패턴 정보를 이용한 효모 관련 문서에서의 이벤트 자동 추출

전홍우[O] 황영숙 임해창
고려대학교
{hwchun[O], yshwang, rim}@nlp.korea.ac.kr

Automatic Event Extraction from the Yeast Literature by Pattern Matching

Hongwoo Chun[0] Youngsook Hwang Haechang Rim
Dept. of Computer Science and Engineering, Korea University

요 약

생명과학 관련 문서에서 자동으로 이벤트를 추출하는 것은 관련 연구자들의 연구에 많은 도움을 줄 수 있다. 본 논문에서는 생명과학 관련 문서 중 특히 효모와 관련된 문서를 대상으로 간단한 자연언어 처리 기술을 적용하여 유의미한 정보를 추출한 결과를 제시하고자 한다. 실험은 효모 관련 문서에서 고빈도의 이벤트 표현 동사에 대한 패턴 정보를 조사한 후, 패턴 정보에 의거하여 이벤트를 추출하였다. 평가 결과, 비교적 간단한 자연언어 처리 기술만으로도 유의미한 정보들을 추출할 수 있었다.

1. 서 론

과학 지식을 획득하기 위한 방법 중 가장 쉽게 접할 수 있는 매체가 문서이다. 그런데 이런 문서자료는 인터넷의 발전과 더불어 날이 갈수록 그 양이 기하급수적으로 늘어나고 있고, 이런 많은 양의 자료로부터 전문가가 특정 정보를 얻어내는 데는 한계가 있다[1].

여러 과학 분야 중 생명과학 분야는 게놈 프로젝트의 성공적인 수행으로 인간의 DNA구조를 파악했고 계속해서 각 유전자 및 단백질의 기능을 파악하고자 할 뿐만 아니라 생물학적 경로(pathway)에 있어서의 역할도 분석하고자 한다. 이러한 분석 결과들은 질병 진단이나 신약개발 등과 같은 생명과학 분야에서 매우 중요한 정보로써 그 중요성이 더해가고 있다. 그런데, 이러한 분석 결과를 도출해내기 위해서는 기초적인 작업으로 유전자와 유전자, 단백질과 단백질, 또는 유전자와 단백질[1]간의 상호작용에 대한 정보를 파악하는 것이 필요하다[2].

정보 추출은 자연어로 된 문서를 분석하여 사용자가 원하는 정보를 선별하고, 그 결과를 정제되고 가공된 형태로 제시하는 것이다[3]. 생명과학 분야에서의 관심대상 정보는 주로 개체들 간의 관계성 또는 상호 작용이라고 볼 수 있다. 그런데 실제 문장에서는 이런 개체들 간의 관계들도 나타나 있지만 특정 작업 또는 작업의 부산물, 개체의 생물학적 기능들과의 관계들도 많이 나타나 있기 때문에 본 논문에서의 이벤트는 개체간의 상호작용 정보뿐만 아니라 개체와 기능간의 관계, 개체와 작용간의 관계성 등도 추출대상에 포함하고자 한다.

본 논문에서 사용하는 방법은 원시 말뭉치에 대한 기본구 인식 결과로부터 정보를 추출하는 것이다. 이 때 모든 문장을 고려하는 것이 아니라 관심대상이 되는 이벤트를 표현하는데 주로 사용된 특정 동사를 미리 정하여 이 동사가 나타나 있는 문장만을 대상으로 하였다. 즉, 각 동사들이 이벤트를 기술하는데 사용되는 패턴을 분석하여 해당 동사에 의한 이벤트를 추출하였다.

위와 같이 동사의 패턴 정보에 의거하여 추출한 결과 정확률은 85.71%, 재현율은 59.26%를 얻었다.

본 논문의 최종 목표는 현재에도 진행 중인 효모 관련 연구의 결과인 문서들로부터 개체간의 상호작용뿐만 아니라 이벤트가 발생하는 상황 정보까지 자동으로 분석하는 것이다.

2. 관련 연구

생물 관련 문서에서의 정보 추출에 대한 연구는 우리나라보다는 외국에서 오래전부터 진행되어 왔다. 대부분의 연구들이 생명과학 관련 요약 문서인 'Medline'을 이용한다.

J.Pustejovsky (2002)의 'Medstract'는 'Medline'의 천만개 요약 문서를 기본구 인식과 부분 구문분석의 결과로부터 특정 8개 동사에 대한 이벤트를 추출하였다. 특히, 이 방법은 명사의 의미적/구조적 특징, 인칭/수 일치 등의 자질을 고려하여 조응 분석(Anaphora resolution)까지도 해결하였다. 실험 결과, 정확률은 90%, 재현율은 59%였다[4].

J.Sluka (2001)의 'PDQ_MED'는 'Medline'에 나타난 모든 단어간의 공기 정보를 이용해 그룹을 짓고 이를 토대로 그 관계를 추출하였다. 이러한 접근방법의 문제는 공기한 적이 없는 개체들도 같은 그룹에 속하여 무관한 개체들간에도 관계성이 발생하는 경우가 있다는 점이다[5].

TK Jenssen (2001)의 'PubGene'은 13,712개의 인간의 유전자간의 관계를 분석했으며 마이크로어레이기술을 이용하여 질병 예방 프로그램에 적용하기도 하였다[6].

L. Tanabe (1999)의 'MedMiner'는 인간 유전자들의 기능 정보를 분석한 'GeneCards'와 'Medline'을 검색하는 'PubMed'를 이용하여 개체들간의 관계성을 추출하였다[7].

3. 본 론

3.1 시스템 구성

본 논문에서 시도한 정보추출 시스템은 그림1과 같이

1) 유전자와 단백질을 개체라고 통칭하겠다.

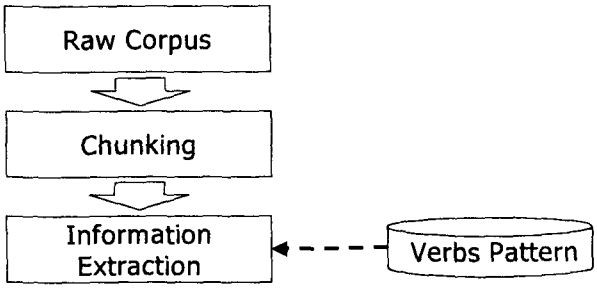

그림 1 시스템 구성도

기본구 인식과 정보추출 2단계로 구성된다. 첫 번째 기본구 인식 단계는 원시말뭉치(Raw Corpus)를 입력받아 기본구 인식을 한다. 기본구 인식에서 사용되는 품사집합 및 구문 범주는 Penn Tree Bank WSJ[2)]에 사용한 것과 동일하다[8].

두 번째 정보 추출 단계는 효모 관련 문서에 나타나는 동사 중 고빈도의 특정 동사만을 대상으로 패턴을 분석하고 분석된 패턴 정보를 사용하여 이벤트를 추출하게 된다. 이때, 이벤트는 이진관계로 제한하였으며 관계 정의에 사용된 각 항은 개체, 기능, 작용, 또는 선행 이벤트가 될 수 있다.

표 1 동사에 따른 패턴 분석

동사	패턴
Inhibit	NP inhibit NP NP be inhibit by NP
Associate	NP associate with NP NP be associate with NP
Induce	NP induce NP NP be induce by NP
Bind	NP bind (to) NP NP binding NP
Activate	NP activate NP NP be activated by NP

표1은 이벤트 추출을 위한 주요 패턴을 보여주고 있는데, 각 패턴은 아래와 같은 정규 표현식으로 정리할 수 있다.

$$((CCNP)^*(PPNP)^*(COMMANP)^*(ADVP)^*)^*(NP)^+(ADVP)^*$$
$$(Verbs)$$
$$(PP)^*(ADVP)^*(NP)^+((ADVP)^*(CCNP)^*(PPNP)^*(COMMANP)^*)^*$$

그림 2 동사 패턴에 대한 정규 표현식

동사를 기준으로 앞쪽은 해당 관계 추출 정보의 첫 번째 항을 인식하는 것으로, 부사구(ADVP)의 삽입은 항상 가능하고 하나이상의 명사구(NP), '명사구 대등구(CC) 명사구', '명사구 전치사구', 또는 '명사구 ,(쉼표) 명사구'의 패턴에 맞는 정보를 해당항으로 추출하게 된다. 동사의 뒷쪽은 관계 추출 정보의 두 번째 항을 추출하는 것으로, 전치사를 고려한다는 점을 제외하고는 앞쪽과 동일하다. 실제 실험에서는 이 정규 표현식에 대한 오토마타를 구성하여 이벤트를 추출하였다.

3.2 실험

실험은 'Medline'에서 효모와 관련된 문서를 대상으로 개체들간 발생하는 상호작용의 추출을 시도하였다. 효모에 대한 연구는 생물의 다른 분야에 비해 오랫동안 진행되고 있어서 체계적인 연구가 이미 어느 정도 진척되어 있다. 그렇기 때문에 추출된 정보의 유용성을 판단할 수 있는 정보가 많고 또한 기존의 정보를 바탕으로 더 많은 유의미한 정보를 추출할 수 있다는 장점이 있다[9].

사용한 말뭉치는 SGD에서 제공하는 효모 관련 요약문서 143개이다. 이 말뭉치로부터 동사 5개가 출현한 문장 중 81개의 이벤트가 발생하는 것을 대상으로 실험하였다. 실제로 추출한 예가 그림3이다.

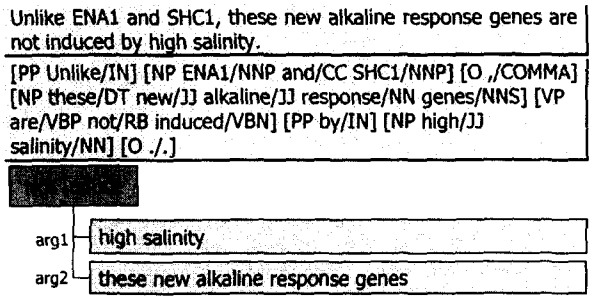

그림 3 예제 1

그림3에서 첫 번째가 원시 말뭉치이고 두 번째가 기본구 인식결과이다. 이로부터 이벤트가 추출되어 'arg1'은 주체, 'arg2'는 객체가 된다. 예제1에서 'induce'는 수동태 동사이기 때문에 실제 문장과의 배치는 반대가 된다. 또한 동사가 포함된 기본구 안에 부정을 의미하는 'not'이 있기 때문에 추출된 관계에 'Not'이 포함되어 있다.

다른 예로 숙어에 대한 결과는 그림 4이다.

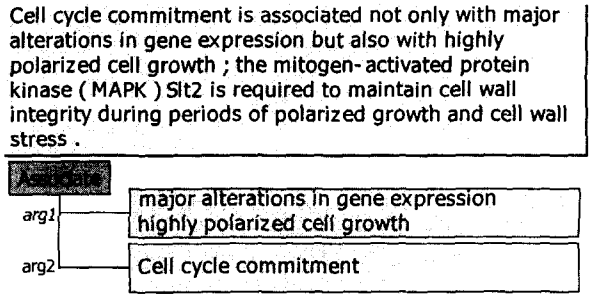

그림 4 예제 2

'not only A but also B'와 같은 숙어의 경우, A와 B는 똑같이 취급되어야 한다. 이 문장에서 A부분에 해당하는 'major alterations in gene expression'이 주체이기 때문에 B부분에 해당하는 'highly polarized cell growth'도

2) NP는 명사구, ADVP는 부사구, CC는 대등 접속사, PP는 전치사, COMMA는 쉼표이다.

주체가 된다.

3.3 평가

평가 척도는 정확률, 재현율과 F-measure를 사용하였고 다음과 같이 정의된다.

$$정확률 = \frac{정답과일치하는이벤트수}{시스템이추출한이벤트수}$$

$$재현율 = \frac{정답과일치하는이벤트수}{정답이벤트수}$$

$$F-measure = \frac{2 \times 정확률 \times 재현율}{정확률 + 재현율}$$

81개의 이벤트에 대한 정답은 수작업으로 분석한 결과이며 평가 결과는 표2와 같다.

표 2 실험 결과

정확률	재현율	F-measure
85.71%	59.26%	70.07%

표2는 전체 동사에 대한 결과이고 각 동사별로 분석한 결과는 표3과 같다.

표 3 동사별 실험 결과

동사	정확률	재현율	F-measure
induce	88.89%	64%	74.42%
associate	100%	76.92%	86.96%
bind	69.23%	64.29%	66.67%
inhibit	80%	57.14%	66.67%
activate	90%	40.91%	56.25%

이 결과를 보면 'associate'의 정확률이 100%로 패턴 정보만으로도 이벤트 추출 결과가 우수한 것을 볼 수 있다. 표2나 표3의 재현율을 보면 비교적 저조한데 이는 패턴 정보만으로는 모든 이벤트를 추출하는데 한계가 있음을 의미한다.

3.4 오류 분석

오류는 크게 두 가지로 구분된다. 하나는 시스템의 이벤트 추출 결과 오류이고 다른 하나는 시스템이 이벤트를 추출하지 못하는 오류이다.

시스템의 이벤트 추출 결과 오류 중 가장 어려운 문제는 조응 분석이 이루어져야 해결할 수 있는 문제이다. 이번 실험은 제약 사항을 한 문장에서의 이벤트 추출로 하였기 때문에 하나의 문장내에서 유추해낼 수 없는 경우에는 올바른 이벤트 추출이 불가능하다. 또한 한 문장내에 정보가 포함되어 있더라도 현재의 패턴만을 이용한 실험으로는 추출해낼 수가 없다. 이는 개체명 인식과 문법적 기능어 부착 등의 추가 작업을 통해서 해결할 것이다.

시스템이 추출하지 못하는 오류는 개체와 동사 간에 부사구외의 구문이 삽입되어 있는 경우에 발생한다. 이것은 어휘 정보를 이용하거나 패턴의 확장으로 해결할 수 있다. 또 다른 오류로는 기본구 인식 오류로 추출되지 않는 경우이다. 이것은 유전자나 단백질의 개체명 등과 같은 신조어가 어휘 사전에 등록되어 있지 않기 때문에 발생하는 것으로써 이것 역시 개체명 인식 등의 추가 작업으로 해결할 수 있다.

4. 결 론

본 논문에서는 현재 수준에서 사용가능한 비교적 간단한 자연언어 처리 기술로 어느 정도의 유의미한 이벤트 정보를 추출할 수 있는가를 알아보고자 하였다. 이를 위해 기본구 인식을 하고, 기본구 인식결과로부터 이벤트 추출을 위한 패턴 정보를 추출하여, 효모 관련 문서에 적용해 보았다. 실험결과는 간단한 패턴 정보만으로도 복잡도 높은 처리를 수행하여 획득한 결과와 유사한 수준의 성능을 획득할 수 있었다.

그러나, 좀더 세밀한 정보 분석 및 높은 성능을 위해서는 개체명 인식이나 문법적 기능어 부착과 같은 고급한 자연언어 처리 기술이 필요하였다. 이에 향후 연구에서는 개체명 인식 및 문법적 기능 등과 같은 추가적인 언어 정보를 활용하는 방안을 연구해 보고자 한다.

참고 문헌

[1] D. Proux et al., "A pragmatic information extraction strategy for gathering data on genetic interactions", In ISMB. 8, 279-285, 2000.

[2] Toshihide Ono et al., "Automated extraction of information on protein-protein interactions from the biological literature", In Bioinformatics, vol 17 no 2, pp. 155-161, 2001.

[3] Ralph Grishman, "Information Extraction : Techniques and Challenges", In Proceedings of the Seventh Message Understanding Conference(MUC-7), Columbia, MD, April 1998.

[4] J.Pustejovsky et al., "Medstract : Creating Large-scale Information Servers for biomedical libraries", In Proceedings of the Workshop on Natural Language Processing in the Biomedical Domain. pp.85-92, 2002.

[5] J. Sluka, "Mining the Biomedical Literature ; A Key Capability for Genomics Research", CAMDA-01, 2001.

[6] TK Jenssen et al., "A literature network of human genes for high-throughput analysis of gene expression". In Nat Genet. vol 28 no 1, pp.21-28, 2001.

[7] L. Tanabe et al., "MedMiner: an Internet text-mining tool for biomedical information, with application to gene expression profiling", In Biotechniques. vol 27 no 6, pp.1210-1214, pp.1216-1217, 1999.

[8] Young-Sook Hwang et al., "Weighted Probabilistic Sum Model based on Decision Tree Decomposition for Text Chunking", In International Journal of Computer Processing of Oriental Languages, vol 16 no 1, 2003.

[9] Hodges, P.E., Payne, W.E. and Garrels, J.I., "The Yeast Protein Database (YPD): a curated proteome database for Saccharomyces cerevisiae", Nucleic Acids Research, vol 26 no 1, pp.68-72, 1998.