

진료서술문의 구조적인 입력을 위한 온톨로지 기반의 통제자연어와 인터페이스 개발

하병현, 김홍기¹, 이재일², 김명기³

부산대학교 공과대학 산업공학과, 서울대학교 치과대학 치과외료와정보기술전공¹,
서울대학교 치과대학 구강병리학교실², 서울대학교 치과대학 치과경영정보학교실³

How to Deal with Medical Narratives based on Controlled Natural Language Operated by an Ontology

Byung-Hyun Ha, Hong-Gee Kim¹, Jae-Il Lee², Myeng-Ki Kim³

Dept. of Industrial Engineering, College of Engineering, Pusan National Univ.;
Dept. of Dental Care and Information Technology, School of Dentistry, Seoul National Univ.¹;
Dept. of Oral Pathology, School of Dentistry, Seoul National Univ.²;
Dept. of Dental Management & Informatics, School of Dentistry, Seoul National Univ.³

Abstract

Objective: Medical narratives entry is a major issue to be solved in developing an electronic medical record system operating in practice, as they are, in large part, described in a free-text format. The issue can be dealt with in three aspects: to improve the reusability by structuring medical narratives, to support clinical pragmatics in medical fields, and to reduce the burden of data entry. With the aspects having in mind, this paper purports to present an ontological method for better way of medical narratives entry. **Methods:** We developed an ontology for which medical knowledge is structurally represented. Then we can enter medical narrative texts with commands of the controlled natural language operable on the ontology model. **Results:** Many theoretical studies on free-text entry were reviewed, based on which an authoring and editing tool for natural language description operable on the ontology model has been developed and tested. The performance of the tool is satisfactory within the limit of the domain models we developed here. **Conclusion:** The results of this paper are contributive for clinicians to make an easy entry of medical narratives as far as the ontology model covers their knowledge domain. It is also expected that the cost in recording medical narratives might be considerably reduced and data quality can be improved. (*Journal of Korean Society of Medical Informatics 12-3,261-271, 2006*)

Key words: EMR, Medical Narrative, Controlled Natural Language, Ontology, User Interface

논문투고일: 2006년 7월 29일, 심사완료일: 2006년 9월 18일

교신저자: 김명기, 110-749, 서울시 종로구 연건동 28번지 서울대학교 치과대학 505호

전화: (02) 740-8791, Fax: (02) 743-7633, E-mail: meeree@snu.ac.kr

* 본 연구는 "의료정보 기술개발사업(0412-MI02-0404-0002)"의 지원에 의하여 연구되었음.

I. 서 론

전자의무기록(EMR: Electronic Medical Records) 시스템은 의료서비스 제공자에게 필수적인 정보시스템이다¹⁾. 전통적으로 사용하던 하드-카피 의무기록과 비교하여 전자의무기록은 가용성(availability)과 재사용성(reusability)이라는 측면에 장점이 있다. 전통적인 의무기록은 한 시점에 한 곳에서만 존재할 수 있는 반면, 전자의무기록은 필요한 모든 사람들이 동시에 열람할 수 있다. 또한 전자의무기록의 자료는 구조적인 형태를 갖추어야 하므로 의사결정을 위한 정보로 변환이 용이하며, 의료 행위와 결과의 상관관계의 도출과 연구용 자료로 활용될 수도 있다.

의무기록을 구성하는 요소는 크게 구조적인 자료와 서술적 자료로 구분된다. 서술문은 구조적인 형태가 아닌 자연어를 사용하여 기록된 자료로, 병력과 검사기록, 경과기록, 특정절차(방사선검사, 병리검사 등)의 보고서를 작성하는데 사용되며, 이와 같은 자료는 의무기록에서 중요한 역할을 담당한다. Sharer는 의무기록의 70% 정도가 진료서술문(medical narratives)으로 구성되어있다고 하였다²⁾. 따라서 전자의무기록의 장점을 최대한 활용하기 위해서는 진료서술문을 체계적으로 획득하고 활용하는 것이 매우 중요하다³⁻⁶⁾. 본 연구에서는 이와 같은 진료서술문을 쉽게 획득하고 구조적으로 저장하기 위한 방법을 제시하였다.

진료서술문을 효과적으로 전산화하기 위해서는, 먼저 하드-카피 의무기록(paper-based medical record)에 익숙한 의료관계자가 쉽게 이용할 수 있는 사용자 인터페이스를 제공할 수 있어야 하며, 동시에 체계적인 표현을 기반으로 획득 자료의 재사용성을 높일 수 있어야 한다⁴⁾⁵⁾. 구조적 자료의 경우 이러한 목표를 달성하기는 비교적 쉬우며, 다양한 방법론이 제시되어 사용되고 있다³⁾⁴⁾⁷⁾. 반면 진료서술문의 처리와 활용은 아직 초보적인 단계로, 사용성 높은 전자의무기록을 구축하는데 주요한 걸림돌이 되고 있는 것으로 파악된다.

저자들은 이전 연구에서 진료서술문을 구조적으로 표현하고 저장하는 방법을 제시하고, 그것을 기반으로 한 입력도구를 개발하였다⁸⁾. 하지만 이전 연구에서 제안된 입력도구는 하드-카피 의무기록을 사

용해온 의료종사자에게 익숙하지 않다는 문제점을 지니고 있다. 따라서 본 연구에서는 동일한 표현 구조를 기반으로 하되, 사용자가 보다 쉽게 접근할 수 있는 통제자연어(CNL: Controlled Natural Language)를 고안하고 그것을 효과적으로 활용할 수 있는 입력도구를 개발하였다. 이러한 통제자연어와 활용도구는 진료서술문의 입력에 드는 노력을 적절한 수준까지 낮출 수 있을 것이며, 결과적으로 보다 효율적인 자료 획득에 기여할 것으로 기대된다.

본 논문의 구성은 다음과 같다. 2절에서는 진료서술문을 전산화하는 기존의 방법들과 그들이 가진 한계를 살펴보고, 사용자에게 친숙한 입력 인터페이스의 필요성을 제시한다. 3절에서는 통제자연어 개발에 바탕이 되는 온톨로지와, 현재 진행중인 통제자연어 관련 연구들을 일람한다. 4절에서 본 연구의 결과물인 온톨로지 기반의 통제자연어와 그것을 효과적으로 활용할 수 있는 도구를 설명한 후, 마지막으로 전자의무기록 시스템에 적용될 때 고려할 요소들을 논의한 후 끝을 맺는다.

II. 진료서술문의 전산화

현재 전자의무기록 시스템을 구축한 일선 병원에서는 진료서술문의 전산화를 위해 대부분 다음의 두 가지의 방법을 사용한다. 먼저, 의무관계자가 수기한 내용을 그림 형태로 획득하여 저장하는 것이다. 이 방법은 하드-카피 의무기록을 사용하던 기존관행을 바꿀 필요가 없어 쉽게 적용할 수 있지만, 획득한 자료를 연구분석의 목적을 위해 재사용하기 힘들다는 단점이 존재한다²⁾. 다른 한 가지 방법은 의무관계자 또는 보조자가 자유 텍스트 형태로 진료서술문을 입력하여 저장하는 것이다. 이와 같은 경우 의무기록을 높은 수준으로 재사용하기 위해서는 자연어처리(NLP: Natural Language Processing)를 통한 구조적인 정보의 추출이 필수적이다. 하지만 의료분야에서 자연어처리는 아직 초보적인 단계로 현재로서는 적극적으로 활용하기 쉽지 않다⁹⁾. 또한, 입력된 자료의 재사용을 위해서는 획득 단계에서부터 용어를 통제할 필요성이 존재한다⁶⁻⁸⁾. 이러한 관점에서 위의 두 방법 모두 자료를 입력하는 시점에서 용어를 통제하기 힘들다는 한계를 지니고 있다.

통제용어(controlled vocabulary)를 바탕으로 진료 서술문을 기록하는 대표적인 방법은 LOINC(Logical Observation Identifiers, Names and Codes), ICNP(International Classification of Nursing Practice) 등의 표준화된 조합형 용어체계를 사용하는 것이다. 예를 들어, ICNP에서는 7개의 축을 사용하여 간호 진단, 중재, 산출을 서술한다. 즉, 하나의 문장을 장소와 시간, 행위 등 7가지 요소로 분해한 후, 각각의 축에 적절한 용어를 부여하는 방식으로 그 문장을 구조적으로 표현할 수 있다. 하지만, 이들 표준용어 체계들은 조합하여 만들어진 사실들을 재조합할 수 있는 방법이 제공되지 않을 뿐 아니라¹⁰⁾, 상세함과 유연함이 상대적으로 부족하여 복잡한 현실을 표현하기에는 적합하지 않다⁴⁾. 결과적으로 적용되는 환경에 적절한 용어체계를 만들어 사용하는 경우가 대부분이다.

앞서 제시한 방법들의 한계로 인해, 현재 진행되고 있는 진료서술문의 구조화된 입력에 관한 연구들에서는 적절한 수준의 통제용어와 조합체계를 자체적으로 개발하여 사용하는 것이 추세이다. 개발된 용어체계는 표준 용어체계와의 상호연관을 통해 그 의미가 정의되며, 용어를 조합하는 방식으로 관계형 데이터베이스(RDB: Relational Database) 기반의 데이터 모형이 사용되거나⁷⁾¹¹⁾, XML(Extensible Markup Language) 등의 트리(tree) 형식의 데이터 모형⁴⁾⁹⁾¹²⁾¹³⁾, 또는 온톨로지 등의 형식용어 시스템 등이 많이 사용된다⁸⁾¹⁴⁾¹⁵⁾.

이러한 방법들은 사용분야에서 필요로 하는 재사용 수준에 따라 표현력과 상세함을 통제할 수 있을 뿐 아니라, 특정 시점까지 입력된 내용을 바탕으로 다음에 입력할 수 있는 내용을 제한하거나 예측할 수 있다는 장점을 가진다. 진료서술문의 입력하는 관점에서 보면, 이와 같은 장점은 쉽고 빠른 입력이 가능한 사용자 인터페이스를 만들 수 있는 기반을 제공한다. 그 결과 많은 연구들은 사용자의 입력 내용에 반응하는 동적 데이터 입력 폼(form)을 구현하거나³⁾⁷⁾¹³⁾, 반복적으로 입력되는 형식에 대해 템플릿 형태로 폼을 설계하고⁵⁾⁸⁾⁹⁾, 이 두 가지 방식을 혼용한 입력 인터페이스를 제공하고 있다⁴⁾. 또한 GALEN 등의 체계적인 의료 지식을 사용하여 사용자에게 선택 가능한 용어를 제시해주는 지능적인 입력 인터페

이스를 구현하거나¹⁵⁾, 유사한 방식으로 그래픽 사용자 인터페이스를 구현하는⁸⁾ 연구도 진행되고 있다.

하지만 위의 연구들을 사용자 편의성의 관점에서 보면 크게 두 가지의 단점이 존재한다. 먼저, 이들은 대부분 서술문을 입력하기 위해 폼 기반의 구조적 입력양식을 채택한다. 이와 같은 양식을 사용하는 경우 본질적으로 하나의 용어를 선택하기 위해 수많은 용어들을 검색할 수밖에 없으며, 또한 폼 기반의 인터페이스가 가지는 다양하게 배치되어있는 항목들을 각기 채워나가야 하는 단점을 지닌다⁷⁾. 다음으로, 사용자는 코드화된 방식을 통해 서술문을 입력한 경우라도, 그것을 다시 볼 때는 자연어 문장으로 기술된 형태를 더 선호할 뿐만 아니라 보다 잘 이해한다는 것이 연구되었다⁶⁾. 따라서 폼 기반의 인터페이스를 통해 입력된 서술문을 자연어 문장으로 재구축해야 하는 쉽지 않은 과제가 필요하다¹¹⁾¹⁶⁾.

결론적으로 진료서술문을 효과적으로 입력하고 저장하기 위해서는 통제된 용어체계를 사용하면서도 하드-카피 의무기록을 사용하는 관행을 크게 벗어나지 않는 사용자 인터페이스가 개발될 것이 요구된다. 이러한 인터페이스는 키보드 또는 마우스를 사용하여 자유롭게 입력하면서도 정해진 용어만 제한하여 입력할 수 있어야 하며, 또한 효율적인 입력을 위해 가능한 용어들 중에서 적절한 것을 추천할 수 있어야 한다⁷⁾. 또한, 사용자가 입력된 자료를 다시 볼 때 내용 파악에 어려움이 없도록, 입력하는 방식 자체가 자유 텍스트를 입력하는 경우와 다르지 않다면 더 쉽게 사용될 수 있을 것이다.

본 연구에서는 이전 연구⁸⁾에서 개발한 데이터 모형을 바탕으로 이러한 요구조건을 만족하는 입력 도구를 개발하였다. 하지만 다른 데이터 모형이라도 구조화의 정도가 충분하다면 본 연구에서 제시하는 방식을 적용하는 데 어려움이 없을 것으로 생각된다.

III. 방 법

이 절에서는 먼저 사용자에게 친숙한 입력 인터페이스를 구축하기 위해 본 연구가 채택한 통제자연어와 관련된 연구를 개괄한 후, 의료환경에서 효과적으로 사용되기 위해서는 온톨로지를 적용할 필요성

을 제시한다. 그리고 온톨로지 언어인 RDF 스키마 (RDF Schema)를 사용하여 진료서술문을 구조적으로 기록하는 방식을 간략하게 설명한다.

1. 지식 표현을 위한 통제자연어

통제자연어는 모호성과 복잡성을 줄이기 위해 문법과 어휘를 제한한 자연어의 부분집합으로, 컴퓨터가 텍스트를 보다 쉽게 처리할 수 있는 기반을 제공한다. 전통적으로 통제자연어는 기업이나 제품의 설명서를 명료하게 작성하기 위해 사용되었다. 이렇게 통제된 언어로 작성된 설명서는 모호성이 줄어들다는 이점 외에도, 타 언어권의 사용자라도 쉽게 이해할 수 있는 장점을 제공한다. 대표적으로 비행기의 유지관리 설명서를 작성하기 위해 AECMA Simplified English 등이 개발되어 활용되고 있다. 또한 컴퓨터가 쉽게 처리할 수 있는 장점을 가지므로 기계번역을 목적으로 다양한 언어를 위한 통제자연어가 개발되어 사용되고 있다.

지식 표현의 영역에서의 통제자연어는 자연어와 형식언어(일차술어논리 등)의 중간 지점에 위치한다고 볼 수 있다. 따라서 통제자연어는 자연어가 가지는 사용자 친숙성과 형식언어가 가지는 잘 정의되고 명확한 의미론과 자동추론기능의 동시에 취한다는 장점을 가진다. 이러한 범주에 해당하는 전형적인 통제자연어로 ACE(Attempto Controlled English)¹⁷⁾, CLCE(Common Logic Controlled English)¹⁸⁾, PENG (Processable ENGLISH)¹⁹⁾, Metalog²⁰⁾ 등이 개발되었다. 통제자연어로 작성된 문장들은 대부분 일차술어논리 또는 그것의 부분집합으로 모호성 없이 변환된다. 예를 들어, 다음의 ACE로 표현된 문장(1)은 (2)의 일차술어논리로 해석된다. 이렇게 명확하게 해석된 지식은 컴퓨터가 보다 쉽게 활용하고 추론하는, 즉 재사용할 수 있는, 기반을 제공한다.

The manager of a department is an employee.
He leads the department. (1)

$(\exists x,y) [Manager(x) \wedge Department(y) \wedge$
 $Of(x,y) \wedge Employee(x) \wedge Lead(x,y)]$ (2)

위에서 제시된 통제자연어 연구를 진료서술문의 표현에 적용하는 데는 크게 두 가지의 어려움이 존재하는 것으로 파악된다. 먼저, 진료서술문은 의료지식이 아닌 개개의 사실들을 구체적으로 기술하기 위한 목적으로 사용된다. 하지만, 일반적으로 통제자연어는 지식의 표현과 기계적 추론을 목표로 하여, 개별 사실들의 표현과 기록에는 많은 중요성을 두지 않는다. 다음으로, 특정 분야에 미리 만들어져 있는 공유된 용어체계인 온톨로지에 대한 충분한 고려가 없다. CLCE의 경우 외부 어플리케이션 시스템과의 통합에 대한 제안은 하고 있지만, 구체적인 스키마의 사용법에 대한 해법을 제시하지 못하고 있으며, PENG의 경우 온톨로지를 구축하는 방법을 제시하고 있지만, 이미 구축되어있는 온톨로지의 활용에 관해서는 아직 연구가 부족한 것으로 보인다.

따라서 이미 구축되어 있는 용어와 지식체계를 사용하여 진료서술문을 기록하기 위해서는 새로운 형태의 통제자연어의 개발이 요구되며, 본 연구에서는 이러한 목적에 적합한 통제자연어인 DESC(Descriptive Element for Semantic Corpus)를 개발하였다. 비록 현재 DESC는 다음 절에서 소개할 RDF 스키마의 지식 체계를 바탕으로 구현되었지만, MUDR의 지식베이스⁷⁾ 또는 Archetypes¹⁴⁾, GALEN¹⁵⁾, ICNP 등의 지식 체계에도 쉽게 적용할 수 있을 것으로 예상된다. 참고로, GALEN의 경우 GRAIL(GALEN Representation and Inference Language)과 IR(Intermediate Representation)이라는 보다 쉽게 표현하기 위한 문법이 존재한다. 하지만, 그 문법이 자연어가 아니라 형식언어에 기반하고 있어 본 연구에서 제시하는 DESC의 사용자 친숙도 수준에는 미치지 못하는 것으로 파악된다. 또한 GALEN은 그 복잡함과 상세함으로 인해 도메인 전문가가 사용하기에 적합하지 않은 것으로 보고되었다⁸⁾.

통제자연어는 그것의 저작을 지원하는 편집도구와 함께 개발되는 것이 일반적이다. 본 연구에서도 DESC의 문장을 쉽게 입력하고 편집할 수 있는 도구인 DESC 편집기를 개발하였다. 정해진 문법과 어휘를 가진다는 점에서 통제자연어 또한 C 또는 Java 등의 컴퓨터 프로그래밍 언어라고도 볼 수 있다. 따라서, 많은 전문 편집기들이 특정 프로그래밍 언어를 지원하여 보다 효율적인 입력을 가능하게 하는

것과 같이²¹⁾, DESC 편집기 또한 추천과 안내의 방식을 통해 DESC로 된 서술문 입력을 지원한다.

2. RDF 기반의 진료서술문의 표현

본 연구에서는 단순한 형태의 온톨로지인 RDF 스키마를 사용하여 진료서술문을 표현하는 수 있는 용어체계를 구축하고, 그 용어체계에 기반하여 서술문을 구조적인 RDF 문서로 기술한다. DESC는 구축되어있는 RDF 스키마를 바탕으로 유효한* RDF 문서를 쉽게 작성하기 위한 통제자연어이다. 이 절에서는 RDF 스키마와 RDF를 간략하게 소개한다. 보다 상세한 내용은 RDF를 소개한 다른 연구를 참조할 수 있다²²⁾.

RDF 스키마에서는 클래스(class)와 프로퍼티(property)의 개념을 사용하여 현실을 기술하기 위한 용어체계를 구축한다. 클래스는 특정 자원(resource: 개체 또는 객체)의 원형(type)을 선언하기 위해 사용하며, 프로퍼티는 자원의 속성(attribute)을 선언하기 위해 사용한다. 특정 클래스를 원형으로 가지는 자원들이 가질 수 있는 속성을 제한하기 위해 도메인(domain)의 개념이 사용되며, 속성이 가질 수 있는 값을 제한하기 위해 레인지(range)의 개념이

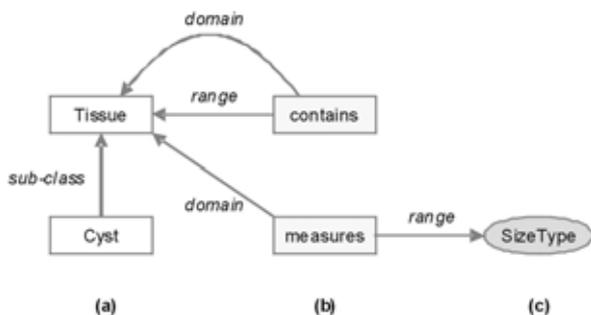


Figure 1. An example of an ontology in the form of RDF Schema. (a) Classes, (b) properties, (c) data types, and the arrows which represent relationship among concepts

* 본 연구에서는 RDF 스키마를 RDF 문서가 지켜야 하는 규칙으로 취급한다. 참고로, 지식표현 시 이러한 방식으로 RDF 스키마와 RDF 문서의 관계를 규정하지 않을 수도 있다.

사용된다. 예를 들어, 육안소견의 부분인 문장(3)을 표현하기 위한 RDF 스키마는 Figure 1과 같다.

Specimen received contains a cyst which measures 2×1cm. (3)

이 예에서 실제적인 개체인 'specimen received'는 Tissue로, 'a cyst'는 Cyst로 개념화되었다. 또한 모든 cyst는 tissue라고 할 수 있으므로 이 두 개념들은 sub-class의 관계를 가지고 있는 것이 표현되었다. 특정 tissue가 가질 수 있는 속성을 개념화하기 위해 프로퍼티인 contains와 measures를 사용한다. 그리고 특정 tissue가 다른 tissue를 포함할 수 있는 것과, 크기를 가진다는 것을 명세하기 위해 각각의 프로퍼티와 적절한 다른 개념들 사이에는 도메인과 레인지의 관계가 표시되어 있다.

Figure 1의 온톨로지를 사용하여 문장(3)을 RDF로 기술하면 Figure 2와 같다. RDF 문서를 구성하는 기본적인 구조는 주어-동사-목적어(subject-predicate-object)의 트리플(triple)이다. Figure 2 (a)는 트리플 형식을 사용하여 위의 문장을 표현한 것이다. 편의상 문장(3)의 'Specimen received'는

```
specimen_received rdf:type Tissue .
a_cyst rdf:type Cyst .
specimen_received contains a_cyst .
a_cyst measures "2x1cm" .
```

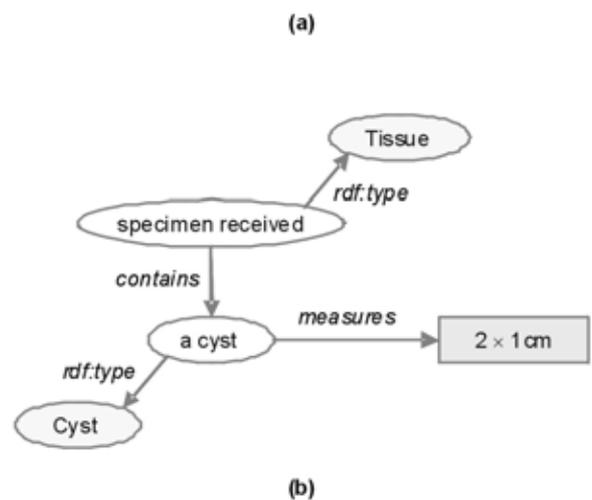


Figure 2. An example of an RDF document. A description is represented with triples (a) and graphical form (b), respectively

Sentence	::= Phrase { "and" Phrase } "."
Phrase	::= NounPhrase VerbPhrase { "and" VerbPhrase } Identifier "is" Class { "and" VerbPhrase }
NounPhrase	::= Subject Subject "which" VerbPhrase Identifier "which" "is" Class
VerbPhrase	::= Property ObjectPhrase { "and" ObjectPhrase }
ObjectPhrase	::= Object Object "which" VerbPhrase Identifier "which" "is" Class Literal
Subject	::= Identifier "it" DefInstance
Object	::= Class Identifier "it" DefInstance
DefInstance	::= "the" Class

Figure 3. DESC grammar in BNF notation

specimen_received로, 'a cyst'는 a_cyst로 각각 나타내었다. specimen_received가 하나의 Tissue라는 것은 동사인 rdf:type을 사용하여 표현한다. 이것은 specimen_received는 Tissue의 인스턴스(instance)임을 의미한다. 위의 온톨로지에 따르면 특정 tissue는 다른 tissue를 포함할 수 있으므로, specimen_received가 a_cyst를 포함하고 있다는 사실을 기술할 수 있다. '2x1cm'는 Figure 1의 SizeType 형식의 값이다. Figure 2의 (b)는 동일한 내용을 그래프 형식을 사용하여 표현한 예이다. RDF를 표현하는 그래프에서 하나의 문장은 두 꼭지점과 그것들을 연결하는 간선으로 표현된다. 간선의 시작점에 위치한 꼭지점이 주어를 의미하며, 간선은 동사를, 간선의 끝점에 위치한 꼭지점은 목적어를 나타낸다.

이 절에서 본 바와 같이 RDF를 사용하면 진료서술문을 구조적으로 표현하는 것이 가능하다. 또한 RDF 스키마는 특정 분야에서 사용하는 개념들과 그들의 연관관계를 나타내는 용어체계이므로 그 체계를 바탕으로 수집한 정보를 다양한 방식으로 재사용할 수 있는 가능성을 보장한다.

IV. 결 과

이 절에서는 RDF 스키마와 통제자연어를 바탕으로 진료서술문을 입력하기 위해 개발된 언어인 DESC를 설명하고, 사용자가 DESC를 쉽게 사용할 수 있도록 고안된 DESC 편집기의 기능을 설명한다. 본 절에서 사용되는 DESC 문장의 예제들은 Figure 1에서 제시된 온톨로지를 기반으로 서술되었다.

1. RDF 기반의 통제자연어

DESC는 RDF 스키마에서 정의한 개념(concept)과 예약어(keyword), 인스턴스를 가리키는 지시자(identifier)를 나열하여 RDF 문서를 표현한다. 현재 버전의 DESC에서 사용하는 예약어들은 다음과 같다: is, the, it, which, and. Figure 3은 DESC의 문법을 BNF(Backus Naur Form) 표기법을 사용하여 형식화된 형태로 보여준다. 본 문법에서 사용되는 Class와 Property는 온톨로지인 RDF 스키마에 정의되어 있는 개념들이다.

예약어 is는 rdf:type을 의미하며 특정 클래스의 인스턴스를 지정할 때 사용된다. 예를 들어 'specimen_received is a tissue'를 표현하기 위해서는 다음의 문장 (4)를 사용할 수 있다. 여기서 specimen_received는 새로 선언한 Tissue를 원형으로 하는 인스턴스를 가리키는 지시자이다.

specimen_received is Tissue. (4)

is를 사용하여 선언한 인스턴스들은 그것의 원형에 따라 적절한 프로퍼티를 사용하여 문장을 표현하는데 이용될 수 있다. 예를 들어 위의 문장(3)을 표현하면 문장(5)와 같다. 이 예에서는 온톨로지서 정의된 개념들인 contains와 measures를 사용하여 인스턴스들인 specimen_received와 a_cyst에 설명을 추가 한 것을 알 수 있다.

specimen_received is Tissue.
a_cyst is Cyst.
specimen_received contains a_cyst.

a_cyst measures 2×1cm. (5)

다양한 예약어를 사용하면 문장(5)를 보다 읽기 쉽게 표현할 수 있다. 이들 예약어의 용법은 자연어에서와 유사하다. 아래의 두 문장(6)과 (7)은 모두 (5)와 동일한 내용을 표현한다. 문장(7)에서는 a_cyst를 명시적으로 선언하지 않은 것을 알 수 있다. 즉 특정 개념이 is를 제외한 동사의 목적어로 사용될 때는 그 개념의 새로운 인스턴스를 선언한 것으로 간주되며, 다음 문장에서 대명사를 사용하여 지칭할 수 있다.

specimen_received which is Tissue.
the Tissue contains a_cyst which is Cyst.
a_cyst measures 2×1cm. (6)

Specimen_received which is Tissue contains
Cyst and it measures 2×1cm. (7)

전형적인 육안소견 (8)을 DESC를 사용하여 표현하면 (9)와 같다. 이 예제를 보면 DESC로 표현한 내용이 원본보다는 복잡하며, 사용자는 기존의 자연어를 사용할 때만큼 자유롭게 진료서술문을 입력하지 못하는 것을 알 수 있다. 하지만, 이것은 미리 정의된 어휘를 사용한 결과이며, 반드시 필요한 내용을 EMR 시스템이 요구하는 형식에 맞추어 제공하기 위한 불가피한 제약임을 알 수 있다.

Specimen received in formalin is a dark brown soft tissue mass measuring 4×3×1cm. It contains cystic area measuring 2×1cm. The cystic wall is thin and transparent. Typical portion is embedded. (8)

specimen_received which is SoftTissue isFixedIn formalin and hasColor darkBrown and measures 4×3×1cm. it contains Cyst which measures 2×1cm. the Cyst contains CystWall which hasThickness thin and hasTransparency transparent. TypicalPortion isProcessed embedded. (9)

참고로 위의 예제에서 사용된 병리 온톨로지는 본 저자들의 이전 연구에서 개발된 것으로, GALEN 온톨로지를 토대로 병리검사 기술에 적절하게 변형하여 만들어졌다. GALEN은 하향식 방식을 통해 체계적으로 구축된 온톨로지로서 확장성과 추론가능성을 담보한다. 본 온톨로지는 약 130개의 클래스와 100개의 프로퍼티, 40개의 데이터 형식, 그리고 230여개의 관계를 사용하여 구축되었다. 상세한 내용은 저자들의 이전 연구를 참고할 수 있다⁸⁾.

DESC를 사용하는 주요한 목적들 중 하나는 모호성을 줄이는 것이다. 하지만 사용자 편의성을 위해 여전히 자연어의 특성을 차용한다는 점에서 일반적으로 CNL은 모호성을 지닐 수밖에 없다. 예를 들어 (10)을 고려한다면 Cyst역시 Tissue이므로, 두 번째 문장의 'the Tissue'는 지시자인 specimen_received 혹은 a_cyst 모두를 지칭한다고 생각할 수 있다.[†]

specimen_received which is Tissue contains
Cyst. the Tissue measures 2×1cm. (10)

이 경우 문장을 작성하는 사람에 따라 지칭하는 지시자가 달라지게 되며, 결과적으로 표현된 내용은 작성자의 의도와 일치하지 않을 수도 있다. 더욱 심각한 것은 이러한 문제들을 CNL의 설계와 사용자 교육만으로 근본적으로 해결할 수 없다는 것이다. 따라서 사용자가 자신이 표현한 내용과 의도한 내용이 정확히 일치하는지를 확인을 가능하게 하는 체계적인 피드백 메커니즘이 필요하다. 다음 절에서는 쉬운 입력뿐만 아니라 정확한 입력을 위해서 반드시 필요한 DESC 편집기를 설명한다.

2. DESC 편집기

DESC 편집기는 DESC를 빠르고 정확하게 입력하기 위한 도구이다. 이것은 Java™ 기술을 사용하여 다양한 종류의 사용자 인터페이스 프레임워크에서 하나의 요소로 동작 가능하며, 인터넷 환경에서도

[†] 지칭의 모호성을 줄이기 위해 GRAIL 등과 같은 형식언어에서는 괄호를 사용한다. 하지만 일반적으로 괄호가 들어간 문장은 자연스럽게 읽혀지지 않는다. 따라서 본 연구에서는 이와 같은 방식을 채택하지 않았다.

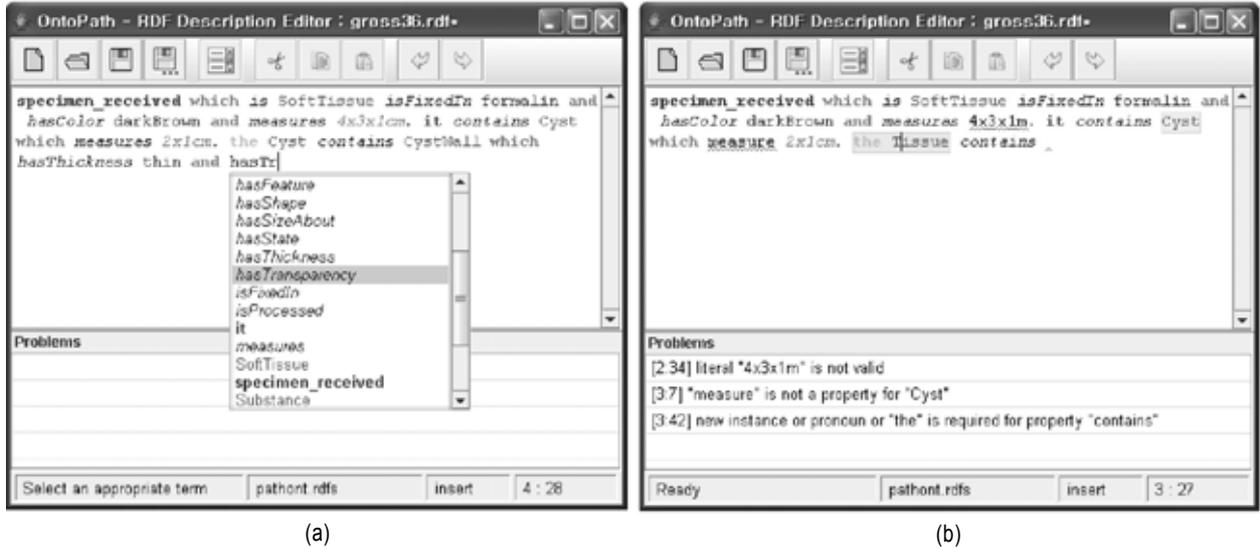


Figure 4. Screenshots of DESC editor

쉽게 활용될 수 있다. DESC 편집기는 일반적인 문서 편집기가 가지고 있는 문자의 복사와 붙여 넣기, 실행취소와 재실행 등의 보편적인 기능들을 기본적으로 지원한다. Figure 4는 독립된 상태로 실행되는 DESC 편집기의 모습을 보여준다. 독립된 상태로 DESC 편집기가 실행될 때는 파일 형태로 저장되어 있는 RDF 스키마를 읽어서 그것을 따르는 RDF 문서를 작성한 후 결과를 파일로 저장하고 읽어올 수 있다.

DESC 언어를 위해 DESC 편집기가 지원하는 기능은 크게 세 가지로 요약할 수 있다. 먼저, 입력하고 있는 컨텍스트에 따라 적절한 용어를 추천하여 빠르고 쉽게 서술문을 완성할 수 있도록 도와준다. 다음으로 실시간으로 구문론적이거나 의미론적인 오류를 확인하고 가능한 해결방안을 제시한다. 마지막으로 사용자에게 적절한 피드백을 제시해 자신이 의도한 내용과 입력된 내용의 동일성 여부를 검증할 수 있도록 한다.

Figure 4(a)는 입력 중인 문장의 내용에 따라 선택 가능한 용어를 제시하는 모습을 보여준다. 사용자는 자신이 의도한 용어를 부분적으로 입력하거나 키보드의 화살표 버튼 또는 마우스를 사용하여 적절한 용어를 선택할 수 있다. 이와 같은 기능은 입력 속도를 높이고 오류를 줄일 수 있다. 또한 현재 입력하고 있는 시점에서의 적절한 용어를 확인할 수 있

어 잘못된 개념화를 입력 단계에서부터 바로잡을 수 있다.

DESC 편집기의 선택 단어 제한 기능으로 사용자가 서술문을 차례로 입력해나갈 때는 오류가 발생할 가능성은 적어진다. 하지만 이미 입력된 문장을 수정하는 상황에서는 오류를 피하기 어렵다. 이러한 경우 DESC 편집기는 오류가 발생한 부분을 붉은 색의 밑줄을 사용하여 알려주며, 추가적인 정보 창에서 오류의 구체적인 내용을 보여준다. Figure 4(b)를 보면 세 곳에서 오류가 있으며 그 이유가 명시되어있다는 것을 알 수 있다.

사용자는 두 가지 방식의 피드백을 통해 정확한 내용이 입력되었는지 확인할 수 있다. 먼저, DESC로 기술된 서술문을 구성하는 용어는 그것의 의미에 따라 각각 다른 색깔과 형태로 보여진다. 예를 들어 프로퍼티는 청색의 기울임 꼴로 보여지며, 클래스는 분홍색으로 보여진다. 이러한 용어의 고유한 형태로 인해 사용자는 정형성을 가지는 서술문을 보다 쉽고 정확하게 파악할 수 있다. 다음으로, 사용자가 특정 용어에 커서를 위치하면 그 용어와 직접적으로 관련된 다른 용어의 배경색이 반전된다. Figure 4(b)에서 사용자는 'the Tissue'를 선택하였고, 그것이 가리키는 Cyst의 인스턴스인 Cyst가 반전되어있는 것을 볼 수 있다.

마지막으로, DESC 편집기에서는 별칭(aliasing)

기능을 통해 서술문의 입력 시 용어가 가진 고유 이름이 대신 다른 명칭을 사용할 수 있다. 이러한 별칭 기능은 크게 두 가지의 목적을 위해 사용된다. 먼저 일반적으로 특정 분야에서 사용되는 용어체계는 서술문과 독립적인 목적으로 개발되는 것이 보통이다. 따라서 보다 사용자에게 자연스러운 문장으로 보이기 위해서는 다른 별칭이 필요해진다. 예를 들어 용어체계에 표본의 색깔을 기술하기 위해 *specimen-Color*라는 개념이 있더라도 이것을 DESC에서 사용할 때는 *hasColor*라는 별칭으로 사용하는 것이 훨씬 자연스러워진다. 반면 둘 이상의 용어체계를 사용하여 서술문을 입력할 때 다른 의미를 가지면서 동일한 이름의 개념이 존재할 수 있다. 이때는 둘을 구분하기 위해 별칭을 사용하거나 혹은 적절한 접두사(prefix)를 사용하여 충돌을 해결해야 한다. 예를 들어 지질 온톨로지와 금융 온톨로지가 동시에 사용될 때 *bank*라는 다의어는 두 온톨로지에 다른 뜻으로 정의되어있을 수 있다. 이러한 경우 별칭을 사용하지 않는다면 전자를 *t:bank*로 또는 후자를 *f:bank*로 표현하여 의미의 혼동을 해결할 수 있을 것이다.

V. 고 찰

본 연구에서는 진료서술문을 구조적이면서도 쉽게 획득하고 저장하기 위해 i) 온톨로지를 사용하고, ii) 통제자연어로 표현하고, iii) 사용자 인터페이스를 활용하는 방식을 제시하였다. 본 절에서는 먼저 온톨로지 기반의 통제자연어를 사용하는 장점을 기술한 후, 실제 임상 현장에 적용할 때 고려해야 하는 요소들을 논의한다. 마지막으로 추후 연구 과제를 제시하고 논문을 마친다.

먼저, 적절하게 구축된 온톨로지가 사용된다면, DESC로 진료서술문을 획득하는 절차는 사용자가 하드-카피 의무기록을 사용하는 기존의 관행과 크게 다르지 않기 때문에, EMR 도입 시 발생하는 저항을 줄일 수 있을 것으로 기대된다. 또한 DESC를 사용하여 기록할 때 사용자 생각의 흐름은 자연어를 사용할 때와 동일하므로 컴퓨터 기반의 인터페이스를 사용할 때 발생하는 정보의 손실과 왜곡을 줄일 수 있을 것이다. 다음으로, 온톨로지에 기반한 정보의 구조적인 획득 방식은 키보드와 마우스뿐 아니라 다

양한 입력 도구의 적용을 가능하게 할 것으로 예상된다. 예를 들어, 기존의 음성인식을 통해 진료서술문을 획득하는 방식은 기술 상의 한계로 인해 활용도가 높지는 않았으나, DESC와 같은 문법을 적용할 경우 입력하는 단어의 자유도를 줄일 수 있어 보다 정확한 인식이 가능할 것으로 생각된다. 그리고 PEN&PAD¹⁵⁾의 예에서도 볼 수 있듯이 현장진단(point-of-care)에서 적절하게 사용할 수 있는 단말기를 개발하는데 활용될 수 있을 것으로 기대된다. 마지막으로, 온톨로지를 활용하는 경우 다국어 처리가 상대적으로 쉬워져²⁶⁾, 적절한 DESC의 문법이 개발된다면 동일한 지식을 사용하는 다국어 기반 시스템의 구현이 보다 용이할 것이다. 결과적으로, 이러한 시스템은 다양한 언어를 가진 사용자가 작성한 보고서를 추가적인 번역의 과정 없이 서로 이해할 수 있게 해줄 것으로 예상된다. 현재 저자들은 영어와 상이한 문법을 가지는 한글을 지원하는 DESC를 개발하고 있다.

진료서술문을 구조적으로 기록해야 하는 필요성은 명확하다. 병력, 가족력, 처치력 등의 경우 그 활용도는 언급할 필요조차 없을 것이며, 본 논문에서 고려하는 병리검사의 경우도 역학, 임상자료와 관련된 조직 정보자원, 처치와 결과의 분석, 의료관리 수준 평가 등의 다양한 분야에서 활용될 수 있다²³⁾. 그러나, 비록 DESC를 사용하여 진료서술문을 획득하는 것이 활용도 높은 의무기록을 구축하는 가능성을 제공하지만, 현재 사용되고 있는 전자의무기록 시스템에 적용하기 위해서는 몇 가지 풀어야 하는 문제들이 존재한다.

먼저, 대부분의 전자의무기록 시스템은 일반적으로 관계형 데이터 모형과 같은 비교적 단순한 형태의 정보 모형을 사용하여 구축되어있다. 하지만 DESC를 사용하기 위해서는 보다 고차원적인 형태인 온톨로지 기반 모형의 사용이 요구된다. 따라서, DESC를 무리 없이 적용하기 위해서는 전자의무기록 시스템에 온톨로지를 지원할 수 있는 기능을 추가해야 한다. 최소 수준은, DESC를 적용한 서술문을 저장할 수 있는 온톨로지 저장소를 구축하고 전자의무기록의 입력 인터페이스와 보고서, 질의 모듈이 DESC를 통합하여 처리할 수 있게 만드는 것이라고 생각된다. 하지만 온톨로지가 가진 장점을 최

대한 활용하기 위해서는 온톨로지의 관점에서 전자 의무기록 전체를 통합하고 활용하는 방법론²⁴⁾의 적용이 필요하다.

다음으로, 다양한 분야에서 사용되는 진료서술문을 위해 개별적인 온톨로지가 만들어졌을 때 이들 간의 의미론적 관계성을 부여해야 통합적인 정보 활용이 가능해진다. 한 분야에서만 사용되는 서술문을 기술하기 위해 적절한 개념들을 찾아내고 온톨로지를 구축하는 작업은 그렇게 어렵지 않을 것으로 생각된다. 하지만 서로 다른 분야에서는 동일한 개념이라도 다른 용어를 사용하여 표현하거나, 같은 용어를 사용하더라도 그 범위가 조금씩 다른 개념들이 많이 존재한다. 따라서 여러 분야에서 작성된 서술문을 모두 활용하여 재사용을 하기 위해서는 이들 간의 상호관련 정보가 필수적이며, 또한 이를 위해 다양한 방법론들이 제시되고 있다²⁵⁾.

앞서 언급한 한글의 지원에 관한 것 외의 시급한 향후 연구계획은 다음과 같다. 먼저 획득한 정보를 의사결정지원(DSS: Decision Support System) 등의 목적을 위해 보다 적극적으로 활용하기 위해서는 현재 DESC가 기반하고 있는 RDF 스키마보다 훨씬 의미론이 풍부한 온톨로지 언어를 사용할 것이 요구된다. 예를 들어 Web Wide Web Consortium에서 표준으로 제정된 OWL이 대표적이며, 현재 DESC에서도 OWL을 지원하기 위한 연구를 진행하고 있다. 하지만 보다 복잡한 형태의 언어를 사용하면 온톨로지를 구축에 보다 많은 노력이 필요하며, 또한 OWL이 의료 분야와 통제자연어를 위해 개발된 체계가 아니므로, 복잡도와 의미의 풍부함의 상관관계를 고찰한 후 적용할 것이 요구된다. 다른 하나의 연구 방향은 DESC를 사용하는데 요구되는 학습량을 최소한으로 줄이는 것이다. 실제로 구조적인 자료를 위한 인터페이스는 사용하는데 불편하기는 하지만 사용자 교육의 필요성은 상대적으로 적다. 만일 DESC가 보다 자연어와 가까운 형태로 개발한다면 역시 요구되는 학습량은 충분히 줄어들 것이다. 하지만, 이 문제도 모호성과의 상관관계를 고려하여 적절한 수준에서 결정되어야 할 것이다.

참고문헌

1. Johnson KB, Ravich WJ, Cowan Jr. JA. Brainstorming about next-generation computer-based documentation: an AMIA clinical working group survey. *International Journal of Medical Informatics* 2004;73(9-10):665-674.
2. Sharer T. Blood sweat and tears (five years of practical experience applying XML/SGML to clinical information). *Proceedings of XML Europe '99:1999 Apr 26-30; Granada, Spain.*
3. van Ginneken AM, Verkoien MJ. A multi-disciplinary approach to a user interface for structured data entry. *MEDINFO 2001, Proceedings of the 10th World Congress on Medical Informatics (Part 1):2001 Sep 2-5; London, UK.*
4. Los RK, van Ginneken AM, van der Lei J. OpenSDE: a strategy for expressive and flexible structured data entry. *International Journal of Medical Informatics* 2005;74(6):481-490.
5. Lovis C, Baud RH, Revillard C, Pult L, Borst F, Geissbuhler A. Paragraph-oriented structure for narratives in medical documentation. *MEDINFO 2001, Proceedings of the 10th World Congress on Medical Informatics (Part 1):2001 Sep 2-5; London, UK.*
6. Tange HJ, Hasman A, de Vries Robbé PF, Schouten HC. Medical narratives in electronic medical records. *International Journal of Medical Informatics* 1997;46(1):7-29.
7. Hanzlicek P, Spidlen J, Heroutova H, Nagy M. User interface of MUDR electronic health record. *International Journal of Medical Informatics* 2005;74(2-4):221-227.
8. Kim HG, Ha BH, Lee JI, Kim MK. A multi-layered application for the gross description using semantic web technology. *International Journal of Medical Informatics* 2005;74(5):399-407.
9. Yamazaki S, Satomura Y. Standard method for describing an electronic patient record template: application of XML to share domain

- knowledge. *Methods of Information in Medicine* 2000;39(1):50-55.
10. Mead CN, Henry SB. Documenting 'what nurses do' - moving beyond coding and classification. *Proceedings of AMIA Annual Fall Symposium:1997 Oct 25-29; Nashville, TN.*
 11. van Mulligen EM, Stam H, van Ginneken AM. Clinical data entry. *Proceedings of the 1998 AMIA Annual Symposium:1998 Nov 7-11; Orlando, FL.*
 12. Berman JJ. Pathology data integration with eXtensible Markup Language. *Human Pathology* 2004;36(2):139-145.
 13. Schweiger R, Brumhard M, Hoelzer S, Dudeck J. Implementing health care systems using XML standards. *International Journal of Medical Informatics* 2005;74(2-4):267-277.
 14. Beale T. Archetypes: constraint-based domain models for future-proof information systems. *Proceedings of the 11th OOPSLA Workshop on Behavioral Semantics:2002 Nov 4; Seattle, WA.*
 15. Heathfield HA, Kirby J, Hardiker NR. Data entry in computer-based care planning. *Computer Methods and Programs in Biomedicine* 1995;48(1-2):103-107.
 16. Wagner JC, Rogers JE, Baud RH, Scherrer JR. Natural language generation of surgical procedures. *International Journal of Medical Informatics* 1999;53(2-3):175-192.
 17. Fuchs NE, Schwertel U, Schwitter R. Attempto controlled English - Not just another logic specification language. *Lecture Notes in Computer Science* 1999;1559:1-20.
 18. Sowa JF. Graphics and languages for the Flexible Modular Framework. *Lecture Notes in Artificial Intelligence* 2004;3127:31-51.
 19. Schwitter R, Tilbrook M. Controlled natural language meets the semantic web. *Proceedings of the 2004 Australasian Language Technology Workshop:2004 Dec 8; Sydney, Australia.*
 20. Marchiori M. Towards a people's web: metalog. *Proceedings of the 2004 International Conference on Web Intelligence:2004 Sep 20-24; Beijing, China.*
 21. des Rivièrers J, Wiegand J. Eclipse: a platform for integrating development tools. *IBM Systems Journal* 2004;43(2):371-383.
 22. Decker S, Mitra P, Melnik S. Framework for the semantic web: an RDF tutorial. *IEEE Internet Computing* 2000;4(6):68-73.
 23. Moore GW, Berman JJ. Anatomic pathology data mining. In: Cios KJ, editors. *Medical data mining and knowledge discovery*. Berlin, Heidelberg: Springer-Verlag;1999. pp.61-108.
 24. Verschelde JL, Santos MCD, Deray T, Smith B, Ceusters W. Ontology-assisted database integration to support natural language processing and biomedical data-mining. *Journal of Integrative Bioinformatics* 2003;1. Online Journal: http://journal.imbio.de/index.php?paper_id=1.
 25. Noy NF. Tools for mapping and merging ontologies. In: Staab S, Studer R, editors. *Handbook on ontologies*. Berlin, Heidelberg: Springer-Verlag;2004. pp.365-384.
 26. Beale S, Nirenburg S, Mahesh K. Semantic analysis in the Mikrokosmos machine translation project. *Proceedings of the 2nd Symposium on Natural Language Processing:1995 May 10-12; Bangkok, Thailand.*

