



생물정보학과 자연언어처리

한국과학기술원 박종철*

1. 서론

생물정보학(Bioinformatics)은 생물학에서 다루는 정보의 양이 급증함에 따라 전산학, 수학, 통계학 등의 분야에서 사용되고 있는 정보처리기술을 응용하여 이를 효율적으로 생산, 관리, 활용하려는 연구분야를 총칭한다.¹⁾ 그리고 자연언어처리(Natural Language Processing-NLP)는 한국어나 영어와 같은 자연언어로 표현된 문장이나 문서들을 컴퓨터를 이용하여 처리하기 위한 연구분야를 총칭하는데 이에 인간과 컴퓨터의 상호작용(Human Computer Interaction-HCI)을 돕기 위한 연구의 측면도 있고 방대한 자연언어 정보를 효율적으로 관리, 활용하기 위한 연구의 측면도 있다. 본고에서는 생물의료 정보 추출(Biomedical Information Extraction)이라는 분야의 연구에 대한 소개를 통해서 이 두 가지 상이한 연구분야가 어떠한 관계를 가지게 되는지에 대한 논의를 제공한다. 최근 생물정보학에 대하여 높아진 일반의 관심을 반영하여 정보과학회지 2000년 8월호 [1]에서는 생물정보학을 주제로 한 특집을 제공하였는데 본고의 내용은 여기에 자연언어처리 응용 분야를 그로부터 일 년후의 시점에서 보완하는 것으로 볼 수 있을 것으로 기대된다.

2. 생물정보학과 자연언어처리의 관계

서로 상이한 역사와 발전과정을 거치고 있는 생물정보학과 자연언어처리가 서로의 방법론과 당면

한 문제에 관심을 가지게 된 것은 비교적 최근의 일인 것으로 보이며, 흥미있는 사실은 이러한 관심이 아직까지는 비대칭적이라는 것이다. 각 분야의 대표적인 연례학술대회에서 발표되고 있는 논문을 보면 생물정보학 분야에서는 자연언어문서에서의 효율적인 정보추출이 필요하다는 인식으로 단순한 키워드 추출로부터 시작하여 자연언어처리기술이 도입된 지 2001년 현재 5년 정도는 경과하였지만 자연언어처리분야에서는 생물정보학 관련한 응용 연구의 필요성이 인식되기 시작한 것은 극히 최근인 것으로 보이며 특히 생물정보학 응용을 위한 별도의 세션이 할당되는 규모의 관심은 아직 보이지 않고 있다.

2.1 생물정보학에서의 자연언어처리

생물정보학 분야에서 독립적으로 유지되고 있는 대표적인 연례학술대회로는 ISMB라고 알려져 있는 Intelligent Systems for Molecular Biology 국제 학술대회, PSB라고 알려져 있는 Pacific Symposium on Biocomputing, 그리고 RECOMB(Research in Computational Biology)이라는 이름으로 보다 잘 알려져 있는 Computational Molecular Biology 국제 학술대회가 있다. 이들 세 학술대회는 모두 국제 계산생물학 연구회(International Society for Computational Biology-ISCB²⁾)에서 후원(ISMB,PSB)하거나 지원(RECOMB)한다.

이들 세 학술대회 중에서 가장 오래된 ISMB는 9차 대회가 2001년 7월에 덴마크의 코펜하겐에서 개최되었다.³⁾ 참고로 10차 대회는 2002년 8월에 캐

* 중신회원

1) 독일 뮌헨 대학의 Rolf Backofen은 Bioinformatics를 solving problems arising from biology using methodology from computer science라고 정의하고 biocomputing, neural computing, genetic algorithms, evolutionary computing 등의 인접 분야와 구별하고 있다.

2) 학회 홈페이지: <http://www.iscb.org>

3) 학술대회 홈페이지: <http://ismb01.cbs.dtu.dk>

나도의 에드먼튼에서 개최될 예정이다.⁴⁾ ISMB 2000과 2001에서는 각각 2편의 자연언어처리 관련 논문이 발표되었으며 이들 논문을 위한 별도의 세션이 할당되지는 않았다[11,14,22,25]. 그런데 ISMB 2001에서는 생물정보학분야에 자연언어처리 기법 도입을 소개하는 튜토리얼이 두 개 선정되어 이 분야에 대한 고조된 관심을 짐작할 수 있게 한다[2,7].

RECOMB은 ACM-SIGACT(Special Interest Group on Algorithms and Computation Theory)에서 후원하는 학술대회로서 5차 대회가 2001년 4월 캐나다의 몬트리올에서 개최되었다.⁵⁾ 그리고 6차 대회는 2002년 4월에 미국의 수도 워싱턴 시에서 개최될 예정인데 2000년 4월과 6월에 인간 Genome의 서열 완전 해독과 조립 완성을 각각 발표한 바 있는 일반 회사 Celera Genomics에서 지원과 함께 조직을 맡는다.⁶⁾ 이 학술대회에서는 아직 자연언어처리관련 논문은 선정되지 않고 있다.

PSB는 2002년 1월에 7차 대회가 열리는데 매년 미국의 하와이에서 개최된다.⁷⁾ 이 학술대회에서는 다른 학술대회와는 달리 세션에 대한 공모를 거친 후, 논문제출요구서 공고, 초청 강사의 선정, 특별 프로그램의 기획 등 각 세션에 대한 운영권을 선정된 세션의 제안자들에게 위임한다. PSB에서는 2000년부터 자연언어처리에 직접 관련된 세션이 2002년 현재까지 매년 제안, 선정되고 있는데 PSB 2000에는 Natural Language Processing for Biology(조직위원: T. Tsunoda & L. Wong) 세션에 다섯 개의 논문[6,15,24,26,28], PSB 2001에는 Natural Language Processing and Information Extraction in Biology(조직위원: J.-I. Tsujii & L. Wong) 세션에 네 개의 논문[8,16,21,29]이 실렸으며 PSB 2002에는 Literature Data Mining for Biology(조직위원: L. Hirschman, J. Park, J.-I. Tsujii, L. Wong & C. Wu) 세션에 여섯 개의 논문[10,13,17,20,23,27,30]이 실릴 예정이다. PSB 2001에서는 하나의 자연언어처리 관련 튜토리얼이 선정되었다[29].

국내에서도 한국생물정보학회(Korean Society

for Bioinformatics⁸⁾)가 1998년 11월에 창립된 후 생물학전문연구정보센터(Biological Research Information Center-BRIC)와 공동으로 2000년 11월에 1차, 그리고 2001년 8월에 2차의 생물정보학 국제 심포지엄(International Symposium on Bioinformatics)을 개최하는 등 활발한 활동을 하고 있다.

2.2 자연언어처리에서의 생물정보학

자연언어처리분야에서 진행되는 연구는 통상적으로 형태소, 통사, 의미, 화용(Pragmatics) 및 문맥(Discourse) 분석 및 생성 단계와 같이 일반적인 언어 현상에 대한 연구를 중심으로 분류되는데, 이 과정에는 한국어나 영어와 같이 특정 언어에 국한되거나 보다 다양한 언어에 대한 종합적인 연구가 병행하여 진행될 수도 있다. 여기에서 예를 들어 의존문법(Dependency Grammar)이나 HPSG(Head-driven Phrase Structure Grammar), 결합범주문법(Combinatory Categorical Grammar) 등과 같은 각종 문법 체계는 주로 통사 단계의 세부 연구를 위한 기초구조로 사용되며, 어휘 단위 의미 체계의 개발이나 도메인 의존적인 온톨로지(Ontology)의 개발 등은 주로 의미 단계의 세부 연구를 위한 것이다. 이러한 세부 연구는 자연언어처리 분야에서 전통적으로 사용되고 있는 심플 단위 접근과 이에 따른 애매성 해소(Ambiguity Resolution) 방안 연구, 그리고 비교적 최근에 제시된 확률적 접근과 이를 위한 말뭉치(Corpus) 구축 방안 연구와 같은 방법론적인 차이를 보일 수 있는데, 이들간의 표면적인 방법론의 차이는 상호 보완적으로 절충되어 보다 높은 수준의 자연언어처리 결과를 기대할 수 있다고 보고 있다. 그리고 이와 같은 자연언어처리의 세부 연구에서의 각종 결과는 정보검색, 정보추출, 문서요약 및 분류, 기계번역, 질의응답 등과 같은 분야에서 사용될 수 있는 범용 응용 시스템의 구현에 직간접적으로 기여를 하게 된다. 마지막으로 이와 같은 범용 응용 시스템은 전자 상거래(Electronic Commerce), 의료 및 생물정보학(Medical and Bioinformatics), 군사(Military), 문화기술(Culture Technology), 오락(Entertainment) 등의 사회 전반에 걸친 다양한

4) 학술대회 홈페이지: <http://www.cs.ualberta.ca/~ismb> 02

5) 학술대회 홈페이지: <http://recomb2001.gmd.de>

6) 학술대회 홈페이지: <http://www.celera.com/recomb2002>

7) 학술대회 홈페이지: <http://psb.stanford.edu>

8) 학회 홈페이지: <http://www.ksbi.or.kr>

도메인에 적용되어 필요한 기능을 수행한다.

자연언어처리 연구분야의 연례학술대회로 가장 대표적인 것이 보통 ACL이라고 알려져 있는 국제 계산언어학회 연례 미팅(Annual Meeting of the Association for Computational Linguistics-ACL⁹⁾이며 이밖에 국제 계산언어학 학술대회(International Conference on Computational Linguistics-COLING), ACL 유럽지부 학술대회(Conference of the European Chapter of the ACL-EACL), 그리고 최근에 발족되어 2001년에 2차 회의가 개최된 ACL 북미지부 학술대회(The Meeting of the North American Chapter of the ACL-NAACL)가 있는데 COLING과 EACL은 격년으로 개최되고 개최 장소에 따라 ACL과 함께 열리는 기회도 가지고 있다.

2001년 7월 프랑스의 툴루스에서 개최된 ACL 39차 연례회의(ACL 2001)에서는 두 개의 초청 강연 중 하나를 생물정보학분야로 선정하여 영국의 연구지원 자선단체인 Wellcome Trust와 왕립의료연구위원회(Medical Research Council)에서 창립하였으며 Human Genome Project의 진행으로 잘 알려진 Sanger 연구센터 소속의 Richard Durbin 박사의 발표를 제공하였는데 Durbin 박사는 통계 문법을 이용한 인간 Genome 서열의 해석(Interpreting Human Genome Sequence using Stochastic Grammar)이라는 주제로 강의하였다. 역시 ACL 2001에서는 튜토리얼 중의 하나로 독일 뮌헨 대학의 Rolf Backofen 박사의 생물정보학 개론(Introduction to Bioinformatics) 강의를 선정하는 등 자연언어처리 관련 연구자들의 생물정보학에 대한 높아진 관심을 시의적절하게 반영하였다. 그러나 ACL에서는 생물정보학과 직접적으로 관련된 논문이 발표되지 않아 생물정보학 학술대회와 비교하여 아직 이른 분위기를 보여준다.

3. 생물의료 정보 추출

현재 인터넷을 통하여 공개되어 있는 수 많은 생물학 관련 데이터베이스들에는 GenBank와 같은 유전자 서열 정보를 가진 것과 함께 MEDLINE과 같은 문헌 데이터베이스들도 포함되어 있다.

MEDLINE은 미국 국립 의료도서관에서 제공하고 있는 공개 데이터베이스로 현재 1억 1천만건 이상의 저널 논문들에 대한 정보를 가지고 있는데 Entrez라고 하는 검색엔진과 PubMed를 통하여 접근할 수 있다.¹⁰⁾ MEDLINE에 대한 검색요구를 키워드의 형태로 PubMed에 요청하면 색인된 저널 논문들에 대한 정보를 받아 볼 수 있는데 예를 들어 면역체계에 관계하는 단백질의 일종인 cytokine에 대한 검색 요구에 2001년 8월말 현재 다음과 같은 결과를 포함하여 305,629건의 논문 정보를 받게 된다.¹¹⁾

Espejo C, Carrasco J, Hidalgo J, Penkowa M, Garcia A, Saez-Torres I, Martinez-Caceres EM. Related Articles

Differential expression of metallothioneins in the CNS of mice with experimental autoimmune encephalomyelitis.

Neuroscience. 2001;105(4):1055-65.

PMID: 11530242 [PubMed - in process]

이에 다시 위의 저자명에 제공되는 링크를 따라 가면 다음과 같은 상세 정보를 볼 수 있다.

Neuroscience 2001;105(4):1055-1065

Related Articles, Books, LinkOut

Differential expression of metallothioneins in the CNS of mice with experimental autoimmune encephalomyelitis.

Espejo C, Carrasco J, Hidalgo J, Penkowa M, Garcia A, Saez-Torres I, Martinez-Caceres EM.

Multiple sclerosis is an inflammatory, demyelinating disease of the CNS. Metallothioneins-I+II are antioxidant proteins induced in the CNS by immobilisation stress, trauma or degenerative diseases which have been postulated to play a neuroprotective role, while the CNS isoform metallothionein-III has been related to Alzheimer's disease. [중략] From these results we suggest that

9) 학회 홈페이지: <http://www.cs.columbia.edu/~acl/home.html>

10) 관련 홈페이지: <http://www.ncbi.nlm.nih.gov>

11) <http://www3.ncbi.nlm.nih.gov/htbin-post/Entrez/query?db=0&form=1&term=cytokine>

metallothioneins-I+II but not metallothionein-III may play an important role during experimental autoimmune encephalomyelitis, and indicate that the pro-inflammatory cytokine interferon-gamma is unlikely an important factor in this response.

PMID: 11530242 [PubMed - in process]

PubMed는 표시창의 선택 사양을 조정하여 이와 같이 단계별로 정보를 검색하는 방안과 함께 관련 저널 논문들에 대한 정보를 한꺼번에 파일에 저장해서 받는 방안도 제공하고 있다.

이와 같이 많은 검색 결과를 수동으로 확인해서 유용한 정보를 찾아 낸다는 것은 거의 불가능에 가까운 일이라고 할 수 있다. 그리고 물론 검색 요구 키워드를 단계별로 구체화하는 방법을 택하는 경우 어느 정도 정보의 양을 제한할 수도 있으나 이러한 키워드에 대한 충분한 사전 지식이 없는 경우에는 역시 이러한 규모의 정보에 적절히 대처하기 어렵다. 따라서 이렇게 대상이 되는 문서가 영어와 같은 자연언어로 기술되어 있다는 점을 감안하면 사용자가 원하는 정보를 자연언어처리기법을 이용하여 자동으로 추출하는 시스템이 절실하게 필요하다는 것을 알 수 있으며 이는 생물정보학 관련 학술대회에서 지금까지 발표된 거의 모든 자연언어처리 논문이 이와 같은 자동 정보 추출 시스템의 구현에 대한 것이라는 점으로 확인할 수 있다.

자연언어처리분야에서 정보추출의 필요성은 많은 분야에서 인식되어 왔으며 특히 MUC라고 알려진 Message Understanding Conference에서는 1987년부터 정보추출 시스템 개발자들이 시스템의 성능을 여러가지 종류의 자연언어문서로부터 제시된 정보를 추출해내는 과제를 수행함으로써 비교할 수 있는 장을 마련해 왔다.¹²⁾ 이러한 시스템의 성능은 정확도(Precision)와 적용도(Recall)라는 측정 기준으로 평가되는데 정확도는 추출해낸 정보의 단위 수에 대한 정확하게 추출해낸 정보의 단위 수의 백분율로 표현되고 적용도는 문서에서 제공하고 있는 정보의 단위 수에 대한 정확하게 추출해낸 정보의 단위 수의 백분율로 표현된다. 예를 들어, 대상이 되는 문서의 집합에 40개의 관련된 사실이 있다

고 할 때, 시스템이 4개의 관련되지 않은 사실과 함께 관련된 사실 중 16개의 사실을 추출해 내었다고 할 경우 이 시스템의 정확도는 $16/(16+4) = 0.8$, 즉 80%가 되고, 적용도는 $16/40 = 0.4$, 즉 40%가 된다. 현재 MUC에서 발표된 시스템의 정확도와 적용도는 named entity 인식 과제에서는 95% 이상 까지도 이르지만 template element 인식 과제에서는 각각 60대와 40대의 백분율에 머무르고 있는 것으로 알려져 있다.

생물정보학 분야에서 요구하고 있는 정보추출 과제도 MUC에서와 같이 named entity 인식과제와 template element 인식 과제로 나뉘어 질 수 있다. Named entity 인식 과제는 주로 유전자나 단백질의 이름의 확인을 통한 해당 온톨로지의 구축에 사용하기 위한 것이며 template element 인식 과제는 주어진 문서에서 제시하고 있는 단백질이나 유전자 간의 상호 작용에 대한 정보를 추출해 내기 위한 것이다. 그런데 생물정보학에서 요구하고 있는 template element 인식은 MUC에서 요구하고 있는 일반적인 template element 인식 과제보다는 훨씬 더 구체적이라는 특징이 있는데, 예를 들어 생물정보학 분야에서는 주어진 문서에서 X와 Y라는 미리 알려지지 않은 단백질의 쌍이 activate 또는 inhibit라는 관계로 연결이 된다는 정도의 정보를 정확하게 추출해야 하는 과제를 가지고 있다. 현재 이와 같은 과제에 대하여 발표되고 있는 시스템의 정확도는 구체적인 측정방법이 통일되어 있지는 않아 상대적인 비교가 어렵지만 특정 관계의 쌍에 대한 추출에서는 각각 90%와 80%대인 것으로 알려져 있다[19,20,21,23]. 그러나 이와 같은 성능은 이 분야에서 아직 최적화된 것이라고 보기는 어려우며 앞으로 확률 기반 기법이나 도메인 의존적인 기법의 개발과 함께 더 높은 성능을 가진 정보추출 시스템이 개발될 것이라고 기대되고 있다.

4. 결 론

자연언어처리 연구분야에서 쓰이고 있는 여러가지 기법은 앞에서 살펴 본 정보 추출에서 뿐만 아니라 문헌 정보를 통한 유전자 온톨로지(Gene Ontology)의 자동 구축 및 관리, RNA 서열 정보의 분석을 통한 서열 문법(Sequence Grammar)의 구축과 이를 이용한 특정 서열 예측 및 검증, 정보추출의 정확도 측정 기준 확보와 기계 학습 등을

12) 관련 홈페이지: http://www.itl.nist.gov/iaui/894.02/related_projects/muc

위한 자동 주석 도구 개발 등 많은 곳에서 생물정보학 분야의 발전을 위하여 응용될 수 있다.

참고문헌

- [1] 정보과학회지. 특집: Bioinformatics, 18(8), 2000.
- [2] S. Ananiadou, L. Hirschman, D. Schuhmann, J. Tsujii & A. Valencia. Mining the Biomedical Literature, a tutorial, ISMB, 2001.
- [3] M. A. Andrade & A. Valencia. Automatic annotation for biological sequences by extraction of keywords from MEDLINE abstracts: Development of a prototype system, ISMB, 25-32, 1997.
- [4] M. A. Andrade & A. Valencia. Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families, Bioinformatics, 14(7): 600-607, 1998.
- [5] C. Blaschke, M. A. Andrade, C. Ouzounis & A. Valencia. Automatic Extraction of Biological Information from Scientific Text: Protein-Protein Interactions, ISMB, 60-67, 1999.
- [6] K. Backawski, J. Cigna, M. M. Kokar, P. Mager & B. Indurkha. Knowledge Representation and Indexing Using the Unified Medical Language System, PSB, 490-501, 2000.
- [7] J. T. Chang & S. Raychaudhuri, Natural Language Processing: The Time is Ripe, a tutorial, ISMB, 2001.
- [8] J. T. Chang, S. Raychaudhuri & R. B. Altman. Including Biological Literature Improves Homology Search, PSB, 374-383, 2001.
- [9] N. Collier, C. Nobata & J. Tsujii. Extracting the Names of Genes and Gene Products with a Hidden Markov Model, COLING, 201-207, 2000.
- [10] J. Ding, D. Berleant, D. Nettleton & E. Wurtele. Mining MEDLINE: Abstracts, Sentences, or Phrases?, PSB, 2002 (to appear).
- [11] C. Friedman, P. Kra, H. Yu, M. Krauthammer & A. Rzhetsky. GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles, ISMB, 74-82, 2001.
- [12] K. Fukuda, T. Tsunoda, A. Tamura & T. Takagi. Toward Information Extraction: Identifying protein names from biological papers, PSB, 705-701, 1998.
- [13] U. Hahn, S. Schulz & H. Schauer. Rich Knowledge Capture from Medical Documents in the MEDSYNDIKATE System, PSB, 2002 (to appear).
- [14] V. Hatzivassiloglou, P. A. Duboue & A. Rzhetsky. Disambiguating proteins, genes, and RNA in text: a machine learning approach, ISMB, 97-106, 2001.
- [15] K. Humphreys, G. Demetriou & R. Gaizauskas. Two Applications of Information Extraction to Biological Science Journal Articles: Enzyme Interactions and Protein Structures, PSB, 502-513, 2000.
- [16] I. Iliopoulos, A. J. Enright & C. A. Ouzounis. TEXTQUEST: Document Clustering of MEDLINE Abstracts for Concept Discovery in Molecular Biology, PSB, 384-395, 2001.
- [17] G. Leroy & H. Chen. Automated Extraction of Medical Knowledge using Underlying Logic from Medical Abstracts, PSB, 2002 (to appear).
- [18] Y. Ohta, Y. Yamamoto, T. Okazaki, I. Uchiyama & T. Takagi. Automatic Construction of Knowledge Base from Biological Papers, ISMB, 218-225, 1997.
- [19] T. Ono, H. Hishigaki, A. Tanigami & T. Takagi. Automated extraction of information on protein-protein interactions from the biological literature, Bioinformatics, 17(2), 155-161, 2001.

[20] J. C. Park. Combinatory Categorical Grammar for Biomedical Information Extraction, IEEE Intelligent Systems, November/December, 2001 (to appear).

[21] J. C. Park, H.-S. Kim & J. J. Kim. Bidirectional Incremental Parsing for Automatic Pathway Identification with Combinatory Categorical Grammar, PSB, 396-407, 2001.

[22] D. Proux, F. Rechenmann & L. Julliard. A Pragmatic Information Extraction Strategy for gathering Data on Genetic Interactions, ISMB, 279-285, 2000.

[23] J. Pustejovsky & J. M. Castano. Robust Relational Parsing over Biomedical Literature: Extracting Inhibit Relations, PSB, 2002 (to appear).

[24] T. C. Rindfleisch, L. Tanabe, J. N. Weinstein & L. Hunter. Biobliometrics: Information Retrieval and Visualization from Co-Occurrences of Gene Names in Medline Abstracts, PSB, 514-525, 2000.

[25] H. Shatkay, S. Edwards, W. J. Wilbur & M. Boguski. Genes, Themes, and Microarrays Using Information Retrieval for Large-Scale Gene Analysis, ISMB, 2000.

[26] B. J. Stapley & G. Benoit. Biobliometrics: Information Retrieval and Visualization from Co-Occurrences of Gene Names in Medline Abstracts, PSB, 526-537, 2000.

[27] B. J. Stapley, L. A. Kelley & M. J. E. Sternberg. Predicting the Sub-Cellular Location of Proteins from Text using Support Vector Machines, PSB, 2002 (to appear).

[28] J. Thomas, D. Milwarrrd, C. Ouzounis, S. Pulman & M. Carrol. Automatic Extraction of Protein Interactions from Scientific Abstracts, PSB, 538-549, 2000.

[29] J.-I. Tsujii & S. Ananiadou. An Introduction to Information Extraction, a tutorial, PSB, 2001.

[30] W. J. Wilbur. A Thematic Analysis of the AIDS Literature, PSB, 2002 (to appear).

박종철



1984. 2 서울대학교 공과대학 컴퓨터공학과 공학사
 1986. 2 서울대학교 대학원 컴퓨터공학과 석사
 1996. 5 미국 펜실바니아 대학교 전산정보학과 박사
 1996. 6~1998. 2 미국 펜실바니아 대학교 박사후과정 및 Research Associate
 1998. 3 한국과학기술원 전산학과 및 전자전산학과 전산학 전공 조교수

E-mail: park@cs.kaist.ac.kr

• 제13회 한글 및 한국어 정보처리 학술대회 •

- 일 자 : 2001년 10월 12일(금) ~ 13일(토)
 - 장 소 : 동국대학교 경주 캠퍼스
 - 내 용 : 논문발표, 초청강연
 - 주 최 : 한국어정보처리연구회
 - 문 의 처 : 동국대학교 전산정보학부 변정용 교수
- Tel. 054-770-2028