Biomedical Event Extraction based on Co-training with Co-occurrence Information and Patterns

공기정보와 패턴 정보의 Co-training에 의한 바이오 이벤트 추출

Hongwoo Chun[1*], Young-Sook Hwang[1], Hae-Chang Rim[1]

[1] Department of Computer Science and Engineering, Korea University, Seoul, Korea

E-mail: {hwchun, yshwang, rim}@nlp.korea.ac.kr

Abstract

생명과학 관련 문서에서의 이벤트 추출은 관련 연구자들의 연구에 많은 도움을 줄 수 있다. 기존의 연구에서는 주로 이벤트 동사에 대해 패턴을 정의한 후에 정의된 패턴에 의해서만 이벤트를 추출하고자하였다. 그러나 모든 패턴을 수동으로 정의하는 것은 너무 많은 비용이 들기 때문에 패턴을 자동 추출 또는 확장하는 방법이 필요하다. 또한 학습을 하기 위해서는 상당수의 학습 말뭉치가 있어야 하는데 그것 또한 충분하지 않은 실정이다.

본 논문에서는 초기 패턴에 의해 생성된 소량의 정답 이벤트로부터 학습한 후 공기정보와 패턴정보를 이용한 Co-training방법으로 패턴 확장 및 이벤트 추출을 시도하였다. 실험 결과, 이벤트 동사의 패턴 정보가 유용한 정보라는 것을 확인할 수 있었고, 후보 이벤트 내의 개체간 공기정보와 문법관계정보 또한 매우 중요한 정보라는 것을 새롭게 보일 수 있었다. GENIA 말뭉치에서 162개의 이벤트 동사에 대해 실험한 결과, 88.02%의 정확률, 79.25%의 재현율을 얻었다.

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].

그 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 신약 개발 등과 같은 생명과학 분야에서 매우 중요한 정보로써 그 중요성이 더해가고 있다. 그런데, 이러한 분석 결과를 도출해 내기 위해서는 기초적인 작업으로 유전자와 유전자, 단백질과 단백질, 또는 유전자와 단백질[1]간의 상호작용에 대한 정보를 파악하는 것이 필요하다[2].

본 논문에서는 개체간의 상호작용에 대한 이벤트 추출 방법을 제안하고자 한다. 이벤

1) 유전자와 단백질 등 GENIA 말뭉치의 개체명 인식결과에서 개체로 분류된 것들을 개체(Entity)라고 통칭하겠다.

트 추출이란, 자연어로 된 문서를 분석하여 사용자가 원하는 정보를 선별하고, 그 결과를 정제되고 가공된 형태로 제시하는 것이다[3]. 생명과학 분야에서의 주요한 이벤트 추출대상은 개체들 간의 관계성 또는 상호작용이라고 볼 수 있다.

현재까지의 연구들은 특정 이벤트 동사에 대해 패턴을 정의하고 그것에 의해서 이벤트를 추출하였다. 이벤트 동사에 대해서 모든 패턴을 구축하는 것은 사실상 불가능하며 많은 비용을 요구한다. 학습을 통한 패턴의 확장을 시도하는 연구도 있지만 현재는 학습말뭉치의 양이 부족하기 때문에 제대로 된 학습이 불가능한 상태이다.

본 논문에서는 원시 말뭉치에 대한 기본구 인식, 개체명 부착과 기본구간의 문법관계 분석의 결과로부터 정보를 추출한다. 제안된 방법에서는 모든 문장을 고려하는 것이 아니라 관심대상이 되는 상호작용을 표현하는데 주로 사용된 특정 동사를 미리 정하여 이 동사가 나타나 있는 문장만을 대상으로 한다.

패턴의 자동 확장과 더불어 올바른 이벤트를 추출하기 위해 제안한 방법은 Co-training으로써 다음의 작업을 반복 수행함으로써 재학습이 이루어지도록 한다. 첫째, 후보 이벤트에 대해 초기 패턴정보, 후보 이벤트 안에 나타난 개체들의 공기 정보와 문법 관계 분석 결과를 토대로 후보 이벤트들의 순위를 결정한다. 둘째, 모든 패턴에 대해 이들이 표현하는 이벤트들의 순위를 이용하여 패턴가중치를 계산하게 된다. 셋째, 패턴가중치에 의해 다시 후보 이벤트의 순위를 결정하게 된다. 위 방법을 후보 이벤트의 순위 변화가 없을 때까지 반복 수행하여 올바른 이벤트를 추출할 수 있게 된다.

2. 관련 연구

생물 관련 문서에서의 이벤트 추출에 대한 연구는 국내보다는 국외에서 오래전부터 진행되어 왔다.

J.Pustejovsky (2002)의 'Medstract'는 'Medline'의 천만 개 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 조응분석(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 상호작용 관계를 추출하였다. 이러한 접근방법의 문제는 공기한 적이 없는 개체들도 같은 그룹에 속하여 무관한 개체들 간에도 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 개체들 간의 관계성을 추출하였다[7].

이들 모두의 문제점은 정의한 패턴에 의존적이어서 재현율이 낮다는 것이고, 그 패턴의 구축이 특정 도메인에 국한되어서 다른 도메인에 적용하기 어렵다는 것이다.

3. 본 론

3.1 시스템 구성

본 논문에서 시도한 정보추출 시스템은 생물학 관련 문서에서 나타나는 이벤트정보를 추출하는 것이다. 이벤트란, 개체간의 유의미한 상호 정보로써 본 논문에서는 하나의 이벤트 동사에 대해 주체가 되는 개체와 객체가 되는 개체간의 이진관계로 제한하였다. 시스템의 구성은 [그림 1]과 같다.

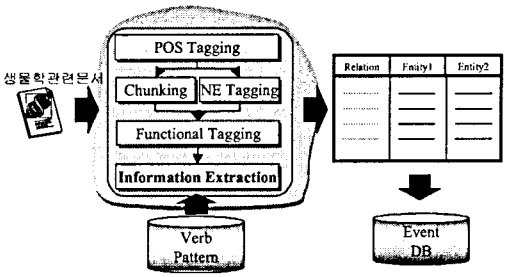

그림 1 시스템 구성도

3.2 전처리 작업

그림 1에서처럼 이벤트 추출 모듈은 세 가지 전처리 작업을 통한 결과를 입력받게 된다.

첫 번째 품사부착(POS Tagging) 및 기본구 인식(Chunking) 모듈은 원시말뭉치(Raw Corpus)를 입력받아 품사를 부착하고 기본구를 인식한다[8].

두 번째 개체명 인식 모듈은 주어진 문장에서 DNA, RNA, 단백질, 유전자 등과 같은 개체명을 인식한다[9].

세 번째 기능 태깅(Function Tagging)모듈은 기본구간의 문법적 관계를 분석한다[10].

3.3 Co-training

일반적으로 두 개의 분류학습기를 사용한 Co-training은 아래와 같이 기술할 수 있다[11].

· 소량의 초기 학습 말뭉치에 의해 두 개의 분류기를 학습시킨다.
· 분류정보가 부착되어 있지 않은 말뭉치를 각각의 분류기를 통해 분류한다.
· 각 분류기의 분류 결과 말뭉치를 학습 말뭉치에 추가시킨다.
· 위 작업을 반복 수행함으로써 각 단계에서 얻어지는 추가된 학습 말뭉치에 의해 각 분류기는 재학습하게 된다.

이와 같은 방법은 소량의 학습말뭉치와 대량의 원시말뭉치를 이용하여 분류기의 성능을 향상시킬 수 있을 뿐만 아니라 다량의 학습 말뭉치도 얻을 수 있다는 장점이 있다.

본 논문에서 Co-training방법을 이벤트 추출에 적용하기 위해 소량의 학습 말뭉치와 두 가지 분류방법을 다음과 같이 고려하였다. 소량의 학습 말뭉치는 초기 패턴에 의해서 추출된 이벤트정보를 부착한 것으로 하였다. 그러므로 초기 패턴은 굉장히 명확하고 간단한 형태를 갖는다[표1]. 분류 방법 하나는 후보 이벤트 내에 나타난 개체들 간의 공기정보, 문법 관계 분석 결과로 계산된 이벤트의 신뢰도와 해당 후보 이벤트를 나타내는 모든 패턴들의 가중치에 의한 이벤트 분류이다. 다른 분류 방법은 현재 생성되어 있는 모든 패턴들 각각에 대해 이 패턴으로 표현된 모든 이벤트들의 신뢰도 계산 결과에 의한 이벤트 분류이다. 두 분류 방법은 서로 다른 분류 방법의 결과에 의해 다시 계산되며 후보 이벤트의 순위는 계속적으로 변한다. 순위의 변화가 없을 때까지 이 작업은 반복되며, 최종적으로 올바른 이벤트를 추출할 수 있다. 자세한 수식은 다음 장에서 설명하였다.

표 1 동사에 따른 초기 패턴

동사	패턴
Inhibit	Entity inhibit Entity
Associate	Entity associate with Entity
Induce	Entity induce Entity
Bind	Entity bind (to) Entity
Activate	Entity activate Entity

3.4 이벤트 추출

이벤트 추출의 일반적인 방법은 패턴을 미리 정의한 후에 이 패턴에 의해서만 이벤트를 추출하는 것이었다. 그러나 이 패턴들을 통해 이벤트를 표현하는 모든 자연어의 문장의 특징을 규정짓기에는 한계가 있다.

이 이유로 본 논문에서는 이벤트 추출시 패턴 정보뿐만 아니라 후보 이벤트 안에 나타난 개체간 공기정보와 문법관계정보까지도 고려하였다. 패턴은 후보 이벤트 생성시 자동으로 구축되도록 함으로써 패턴 구축에 들어가는 비용을 줄일 수 있었다. 이번 장에서는 입력으로부터 이벤트와 패턴의 추출에 이르기까지의 과정을 설명한다.

3.4.1 패턴구성을 위한 문장의 인자화

품사, 개체명, 기본구 정보가 부착되어 있는 문장에서, 이벤트 추출을 위해 원시 말뭉치의 모든 어휘 정보가 필요한 것은 아니다. 즉, 이벤트 추출에 필요한 요소가 있으며 이 요소들만을 분석하여 이벤트를 추출할 수 있다.

개체명과 기본구는 하나의 덩어리로 더 이상 세세하게 분석할 필요가 없다. 동사에 대해서는 생물학 문서에서 이벤트를 나타내는 동사들을 생물학자들의 도움으로 162개를 미리 정의한 후 이 동사들에 대해서만 고려하였고 이 때 특정 동사들의 활용(Conjugation)까지도 고려한다. 전치사는 동사와 더불어 이벤트추출에 중요한 단서를 제공한다. 그렇기 때문에 특정 동사와 자주 공기하는 전치사에 대해서는 사전에 정의하여 실제 이벤트 추출 모듈에 사용하였다. 또 다른 중요한 요소로 문장부호, 접속사가 있다. 이 요소들은 특히 안긴문장과 안은문장의 복문형태, 이중 주어와 이중 목적어의 중문형태로 구성된 문장에서 많이 나타난다. 이런 겹문장에서 올바른 이벤트를 추출하기 위해서는 이 정보들의 분석이 반드시 이루어져야 한다. 반면, 부사는 이벤트를 구성하는데 있어서 영향력이 크지 않기 때문에 고려하지 않는다. 단, 'negatively'와 같은 이벤트의 의미를 바꾸는 부사에 대해서는 어휘 자질을 그대로 고려한다. 이와 같이 이벤트를 추출하기 위해 고려해야 하는 요소는 문장 내의 모든 단어가 아니며 필요한 요소만으로 문장을 재구성할 수 있게 된다. 이를 본 논문에서는 '문장의 인자화' 라 한다.

실제로 원시 문장에 대한 인자화의 예는 그림2와 같고 구성할 때 고려하는 인자들은 표2에 나타나 있다.

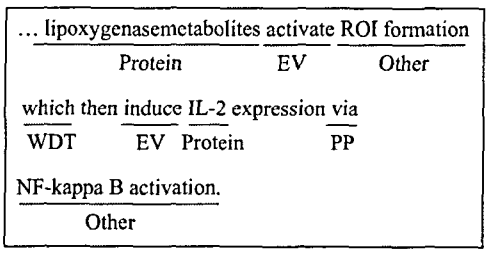

그림 2 문장의 인자화

표 2 문장을 인자화할 때 고려하는 것들

고려하는 인자	예제
개체명	Protein, Gene, RNA, DNA
이벤트 동사	activate, bind, induce
일반 동사	be, 이벤트 동사외 동사
품사 정보	관계대명사, 관계부사
기본구 정보	개체가 아닌 명사구(NP)
전치사	of, by, with, in
접속사	and, but, or
기호	(,), :, ;

3.4.2 후보 이벤트와 패턴의 생성

본 논문에서 고려하는 이벤트는 하나의 이벤트 동사에 대해 주체가 되는 개체와 객체가 되는 개체간의 이진관계로 하였다.

후보 이벤트는 문장의 인자화 단계에서 한 문장 내에 한 개의 이벤트 동사와 두 개 이상의 개체를 포함한 문장으로부터 추출한다.

선택된 문장에 대해 후보 이벤트와 이에 해당하는 패턴을 추출하는 방법은 다음의 두 단계로 이루어진다. 첫 번째 단계는 문장 내에서 이벤트 동사의 위치를 파악하여 이벤트 동사를 기준으로 앞뒤에 존재하는 개체들에 대해 모든 가능한 쌍을 구성하는 것이다. 두 번째 단계는 선택된 두 개체 사이에 존재하는 모든 인자들로 패턴을 구성

하는 것이다[표3]. 이때 수일치나 시제일치와 같은 기본적인 영어문법을 적용하여 조건을 만족하는 것들만으로 후보 이벤트의 패턴을 구성한다. 또한, "be found/shown /able/sufficient to"의 경우는 일반적으로 이 관용어구 뒤에 이벤트 동사가 출현하기 때문에 이와 같은 관용어구들은 부사와 같이 패턴 구성 시 고려대상에서 제외하였다.

그 결과, 첫 번째 단계에서 구성된 개체쌍 중 두 번째 단계에 의해 패턴이 구축된 것들만이 최종적으로 후보 이벤트가 된다.

표 3 이벤트 동사에 따른 패턴의 예

동사	패턴
bind	Entity WDT EV bind Entity
activate	Entity EV activate PP with Entity
induce	Entity EV induce Entity AND Entity
contain	Entity AND Entity EV contain Entity
phosphorylate	Entity EV phosphorylate PPto Entity

3.4.3 이벤트와 패턴의 순위 결정

구성된 후보 이벤트 중 정답이벤트를 선별하기 위해서 신뢰도를 계산하여 판단 기준으로 하였다. 이에 본 절에서는 후보 이벤트들에 대해 신뢰도 계산 방법 및 패턴 정보를 어떻게 활용했는지에 대해 설명하고자 한다.

3.4.3.1 이벤트의 신뢰도 측정

이벤트의 신뢰도는 후보 이벤트의 인자간 분석 가중치와 후보 이벤트를 표현하는 모든 패턴들의 가중치에 의해 계산되어진다. 초기 단계에서는 패턴의 가중치가 계산되어 있지 않기 때문에 초기 패턴의 가중치를 기본값으로 주고 사용하였다.

이벤트의 신뢰도계산에서 후보 이벤트의 인자간 분석 가중치 측정 방법은 수식(1)과 같다.

후보이벤트의인자간분석가중치
=α공기정보+β문법관계+γ주성분수(1)

수식(1)에서 첫 번째 척도는 이벤트 동사, 인자1, 인자2에 대해 개체간의 공기 정보와 이벤트 동사와 개체간의 공기정보이다. 공기정보는 χ^2분포를 사용하여 측정하였다[수식(2)]. 수식(2)에서 'O'는 말뭉치 전체에서 얻을 수 있는 실제 정보로써, 공기 정보를 알고자 하는 두 인자간의 공기빈도와 두 인자가 나타나지 않는 빈도의 곱으로 계산되어 지고 'E'는 실제 정보를 토대로 예측할 수 있는 빈도를 일컫는다.

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad (2)$$

두 번째와 세 번째 척도는 기본구간의 문법관계분석 결과로부터 얻을 수 있는 것이다. 두 번째 척도는 개체들과 이벤트 동사와의 문법적 의존관계를 분석하는 것으로 하나의 이벤트 후보에서 동사의 앞뒤의 개체들이 각각 주어역할과 목적어 역할을 한다면 이들은 이벤트일 가능성이 매우 높다고 본다. 세 번째 척도는 각 개체 및 기본구들의 문법적 기능을 참고하는 것으로 해당 인자가 문장 구성상 주성분인지를 파악하는 것이다. 문장 내에서 주성분일수록 이벤트에 속할 확률이 높을 것이라는 가정 때문이다.

이벤트의 인자간 분석 가중치에 해당 이벤트를 표현하는 모든 패턴 가중치까지 고려하여 이벤트의 신뢰도를 계산한다[수식(3)]. 이는 해당 후보 이벤트가 하나의 패턴으로만 표현되는 것이 아니라 여러 형태의 패턴으로 표현될 수 있기 때문에 고려된 것으로 해당 후보 이벤트를 표현하는 모든 패턴들의 가중치들의 평균을 해당 후보 이벤트 신뢰도 측정에 추가하였다.

$$Score(E) = [\text{수식(1)}] \times \frac{\sum_i Score(P_i^E)}{\sum_j P_j^E} \quad (3)$$

'E'는 후보 이벤트, 'P'는 패턴, 'P^E_i'는 특정 후보 이벤트를 표현하는 패턴을 의미한다.

3.4.3.2 패턴의 가중치 측정

이벤트를 나타내는 패턴의 가중치는 각 패턴이 후보 이벤트들을 표현하는데 사용된 빈도의 합과 패턴에 의해 생성된 후보 이벤트 중에서 신뢰도가 특정 임계치를 넘는 것들의 빈도의 합을 사용하여 측정하였다[수식(4)].

$$Score(P) = \frac{\sum_{Score(E_i) > \delta} Score(E_i)}{\sum_j E_j^P} \quad (4)$$

수식(4)에서 사용된 임계치 δ는 패턴의 가중치를 계산하는 시점의 이전 단계에서 계산되어진 이벤트들 신뢰도의 평균값으로 하였다. 'E^P_j'는 특정 패턴에 의해 생성된 후보 이벤트를 의미한다.

3.4.3.3 이벤트와 패턴의 순위

후보 이벤트에 대한 신뢰도와 그로부터 생성된 패턴들의 가중치에 의해 후보 이벤트들의 순위를 매길 수 있다. 수식(3)과 수식(4)를 보면 각 수식은 서로의 결과를 사용하고 있다. 초기 이벤트 신뢰도는 초기 패턴에 의해 계산하였고 그 결과가 패턴의 가중치 계산에 적용되었다. 이 결과를 다시 이벤트의 신뢰도 계산에 사용하는 것이다.

이와 같은 반복 수행으로 이벤트의 신뢰도와 패턴의 가중치가 갱신되고, 이에 따라 순위도 변화하게 된다. 이 반복은 이벤트 간 순위 변화가 없을 때까지 수행된다. 결과적으로 얻어진 신뢰도와 순위에 의해서 올바른 이벤트를 추출하게 된다.

3.5 실험

사용한 말뭉치는 GENIA 말뭉치로써 총 18,544의 문장으로 구성되어 있다[12]. 이 문장 중 현재 실험에 사용한 문장은 241개의 이벤트를 갖는 372개 문장이다. 이벤트 동사로 정해 놓은 동사는 총 162개로써 이들의 활용(Conjugation)까지 고려하였다. 실제로 추출한 예는 [그림 3]과 [표4]에서 보는 바와 같다.

예제에서 볼 수 있듯이 간단한 형식의 이벤트도 추출할 수 있지만 겹문장에 대해서도 이벤트 추출이 가능하다.

> Our data suggest that lipoxygenase metabolites activate ROI formation which then induce IL-2 expression via NF-kappa B activation.

그림3 예제 문장

표 4 예제문장의 이벤트 추출 결과

	Event1	Event2
Argument1	Lipoxygenase metabolites	ROI formation
Argument2	ROI formation	IL-2Effectactivateinduce
Effect	activate	induce

3.6 평가

평가 척도는 정확률, 재현율과 F-measure를 사용하였고 다음과 같이 정의된다[수식(5),(6),(7)].

$$\text{정확률} = \frac{\text{정답과 일치하는 이벤트수}}{\text{시스템이 추출한 이벤트수}} \quad (5)$$

$$\text{재현율} = \frac{\text{정답과 일치하는 이벤트수}}{\text{정답 이벤트수}} \quad (6)$$

$$\text{F-measure} = \frac{2 \times \text{정확률} \times \text{재현율}}{\text{정확률} + \text{재현율}} \quad (7)$$

241개의 이벤트에 대한 결과는 표5와 그림 4 이다.

표 5 임계치에 따른 실험 결과

	평가 임계치 (상위등수)	정확률(%)	재현율(%)	F-measure (%)
인자간 분석 가중치	177	81.36	59.75	68.90
인자간 분석 가중치 + 패턴가중치	251	83.67	87.14	85.37
	217	88.02	79.25	83.41
	213	87.79	77.59	82.38
	177	89.27	65.56	75.60

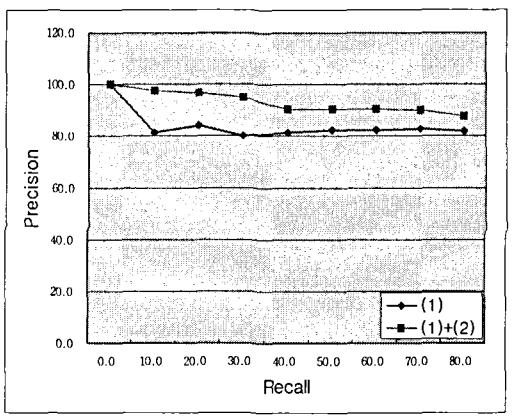

그림 4 정확률과 재현율

총 생성된 후보 이벤트의 수는 258개이고 생성된 패턴의 수는 204개였다. 표5는 평가 임계치에 따른 시스템의 정확률과 재현율의 변화를 보여주는 것으로, 평가 임계치란 생성된 후보 이벤트들에 대한 상위 순위를 일컫는다. 그림4는 재현율에 따른 정확률을 나타내는 그래프로 '(1)'은 인자간 분석 가중치만을, '(1)+(2)'는 인자간 분석 가중치와 패턴 가중치를 함께 고려한 결과이다.

분석 결과, 패턴 정보 없이 인자간 분석 가중치만으로도 이벤트 추출의 성능은 기존의 연구결과와 견줄 만 했다. 여기에 자동으로 구축된 패턴의 적용은 이벤트 추출에서 높은 정확률은 유지한 채 재현율을 향상시키는 중요한 역할을 한다는 것을 알 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. 기존 연구의 패턴 구축의 어려움을 자동 확장으로 극복하였고 이 확장된 패턴 정보와 이벤트내의 정보까지 고려하여 이벤트를 추출하였다. 기존의 연구방법에 비해 정확률은 다소 낮지만 재현율은 기존의 방법보다 높은 것을 알 수 있다. 낮은 정확률은 보다 정교한 신뢰도 측정방법으로 개선할 수 있으리라 생각된다.

지금은 실험한 이벤트수의 양이 매우 적기 때문에 앞으로 많은 실험량에 의해 보다 신뢰할 수 있는 결과를 얻어야 할 것이다.

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] Ki-joong Lee, Young-Sook Hwang, Hae-Chang Rim, "Two-Phase Biomedical NE Recognition based on SVMs", Proc.of the ACL 2003 Workshop on Natural Language Processing in Biomedicine, pp33-40, 2003.
[10] Kyung-Mi Park et al., "Grammatical Relation Analysis Using Support Vector Machine in Biotext" , Proc. of the 15th Conference of Hangul and Korean Information Processing, (To be appear), 2003.
[11] A. Blum & T. Mitchell, "Combining labeled and unlabeled data with co-training", In Proceedings of the 11th Annual Conference on Computational Learning Theory, pp224-231, 1998.
[12] GENIA Corpus 3.0p. 2003. available at http://www-tsujii.is.s.u-tokyo.ac.jp/genia/topics/Corpus/3.0/GENIA3.0p.intro.html