평면적 어휘 자질들을 활용한 확장 혼합 커널 기반 관계 추출

(Relation Extraction based on Extended Composite Kernel using Flat Lexical Features)

최 성 필 † 정 창 후 † 최 윤 수 †† 맹 성 현 †††
(Sung-Pil Choi) (Chang-Hoo Jeong) (Yun-Soo Choi) (Sung-Hyon Myaeng)

요 약 본 논문에서는 기존의 관계 추출 성능을 향상시키기 위해서 기존의 자질 기반 방법에서 추구하였던 개체 주변 문맥 다양성 정보의 추출 및 적용과 커널 기반 방법의 강점인 관계 인스턴스에 대한 구문 구조적 자질 정보의 통합 활용을 통한 확장된 혼합 커널을 제안한다. ACE RDC 코퍼스[1]를 활용한 실험에서, 기존의 합성곱 구문 트리 커널 기반 혼합 커널을 기반으로 총 9 종류의 평면적 어휘 자질 집합을 정의하고 이를 적용함으로써 성능 향상에 기여하는 어휘 자질 유형을 파악할 수 있었으며, 적은 규모의 학습 집합으로도 현재 최고 수준의 성능에 필적하는 결과를 얻을 수 있었다. 결론적으로 관계 추출을 위한 세 가지 핵심 정보, 즉 개체 자질, 구문 구조적 자질, 주변 문맥 어휘 자질을 통합 적용하면 관계 추출의 성능을 향상시킬 수 있음을 알 수 있었다.

키워드 : 정보 추출, 관계 추출, 기계 학습, 지지벡터기계, 혼합 커널, 합성곱 구문 트리 커널, 평면적 어휘 자질

Abstract In order to improve the performance of the existing relation extraction approaches, we propose a method for combining two pivotal concepts which play an important role in classifying semantic relationships between entities in text. Having built a composite kernel-based relation extraction system, which incorporates both entity features and syntactic structured information of relation instances, we define nine classes of lexical features and synthetically apply them to the system. Evaluation on the ACE RDC corpus shows that our approach boosts the effectiveness of the existing composite kernels in relation extraction. It also confirms that by integrating the three important features (entity features, syntactic structures and contextual lexical features), we can improve the performance of a relation extraction process.

Key words : Information Extraction, Relation Extraction, Machine Learning, Support Vector Machines, Composite Kernels, Convolution Parse Tree Kernels, Flat Lexical Features

† 정 회 원 : 한국과학기술정보연구원 정보기술연구실
spchoi@kisti.re.kr
chjeong@kisti.re.kr
†† 비 회 원 : 한국과학기술정보연구원 정보기술연구실
armian@kisti.re.kr
††† 종신회원 : 한국과학기술원 전산학과 교수
myaeng@kaist.ac.kr
논문접수 : 2009년 6월 3일
심사완료 : 2009년 6월 24일

정보과학회논문지: 소프트웨어 및 응용 제36권 제8호(2009.8)

1. 서 론

자연어 처리 및 텍스트 마이닝 분야에서 정보 추출(Information Extraction)은 핵심적인 영역으로 인식되고 있다. 정보 추출의 최종 목표는 비정형적인 텍스트 데이터 내에서 테이블화된 정형 데이터를 추출, 변환하기 위해서 텍스트 내에 존재하는 중요하고 연관성 있는 정보를 식별하는 것이다[1]. [2]에서는 이러한 정보 추출 기술을 구성하는 요소 기술로서 (1) 개체명 인식(Named-Entity Recognition), (2) 관계 추출(Relation Extraction), (3) 대용어 참조 해소(Coreference Resolution) 등을 언

1) Automatic Content Extraction, ACE Project http://www.ldc.upenn.edu/Projects/ACE/

급하고 있다. 이 중 관계 추출은 현재까지도 가장 난이도가 높은 미해결 분야로 인식되고 있다[1-3].

현재까지 관계추출의 성능을 높이기 위해서 다양한 지도 학습(Supervised Learning) 기반의 관계 추출 기법이 소개되었다. 이들은 (1) 규칙기반 방법(Rule-based Methods), (2) 자질기반 방법(Feature-based Methods), (3) 커널기반 방법(Kernel-based Methods)의 세가지 유형으로 분류될 수 있다. 이들 중 비교적 최근에 개발된 관계추출에 특화된 커널 함수를 새롭게 구성하여 이를 기반으로 지지벡터기계(Support Vector Machines)에 적용하는 커널기반 방법의 효과가 주목받고 있다. 관계추출 분야에서의 커널기반 방법의 특징은 한 문장에 존재하는 두 개체간의 관계를 가장 잘 표현하고, 이를 포함하는 두 관계 포함 문장들 간의 유사도를 가장 효과적으로 계산하는 커널을 구성하기만 하면, 그 성능이 매우 높게 나타난다는 것이다. 특히 개체쌍 및 그들 간의 관계가 포함된 문장을 구문 분석하여 얻어지는 구문 트리 기반의 커널을 기반으로 한 합성곱 구문 트리 커널[2](Convolution Parse Tree Kernel, 이하 CPTK)과 대상 개체 고유의 속성자질을 활용한 혼합 커널(Composite Kernel, 이하 CK) 기반 관계추출 기법의 성능이 매우 뛰어난 것으로 알려져 있다[4,5].

본 논문에서는 다양한 평면적 어휘 자질들을 활용하여 기존의 혼합 커널 기반 관계 추출 성능을 향상시키는 방안을 제시한다. 이를 위해, 개체 구성 단어, 주변 문맥 단어 등 총 9 종류의 어휘자질유형을 정의하고 이를 기존의 혼합 커널 기반 관계추출 기법에 다각적으로 적용하여 성능 향상의 중요한 단서가 될 수 있는 자질 유형을 선별하였다.

논문의 구성은 다음과 같다. 우선 2장에서 관계추출과 관련한 선행 연구에 대해서 살펴본다. 이어서 3장에서는 본 논문의 기저 시스템(base system)이 되는 혼합 커널 기반 관계 추출 기법에 대해서 살펴보고, 이를 개선하기 위해 본 논문에서 활용한 평면적 어휘 자질들을 4장에서 소개한다. 5장에서 본 논문에서 제시한 시스템의 성능을 평가하고, 평가 결과에 대한 분석을 제시한다. 마지막으로 6장에서 결론과 향후 연구 방향을 논한다.

2. 관련 연구

지도학습기반 관계추출(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 관계추출을 위한 학습 집합을 제공함으로써 이 분야 연구의 단초를 제공하였다. 그 당시에 공개된 최고 성능은 F1 기준으로 75% 였다[6].

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[7]. 이를 기반으로 GuoDong et al.(2005)는 지지벡터기계(Support Vector Machines)를 활용하여 더 확장되고 세분화된 자질 정보를 관계 추출에 적용하였다[8]. 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 고안하고 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[9]. 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[1]. 또한 Bunescu et al.(2005)는 [1]의 결과를 확장하여 의존 구문 트리를 부분 트리로 분할하고, 문장 내의 의존 관계 경로를 대상으로 커널 함수를 구성하여 [1]에서보다 더 나은 결과를 얻었다[2].

최근에는 Zhang et al. (2006)이 Collins and Duffy (2001)에서 새롭게 고안한 합성곱 구문 트리 커널(convolution parse tree kernel)을 기반으로 다양한 구조적 자질 정보와 기존의 개체 자질 정보를 결합한 혼합 커널(composite kernel)을 개발하였다[4,5]. 또한

2) "Convolution"을 한글로 번역함에 있어서 수학, 물리학 분야의 다양한 사전 및 문헌을 참조하였으며 결론적으로 "합성곱"이라는 용어가 본 연구의 문맥에 가장 적합하다고 판단되어 이를 채택하였음.

GuoDong et al.(2007)은 [5]에서 제안한 구문 트리 커널의 가지치기 기법 및 커널 계산 기법을 확장하여 개체 쌍 주변 문맥까지도 포괄하는 새로운 트리 커널을 개발하였다[4]. 그러나 두 기법 모두 관계 인스턴스의 구문 구조 정보와 컬렉션에서 제공하는 개체 자질 정보만을 활용함으로써 일반적으로 자질 기반 기법에서 추구하고자 했던 관계 추출을 위한 인스턴스내 문맥 다양성 정보의 집약적 적용에 대한 연구는 수행되지 않았다.

본 논문에서는 이러한 기반 연구에 기초하여 자질 기반 기법에서 시도되었던 문맥 다양성 정보(contextual diversity information)에 대한 통합적 적용 방법과 커널 기반 기법의 구조 및 개체 자질 정보를 결합한 확장된 혼합 커널을 구성하여 기존 시스템의 성능을 개선하고자 한다.

3. 혼합 커널 기반 관계 추출

3 장에서는 관계 추출에 적용되어 높은 성능을 나타내고 있는 혼합 커널(Composite Kernel)을 소개하고, 그 장점 및 한계점에 대해서 살펴본다. 관계 추출 측면에서 볼 때, 혼합 커널은 (1) 합성곱 구문 트리 커널(Convolution Parse Tree Kernel)과 (2) 개체 커널(Entity Kernel, 이하 EK)로 구성되어 있다[4,5].

3.1 합성곱 구문 트리 커널

CPTK는[10]에서 최초로 고안되었으며, [4,5]에서 EK와 복합적으로 활용되어 관계 추출에 적용되었다. 일반적인 커널 기반 기계 학습에서 두 개의 구문 분석 트리의 유사도를 측정하기 위한 특수한 커널 함수로서 이를 기반으로 다양한 향상된 기법들이 개발되기도 하였다. 특히 [11]에서는 [10]에서 고안된 기존의 "부분 집합 트리 커널"(SubSet Tree Kernel, 이하 SST)에 새롭게 "부분 트리 커널"(SubTree Kernel, 이하 ST)이라는 새로운 트리 커널을 제안하였으며, 더불어 두 가지 트리 커널의 계산 속도를 높이기 위한 알고리즘도 함께 제시하였다. 아래 그림은 두 가지 방법에 의해서 구문 트리 내의 부분 트리가 분리되는 모습을 보여준다.

그림 1은 "*at Ramstein, Germany*"라는 전치사구에 대해서 ST 기법을 적용하여 분리된 부분 트리집합을 보여준다. ST 기법은 트리 내에서 특정 노드의 모든 자식 노드로 구성된 부분 트리를 구성하는 것이다. 따라서 모든 부분 트리는 최말단 자식 노드로서, 전체 트리의 잎 노드(leaf node)를 가져야 하며, 구문 생성 규칙에 위배되지 말아야 한다. 아래 그림은 "*at Ramstein*"이라는 전치사구가 SST 기법에 의해서 분리되는 모습을 보여준다.

위 그림에서 보는 바와 같이 SST 기법은 ST 기법보다 좀더 일반화된 방법으로서, 특정 부분 트리가 반드시 전체 트리의 잎 노드(leaf node)를 가질 필요는 없다. 다시 말해서, 구문 생성 규칙에 위배되지만 않는다면, 특정 노드에서 출발하여 그 노드의 자식 노드 중 일부분을 포함할 수 있으며, ST 기법보다 훨씬 많은 부분 트리를 생성한다.

두 입력 구문 트리의 유사도를 효과적으로 측정하기 위해서, [10,11]에서는 한 구문 트리를 구성하는 요소 트리 집합을 추출하고 이를 벡터 공간의 하나의 축으로 지정함으로써, 구문 트리가 공간 내의 특정 벡터로 사상(mapping)될 수 있도록 하였다. 따라서 입력 구문 트리 $\vec{x}_{pt}$는 다음과 같이 함수 φ에 의해서 새로운 자질 공간 Φ로 전사된다.

$$\varphi : \vec{x}_{pt} \in \mathrm{X} \mapsto \varphi\left(\vec{x}_{pt}\right) \in \Phi \subseteq \mathbb{R}^{N} \quad (1)$$

식 (1)에서 사상된 자질 공간은 N-차원의 유클리드 공간이며, 자질 공간 내에서의 $\varphi\left(\vec{x}_{pt}\right)$은 다음과 같이, 구성 요소 트리의 출현 빈도 벡터로 표현된다.

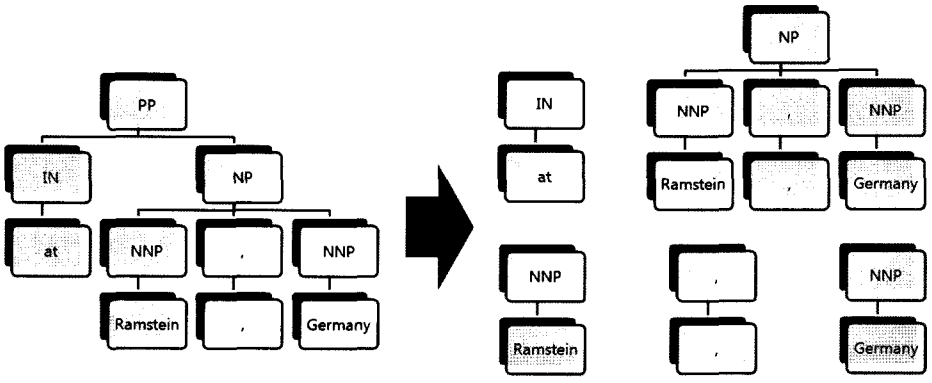

그림 1 ST 기법으로 분리된 구문 트리

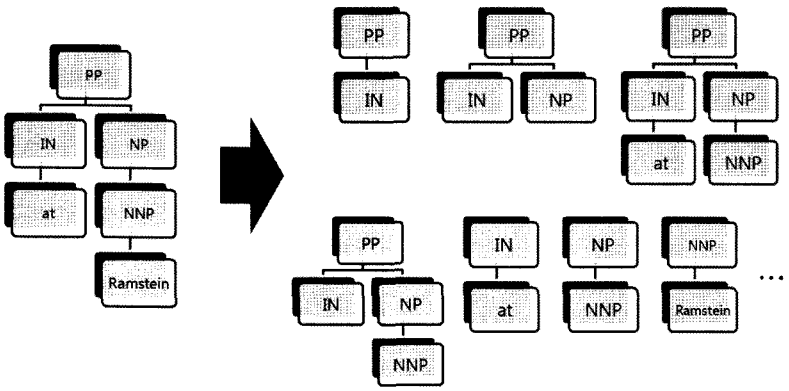

그림 2 SST 기법으로 분리된 구문 트리

$$\varphi(\vec{x}_{pt}) = \left(f_1(\vec{x}_{pt}), f_2(\vec{x}_{pt}), ..., f_N(\vec{x}_{pt})\right)$$
$$f_i(T) = \text{the number of } subtree_i \in S, \text{ appearing in } T$$
$$S = \text{a set of all the unique subtrees of the entire tree set.} \quad (2)$$

식 (2)에서 함수 $f_i(T)$는 구문 트리 T 내에 존재하는 i 번째 부분 트리의 의 출현 빈도를 계산한다. 따라서 $\varphi(\vec{x}_{pt})$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.

$$K_{pt}(\vec{x}_{pt}, \vec{x}'_{pt}) = \left\langle \varphi(\vec{x}_{pt}), \varphi(\vec{x}'_{pt}) \right\rangle = \sum_{i=1}^{N} \left[f_i(\vec{x}_{pt}) \cdot f_i(\vec{x}'_{pt}) \right] \quad (3)$$

그러나 특정 입력 구문 트리 집합 내에 존재하는 모든 부분 트리를 추출하고, 이를 이용하여 개별 입력 구문 트리에 대해서 벡터를 구성하는 작업은 상당히 비효율적이다. 따라서 [10]에서는 $K_{pt}(\vec{x}_{pt}, \vec{x}'_{pt})$를 직접 계산하는 재귀적 방법을 고안해 냄으로써 처리 속도를 향상시켰다. 본 논문에서는 세부 알고리즘에 대한 내용은 지면상 생략하며, [10,11]에 자세하게 기술되어 있다.

3.2 개체 커널

특정 문장 내에서 서로 연관성이 있는 개별 개체명은 각각의 특성을 가지고 있다. 따라서, 이러한 특성 자질들을 반영하면 관계 추출 성능이 매우 높아질 수 있다. 본 논문의 실험 데이터인 ACE 2003에는 총 4가지 종류의 개체 자질[3]이 정의되어 있다. [4,5]에 의하면, 이러한 개체 특성 자질들은 개체 간의 관계 설정에 큰 영향을 주고 있으므로, 다음과 같이 선형 커널(linear kernel)을 구성함으로써 직접적인 분별 효과를 나타낼 수 있도록 하였다.

$$K_{entity}(E_1, E_2) = \sum_i I(E_1.f_i, E_2.f_i)$$
$$I(f_1, f_2) = \begin{cases} 1, & \text{if } f_1 = f_2 \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

여기서 $E_k.f_i$는 개체 E_k의 i 번째 자질을 나타낸다. 본 논문에서는 ACE 2003에서 제공하는 개체 자질 중에서 관계 추출에 가장 영향력이 강한 개체 유형(Entity Types), 개체 하부 유형(Entity Subtypes)을 커널에 적용하였다.

3.3 평면 자질 기반 선형 커널

본 논문에서는 CPTK가 표현하는 구문 구조적 자질(syntactic structural feature)과 대비되는 개념으로서 평면 자질(flat feature)이라는 개념을 도입한다. 이미 [5]에서 이러한 두 종류의 자질을 구분하여 관계 추출에 적용하였으나, 본 논문에서는 이러한 평면 자질의 개념을 확장하여, 다양한 주변 문맥 어휘 자질들을 포괄할 수 있는 개념으로 변형시킨다. 따라서 식(4)에서 정의한 개체 커널을 기반으로 다음과 같은 평면 자질 기반 선형 커널을 구성할 수 있다.

$$K_{flat}(R_1, R_2) = \left[\sum_{i=1,2} K_{entity}(R_1.E_i, R_2.E_i) \right] + K_{lexical}(R_1, R_2)$$
$$K_{entity}(E_1, E_2) = \sum_i I(E_1.f_i, E_2.f_i)$$
$$K_{lexical}(R_1, R_2) = \sum_i I(R_1.f_i, R_2.f_i)$$
$$I(f_1, f_2) = \begin{cases} 1, & \text{if } f_1 = f_2 \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

K_{flat}은 크게 두 가지의 커널로 구성되어 있다. 하나는 앞에서 설명한 개체와 직접적으로 연관된 개체 자질, K_{entity}이고, 나머지 하나는 4장에서 설명할 다양한 평면적 어휘 자질이 적용되는 $K_{lexical}$이다. 평면적 어휘 자질에 대한 상세한 내용은 4장에서 다룬다.

3) 개체 표제어 정보(entity headwords), 개체 유형(entity types), 개체 하부 유형(entity subtypes, GPE에만 해당), 언급 유형(mention type)

3.4 관계 인스턴스 가지치기

관계 인스턴스 가지치기란 [5]에서 소개된 개념으로서, 커널 함수의 유사도 측정 성능을 향상시키고, 관계 추출을 위한 학습 시에 불필요한 문맥 정보들을 제외시키기 위해서 구문 트리의 일부분만을 남기고 나머지는 제거하는 작업을 말한다. [5]에서는 구문 트리 내에서 제거되는 구문 정보 및 단말 노드 정보의 종류에 따라 5 가지의 가지치기 방법을 제시하였고, [12]에서는 이를 확장하여 총 7 가지의 가지치기 방법을 고안하였다. 다음 표 1에 총 7 가지 가지치기 방법에 대한 세부 내용을 정리하였다.

위 표 1에서는 각 가지치기에 대한 설명과 더불어 ACE 2003 검증 컬렉션을 이용하여 CPTK 단독으로 실험한 결과의 성능 순위를 함께 보여준다. 보는 바와 같이 PT 방법이 가장 좋은 성능을 보이고 있다. 따라서 전체 문장에서 좌, 우 개체를 포함하는 문장 일부의 구문 구조가 두 개체간 관계를 추정하는데 있어 가장 좋은 단서가 될 수 있음을 알 수 있다.

위의 결과에 따라, 본 논문에서는 기저 시스템으로서 이들 방법 중에서 가장 성능이 높은 "경로 포함 트리(Path-enclosed Tree: PT)"를 이용한 CPTK를 구성하였다. 그림 3은 ACE 2003 컬렉션에 출현한 "*The espionage trial of U.S. businessman Edmund Pope is under way in Moscow*" 라는 문장에 대한 구문 분석 트리와 PT 방법을 사용했을 때, 생성되는 부분 트리를 점선으로 표시하였다. 그림에서 보는 바와 같이, PT는 두 개체와 그 개체 사이에 존재하는 단어들로 구성된 부분 문장들에 대한 구문적 구조 정보만을 남기고 나머지 정보들은 제거하게 된다. 따라서 위의 원본 문장이 "*U.S. businessman Edmund Pope is under way in Moscow*"로 축소되므로, 두 개체의 관계를 분류할 때, 불필요한 구조 자질들의 영향을 최소화하게 된다.

3.5 관계 추출을 위한 확장된 혼합 커널

본 절에서는 앞에서 설명한 세 가지 커널을 결합한

표 1 구문 트리 가지치기 방법[4,5,12]

Tree Pruning Methods	Details	F1 Ranking
Minimum Complete Tree(MCT)	구문 트리 내에서 두 개체를 포함하고 있는 최소 완전 부분 트리	7
Path-enclosed Tree(PT)	두 개체를 연결하는 최소 경로 내에 포함된 부분 트리	1
Chunking Tree(CT)	PT에서 기저구(Base Phrase) 및 품사정보를 제외한 모든 내부 노드들을 제거한 트리	5
Context-sensitive PT(CPT)	PT에서 좌측 개체의 좌측 노드 하나, 우측 개체의 우측 노드 하나를 추가한 트리	3
Context-sensitive CT(CCT)	CT에서 좌측 개체의 좌측 노드 하나, 우측 개체의 우측 노드 하나를 추가한 트리	6
Flattened PT(FPT)	PT에서 부모 노드 및 자식 노드가 각각 1개 뿐인 노드들을 제거(품사 노드 제외)	2
Flattened CPT(FCPT)	CT에서 부모 노드 및 자식 노드가 각각 1개 뿐인 노드들을 제거(품사 노드 제외)	4

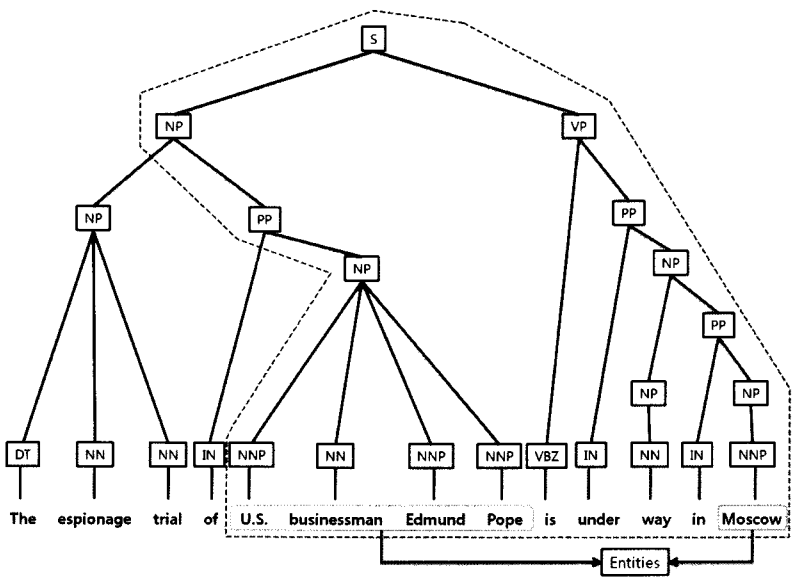

그림 3 전체 구문 트리 및 PT 기반 가지치기된 부분 트리

표 2 관계 포함 문장에 속하는 개체 타입 및 관계 타입

Sentences	1st entity's class	2nd entity's class	Relation (major)
They have **biological weapons.**	GPE	WEA	ART
And how'd **they** get **them.**	GPE	WEA	ART
They went through **a foreign subsidiary.**	ORG	ORG	PART-WHOLE
Congressman Peter King from **New York.**	PER	GPE	GEN-AFF
My question is not **who** votes at **the U.N.**	GPE	ORG	ORG-AFF

새로운 혼합 커널을 정의한다. 기존의 혼합 커널은 관계 분류에 필요한 자질들 중에서 구문 구조 자질과 개체 자질만을 사용하였다[4,5,12]. 그러나 관계 포함 문장의 구문적 자질과 개체 자체의 속성 자질로는 분별이 힘든 다양한 형태의 문장들이 존재한다.

만일 "*They have biological weapons*"라는 문장에 존재하는 두 개의 개체, "*They*"와 "*biological weapons*"에 대한 관계를 설정한다고 할 때, 기존의 방법은 "*NP + VP + NP*" 형태의 구문적 정보와 두 개체의 속성 정보(개체 클래스, 위 표에서 "*GPE*", "*WEA*") 만을 이용하여 분류를 하게 된다. 그러나, 세 번째 문장인 "*They went through a foreign subsidiary*"도 역시 넓은 의미에서 위의 문장과 동일한 구문적 구조("*NP + VP + NP*")를 가진다. 이러한 구문적 구조는 매우 빈번하게 발행하므로, CPTK의 변별력이 매우 떨어지게 된다. 따라서 이 경우에는 혼합 커널 내의 개체 커널이 주로 관계 인스턴스 유사도 측정의 책임을 맡게 된다. 비록 ACE 컬렉션 내부에서는 두 개체의 클래스 쌍과 대상 관계 사이에 매우 긴밀한 의존성이 존재하지만[5], 정확한 분류를 위해서는 반드시 개체 주변에 존재하는 "*have*"(첫째 문장인 경우)와 "*went through*"(셋째 문장인 경우)와 같은 문맥 자질이 동반되어 이들의 의미적 분별력이 활용되어야 한다.

위와 같은 기존 방법의 한계를 극복하기 위해서 본 논문에서는 다음과 같은 확장 혼합 커널을 정의하고, 여기에 4장에서 제시되는 다양한 형태의 문맥기반 평면적 어휘 자질을 적용하였다.

$$\hat{K}_c(R_1,R_2)=\tau\times K_{pt}(R_1\tilde{t},R_2\tilde{t},\lambda)+K_{flat}(R_1,R_2)$$
$$R_i\tilde{t}=\Pr(R_i t) \quad (6)$$

식 (6)에서 K_{pt}는 식 (3)의 CPTK를, $R_i t$는 관계 인스턴스 R_i의 구문 트리를 나타낸다. 또한 $\Pr(t)$는 구문 트리 가지치기 함수를 나타낸다. 또한 τ는 K_{pt} 값의 적용 정도를 나타내는 가중치 값이며, λ는 두 구문 트리의 유사도 계산에서 트리 내의 노드 깊이(depth)에 따라서 노드간 유사도의 적용 정도를 조절하는 소멸 인자(decay factor)이다. 위에서 언급하였듯이, 여기서는 관계 추출 성능이 가장 뛰어난 것으로 알려진 *PT* 기반의 가지치기만을 수행하였다.

4. 평면적 어휘 자질

이 장에서는 3.4에서 제시한 식 (6)의 K_{flat} 커널 내부의 $K_{lexical}$ 요소 커널에 적용되는 어휘 자질에 대해서 설명한다. 본 논문에서 관계 인스턴스를 기반으로 자동으로 추출한 평면적 어휘 자질은 총 9가지이다.

4.1 개체 어휘 자질

개체 어휘 자질이란 관계 인스턴스 내부에서 단일 혹은 다중 단어로 구성된 개체에 국한된 어휘 자질을 말한다. 총 3가지의 자질이 존재한다.

- *Entity Mention*(**EM**): 단일 혹은 다중 단어로 구성된 개체명 전체(개별 관계 인스턴스 당 2개)를 직접 자질로 활용한다. 이는 특정 개체가 특정 관계와 빈번하게 연관되는 현상을 활용하기 위한 자질이다.
- *Entity Words*(**EW**): 개체를 구성하는 요소 단어 집합. ACE 컬렉션 내의 상당수 개체가 매우 긴 다중 단어 개체로 구성되어 있다.(예: "*the streets in Britain, Spain and Italy, whose governments approved of the war*") 따라서 이를 구성하는 단어들은, 통계적으로 볼 때, 관계 추출의 중요한 단서가 될 수 있다.
- *Entity Words Bigram*(**EB**): 단일 혹은 다중 단어로 구성된 개체명의 단어 바이그램.

개체 어휘 자질 적용의 또 다른 부차적 목적은 중첩 개체 쌍[4](Overlapped Entity Pair)에 대한 관계 추출 성능 개선이다. 이러한 중첩 개체 쌍이 포함된 문장을 구문분석하고 인스턴스 가지치기(PT)를 수행하면, 남는 것은 하나의 개체에 대한 구문분석 트리 일부 뿐이다. 따라서, 관계 분류 시에 문맥 구문 구조 정보가 적용되는 것이 아니라, 단순히 개체 자체에 대한 구문 정보만이 활용된다. 이를 개선하여 더 많은 특정 정보가 적용될 수 있도록 개체 구성 단어 정보를 어휘 자질로서 추가하였다.

4.2 문맥 어휘 자질

문맥 어휘 자질이란 개체를 둘러싸고 있는 주변 문맥 어휘들을 기반으로 구성된 자질이다. 주변 문맥 어휘 적용의 가장 큰 목적은 표 2에서 제시한 예제 문서에서

4) 두 개체 쌍에서 특정 개체가 다른 개체의 내부에 포함되어 있는 경우. (예) Iraqi troops → Iraqi + Iraqi troops, his lawyers → his + his lawyers, the former Iraqi leader → the former Iraqi leader + Iraqi

보듯이 두 개체의 의미 관계를 효과적으로 표현하는 단어들을 관계 추출에 적용하기 위함이다. 관련 연구에서도 언급하였지만, 이러한 주변 문맥 자질들은 [6,8]에서 자질기반 관계 추출 기법의 핵심 자질로서 활용되어 그 적용 효과가 입증된 바 있다.

문장 내에서 두 개체를 중심으로 분포된 주변 문맥 어휘들의 상대적 위치 및 특성에 따라서 다음의 10 가지 종류로 세분화될 수 있으며, 이들 각각의 구성 어휘들을 WordNet에 매핑시켜 추상화시킨 자질 집합이 부가적으로 생성된다.

- *Context Words*(**CW**): 개체 주변에 존재하는 단어들에 대한 위치 특성적 자질
  - BOOW(Between Only One Word): 두 개체 사이에 존재하는 단어가 1개 뿐일 때, 그 단어를 지정
  - BFW(Between First Word): 두 개체 사이에 존재하는 단어가 2개 이상일 때, 그 단어들 중 첫 번째 단어를 지정
  - BLW(Between Last Word): 두 개체 사이에 존재하는 단어가 2개 이상일 때, 그 단어들 중 마지막 단어를 지정
  - BOW(Between Other Words): BFW와 BLW를 제외한 나머지 단어들
  - PFW1(Previous First Word 1): 첫째 개체 이전에 출현한 단어
  - PSW1(Previous Second Word 1): PFW1 이전에 출현한 단어
  - NFW2(Next First Word 2): 둘째 개체 다음에 출현한 단어
  - NSW2(Next Second Word 2): NFW2 다음에 출현한 단어
  - NPP1(Next Proposition 1): 첫째 개체의 다음 단어가 전치사이면 이 단어를 저장
  - PPP2(Previous Proposition 2): 둘째 개체의 이전 단어가 전치사이면 이 단어를 저장
- *Context Word* Synsets(**CWS**): (1)의 **CW**에 해당되는 단어 리스트에 대한 Wordnet Synset 매핑. CW 내의 총 8가지 단어들에 해당하는 Wordnet Synset 식별자(Synset ID)를 자질로 활용한다. 이는 자료 희소성(data sparseness)을 극복하기 위한 한 방법으로 채택하였으며, 본 논문에서는 MIT Java Wordnet Interface(JWI 2.1.5)[5]를 이용하여 단어를 Wordnet에서 검색하고 그 첫 번째 검색 결과를 이용하였다. Wordnet에서 검색되는 대상 단어는 명사, 동사, 부사로 한정시켰다. 결과적으로 CWS 내에서도 역시 CW와 마찬가지로 10 종류의 자질이 생성된다.

CWS는 앞에서도 언급하였듯이 포함 어휘 자질들의 의미적 추상화를 통해서 CW의 출현 분포를 축소시킴으로써 핵심 문맥 개념을 적용시키려는 시도에서 본 논문에서 최초로 고안되었다. 그러나 개별 문맥 어휘를 WordNet 신셋 기반으로 추상화시키는 과정에서 발생하는 다중 매핑 문제는 추후 연구 과제로 개선되어야 할 사항이다. 이러한 문제점에도 불구하고 아래 실험 결과에서는 이 CWS의 성능 개선 효과가 두드러짐을 알 수 있다. 만일 위의 문맥 어휘 추상화에 따른 정확도를 향상시킨다면 관계 추출에서 더 높은 성능 개선 효과가 기대된다.

4.3 개체-문맥 통합 어휘 자질

이 절에서 설명하는 자질은 구문분석 완료된 관계 포함 문장을 가지치기한 후에 생성되는 부분 구문 트리 내의 모든 어휘를 대상으로 생성된 자질이다. 따라서 이 자질의 범주에는 문맥 어휘와 개체 구성 어휘가 결합된 형태의 통합 어휘 자질이 포함된다.

- *PT Leaf Node Bigram*(**PTB**): 경로 포함 트리(Path-enclosed Tree)의 모든 단말 노드(어휘 노드)들에 대한 바이그램(Word Bigram)
- *PT Preterminal Node Bigram*(**PTPB**): 경로 포함 트리(Path-enclosed Tree)의 모든 선행 단말 노드(품사 노드)들에 대한 바이그램(POS Bigram)
- *PT Leaf Node Synset Bigram*(**PTSB**): 경로 포함 트리(Path-enclosed Tree)의 개별 단말 노드(어휘 노드)에 대한 Wordnet synset 바이그램(Synset Bigram)
- *PT Leaf Node Synset Trigram*(**PTST**): 경로 포함 트리(Path-enclosed Tree)의 개별 단말 노드(어휘 노드)에 대한 Wordnet synset 트라이그램(Synset Trigram)

4.1과 4.2에서 제시한 개체 어휘 자질 및 문맥 어휘 자질과의 중복을 피하기 위해서 위 네 가지 통합 어휘 자질은 모두 바이그램이나 트라이그램으로 구성된다. 자질 기반 관계 추출 기법을 활용한 Zhao et al.(2005)의 연구 결과에서 보듯이 주변 문맥이나 개체를 구성하는 어휘들에 대한 바이그램은 관계 추출의 성능을 높이는데 도움을 준다[6]. 따라서 [6]에서 사용된 일반적인 어휘 자질의 범위를 세분화시키고 WordNet을 이용한 추상화 기법을 적용함으로써 새로운 어휘 자질 집합을 구성하였다. 본 논문에서는 CPTK 방법을 기반으로 위에서 제시한 세 가지 종류(총 9 가지의 어휘 자질)의 평면적 어휘 자질을 복합적으로 혼합 커널에 확장 적용하여 관계 추출 성능 평가 실험을 수행하였다.

5. 실험 및 토의

이 장에서는 4장에서 설명한 평면적 어휘 자질들이

5) http://projects.csail.mit.edu/jwi/

기존 혼합 커널에 주는 영향을 분석하기 위해서 수행한 성능 실험에 대해서 다루고 실험 데이터 및 방법 그리고 최종적으로 실험 결과 및 분석을 제시한다.

5.1 실험 데이터

제시한 혼합 커널의 성능을 측정하기 위해서, 본 논문에서는 LDC[6]에서 제공하는 ACE 2003(LDC2004T9)를 활용하였다[16]. 이 자료는 총 252건의 신문 기사와 뉴스 스크립트로 구성되어 있으며, 총 4,446개의 개체간 관계 쌍이 설정되어 있다. 개체 타입은 총 5개로서 "Persons (인명)", "Organizations(기관명)", "Locations(지명)", "Facilities(시설물)", "GPE[7]"로 구성되어 있다. 개체간 관계는 총 5가지의 주요 관계 유형(major relation type), 그 하부에 총 24 가지의 상세 관계 유형(relation subtype)으로 구성된다. 아래 표는 관계 정보와 개별 관계에 해당하는 관계 인스턴스[8]의 수를 보여준다.

표 3 ACE 2003에서의 관계 유형 종류 및 분포

Major Type	Subtype	#Instances
AT (1,225)	BASED-IN	107
	LOCATED	1,012
	RESIDENCE	106
NEAR (97)	RELATIVE-LOCATION	97
PART (561)	OTHER	3
	PART-OF	322
	SUBSIDIARY	236
ROLE (1,999)	AFFILIATE-PARTNER	44
	CITIZEN-OF	153
	CLIENT	67
	FOUNDER	14
	GENERAL-STAFF	626
	MANAGEMENT	410
	MEMBER	488
	OTHER	136
	OWNER	61
SOC (326)	ASSOCIATE	31
	GRANDPARENT	4
	OTHER-PERSONAL	19
	OTHER-PROFESSIONAL	120
	OTHER-RELATIVE	48
	PARENT	46
	SIBLING	9
	SPOUSE	49
Total		4,208

위의 표에서는 ACE 2003의 모든 개체 관계 쌍 4,446개 중에서 4,208개만 나열되어 있다. 비록 현재까지 발표된 문헌에서 명확하게 알려지지는 않았으나, 본 연구에서 분석한 바로는 컬렉션 내부에 다중 문장에 걸쳐서 나타나는 전역 관계 쌍(global relation pairs)이 약 200개 이상 존재한다. 일반적으로 관계 추출의 정의는 단일 문장 내의 개체 쌍에 대한 관계 추정을 의미하므로, 이 논문에서는 이들 전역 관계 집합은 처리 대상에서 배제하였다. 부가적으로 구문 분석 트리의 말단 노드와 개체 문자열이 완전히 일치하지 않는 관계 인스턴스 역시 제외하였다.

표 3에서 알 수 있듯이, 각 관계 유형 별 인스턴스의 개수의 편차가 매우 크다. 특히 주요 관계 유형 중에서 "*NEAR*"에 속하는 인스턴스의 개수는 97개로, 가장 많은 "*ROLE*"의 약 4.9% 밖에 되지 않는다. 상세 관계 유형으로 들어가면 문제는 더 심각해진다. "*PART/OTHER*"가 3개, "*SOC/GRANDPARENT*"가 4개로, 이 관계들은 학습 단계에서 그 식별력(discriminative power)를 거의 상실하게 된다. 그럼에도 불구하고 본 논문에서 이 컬렉션을 기반으로 실험을 수행한 가장 큰 이유는, 전 세계적으로 관계 추출 분야에서 가장 권위 있고, 많이 활용되고 있으며, 지속적으로 구축 및 평가되고 있는 컬렉션 중의 하나이기 때문이다. 또한 본 논문에서 고안된 방법론과 기존의 방법론을 동일 컬렉션 기반 하에서 비교 및 분석하기 위한 목적도 있다.

5.2 실험 방법[9]

우선 ACE 2003 내의 모든 문서에 대해서 수동으로 문장 분리 작업을 수행하였으며, 분리된 문장은 Stanford Parser[10]를 이용하여 구문 분석하였다. 생성된 구문 트리는 다시 관계 인스턴스 가지치기 과정을 거쳐서 최종 관계 인스턴스로 변환된다.

관계 분류기 학습을 위해서 [11,13]에서 개발한 SVM^{ligth} *TK* 1.2[11]를 활용하였다. 이 시스템은 SVM^{light}[14]를 기반으로 이에 트리 유사도 계산에 필요한 *SST*(SubSet Tree), *ST*(SubTree) 커널 함수가 내장되어 있으며, 기존의 벡터 자질도 함께 포괄적으로 학습될 수 있도록 구성되어 있다. 본 논문에서는 이 중에서 ST를 적용하였으며, 트리 커널 내의 소멸 인자(decay factor)는 0.75로 설정하였다. 또한 트리 커널의 가중치 τ는 1.0으로 지정하여, 평면적 어휘 자질과 동일한 적용 효과를 나타내도록 하였다. 다중 관계 분류(multi-class classification)를 위해서 모든 관계 쌍에 대해서 $K(K-1)/2$개

6) Linguistic Data Consortium (http://www.ldc.upenn.edu/)

7) Geographical/Social/Political Entities(지리학적, 사회적, 정치적 개체)

8) 관계 인스턴스란, 쉽게 말해서, 두 개체와 그 관계가 포함되어 있는 문장을 의미한다. 특정 문장에 여러 쌍의 개체가 존재하고, 그들 간의 관계도 다중으로 설정되어 있으므로, 동일한 문장이 중복될 수도 있으며, 실제로 ACE 컬렉션 내의 상당수 문장이 다중 관계 쌍을 포함하고 있다.

9) 본 논문에서는 관계 미포함 인스턴스와 관계 포함 인스턴스를 식별하는 관계 탐지(relation detection)에 대한 실험은 하지 않았다.

10) http://nlp.stanford.edu/software/lex-parser.shtml

11) http://dit.unitn.it/~moschitt/Tree-Kernel.htm

의 이진 분류기(binary classifiers)를 구성하였으며(K: 분류 개수), 이를 기반으로 다수결 원칙(majority vote principle)에 의해서 최종 관계 분류를 수행하였다. 관계 분류기의 성능 평가는 5겹 교차 평가(5-fold cross validation)를 기반으로 수행되었다.

비록 [4]와 [5]에서는 24가지의 상세 관계 유형(relation subtype)에 대한 성능 평가를 위해서 LDC2003T11과 LDC2004T9를 하나로 합친 9,683개의 학습집합을 활용하여 관계 분류 실험을 하였으나, 본 논문에서는 4,208개의 LDC2004T9만을 이용하는 관계로 5가지의 주요 관계 유형(major relation type)에 대한 실험만 수행하였다. 본 논문에서 제시된 방법론의 유효성을 보다 객관적으로 검증하기 위해서 향후 연구로 ACE 2004 및 2005에 대한 성능 평가를 진행 중에 있다.

5.3 실험 결과 및 분석

아래 표는 ACE 2003의 주요 관계 유형 5가지에 대한 관계 분류 성능 실험 결과이다. 기본적인 CPTK 기반 혼합 커널 시스템(Baseline, CK)과 함께, 본 논문에서 제시한 총 9가지의 평면 어휘 자질들에 대한 다양한 조합에 대한 평가 결과 표 4와 같이 특정 자질 조합에서 의미 있는 성능 향상이 이루어짐을 알 수 있었다.

표 4의 실험 결과를 바탕으로 관계 추출을 위한 혼합 커널과 평면적 어휘 자질 적용에 대해 다음과 같은 세부 분석을 할 수 있다.

- (10)에서 기존의 혼합 커널에 개체 자체 자질, PT 내의 단어들에 대한 바이그램 자질, PT 내의 단어들의 품사정보에 대한 바이그램 자질, 그리고 주변 문맥 단어들에 대한 WordNet 신셋 정보가 결합되어, 기존 혼합 커널에 비해 성능이 약 3.1% 정도 높게 나타난다.
- (1)과 (2)의 성능에서 보듯이, CPTK 자체로는 낮은 성능을 보이지만, 이에 개체 유형 자 질이 추가되어 혼합 커널로 변경되면 71.23%의 매우 높은 성능을 나타내고 있다. 이는 앞에서도 지적하였듯이 개체 유형 정보와 개체 간 관계에는 매우 밀접한 의존성이 존재하는 것을 증명한다.
- 전체적으로 EM과 PTB 자질은 효과적으로 성능을 개선시키고 있다. 위의 실험 결과는 약 30가지 정도의 자질 결합 형태에 대한 초기 실험 결과에서 가장 성능이 뛰어난 8개의 시스템을 선별한 결과로서 그 중 7개의 시스템에 PTB가 포함되어 있다.
- (4), (6), (10)에서 보듯이, CW는 자질로서 효과적이지 못하지만, CWS는 다른 자질들과 결합되어 최고 성능을 나타내고 있다. 이로서 WordNet과 같은 어휘 자원이 관계 추출 성능 향상에 도움을 줄 수 있음을 보여준다.
- (3)의 성능에서, EB 단독으로도 (2)의 성능을 개선하고 있음으로 보아, ACE 2003 내에 존재하는 길이가 긴 개체(포함 단어 수가 많은 개체)에 대한 관계 설정에 이 자질이 도움을 주고 있음을 유추할 수 있다.
- (7)에서 PTB 단독으로도 기존의 혼합 커널과 결합하여 기본 시스템에 비해 약 1%의 성능 개선을 나타내고 있다. 앞에서도 지적하였듯이 EM과 함께 PTB가 매우 유용한 자질임에 틀림없으나 중첩 개체 쌍에 대해서는 그 범위가 EB와 동일한 효과를 가져오므로 한계가 있다.
- (3)과 (7)의 성능 비교에서 위에서 지적한 사항이 드러난다. (7)이 더 넓은 범위에서 다양한 자질들을 수용함으로써 (3)의 한계를 극복하고 있다.

평가 대상 컬렉션과 기반 언어처리 기법의 상이함으로 인해 비록 정확한 성능 비교는 될 수 없으나, 표 5에서는 본 논문에서 제시한 시스템과 기존 시스템에 대해서 문헌에 발표된 수치를 기반으로 성능 비교를 해 보았다.

본 논문에서 제안한 시스템 외에는 모두 동일한 컬렉션을 활용하였으며, 교차 평가가 아닌 1,386개의 검증 집합을 대상으로 실험을 수행하였다. 이에 반하여 본 시

표 4 ACE 2003의 5가지 주요 관계 유형에 대한 적용 어휘 자질별 실험 결과

Methods		Precision	Recall	F1
(1)	Convolution Parse Tree(PT) Only	0.5702	0.4211	0.4845
(2)	CK(Baseline, PT + EK)	0.7545	0.6746	**0.7123**
(3)	+ EB	0.7624	0.6712	0.7139
(4)	+ EM + PTB + PTPB + CW	0.7927	0.6560	0.7179
(5)	+ EM + PTB + PTST	0.7533	0.6873	0.7188
(6)	+ EM + PTB + PTPB	0.7848	0.6631	0.7188
(7)	+ PTB	0.7624	0.6826	0.7203
(8)	+ EM + PTB	0.7720	0.6912	0.7294
(9)	+ EM + PTB + PTSB	0.7929	0.6873	0.7363
(10)	+ EM + PTB + PTPB + CWS	0.8303	0.6731	**0.7435**

표 5 기존 시스템들과의 성능 비교(ACE 2003, 5 Major Types)

System	Corpus Size	Approaches	P(%)	R(%)	F1
Ours	4,208 (LDC2004T9)	Composite kernel enriched by lexical features	83.0	67.3	74.4
GuoDong et al. (2007)[4]	9,683 (LDC2003T11) (LDC2004T9)	Composite kernel with context sensitive convolution parse tree kernel	80.8	68.4	74.1
Zhang et al. (2006)[5]	9,683 (LDC2003T11) (LDC2004T9)	Composite kernel	77.3	65.6	70.9
Zhang et al. (2006)[5]	9,683 (LDC2003T11) (LDC2004T9)	Convolution tree kernel	76.1	62.6	68.7
Zhou et al. (2005)[8]	9,683 (LDC2003T11) (LDC2004T9)	Feature-based SVM	77.2	60.7	68.0
Bunescu and Mooney (2005)[1]	9,683 (LDC2003T11) (LDC2004T9)	Shortest path dependency kernel	65.5	43.8	52.5
Culotta and Sorensen (2004)[2]	9,683 (LDC2003T11) (LDC2004T9)	Dependency kernel	67.1	35.0	45.8

스템은 5겹 교차 평가(5-fold cross validation)를 수행하였으므로 더 신뢰도가 높은 결과라고 말할 수 있다. 특히 기존의 시스템에 비해서 재현율을 크게 떨어뜨리지 않고도 정확도가 약 3% 향상되었음을 알 수 있다. 또한 선택된 어휘 자질들의 또한 학습 집합의 크기가 기존 시스템들의 50% 이하임에도 불구하고 기존의 최고 성능과 동일한 성능을 나타내고 있다. 이는 좀더 분석이 필요하지만 본 논문에서 제시한 다양한 자질들이 비교적 효과적으로 적용됨을 보여준다.

6. 결론 및 향후 연구 방향

이 논문에서는 기존의 혼합 커널 기법과 정밀한 자질 공학적 접근 방법 두 가지의 장점을 통합 적용하여 관계 추출의 성능을 향상시켰다. 특히 관계 추출을 위한 혼합 커널의 성능 향상에 도움을 줄 수 있는 평면적 어휘 자질 집합(PTB, EB, CWS 등)을 파악할 수 있었으며, 그들 간의 결합 방법이 성능 변화에 매우 민감하게 영향을 줄 수 있음을 알 수 있었다.

ACE 2003을 활용한 성능 평가에서 F1 기준으로 74.4%의 높은 성능을 보이고 있다. 이는 기존의 시스템에서 평가한 컬렉션의 절반 수준인 4,208개의 관계 인스턴스만을 활용한 수치이다. 일반적으로 커널 기반 기법은 기존의 기계 학습 기법과는 달리, 내부적으로 정규화(regularization) 및 최적화(optimization) 기법이 적용되므로 과적합(overfitting)에 매우 강건하며, 대체적으로 학습 집합의 규모가 커질수록 성능이 더 높아지는 양상을 보인다[15]. 물론 보다 정확한 검증을 위해서 자질 규모 중심적 실험이 필요하겠지만 결론적으로 본 논문에 제시된 시스템에 대한 성능 평가 결과, 평면적 어휘 자질의 긍정적 효과가 소규모의 학습 집합으로도 기존 방법과 동일한 성능을 나타냈다고 판단할 수 있다.

향후 연구로서 시스템의 오류 패턴을 분석하여 보다 면밀한 관계 추출 기법을 고안해야 한다. 또한 본 논문에서 적용한 WordNet과 같은 언어 자원들을 더 효과적으로 활용하기 위해서, 단어 의미 중의성 해소 차원에서의 접근 방법도 필요하다. 자질의 중요도를 판정할 수 있는 기준을 설정하여 이를 학습 모델에 반영하는 작업도 준비 중이다. 기존의 혼합 커널에 의존 문법 트리 커널을 추가하여 나타나는 성능의 변화도 분석해볼 만 하다. 마지막으로 다양한 관계 패턴을 처리하기 위해서 지수모델(exponential model) 기반의 기계학습 모델과의 결합 기법도 고려해야 한다.

부가적으로 한국어에 대한 지도학습기반 관계 추출을 위해서, ACE와 같은 정밀한 학습 및 검증 컬렉션이 개발되어야 한다. 이를 기반으로 다양한 형태의 관계 인스턴스 생성 및 자질 추출 기법을 통해서 최적의 관계 추출 방법을 찾아내어야 한다. 본 논문에서 나타나듯이 영어에서는 구문분석을 통한 문장 구조 자질 등을 통합 적용하여 좋은 성능을 나타내었으나 한국어에 대해서는 아직까지 시도된 바가 없으므로, 이를 위한 기반 연구가 반드시 필요하다.

참 고 문 헌

[1] Culotta, A., Sorensen, J., "Dependency Tree Ker-

nels for Relation Extraction," *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, 2004.
[2] Bunescu, R. C., 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, Vancouver, B.C.*, pp.724-731, 2005.
[3] Bunescu, R. C., Mooney, R. J., "Subsequence Kernels for Relation Extraction," *Advances in Neural Information Processing Systems*, 2006.
[4] GuoDong Z., Min Z., Dong H. J., 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, Prague*, pp.728-736, June 2007.
[5] Zhang, M., Zhang, J., Su, J., 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.
[6] Zhao, S. B., Grishman, R., "Extracting Relations with Integrated Information Using Kernel Methods," *ACL-2005*, 2005.
[7] Kambhatla N., "Combining lexical, syntactic and semantic features with Maximum Entropy models for extracting relations," *ACL'2004 (Poster)*, pp. 178-181. 21-26 July 2004, Barcelona, Spain.
[8] 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, Michgan, USA, 2005.
[9] Zelenko, D., Aone, C., Richardella, A., "Kernel Methods for Relation Extraction," *Journal of Machine Learning Research 3*, pp.1083-1106, 2003.
[10] Collins, M., Duffy, N., "Convolution Kernels for Natural Language," *NIPS-2001*, 2001.
[11] Alessandro Moschitti, "Making tree kernels practical for natural language learning," *Proceedings of EACL'06*, Trento, Italy.
[12] Zhang, M., GuoDong, Z., Aiti, A., "Exploring syntactic structured features over parse trees for relation extraction using kernel methods," *Information Processing and Management*, vol.44, pp.687-701, 2008.
[13] Alessandro Moschitti, "Efficient Convolution Kernels for Dependency and Constituent Syntactic Trees," *Proceedings of the 17th European Conference on Machine Learning, Berlin, Germany*, 2006.
[14] Thorsten Joachims, "SVM Light," http://svmlight.joachims.org/, 2008.
[15] Bernhard Schölkopf, Alexander J. Smola, *Learning with Kernels*, The MIT Press, Cambridge, Massachusetts, London, England, 2002.
[16] Linguistic Data Consortium (LDC), "Automatic Content Extraction," http://www.ldc.upenn.edu/Projects/ACE/

최 성 필
1996년 부산대학교 전자계산학과 졸업(학사). 1998년 부산대학교 대학원 전자계산학과 졸업(석사). 2009년 한국과학기술원 대학원 정보통신공학과(박사 수료) 1998년~현재 한국과학기술정보연구원 정보기술연구실. 관심분야는 기계학습, 정보검색, 자연어처리, 정보추출, 텍스트마이닝

정 창 후
1999년 충남대학교 컴퓨터과학과 졸업(학사). 2002년 충남대학교 대학원 컴퓨터과학과 졸업(석사). 2003년~현재 한국과학기술정보연구원 정보기술연구실. 관심분야는 정보검색 및 추출, 분산 데이터마이닝

최 윤 수
1993년 충남대학교 컴퓨터공학과 졸업(학사). 1995년 충남대학교 대학원 컴퓨터공학과 졸업(석사). 1995년~현재 한국과학기술정보연구원 선임연구원. 관심분야는 데이터베이스, 정보검색

맹 성 현
정보과학회논문지 : 소프트웨어 및 응용 제 36 권 제 3 호 참조