대한진단검사의학회지 제28권 제1호 2008
Korean J Lab Med 2008;28:79-87
DOI 10.3343/kjlm.2008.28.1.79

의생물학 문헌에 보고된 후보 암표지자 정보추출시스템 개발

채정민[1] · 오흥범[2] · 최성은[2] · 차충환[2] · 김명희[2] · 정순영[1]

고려대학교 컴퓨터교육학과[1], 울산의대 서울아산병원 진단검사의학과[2]

Development of a System for Extracting the Information of Candidate Tumor Markers Reported in Biomedical Literatures

Jeong-Min Chae[1], Heung-Bum Oh, M.D.[2], Sung-Eun Choi[2], Choong-Hwan Cha, M.D.[2], Myung-Hee Kim, M.D.[2], and Soon-Young Jung, Ph.D.[1]

Department of Computer Science Education[1], Korea University, Seoul; Department of Laboratory Medicine[2], University of Ulsan College of Medicine and Asan Medical Center, Seoul, Korea

Background : Since the human genome project was completed in 2003, there have been numerous reports on cancer and related markers. This study was aimed to develop a system to extract automatically information regarding the relationship between cancer and tumor markers from biomedical literatures.

Methods : Named entities of tumor markers were recognized by both a dictionary-based method and machine learning technology of the support vector machine. Named entities of cancers were recognized by the MeSH dictionary.

Results : Relational and filtering keywords were selected after annotating 160 abstracts from PubMed. Relational information was extracted only when one of the relational keywords was in an appropriate position along the parse tree of a sentence with both tumor marker and disease entities. The performance of the system developed in this study was evaluated with another set of 77 abstracts. With the relational and filtering keyword used in the system, precision was 94.38% and recall was 66.14%, while without the expert knowledge precision was 49.16% and recall was 69.29%.

Conclusions : We developed a system that can extract relational information between a tumor and its markers by incorporating expert knowledge into the system. The system exploiting expert knowledge would serve as a reference when developing another information extraction system in various medical fields. *(Korean J Lab Med 2008;28:79-87)*

Key Words : *Tumor, Tumor marker, Information extraction*

서 론

인간유전체사업이 2003년에 완료된 이후 이를 토대로 생물학 및 의학 분야의 많은 연구결과물이 문헌을 통해 발표되고 있다[1]. 현재 NCBI의 PubMed (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi) 데이터베이스에는 총 1,500만 건의 생물·의학 관련 논문이 저장되어 검색서비스를 제공하고 있는데, 그 저장된 문헌의 수가 기하급수적으로 증가하고 있다. 이러한 문헌의 증가는 필요한 정보를 찾기 위해 관련 연구자에게 점점 더 많은 노력과 시간을 요구하고 있다.

최근 생물정보학 분야에서는 텍스트마이닝 기법을 이용하여 문헌속에 포함되어 있는 DNA, RNA, 단백질 등과 같은 생물학적

접 수 : 2007년 8월 16일 접수번호 : KJLM2064
수정본접수 : 2007년 11월 12일
게재승인일 : 2007년 11월 15일
교 신 저 자 : 오 흥 범
우 138-736 서울시 송파구 풍납2동 388-1
울산의대 서울아산병원 진단검사의학과
전화 : 02-3010-4505, Fax : 02-478-0884
E-mail : hboh@amc.seoul.kr

*본 연구는 아산생명과학연구소 연구비(2007-219) 지원으로 이루어졌음.

엔터티(biomedical entity)들 간의 관계정보를 자동으로 추출하는 정보추출시스템에 대한 연구가 활발히 진행되고 있다[2]. Temkin 등[3]은 context free grammar 등의 규칙기반 방법을 이용하여 단백질 간의 관계정보를 추출하는 시스템을 개발하였으며, Friedman 등[4]은 세포 내 신호전달 과정을 밝혀내기 위해 단백질과 단백질 사이의 관계정보를 추출하는 시스템을 개발하였고, Ono 등[5]은 관계정보 키워드를 정의하고, 관계 키워드(relation keyword)의 유무에 따라 단백질들 사이의 관계정보를 추출하는 시스템을 개발하였다. 이러한 기초생물학 분야에서의 정보추출시스템의 개발은 연구자들에게 다량의 문헌으로부터 필요한 정보를 손쉽게 얻을 수 있는 길을 제공하고 있다. 지금까지 구현된 정보추출시스템 연구의 대부분은 DNA-단백질과 단백질-단백질 간의 상호작용과 같이 분자생물학 혹은 기초생물학 분야에 국한되어 있었다. 따라서 본 연구에서는 의학의 한 세부분야인 암표지자 연구에 있어 연구자들이 관심을 갖는 암표지자와 암과의 관계정보를 PubMed 문헌으로부터 자동추출하는 시스템을 개발하고자 하였다. 의학분야는 DNA-단백질 혹은 단백질-단백질 간의 상호작용보다는 유전자 혹은 단백질과 질병과의 관계정보에 더 많은 관심이 있다. 그런데 질병과 이들 생물학적 엔터티 간의 연관성은 모호한 추측에서부터 직접적인 증거가 있는 경우까지 다양하다. 따라서 이를 기술하는 방식 또한 매우 다양한 편이어서 정보추출 과정이 기초생물학 분야보다는 다소 어려운 점이 있다. 본 연구에서는 다양한 관계정보 기술방식을 관계 키워드로 정의하고 이를 토대로 기초분야에서 주로 이루어졌던 텍스트마이닝 기법을 임상의학 분야에 적용해보고자 하였다.

재료 및 방법

1. 시스템 설계

관계정보추출시스템은 도메인 지식을 범용적으로 이용할 수 있도록 설계하였는데, 크게 named entity recognition (NER)부분과 관계정보추출(information extraction, IE)부분으로 나누어 구성하였다. IE부분은 다시 tokenization 컴포넌트, part-of-speech (POS) 컴포넌트, syntactic analysis 컴포넌트, 엔터티 인식 컴포넌트, semantic processing 컴포넌트 등으로 모듈화하였다. 새로운 문헌이 들어오게 되면 위에서 설명한 5단계를 거쳐 관계정보가 자동으로 추출되며, 추출된 정보는 의학전문가에 의해서 타당성을 검증받도록 하였다(Fig. 1).

후보 암표지자로는 단백질, DNA, RNA, 탄수화물, 지질 등 크게 5종류의 생물질이므로 이들 엔터티를 인식할 수 있는 NER 모듈은 support vector machine (SVM)이라는 기계학습 방법[6]과 NCI Thesaurus (http://nciterms.nci.nih.gov/NCIBrowser/Dictionary.do) 사전을 이용한 패턴매칭 방법[7]을 동시에 이용하였다. SVM은 GENIA Corpus 3.0 (http://www-tsujii.is.s.u-tokyo.ac.jp/~genia/index.html)을 이용하여 학습시켰다[8]. 암

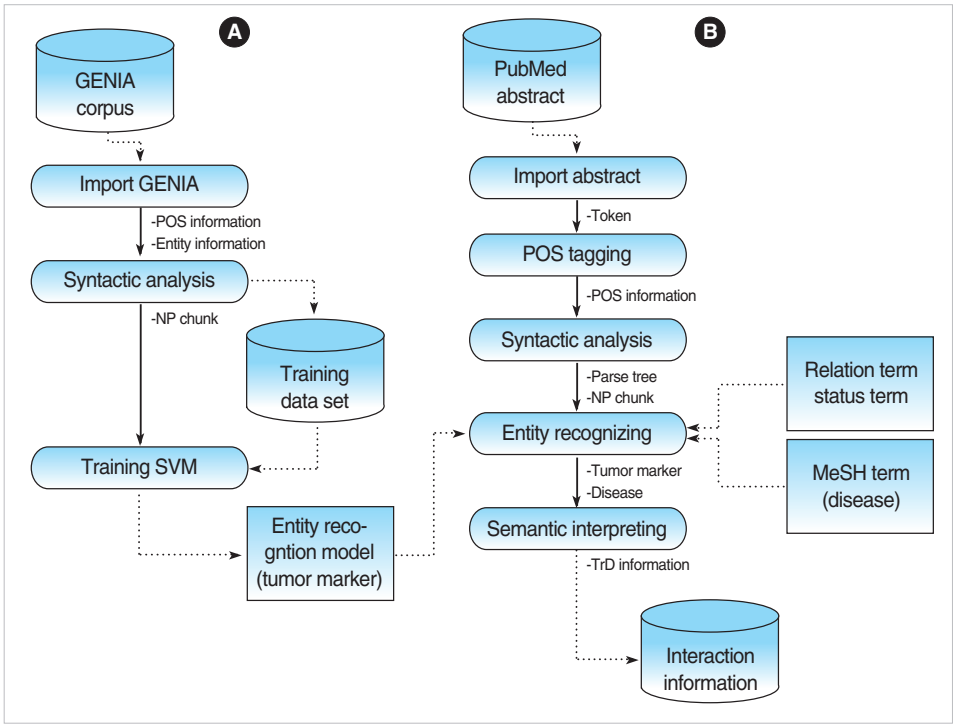

Fig. 1. System overview for the relational information extraction. (A) Named entity recognition module. (B) Information extraction module.

명칭은 Medical Subject Heading (MeSH) (http://www.nlm.nih.gov/mesh/)을 이용한 패턴매칭 방법을 이용하였다(Table 1).

정보추출모듈 즉 IE모듈은 도메인의 변경에 유연한 범용적인 관계정보추출시스템을 만들기 위해서 전문가의 지식을 통합할 수 있도록 구현하였고, 각 부분을 최대한 나누어 서브모듈화하였다. 크게 단어분리, 품사태깅, 문장구조해석, 엔터티인식, 관계정보해석 등의 단계를 거쳐 관계정보가 추출되도록 하였다. IE모듈의 서브모듈별은 다음과 같이 구현하였다. 단어분리 모듈은 PubMed의 논문초록을 XML형태로 다운로드받아 로컬시스템 데이터베이스에 저장할 때 이용되도록 하였다. 괄호에 둘러싸인 문장은 원래 문장의 엔터티 간 관계정보를 해석하는데 크게 영향을 주지 않고, 약어로 사용되는 경우가 많기 때문에 괄호 시작부터 끝까지 다른 문장으로 저장하여 원래 문장과 구분하였다. 품사태깅 모듈에서는 Fast Transformation-Based Learning (fnTBL) POS Tagger (http://nlp.cs.jhu.edu/~rflorian/fntbl/)를 사용하여 논문초록의 모든 단어의 품사를 찾아내도록 하였다. 문장구조해석 모듈에서는 Collins Parser[9]를 이용하여 문장의 파스트리와 명사구를 찾아내었다. 찾아낸 명사구는 엔터티인식 모듈에서 암표지자인지 아닌지 확인될 수 있도록 하였다. 마지막으로 관계정보해석 단계에서는 파스트리를 검색하여 관계정보가 있는지 찾아내도록 하였다.

Table 1. Recognizing methods for named biological entities

Named entities	Recognizing methods
Tumor marker	Suppot vector machine, NCI Thesaurus
Tumor name (disease)	MeSH
Relation keywords	Direct extraction by medical experts
Filtering keywords	Direct extraction by medical experts

NCI Thesaurus (http://nciterms.nci.nih.gov/NCIBrowser/Dictionary.do).
Medical Subject Heading (MeSH) (http://www.nlm.nih.gov/mesh/).

2. SVM 학습

후보 암표지자를 인식하기 위해서 학습데이터로 GENIA V3.0을 사용하였고 이를 SVM으로 학습시켰다(Fig. 2). 후보 암표지자에 해당하는 단백질, DNA, RNA, 탄수화물, 지질 등 5가지 종류의 명칭을 Inner특성으로 하고, 이를 기준으로 왼쪽 4단어와 오른쪽 4단어를 Outer특성으로 정의하였다. 즉 연속적인 단어의 배열에서 Inner특성은 후보 암표지자 이름에서 추출하였고 Outer특성은 그 주변단어에서 추출하도록 하였다. Inner 특성은 단어 개수를 5개 이하로 제한하였다. Inner 및 Outer 특성은 각각 품사정보(POS)[10, 11], 단어형성양상(word formation pattern, WFP)[12], 접두사 및 접미사(prefix/suffix), 어휘정보(word frequency)[13] 등 4가지 특성으로 구분하였다. 이러한 과정을 통하여 후보 암표지자 이름뿐만 아니라 주변 단어의 특성을 통해 암표지자를 인식할 수 있도록 하였다. 두 번째 단계에서는 GENIA V3.0에 있는 2,000개의 초록을 9:1의 비율로 트레이닝 세트(training set)과 테스트 세트(test set)로 나눈 후 기계학습과 테스트 과정을 거쳐 NER 모듈을 완성하였다.

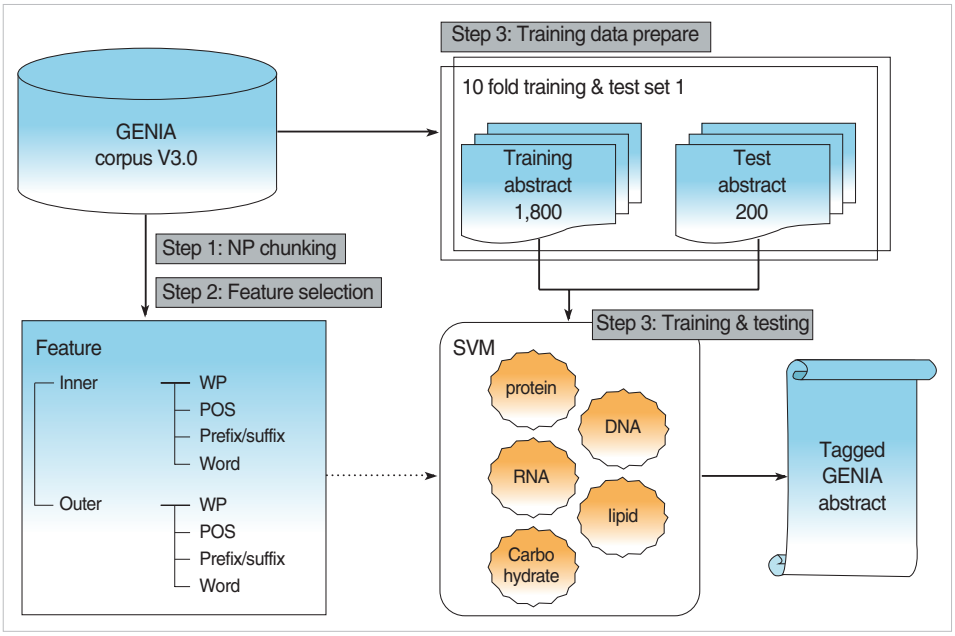

Fig. 2. Named entity recognition (NER) module by support vector machine (SVM)

3. 전문가 지식의 도출

암표지자와 암이 동일한 문장에 나타나면서 둘 간의 관계정보를 분명히 표현하는 관계키워드를 추출하였다. 그러나 두 엔터티와 관계키워드가 동일한 문장에 있지만 두 엔터티 간에 관계가 있다고 말할 수 없는 경우를 가려내기 위해 필터링 키워드(filtering keyword)를 추출하였다. 예를 들어 “A는 B의 암표지자인지 알아보고자 하였다”라는 가설적 수준의 서술문장은 선정되지 않도록 하였다. 암표지자-암 간의 관계키워드, 필터링 키워드를 만들기 위해 PubMed로부터 암표지자와 관련된 160개의 논문초록을 다운로드 받고 주석을 달 수 있는 인터페이스를 만든 후(Fig. 3) 직접 암표지자와 암 명칭, 암표지자 관계키워드, 암표지자 필터링 키워드를 표시하였다. 표시된 각 키워드는 시스템에 자동으로 등록되도록 하였다.

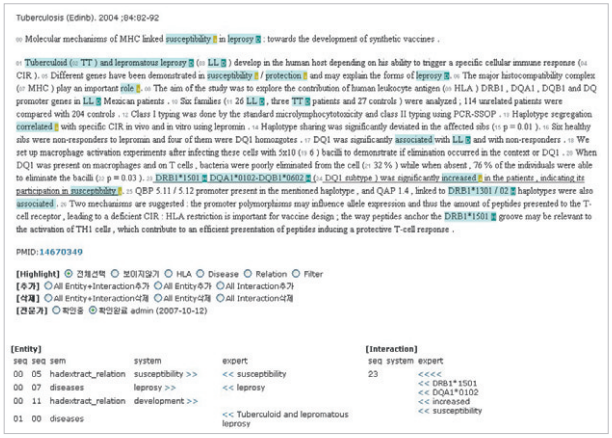

Fig. 3. Sentence annotation by medical experts for extracting relational and filtering keywords.

4. 문헌으로부터의 정보추출

PubMed의 논문초록을 XML형태로 다운로드받아 관계정보 추출 시스템의 각 단계를 거쳐 최종적으로 암표지자-암 간의 관계정보를 추출하였다. 논문초록은 PubMed에서 1990년부터 2004년까지 종양(tumor)과 표지자(marker)를 키워드로 검색된 것으로 하였는데, 1990년(4,245건)부터 2004년(11,726건)까지 총 79,845개의 논문을 시스템에 저장하였다. 초록을 다운로드받은 후 이를 우선 문장과 단어 단위로 분해하였다. 하나의 문장은 마침부호(예: “.”)를 기준으로 구분하였으며, 각 단어는 white space를 기준으로 구분하였다. 문장 중간에 괄호가 있으면 그 문장의 하위문장으로 따로 저장하였다. 논문초록 도중에 ‘1)’, ‘2)’와 같이 번호가 나타나거나 ‘i.e.’, ‘e.g.’, ‘vs.’ 등과 같이 마침 부호가 문장 중간에 나타나는 경우는 특별처리를 하여 문장단위로 구분되지 않도록 하였다.

논문초록의 각 문장은 fnTBL POS Tagger를 사용하여 품사를 결정하였다. fnTBL은 일반적인 언어를 다루는 품사태거이므로 의생물 분야에 맞게 학습을 시킬 필요가 있어, GENIA Corpus 3.0을 이용하여 학습시켰다. 문장구조를 해석하기 위한 파스트리(parse tree) 생성을 위해 Head-driven Dependency Parser인 Collins Parser를 사용하였다[9]. 생성된 파스트리는 로컬 데이터베이스에 저장하였다. 파스트리상에서 관계정보를 추출하는 과정은 postorder traversal 알고리즘을 이용하였는데, 원리를 간략히 소개하면 다음과 같다. 먼저 각 노드를 검색하다가 암표지자(T) 및 암 엔터티(D), 관계키워드(r) 키워드를 만나게 되면 자신의 정보를 상위노드에 복사한다. 상위노드에서 TrD (tumor_ marker relation disease) 관계가 모두 만족되면 관계정보를 추출한다(Fig.

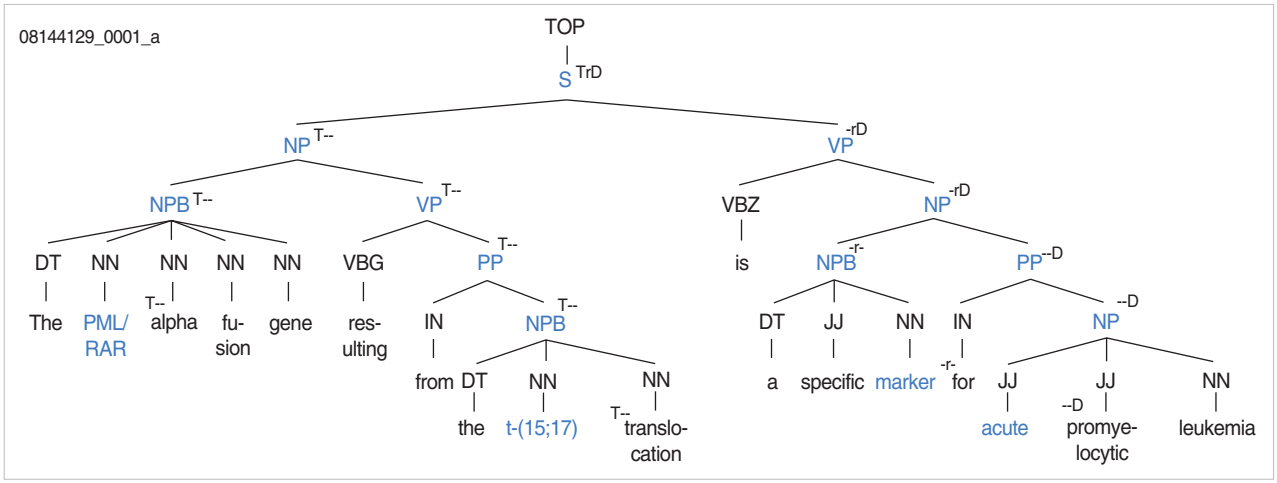

Fig. 4. Postorder traversal algorithm for extracting relational information using a parse tree.

Tumormarker	Relation	Disease	Sentence
p53	indicator	breast cancer	In breast cancer, it has been noted that the overexpression of p53 protein in the nucleus is an indicator of poor prognosis, although there is a high degree of variability, which may be due to different immunohistochemical techniques, varying assessment of results and the type of monoclonal antibody used.
HER2	overexpression	breast cancer	Response to cyclophosphamide, methotrexate, and fluorouracil in lymph node-positive breast cancer according to HER2 overexpression and other tumor biologic variables.

Fig. 5. Example of extracted information.

4). 추출된 정보는 필터링 과정을 거친 후 relation ([T], [D])의 형태로 저장하였고, Fig. 5와 같이 스크린상에 출력되도록 하였다.

5. 시스템 성능평가

시스템 개발에 사용하였던 160개의 암표지자 관련문헌 외에 "tumor marker"를 키워드로 PubMed에서 임의 선정한 77개 초록을 테스트 세트로 하여 시스템의 성능을 평가하였다. 시스템 성능은 전문가가 찾아낸 관계정보와 시스템이 찾은 관계정보를 비교하여 정확률(precision)과 재현율(recall), F-score로 표현하였다. 정확률은 시스템이 찾아낸 것 중에서 정답인 분율(P)로 정의하였고, 재현율은 정답 중에서 시스템이 정확히 찾아낸 분율(R)로 정의하였다. F-score는 다음의 수식을 이용하여 구하였다.

$$F\text{-}score = \frac{2PR}{(P+R)^2}$$

결 과

1. 관계정보추출시스템 개발

본 연구를 통하여 NER 모듈과 IE 모듈로 구성된 관계정보추출 시스템을 개발하였다[14]. 엔터티 인식모듈 즉 NER 모듈은 논문 초록으로부터 암표지자 엔터티를 인식하는 부분으로, IE 모듈의 엔터티 인식단계에 사용되도록 구성하였다. 암표지자는 단백질, DNA, RNA, 탄수화물, 지질 등 5개 종류였으므로, 암표지자를 인식하는 기계학습모듈은 각 종류에 따라 서로 다른 엔터티 인식 모듈을 개발하였다. 정보추출모듈 즉 IE 모듈은 도메인의 변경에 유연한 범용적인 관계정보추출시스템을 만들기 위해서 전문가의 지식을 통합할 수 있도록 하였고, 필요에 따라 최소한 부분만을 수정하여도 기능을 지속적으로 수행할 수 있도록 단어분리, 품사 태깅, 문장구조해석, 엔터티인식, 관계정보해석 등으로 서브모듈화하였다. Table 2는 암표지자-암 간의 관계정보를 추출하기 위해 본 연구에서 의학전문가가 찾아낸 관계키워드와 필터링 키워드로서 이를 관계정보추출 과정에서 활용할 수 있도록 하였다.

Table 2. An example of relation and filtering keywords for extracting information relating between tumors and tumor markers

Type	Keyword
Relation	marker, biomarker, associate, expression, coexpression, overexpression, correlate, distinguish, elevation, factor, identification, increase, indicator, parameter, screening tool, upregulation, usefulness, utility, value
Filtering	additional studies, aim, analysed, antibody, assessed, cell lines, clarify, compared, design, evaluate, examined, further study, hypothesis, immunochemical, investigate, measured, methods, no studies, objective, performed, studied, tested, unclear, unknown, variable, whether

2. 시스템 성능

임의로 선택된 77개의 논문 중에서 시스템은 938개의 암표지자, 341개의 암, 678개의 관계키워드, 460개의 필터링 키워드를 찾아내었다. 전문가 지식을 사용하지 않을 경우 총 179개의 문장에서 암표지자와 암 엔터티가 동시에 발견되었으며, 전문가 태깅 결과와 비교한 결과, 정확률 49.16%, 재현율 69.29%, F-score 57.51의 성능을 보였다. 반면 암표지자-암 관계키워드와 필터링 키워드를 동시에 사용하였을 경우에는 총 89개의 문장에서 관계정보가 있다고 하였으며, 정확률 94.38%, 재현율 66.14%, F-score 77.78의 성능을 보였다. 160개의 트레이닝 세트(training set)에 대해 동일한 성능실험을 시행한 결과에서는 전문가 지식을 사용하지 않은 경우 정확률 50.71%, 재현율 91.12%, F-score 65.16의 성능을 보였고, 관계키워드와 필터링 키워드 등의 전문가 지식을 사용하는 경우에는 정확률 90.49%, 재현율 84.52%, F-score 87.40의 성능을 보였다.

고 찰

비구조화된 문헌정보로부터 구조화된 정보를 추출하고 이를 토대로 새로운 지식을 발굴해내는 과정을 텍스트마이닝이라 한다[15]. 텍스트마이닝은 크게 information retrieval (IR), IE, data mining (DM)으로 구성된다. IR은 Google (http://www.google.co.kr/), PubMed 등을 통해 자료를 불러오는 기능이며, IE는 불러온 자료로부터 정보를 추출하여 구조화하는 것이고, DM은 구조화된 자료로부터 새로운 지식을 형성해 나가는 것이다. 정보추

출을 위해서는 우선 문헌 속에 있는 단백질, 유전자 및 질병의 이름을 찾아낼 수 있어야 하는데, 이를 NER이라 한다. NER은 뉴스 분야에서 먼저 시작되었으며 최근에는 바이오 메디컬 분야에서도 NER을 효율적으로 처리하기 위한 연구가 활발히 진행되고 있다[16]. 신문으로부터 테러분자를 찾아내기 위한 message understanding conference (MUC) 콘테스트를 통하여 NER의 성능은 매우 향상되었는데, 1995년에 있었던 MUC-6에서 이미 96%의 recall과 97%의 precision이라는 놀라운 성능을 보였다[17]. 1998년부터 의생물학 영역에 NER을 적용하려는 시도가 있었는데 초기의 성능은 그리 좋은 것이 아니었다. 이는 MUC에서 다루었던 것이 신문방송 자료에서의 사람, 지역, 소속기관 등인 반면 의생물학 분야에서는 유전자, 단백질 등 기존의 엔터티와는 매우 다른 것이기 때문이었다.

단백질이나 유전자의 이름은 일반적으로 해당 유전자나 단백질의 특성을 서술하는 형식으로 만들어진다. 주로 사용되는 특성으로는 기능(예: growth hormone), 위치(예: nuclear protein), 기원한 종의 이름(예: HIV-1 envelope glycoprotein), 물리성상(예: salivary acidic protein), 기존 단백질과의 유사성(예: Rho-like protein) 등이 있다. 이런 서술형 방식은 엔터티의 단어 수를 많게 하여 좌측 경계가 어디인지를 결정하는데 어려움을 준다. 예를 들어 “inhibitor of p53”라는 엔터티가 있을 경우 “p53” 외에도 “inhibitor of p53” 자체도 별개의 엔터티로 인식할 수 있어야 하는 어려움이 있다는 것이다. 더구나 GENIA corpus V3.0에는 19개의 단어로 이루어진 단백질 이름이 존재하고 있다[6]. 의생물학 분야의 NER에서 또 다른 어려움은 많은 경우 엄격히 적용되는 표준화된 명명법이 없어 용어변이(term variation)가 심하다는 점과 단백질과 유전자 심지어는 질병명이 혼용되어 사용될 수 있다는 점이다. 예를 들어 N-acetylcystein을 N-acetyl-cystein 혹은 NAcetylCystein 등과 같이 약간씩 다르게 기술한다거나, NAC처럼 고유한 축약형을 쓰는 경우들이 있다[18]. 또한 major histocompatibility complex (MHC)가 유전자를 뜻하기도 하지만 해당 유전자의 산물인 단백질을 뜻하기도 한다거나, “neurofibromatosis 2”라는 질병명의 원인 유전자 이름 또한 “neurofibromatosis 2”라 하므로 엔터티 인식은 쉬운 일이 아니다[19]. 일부의 연구에서는 동일한 문장 내의 엔터티가 유전자인지 혹은 단백질인지를 조사한 결과 69-77%의 일치율 밖에 보이지 않았다[20, 21].

의생물 분야의 NER을 위해 사전기반, 규칙기반, 기계학습 기반의 방법 등이 사용되어 왔다. 사전을 이용하여 단백질이나 유전자를 찾아보려면 HUGO Gene Nomenclature Committee (http://www.gene.ucl.ac.uk/nomenclature/), GenBank (http://www. ncbi.nlm.nih.gov/Genbank), UniProt (http://www.expasy.org/sprot), IPI (http://www.ensemble.org/IPI/), NCI Thesaurus 등을 사용할 수 있다. 그러나 사전기반 방법에서는 새로 명명되는 NE를 찾을 수 없다는 것, 모호한 명명 때문에 위양성이 있을 수 있다는 것, 동의어, 단어변이 때문에 위음성을 일으킬 수 있다는 단점이 있다[17]. 그러나 질병 이름의 경우는 새로 만들어지는 경우가 매우 드물기 때문에 MeSH의 “disease category”를 이용하는 것이 보편적이다. MeSH에는 질병관련 서술자가 일만개 가량 있으며 해당 서술자에 대해 여러 개의 검색용 키워드가 있기 때문에 총 검색용 키워드는 4만개 가량 된다. 또한 MeSH에는 온톨로지 정보를 가지고 있기 때문에 상위개념의 용어는 하위개념의 용어를 포함하여 검색하는 장점을 가지고 있다.

규칙을 이용하는 경우는 조직적합성 항원에서처럼 명명 규칙이 정립되어 오랫동안 잘 지켜져 왔던 경우에 사용될 수 있다[22]. 일반적인 단백질의 경우에는 core term (Ras, Sos 등)과 feature term (receptor, protein 등)을 골라낸 후 그 사이에 있는 단어 모두를 엔터티로 인식하는 연구가 있었는데, precision 94.7%, recall 98.8%의 성능을 보였다[17]. Core term은 문장 중에 있으면서 대문자로 시작하는 경우, 단어에 숫자가 섞여 있는 경우, 단어 사이에 쉼표, hypen, slash가 있는 경우 등이다. 예를 들어 “Ras guanine nucleotide exchange factor Sos”의 경우 Ras, Sos 등은 문장 가운데 있으면서 대문자로 시작하기 때문에 core term으로 인지되므로 의생물 분야의 엔터티 여부가 쉽게 판정될 수 있다.

기계학습을 이용하는 경우는 Hidden Markov Model (HMM), SVM 등이 사용되어 왔다[16, 20]. 기계학습의 경우는 유전자 혹은 단백질을 구성하는 각 단어의 특성과 엔터티 전후의 수 개 단어까지의 정보를 벡터로 구성하여 이를 학습한 후 새로운 문장에서 유전자 혹은 단백질 이름을 찾아내는 방식을 취한다. 단어의 특성으로 사용되는 것은 품사특성(POS), 단어형성양상(WFP), 접두사 혹은 접미사 정보(prefix/suffix), 어휘정보 등이다. 의생물학 분야의 NER 연구에 많이 사용되는 GENIA corpus의 경우 엔터티의 90% 이상이 명사이거나 “명사+형용사+복수명사”의 결합으로 이루어져 있기 때문에 POS 정보는 단백질 혹은 유전자 이름과 일반단어를 경계짓는 중요한 정보가 된다. WFP 특성은 단백질이나 유전자 이름의 경우 숫자, 대문자, 로마자, 그리스어, 특수기호 등을 가지는 특성을 이용하는 것이다. 예를 들어 “for activation of I kappa B-alpha proteolysis”이란 문장이 있을 때, 대문자인 “I”와 그리스문자인 “alpha”를 찾아내어 단백질 이름인 “I kappa B-alpha”로 찾아낼 수 있다. 따라서 이 WFP 특성은 여러 개의 단어로 이루어진 단백질 이름을 찾는데도 매우 유용하다. 단백질이나 유전자의 경우에는 명칭 내에 단백질이나 유전자

의 성질을 함축하는 prefix 혹은 suffix를 사용하기 때문에 이를 이용할 수 있다. Protein, gene, receptor, factor 등과 같은 어휘들은 생물학적 엔터티를 표현하는데 자주 사용되기 때문에 이들의 사용빈도를 조사하면 엔터티를 인식하는데 좋은 정보를 제공한다.

엔터티를 찾아낸 후 이들 사이의 관계정보를 추출하는 방법에는 패턴매칭, context free grammar (CFG) 및 파스트리를 이용할 수 있다. 패턴매칭이란 파싱을 하지 않고 엔터티들과 관계 키워드가 일정한 template 형태로 존재하는지를 판단하여 정보를 추출하는 단순한 방법이다. 여기서 관계 키워드는 전문가에 의해 작성된다. 패턴매칭을 이용한 사례로는 Horn 등[23]이 유전자의 점 돌연변이 정보를 자동으로 추출하여 저장하는 연구에서 49.3-64.5%의 재현율과 85.8-87.9%의 정확률을 보고하였다. 점 돌연변이는 표현형태가 일정하고 단순하기 때문에 정규표현식으로도 원하는 엔터티를 효율적으로 찾아낼 수 있었다. Temkin 등[3]은 CFG를 이용하여 관계정보를 추출하는 방법을 제시하였으며 재현율 63.9%, 정확률 70.2%의 성능을 보고하였다. 그러나 규칙기반 방법으로 성능을 향상시키기 위해서는 복잡한 문법을 계속 만들어야 하는 문제가 있고, 엔터티와 엔터티 사이의 관계를 설명하는 키워드의 정확한 수식위치를 찾을 수 없어 해석의 중의성을 해결할 수 없는 단점이 있다. 이를 해결하기 위해서는 파스트리를 이용하여 관계키워드의 위치 정보를 점검한 후 관계정보를 추출하는 작업이 필요하다. Friedman 등[4]은 관계정보 키워드를 직접 설정한 후 파스트리를 이용하여 단백질과 단백질 사이의 관계정보를 추출한 연구에서 96%의 정확률을 보고하였다. 또한 Novichkova 등[24]은 이러한 관계정보를 추출하는 시스템이 너무 작은 문장단위에서만 의미를 해석하려고 하기 때문에 실제로 중요한 관계정보를 놓치거나 잘못 해석하는 경우가 있다고 경고하면서 파스트리를 이용해 다양한 도메인에 적용되더라도 유연한 특성을 가지는 시스템을 구현하였다.

인구 4명 중 1명 이상은 생애 중에 어떤 형태로든 암이 발생한다. 의료기술의 발달은 심혈관계 질환이나 감염질환으로 인한 사망률을 현격히 감소시켰으나 암에 의한 사망률은 크게 감소시키지 못하고 있다. 선진국 및 우리나라의 주요 사망원인의 1위는 암에 의한 것이다. 1998-2000년도 한국인 암발생률은 남자 10만 명당 243.9명, 여자 10만 명당 175.2명이며, 0세에서 74세까지의 누적 암발생률은 남자 35%, 여자 17%라는 보고가 있었다[25]. 높은 암사망률의 원인은 치료방법의 부재가 원인이기 보다는 치료가 효과적으로 시행될 수 있는 시기에 암을 진단하지 못하기 때문이다. 암표지자란 암의 존재 혹은 진행과 관련된 생체물질로써 암의 조기 진단에 매우 긴요한 것이다. 전립선암의 표지자로써 가장 많이 사용되고 있는 전립선특이항원(prostate-specific antigen, PSA)의 도입으로 진단 당시 말기 전립선암의 비율이 급격히 감소한 것은 암표지자의 중요성을 잘 보여주는 것이다[26]. 또한 전립선암의 생존율이 1970년대 67%에 불과하던 것이 1990년대에는 99%로 더 높은 생존율을 보이게 된 것은 90년대 도입된 암표지자 선별검사가 큰 몫을 했기 때문으로 평가되고 있다[27].

본 논문에서는 기하급수적으로 증가하는 바이오 관련 문헌을 대상으로 암과 암표지자의 관련 정보를 자동으로 추출하는 시스템을 구축하였다. 본 연구에서는 전문가 지식을 시스템에서 활용할 수 있도록 고안하였는데, 관계정보 키워드, 필터링 키워드를 선정하고 이를 시스템에서 활용하는 방법을 통해 전문가 지식을 활용하였다. 본 연구를 통해 개발한 시스템의 재현율은 66.14%, 정확률은 94.38%으로서 암표지자 연구분야에 실질적으로 사용할 수 있을 정도의 성능을 가지는 정보추출 시스템을 개발할 수 있었다. 또한 시스템에서 전문가 지식을 사용한 경우는 그렇지 않은 경우에 비해 F-score가 20% 가량 증가하는 것을 관찰할 수 있었다. 따라서 전문가 지식을 활용하는 본 연구의 방식은 의학의 다른 연구분야에도 참고할 수 있을 것으로 사료된다.

요 약

배경 : 2003년 완료된 인간유전체사업 이후 암발현과 관련된 단백질 및 유전자 정보가 대량으로 만들어지고 있다. 본 연구에서는 암표지자와 암과의 관계정보를 의생물학 문헌으로부터 자동추출하는 시스템을 개발하고자 하였다.

방법 : 암표지자 인식은 사전검색 방법과 기계학습 방법의 하나인 support vector machine (SVM)을 이용하였다. 단백질, DNA, RNA, 탄수화물 및 지질 등 5종류의 후보 암표지자를 인식하기 위하여 5개의 SVM을 구성하였다. 암의 명칭은 MeSH 사전을 이용하였다.

결과 : 160개의 문헌에서 전문가 지식을 사용하여 관계키워드 및 필터링 키워드를 선정하였고, 시스템에서 이를 이용하여 관계정보를 추출하도록 하였다. 관계정보는 파스트리에서 암표지자, 암 및 관계 키워드가 적절한 위치에 있는 경우에 추출하였다. 별도의 77개 초록으로부터 시스템의 성능을 평가한 결과 관계키워드 및 필터링 키워드를 사용한 경우 정확률 94.38%, 재현율 66.14%인 반면, 전문가 지식을 사용하지 않은 경우 정확률 49.16%, 재현율 69.29%의 성능을 보였다.

결론 : 본 연구를 통하여 암표지자 관련 연구문헌으로부터 전문가 지식을 접목하여 실제 연구에 활용 가능한 성능을 가지는 관계정보 자동추출시스템을 개발할 수 있었다. 본 시스템에 구현된 전문가지식 활용 방법은 향후 다른 의학분야의 정보추출시스템을

개발하고자 할 때 참고가 될 수 있을 것이다.

참고문헌

1. Collins FS, Green ED, Guttmacher AE, Guyer MS, US National Human Genome Research Institute. A vision for the future of genomics research. Nature 2003;422:835-47.
2. Jensen LJ, Saric J, Bork P. Literature mining for the biologist: from information retrieval to biological discovery. Nat Rev Genet 2006; 7:119-29.
3. Temkin JM and Gilder MR. Extraction of protein interaction information from unstructured text using a context-free grammar. Bioinformatics 2003;19:2046-53.
4. Friedman C, Kra P, Yu H, Krauthammer M, Rzhetsky A. GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. Bioinformatics 2001;17(S):S74-82.
5. Ono T, Hishigaki H, Tanigami A, Takagi T. Automated extraction of information on protein-protein interactions from the biological literature. Bioinformatics 2000;17:155-61.
6. Cristianini N and Shawe-Taylor J. eds. An introduction to support vector machines and other kernel based learning methods. 1st ed. Cambridge: Cambridge University Press, 2000.
7. McNaught J and Black WJ. Information extraction. In: Ananiadou S and McNaught J, eds. Text Mining for biology and biomedicine. 1st ed. Norwood: Artech House, 2006:143-77.
8. Kim JD, Ohta T, Tateisi Y, Tsujii J. GENIA corpus- semantically annotated corpus for bio-textmining. Bioinformatics 2003;19(S): i180-2.
9. Collins M, Head-Driven Statistical Models for Natural Language Parsing [Dissertation]. Philadelphia (PA): Pennsylvania Univ.; 1995.
10. Tanabe L and Wilbur WJ. Tagging gene and protein names in biomedical text. Bioinformatics 2002;18:1124-32.
11. Kazama J, Makino T, Ohta Y, TsujiiJ J. Tunning support vector machines for biomedical named entity recognition. In: Association for Computational Linguistics, ed. ACL 2002 Workshop. Proceedings of the ACL 2002 Workshop on Natural Language Processing in the Biomedical Domain; 2002 July 11; Philadelphia, PA, USA; 2002. p. 1-8.
12. Zhou G, Zhang J, Su J, Shen D, Tan C. Recognizing names in biomedical texts: a machine learning approach. Bioinformatics 2004;20: 1178-90.
13. Proux D, Rechenmann F, Julliard L, Pillet V V, Jacq B. Detecting Gene Symbols and Names in Biological Texts: A First Step toward Pertinent Information Extraction. Genome Inform Ser Workshop Genome Inform 1998;9:72-80.
14. Chae JM, Jung SY, Oh HB (Eds.). Tumor marker information extraction system. http://medtextmining.net/ (Updated on Aug 2006).
15. Ananiadou S and McNaught J. Introduction. In: Ananiadou S, McNaught J, ed. Text mining for biology and biomedicine. 1st ed. Norwood: Artech House, 2006:1-12.
16. Lee KJ, Hwang YS, Kim S, Rim HC. Biomedical named entity recognition using two-phase model based on SVMs. J Biomed Inform 2004;37:436-47.
17. Park JC and Kim JJ. Named entity recognition. In: Ananiadou S, McNaught J, ed. Text mining for biology and biomedicine. 1st ed. Norwood: Artech House, 2006;121-42.
18. Ananiadou S and Nenadic G. Automatic terminology management in biomedicine. In: Ananiadou S, McNaught J, ed. Text mining for biology and biomedicine. 1st ed. Norwood: Artech House, Inc., 2006; 67-98.
19. Bodenreider O. Lexical, terminological, and ontological resources for bilogical text mining. In: Ananiadou S, McNaught J, ed. Text mining for biology and biomedicine. 1st ed. Norwood: Artech House, 2006;43-67.
20. Hatzivassiloglou V, Duboue PA, Rzhetsky A. Disambiguating proteins, genes, and RNA in text: a machine learning approach. Bioinformatics 2001;17(S):S97-106.
21. Krauthammer M, Rzhetsky A, Morozov P, Friedman C. Using BLAST for identifying gene and protein names in journal articles. Gene 2000; 259:245-52.
22. Marsh SG, Albert ED, Bodmer WF, Bontrop RE, Dupont B, Erlich HA, et al. Nomenclature for factors of the HLA system, 2004. Tissue Antigens 2005;65:301-69.
23. Horn F, Lau AL, Cohen FE. Automated extraction of mutation data from the literature: application of MuteXt to G protein-coupled receptors and nuclear hormone receptors. Bioinformatics 2004;20:557-68.
24. Novichkova S, Egorov S, Daraselia N. MedScan, a natural language processing engine for MEDLINE abstracts. Bioinformatics 2003;19: 1699-706.
25. Shin HR, Won YJ, Jung KW, Park JG, Ahn YO. Cancer Registration and Statistics in Korea. J Korean Assoc Cancer Prev 2004;9:49-55. (신해림, 원영주, 정규원, 박재갑, 안윤옥. 우리나라 암등록사업과 암통계. 대한암예방학회지 2004;9:49-55.)
26 Hernandez J and Thompson IM. Prostate-specific antigen: a review

of the validation of the most commonly used cancer biomarker. Cancer 2004;101:894-904.
27. Herbst RS, Bajorin DF, Bleiberg H, Blum D, Hao D, Johnson BE, et al. Clinical Cancer Advances 2005: major research advances in cancer treatment, prevention, and screening--a report from the American Society of Clinical Oncology. J Clin Oncol 2006;24:190-205.