바이오 이벤트 추출을 위한 피쳐 개발

이석준○, 김영태, 황민국, 임수종†, 나동열
연세대학교, 컴퓨터정보통신공학부
†한국전자통신연구원
seokjun88@naver.com, wolfnamu@gmail.com, peacsid@nate.com, isj@etri.re.kr, dyra@yonsei.ac.kr

Developing Features for Bio Event Extraction

Seok-Jun Lee○, Young-Tae Kim, Min-Kook Hwang, Soo-Jong Lim†, Dong-yul Ra
Yonsei University, Computer & Telecommunications Engineering Division
†Electronics & Telecommunications Research Institute

요 약

본 논문은 바이오 문서에서의 정보추출 시스템 개발에 대한 것이다. 이 시스템의 목표는 바이오 관련 문서에서 바이오 이벤트의 발생을 탐지하고 이벤트의 타입 및 이벤트에 관여된 필수 논항을 채우는 구문요소를 인식하는 것이다. 우리는 두 개의 별도의 단계를 이용하는 시스템 구성을 사용한다. 첫 단계에서는 SVM을 사용하여 이벤트의 발생 및 이벤트의 타입을 결정한다. 두 번째 단계에서는 이벤트의 논항을 채우는 참여자를 인식하는 작업을 한다. 본 논문은 단계 1에서 사용되는 SVM의 피쳐 리스트의 개발에 대한 문제를 다룬다. 본 논문에서 제안하는 피쳐 리스트를 사용하여 좋은 성능을 가지는 첫 단계에 대한 모듈을 얻을 수 있음을 관찰하였다.

주제어: 정보추출, 이벤트추출, 바이오 이벤트, 바이오 문서, 이벤트 탐지.

1. 서론

구글이나 네이버 등에서 볼 수 있듯이 정보검색 (information retrieval; IR) 기술의 사회에 대한 기여는 막대하다. 이와 마찬가지로 정보추출 (information extraction; IE) 기술의 발전이 사회에 제공할 기여 또한 매우 클 것으로 관측된다. 그러나 IE 의 가장 큰 문제인 텍스트 마이닝 (text mining) 기술의 경우 상용화에 이르기 위해서는 아직 많은 발전을 필요로 하는 실정이다. 이러한 기술의 발전을 위해 자연어처리 연구분야에서는 여러 텍스트 마이닝 관련 학술 대회가 개최되었다.

이들 중에서도 대표적인 것으로 BioNLP 학술대회를 들 수 있다. 이미 여러 해 째 진행되고 있는 이 학술대회는 생명공학과 관련된 문서를 대상으로 하는 정보추출 시스템 개발 기술의 발전을 목표로 한다. 시스템 개발 참가팀들은 미리 공통으로 주어진 작업(shared task)을 수행하는 시스템의 개발을 위해서 미리 제공되는 학습데이타(training data)를 이용할 수 있다.

본 논문에서는 BioNLP-2009 의 ST(shared task) 1에 대한 시스템 개발을 목표로 한다. BioNLP-2009 ST-1은 바이오 문서에서 단백질과 관련된 이벤트의 추출을 목표로 한다[1]. 즉 문서에서 미리 정해진 종류의 이벤트 및 이벤트에 관여된 참여자(participant) 즉 논항(argument)을 인식하는 것을 목표로 한다. 이 작업은 바이오 정보추출의 가장 기본적인 형태로서 정보추출과 관련된 가장 핵심(core)이 되는 기술이라고 생각된다.

이 작업을 위한 시스템의 개발은 여러 가지 접근 방법이 가능하다. 본 논문에서는 다음과 같이 구분된 2 단계로 구성된 시스템을 사용하는 것을 가정한다. 즉 단계 1에서 문서에 존재하는 이벤트 발생 (trigger) 탐지 및 그것의 타입(type)을 결정한다. 그 다음 별도의 다음 단계 2에서 앞에서 탐지된 각 이벤트에 대하여 그것의 논항을 채우는 참여자를 찾아서 이벤트를 완성하도록 한다.

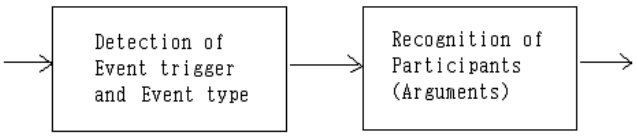

그림 1: 바이오 이벤트 추출 시스템 구성.

그러나 BioNLP-2009 ST-1에 참여한 팀들은 여러 형태의 다른 방식의 접근을 선택한 경우도 많다. 예를 들면 위의 두 단계를 별도로 구분하지 않고 하나의 단계로 하여 이벤트의 존재 및 종류의 결정을 논항의 탐지와 병행하는 기법을 사용하는 것이다.

우리는 단계 1을 위한 모듈의 개발을 위하여 다중-클래스(multiple class) SVM 을 사용하기로 하였다. 본 논문에서는 이를 위한 효과적인 피쳐(feature) 집합의 개발을 목표로 한다. 실험 결과 본 연구에서 제안하는 피쳐 집합을 사용하는 경우 높은 성능의 단계 1 모듈의 개발이 가능한 것으로 생각된다.

단계 2에 대한 모듈의 개발을 위해서 먼저 우리는 기계학습 방법을 사용하는 대신 의존트리 탐색을 통하여 이벤트의 논항을 찾는 휴리스틱 기반 탐색 기법을 시도해 보았으나 이 기법으로는 좋은 성능을 얻지 못하였다. 이 모듈에 대해서는 결국 기계학습을 포함한 다른 기법을 탐구할 필요가 있다.

2. 관련 연구

자연어처리 기술의 발전을 위하여 공유 작업(shared task) 문제를 설정하고 이에 대한 시스템의 개발 후 그 경험을 교류하는 형태의 학술대회가 많이 이용되었다. 특히 텍스트 마이닝의 경우 MUC, TREC, ACE 등은 관련 기술의 발전에 많은 기여를 한 것으로 생각된다. 바이오-텍스트 마이닝의 경우도 예외는 아니어서 TREC Genomics track, JNLPBA, LLL, BioCreative 등과 같은 학술대회가 이에 속한다. 이 중에서도 가장 대표적인 것으로는 BioNLP 를 들 수 있다[1]. 이 학술대회는 2006년 이래로 매년 열리고 있는데 바이오 정보추출 기술의 발전에 매우 중요한 역할을 한 것으로 평가된다. 이러한 학술대회에서는 shared task 에 대한 시스템의 개발에 도움이 되는 자원(resource)을 제공한다. 이의 대표적인 것은 학습에 이용될 수 있는 학습데이타이다. 대부분의 개별 연구자들은 비용과 시간상의 문제로 인해 독자적으로 이러한 학습 데이타를 구축하여 사용할 형편이 되지 못한다. 따라서 시스템 개발에 필수적인 학습데이타의 제공은 많은 연구집단으로 하여금 연구 및 개발을 가능케 하므로 기술의 발전에 매우 중요한 역할을 한다.

3. 문제정의 및 데이타

바이오 정보 추출에 대한 기본적인 기술의 개발을 목적으로 우리는 BioNLP-2009 의 문제 및 데이타를 이용한다. 이렇게 문제를 선정한 이유는 여기에서 주어진 문제를 연구함으로서 가장 기초적이고 핵심적인 이벤트 추출 기술에 대한 연구가 가능한 것으로 판단하였기 때문이다.

3.1. 문제정의

BioNLP-2009 ST(shared task) 1의 문제는 테스트 데이타로 제공된 바이오 문서 안에서 나타난 모든 이벤트에 대하여 다음과 같은 정보를 추출하여야 한다:

- 이벤트의 발생을 알리는 트리거 (스트링) 탐지(ETD),
- 이벤트의 타입 분류(ETC),
- 이벤트에 관계된 필수 논항 참여자 인식(EPR).

3.2. 데이타

시스템의 개발 및 실험을 위해 다음과 같은 데이타가 제공된다. 이와 같은 데이타는 바이오 이벤트 태깅 데이타 자원인 GENIA에서 추출한 데이타를 가공한 것이다[2].

- 훈련 데이타(training data): 시스템의 학습에 이용되는 데이타.
- 개발 데이타(development data): 시스템의 튜닝(tuning)에 이용되는 데이타.
- 테스트 데이타(test data): 시스템의 성능 측정에 사용되는 데이타.

위의 각 데이타는 GENIA 바이오 문서 집단으로 구성되며 문장에 대하여 다음과 같은 정보가 태깅(tagging) 되어 있다:

- 단백질: 단백질을 나타내는 스트링.
- 이벤트 트리거: 이벤트 발생을 알리는 스트링.
- 이벤트: 고유 번호가 부여된 이벤트.
- 이벤트 논항: 각 이벤트에 참여하는 논항(들).

훈련 데이타와 개발 데이타는 위의 4 가지 정보가 모두 태깅되어 있다. 테스트 데이타의 경우에는 단백질 정보만이 태깅되어 있다.

바이오 이벤트의 종류는 다음 표 1과 같이 9가지이다. 이 표에는 각 타입의 이벤트가 가질 수 있는 주요논항(primary arguments)이 표시되어 있다.

표 1. 추출 대상 이벤트의 종류.

타입(type)	주요논항(primary arguments)
Gene expression	Theme(Pro)
Transcription	Theme(Pro)
Protein catabolism	Theme(Pro)
Phosphorylation	Theme(Pro)
Localization	Theme(Pro)
Binding	Theme (Pro), ...
Regulation	Theme(Pro/Ev), Cause(Pro/Ev)
Positive regulation	Theme(Pro/Ev), Cause(Pro/Ev)
Negative regulation	Theme(Pro/Ev), Cause(Pro/Ev)

(Pro: protein, Ev: event)

위 표에서 괄호 속은 이러한 논항을 채울 수 있는 개체의 타입을 나타낸다. 추출 대상이 되는 9가지의 이벤트 종류는 크게 3 종류로 다음과 같이 다시 구분될 수 있다.

(1) 단순 이벤트	논항 1 개만을 가지는 이벤트들로서 그림 2 에서 처음 5가지가 이에 해당한다.
(2) 다중 이벤트	그림2 의 binding 타입 이벤트로서 1개 이상의 Theme 을 가질 수 있다.
(3) 복합 이벤트	Theme 과 Cause 논항을 한 개씩 가질 수 있으며 이 논항들은 단백질 뿐 아니라 이벤트에 의해서도 채워질 수 있다. 결국 이벤트가 이벤트를 논항으로 포함하는 재귀적인 구조가 가능하므로 매우 복잡한 구조를 가지는 이벤트의 생성이 가능하다.

제공되는 훈련, 개발, 테스트 데이타 이외에 별도로 문장분리, 품사태깅, 의존관계트리 정보도 제공된다. 개발자는 시스템의 개발에 이러한 정보를 최대한 이용할 수 있다.

학습용 데이타의 한 문장을 예로 들면 다음과 같다:

"Phosphorylation of TRAF2 inhibits binding to the CD40 cytoplasmic domain."

이 문장에 대하여 제공되는 이벤트 태깅 정보는 다음 네모 안의 내용과 같다. 여기에서 T1, T2 는 단백질을 나타내며 T15, T16, T17 은 이벤트 트리거를 나타낸다. E1, E2, E3 는 이벤트를 나타내며 각 이벤트의 논항정보도 포함한다.

T1	Protein 19 24	TRAF2
T2	Protein 49 53	CD40
T15	Phosphorylation 0 15	Phosphorylation
T16	Negative_regulation 25 33	inhibits
T17	Binding 34 41	binding
E1	Phosphorylation:T15 Theme:T1	
E2	Negative_regulation:T16 Theme:E3 Cause:E1	
E3	Binding:T17 Theme:T1 Theme2:T2	

예를 들면 E3 는 E2 의 Theme 논항으로 참여하고 있다. 이 태깅 정보는 텍스트 형태로 되어 있어 전체적인 이벤트들의 파악이 어렵다. 이를 다음과 같이 그림으로 표시하면 보다 이해하기 쉽다.

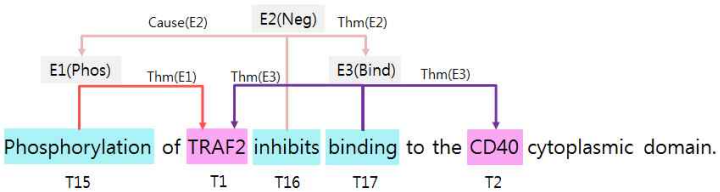

그림 2. 문장의 이벤트 태깅 정보.

4. 바이오 이벤트 시스템 개발

우리는 바이오 이벤트 추출 문제에 대한 시스템의 개발을 위하여 그림 1에 표시된 바와 같이 별도로 분리된 2 단계의 처리를 거치는 기법을 사용한다:

- 단계 1: 이벤트의 발생을 나타내는 단어(열)의 탐지 및 이벤트의 타입의 결정 (ETD-ETC),
- 단계 2: 이벤트의 주요논항을 채우는 참여자의 인식(EPR).

4.1. 이벤트 트리거-타입 탐지를 위한 피처 개발

단계 1을 이벤트 트리거 탐지(event trigger detection; ETD)와 이벤트의 타입 결정 (event type classfication; ETC)라는 별도의 두 세부 작업(sub-task)로 나누어 [3] 에서와 같이 접근할 수도 있으나, 우리는 하나의 모듈이 이 두 세부작업을 동시에 처리하는 방법을 택한다. 우리는 이 모듈의 개발을 위해 그림 3 과 같이 다중-클래스(multi-class) SVM을 사용한다.

그림 3에서 보듯이 테스트 데이타 문장의 각 단어에 대하여 이 단어가 이벤트 트리거인지 그리고 그런 경우 그 이벤트의 타입이 무엇인지를 분류해 주는 SVM 기반의 분류기를 사용한다. 이 분류기의 입력은 분류하고자 하는 현재 토큰에 대하여 생성되는 피처 리스트 (feature vector)이다. 출력은 1에서 10 사이의 정수로서 10은 트리거가 아님을 나타내며 다른 값은 이벤트의 종류를 나타낸다. 본 논문은 이 ETD/ETC SVM을 위한 효과적인 피처집합의 생성에 대하여 자세히 살펴보고자 한다.

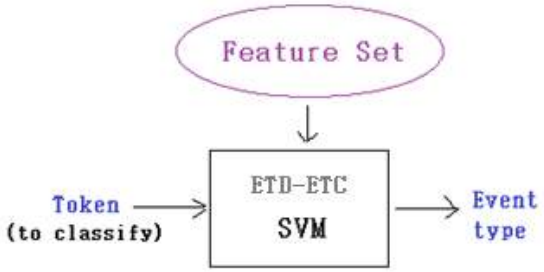

그림 3: SVM에 기반한 이벤트탐지 및 타입인식.

실제로 BioNLP-2009 에 참여한 팀들의 경우 단계 1의 성능이 별로 높지 않았다. [4, 5] 의 경우 단계 1을 위하여 CRF 모델을 사용하였다. 그러나 이 단계에 대한 성능은 F-score 43.3 에 머물고 있다. [3] 의 경우 단계 1은 이벤트 트리거의 존재 여부를 탐지하고(event trigger detection), 그 다음 이벤트의 타입을 결정하는(event type classification) 2 개의 별도의 세부단계로 구성되어 있다. 이들 세부단계는 각각 Maximum entropy 모델을 이용하여 구현되었다. 이 시스템의 성능을 보면 ETD 의 경우 F-score 68.5, ETC 의 경우 85.20 정확도(accuracy)를 가져 단계 1의 결합 성능이 58.32 F-score 를 가지는 것으로 보고되었다.

4.2. ETD-ETC SVM 피처 리스트

우리는 다음과 같이 매우 기본적인 피처 종류에서 출발하여 점차 다양한 피처를 추가함으로써 결과적으로 좋은 성능을 가능케 하는 피처 집합을 이용한다.

- BF (Basic features)

대부분의 분류 문제에서 항상 사용하는 가장 기본적인 피처집합이다. 현재 토큰 인덱스가 i 라 하자. 즉 Wi 는 현재 토큰(current token)을 나타낸다. 이를 기준으로 좌우-윈도우(Wi, Wi-1, Wi-2, Wi+1, Wi+2) 안의 각 토큰에 대하여 다음과 같이 3 가지의 피처를 생성한다.

BF	For each of Wi, Wi-1, Wi-2, Wi+1, Wi+2: prepare features of **String, Stem, POS**

- PF (Protein features)

좌우-윈도우 내의 각 토큰에 대하여 단백질 여부를 나타내는 피처이다.

PF	For each of Wi, Wi-1, Wi-2, Wi+1, Wi+2: prepare features of **Is_Protein**

• DF (Dependency features)

실험에 사용되는 모든 데이타 문장에 대한 의존 파스트리가 제공된다. 따라서 구문구조적인 정보를 이용하기 위하여 우리는 다음과 같은 피쳐를 준비한다. 이 피쳐 집합과 관련하여 주의할 점은 단백질 여부에 대한 피쳐도 생성한다.

DF	For each of LC(closest left child of Wi) RC(closest right child of Wi) LLC(2nd closest left child of Wi) RRC(2nd closest right child of Wi) Par(parent of Wi) Gpar(grandparent of Wi): prepare features of **String, Stem, POS, Is_Protein**

• CCLRC (closest child of left or right child)

현재 토큰의 좌/우 자식의 자식 중 가장 가까운 것이 단백질인지의 여부에 대한 피쳐이다.

CCLRC	For each of CCLC(closest protein child of left child of Wi), CCRC(closest protein chjild of right child of Wi): prepare features of **String, Stem, POS, Is_Protein**

• WCP (wide context protein features)

문장내의 단백질의 개수, 현재 단어 주변의 정해진 크기(현재는 좌우 각 4 단어)의 윈도우 내의 단백질의 개수를 피쳐로 한다.

WCP -w -s	 **Number of proteins in the window** **Number of proteins in the sentence**

• ET (event type of current token)

우리는 훈련 데이타로부터 이벤트 트리거로 사용된 모든 단어의 어근을 수집한다. 그리하여 각 어근에 대하여 이벤트 트리거로 가장 많이 사용된 이벤트 종류를 파악한다. 이에 기반하여 다음 피쳐를 준비한다.

ET	event class number for which current token's stem was used most

4.3. 이벤트 참여자 인식 모듈 개발

단계 2의 작업은 단계 1에서 탐지한 각 이벤트에 대하여 이의 논항이 되는 참여자를 찾는 작업(participant searching; PS)이다. 결국 이 단계에 대한 입력은 "이벤트 트리거" 및 이를 포함하는 "문장"이며, 출력은 이 이벤트의 필수 논항을 채우는 모든 "참여자"이다. 참여자는 이벤트의 타입에 따라서 단백질 또는 다른 이벤트가 될 수 있다.

단계 2의 작업에 대한 접근 방법은 여러 가지가 가능하다. 예를 들면, 단계 1과 마찬가지로 분류의 문제로 접근할 수도 있다. <이벤트트리거, 단백질>, <이벤트트리거, 이벤트> 등과 같이 문장으로부터 가능한 쌍을 구하여 이것을 "참" 또는 "거짓"의 하나로 분류하는 것이다. 이러한 접근 방법의 장점은 기계학습 기법이 제공하는 특징 중의 하나인 높은 recall을 얻을 수 있다는 점이다.

그러나 일차적으로 먼저 우리는 단계 2를 분류 문제로 접근하지 않고 의존트리에 기반한 휴리스틱 규칙을 사용하여 참여자를 탐색하는 기법을 사용하는 접근 방법을 먼저 시도해 보았다. 그 이유는 첫째로 단계 2를 시작하는 시점에서 이벤트 트리거 단어, 이 이벤트를 포함하는 문장, 이 문장의 의존트리, 이 문장 내에 존재하는 모든 단백질에 대한 정보가 사용이 가능하며, 둘째로 술어와 논항 사이에 의존관계가 존재한다는 것이 일반적으로 알려져 있기 때문이다. 즉 이벤트 추출 문제의 경우 다음과 같은 가정이 그럴듯하다고 보기 때문이다:

하나의 이벤트와 이 이벤트의 논항을 채우는 참여자는 의존관계 경로로 연결된다.

우리는 학습데이타에 대한 많은 관찰을 통하여 이 가정이 상당히 타당성이 있음을 관찰하였다. 이 가정에 기반한 단계 2의 처리는 결국 이벤트의 종류에 따라 다음과 같은 방법을 사용하기로 한다.

①단순 이벤트	하나의 Theme 논항을 가지며 이는 단백질에 의해 채워진다. 이벤트와 의존관계 경로로 연결된 단백질 하나를 탐색하는 휴리스틱을 개발한다.
②다중 이벤트	1개 또는2개의 Theme 논항을 가지는 경우만을 고려한다. 트리거 단어에 따른 논항의 개수 추정, 의존 경로로 연결된 단백질 중에서 참여자를 탐색하는 휴리스틱을 개발한다.
③복합 이벤트	Theme, Cause 두 가지의 논항이 가능하며 단백질 또는 다른 이벤트가 이를 채울 수 있다. 문장의 의존구조 트리에서 이벤트 노드 주위의 구조를 고려하여 참여자를 찾는 휴리스틱을 개발한다.

[5]의 연구에 따르면 의존관계 구문구조를 효과적으로 이용한다면 단계 2 만의 성능은 상당히 높을 것으로 추정된다. 따라서 우리의 전략은 좋은 SVM 피쳐집합의 선정을 통하여 가능하면 단계 1의 성능을 높이고자 하였다.

5. 실험 결과 및 평가

5.1. 이벤트 트리거 및 타입 탐지 (ETD-ETC) 성능측정

우리는 먼저 전체 시스템의 성능을 측정하는 대신에 단계 1 즉 이벤트 발생 탐지 및 이벤트 타입 결정 작업에 대하여 성능을 측정하여 보았다 (표 2 참조). 우리의 가정은 단계 1의 성능이 전체 시스템의 성능에 큰 영향을 미친다고 보기 때문이다. 앞에서 설명한대로 단계 1에 대한 모듈은 SVM 을 사용하였다. 그래서 이 모듈의 성능은 SVM에 사용되는 피쳐에 매우 큰 영향을 받게 된다. 앞에서 소개한 여러 피쳐 집합의 효과를 측정하기 위해 우리는 점증적인 실험을 수행하였다. 즉 맨 처음에는 가장 기본적인 피쳐집합인 BF를 사용하여 시스템을 구성하고 이에 의한 단계 1 모듈의 성능을 측정하였다 (표의 첫 번째 행). 다음 행은 앞의 시스템에 대하여 PF 피쳐를 추가하여 보다 증가된 피쳐집단으로 시스템을 구성하고 성능을 측정하였다. 이와 같은 방법으로 점진적으로 피쳐를 추가하여 가면서 성능을 측정하였다.

표 2. 단계 1에 대한 성능 측정 결과.

피쳐	Recall	Precision	F-score
BF	43.63	52.48	47.65
BF+ PF	47.76	55.37	51.28
BF+PF+ DF	46.53	62.13	53.21
BF+PF+DF+ CCLRC	52.17	61.31	56.37
BF+PF+DF+CCLRC+ WCP	50.07	68.45	57.84
BF+PF+DF+CCLRC+ WCP+ ET	59.19	64.77	61.85

위의 단계 1의 모듈에 대한 평가는 개발 데이타 (development data)를 테스트 데이타로 이용하였다. 그 이유는 배포한 테스트 데이타는 이벤트에 대한 태깅이 되어 있지 않기 때문이다. 배포된 테스트 데이타를 이용하지 못한 이유는 테스트 데이타를 이용한 시스템 평가는 전체 시스템에 대한 평가만 가능하기 때문에 단계 1만에 대한 평가는 수행할 수 없기 때문이다. 이러한 문제로 인하여 우리는 훈련 데이타를 두 부분으로 나누어 일부를 훈련에 나머지 일부를 튜닝에 사용하고, 배포된 개발 데이타를 테스트 데이타로 이용하였다.

실험결과 우리 시스템의 단계 1 모듈의 성능은 61.85의 F-score로 측정되었다. 우리 시스템과 유사하게 구분된 단계 1을 사용하는 두 개의 시스템으로 [4] 와 [5]가 있다. 이들 시스템의 경우 CRF를 사용하여 이 모듈을 개발하였다. 그들 시스템의 단계 1에 대한 성능은 43.3의 F-score로 관찰되었다. [3] 의 경우 단계 1을 두 개의 세부 단계로 나누어 구현하였는데 단계 1에 해당하는 성능이 58.32 F-score를 가지는 것으로 보고되었다. 이러한 결과에 따르면 우리 시스템의 단계 1 모듈이 다른 시스템에 비해 좋은 성능을 보임을 알 수 있다. 즉 우리가 제시하는 피쳐들을 사용하는 경우 좋은 시스템을 달성할 수 있다고 할 수 있다.

5.2. 이벤트 참여자 인식(EPR) 성능측정

우리는 단계 2 즉 EPR 에 대한 성능을 별도로 측정하지 않고 단계 1과 단계 2를 합친 전체 시스템의 성능을 측정하여 그 결과를 표 3에 나타내었다. 특히 이벤트의 종류에 따른 이벤트 추출 문제의 난이도를 알아보기 위해 각 종류에 따른 성능도 표시하였다.

표 3. 전체 시스템 성능.

	GS	Sys	GS∩ Sys	R	P	F-s
단순 이벤트	559	435	269	48.12	61.84	54.12
다중 이벤트	249	265	76	30.52	28.68	29.57
복합 이벤트	987	663	19	1.93	2.87	2.30
전체 평균	1795	1363	364	20.28	26.71	23.05

GS: 정답집합, Sys: 시스템 출력, GS∩Sys: true positive, R: recall, P: precision, F-s: F-score.

위 표를 보면 단순 이벤트의 경우는 단계 1의 성능에 별로 떨어지지 않는 전체 성능을 보이는 것으로 볼 때 EPR 모듈만의 성능은 상당히 좋은 것을 알 수 있다. 그러나 복합 이벤트의 경우 2.30 F-score의 성능을 가지는데 이는 이 종류의 거의 모든 이벤트의 추출에 실패함을 알 수 있다. 특히 복합 이벤트의 수가 전체 이벤트들에서 차지하는 비중이 55%를 차지하므로 전체 시스템의 성능을 23.05로 저하시키게 되었다.

이벤트 추출 시스템의 전체 성능 향상을 위해서는 각 이벤트 종류별로 성능 향상을 추구하여야 하지만 특히 아주 낮은 성능을 가지는 복합 이벤트와 관련하여 보다 개선된 의존트리 기반 휴리스틱의 기법을 개발하여야 할 것이다.

6. 결론

본 논문에서는 바이오분야 문서에서 바이오 이벤트의 발생 및 그 타입 그리고 이벤트의 논항을 채우는 참여자의 인식을 수행하는 시스템 개발에 대한 문제를 다룬다. 우리가 제안한 시스템은 이벤트 발생의 탐지 및 그 타입의 인식 (ETD-ETC)을 위한 별도의 단계를 가지는데 여기에서는 다중-클래스 분류에 적합한 SVM을 이용한다. 우리는 여러 가지 다양한 피쳐를 사용함으로써 높은 성능을 가지는 ETD-ETC 단계를 위한 모듈의 개발이 가능함을 보였다. 이벤트 논항의 참여자 인식 (EPR) 모듈의 경우

현재로서는 의존트리 기반 휴리스틱에 의한 탐색 기법을 사용하여 보았다. 그러나 보다 더 높은 성능을 위해서는 기계학습 기법 등 다른 기법을 시도할 필요가 있음을 알게 되었다.

참고문헌

[1] J. Kim, T. Ohta, S. Pyysalo, Y. Kano and J. Tsujii, "Overview of BioNLP' 09 Shared Task on Event Extraction," Proceedings of the Workshop on BioNLP: Shared Task, pages 1-9, 2009.

[2] J. Kim, T. Ohta and J. Tsujii, "Corpus annotation for mining biomedical events from literature," BMC Bioinformatics 9:10, 2008.

[3] H. Lee, H. Cho, M. Kim, J. Lee, G. Hong and H. Rim, "A Multi-Phase Approach to Biomedical Event Extraction," Proceedings of the Workshop on BioNLP: Shared Task, pages 107-110, 2009.

[4] G. Georgiev, K. Ganchev, V. Momtchev, D. Peychev, P. Nakov and A. Roberts, "Tunable Domain-Independent Event Extraction in the MIRA Framework," Proceedings of the Workshop on BioNLP: Shared Task, pages 95-98, 2009.

[5] F. Sarafraz, J. Eales, R. Mohammadi, J. Dickerson, D. Robertson and G. Nenadic, "Biomedical Event Detection using Rules, Conditional Random Fields and Parse Tree Distances," Proceedings of the Workshop on BioNLP: Shared Task, pages 115-118, 2009.