약어와 두문자어의 모호성 해결을 통한 임상 의무기록의 정규화

배인호*, 김진상*, 김윤년**
*계명대학교 컴퓨터공학과
**계명대학교 의과대학 내과학 및 의료정보학교실
e-mail:inobae@kmu.ac.kr

Normalization of Clinical Medical Records by Disambiguating Abbreviations and Acronyms

Inho Bae*, Jin-Sang Kim*, Yoon-Nyun Kim**
*Dept of Computer Engineering, Keimyung University
**Medical School, Internal Medicine, Keimyung University

요 약

임상 의무기록에 나타나는 많은 두문자어들은 기계적인 처리과정에서 의무기록의 모호성을 크게 증가시키기 때문에, 정보추출이나 텍스트 마이닝을 하기 전에 전처리 과정으로 의무기록이 정규화 되어야 한다. 본 연구에서는 임상 의무기록 중 하나인 퇴원요약지에 사용된 약어와 두문자어들의 모호성을 제거하기 위한 정규화 시스템을 설계하고 구현했다. 정규화를 위해 문맥정보를 이용하여 의무기록의 종류와 기록내 위치정보를 파악하였고 이를 이용하여 약어와 두문자어의 의미를 학습하고 분류하였다. 본 연구에서 구현한 정규화 시스템은 실험에서 6가지 두문자어들이 가지는 16가지 의미들에 대해 94.7%의 정확률을 얻었다.

1. 서론

최근 지능형 웹 검색 환경과 전자문서로부터 지식을 찾는 연구가 활발히 진행되고 있다. 전자문서나 웹문서들로부터 지식을 발견하기 위해서는 해당 문서들로부터 원하는 정보를 추출해야 하는데, 이 과정에서 문서에 사용된 단어들의 모호성은 전체 시스템의 성능에 많은 영향을 미치게 된다. 따라서 전자문서 소스로부터 정보를 추출하기 전에 전처리 과정으로 문서 정규화에 대한 필요성과 중요성이 부각되고 있다. 문서의 정규화는 단어의 철자 오류 수정과 약어(abbreviation) 및 두문자어(acronym)를 원래 단어 및 구(phrase)로 복원하여 모호성을 해결하는 과정들을 포함한다.

임상 의무기록은 환자의 진단과 치료 과정을 기술한 문서로서 입력의 편의성을 위해 약, 검사, 진단명, 수술, 의료행위 등에 관련된 많은 용어들을 약어나 두문자로 표현한다. 약어와 두문자의 사용은 의사들의 문서 입력을 신속하게 하고 필요에 따라 리뷰를 할 때도 간결성을 제공하지만, 다양한 의미를 가진 두문자어들로 인해 해당 분야의 전문의가 아니면 모호성으로 인해 의미의 식별이 어렵다. 더욱이 의무기록에서 정보추출이나 텍스트 마이닝과 같은 기계적 처리를 하고자 하는 경우, 약어와 두문자어가 갖는 모호성으로 신뢰도와 정확도는 현저하게 떨어질 수 밖에 없다. 이러한 이유로 인해 의무기록과 같은 의료문서나 일반 문서에서 단어들의 모호성을 없애는 문서 정규화에 관한 연구가 다양하게 진행되고 있다[1-4].

본 연구에서는 임상 의무기록 중 퇴원요약지에서 약어와 두문자어들이 갖는 의미의 모호성을 해결하는 방법에 대해 기술한다.

2. 연구방법

퇴원요약지는 그림 1과 같이 주 증상(Chief Complaint), 과거 병력(Past History), 가족 병력(Family History), 현재 상태(Present Illness), 신체 검사(Physical Examination), 검사실 검사(Lab Test), 입원 경과(Hospital Course), 치료 과정(Plan) 등과 같이 여러 종류의 섹션으로 구분되어 기술된다.

```
<CHIEF COMPLAINT>-----------------------------------------------------------
        general weakness

<PRESENT ILLNESS>-----------------------------------------------------------
 known RA, osteoporosis, bronchiectasis, pacreas papillary mucin-productin(
 입원 11년전
both ankle and shoulder의 pain and swelling으로 세브란스병원 방문 RA진단
 입원  6년전
1시간 이상지속되는 mornig stiffness, polyarthralgia로 본원 방문 RA and ost(
 입원  4년전
cough c sputum, dyspnea를 주소로 내원 bronchiectasis진단받고 본원 PL OPD r(
 입원 14개월전
dyspnea로 내원 bronchiectasia c 2nd. infection  assess하에 managment받았고,
 입원 당일
1개월간 지속된 dyspnea, general weakness and general ache를 주소로 further

<PHYSICAL EXAMINATION>------------------------------------------------------
        G/A      : chronic ill appearance
        HEENT    : isocoric pupil with light reflex (+/+)
                   not pale conjunctiva
                   anicteric sclera
```

(그림 1) 퇴원요약지: 섹션별로 두문자어와 약어를 포함

임상 의무기록 문서의 섹션 정보는 약어와 두문자어의 의미를 결정하는데 중요한 단서가 된다. "AB"라는 약어가 나타날 때 검사 영역에서 이 약어가 나타난다면 다른 섹션에 비해 "antibody"를 줄인 용어일 가능성이 높다. 이와 같이 섹션의 정보는 특정 섹션에 사용되는 약어와 두문자어의 의미를 알아내는데 유용한 단서가 될 수 있기 때문에, 섹션 정보를 약어와 두문자어의 의미를 밝히는 속성으로 이용했다. 섹션 정보를 얻기 위한 방법은 섹션의 제목에 해당하는 문자열에 대한 정규형을 이용하여 섹션을 구분했다.

의료용어로서의 약어와 두문자어들은 특정 질환이나 진료과 관한 컨택스트가 주어지면 훨씬 더 그 의미를 명확하게 결정할 수 있다. 예를 들어, "RA"라는 두문자어가 류머티스 내과에서 나타났을 경우 "rheumatoid arthritis"라는 의미로, 심장내과의 경우 "right atrium"로 이용될 가능성이 높다. 이와 같은 컨택스트는 해당 문서를 어느 진료과에서 작성한 것인지 분류를 하면 결정할 수 있다. 본 연구에서는 의무기록 문서를 작성한 진료과를 결정하기 위해 단순 베이지안(Naive Bayesian) 분류기를 이용하였다.

본 연구에서는 또한 약어와 두문자어를 분류하기 위해 최대 엔트로피 모델(Maximum Entropy Model)[5,6]을 이용하였다. 최대 엔트로피 모델은 최근 자연어처리 분야에서 문서번역 시스템, 구문 분석, 문장 경계 분석[7], 형태소 태깅[8] 등 여러 가지 문제해결을 위해 많이 이용되어 왔다. 그 이유는 분류 문제를 확률적 모델링으로 해결하는데 있어서 최대 엔트로피 모델과 같은 조건부 확률 모델이 HMM과 같은 결합 확률 모델 보다 더 좋은 방법이 될 수 있기 때문이다.

본 연구에서 사용한 최대 엔트로피 모델을 학습시키기 위해 이용한 속성은 표 1과 같다.

(표 1) 최대 엔트로피 모델을 학습하기 위해 사용된 속성

Attribute	Description
Word(i-2)	약어 전 w_{i-2}의 단어
Word(i-1)	약어 전 w_{i-1}의 단어
Word(i+1)	약어 후 w_{i+1}의 단어
Word(i+2)	약어 후 w_{i+2}의 단어
Dept	해당 문서의 분류정보
Section	약어가 위치한 섹션

문서 내에서 약어와 두문자어를 찾은 후 해당 약어 및 두문자어의 의미를 알아내기 위해 약어/두문자어 w_i의 전 (w_{i-2}, w_{i-1}), 후(w_{i+1}, w_{i+2}) 각각 두 단어, 진료과별 정보, 그리고 섹션 정보를 이용하여 훈련 데이터를 구성하였다. 예를 들어, 최대 엔트로피 모델을 이용하여 두문자어 "RA"는 식 (1)과 (2)를 통해 "rheumatic arthritis"와 "right atrium" 두 가지 의미에 대한 확률을 각각 구하여 "RA"가 어떤 의미에 해당하는지를 결정할 수 있다.

$$P(abbr = rheumatic\ arthritis | w_{i-2} \rightarrow joint, w_{i-1} \rightarrow pain, w_{i+1} \rightarrow oa, w_{i+2} \rightarrow ana, dept \rightarrow rh, section \rightarrow PI) \quad (1)$$

$$P(abbr = right\ atrium | w_{i-2} \rightarrow joint, w_{i-1} \rightarrow pain, w_{i+1} \rightarrow oa, w_{i+2} \rightarrow ana, dept \rightarrow rh, section \rightarrow PI) \quad ----(2)$$

3. 결과

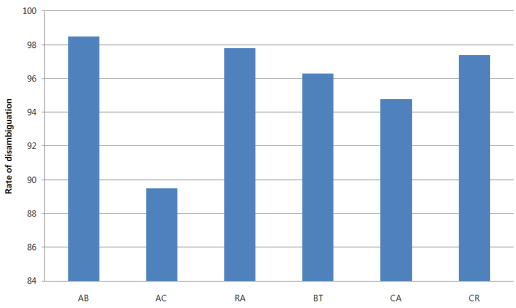
실험은 D 의료원의 퇴원요약지 데이터에 이용된 약어와 두문자어의 의미 분류에 대한 정확도를 측정하였다.

표 2는 8개의 진료과에 대한 베이지안 분류기의 과별 분류 정확도를 나타내고 있다. 진료과별 분류를 위한 베이지안 분류기를 학습시키는데 12,000건의 퇴원요약지를 이용하였다. 진료과에 대한 분류에는 2,700건의 퇴원요약지 자료를 이용하였고, 이 가운데 2,603개의 문서가 진료과에 맞게 분류되었으며, 정확률은 92.47%로 나타났다.

<표 2> 베이시안 분류기를 이용한 문서 분류 결과

	BE	CR	CV	DM	GE	GI	GU	NE	%
BE	83	0	0	0	0	3	1	0	95.4023
CR	0	35	1	0	2	3	2	0	81.39535
CV	0	0	1307	0	1	27	0	1	97.82934
DM	0	0	0	35	1	1	0	0	94.59459
GE	0	2	1	0	46	7	0	0	82.14286
GI	1	2	21	0	1	554	0	2	95.35284
GU	4	0	0	0	0	0	93	0	95.87629
NE	0	0	2	0	3	7	1	450	97.19222

그림 2는 6개의 약어에 16개의 의미를 이용한 약어와 두문자어의 모호성 해결 실험에 대한 결과를 보여주고 있으며, 정확률은 94.71667% 정도로 나타났다.



(그림 2) 약어/두문자어의 의미 분류 정확률

약어와 두문자의 모호성 해결에 있어, 과거 병력과 같은 문서정보가 제한적인 섹션에 약어가 나타나거나 독립되어 나타나는 경우 분류에 이용되는 속성에 대한 신뢰도가 떨어지게 되며, 이는 정확률을 떨어뜨리는 요인이 된다.

4. 결론

자연어 문서로부터 약어와 두문자어의 의미를 파악하는 것은 문서로부터 정보를 추출할 때 반드시 필요하며 유용한 전처리 과정이다. 다양한 의미로 사용될 수 있는 약어와 두문자어들이 나타날 때, 문서의 의미는 모호해지므로 유용한 정보의 추출이나 텍스트 마이닝을 통한 의미 있는 패턴의 발견은 어려워진다. 본 연구에서는 정보추출에 앞서 문서를 정규화하는 과정의 하나인 약어와 두문자어의 의미 모호성 해결을 위한 방법을 제안하고, 실험을 통해 94.7%의 정확률을 얻었음을 밝힌다.

Acknowledgment

본 연구는 산업자원부 지방기술혁신사업(RTI04-01-01) 지원으로 수행되었음.

참고문헌

[1] Serguei Pakhomov, “Semi-Supervised Maximum Entropy Based Approach to Acronym and Abbreviation Normalization in Medical Tests”, ACL, pp. 160-167, July 2002

[2] Adam L. Berger, Stephen Della Pietra, and Vincent Della Pietra, “A maximum entropy approach to natural language processing”, Computational Linguistics 22(1), 1996

[3] Youngja Park, Roy J. Byrd, ″Hybrid text Mining for Finding Abbreviations and their Definitions″, In Proc. EMNLP 2001.

[4] Naoaki Okazaki, Sophia Ananiadou, “Building an abbreviation dictionary using a term recognition approach”, Bioinformatics, 22(24):3089-3095, 2006

[5] E.T.Jaynes, 1957, “Information Theory and Statistical Mechanics”, in Physical ReviewVolume 106 #4(p.620-630), May 1957

[6] Charles Sutton, Andrew McCallum, “An introduction to conditional random fields for relational learning.” In Getoor, L., Taskar, B. (Eds.), Introduction to Statistical Relational Learning. MIT Press, 2006.

[7] Jeffrey C. Reynar, A. Ratnaparkhi, “A Maximum Entropy Approach to Idenfying Sentence Boundries,” In Proceeding of the Fifth Conference on Applied Natural Language Processing, 1997 <- 형식 체크할 것.

[8] Adwait Ratnaparkhi, “A Maximum Entropy Model for part-of-speech tagging,” In Conference of Empirical Methods in Natural Language Processing, pp. 133-142, 1996