반도체디스플레이기술학회지 제13권 제1호(2014년 3월)
Journal of the Semiconductor & Display Technology, Vol. 13, No. 1. March 2014.

텍스트기반 임상데이터의 인터페이스 용어 매핑 방법

유돈식† · 배인호*
†한국전자통신연구원, *카이랩

Method of The Interface Terminology Mapping based Free Text Medical Data

Done Sik Yoo† and Inho Bae*
†ETRI, *KAILab corp.

Abstract

Since 2010, issues for data sharing and data exchanging in hospital information systems have been emerged. In order to solve the issues, standards should be applied to develop the systems and there should be no ambiguities between terminologies in the systems. In this paper, the terminology mapping system for narrative clinical records was implemented. The term mapping precision was 83.4%. This system could help to upgrade the text based clinical system and it would be expected to support for high quality clinical services.

Key Words : Interface terminology, Terminology mapping, Electronic medical record

1. 서 론

2000년대 중반부터 활발히 도입이 이루어진 병원정보화시스템(HIS, Hospital Information System) 을 통해, 의료기관의 전산화가 이루어져왔고 이를 통해 병원의 모든 데이터들이 데이터베이스 시스템에 쌓이게 되었다. 그리고, 2010년 이후 병원시스템의 데이터의 공유/교환에 대한 이슈가 전세계적으로 대두되고 있다. 그러나 국내 시스템들은 표준에 대한 준비 미흡과 기관마다 사용하는 용어의 차이로 인해 교환과 공유, 통합이라는 관점에 대해 접근하기 쉽지 않은 것이 현실이다. 다기관의 임상데이터를 통한 임상연구를 위해서는 임상데이터의 구조화가 필수적인데, 이 또한 임상기록 데이터들의 구조화된 저장체계 등의 부재로 쉽지 않으며 임상데이터 중 가장 방대한 양의 데이터를 담고 있는 임상기록지의 경우는 더욱 그 처리가 쉽지 않다. 이를 위해서는 용어를 추출하고 구조화하는 방법과 다기관 임상데이터 의미호환성 유지를 위한 매핑이 필요하다. 그래서 다수의 기관간에 용어데이터를 매핑해 주기 위해 인터페이스 용어라는 개념이 등장하게 되었다. 인터페이스 용어는 용어들 간의 인터페이스를 제공하기 위한 방법으로 활용되며, 매핑을 통해 용어들간의 의미관계를 유추할 수 있게 해준다.

본 논문에서는 텍스트로 작성된 임상데이터를 기준 용어체계와 자동으로 매핑하는 Mapper을 개발함으로서 임상데이터 교환에서 의미호환성을 유지할 수 있는 방법을 제공하고자 한다.

†E-mail : dsyoo@etri.re.kr

2. 임상데이터와 처리방법

2.1. 임상데이터

임상데이터는 다음과 같이 3가지 형태로 나누어볼 수 있다.

1. 코드 형태의 데이터
2. 텍스트 형태의 데이터
3. 멀티미디어 데이터

코드형태의 데이터는 해당 코드와 타 기관의 코드와의 매핑이 비교적 쉽고, 이를 통해 의미호환성을 유지할 수 있다. 멀티미디어 형태의 데이터는 DICOM 등의 표준을 따르기 때문에 처리하는데 어려움이 적다.

그러나 텍스트 형태의 임상기록지 데이터의 경우, 병원마다 작성양식이 다르고 해당 교육기관마다 상이한 용어를 많이 사용한다. 또한, 다수의 약어와 유의어들이 사용되기 때문에 이를 처리하는 것이 쉽지 않다.

본 연구에서는 S병원의 환자의 개인정보가 제거된 임상기록지 5,000건을 바탕으로 연구하였다.

2.2. 인터페이스 용어와 참조용어

인터페이스 용어는 의료기관간의 상이한 용어 사용 및 체계들을 의미기반 매핑하기 위해, 사용되는 용어로서 의료기관의 로컬용어와 참조용어를 포함한다. 인터페이스 용어를 기반으로, 기관간의 용어들이 연결구조를 가지고, 의미적 호환성을 추적할 수 있는 방법을 제공할 수 있다. 예를 들어, 흔히 “맹장염”이라고 부르는 단어는 “충수염”, “막창자꼬리염”, “appendicitis”, “epityphlitis” 등 다양한 이름으로 불려진다. 이렇게 텍스트로 기록된 기록들은 챠트 간 의미호환성을 유지하기 어렵게 만들고 읽기 어렵게 만드는 요인이 된다. 이들을 하나의 참조용어에 매핑하여 사용하면, 문서레벨에서 용어들이 상호 의미호환성을 유지할 수 있는 방법을 제공해줄 수 있게 된다.

본 연구에서 참조용어는 SNOMED-CT, KOSTOM, KCD, UMLS, ICD를 이용하여 참조용어 데이터베이스로 구축하고, 여기에 병원의 로컬용어들을 매핑하여 인터페이스 용어 사전을 구축하였다.

2.3. 텍스트기반 임상기록지 처리

임상기록지는 자연어 형태로 기록되는데, 해당 데이터들은 다양한 약어, 전문의학용어들이 한국어와 영어가 혼재되어 기록된다. 이들을 처리하기 위한 기본 사전은 인터페이스 용어 사전을 활용하며, 그 외에 텍스트 자연어처리를 위한 사전과 약어 확장을 위한 사전이 활용하였다. 약어의 경우 진료과나 상병, 함께 사용된 용어 등에 따라 다르게 사용되는 약어들이 존재한다. 진료과는 진료시 정해지지만, 상병이나 함께 사용된 용어는 경우에 따라 다양한 형태로 활용된다. 따라서, 전처리과정에서 분류기를 통해 이를 분류하고, 이를 바탕으로 약어를 구분하는 방법을 사용하였다.

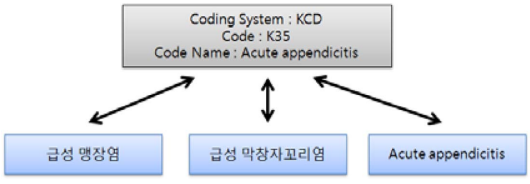

Fig. 1. Relation of the Reference terminology and the Interface terminology.

사용된 분류기는 텍스트 마이닝에서 많이 사용되는 NBC(Navie Bayes Classifier)를 이용하였다.전처리 과정은 Fig. 3과 같다. 약어 사전은 수집된 임상기록지로부터 대문자로만 이루어진 단어들을 추출하고, 해당 용어들에 대해 임상의들의 도움을 받아 구축하였다.

3. 인터페이스 용어 매핑

3.1. 용어분리

용어의 분리는 인터페이스 용어의 매핑에 있어서 중요한 부분으로서, 임상데이터의 특성에 맞춰 문서를 분리할 필요가 있다. 기본적인 임상데이터 기록은 다음과 같이 분류해 볼 수 있다.

1. 템플릿 기반 작성
2. 자연어 기술

템플릿 기반 작성은 주로 검사결과 등 결과를 기록할 때, 특정한 서식에 따라 기술하는 것을 뜻한다. 주로 검사의 결과나 습관의 기록 등에 비슷한 형태로 기록이 된다.

Smoking : 20
Drink : 1/1w
HBP : 120
SBP : 85

이런 데이터들은 문서상 위치와 템플릿을 통해 의미를 구분할 수 있다.

자연어 기술 데이터들은 한국어와 영어가 혼용되어 사용되고 다양한 약어들이 사용되어지며, 문장의 구성

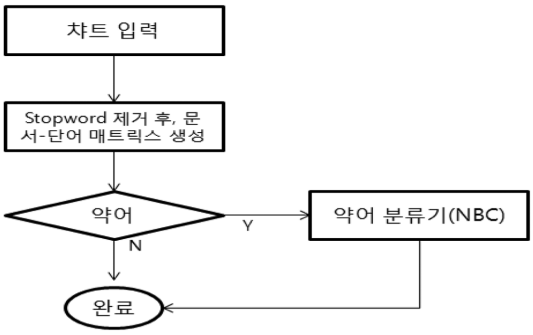

Fig. 2. Pre-processor.

이 완벽하지 않은 경우가 대부분이다. 본 연구에서는 명사구 단위로 문장을 분리하는 방법을 통해 명사구 단위로 인터페이스 용어를 검색하는 방법을 이용하였다. 명사구의 분리는 비교적 간단한 방법으로 문장규칙을 활용한 유한상태오토마타(Finite State Automata)를 이용하는 방법이 사용되거나, 기계학습방법을 이용한 분류기 모델을 만들고 이를 통해 처리하는 방법으로 나뉜다. 본 연구에서는 임상기록이 장문보다는 단문형태의 기술이 많고 문장구조가 복잡하지 않고 실제 시스템에 활용되기 위해서는 빠른 처리속도를 요구하기 때문에, 유한상태오토마타를 이용하여 명사구를 추출하였고, F-measure를 통해 약 91.3%정도의 성능을 보여주었다.

3.2. 인터페이스 용어 매핑

추출된 명사구는 인터페이스 용어와 매핑하기 위하여, TF/IDF(Term Frequency/Inverse Document Frequency)로 구성된 인터페이스 용어 인덱스와 거리를 계산하였다.

$$\mathrm{tfidf}(t, d, D) = \mathrm{tf}(t, d) \times \mathrm{idf}(t, D) \quad (1)$$

$$\mathrm{sim}(d_j, q) = \frac{\mathrm{d_j} \cdot \mathrm{q}}{|\mathbf{d_j}||\mathbf{q}|} = \frac{\sum_{i=1}^{N} \omega_{i,j}\omega_{i,q}}{\sqrt{\sum_{i=1}^{N} \omega_{i,j}^2}\sqrt{\sum_{i=1}^{N} \omega_{i,q}^2}} \quad (2)$$

TF/IDF는 식(1)을 이용하여 구성된다. 용어 매핑을 위한 계산은 벡터공간모델(Vector Space Model)을 이용하였다. 매핑 될 인터페이스 용어와의 거리계산을 통해 최소 거리의 용어를 선택하게 된다. 이 계산은, 식 (2)의 2개의 벡터간에 거리를 구하는 함수를 통해 간단히 계산된다. 인터페이스 용어에 대한 매핑 정확률은 수집된 데이터에 대해 84.3%로 나타났다.

검색 성능의 개선을 위해 키워드 확장을 일부 단어에 대해 적용하였으나, 사전의 사이즈로 인해 큰 효과는 볼 수 없었다.

4. 결 론

본 논문에서는, 텍스트 기반 임상기록지와 인터페이스 용어를 매핑하는 방법에 대해서 연구하였다. 임상기록지를 문서를 구조화하고, 임상용어를 매핑하는 것은 임상기록 기반 연구에 있어서 그 가치가 매우 높다. 임상기록지에는 환자의 입원부터 퇴원시까지 상태의 변화와 치료과정이 기록되며, 그 과정들은 임상데이터 기반 연구에 있어서 중요한 데이터로서 활용될 수 있으나, 국내에서는 해당 방법에 대한 연구가 많이 이루어지지 않고 있다. 이 연구의 결과는, 근거기반 임상연구 외에도 임상기록지에 대한 개체명인식기술(Named Entity Recognition), 기록지 익명화 기술 등의 개발에도 기반기술로서 활용이 될 수 있을 것이다.

이후 연구는, 용어의 추출방법과 용어 매핑 방법의 개선을 통해 정확률을 높이는 연구를 추가적으로 진행하고, 실제 임상기록지 작성에 적용할 수 있는 방법과 그 효율성에 대한 입증을 진행할 예정이다.

감사의 글

이 연구는 산업통상자원부 바이오의료기기 산업원천기술개발사업의 “디지털병원 전자건강기록 적용을 위한 Interface 용어 기반 임상데이터 구조화 기술 개발(10033187)과제의 지원을 받아 수행하였음.

참고문헌

1. Li, Z, George, H, “Temporal reasoning with medical data-A review with emphasis on medical natural language processing,” J. of Biomedical Informatics, Vol.40, Issue2, pp.183-202, 2007.
2. Li, L, Herbert, C, Chintan, P, Carol, F, Chunhua, W, “Comparing ICD-9 Encoded Diagnoses and NLP-Processed Discharge Summaries for Clinical Trials Pre-Screening: A Case Study,” AMIA, Annu Symp Proc, 404-408, 2008.
3. Diaz, G, Martin, V, Urena, L, “Query expansion with a medical ontology to improve a multimodal information retrieval system,” Computer in Biology and Medicine, Vol. 29, Issue 4, pp.396-403, 2009.
4. Ping, C, “A Query-Based Medical Information Summarization System Using Ontology Knowledge,” CBMS, pp.37-42, 2006.
5. Rose, D, David, M, Marek, R, “Building and using a medical Ontology for Knowledge Management and Cooperative Work in a Health Care Network,” Computer in Biology and Medicine, Vol.36, Issues 7-8, pp.871-892, 2006.
6. Young Seop. K., Jae hoon, Cho, “3D Volumetric Medical Image Coding Using Unbalanced Tree,” Journal of the Semiconductor & Display Equipment Technology, Vol.5-2, pp.19-25, 2006.

접수일: 2014년 3월 4일, 심사일: 2014년 3월 13일,
게재확정일: 2014년 3월 20일