탐지하는 방법을 제안하였다. 연구 [11]은 비격자 서브 그래프 내 개념들 간의 누락된 개념을 탐지하기 위한 기존의 방법[8] 대신 텍스트 요약 모델(Text Summarization Model)[12]을 이용하여 누락된 개념과 그 용어를 생성하는 방법을 제안하였다. 연구 [13]은 계층 구조를 그래프 합성곱 신경망(Graph Convolutional Network, GCN)으로 학습하여 누락된 계층 관계를 탐지하는 모델을 제안하였으며, 지식 그래프 임베딩(Knowledge Graph embedding)을 이용하여 노드와 관계의 임베딩 표현을 학습하고 벡터 유사성 예측을 통해 누락된 계층 관계를 예측하는 연구도 수행되었다[14].

기존의 연구들은 공통적으로 개념에 정의된 용어가 가지는 어휘적 특성 및 계층 관계를 이용하여 오류를 탐지하는데 초점을 맞추고 있다. 일부 기존 연구에서 개념에서 어휘적 특성을 추출하기 위해 속성에 정의된 용어를 이용하기도 하였으나, 다양한 속성들 사이에 발생할 수 있는 구조적 특성은 반영하지 못했다[6,13]. 개념의 속성을 활용하여 누락된 개념을 제안하는 연구는 본 연구의 선행연구에서 처음 수행되었다[15]. 이 연구에서는 개념에 정의된 속성 중 계층적 특성이 있는 두 속성을 이용하여 누락된 개념을 탐지하는 아이디어를 제안하였는데, 구체적인 알고리즘 제시는 이루어지지 않았으며, 유사한 개념들이 가지는 여러 유형에 대한 사례 분석도 포함되지 않았다. 본 연구는 선행 연구를 확장하여 속성의 계층 구조에 따라 유사한 특성을 가지는 개념들을 유형화하고 각 유형에 따라 개념의 누락 여부를 판단하는 알고리즘을 수식과 순서도를 이용하여 명확하게 제시하였다. 또한, SNOMED CT 배포판 데이터 세트(Dataset)에 대해 제안한 알고리즘에 따른 유형별 분석 결과와 사례를 제시함으로써 제안한 방법의 유용성을 실증하였다.

III. The Proposed Scheme

이 장에서는 SNOMED CT에서 개념의 속성을 이용하여 누락된 개념을 탐지하는 방법을 제안한다. 제안하는 기법의 기본 아이디어는 유사한 특성을 가지는 형제 개념들을 그룹화 하여 잠재적으로 누락된 부모 개념을 탐지하는 것이다. 이때의 유사한 특성은 개념들이 가지는 속성을 이용하여 도출하는데 본 연구에서는 '해석한다'를 의미하는 interprets 및 '해석된 값'을 의미하는 has interpretation 속성을 이용한다. has interpretation 속성은 interprets 속성에 대한 한정적 의미를 제공하므로 interprets 속성과 함께 정의되는 특성이 있다. 또한 이 속성들은 계층적 특성을 가지고 있는데 부모 개념이 interprets 속성을 가지고 있다면 자식 개념은 동일한 interprets 속성과 추가적으로 has interpretation 속성을 가질 수 있다. 그림 2의 예시를 보면 영상의학적 소견(|Radiologic finding|) 개념의 경우 두 자식 개념들도 동일한 interprets 속성을 가지고 있으며, has interpretation 속성으로 그 의미를 한정하고 있음을 확인할 수 있다.

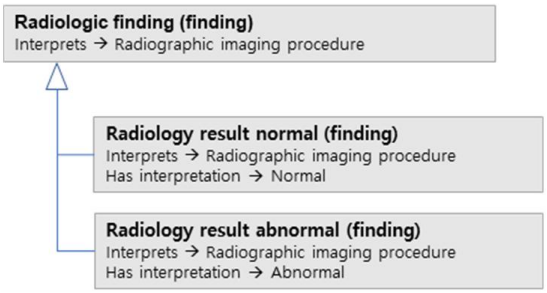


Fig. 2. An example of hierarchical characteristics of interprets and has interpretation attributes

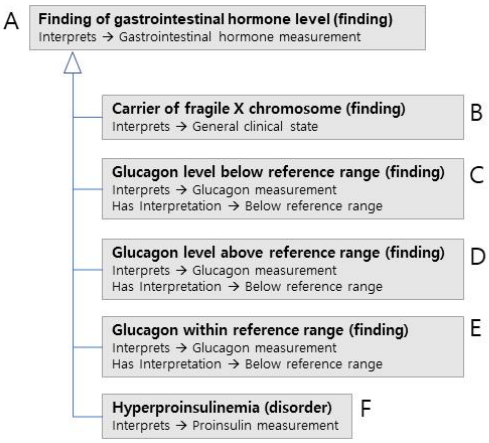
이를 자식 개념 관점에서 표현하면 interprets 속성과 has interpretation 속성을 모두 가진 개념이 있다면 그 개념의 부모 개념은 동일한 interprets 속성을 가진다는 규칙을 생성할 수 있다. 이 규칙을 수식으로 표현한 결과는 수식 (1)과 같다. 수식 (1)에서 *c*는 개념, *P(c)*는 *c*의 부모 개념 집합을 의미하며, *i, hi*로 표기된 화살표 기호는 각각 *interprets* 속성, *has interpretation* 속성을 의미하며 x, y는 각각의 속성 값을 의미한다. 이 규칙의 타당성은 4장에서 데이터 분석을 통해 제시한다.

$$\left((c \xrightarrow{i} x) \wedge (c \xrightarrow{hi} y)\right) \Rightarrow \exists\, c_p \in P(c)\left(c_p \xrightarrow{i} x\right). \quad (1)$$

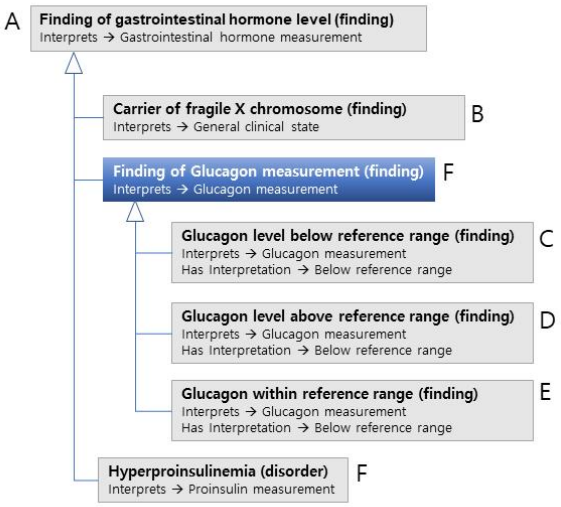
본 연구에서 제안하는 개념의 누락 여부 판별 기준은 수식 (1)에 해당하는 부모 개념이 존재하지 않을 때이다. 즉, 두 속성을 가진 개념의 부모 개념 모두가 동일한 *interprets* 속성을 가지고 있지 않다면 해당 속성을 가진 개념이 잠재적으로 누락되어 있다고 판단한다. 이 조건은 수식 (2)와 같이 표현할 수 있다.

$$\left((c \xrightarrow{i} x) \wedge (c \xrightarrow{hi} y)\right) \Rightarrow \forall\, c_p \in P(c)\left(\neg(c_p \xrightarrow{i} x)\right). \quad (2)$$

그림 3 (a) 의 예시로 볼 때 다섯 개의 자식 개념들 중 C~E 세 개념들은 interprets 속성으로 'Glucagon measurement'를 모두 가지고 있으며 각각 has interpretation 속성도 가지고 있다. 하지만, 그 부모 개념인 A는 동일한 interprets 속성 값을 가지고 있지 않아서 수식 (2)의 조건을 충족한다. 이에 따라 글루카곤 측정(Glucagon Measurement)이라는 interprets 속성을 가진 부모 개념이 잠재적으로 누락되었다고 판단할 수 있으며, 그림 3(b)의 F와 같이 이 속성 값을 가진 새로운 부모 개념을 생성하여 용어체계의 일관성을 높일 수 있다.



(a) An example of sibling concepts



(b) An example of adding a missing concept

Fig. 3. An illustrated example of detecting and adding a potentially missing concept

제안하는 누락 개념 탐지 알고리즘은 두 가지 단계로 수행한다. 첫 번째 단계는 수식 (2)의 조건을 만족하는 개념들에 대해 동일한 부모를 가지고 있으며 interprets 속성의 값이 동일한 형제 개념들을 묶어서 형제 개념 그룹을 생성한다. 이때, 형제 개념 그룹은 그 그룹에 포함되는 개념의 수가 2개 이상인 것으로 제한한다. 그 이유는 동일한 interprets 속성 값을 가진 형제 개념이 없는 경우에도 부모 개념을 생성하면 계층 구조가 지나치게 깊어져서 오히려 용어체계의 효율성을 떨어뜨릴 수 있기 때문이다. 수식 (2)를 만족하는 개념의 전체 집합을 C_c라고 할 때 *x*라는 interprets 속성 값을 가지고 *p*라는 부모 개념을 가지는 형제 개념 그룹 *S(x, p)*는 수식 (3)과 같이 정의할 수 있다.

$$S(x,p) = \left\{ c_i | c_i \in C_C \wedge p \in P(c_i) \wedge c_i \xrightarrow{i} x \right\}, \quad where\ |S(x,p)| \geq 2. \tag{3}$$

예를 들어 그림 3 (a)의 개념 C~E는 형제 개념 그룹으로 묶어서 *S("Glucagon measurement", A)* = {C,D,E}로 표기할 수 있다. 이때, 형제 개념 그룹 *S(x,p)*에 대해 생성할 수 있는 잠재적으로 누락된 부모 개념 c_x^p는 부모 개념 *p*, *S(x,p)*에 포함된 개념들을 자식 개념, 그리고 interprets 속성 값 *x*를 가지는 형태로 정의될 수 있다. 이를 수식으로 표현하면 수식 (4)와 같다. 수식에서 *P(c)*와 *C(c)*는 각각 개념 c의 부모 개념, 자식 개념을 의미한다.

$$c_x^p \ where\ P(c_x^p) = \{p\} \wedge C(c_x^p) = S(x,p) \wedge c_x^p \xrightarrow{i} x. \tag{4}$$

두 번째 단계는 SNOMED CT 용어체계가 가지는 고유한 특성인 다중 부모 개념을 고려하여 형제 개념 그룹들을 군집화 하는 것이다. 이 용어체계는 다중 부모를 허용하므로 interprets 값이 x_1로 동일하지만 부모는 p_1, p_2로 서로 다른 형제 개념 그룹 $S(x_1,p_1)$, $S(x_1,p_2)$가 존재할 수 있다. 어떤 개념이 interprets 속성이 동일한 두 형제 개념 그룹에 모두 포함되어 $S(x_1,p_1) \cap S(x_1,p_2) \neq \{\}$ 을 만족하는 경우 각각의 그룹에 대해 누락된 부모 개념을 생성하는 것은 불필요하게 중복되거나 유사한 개념들이 생성될 수 있는 문제가 있다. 그림 4는 이러한 동일한 개념을 포함하고 있는 두 형제 개념 그룹과 그 부모 개념들 간의 가능한 경우를 4가지 케이스로 분류한 예시이다. 그림에서 형제 노드 집합은 각각 S_1, S_2로 표기하고 점선 사각형으로 표시하였으며, 각각의 부모 개념을 p_1, p_2로 표시하였다. 그림에서 p_1, p_2의 자식 개념 중 형제 노드 집합에 포함되지 않은 개념들은 생략하였다.

그림 4 (a)는 S_1과 S_2가 $S_1 = S_2$ 로 동일(Equal)한 개념들을 가지는 경우이며, 이때는 두 그룹 각각에 대해 누락된 부모 개념을 생성하는 대신 하나의 누락된 부모 개념을 생성하고 이를 공유할 수 있다. 즉, S_1의 누락된 부모 개념 c_p을 생성하고 $P(c_p)$ 를 $\{p_1, p_2\}$로 지정함으로써 누락 개념의 중복 생성을 피하고 용어체계의 구조화를 강화할 수 있다. 그림 4 (b)의 포함관계(Containment) 및 그림 4 (c)의 교차관계(Intersection)에 해당하는 경우에는 형제 개념 그룹 각각에 대해 누락 부모 개념을 생성하면 그 의미가 유사해 질 수 있으므로 누락된 부모 개념을 생성하는 형제 노드 집합은 $S_1 \cap S_2$에 해당하는 개념으로 한정하여 공통 부모 개념을 생성할 수 있다. 그림 4 (b)와 그림 4 (c)의 예시로 볼 때 각각 $\{c_1, c_2\}$, $\{c_2, c_3\}$에 대해 누락된 부모 개념을 생성할 수 있다.

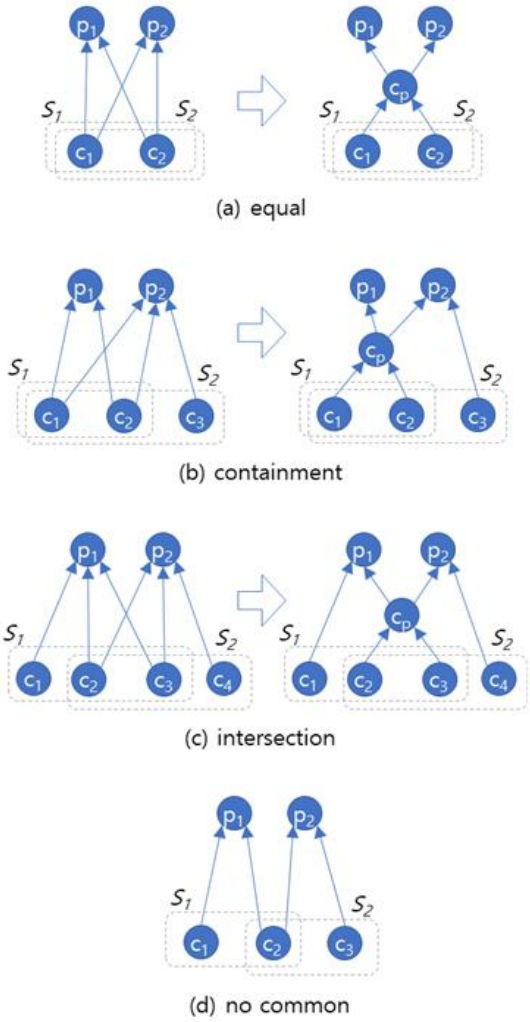


Fig. 4. Four possible cases of two sibling concept groups with two parents

마지막 경우는 개별 관계(No Common)로서 그림 4 (d)와 같이 $|S_1 \cap S_2| \geq 2$를 만족하지 않는 경우이다. 이는 수식 (3)에서 제시한 바와 같이 누락된 부모 개념을 생성할 최소 조건을 만족하지 않는다. S_1과 S_2에 공통적으로 존재하는 c_2에 대해서만 누락된 부모 개념을 생성하는 것은 적절하지 않으며, 또한 각각 부모 개념을 생성하더라도 의미적으로 유사성을 띌 수가 있다. 이 경우는 전문가의 수동 검사(Manual Inspection)를 통해 누락 여부를 판별하는 것이 필요할 것으로 판단되며, 본 연구에서는 이 경우에 대해 자동화된 누락 개념 생성을 수행하지 않는다. 대신 수동 검사를 수행할 수 있도록 4장에서 해당 케이스의 건수를 분석하고 예시를 제시한다.

위와 같이 *interprets* 값이 동일한 여러 형제 개념 그룹

들 간에 공통적으로 존재하는 개념들이 있을 때 누락 부모 개념을 생성할 대상이 되는 형제 개념 그룹 $S_{int}(x)$는 수식 (5)와 같이 정의한다. 여기서 $G(x)$는 interprets 값이 x로 동일한 형제 개념 그룹 $S(x,p_i)$의 집합을 의미한다. 그러면 이를 수식 (4)에 적용하여 $S_{int}(x)$에 대해 누락된 부모 개념을 생성할 수 있다.

$$S_{int}(x) = \{c_i | c_i \in \bigcap S(x,p_i) \wedge S(x,p_i) \in G(x)\}, \quad where\ |S_{int}(x)| \geq 2. \tag{5}$$

그림 5는 앞에서 수식으로 제시한 속성 기반의 누락 개념 탐지 및 생성 알고리즘을 흐름도로 도시한 결과이다. 알고리즘의 순서는 첫째, 수식 (2)를 만족하는 후보 개념들 C_C를 찾아내고, 둘째, 수식 (3)에 따라 interprets 값과 부모 개념이 동일한 형제 개념 그룹 $S(x, p)$를 생성한다. 셋째, $S(x,p)$를 x가 동일한 $G(x)$로 묶은 후, 넷째, 각 $G(x)$ 별로 포함된 형제 개념 그룹의 수에 따라 누락 개념을 바로 생성하거나, 다섯째, 수식 (5)에 따라 형제 개념 그룹에 포함된 개념들의 교집합 S_{int}를 생성하고 이에 대해 누락 개념을 탐지하고 생성한다. 본 알고리즘을 통해서 interprets 속성을 이용하여 계층 구조에 잠재적으로 누락된 개념을 자동으로 탐지하고 생성할 수 있다.

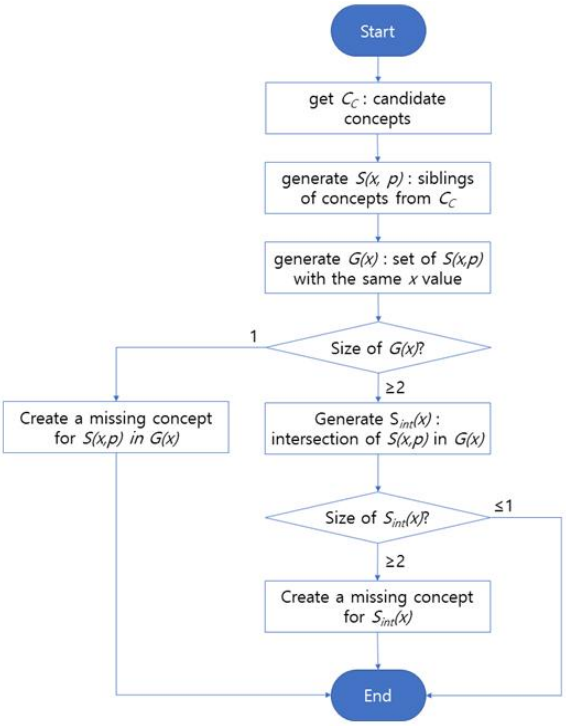


Fig. 5. Flowchart of the Proposed Algorithm

IV. Evaluation

1. Evaluation Environment

본 연구에서 제안한 누락 개념 탐지방법을 평가하기 위해 사용한 데이터 세트는 2023년 3월에 발표된 SNOMED CT 국제 배포판이다[16]. 이 데이터 세트에는 최상위 개념인 SNOMED CT 개념(138875005 |SNOMED CT Concept|) 아래에 19개의 하위 계층(sub-hierarchy)이 존재한다. 각 하위 계층의 이름, 하위 계층 별 개념의 수, 전체 개념의 수는 표 1에 제시하였다.

Table 1. Subhierarchies of SNOMED CT and their descendants (SNOMED CT 2023/03/31 International Release)

Sub-hierarchy	# of concepts	Ratio (%)
Clinical finding	118,729	32.94
Procedure	59,071	16.39
Body structure	41,037	11.39
Organism	33,115	9.19
Substance	27,319	7.58
Pharmaceutical/biologic product	25,195	6.99
Physical object	13,788	3.83
Qualifier value	10,891	3.02
Observable entity	10,336	2.87
Situation with explicit context	4,892	1.36
Social context	4,410	1.22
Event	3,230	0.90
SNOMED CT Model Component	1,868	0.52
Environment or geographical location	1,858	0.52
Specimen	1,766	0.49
Staging and scales	1,570	0.44
Special concept	635	0.18
Record artifact	510	0.14
Physical force	170	0.05
Total	360,390	100.00

데이터 세트에 포함된 총 360,390개의 활성 개념 중 연구 대상으로 설정한 개념은 임상적 발견(404684003 |Clinical finding|) 개념의 하위 계층에 포함된 개념들이다. 이는 질병, 진단 등 임상 진료 과정에서 발견되는 개념들이며 19개의 하위 계층 중 가장 많은 32.94%에 해당하는 개념을 가지고 있다. 또한 이 개념들에서만 interprets, has interpretation 속성들이 정의되어 있다. 본 연구에서는 이 데이터 세트를 MySQL 8.0에 적재시킨 후 SQL을 사용하여 본 알고리즘을 구현하고 분석하였다.

2. Evaluation Results

연구 대상 데이터 세트에 대해 두 속성의 특성과 형제 개념 그룹을 분석한 결과는 표 2와 같다. 임상적 발견 개념의 하위 개념들에 포함된 수는 118,729개이며, 그 중

interprets 속성과 has interpretation 속성을 가진 개념은 각각 31,764개와 15,300개로 나타났다. 두 속성을 모두 가지고 있는 개념은 15,198개로서 has interpretation 속성을 가진 개념 중 99.3%에 해당하는 개념이 interprets 속성을 함께 가지고 있음을 확인할 수 있다.

두 속성을 모두 가진 개념 중 그 부모 개념이 동일한 interprets 속성 값을 가진 경우는 12,798건으로서 수식 (1)의 규칙을 만족하는 경우의 비율이 84.2%에 해당하는 것을 확인할 수 있다. 나머지에 해당하는 2,400건은 자식 개념의 interprets 속성이 그 부모 개념과 일치하지 않는 경우이며, 이를 수식(3)에 따라 두 개 이상의 형제 개념으로 구성된 형제 개념 그룹으로 분류한 결과 그 그룹의 수는 총 564건으로 나타났다.

Table 2. Results of the detection of sibling concept groups

Description	Count	Remark
Target concepts (Clinical finding)	118,729	
Concepts having *interprets* attribute	31,764	
Concepts having *has interpretation* attribute	15,300	
Concepts having both *interprets* and has *interpretation* attributes	15,198	
Concepts having a parent with the same *interprets* attribute	12,798	Eq. (1)
Concepts does not having parents with the same *interprets* attribute	2,400	Eq. (2)
Sibling concept group	564	Eq. (3)

앞서 확인된 564건의 형제 개념 그룹인 $S(x,p)$들에 대해 수식 (5)에 따라 부모 개념이 다르나 공통적으로 존재하는 형제 개념 그룹들을 interprets 속성 값을 기준으로 군집화 한 결과는 총 399건으로 나타났는데 이를 그림 4에 따라 분류한 결과는 표 3과 같다. 그 중 298건은 다중 부모가 아닌 경우로 이 경우는 수식 (4)에 따라 형제 개념 그룹의 공통 속성인 interprets 속성을 이용하여 새로운 부모 개념을 생성하고 이를 기존 부모 개념의 자식으로 설정하는 방법으로 누락 개념을 생성할 수 있다. 본 연구를 통해 분석된 형제 개념 그룹의 사례는 표 4에 제시하였다.

Table 3. Classification of sibling concept groups based on common attributes

# of Parents	Case	Count	Missing concepts
Single	-	298	298
Multiple	equal	53	53
	containment	31	31
	intersection	2	2
	no common	15	-
Total		399	384

그 외의 경우는 두 개 이상의 부모 개념이 존재하는 경우로서 표3에서 두 개 이상의 부모 개념이 존재하는 경우는 총 101건이다. 형제 개념 그룹이 여러 부모에 동일하게 연결되어 있는 경우가 53건으로 가장 많으며 형제 개념 그룹들 간에 서로 포함관계에 있는 포함 경우가 31건 발견되었다. 그리고 형제 개념 그룹들에 포함된 개념들 중 일부만 공통적으로 존재하는 교차 경우가 2건 발견되었다. 이 경우에는 수식 (5)와 같이 형제 개념 그룹들의 교집합에 대해 새 부모 개념을 생성할 수 있다.

형제 개념 그룹들이 서로 공통되는 개념을 두 개 이상 갖고 있지 않은 개별 관계는 총 15건 발견되었다. 표 4의 예시에서 interprets 속성의 값이 감각 인지 기능(Sensory perception function)인 형제 개념 그룹의 집합은 총 3개이나 서로 공통적인 개념을 갖고 있지는 않다. 또 다른 예시로 interprets 속성의 값이 신장 측정(Body height measure)인 형제 개념 그룹의 집합은 형제 개념 그룹의 수가 29개에 달하고 공통으로 존재하는 개념이 발견되지 않았다. 이와 같이 개별 관계에 해당하는 경우에는 3장에서 제시한 바와 같이 본 알고리즘을 적용하기는 어려우며 전문가의 검사 또는 개념의 어휘적, 구조적 특성을 추가로 고려하여 누락 여부를 판단할 필요가 있다. 최종적으로 본 누락 개념 탐지방법을 이용하면 개별 관계를 제외하고 총 384개의 잠재적 누락 개념을 자동으로 탐지할 수 있었다.

V. Conclusions

본 논문에서는 SNOMED CT 용어체계에서 속성 관계를 이용하여 누락된 개념을 탐지하는 방법을 제안하였다. 제안한 방법은 interprets 속성과 has interpretation 속성이 개념의 계층 관계 내에서 규칙을 띄고 있음에 착안하여 interprets 속성이 제대로 설정되어 있지 않은 형제 개념 집합을 탐지하고 이를 통해 누락된 개념을 생성할 수 있는 방법을 제안하였다. 그리고 이를 용어체계에 적용하여 분석한 결과 총 384개의 잠재적 누락 개념을 탐지할 수 있었다. 본 연구에 따라 탐지된 잠재적 누락 개념을 생성하면 용어체계 내에 내재된 규칙에 따른 계층 구조를 보다 일관성 있게 표현할 수 있으며, 이를 통해 각 개념의 범주와 그 의미를 보다 일관적이고 명확하게 제시할 수 있다. 계층 구조의 일관성을 높이기 위해 개념의 어휘적 특성 및 계층 관계를 주로 고려한 기존 연구와 달리 속성에서 특성을 도출한 점, 그리고 이를 이용하여 자동화된 탐지방법을 제안

Table 4. An example of classification of sibling concept groups with multiple parents

Case	Parents	Children (sibling concept group)	Common interprets attribute of siblings
equal	Stomach finding	**Abnormal gastric motility** **Increased gastric motility**	Gastric motility, function
	Functional finding	**Abnormal gastric motility** **Increased gastric motility**	
containment	Stool finding	Abnormal feces **Feces normal** **Occult blood detected in feces**	Evaluation of stool specimen
	Evaluation finding	**Feces normal** **Occult blood detected in feces**	
intersection	Functional finding	Increased hormonal activity Normal endocrine system function **Absence of hormonal activity** **Endocrine system alteration**	Endocrine function
	Endocrine finding	**Absence of hormonal activity** **Endocrine system alteration** Decreased hormonal activity	
no common	Sensory nervous system finding	Disturbed sensory perception Normal sensory perception	Sensory perception function
	Altered perception	Disturbed sensory perception Illusion	
	Cognitive function finding	Normal sensory perception Numbness	

한 점에서 본 연구의 의의가 있다. 이에 기존 연구와 직접적인 성능 비교는 어려우나 용어체계의 완성도를 높이는 방법적 측면에서 다른 연구와의 차별성을 제시하였다. 다만, 분석 결과에서 제시한 15건의 개별 관계는 제안한 자동화 알고리즘을 통해 누락 여부를 판별하기에는 한계가 있고 전문가의 수동 검사가 필요한 영역이나 향후 연구를 통해 개념의 위상 구조, 어휘적 특성을 종합적으로 고려하여 누락 여부를 판단하는 보다 개선된 자동화 알고리즘으로 고도화될 필요가 있다.

ACKNOWLEDGEMENT

This paper was supported by RESEARCH FUND offered from Catholic University of Pusan (2022).

REFERENCES

[1] H. A. Park, S. J. Yu, and H. Jung, "Strategies for Adopting and Implementing SNOMED CT in Korea," Healthcare Informatics Research, Vol. 27, No. 1, pp. 3-10, Jan. 2021. DOI: 10.4258/hir.2021.27.1.3

[2] SNOMED International, Data Analytics with SNOMED CT, *http://snomed.org/analytics*

[3] F. A. Navarro, M. Q. Martinez, A. D. Ramos, et al, “Analysis of readability and structural accuracy in SNOMED CT,” BMC Medical Informatics and Decision Making, Vol. 20, pp. 1-21, Dec. 2020. DOI: 10.1186/s12911-020-01291-y

[4] M. Amith, Z. He, J. Bian, et al, “Assessing the practice of biomedical ontology evaluation: Gaps and opportunities,” Journal of biomedical informatics, Vol. 80, pp. 1–13, Apr. 2018. DOI: 10.1016/j.jbi.2018.02.010

[5] X. Hao, R. Abeysinghe, J. Shi, and L. Cui, "A substring replacement approach for identifying missing IS-A relations in SNOMED CT," 2022 IEEE International Conference on Bioinformatics and Biomedicine, pp. 2611-2618, Las Vegas, USA, Dec. 2022. DOI: 10.1109/BIBM55620.2022.9995595

[6] F. Zheng, J. Shi, and L. Cui, “A lexical-based approach for exhaustive detection of missing hierarchical IS-A relations in SNOMED CT,” AMIA Annual Symposium Proceedings, Vol. 2020, pp. 1392-1401, Nov. 2020.

[7] R. Hu, J. Shi, L. Cui, et al, “An Automated Approach for Identifying Errornous IS-A Relations in SNOMED CT,” AMIA Summits on Translational Science Proceedings 2024, Vol. 2024, pp. 545-554, May 2024.

[8] R. Abeysinghe, F. Zheng, J. Shi, et al “Leveraging logical definitions and lexical features to detect missing IS-A relations in biomedical terminologies,” Journal of Biomedical Semantics, Vol. 15, pp. 1-12, May 2024. DOI: 10.1186/s13326-024-00309-y

[9] R. Burse, G. McArdle, and M. Bertolotto, “Targeting stopwords for quality assurance of SNOMED-CT,” International Journal of Medical Informatics, Vol. 167, pp. 1-8, Nov. 2022. DOI:

10.1016/j.ijmedinf.2022.104870

[10] A. Agrawal, and K. Qazi, “Detecting modeling inconsistencies in SNOMED CT using a machine learning technique,” Methods, Vol. 179, pp. 111-118, Jul. 2020. DOI: 10.1016/j.ymeth.2020.05.019

[11] X. Hao, R. Abeysinghe, K. Roberts, et al, “Logical definition-based identification of potential missing concepts in SNOMED CT,” BMC Medical Informatics and Decision Making, Vol. 23, No. 87, pp. 1-17, May 2023, DOI: 10.1186/s12911-023-02183-7

[12] Zhang J, Zhao Y, Saleh M, et al, “Pegasus: Pre-training with extracted gap- sentences for abstractive summarization,” Proceedings of the 37th International Conference on Machine Learning, PMLR 2020, Vol 119, pp. 11328−11339, Jul. 2020.

[13] R. Abeysinghe, F. Zheng, E. V. Bernstam, et al, “A deep learning approach to identify missing is-a relations in SNOMED CT,” Journal of the American Medical Informatics Association, Vol. 30, No. 3, pp. 475-484, Mar. 2023. DOI: 10.1093/jamia/ocac248

[14] X. Huang, F. Zheng, L. Jiang, et al, “A Vector Similarity-based Knowledge Graph Embedding Method for Predicting Missing Hierarchical Relationships in SNOMED CT,” 2024 4th International Conference on Neural Networks, Information and Communication (NNICE), pp. 16-20, Guanghou, China, Jan. 2024. DOI: 10.1109/NNICE61279.2024.10498282

[15] W. Ryu, “Identifying missing concepts in SNOMED CT through similarity analysis of sibling nodes,” Proceedings of the 16th International Conference on Future Information & Communication Engineering, pp. 230-232, Kota Kinabalu, Malaysia, Jan. 2024.

[16] National Library of Medicine, SNOMED CT International Edition, *https://www.nlm.nih.gov/healthit/snomedct.*

Authors

Wooseok Ryu received the Ph.D. degrees in Computer Engineering from Pusan National University, Korea in 2012. Dr. Ryu joined the faculty of the Department of Health Care Management at Catholic University of Pusan, Busan, Korea, in 2013. He is currently a professor in the Department of Health Care Management at Catholic University of Pusan. He is interested in clinical terminologies, distributed processing, and machine learning.