바이오 텍스트 마이닝 시스템 개발

고려대학교 임해창* · 황영숙 · 박경미

1. 서 론

바이오 인포매틱스는 생물학에서 다루는 정보의 양이 급증함에 따라 전산학, 수학, 통계학 등의 분야에서 사용되고 있는 정보 처리 기법을 응용하여 이를 효율적으로 생산, 관리, 활용하려는 연구 분야를 총칭한다[1]. 본 논문에서는 바이오 인포매틱스 분야 중 바이오 텍스트 마이닝에 대해 살펴보고자 한다. 본 논문에서 의미하는 바이오 텍스트 마이닝은 바이오 관련 텍스트에 데이터 마이닝 기술을 적용하는 것으로, 폭발적으로 증가하고 있는 바이오 관련 문헌의 효과적인 처리를 통해 필요한 지식을 획득하기 위한 것이다.

즉, 게놈 프로젝트의 성공 이후 생물학 등 바이오테크놀로지(BT)와 관련된 다양한 연구 결과가 발표되고 있으며, 관련 문헌수의 양적인 증가는 점차 가속화 되고 있다. 이처럼 생물학 분야에서는 새로운 형태의 단백질 혹은 유전자 명칭들이나, 이들간의 관계에 관한 새로운 연구 관련 문헌이 끊임없이 쏟아지고 있기 때문에 일선 분야의 학자들이나 연구자들은 점차 원하는 정보를 얻기가 어려워지고 있다. 따라서 BT관련 문헌 데이터베이스에서 유의미한 정보를 추출해 내는 바이오 텍스트 마이닝 기술의 중요성은 점점 더 강조되고 있다. 때문에 텍스트 마이닝 기술을 정보 과다로 정보 습득의 효율성과 신개발 정보에 대한 접근에 많은 문제를 겪고 있는 생물학 관련 분야에 활용한다면 생물학 연구 분야의 연구 효율성 제고에 많은 기여를 할 수 있을 것으로 예상된다. 특히 최근의 연구 결과들이 대부분 온라인 접근이 가능한 전자 문서나 데이터베이스 형태로 존재하기 때문에 이러한 마이닝 기술을 보다 효율적으로 활용할 수 있는 기반 환경은 이미 조성되어 있다고 할 수 있다.

바이오 분야에 적용되는 데이터 마이닝 기술은 생물학 관련 문서들이 자연언어로 되어 있기 때문에 1차적으로 텍스트를 분석하기 위한 자연언어처리 기술과, 고차원적인 관계를 추론하는 데이터 마이닝 관련 학습 알고리즘을 필요로 한다. 이 기술들을 적용해 바이오 관련 텍스트로부터 추출하는 유용한 정보는 단백질과 단백질, 또는 유전자와 유전자 사이의 상호작용(interaction)관계[1]이다. 텍스트로부터 개체들간에 상호작용관계를 자동으로 추출해 데이터베이스에 저장하면, 특정 단백질이나 유전자에 대한 검색을 통해 그와 상호작용관계를 갖는 모든 단백질 및 유전자 정보를 그래프 등으로 볼 수 있고 각각이 어떤 관계를 갖는지를 알 수 있게 된다. 이를 통해 생물학 관련 연구자는 여러가지로 도움을 받을 수 있다.

앞으로, 본 논문의 2장에서는 기존의 시스템에 대해 알아보고, 3장에서는 텍스트 마이닝 기술로 바이오 관련 정보를 추출하기 위해 어떤 리소스들이 구축되어야 하는지 살펴본다. 그리고 4장에서는 바이오 텍스트 마이닝을 구성하는 관련 기술들에는 어떠한 것들이 있는지, 현재 고려대학교 자연어처리연구실[2]에서 수행중인 과제를 중심으로 살펴본다. 5장에서는 바이오 텍스트 마이닝을 통해 얻을 수 있는 기대효과와 활용방안에 대해 살펴보고 6장에서 결론을 맺고자 한다.

2. 기존의 시스템

이 장에서는 기존에 구축된 시스템들에 대해서 살펴보겠다.

* 종신회원

1) 상호작용관계의 예: '활성화하다(activate)', '억제하다(inhibit)'

2) 홈페이지 http://nlp.korea.ac.kr

먼저, MedStract[3]는 개체간의 상호작용 정보를 자동으로 추출하는 시스템으로 추론 과정없이 견고한 자연언어처리 기술만을 사용한다[2]. 이 시스템에서는 바이오 텍스트에 특징적으로 나타나는 단어들을 고려해 UMLS[4] 시소러스에 수반된 사전을 사용하여 개체명과 품사를 인식한다. 그리고, 독립된 몇가지 오토마타를 단계적으로 적용해 명사구와 동사구 등을 인식하고[3, 4, 5], 정의한 패턴[5]에 따라 상호작용정보를 추출한다. 그리고 웹에 기반한 사용자 인터페이스[6]를 제공하는데 키워드 검색결과가 테이블과 그래프 형태로 주어진다.[7]

다음으로, GENIES는 생물학 관련 문헌들에서 molecular pathway를 추출하는 자연언어처리 시스템이다[6]. 이것은 GeneWay[8]를 구성하는 모듈 중 하나로 MedLEE[9]를 변형한 것이다. GENIES는 단백질 또는 유전자 명칭을 확인하는 Term tagger와 문장, 단어, 구를 결정하는 Preprocessor, 그리고 제약 규칙과 의미적 패턴으로 되어 있는 문법을 사용해 적절한 상호작용관계를 확인하는 Parser, 구문분석의 오류를 여러가지 휴리스틱을 사용해 처리하는 Error recovery 모듈로 구성되어 있다. 이 시스템은 추출된 개체간의 상호작용 정보를 이용해 pathway를 구성한다.

BIOBIBLIOMETRICS[10]는 유전자 이름을 사용해 생물학 관련 문헌 DB에서 정보를 검색하고 가시화하는 시스템이다[7]. 이 시스템은 두 유전자가 서로 관련된 생물학적 기능을 갖는다면, 생물학 문헌들 안에서 자주 공기한다는 가정에서부터 개발이 시작되었다. 문헌 DB에서 두 유전자가 공기하는 정도로부터 유사도를 구하고 특정 임계값 이상이면 관련이 있다고 판단하였다. 실제로 생물학 관련 연구자가 특정 유전자를 검색하면 그 유전자와 관련된 다른 유전자들이 검색되고 이것을 가시화해 보여주는데, 이 결과로부터 유전자와 유전자 사이의 관계를 연구하는 데 도움을 준다.

3. 바이오 텍스트 마이닝을 위한 리소스 구축

생물학 관련 분야에서 끊임없는 연구 결과로 많은 새로운 단백질 및 유전자 명칭이 생겨나고 있다. 따라서 기존의 사전을 보완하지 않는다면 새로운 개체에 관련한 정보를 텍스트로부터 자동으로 추출하기가 어렵다. 그러나 사람이 수동으로 계속해서 사전을 갱신하는 것은 너무 많은 인적, 시간적 비용이 소요된다. 그러므로, 사전에 없는 개체라 할지라도 자동으로 인식하고 개체에 대한 정보를 추출하기 위해 기계학습 방법을 적극적으로 활용한다. 그러나 기계학습을 효과적으로 수행하기 위해서는 개체와 개체들 간의 상호작용 정보가 부착된 말뭉치를 필요로 한다.

현재, 바이오 텍스트 분석을 위한 리소스 구축에 대한 대표적인 연구인 일본의 GENIA[11] 프로젝트의 경우, 관련 연구 문헌에서 자동으로 단백질 및 유전자의 이름을 추출하고 이들 간의 관계를 인식하기 위해 학습 말뭉치를 구축하고 있는데, 아직까지는 단백질 및 유전자의 이름을 추출하기 위한 소규모의 학습 말뭉치를 구축한 상태이다[8, 9]. GENIA 프로젝트에서 학습 말뭉치는 완전히 수동적인 방법에 의존하여 구축되고 있으므로 실용적으로 사용할 수 있을 만한 수준의 말뭉치를 구축하기 위해서는 오랜 시간과 노력이 요구된다. 따라서 대량의 학습 말뭉치를 수작업을 최소화 하면서 효과적으로 구축하기 위한 방법론을 필요로 하고 있다[10].

본 연구실에서 수행중인 과제에서도 생물학 분야 중 특정 주제와 관련된 문헌만을 대상으로 생물학 분야의 전문가들에 의해서 리소스가 구축된다. 전문가

3) 홈페이지 http://www.medstract.org
4) 홈페이지 http://www.nlm.nih.gov/research/umls
5) 패턴은 [argument1 relation argument2]로써 argument1은 개체명을 포함한 주격의 명사구, argument2는 개체명을 포함한 목적격의 명사구를 의미한다. 그리고 relation은 개체들간의 상호작용관계를 나타내는 'activate', 'inhibit' 등과 같은 동사들을 말한다.
6) 홈페이지 http://scylla.cs.brandeis.edu/~weiluo/relation/main.htm
7) MedStract에서 상호작용관계를 나타내는 것으로 고려한 동사의 예: diminish, inhibit, reduce, decrease, block, regulate, lessen
8) 생물학 관련 문헌에서 정보의 자동 추출과 지식 베이스의 자동 유지보수를 위한 시스템
signal-transduction pathway에 대한 정보, pathway와 관련된 질병 등에 대한 정보를 포함함
9) 홈페이지 http://leo.cpmc.columbia.edu/medleexml/
10) 홈페이지 http://www.bmm.icnet.uk/~stapleyb/biobib/
11) 홈페이지 http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/

들은 먼저, 관련문헌에서 관심의 대상이 되는 개체들을 찾아내고, 그것들이 속하는 의미 부류를 결정한다. 그리고, 개체들 간의 상호작용관계인 이벤트를 나타내는 이벤트성 동사가 무엇인지 정의하고, 문헌에 개체들 간의 관계가 이벤트성 동사로 표현되어 있으면 태그를 붙인다[그림 1]. 이 때, 한 문헌을 두사람 이상이 검토하도록 하여 최대한 객관성을 유지할 수 있게 한다. 이렇게 태깅된 문헌들은 개체명 인식과 상호작용관계 추출단계에서 학습 말뭉치로 사용된다.

본 과제에서는 효과적인 리소스 구축을 위해 공개된 바이오 관련 데이터베이스를 활용한다. 정보추출의 대상이 되는 바이오 관련 문헌은 미국 국립의료 도서관에서 제공하는 공개 DB인 MEDLINE[12]으로부터 획득한다. 유전자 명칭과 관련해서는 REBASE[13], GenBank[14] 등을 이용하고, 단백질 명칭과 관련해서는 PDB[15], PIR[16], SWISS-PROT[17] 등을 이용한다. 또한, 각 개체의 의미 분류를 위해 UMLS와 같은 바이오 관련 용어들의 의미 분류 체계를 이용한다.

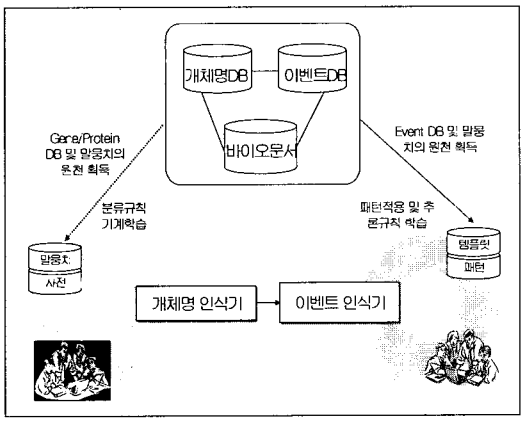

그림 1 리소스 구축 및 관리 개요도

12) 홈페이지 http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed
13) 홈페이지 http://rebase.neb.com/rebase/rebase.html
14) 홈페이지 http://www.ncbi.nlm.nih.gov/Genbank/GenbankSearch.html
15) 홈페이지 http://www.rcsb.org/pdb/
16) 홈페이지 http://pir.georgetown.edu/
17) 홈페이지 http://www.ebi.ac.uk/swissprot/access.html

4. 시스템 구성

바이오 텍스트 마이닝 시스템의 최종 목표는 바이오 텍스트 문서를 분석하여 생물학적 요소들 간의 정형화된 상호작용관계를 추출하는 시스템의 구현이다[11, 12, 13]. 이를 위해 시스템은 텍스트 분석, 관계 추론, 네트워크 가시화 모듈로 구성될 수 있다[그림 2].

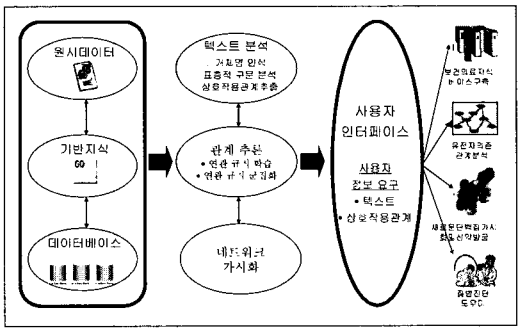

그림 2 시스템 개요도

4.1 텍스트 분석 모듈

텍스트 분석 모듈은 바이오 텍스트 마이닝 시스템의 시작부분으로서 마이닝을 수행 할 텍스트 데이터를 대상으로 자연어처리 기술을 사용하여 텍스트 분석을 수행한다. 이 모듈에서는 바이오 텍스트 마이닝의 입력으로 사용하는 문서에 대해 품사 및 통사 정보를 부착하고, 유전자 이름 등의 개체명을 인식하고 이들 간의 1차적 상호작용관계 정보를 추출한다[그림 3].

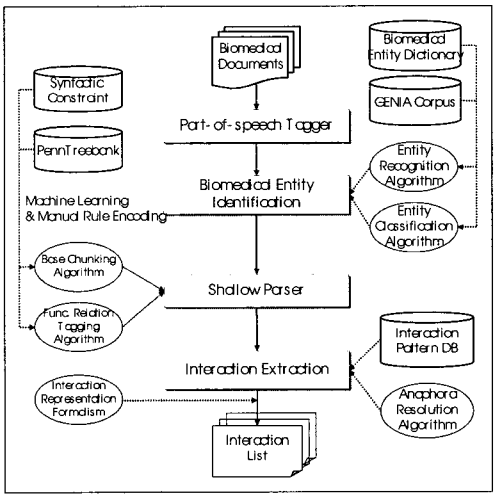

그림 3 텍스트 분석 단계 구조도

이렇게 추출된 유전자나 단백질 등과 같은 생물학 개체들 간의 상호작용 정보는 보다 확장된 상호작용 정보를 얻기 위해 다음 단계인 관계 추론 모듈로 전달되게 된다.

4.1.1 개체명 인식

개체명 인식은 개념적으로 볼 때 (1) 개체명의 경계를 구분하는 개체명의 경계 인식과 (2) 인식된 개체명의 의미적 부류를 결정하는 의미부류 결정의 두 가지 문제로 나누어 생각할 수 있다. 일반적으로 이 문제들은 분류 문제로 간주되고 은닉 마르코프 모형(HMM : Hidden Markov Model)[14], 지지 벡터 기계(SVM : Support Vector Machine)[15], 최대 엔트로피(ME : Maximum Entropy) 모델 등과 같은 방법을 사용하여 개체명 인식 모듈이 개발된다. 그러나 학습 말뭉치의 부족으로 인해 현재까지 개발된 시스템들은 만족할 만한 성능을 보이지 못하고 있는 실정이다. 특히 개체명의 경계 인식과 부류 결정을 하나의 문제로 통합하여 풀고자 하는 경우 학습 자료 부족의 문제는 더욱 심각해지는 경향이 있다. 이는 성능으로 곧바로 연결되어 처리 효율뿐만 아니라 정확도를 저하시키는 원인이 된다. 이에 두 작업을 개별 작업으로 분리해 접근할 필요가 있다. 또한 전문 용어 사전이나 수동으로 작성한 전문용어 인식 규칙을 함께 사용하여 인식 정확도를 향상시키는 방법이 시도되기도 한다.

이 두 가지 문제에 대해 본 연구에서는 개체명의 경계 인식과 개체명의 의미부류 결정을 분리하고, 기 구축된 개체명 사전과 기계학습 방법을 결합하는 방법을 사용한다. 기계학습 방법은 기존의 많은 분류 문제에서 뛰어난 성능을 보인 SVM을 활용하고, 품사, 철자형태, 내 외부 어휘 문맥 정보들 가운데에서 각 작업에 적합한 자질을 선택하여 인식기와 의미 분류기를 개발하는데 사용하고 있다.

세부적으로 SVM을 사용해 경계 인식을 하는 경우 학습 문서에 TO(Term, Others) 표기법을 사용하여 개체명의 경계 인식 태그를 부착하고 T/O를 경계 부류로 간주해 one-vs-rest 방식을 이용해 하나의 SVM 모델을 생성, 경계를 인식한다. 개체명의 경계가 인식되고 난 뒤에는 인식된 개체명들 만을 대상으로 하여 개체명의 의미 분류를 수행한다. 이때 GENIA 말뭉치의 22개 부류를 대상으로 22개의 SVM 분류기를 one-vs-rest 방식으로 학습하고 개체명 분류를 수행한다[그림 4].

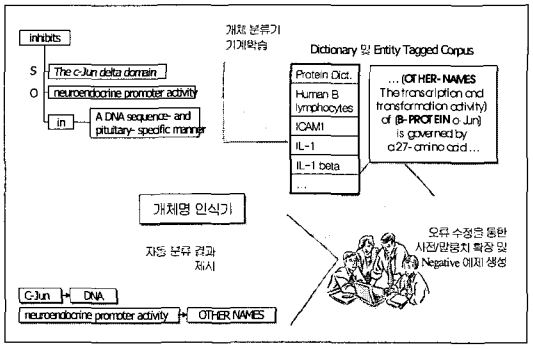

그림 4 개체명 인식 개요도

그러나 적은 양의 학습집합으로부터 기계학습을 통해 개체명 인식 및 분류기를 개발하게 되면 오류가 발생하기 쉽다. 특히 본 연구에서와 같이 두단계로 개체명을 인식하는 경우 경계 인식 단계에서의 오류는 의미부류 단계에 치명적인 영향을 미치게 된다. 그러므로 본 연구에서는 기 구축된 개체명 사전 정보를 이용하여 경계 인식기의 오류를 보정하여 성능을 향상시킨다.

4.1.2 표층적 구문분석

전처리가 끝난 문장의 구조를 분석하기 위해 완전한 구문분석(full parsing)을 수행 할 경우, 정확도가 떨어지고 분석 속도가 느려질 수 있다. 따라서 다음 단계인 이벤트 분석에서 필요로 하는 정도의 정보만을 추출하기 위해 최소한의 구문분석을 수행할 필요가 있다[16, 17, 18, 19, 20]. 여기서 말하는 최소한의 구문분석은 문장에서 기본구[18]를 인식하고 그들사이의 의존관계를 결정하는 것이다. 기본구들은 자동으로 학습된 구문 제약 규칙 및 수동으로 작성된 제약 규칙을 통하여 다른 기본구들과 결합되어 문장 구조를 생성하며, 기본구들 사이의 문법적 관계도 이 과정에서 함께 결정된다[21, 22, 23].

구문 제약 규칙만으로는 해결이 힘든 문장의 구조 분석 문제는 대량의 말뭉치로부터 학습한 통계정보를 함께 이용하여 해결하기도 한다. 위에서 설명한 제약 규칙은 Penn Treebank[24]로부터 상당한 수준의 자동 획득이 가능하다. 또, 생물학 분야에서 자주 사용되는 어휘들에 대해서는 자동으로 학습된 규칙

18) 명사구, 동사구 등을 말함

집합에 수동으로 규칙을 추가하거나 수정하여 정확한 분석을 하도록 할 수 있다. 제약 규칙을 적용하면 복수개의 구문분석 후보가 생길 수도 있는데[19], 이 때는 말뭉치로부터 학습한 통계정보를 함께 이용하여 구조적 중의성을 해결할 수 있다.

본 연구에서는 자질 기반의 통계적 접근 방법을 사용하기 위해, 대용량의 구문분석된 학습 자료인 Penn Treebank로부터 자질 및 통계 정보를 추출한다. 그리고 기본구 인식(base chunking) 결과와 통계 정보로부터 기본구들 사이의 의존 관계를 결정하고, 동시에 문법적 기능 태그(function tag)[20]를 부착한다[그림 5].

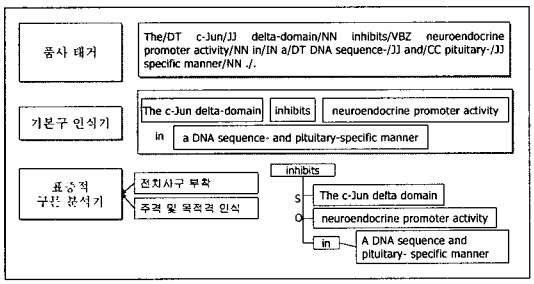

그림 5 표층적 구문분석 개요도

이 때, 기존 자질을 빼거나 새로운 자질을 추가해 적합 자질 집합을 만들어 가고, 학습 말뭉치를 사용한 통계 기반의 모듈을 구축하는 과정에서 발생하는 자료 부족 문제를 해결하기 위해 영어 단어 온톨로지인 워드넷을 활용하고 있다.

4.1.3 상호작용관계 추출

이 단계에서는 표층적 구문분석 결과로부터 상호작용관계를 추출해낸다. 구문분석 결과로부터 상호작용관계를 추출하는 기초적인 방법은 수동으로 작성된 규칙을 이용하는 것이다. 이 경우 상호작용관계 추출을 위한 패턴을 구성하는 것이 필요한데, 이는 생물학 관련 전문가에 의해 구축된다. 그러나 수동으로 필요한 패턴들을 구축하는 데는 한계가 있기 때문에 자동으로 확장할 수 있는 방법의 개발이 요구된다[25, 26, 27].

19) 예를들어, 전치사구 부착 문제

20) 학습 말뭉치로 사용하는 Penn Treebank에는 20개의 기능 태그가 있다. 이 중에서 이벤트를 추출하는데 유용한 기능 태그만을 선별할 필요가 있다.

본 연구에서는 생물학 관련 문서에서 사용자가 궁극적으로 얻고자 하는 상호작용관계를 추출하기 위해 구문분석 정보를 이용하는 상호작용관계 추출패턴을 작성한다. 즉, 생물학관련 문서에서 유전자, 혹은 단백질 개체명과 함께 고빈도로 나타나는 동사들 중 'activate'나 'inhibit'과 같이 상호작용관계를 나타내는 이벤트성 동사들을 추출해 패턴을 분석하고 분석된 패턴 정보를 활용하여 수동으로 상호작용관계 추출패턴을 작성하게 된다. 아래의 표 1은 작성된 패턴을 보여준다[28].

표 1 동사에 따른 패턴의 예

동사	패 턴
activate	NP activate NP NP be activated by NP
inhibit	NP inhibit NP NP be inhibit by NP
associate	NP associate with NP NP associate with by NP
bind	NP bind (to) NP NP binding NP

그러나 수동으로 패턴 정보를 추출하는 데는 한계가 있다. 이에 상호작용관계 추출을 위한 패턴을 반자동으로 획득할 수 있는 방법을 사용할 수 있는데, 그림 6은 상호작용관계 추출패턴을 전문가의 도움을 빌어 반자동으로 확장하는 방법을 보여준다. 이 방법은 이벤트 인식기의 결과로부터 이벤트 추출을 위한 패턴 후보를 자동으로 추출하고 추출된 패턴들을 사용하여 이벤트를 추출한 결과를 전문가에게 제시하고 검증을 받음으로써 패턴을 확장하는 방법이다.

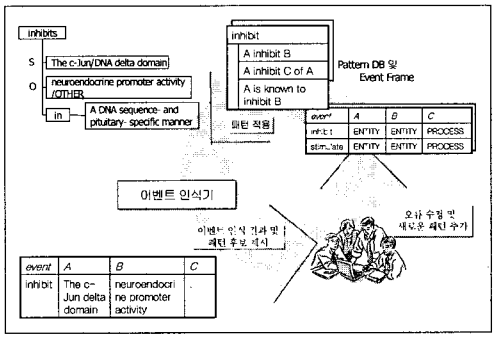

그림 6 상호작용관계 추출 개요도

또한, 조응현상 애매성 해소(Anaphora Resolution)를 통해 상호작용관계 추출의 성능을 향상시키고, 최종적으로 상호작용을 정형적으로 표현하는 형식을 결정한다. 이것은 생물학 관련 문서에서 개체간의 상호작용을 어떻게 표현할 것인지에 관한 것으로 개체간의 정적 관계 및 동적 관계를 표현하는 방법을 정의하는 것으로 실제로 생물학자들이 필요한 정보를 포함하고 있는지 검증을 받아야 한다.

4.2 관계 추론 모듈

본 모듈은 실제로 데이터 마이닝 알고리즘 및 기계학습 알고리즘을 이용하여 연관 규칙에 대한 추론 및 클러스터링 등을 수행함으로써 텍스트 분석 결과 추출된 1차적 상호작용 정보들로부터 고차원적 상호작용 정보를 추론하는 단계이다. 본 모듈에서는 이러한 상호작용 정보들을 네트워크로 표현하고 표현된 네트워크에 대한 가설을 생성하는 기능을 지원한다.

본 모듈에서 생성하는 가설은 상호작용관계 추출 규칙 및 추론된 규칙을 포괄적으로 표현할 수 있는 보다 일반화된 연관성 규칙을 의미한다. 이러한 연관성 규칙은 베이지안망(Bayesian network)과 단순 베이즈 분류기(Naive Bayes Classifier)를 이용하여 학습한다. 그리고 SVM을 사용하여 각 상호작용관계들의 부류를 결정한다. 이때 계층적 혹은 비계층적 클러스터링 알고리즘을 개발하여 연관 규칙의 대분류 모델을 개발하고, 대분류 모델에 따라 연관 규칙의 부류를 정한다. 그리고 이 클러스터링 정보 및 연관 규칙 정보를 사용하여 상호작용관계의 추론 범위를 결정하고 단백질-단백질 또는 유전자-유전자 등과 같은 생물학적 개체들 간의 고차원적 연관성 추론을 수행한다.

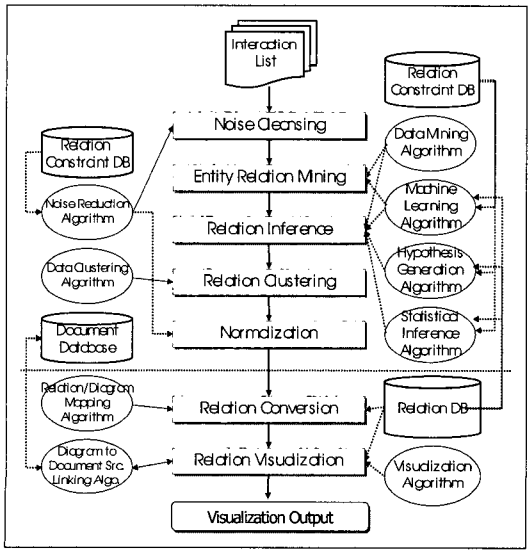

그림 7 관계 추론 및 시각화 단계 구조도

이처럼 다양한 알고리즘을 이용하여 추론한 관계를 포함하여 확장된 전체 상호작용관계 정보는 우선 로컬 DB에 별도로 체계화 되어 저장되며, 보다 효율적인 정보 전달을 제공하기 위한 모듈인 네트워크 가시화 모듈로 전달되게 된다[그림 7].

4.2.1 연관 규칙 학습

상호작용관계 분석을 통한 고차 관계 추론을 위해 기계학습 방법을 이용해 접근하는 단계이다. 이를 위해 기계학습 기법 중 베이지안망과 단순 베이즈 분류기를 이용해 연관 규칙 추론기를 학습한다. 그리고 SVM을 이용하여 각 상호작용관계 그룹에 대한 분류 모델을 개발한다. 그러나 SVM이나 베이지안망과 같은 기계학습 알고리즘은 학습 자료의 부족 등으로 인해 성능 향상에 제한을 받게 된다. 이에 이를 보완하기 위해 KL clustering과 같은 데이터 마이닝 기법을 사용하여 유전자 혹은 단백질 그룹 사이의 연관성 발견과, 상호작용관계와 유전자 혹은 단백질 사이의 연관성을 발견해 내고 이를 기계학습과 결합하는 것이다. 결국, 이 단계에서는 데이터 마이닝 기법을 이용한 연관 규칙 발견(Association rule discovery) 기법과 텍스트 마이닝에 사용되는 기계학습 기법을 결합함으로써 고차원의 상호작용관계를 추출하게 된다.

4.2.2 연관 규칙 군집화

계층적, 비계층적 군집화를 통한 연관 규칙 대분류 모델을 개발하고, 대분류 그룹간의 상호 연관성 추론을 위한 클러스터링 모델을 개발하는 단계이다. 또한 여기서 더 나가 Dynamic Bayesian Network(DBN) 모델을 이용해 추출된 관계의 분석을 통한 추론 연관 관계의 통계적 가설 생성 모델을 개발한다.

4.3 네트워크 가시화 모듈

본 모듈은 시스템에서 추출하고 추론한 단백질과 유전자등 개체들 간의 상호작용관계를 시각화 하여 보여주는 모듈이다. 이 모듈에서는 이전의 텍스트 분석 모듈과 관계 추론 모듈에서 추출된 연관성 정보

와, 추출 작업을 수행한 원본 문서를 연결하거나 문서에서의 연관성 출현 빈도 등을 이용하여 계산된 생물학 개체들 간의 상호작용 가중치(weight)등 여러 가지 정보를 다양한 방법으로 표현하여 사용자로 하여금 시스템이 제공하는 연관성 정보에 대한 신뢰성 정도를 확인할 수 있도록 한다.

이 모듈에서는 그래프나 다이어그램, 네트워크 구조를 이용하여 추출된 관계를 시각화 하여 표현하는데, 추출 및 추론된 개체들 간의 상호관계를 시각화할 수 있는 통합 인터페이스를 제공한다[그림 8].

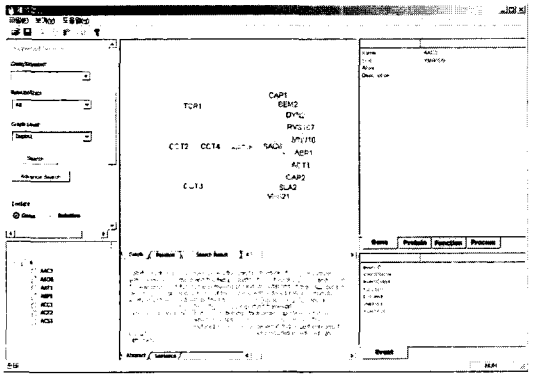

그림 8 가시화를 위한 인터페이스의 예

또한 대상과 관계에 대한 출처 검색 및 참조를 통한 신뢰도 확인 기능을 제공하고 사용자의 연관성 정보 요청에 대한 근거 문서 제시 기능을 제공할 수 있도록 한다.

5. 기대성과 및 활용 방안

향후 생물학 연구와 유전자 정보가 급증하고 관련 연구 결과가 폭발적으로 증가할 것으로 예상되는데, 이는 생물학 관련 연구 문헌의 연구 결과를 자동으로 요약 및 정리할 수 있는 바이오 텍스트 데이터 마이닝 관련 기술의 필요성 증가를 가져올 것이다. 따라서 바이오 텍스트 마이닝 관련 기술은 실제로 바이오 테크놀로지 산업이 활성화 되기 위한 필수기술로서 그 중요성이 증가할 것으로 예상된다. 바이오 텍스트 마이닝 기술은 바이오 인포매틱스 기술을 활용한 구체적인 결과물 획득 시기를 크게 앞당길 수 있는 것으로 알려져 있다[29, 30].

우선, 바이오 텍스트 마이닝 기술을 기반으로 신약 개발 등 BT관련 연구 개발에 소요되는 시간적, 금전적 비용을 획기적으로 줄여, BT관련 제반 기술이 급속도로 성장할 수 있을 것으로 예상되며 생명공학 관련 문헌으로부터의 개체명 자동 인식을 통한 GeneBank등 관련 리소스의 자동 확장 기술의 개발이 앞당겨 질 수 있을 것이다. 또한, 사용자가 제시한 키워드로부터 검색된 생명 공학 관련 문헌 집합의 클러스터링 및 실시간 자동 분류를 통한 문헌 검색 서비스의 지능화가 가능해지고 표층 구문분석이나 조응어 애매성 해소 기술 등을 활용한 언어처리 기술의 발달은 다국어 바이오 텍스트 문헌검색시스템 연구 및 번역시스템 연구 활성화에도 기여할 것으로 예상된다.

6. 결 론

앞으로 생물학의 다양한 분야에 걸쳐 실험 결과나 연구 문헌을 집약·정리한 데이터베이스들이 탄생될 것이다. 따라서, 더욱 방대해지는 데이터를 효율적으로 요약·정리 및 추론을 하여 가시화 하여 주는 바이오 텍스트 마이닝 시스템의 개발은 필수적이다. 그러나, 아직까지 바이오 인포매틱스의 전반적인 과정에서 볼 때, 바이오 텍스트 마이닝 관련 기술은 초기 단계에 머물러 있다. 이는, 생물학 관련 문서에서의 마이닝 전반을 자동화 하거나 기계학습과 같은 학습 알고리즘을 이용한 고급 마이닝 기법에 관한 연구가 아직 미비하기 때문이다. 그러므로, 이 분야에 대한 지속적인 연구와 투자가 필요하고 그 결과로 생물학 분야의 유용한 시스템이 개발된다면, 생물학 연구자들의 연구 능률을 향상시키고 최신 연구정보의 효율적 제공을 가능하게 할 것이다.

참고문헌

[1] 박종철, "생물정보학과 자연언어처리", 정보과학회지, 19(10), 2001.

[2] J. Pustejovsky, J. Castano, R. Sauri, A. Rumshinsky, J. Zhang and W. Luo, "Medstract: Creating large-scale information servers for biomedical libraries," Proceedings of the Association for Computational Linguistics (the Workshop on Natural Language Processing in the Biomedical Domain), pp. 85-92, 2002.

[3] D. Hindle, "Deterministic parsing of syntactic non-fluencies," In Proceedings of the 21st

Annual Meeting of the Association for Computational Linguistics, 1983.

[4] D. McDonald, "Robust partial parsing through incremental multi-algorithm processing," In P. Jacobs, editor, Text-based Intelligent Systems, 1992.

[5] J. Pustejovsky, B. Boguraev, M. Verhagen, P. Buitelaar, and M. Johnston, "Semantic indexing and typed hyperlinking," In AAAI Symposium on Language and the Web, Stanford, CA, 1997.

[6] C. Friedman, P. Kra, H. Yu, M. Krauthammer and A. Rzhetsky, "GENIES : A Natural-Language Processing System for the Extraction of Molecular Pathways from Journal Articles," Bioinformatics 2001.

[7] B. Stapley and G. Benoit, "BIOBIBLIOMETRICS : Information Retrieval and Visualization from Co-occurrences of Gene Names in MEDLINE Abstracts," PSB 2000.

[8] Ohta, Tomoko, Yuka Tateisi, Hideki Mima and Jun'ichi Tsujii. (2002). GENIA Corpus: an Annotated Research Abstract Corpus in Molecular Biology Domain. In the Proceedings of he Human Language Technology Conference (HLT 2002).

[9] Ohta, Tomoko, Yuka Tateisi, Jin-Dong Kim and Jun'ichi Tsujii. (2002). The GENIA Corpus: an Annotated Corpus in Molecular Biology Domain. In the Proceedings of the 10th International Conference on Intelligent Systems for Molecular Biology (ISMB 2002) poster session.

[10] U. Hahn, M. Romacker and S. Schulz, "Creating Knowledge Repositories from Biomedical Reports : The MEDSYNDIKATE Text Mining System," PSB 2002.

[11] T. Rindflesch, L. Tanabe, J. Weinstein and L. Hunter, "EDGAR : Extraction of Drugs, Genes and Relations from the Biomedical Literature," PSB 2000.

[12] L. Wong, "PIES, a Protein Interaction Extraction System," PSB 2001.

[13] C. Aone, L. Halverson, T. Hampton and M. Ramos-Santacruz, "SRA : Description of the IE System used for MUC-7," MUC-7 1998.

[14] N. Collier, C. Nobata and J. Tsujii, "Extracting the Names of Genes and Gene Products with a Hidden Markov Model," COLING 2000.

[15] J. Kazama, T. Makino, Y. Ohta, and J. Tsujii, "Tuning support vector machines for biomedical named entity recognition," 2002, pp. 1-8.

[16] S. Abney, "Partial Parsing via Finite-State Cascades," In Proceedings of the ESSLLI '96 Robust Parsing Worksop, 1996.

[17] S. Ait-Mokhtar and J-P Chanod, "Subject and object dependency extraction using finite-state transducers," In Proceedings of the ACL/EACL '97 Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources, 1997.

[18] S. Argamon, I. Dagan and Y. Krymolowski, "A memory-based approach to learning shallow natural langugae patterns," In Proceedings of the 36th ACL, 1998.

[19] G. Grenfenstette, "Light parsing as finite-state filtering," In Proceedings of Workshop on Extended Finite State Models of Language, ECAI '96, Budapes, Hungary, 1996.

[20] H. Jerry, R. Douglas, E. Appelt, J. Bear, D. Israel, M. Kameyama, M. Stickel, M. Tyson, "FASTUS: A Cascaded Finite-State Transducer for Extracting Information from Natural-Language Text," In Finite State Devices for Natural Language Processing. MIT Press, Cambridge, MA, 1996.

[21] T. Brants, W. Skut, B. Krenn, "Tagging grammatical fuctions," In Proceedings of the 2nd Conference on EMNLP, 1997.

[22] A. Voutilainen and J. Heikkila, "An English Constraint Grammar(ENGCG), a surface-syntactic parser of English," In "Creating and using English language corpora," 1993.

[23] D. Blaheta, E. Charniak, "Assigning function tags to parsed text," In Proceedings of the 1st Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), 2000.

[24] M. Marcus, B. Santorini, M. A. Marcinkiewicz, "Building a Large Annotated Corpus of English:The Penn Treebank," Computational Linguistics, Vol 19, No 2, 1993.

[25] G. Leroy and H. Chen, "Filling Preposition-based Templates to Capture Information from Medical Abstracts," PSB 2002.

[26] M. Andrade and A. Valencia, "Automatic Extraction of Keywords from Scientific Text : Application to the Knowledge Domain of Protein Families," Bioinformatics 1998.

[27] E. Marcotte, I. Xenarios and D. Eisenberg, "Mining Literature for Protein-Protein Interactions," Bioinformatics 2001.

[28] 전홍우, 황영숙, 임해창, "패턴 정보를 이용한 효모 관련 문서에서의 이벤트 자동 추출", 춘계 정보과학회 학술대회, 2003.

[29] 정재훈, "제6장 생물정보학과 인터넷 자원", 한유전학회지, 2000.

[30] 김창훈, "제8장 염기서열을 분석하거나 이용하는 기능유전체학을 위한 생물정보학", 한국유전학회지, 2000.

임 해 창

1990 Texas주립대학 컴퓨터학과 박사
1991~현재 고려대학교 컴퓨터학과 교수
1998. 5~2000. 5 정보과학회 한국어정보처리연구회 운영위원장
2001~현재 ACM Transaction on Asian Language Information Processing Associate Editor
관심분야 : 자연어처리, 정보검색, 생물정보학
E-mail : rim@nlp.korea.ac.kr

황 영 숙

1991 고려대학교 전산과학과 학사
1991~1995 쌍용정보통신 근무
1998 고려대학교 컴퓨터학과 석사
2003 고려대학교 컴퓨터학과 박사
2003~현재 고려대학교 정보통신공동기술연구소 연구 교수
관심분야 : 자연어처리, 기계학습, 생물정보학
E-mail : yshwang@nlp.korea.ac.k

박 경 미

1998 연세대학교 식품영양학과 학사
2000 연세대학교 기계전자공학 부학사
2002 연세대학교 컴퓨터과학과 석사
2002~현재 고려대학교 컴퓨터 학과 박사과정
관심분야 : 자연어처리, 생물정보학
E-mail : kmpark@nlp.korea.ac.kr