# 시맨틱 구문 트리 커널 기반의 단백질 간 상호작용 식별

정창후○, 전홍우, 최윤수, 최성필
한국과학기술정보연구원
{chjeong○, hw.chun, armian, spchoi}@kisti.re.kr

# Protein-Protein Interaction Recognition based on Semantic Parse Tree Kernel

Chang-Hoo Jeong○, Hong-Woo Chun, Yun-Soo Choi, Sung-Pil Choi
Korea Institute of Science and Technology Information

**요 약**

본 논문에서는 단백질 간 상호작용 자동 식별을 위해서 구문 트리 커널을 확장한 시맨틱 구문 트리 커널을 제안한다. 기존의 구문 트리 커널은 구문 트리의 단말 노드를 구성하는 개별 어휘에 대해서 단순하게 외형적 비교를 수행하기 때문에 실제 의미적으로는 유사한 두 구문 트리의 커널 수치가 상대적으로 낮아져서 단백질 간 상호작용 식별의 성능이 떨어지는 문제점이 발생한다. 이를 극복하기 위해서 두 구문 트리의 구문적 유사도(syntactic similarity)와 어휘 의미적 유사도(lexical semantic similarity)를 동시에 효과적으로 계산하여 이를 결합하는 새로운 커널을 고안하였다. 그리고 제안된 시맨틱 구문 트리 커널을 활용하여 단백질 간 상호작용 식별 성능을 향상시킬 수 있음을 실험을 통하여 보여주었다.

## 1. 서 론

단백질 간 상호작용(Protein-Protein Interaction, PPI) 자동 추출은 텍스트 내에 표현된 단서 어휘 및 구문 구조 혹은 기타 부가 자질들을 활용하여 그 텍스트 내에 출현한 다수의 단백질들 간의 상호작용에 관한 정보를 자동으로 추출하는 기술이며, PPI의 존재 여부를 판단하는 식별(PPI Identification, PPII)과 PPI의 구체적인 상호작용의 종류를 결정하는 분류(PPI Classification, PPIC)로 구성된다.

본 논문에서는 PPI 자동 추출 중에서도 PPI 식별에 관해서 다루는데, PPI 식별을 위해서 구문 트리 커널을 확장한 시맨틱 구문 트리 커널을 제안한다. 기존의 구문 트리 커널은 구문 트리의 단말 노드를 구성하는 개별 어휘에 대해서 단순하게 외형적 비교를 수행한다. 따라서 실제 의미적으로는 유사한 두 구문 트리의 커널 수치가 상대적으로 낮아지는 문제점이 발생하여 PPI 자동 추출의 전체적인 성능에 악영향을 줄 수 있다. 이를 극복하기 위해서 본 논문에서는 두 구문 트리의 구문적 유사도(syntactic similarity)와 어휘 의미적 유사도(lexical semantic similarity)를 동시에 효과적으로 계산하여 이를 결합하는 새로운 커널을 고안하였다.

## 2. 관련 연구

PPI 자동 추출에 관한 연구는 그 중요성으로 인하여 매우 활발하게 진행되어 왔으며 다양한 기법들이 소개되어왔다. Zhou and He (2008)은 최근까지 연구된 PPI 추출 기법을 전산언어학 기반 기법, 규칙 기반 기법, 그리고 기계학습 및 통계적 기법의 세 종류로 구분하여 설명하고 있다[1].

우선 전산언어학 기반 기법에서는 PPI를 표현할 수 있는 대표적인 문장 구조를 분석하여 이를 문법으로 구성한다. 이러한 요소 문법들은 상호작용 추출을 위한 특화된 언어분석 시스템(품사 부착기, 기저구 인식기, 구문 분석기 등)의 기반 문법으로 활용된다. PPI 자동 추출 기술의 특성에서 볼 때, 문장에 대한 심층 분석은 필수적이며 그 분석정도에 따라 shallow parsing 기반 기법[2, 3]과 full parsing 기반 기법[4, 5]으로 나눌 수 있다.

두 번째로 규칙 기반 기법은 상호작용 표현의 단서가 될 수 있는 어휘적 패턴 집합을 수작업으로 정의하고 이를 기반으로 문장에서 이들 패턴과 일치하는 부분을 찾는다. 이 범주에 속하는 방법의 하나로서 Blaschke et al. (1996)은 상호작용 단서 어휘집합을 수집하고 이를 기반으로 어휘적 규칙(lexical rules)을 고안하여 PPI에

적용하였다. 그리고 문장 내에서 발견한 어휘적 규칙에 대한 신뢰도를 자동으로 계산하여 이를 추출된 PPI의 신뢰도로 활용하였다[6].

마지막으로 기계학습 및 통계적 기법은 가장 최근에 도래한 기법으로서, 지도학습 혹은 반지도 학습 기반의 기계학습 모델을 적용하여 미리 수작업으로 구성된 학습 집합을 기반으로 관계 및 상호작용을 표현하는 핵심 단서인 자질 집합을 자동으로 추출하고 이를 실제 시스템에 적용한다. 확장성 및 효율성 측면에서 가장 높은 성능을 나타내고 있으며, 현재도 활발하게 연구가 진행되고 있다[7, 8, 9].

## 3. 시맨틱 구문 트리 커널

### 3.1 구문 트리 커널의 문제점

합성곱 구문 트리 커널의 기본적인 개념은 구문 트리를 요소 하부 트리로 분리하고 이들 하부 트리를 벡터 공간의 개별 축(axis)으로 전사시킴으로써 M개의 하부 트리에 대해서 M차원의 벡터 공간을 구성하는 것이다. 이때 개별 구문 트리는 벡터공간의 특정 벡터로 전사된다. 벡터 공간으로 전사된 구문 트리 집합 쌍은 그들 간의 내적을 계산함으로써 유사도를 측정할 수 있으며, 이 내적 값이 바로 구문 트리 커널의 출력이다. 이러한 장점에도 불구하고, 구문 트리 커널은 두 개의 구문 트리를 비교함에 있어서 각 단말 노드를 구성하는 개별 어휘에 대한 단순 비교로 인해 실질적으로는 매우 유사한 두 구문 트리의 유사성을 반영하지 못하는 결과를 초래하였다. 비록 매우 단순한 예제이지만 그림 1에서 이와 같은 현상을 구체적으로 보여준다.

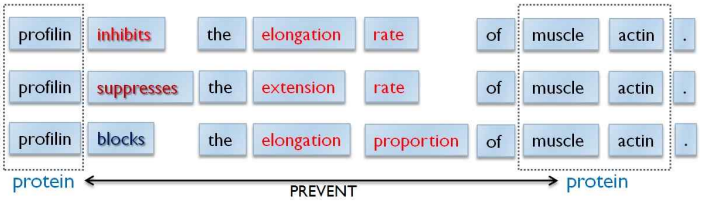


그림 1 동일한 의미의 상이한 표현에 따른 구문 트리 커널의 한계점

그림 1의 세 문장은 모두 동일한 의미를 표현하고 있으며, PPI의 관점에서도 모두 "profilin"과 "muscle actin"을 "PREVENT" 관계로 표현하고 있다. 각 문장의 구문 구조도 모두 동일하므로 개별 문장 쌍에 대한 커널 값(유사도)은 매우 높게 나타나야 한다. 그러나 이 상황에서 Moschitti (2006)가 개발한 알고리즘을 그대로 적용한다면, 단말 노드를 구성하는 개별 어휘들에 대한 단순 외형적 비교로 인해서 우리가 예상하는 수치보다 낮은 커널 값을 도출하게 된다. 따라서 위 그림에서 보듯이 "inhibit", "suppress", "block" 등의 동사들과 "elongation", "extension", "rate", "proportion"과 같은 명사들에 대한 의미 분석을 통해서 가능한 모든 어휘들에 대한 개념화(conceptualization)가 이루어지면 위와 같은 문제를 해결하고 더 나아가서 높은 성능의 PPI 자동 추출 시스템을 구성할 수 있다.

### 3.2 시맨틱 구문 트리 커널

본 논문에서 제안하는 확장된 구문 트리 커널의 기본적인 개념은 기존의 구문 트리 커널에서 강조되었던 두 문장 간의 구문적 유사도뿐만 아니라, 개별 어휘들의 개념화를 통해 얻어진 의미적 유사도를 추가적으로 활용하는 것이다. 다시 말해서, 두 구문 트리의 구문적 유사성과 의미적 유사성을 동시에 계산하여 이를 결합하는 시맨틱 구문 트리 커널을 새롭게 고안하였다. 두 구문 트리의 의미적 유사도 계산은 개별 구문 트리를 구성하는 어휘들에 대해서 문맥 기반의 개념 정보를 생성하고 이것들을 의미적으로 비교함으로써 수행된다. 문맥 기반 어휘 개념은 구문 트리의 단말 노드를 구성하는 어휘들에 대한 중의성 해소(Word Sense Disambiguation, WSD)를 통해 이루어지며 이를 위해서 워드넷(WordNet) 기반의 어휘 중의성 해소 알고리즘을 사용한다. 결론적으로 두 문장의 유사도는 구문적 유사성과 어휘 의미적 유사성을 동시에 측정하여 결합함으로써 계산된다. 따라서 커널 함수는 식 1과 같이 구성될 수 있다.

$$K_{sem}(T_1,T_2,\lambda,\sigma,\alpha)=\sum_{n_1\in N_{T_1}}\sum_{n_2\in N_{T_2}}\Delta_{sem}(n_1,n_2,\lambda,\sigma,\alpha)$$
$$=sim_{lex}(T_1,T_2,\alpha)+sim_{syn}(T_1,T_2,\lambda,\sigma)$$

식 1 시맨틱 구문 트리 커널 계산식

식 1은 시맨틱 구문 트리 커널이 (1) 어휘 의미적 유사도(lexical semantic similarity, $sim_{lex}(T_1,T_2,\alpha)$)와 (2) 구문적 유사도(syntactic similarity, $sim_{syn}(T_1,T_2,\lambda,\sigma)$)를 합산하여 커널 값을 계산하고 있음을 나타낸다. 어휘 의미적 유사도는 구문 트리의 단말 노드만을 대상으로 계산되며, 구문적 유사도 계산은 그 외의 노드를 대상으로 이루어진다. 본 논문에서 제안하는 어휘 의미적 유사도에 대한 세부적인 계산 방법은 식 2와 같다.

$$sim_{lex}(T_1,T_2,\alpha) \equiv sim(W_{T_1},W_{T_2},\alpha) \quad \text{(식 2-1)}$$

$$\approx \sum_{w_1 \in W_{T_1}, c_1 \subset W_{T_1}^{w_1}} \left( \sum_{w_2 \in W_{T_2}, c_2 \subset W_{T_2}^{w_2}} sim(w_1,w_2,c_1,c_2,\alpha) \right) \quad \text{(식 2-2)}$$

$$\approx \sum_{w_1 \in W_{T_1}, c_1 \subset W_{T_1}^{w_1}} \left( \sum_{w_2 \in W_{T_2}, c_2 \subset W_{T_2}^{w_2}} I(concept(w_1,c_1,\alpha), concept(w_2,c_2,\alpha)) \right) \quad \text{(식 2-3)}$$

$$\approx \sum_{w_1 \in W_{T_1}, c_1 \subset W_{T_1}^{w_1}} \left( \sum_{w_2 \in W_{T_2}, c_2 \subset W_{T_2}^{w_2}} I(synset(w_1,c_1,\alpha), synset(w_2,c_2,\alpha)) \right) \quad \text{(식 2-4)}$$

$$= \sum_{w_1 \in W_{T_1}, c_1 \subset W_{T_1}^{w_1}} \left( \sum_{w_2 \in W_{T_2}, c_2 \subset W_{T_2}^{w_2}} I(pos(synset(w_1,c_1,\alpha)), pos(synset(w_2,c_2,\alpha))) \right) \quad \text{(식 2-5)}$$

식 2 어휘 의미적 유사도 계산식

$W_T$는 구문 트리 $T$를 구성하는 단어 집합이며, $W_T^{w_i}$는 이 중에서 $w_i$ 주변의 문맥 단어 집합이다. 식 2-1에서의 $W_{T_1}$와 $W_{T_2}$의 유사도 계산식 $sim(W_{T_1}, W_{T_2}, \alpha)$은 식 2-2와 같이 문장 구성 단어 $w$와 주변 문맥 정보 $c$를 이용한 유사도 계산식 $sim(w_1, w_2, c_1, c_2, \alpha)$으로 대체된다. 단어 간 유사도를 수치화하여 정확하게 계산하기는 어려우므로, 식 2-3에서는 개별 단어를 개념화함으로써 두 단어의 유사도 계산을 단순화하였다. *concept* 함수는 $w$가 가지고 있는 여러 의미들을 주변 문맥 단어 집합 $c$를 이용하여 정확한 하나의 의미로 식별하는 역할을 수행한다. 식 2-3에서 $I(A,B)$는 $A$와 $B$가 동일하면 1을 반환하고 아니면 0을 반환하는데, 결론적으로 식 2-3은 두 문장을 교차 비교하면서 동일한 개념을 나타내는 단어 쌍의 개수를 계산한다. 본 논문에서는 특정 단어의 의미 식별을 그 단어에 대한 워드넷에서의 synset 집합 중에서 주변 문맥 단어들과 가장 일치하는 synset을 식별하는 것으로 정의한다. 따라서 식 2-4에서는 *concept* 함수가 워드넷에서의 가장 유사한 synset을 반환하는 *synset* 함수로 대체된다. 마지막으로 워드넷에서 특정 synset은 전체 파일 내에서의 위치 정보로 식별될 수 있으므로 이들의 위치 정보를 반환하는 *pos* 함수를 도입하여 실질적인 비교 작업을 수행하였다. 모든 식에 포함된 추상화 수준 지정 인자 $\alpha$는 개별 어휘의 의미적 개념들의 추상화 수준을 지정하기 위해서 사용한다.

한편, 두 문장에 대한 구문적 유사도를 나타내는 $sim_{syn}(T_1, T_2, \lambda, \sigma)$은 Moschitti (2006)가 제안한 것과 동일하게 다음과 같이 표현될 수 있다.

$$sim_{syn}(T_1,T_2,\lambda,\sigma) \equiv \sum_{n_1 \in N_{T_1}, n_1 \notin L_{T_1}} \left( \sum_{n_2 \in N_{T_2}, n_2 \notin L_{T_2}} \Delta(n_1,n_2,\lambda,\sigma) \right)$$

식 3 구문적 유사도 계산식

$N_{T_i}$는 구문 트리 $T_i$의 전체 노드 집합이고, $L_{T_i}$는 $T_i$의 모든 단말 노드 집합을 나타낸다. 그리고 $\Delta(n_1,n_2,\lambda,\sigma)$은 Moschitti (2006)가 제안한 특정 노드 $n_1$과 $n_2$를 최상위 노드로 가지는 트리의 공통 하부 트리 개수 계산 알고리즘으로서, $\lambda$는 트리 커널 소멸 인자이고 $\sigma$는 트리 커널 계산 방법 지정 인자이다. 식 3에 대한 자세한 내용은 [10, 11]에서 상세히 기술하고 있다. 식 2와 식 3에 의거하여, 시맨틱 구문 트리 커널은 최종적으로 다음과 같이 표현될 수 있다.

$$K_{sem}(T_1,T_2,\lambda,\sigma,\alpha) \equiv \sum_{w_1 \in W_{T_1}, c_1 \subset W_{T_1}^{w_1}} \left( \sum_{w_2 \in W_{T_2}, c_2 \subset W_{T_2}^{w_2}} I(pos(synset(w_1,c_1,\alpha)), pos(synset(w_2,c_2,\alpha))) \right) + \sum_{n_1 \in N_{T_1}, n_1 \notin L_{T_1}} \left( \sum_{n_2 \in N_{T_2}, n_2 \notin L_{T_2}} \Delta(n_1,n_2,\lambda,\sigma) \right)$$

식 4 시맨틱 구문 트리 커널 계산식 구체화

식 4는 시맨틱 구문 트리 커널의 세부적인 계산 방법을 나타낸다. 식에서도 알 수 있듯이 우변의 첫째 항은 어휘 의미적 유사도를, 둘째 항은 구문적 유사도를 나타내며, 최종 커널 값은 이 두 유사도를 더한 값이다.

## 4. 실험 및 분석

### 4.1 실험 대상 말뭉치

본 논문에서 제안한 시맨틱 구문 트리 커널의 객관적인 성능 수준을 파악하기 위해서, 기존의 연구 결과에서 사용된 다양한 말뭉치 기반의 단백질 간 상호 작용 식별에 관한 실험을 수행하였다. 통상적으로 "Five PPI Corpora"[12]라고 불리는 이 말뭉치 집합은 Almed[13], BioInfer[14], HPRD50[15], IEPA[16], 그리고 LLL[17]을 단일화된 XML 형식으로 변환해 놓은 컬렉션으로서, 단백질 간 상호작용 식별 기법의 평가 컬렉션으로 활용되고 있다.

### 4.2 성능 측정 기준

본 논문에서 사용한 성능 측정 기준은 거시평균 기반 F-점수(macro-averaged F-score)와 미시평균 기반 F-점수(micro-averaged F-score)이다. 우선 거시평균 기반 방법은 m개의 클래스에 대해서 개별적으로 정확도와 재현율이 합산된 F-점수를 계산하고 이를 m으로 나눈 평균을 계산하는 방법이다. 이에 반해 미시평균 기반 방법은 전체 검증 데이터를 기반으로 옳게 분류된 데이터와 그르게 분류된 데이터를 누산하고 이를 기반으로 F-점수를 계산하는 방법이다. 거시평균 기반 방법은 학습 모델의 모든 클래스에 대한 분류 능력을 전체적으로 살

펴볼 수 있는 장점이 있으나, 학습 집합의 클래스별 분포가 고르지 않을 경우 상대적으로 낮은 성능측정 결과를 가져온다. 미시평균 기반 방법은 학습 모델의 특정 클래스에 대한 분류 능력이 상대적으로 낮을 경우, 이를 제대로 반영하지 못한다는 단점이 있다. 학습 집합의 클래스별 분포가 차이가 나는 경우나 학습 모델의 특정 클래스 예측 성능이 낮게 나타날 경우에는 두 평가 방법의 수치가 상당히 차이나는 경우도 있다.

### 4.3 단백질 간 상호작용 식별 실험

본 논문에서 수행한 실험은 총 세 가지의 매개변수가 필요하며, 이 값의 지정 방법에 따라서 성능의 차이가 발생할 수 있다. 세부적인 내용은 표 1과 같다.

표 1 실험 대상 시스템 설정 방법 및 개수

| 매개변수 | 설명 (details) | 범위 (range) | | 설정개수 |
|---|---|---|---|---|
| $\lambda$ | 구문트리커널 소멸인자 | 0.1 ~ 1.0 (단위: 0.1) | | 10 |
| $C$ | SVM 정규화 매개변수 | 1.0 ~ 7.0 (단위: 1.0) | | 7 |
| $\alpha$ | 시맨틱 구문트리 커널에서의 어휘 개념에 대한 추상화 수준 지정 인자 (generalization level) | **0** | node concept 그대로 사용 | 4 |
| | | **1** | 현재 node concept의 부모를 사용 | |
| | | **2** | 현재 node concept의 조부모를 사용 | |
| | | **N** | 기존 구문 트리 커널 | |
| 총 시스템 수 | | | | **280** |

구문 트리 커널 소멸인자는 최소 0.1에서 최대 1.0까지 0.1 단위로 10개를 지정하였다. 그리고 SVM의 정규화 인자는 1.0에서 7.0까지 7가지로 한정하였다. 마지막으로 어휘 개념 추상화 수준 지정 인자는 총 4가지로 설정하였다. 따라서 이들 세 가지 매개변수를 모두 적용하면 총 280가지의 설정이 도출된다. 본 논문에서는 이 설정 집합 각각에 대해서 10겹 교차평가를 수행하여 성능을 측정하였다.

표 2 각 말뭉치별 최고 성능을 나타내는 설정 값 및 성능 세부 정보

| Collection | Tree Kernels | Abstraction Level ($\alpha$) | DF ($\lambda$) | Regularization Factor ($C$) | mi-F1 | Precision | Recall | ma-F1 |
|---|---|---|---|---|---|---|---|---|
| **AImed** | SPTK | 1 | 0.5 | 7.0 | **89.33** | 84.86 | 77.45 | 80.99 |
| **BioInfer** | SPTK | 0 | 0.5 | 5.0 | 89.00 | 87.22 | **84.81** | 86.00 |
| **IEPA** | PTK | - | 0.4 | 7.0 | 79.17 | 78.51 | 78.30 | 78.41 |
| **HPRD50** | SPTK/PTK | 0/1/2 | 0.7 | 6.0 | 85.22 | 84.74 | 83.41 | 84.07 |
| **LLL** | SPTK | 1 | 0.4 | 4.0 | 88.48 | **88.64** | 88.47 | **88.55** |

표 2는 총 280개의 시스템 중에서 최고 성능을 보이는 설정 및 상세 성능 수치를 5개의 말뭉치별로 보여준다. 총 5개의 말뭉치를 대상으로 한 실험에서 IEPA를 제외하고 시맨틱 구문 트리 커널(SPTK)이 기존의 일반 구문 트리 커널(PTK)보다 좀 더 좋은 성능을 나타내고 있다. HPRD50은 일반 구문 트리 커널과 시맨틱 구문 트리 커널의 성능이 동일하였다. IEPA를 제외하고는 거시평균 기반 F-점수 기준으로 대부분 80.0 이상을 나타내고 있다.

표 3 ma-F1 기준 상위 20개 시스템에서의 설정 분포

| Collections | PTK | SPTK | | | SPTK (total) | Coverage rate |
|---|---|---|---|---|---|---|
| | | α = 0 | α = 1 | α = 2 | | |
| **AImed** | 7 | 4 | 5 | 4 | 13 | 65% |
| **BioInfer** | 3 | 7 | 5 | 5 | 17 | **85%** |
| **IEPA** | 12 | 4 | 3 | 1 | 8 | 40% |
| **HPRD50** | 6 | 4 | 4 | 6 | 14 | 70% |
| **LLL** | 4 | 5 | 5 | 6 | 16 | 80% |

표 3은 거시평균 기반 F-점수 기준으로 $\alpha$ 값을 다양하게 변화시키면서 실험한 결과 중에서 상위 20등까지의 시스템을 선정하여 구문 트리 커널과 시맨틱 구문 트리 커널의 분포를 분석한 자료이다. IEPA를 제외한 나머지 자료에서 대부분 시맨틱 구문 트리 커널의 출현 횟수가 많았으며, BioInfer 말뭉치에 대해서는 20개의 시스템 중에서 총 17개가 시맨틱 구문 트리 커널이었다. 따라서 시맨틱 구문 트리 커널 기반 기법이 전반적으로 높은 성능을 유지하고 있음을 알 수 있다.

## 5. 결론 및 향후 연구

본 논문에서는 두 문장의 구문적 유사도에만 관심을 두었던 기존의 구문 트리 커널을 확장하여 어휘 의미적 유사도를 함께 고려할 수 있는 시맨틱 구문 트리 커널을 새롭게 개발하였다. 그리고 제안된 시맨틱 구문 트리 커널을 활용하여 단백질 간 상호작용 식별 성능을 향상시킬 수 있음을 실험을 통하여 보여주었다.

향후 연구로 문맥기반 어휘 중의성 해소 시스템의 성능 향상이 시급하다. 특정 문장 내의 특정 어휘에 대한 의미를 파악함에 있어서 문맥 정보의 중요도는 매우 높다. 따라서 보다 심층적인 언어 분석을 통해서 대상 문맥을 확대시킴으로써 정확한 어휘 개념 추출이 이루어져야 한다. 이를 위해서 추가적인 언어 자원이나 언어 분석 시스템의 도입이 필요하다. 이와 더불어 기존 커널과 본 논문의 시맨틱 구문 트리 커널을 결합한 혼합 커널(composite kernel) 개발을 생각해 볼 수 있다. [18]에서 제안한 의존 그래프 커널 등과 같은 기존 커널과의 밀결합을 통해서 새로운 커널을 구성할 수 있을 것이다.

## 참 고 문 헌

[1] Zhou D. and He Y., "Extracting interactions between proteins from the literature," Journal of Biomedical Informatics, Vol.41, pp.393-407, 2008.

[2] Sekimizu T., Park H. S., and Tsujii J., "Identifying the interaction between genes and gene products based on frequently seen verbs in MEDLINE abstracts," Workshop on genome informatics, Vol.9, pp.62-71, 1998.

[3] Gondy L., Hsinchun C., and Martinez Jesse D., "A shallow parser based on closed-class words to capture relations in biomedical text," Journal of Biomedical Informatics, Vol.36, No.3, pp.145-158, 2003.

[4] Temkin Joshua M., and Gilder Mark R., "Extraction of protein interaction information from unstructured text using a context-free grammar," Bioinformatics, Vol.19, No.16, pp.2046-2053, 2003.

[5] Nikolai D., Anton Y., Sergei E., Svetalana N., Alexander N., and Ilya M., "Extracting human protein interactions from MEDLINE using a full-sentence parser," Bioinformatics, Vol.20, No.5, pp.604-611, 2004.

[6] Blaschke C., Andrade M., Ouzounis C., and Valencia A., "Automatic extraction of biological information from scientific text: protein-protein interactions," Proceedings of the 7th International Conference on Intelligent Systems for Molecular Biology, pp.60-67, 1999.

[7] Andrade Miguel A., and Valencia A., "Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families," Bioinformatics, Vol.14, No.7, pp.600-607, 1998.

[8] Marcotte Edward M., Xenarios I., and Eisenberg D., "Mining literature for protein-protein interactions," Bioinformatics, Vol.17, No.4, pp.359-363, 2001.

[9] Craven M. and Kumlien J., "Constructing biological knowledge bases by extracting information from text sources," Proceedings of the 7th International Conference on Intelligent Systems for Molecular Biology, pp.77-86, 1999.

[10] Collins M. and Duffy N., "Convolution Kernels for Natural Language," NIPS-2001, 2001.

[11] Moschitti A., "Making tree kernels practical for natural language learning," Proceedings of EACL'06, 2006.

[12] Pyysalo S., Airola A., Heimonen J., Björne J., Ginter F., and Salakoski T., "Comparative analysis of five protein-protein interaction corpora," BMC Bioinformatics, Vol.9, No.S6, 2008.

[13] Bunescu R., Ge R., Kate R., Marcotte E., Mooney R., Ramani, A., and Wong, Y., "Comparative Experiments on Learning Information Extractors for Proteins and their Interactions," Artif. Intell. Med., Summarization and Information Extraction from Medical Documents, Vol.33, pp.139-155, 2005.

[14] Pyysalo S., Ginter F., Heimonen J., Bjorne J., Boberg J., Jarvinen J., and Salakoski T., "BioInfer: a corpus for information extraction in the biomedical domain," BMC Bioinformatics, Vol.8, No.50, 2007.

[15] Fundel K., Küffner R., and Zimmer R., "RelEx - Relation extraction using dependency parse trees," Bioinformatics, Vol.23, pp.365-371, 2007.

[16] Ding J., Berleant D., Nettleton D., and Wurtele E., "Mining MEDLINE: abstracts, sentences, or phrases?" Proceedings of PSB'02, pp.326-337, 2002.

[17] Nedellec C., "Learning language in logic - genic interaction extraction challenge," Proceedings of LLL'05, pp.31-37, 2005.

[18] Airola A., Pyysalo S., Bjorne J., Pahikkala T., Ginter F., and Salakoski T., "All-paths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning," BMC Bioinformatics, Vol.9, No.S2, 2008.