술어-논항 구조의 패턴 유사도를 결합한 혼합 커널 기반 관계 추출

Relation Extraction based on Composite Kernel combining Pattern Similarity of Predicate-Argument Structure

정 창 후* 최 성 필** 최 윤 수*** 송 사 광**** 전 홍 우*****
Chang-Hoo Jeong Sung-Pil Choi Yun-Soo Choi Sa-Kwang Song Hong-Woo Chun

요 약

문헌에 존재하는 핵심개체 간의 관계를 자동으로 추출할 때 다양한 형태의 문서 분석 결과를 활용할 수 있다. 본 논문에서는 기존에 개발되어 비교적 높은 성능을 보여준 합성곱 구문 트리 커널의 구절 구조 유사성 정보와 두 개체 사이의 유의미한 연관관계를 표현해주는 술어-논항 구조 패턴의 유사성 정보를 동시에 활용하는 혼합 커널을 제안한다. 구문적 구조를 이용하는 기존의 합성곱 구문 트리 커널에 술어와 논항 간의 의미적 구조를 활용하는 술어-논항 구조 패턴 유사도 커널을 결합하여 상호보완적인 혼합 커널을 구성하였고, 다양한 테스트컬렉션 기반의 실험을 통하여 개발된 커널의 성능을 측정하였다. 실험 결과 구절 구조 정보를 이용하는 합성곱 구문 트리 커널만을 단독으로 사용했을 때보다 술어-논항 구조의 패턴 정보를 결합한 혼합 커널을 사용했을 때에 더 좋은 성능을 보이는 것을 확인할 수 있었다. 또한 기존의 시스템보다 우수한 성능을 보이는 것도 함께 확인할 수 있었다.

ABSTRACT

Lots of valuable textual information is used to extract relations between named entities from literature. Composite kernel approach is proposed in this paper. The composite kernel approach calculates similarities based on the following information: (1) Phrase structure in convolution parse tree kernel that has shown encouraging results. (2) Predicate-argument structure patterns. In other words, the approach deals with syntactic structure as well as semantic structure using a reciprocal method. The proposed approach was evaluated using various types of test collections and it showed the better performance compared with those of previous approach using only information from syntactic structures. In addition, it showed the better performance than those of the state of the art approach.

☞ keyword : 합성곱 구문 트리 커널(Convolution Parse Tree Kernel), 술어-논항 구조 패턴(Predicate-Argument Structure Pattern), 관계 추출(Relation Extraction), 혼합 커널(Composite Kernel)

1. 서 론

비정형적인 텍스트 내에서 중요하고 연관성 있는 정보를 식별하는 정보 추출은 자연어 처리 및 텍스트 마이닝 분야에서 핵심적인 영역으로 인식되고 있다. 이러한 정보 추출 기술을 구성하는 요소 기술로서 (1) 개체명 인식(Named-Entity Recognition), (2) 관계 추출(Relation Extraction), (3) 대용어 참조 해소(Co-reference Resolution) 등이 있는데[1], 이 중에서 문서 내에 존재하는 중요한 개체 간의 관계를 자동으로 추출하는 관계 추출은 정보 추출 중에서 핵심적인 작업으로 꼽히면서도 가장 어려운 작업으로 알려져 있다[1-3].

* 정 회 원 : 한국과학기술정보연구원 선임연구원
chjeong@kisti.re.kr
** 정 회 원 : 한국과학기술정보연구원 선임연구원
spchoi@kisti.re.kr
*** 정 회 원 : 한국과학기술정보연구원 선임연구원
armian@kisti.re.kr
**** 정 회 원 : 한국과학기술정보연구원 선임연구원
esmallj@kisti.re.kr
***** 정 회 원 : 한국과학기술정보연구원 선임연구원
hw.chun@kisti.re.kr(교신저자)
[2011/07/01 투고 - 2011/07/05 심사 - 2011/08/08 심사완료]

관계 추출 문제를 해결하기 위한 방법론으로 관계 추출에 특화된 커널 함수를 새롭게 구성하여 이를 기반으로 지지벡터기계(Support Vector Machines)에 적용하는 커널기반 방법의 효과가 주목을 받고 있다. 관계 추출 분야에서의 커널기반 방법의 특징은 한 문장에 존재하는 두 개체 간의 관계를 가장 잘 표현하는 특징을 선별해서 유사도를 가장 효과적으로 계산하는 커널을 구성하면 성능이 매우 높게 나타난다는 것이다. 개체 간의 관계를 추출할 때 문서 내에 존재하는 다양한 특징을 활용할 수 있는데, 본 논문에서는 두 개체 간의 관계를 핵심적으로 표현하고 있는 술어-논항 구조 패턴을 추출하여 이를 기존의 합성곱 구문 트리 커널(Convolution Parse Tree Kernel) 기법과 결합한 혼합 커널 기반의 관계 추출 방법을 제안한다.

본 논문의 구성은 다음과 같다. 우선 2장에서 관계 추출과 관련한 선행 연구에 대해서 살펴본다. 이어서 3장에서는 혼합 커널 기반 관계 추출 기법에 대해서 살펴보고, 이를 구성하는 합성곱 구문 트리 커널과 술어-논항 구조 패턴 유사도 커널에 대해서 좀 더 상세히 설명한다. 4장에서는 본 논문에서 제시한 시스템의 성능을 평가하고, 평가 결과에 대한 분석을 제시한다. 마지막으로 5장에서 결론과 향후 연구 방향을 논의한다.

2. 관련 연구

지도학습 기반 관계 추출(Supervised Relation Extraction)은 1997년도에 개최된 MUC-7(Message Understanding Conference 7)에서 처음으로 도입된 '템플릿 기반 관계 추출(Template Relation Extraction)' 태스크에서 본격적으로 기계학습 기반의 관계 추출을 위한 학습 집합을 제공함으로써 이 분야 연구의 단초를 제공하였다.

그 이후로 많은 관계 추출 기법들이 개발되었으며, 이를 처리 기법에 따라 분류하면 크게 (1) 규칙 기반 방법(rule-based methods), (2) 자질 기반 방법(feature-based methods), 그리고 (3) 커널 기반 방법(kernel-based methods)으로 구분된다.

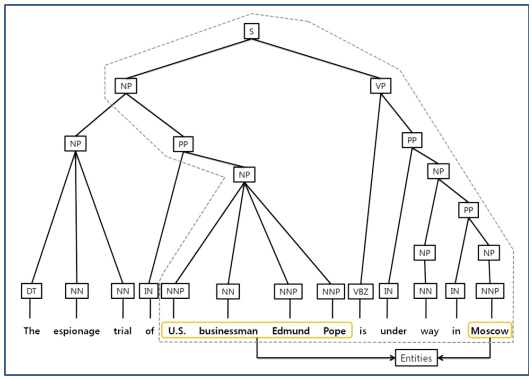
자질 기반 방법으로서 Kambhatla (2004)는 최초로 최대 엔트로피 모델(Maximum Entropy Model)을 기반으로 다양한 형태의 어휘적, 구문적, 의미적 자질들을 이용하여 관계 추출을 시도하였다[4]. 이를 기반으로 GuoDong et al. (2005)는 지지벡터기계를 활용하여 더 확장되고 세분화된 자질 정보를 관계 추출에 적용하였다[5]. 이와 유사하게 Zhao et al. (2005)는 모든 세부 자질을 종류별로 구분하고 이를 개별적인 선형 커널로 구성하여 최종적으로 혼합 커널로 결합하는 기법을 제안하였다[6]. 이 방법은 커널 함수를 직접 고안하여 적용하였다는 점에서 커널 기반 기법으로 분류될 수도 있으나, 커널의 구조가 단순하고 대부분 자질 벡터로 변환될 수 있는 점에 근거하여 자질 기반 방법으로 분류하였다. 기본적으로 위의 논문들 모두 관계 추출을 위한 자질 선정이나 구성 방법에 준거하여 자질 공학적 시도에 국한하여 접근하였으며, 관계 인스턴스의 구문 구조에 대한 적용은 매우 제한적으로 이루어졌다.

커널 기반 기법의 단초는 Zelenko (2003)에서 제시하였다. 최초로 두 개의 구문 분석 트리에 대한 유사도를 재귀적으로 측정하는 연속 부분 트리 커널(contiguous subtree kernel)과 희소 부분 트리 커널(sparse subtree kernel)의 두 가지 구문 트리 커널을 고안하고, 이를 두 가지 이진 관계에 적용하여 매우 높은 성능을 보였다[7]. 이 연구를 기반으로 Culotta et al. (2004)는 의존 구문 트리(dependency parse tree)의 유사도를 측정할 수 있는 커널을 개발하였으며, 최초로 ACE 컬렉션을 대상으로 실험하였으나 그 성능은 비교적 낮았다[2]. 또한 Bunescu et al. (2005)는 [2]의 결과를 확장하여 의존 구문 트리를 부분 트리로 분할하고, 문장 내의 의존 관계 경로를 대상으로 커널 함수를 구성하여 [2]에서보다 더 나은 결과를 얻었다[1].

한편 Zhang et al. (2006)은 Collins and Duffy (2001)에서 새롭게 고안한 합성곱 구문 트리 커널을 기반으로 다양한 구조적 자질 정보와 기존의 개체 자질 정보를 결합한 혼합 커널(composite kernel)을 개발하였다[8]. 또한 GuoDong et al. (2007)은 [8]

에서 제안한 구문 트리 커널의 가지치기 기법 및 커널 계산 기법을 확장하여 개체 쌍 주변 문맥까지도 포괄하는 새로운 트리 커널을 개발하였다[9].

본 논문에서는 기존에 연구되어 비교적 높은 성능을 나타낸 합성곱 구문 트리 커널에 술어-논항 구조 패턴 유사도 커널을 결합한 혼합 커널을 제안한다. 구절 구조의 유사도를 측정하는 합성곱 구문 트리 커널과 술어와 논항 간의 의미적 연결고리를 활용하는 술어-논항 구조 패턴 유사도 커널을 결합하여 보다 성능 좋은 혼합 커널을 구성할 수 있음을 보인다.



(그림 1) 전체 구문 트리에서 가지치기된 문맥 구문 트리

3. 혼합 커널 기반 관계 추출

3.1 합성곱 구문 트리 커널

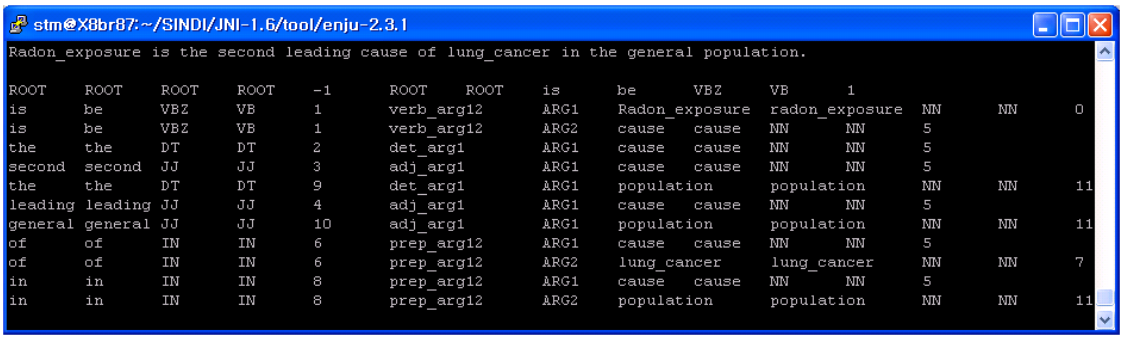
합성곱 구문 트리 커널(Convolution Parse Tree Kernel, 이하 트리 커널)의 기본적인 개념은 구문 트리를 요소 하부 트리로 분리하고 이들 하부 트리를 벡터 공간의 개별 축으로 전사시킴으로써 M개의 하부 트리에 대해서 M차원의 벡터 공간을 구성하는 것이다. 이때 개별 구문 트리는 벡터 공간의 특정 벡터로 전사된다. 벡터 공간으로 전사된 구문 트리 집합 쌍은 그들 간의 내적을 계산함으로써 유사도를 측정할 수 있으며, 이 내적 값을 이용하여 관계 추출을 수행할 수 있다. 이때 관계 추출을 위한 커널 함수의 유사도 측정 성능을 향상시키기 위해서 불필요한 문맥 정보들을 제외시키는 트리 가지치기 작업을 수행해야 하는데, 다양한 트리 가지치기 방법들 중에서 PT(Path-enclosed Tree) 방법이 가장 우수하다고 알려져 있다[10]. 따라서 본 논문에서도 PT 방법을 사용하여 두 개체를 연결하는 최소 경로 내에 포함된 부분 트리를 사용하도록 한다.

문장의 전체 구문 트리에서 두 개체 간의 관계 추정을 위한 자질로 활용하기 위해서 가지치기된 문맥 구문 트리의 예는 (그림 1)과 같다.

트리 커널은 하부 트리 분리 방법에 따라 Vishwanathan and Smola (2003)이 [11]에서 제안한 부분트리 커널(SubTree Kernel)과 Collins and Duffy (2001)가 [12]에서 고안한 부분집합트리 커널(SubSet Tree Kernel)로 나뉜다. 부분트리 기법은 트리 내에서 특정 노드의 모든 자식 노드로 구성된 부분 트리를 구성하는 것이다. 따라서 모든 부분트리는 말단 자식 노드로서 전체 트리의 잎 노드를 가져야 하며, 구문 생성 규칙에 위배되지 말아야 한다. 이에 반해서 부분집합트리 기법은 부분트리 기법보다 더 일반화된 방법으로서, 특정 부분트리가 반드시 전체 트리의 잎 노드를 가질 필요는 없다. 다시 말해서, 구문 생성 규칙에 위배되지만 않는다면, 특정 노드에서 출발하여 그 노드의 자식 노드 중 일부분을 포함할 수 있으며, 부분트리 커널 기법보다 훨씬 많은 부분트리를 생성한다. Moschitti (2006)는 [13]에서 이들 두 가지 커널을 빠르게 계산할 수 있는 알고리즘을 개발하고, 이를 의미역 결정(Semantic Role Labeling)에 활용하여 괄목할 만한 성능을 보여주었다. 또한 Moschitti (2006)에 의하면, 부분트리 커널은 부분집합트리 커널에 비해서 성능이 매우 저조하게 나타났다[13]. 따라서 본 논문에서도 구문 트리 분리 방법은 부분집합트리를 이용하였다.

구문 트리의 유사도를 측정하기 위해서 입력 구문 트리 $\overrightarrow{x_{pt}}$는 다음과 같이 함수 ϕ에 의해서 새로운 자질 공간 Φ로 전사된다.

$$\phi : \overrightarrow{x_{pt}} \in X \mapsto \phi\left(\overrightarrow{x_{pt}}\right) \in \Phi \subseteq R^N \qquad \text{식 (1)}$$



(그림 2) HPSG 파서의 문장 분석 결과

식 (1)에서 사상된 자질 공간은 *N*-차원의 유클리드 공간이며, 자질 공간 내에서의 $\phi(\overrightarrow{x_{pt}})$은 다음과 같이 구성 요소 트리의 출현 빈도 벡터로 표현된다.

$$\phi(\overrightarrow{x_{pt}}) = (f_1(\overrightarrow{x_{pt}}), f_2(\overrightarrow{x_{pt}}), ..., f_N(\overrightarrow{x_{pt}})) \quad \text{식 (2)}$$

$f_i(T)$ = the number of $subtree_i \in S$, appearing in T

S = a set of all the unique subtrees of the entire tree set.

식 (2)에서 함수 $f_i(T)$는 구문 트리 T 내에 존재하는 i 번째 구성 요소 트리의 출현 빈도를 계산한다. 따라서 $\phi(\overrightarrow{x_{pt}})$는 그 내부 구조에 따라 *N*-차원의 희소 벡터(sparse vector)로 표현될 수 있으며, 이들 간의 유사도, 즉 커널 값은 다음과 같이 내적을 통해서 계산할 수 있다.

$$K_{pt}(\overrightarrow{x_{pt}}, \overrightarrow{x'_{pt}}) = \langle \phi(\overrightarrow{x_{pt}}), \phi(\overrightarrow{x'_{pt}}) \rangle = \sum_{i=1}^{N} [f_i(\overrightarrow{x_{pt}}) \cdot f_i(\overrightarrow{x'_{pt}})] \quad \text{식 (3)}$$

그러나 특정 입력 구문 트리 집합 내에 존재하는 모든 구성 요소 트리를 추출하고, 이를 이용하여 개별 입력 구문 트리에 대해서 벡터를 구성하는 작업은 상당히 비효율적이다. 따라서 [12]에서는 $K_{pt}(\overrightarrow{x_{pt}}, \overrightarrow{x'_{pt}})$를 직접 계산하는 재귀적 방법을 고안해 냄으로써 처리 속도를 향상시켰다.

3.2 술어-논항 구조 패턴 유사도 커널

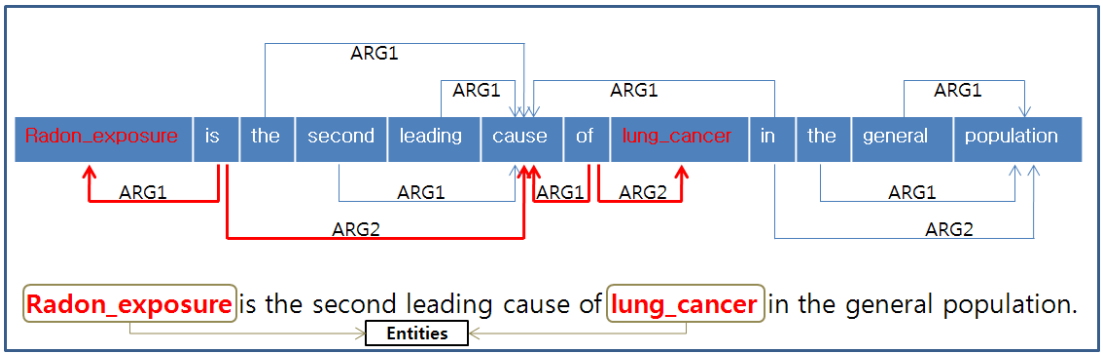
술어-논항 구조는 술어와 논항 관계를 이용하여 문장 내에 존재하는 각 단어 간의 유의미한 연관관계를 표현하는 구조이다. 그리고 술어-논항 구조 패턴은 문장을 구성하는 모든 단어에 대한 술어-논항 관계 그래프에서 중요하게 지정된 두 개체를 연결하는 최소 집합의 술어-논항으로 구성된 순서 열을 의미한다. 이러한 특성 때문에 술어-논항 구조 패턴은 문장 내에서 상호작용하는 두 개체 간의 연관관계를 표현해주는 중요한 단서 정보가 된다. 따라서 한 개체로부터 시작해서 다른 개체로까지의 의미적 연결고리를 제공해주는 술어-논항 구조 패턴을 이용하여 관계 추출을 수행할 수 있다.

본 논문에서는 술어-논항 구조 패턴을 추출하기 위해서 HPSG* 파서를 이용하였다. CFG**를 사용하는 전통적인 파서와 달리 HPSG를 사용하는 파서는 효과적으로 문장의 구문적/의미적 구조를 분석하여 술어-논항 관계를 제공한다. 따라서 사용자는 파싱 결과로부터 직접적으로 문장에 있는 단어들 사이의 의미적 연관관계를 파악할 수 있다.

HPSG 파서를 이용한 문장 분석 결과의 예는 (그림 2)와 같다.

* Head-driven Phrase Structure Grammar, http://en.wikipedia.org/wiki/HPSG

** Context Free Grammar, http://en.wikipedia.org/wiki/Context-free_grammar



(그림 3) 술어-논항 관계 그래프

(표 1) HPSG 파서 분석 결과의 각 열에 대한 설명

열 번호	상세 설명
1	술어
2	술어의 기본형
3	술어의 품사
4	술어의 기본형의 품사
5	문장에서 술어의 위치
6	술어의 종류
7	술어와 논항 사이의 관계 레이블
8	논항
9	논항의 기본형
10	논항의 품사
11	논항의 기본형의 품사
12	문장에서 논항의 위치

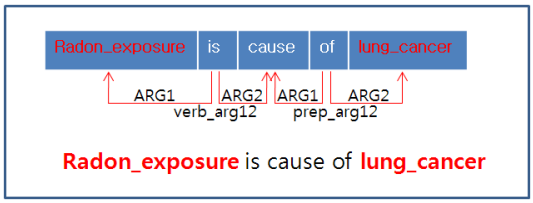
(그림 2)에서 보는 바와 같이 HPSG 파서는 문장을 입력으로 받아서 문장을 구성하는 각 단어의 술어-논항 관계를 분석하여 제공한다. 행으로 나열된 술어-논항 분석 결과의 각 필드에 대한 설명은 (표 1)과 같다.

(표 1)에서 설명한 내용을 바탕으로 "Radon_exposure is the second leading cause of lung_cancer in the general population." 문장에 대한 분석 결과인 (그림 2)의 2번째 행과 3번째 행을 설명하면, 우선 단어 'is'는 동사로서 논항 1과 2를 갖는데 그 중 논항 1은 명사인 단어 'radon_exposure'를 지칭하고 논항 2는 또 다른 명사인 단어 'cause'를 지칭한다는 사실을 나타낸다. 분석 결과의 1번째 행은 단순히 문장의 기본 술어(root predicate)를 표현하는 것이고, 4번째 행부터는 2번째와 3번째 행을 해석한 것과 같은 방식으로 해석하면 된다.

HPSG 파서에서 제공된 결과를 이용하여 각 단어의 술어-논항 관계 그래프를 그리면 (그림 3)과 같이 표현된다.

(그림 3)에서 실제적으로 문장 내에 존재하는 두 개체 간의 유의미한 관계를 표현하는 술어-논항 구조만을 추출하여 패턴을 구성하면 (그림 4)와 같다. 화살표의 연결은 한 개체로부터 상호작용하는 다른 개체로까지의 술어-논항 관계를 추적할 수 있다는 것을 의미한다. 따라서 'radon_exposure'와 'lung_cancer' 사이의 관계를 추적해보면 'is cause of'와 같은 중요한 패턴을 기반으로 관계가 형성되어 있음을 알 수 있다. 다시 한 번 말하지만, 이러한 패턴은 두 개체 간의 상호작용을 식별하는데 중요한 자질로 사용될 수 있다.

결과적으로 개체 1과 개체 2의 관계는 두 개체를 유의미한 관계로 연결해주는 술어-논항 구조 패턴에 의하여 식별될 수 있다. 술어-논항 구조 패턴을 자질로 활용하기 위해서 술어-논항 구조, 즉 (그림 4)에서 화살표로 연결되는 술어와 논항, 술어의 종류, 그리고 술어와 논항 사이의 관계 레이블을 이용하여 벡터 값을 생성하였다. 그리고 이 벡터 값을 SVM의 내장 커널 중 하나인 RBF(Radial Basis



(그림 4) 술어-논항 구조 패턴

1) ENTITY1 recognizes and activates ENTITY2.
2) ENTITY2 activated by ENTITY1 are not well characterized.
3) The herpesvirus encodes a functional ENTITY1 that activates human ENTITY2.
4) ENTITY1 can functionally cooperate to synergistically activate ENTITY2.
5) The ENTITY1 plays key roles by activating ENTITY2.

(그림 5) "Entity1 activate Entity2"를 표현하는 서로 다른 문장 구조

Function) 커널을 이용하여 술어-논항 구조 패턴의 유사도를 측정하였다.

3.3 혼합 커널

문장 내에 존재하는 두 개체 간의 구절 구조 정보를 이용하여 유사도를 계산하는 트리 커널과 두 개체 사이의 유의미한 연관관계를 표현하는 술어-논항 구조 패턴을 기반으로 유사도를 계산하는 커널을 선형 관계로 결합하여 혼합 커널을 구성하였다. 트리 커널은 그 자체만으로도 좋은 성능을 발휘하기 때문에 이전의 많은 연구에서 활용되어 왔지만, 동일한 의미를 반영하는 문장의 다양한 변형에 대해서는 유사도 측정에 잡음이 발생할 수밖에 없다. 하지만 술어-논항 구조 패턴의 경우에는 능동형, 수동형, to 부정사, that 절과 같은 다양한 변형이 발생하더라도 두 개체 간의 연관관계를 표현해 주는 패턴을 정규화해서 일관된 형태로 기술할 수 있다. (그림 5)는 'Entity1'이 'Entity2'를 'activate'시킨다는 의미를 서로 다른 문장 구조로 표현한 예이다. (그림 5)의 예문들은 구절 구조를 이용하는 트리 커널에서는 상이한 유사도를 나타내지만 술어-논항 구조 패턴을 이용할 경우에는 모두 동일한 유사도를 나타낸다.

그렇기 때문에 이 두 가지 방법을 결합하면 기존의 트리 커널 기반의 방법보다 더 나은 성능을 얻을 수 있다.

$$K_{composite} = (1-\tau)K_{tree}(I_1, I_2) + \tau \times K_{pas}(I_1, I_2)$$

(수식 1) 혼합 커널 구성

(수식 1)에서 $K_{tree}(I_1,I_2)$는 인스턴스 I_1과 I_2 간의 구절 구조 유사도를 계산하는 트리 커널이고, $K_{pas}(I_1, I_2)$는 인스턴스 I_1과 I_2 간의 술어-논항 구조 패턴 유사도를 계산하는 커널 함수이다. τ는 두 커널 사이의 가중치 조절 역할을 수행한다.

본 논문에서는 트리 커널을 빠르게 계산하기 위하여 [13]에서 개발한 트리 커널 도구를 활용하였고, 지지벡터기계 학습을 위해서는 LIBSVM*을 활용하였다. 그리고 구절 구조 트리 생성을 위해서 Charniak parser**를 이용하였고, 술어-논항 구조 분석을 위해서 Enju parser***를 이용하였다.

4. 실험 및 분석

본 장에서는 논문에서 제안한 술어-논항 구조의 패턴 유사도를 결합한 혼합 커널의 성능을 파악하기 위해서 다양한 테스트컬렉션 기반의 실험을 수행하고 결과를 분석한다. 먼저 기존의 트리 커널만을 사용한 방법과 술어-논항 구조의 패턴 유사도를 결합한 혼합 커널을 사용한 방법의 성능에 대해서 비교 실험을 수행한다. 첫 번째 실험을 통해서 술어-논항 구조의 패턴 유사도 정보를 결합하여 혼합 커널을 구성하는 것이 관계 추출에 유용한 작업임을 보인다. 다음으로 최근에 발표된 시스템들과의

* http://www.csie.ntu.edu.tw/~cjlin/libsvm/
** http://www.cs.brown.edu/~ec/#software
*** http://www-tsujii.is.s.u-tokyo.ac.jp/enju/

성능 비교 실험을 수행한다. 두 번째 실험을 통해서 본 논문에서 제안한 방법론과 기존 시스템과의 객관적인 성능 비교 평가를 수행할 수 있다.

실험에 사용된 테스트컬렉션에 대해서 살펴보면, 첫 번째 실험에서는 녹색기술문헌에 존재하는 PLOT 간의 구체적인 상호작용의 종류를 판별하는 성능 평가에 사용되는 KREC 2010 테스트컬렉션을 이용하였다. 그리고 두 번째 실험에서는 바이오 분야의 단백질 간 상호작용 식별 실험에 대표적으로 사용되는 Five PPI Corpora라고 불리는 테스트컬렉션을 이용하였다.

본 논문에서 사용한 성능 측정 기준은 거시 평균 기반 F-점수(macro-averaged F-score)와 미시 평균 기반 F-점수(micro-averaged F-score)이다. 우선 거시 평균 기반 방법은 m개의 클래스에 대해서 개별적으로 정확율과 재현율이 합산된 F-점수를 계산하고, 이를 m으로 나눈 평균을 계산하는 방법이다. 이에 반해 미시 평균 기반 방법은 전체 검증 데이터를 기반으로 옳게 분류된 데이터와 그르게 분류된 데이터를 누산하고 이를 기반으로 F-점수를 계산하는 방법이다. 전자는 학습 모델의 모든 클래스에 대한 분류 능력을 전체적으로 살펴볼 수 있는 장점이 있으나, 학습 집합의 클래스별 분포가 고르지 않을 경우 상대적으로 낮은 성능측정 결과를 가져온다. 미시 평균 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 수치 차이가 상당한 경우도 있다. 본 논문에서는 10겹 교차평가(10-fold cross validation)를 수행하여 각 성능을 측정하였다.

4.1 녹색기술문헌에 존재하는 PLOT 간의 상호작용 추출 실험

본 절에서는 술어-논항 구조의 패턴 유사도를 활용하는 방법이 기존의 트리 커널 방법과 결합됐을 때 얼마만큼의 성능 향상 효과를 발휘하는지를 알아보기 위한 실험을 수행한다.

4.1.1 실험 대상 테스트컬렉션

과학기술 문헌에 존재하는 PLOT 간 연관관계 추출 성능 평가를 위해서 자체적으로 구축한 관계 추출 테스트컬렉션 KREC 2010을 활용하였다. PLOT간 연관관계 추출 대상은 과학기술문헌 중에서도 최근 들어 활발히 언급되고 있는 녹색기술 관련 분야로 한정하였다. 그리고 실제 문헌은 과학기술 뉴스*와 NDSL**에서 보유하고 있는 해외학술지에서 선정하였다. KREC 2010 구축 과정에 대해서 좀 더 구체적으로 살펴보면, 우선 과학기술 뉴스 데이터는 녹색기술 분야 중에서도 2000년도 이후의 문서를 대상으로 문서크기가 상위 80% 이상에 해당되는 것들만을 연도별로 임의로 선정하여 전체 11,185건을 수집하였다. 그리고 해외학술지 데이터는 SCI급, 인용지수, 초록크기 등 여러 가지 요소들을 고려하여 수집하였다. 그 기준으로는 첫째, 해외학술지 중에서 인용지수(impact factor)를 기준으로 상위 50종을 우선 선별하였다. 둘째, 동일한 종에서 개별 초록의 크기가 평균초록 크기의 90% 이상인 문서를 선정하였다. 셋째, 발행연도가 2000년 이후인 최신 문서를 선정하였다. 넷째, 선정된 종에서 종별로 각 25%의 문서를 선정하여 최종적으로 10,310건의 문서를 수집하였다. 선정된 문서는 미리 정의된 연관관계 집합을 참조하여 관계태깅 작업을 수행하였다. 실제 테스트컬렉션에 존재하는 연관관계의 수는 39개이지만, 본 실험에서는 비슷한 성질의 연관관계를 통합하여 (표 2)와 같이 총 7개의 연관관계로 재구성하여 실험을 진행하였다.

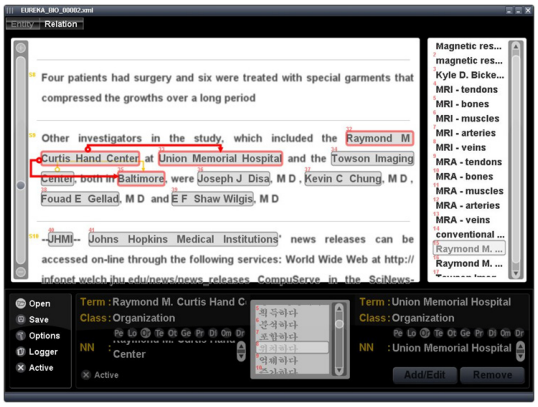
테스트컬렉션 구축은 전문가 2인에 의해 수행되었고, 서로 교차 검토하여 오류를 최소화하였다. 테스트컬렉션 구축 시 발생되는 철자오류 및 태깅오류 등을 방지하고 작업 속도를 높이기 위하여, (그림 6)과 같은 테스트컬렉션 구축 도구를 자체적으로 개발하여 사용하였다.

* http://www.eurekalert.org/

** http://www.ndsl.kr/index.do

(표 2) 실험대상 관계 종류

관계 종류	의미
relate	관계있다
change	변경하다
produce	생산하다
own	소유하다
connect	연결하다
analyze	분석하다
cause	야기하다



(그림 6) 테스트컬렉션 구축 도구

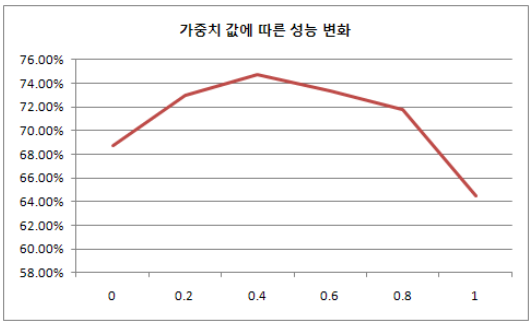
(표 3) KREC 2010 통계

문서	문장	핵심개체	연관관계
1,090	14,341	22,125	2,441

테스트컬렉션 구축 도구는 문장 분리 및 합병, 핵심개체 지정 및 취소, 핵심개체 추천, 연관관계 태깅, 분류코드 지정, 오류 검증 등의 기능을 제공하고, 완성된 문서를 테스트컬렉션의 DTD에 맞는 XML문서로 저장한다. (표 3)은 본 연구에서 1차적으로 구축한 테스트컬렉션의 통계 정보이다.

4.1.2 실험 결과 및 분석

본 절에서는 앞의 4.1.1에서 소개한 KREC 2010 테스트컬렉션을 기반으로 수행한 PLOT 간 연관관계 자동 분류에 대한 성능 평가 결과를 보인다. 혼합 커널의 보다 정확한 성능 비교를 위해서 우선 일반 구문 트리 커널과 술어-논항 구조의 패턴 유사도 커널을 각각 단독으로 사용한 경우에 대해서 살펴본다. 그리고 나서 최종적으로 두 방법을 결합한 상태로 사용한 경우의 성능 측정 결과를 비교해 본다.



(그림 7) τ 값에 따른 성능 변화 그래프

우선 본 실험에서는 두 커널 사이의 가중치 변수 τ 값에 따라 성능 평가 결과가 달라지는데, (그림 7)에서와 같이 최적의 τ 값은 0.4로 측정되었다. τ 값이 증가함에 따라 전체적인 성능이 좋아지다가 0.4를 넘어서면서 전체 성능이 나빠지는 것을 확인할 수 있다. 따라서 술어-논항 구조의 유사도를 활용한 방법론의 기여도는 40% 정도로 제한하도록 한다.

(표 4)는 PLOT 간 연관관계 추출 실험에 대한 성능 평가 결과를 보여준다. (표 4)에서 보이는 것처럼 트리 커널과 술어-논항 구조 패턴 유사도 커널을 단독으로 수행했을 때는 트리 커널의 성능이 술어-논항 구조 패턴 유사도 커널보다 더 좋은 것을 확인할 수 있다. 하지만 트리 커널 단독으로 사용하는 것보다는 술어-논항 구조 패턴 유사도 커널을 결합하여 혼합 커널을 구성하였을 때에 더 나은 성능을 보이는 것을 확인할 수 있다. 따라서 기존의 트리 커널은 술어-논항 구조 패턴 유사도 커널과 결합하여 더 나은 성능을 발휘한다는 사실을 알 수 있다.

(표 4)에서 미시 평균 기반 F-점수(mi-F1)보다 거

(표 4) PLOT 간 연관관계 추출 성능

커널 종류	mi-F1(%)*	ma-F1(%)**
술어-논항 구조 패턴 유사도 커널	64.60	33.69
구문 트리 커널	68.78	38.09
혼합 커널	74.72	42.33

시 평균 기반 F-점수(ma-F1)의 성능이 낮게 나타나는 이유는 테스트컬렉션에 사용된 7가지 관계의 인스턴스의 분포 때문이다. 7개의 연관관계 중에서 인스턴스의 개수가 100개 미만인 연관관계가 3개 존재하는데, 이 3개의 연관관계는 전체 학습 인스턴스의 단지 6.3%(2,441 중에서 152개)만을 차지하고 있다. 다시 말해서 나머지 4개의 관계가 학습 집합의 93.7%를 차지한다는 사실로 그 편중현상이 매우 심함을 알 수 있다. 따라서 관계 추출의 정답과 오답만을 검사하는 미시 평균 기반 F-점수와는 달리 거시 평균 기반 F-점수는 각 관계 별 정답과 오답을 따로 검사한 후에 그것의 평균 값을 성능으로 취하기 때문에 관계별 인스턴스 개수의 분포가 고르지 못한 경우에는 성능이 낮게 나오는 경향이 있다. 이는 관계 인스턴스가 적은 연관관계의 경우에 학습 집합의 부족 현상이 발생하기 때문이다.

4.2 단백질 간 상호작용 식별 실험

본 절에서는 술어-논항 구조의 패턴 유사도를 활용하는 혼합 커널의 보다 객관적인 성능 비교를 위해서 비교적 선행 연구가 많이 수행되어진 바이오 분야의 단백질 간 상호작용 식별 실험을 수행하여 기존 시스템과 비교를 수행한다. 본 논문에서 제안하는 방법론은 분야에 의존적이지 않기 때문에 테스트컬렉션만 존재하면 다양한 분야에 적용해볼 수 있는 장점이 있다.

4.2.1 실험 대상 테스트컬렉션

단백질 상호작용 식별 실험은 [14]에서 구성한 5가지의 PPI(Protein-Protein Interaction) 관련 테스트컬렉션을 대상으로 수행하였다. 통상적으로 Five PPI Corpora***라고 불리는 이 테스트컬렉션 집합은 AIMed[15], BioInfer[16], HPRD50[17], IEPA[18] 그리고 LLL[19]을 단일화된 XML 형식으로 변환해 놓은 컬렉션으로서, 현재 단백질 간 상호작용 추출 기법의 준거 평가 컬렉션으로 활용되고 있다.

(표 5)는 [14]에서 구성한 Five PPI Corpora에 포함된 개별 컬렉션의 규모와 상호작용 포함 문장 및 불포함 문장에 대한 통계 정보이다.

특정 문장에 2개 이상의 단백질 이름이 출현하고 그것들 간의 상호작용 관계가 설정되어 있으면, 단일 문장에 대해서도 여러 개의 상호작용 포함 문장이 구성된다. 또한 문장 내에 단백질 이름이 존재하더라도 상호작용 관계가 설정되어 있지 않다면 상호작용 포함 문장도 불포함 문장으로 동시에 설정될 수 있다. 이를 기반으로 단백질 간 상호작용 추출은 개별 인스턴스(상호작용 포함/불포함 문장)에 대한 이진 분류 작업으로 규정할 수 있다.

(그림 8)은 Five PPI Corpora에 포함된 BioInfer 테

(표 5) Five PPI Corpora 규모 및 내용

테스트컬렉션	AIMed	BioInfer	HPRD50	IEPA	LLL
문장 개수	1,955	1,100	145	486	77
단백질 간 상호작용 포함 문장 (Positive instance)	1,000	2,534	163	335	164
단백질 간 상호작용 불포함 문장 (Negative instance)	4,834	7,132	270	482	166

* micro-averaged F-score

** macro-averaged F-score

*** http://mars.cs.utu.fi/PPICorpora/eval-standard.html

```
〈sentence id="BioInfer.d0.s0" origId="2" text="_____ inhibits _____ signaling by
                    preventing formation of a _____*_____*DNA complex."〉
        〈entity charOffset="88-100" id="BioInfer.d0.s0.e0" origId="e.2.2" type="Individual_protein" /〉
        〈entity charOffset="0-12" id="BioInfer.d0.s0.e1" origId="e.2.3" type="Individual_protein" /〉
        〈entity charOffset="23-34" id="BioInfer.d0.s0.e2" origId="e.2.4" type="Individual_protein" /〉
        〈entity charOffset="75-86" id="BioInfer.d0.s0.e3" origId="e.2.5" type="Individual_protein" /〉
        〈pair e1="BioInfer.d0.s0.e0" e2="BioInfer.d0.s0.e1" id="BioInfer.d0.s0.p0" interaction="True" /〉
        〈pair e1="BioInfer.d0.s0.e0" e2="BioInfer.d0.s0.e2" id="BioInfer.d0.s0.p1" interaction="True" /〉
        〈pair e1="BioInfer.d0.s0.e0" e2="BioInfer.d0.s0.e3" id="BioInfer.d0.s0.p2" interaction="True" /〉
        〈pair e1="BioInfer.d0.s0.e1" e2="BioInfer.d0.s0.e2" id="BioInfer.d0.s0.p3" interaction="True" /〉
        〈pair e1="BioInfer.d0.s0.e1" e2="BioInfer.d0.s0.e3" id="BioInfer.d0.s0.p4" interaction="True" /〉
        〈pair e1="BioInfer.d0.s0.e2" e2="BioInfer.d0.s0.e3" id="BioInfer.d0.s0.p5" interaction="True" /〉
〈/sentence〉
```

(그림 8) Five PPI Corpora 내에서의 BioInfer 테스트컬렉션 첫 번째 문장

스트컬렉션 내에 존재하는 첫 번째 인스턴스를 보여준다. 단백질 간의 알려진 상호작용의 비정상적 적용을 방지하기 위해서 문장 내의 모든 단백질 이름은 블라인드 처리가 되어 있음을 알 수 있다. 또한 총 4개의 단백질 명이 존재하며, 이들 간의 상호작용 쌍은 총 6가지이다. 결론적으로 위의 문장에서는 총 6개의 단백질 간 상호작용 포함 문장이 구성될 수 있으며, 이들 각각은 동일한 문장을 공유하게 된다.

4.2.2. 실험 결과 및 분석

(표 6)은 각 테스트컬렉션별로 가장 높은 성능을 나타내는 단백질 간 상호작용 식별 실험 결과와 매개변수 값을 보여준다. 실험에 필요한 학습 매개변수 C는 SVM 정규화 인자를 나타내고, λ는 비교 대상이 되는 구문 트리들의 깊이(tree depth)가 서로 상이함에 따라 발생하는 커널 값의 불일치성을 해결하기 위해서 사용되는 트리 커널 소멸 인자를 나타낸다. 실험 결과, 대체적으로 80% 중ㅡ후반대의 높은 성능을 보여주고 있다.

다음으로 본 논문에서 구현한 접근 방법과 [20, 21]에서의 접근 방법에 대한 성능 비교를 표 7에 나타내었다.

(표 6) 각 테스트컬렉션별 최고 성능

Corpus	λ	C	micro-F1(%)	macro-F1(%)
AIMed	0.6	6.0	89.5125	81.239
BioInfer	0.5	7.0	88.9233	85.9112
HPRD50	0.7	6.0	84.5266	83.3478
IEPA	0.3	4.0	78.799	77.8944
LLL	0.3	7.0	86.9697	86.9992

(표 7) 거시 평균 기반 F-점수 기준 성능 비교

	AIMed	BioInfer	HPRD50	IEPA	LLL	평균
Airola et al. (2008) [20]	56.4	61.3	63.4	75.1	76.8	66.60
Miwa et al. (2009) [21]	60.8	68.1	70.9	71.7	80.1	70.32
Our system	77.1	82.0	79.0	76.3	85.6	80.00

학습 매개변수 중의 하나인 SVM 정규화 인자는 [21]과의 객관적인 비교를 위해서 1.0으로 일치시켰고, 트리 커널 소멸 인자는 모두 0.5로 동일하게 적용하였다. 실험 결과, 모든 테스트컬렉션에 대해서 본 논문에서 제안한 시스템이 우수한 성능을 보이고 있다. 특히 학습 집합의 규모가 큰 AIMed와 BioInfer 테스트컬렉션에서의 성능 향상이 두드러진다.

5. 결론 및 향후 연구

본 연구에서는 문장 내에 존재하는 두 개체 간의 구절 구조 정보를 이용하여 유사도를 계산하는 합성곱 구문 트리 커널과 두 개체 사이의 유의미한 연관관계를 표현해주는 술어-논항 구조 패턴을 기반으로 유사도를 계산하는 커널을 선형 관계로 결합하는 혼합 커널을 제안하였다. 그리고 이것을 과학기술 문헌의 PLOT 간 연관관계 추출 및 바이오 분야의 단백질 간 상호작용 식별 문제에 적용하여 성능 향상을 입증하였다. 결과적으로 문장 내에 존재하는 술어와 논항 간의 의미적 구조를 활용하는 술어-논항 구조 패턴 유사도 커널은 기존의 합성곱 구문 트리 커널과 결합됨으로써 관계 추출의 성능을 향상시킬 수 있음을 보여주었다.

향후 연구로는 본 논문에서 제안한 아이디어를 또 다른 형태로 적용하는 방법을 생각해볼 수 있다. 본 논문에서는 술어-논항 구조의 패턴 유사도 정보를 트리 커널에 혼합 커널 형태로 적용하였지만, 또 다른 방법으로 트리 커널의 가지치기 수행 시에 본 아이디어를 적용해 볼 수 있을 것이다. 기존에 제안된 다양한 형태의 트리 가지치기 방법이 있지만, 술어-논항 구조의 패턴 정보에 근거하여 정말 중요한 노드들을 인식하고 그 정보에 기반하여 가지치기를 수행한다면 혼합 커널 기법이 아닌 트리 커널 단독으로도 충분히 성능을 향상시킬 수 있을 것으로 예상한다. 이 경우 트리 커널은 또 다른 종류의 커널과 결합하여 더욱 정교한 혼합 커널을 구성할 수 있을 것이다.

참 고 문 헌

[1] Bunescu, R. C. and Mooney, R. J., “A Shortest Path Dependency Kernel for Relation Extraction,” Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, pp.724-731, Vancouver, B.C., 2005.

[2] Culotta, A. and Sorensen, J., “Dependency Tree Kernels for Relation Extraction,” Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, 2004.

[3] Bunescu, R. C. and Mooney, R. J., “Subsequence Kernels for Relation Extraction,” Advances in Neural Information Processing Systems, 2006.

[4] Kambhatla N., “Combining lexical, syntactic and semantic features with Maximum Entropy models for extracting relations,” ACL’2004 (Poster), pp.178-181, 21-26 July, Barcelona, Spain, 2004.

[5] GuoDong Z., Su J. Zhang J. and Zhang M., “Exploring various knowledge in relation extraction,” ACL’2005, pp.427-434, 25-30 June, Ann Arbor, Michigan, USA, 2005.

[6] Zhao, S. B. and Grishman, R., “Extracting Relations with Integrated Information Using Kernel Methods,” ACL-2005, 2005.

[7] Zelenko, D., Aone, C. and Richardella, A., “Kernel Methods for Relation Extraction,” Journal of Machine Learning Research 3, pp.1083-1106, 2003.

[8] Zhang, M., Zhang, J., Su, J. and Zhou, G., “A Composite Kernel to Extract Relations between Entities with both Flat and Structured Features,” 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, pp.825-832, 2006.

[9] GuoDong Z., Min Z., Dong H. J. and QiaoMing Z., “Tree Kernel-based Relation Extraction with Context-Sensitive Structured Parse Tree Information,” Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning,

pp.728-736, Prague, June 2007.

[10] Zhang, M., GuoDong, Z. and Aiti, A., "Exploring syntactic structured features over parse trees for relation extraction using kernel methods," Information Processing and Management, v.44, pp.687-701, 2008.

[11] Vishwanathan S. V. N. and Smola A. J., "Fast Kernels for String and Tree Matching," Advances in Neural Information Processing Systems, MIT Press, vol.15, pp.569-576, 2003.

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

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

[14] Pyysalo S., Airola A., Heimonen J., Bjorne J., Ginter F. and Salakoski T., "Comparative analysis of five protein-protein interaction corpora," BMC Bioinformatics, vol.9, no.S6, 2008.

[15] 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.

[16] 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.

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

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

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

[20] 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.

[21] Miwa M., Sætre R., Miyao Y. and Tsujii J., "Protein-protein interaction extraction by leveraging multiple kernels and parsers," International Journal of Medical Informatics, 2009.

◐ 저 자 소 개 ◑

정 창 후 (Chang-Hoo Jeong)

1999년 충남대학교 컴퓨터과학과 졸업(학사)
2002년 충남대학교 대학원 컴퓨터과학과 졸업(석사)
2003년~현재 한국과학기술정보연구원 선임연구원
관심분야 : 정보검색 및 추출, 텍스트마이닝
E-mail : chjeong@kisti.re.kr

◐ 저 자 소 개 ◑

최 성 필 (Sung-Pil Choi)

1996년 부산대학교 전자계산학과 졸업(학사)
1998년 부산대학교 대학원 전자계산학과 졸업(석사)
2009년 한국과학기술원 대학원 정보통신공학과(박사 수료)
1998년~현재 한국과학기술정보연구원 선임연구원
관심분야 : 기계학습, 정보검색, 자연어처리, 정보추출, 텍스트마이닝
E-mail : spchoi@kisti.re.kr

최 윤 수 (Yun-Soo Choi)

1993년 충남대학교 컴퓨터공학과 졸업(학사)
1995년 충남대학교 대학원 컴퓨터공학과 졸업(석사)
1995년~현재 한국과학기술정보연구원 선임연구원
관심분야 : 정보검색, 텍스트마이닝
E-mail : armian@kisti.re.kr

송 사 광 (Sa-Kwang Song)

1997년 충남대학교 통계학과 졸업(학사)
1999년 충남대학교 대학원 컴퓨터과학과 졸업(석사)
2011년 한국과학기술원 대학원 전산학과 졸업(박사)
2005년~2010년 한국전자통신연구원 바이오인포매틱스팀 연구원
2010년~현재 과학기술정보연구원 선임연구원
관심분야 : 텍스트마이닝, 자연어처리, 정보검색, 시맨틱 웹
E-mail : esmallj@kisti.re.kr

전 홍 우 (Hong-Woo Chun)

2002년 고려대학교 컴퓨터학과 졸업(학사)
2004년 고려대학교 대학원 컴퓨터학과 졸업(석사)
2007년 일본 동경대학 대학원 컴퓨터과학 전공 졸업(박사)
2009년~현재 한국과학기술정보연구원 선임연구원
2008년~2009년 Japan Research Organization of Information Systems, Database Center for Life Science, Project researcher
2007년~2008년 Japan National Institute of Advanced Industrial Science and Technology (AIST), Japan Biological Information Research Center (JBIRC), 박사후과정
관심분야 : 자연어처리, 기계학습
E-mail : hw.chun@kisti.re.kr