인지과학, 제19권 제1호
Korean Journal of Cognitive Science
2008. Vol 19, No. 1, 1~15.

복합 커널을 사용한 한국어 종속절의 의존관계 분석

김 상 수† 박 성 배 박 세 영 이 상 조
경북대학교 컴퓨터공학과

한국어에서 절들의 의존관계를 밝히는 작업은 구문 분석 작업에서 가장 어려운 작업들 중에 하나로 인식되고 있다. 절의 의존관계를 파악하는 일은 표면적으로 나타나는 정보만을 가지고 처리할 수 없고, 의미정보와 같은 추가적인 정보가 필요할 것으로 판단하고 처리해 왔다. 본 논문에서는 추가적인 정보를 사용하지 않고, 문장에서 얻을 수 있는 표면적인 정보만을 사용하여 절들 간의 의존관계를 파악하는 방법을 제안한다. 문장에서 얻을 수 있는 표면적인 정보는 문장의 구문 정보(tree structure information)와 어휘 및 거리 정보를 가지고 있는 정적인 정보(static information)로 나누어 볼 수 있다. 본 논문에서는 절들 간의 의존 관계 파악을 위하여 구문 정보와 정적 정보를 다루는 하나 이상의 커널의 결합해서 사용하는 복합 커널(composite kernel)을 제안하고, 이 커널에 맞는 다양한 인스턴스 공간의 설정을 제안한다. 실험은 최적화된 인스턴스 공간을 절들 간의 의존관계 파악 및 문장 수준에서 성능을 검정하였다. 관계 인스턴스 공간은 절들 간의 연결 및 하부절의 표현 유무로 나누었고, 결정된 인스턴스 공간에서 복합커널을 사용한 방법이 좋은 성능을 발휘함을 보였다.

주제어 : 한국어 구문분석, 파스 트리 커널, 복합 커널, SVM

---

† 교신저자: 김상수, 경북대학교 컴퓨터공학과, 연구 분야: 자연어처리 및 기계학습
E-mail: sskim@sejong.knu.ac.kr

한국어에서 절들의 의존관계를 밝히는 작업은 구문 분석 작업에서 가장 어려운 작업들 중에 하나로 인식되고 있다. 특히 한국어는 문장의 어순이 자유롭게 사용되고 동사가 다양하게 활용되고, 의존 관계가 성립되는 대상이 절, 명사구 등이 다양하게 나타난다. 따라서 한국어에서 절의 의존관계를 파악하는 일은 표면적으로 나타나는 정보만을 가지고 처리할 수 없고, 의미 정보 같은 추가적인정보가 필요할 것으로 판단하고 처리해왔다.

절들 사이의 의존관계를 인식하는 연구는 주로 절을 인식(clause identification)하는 연구와 문장의 의존 구조를 밝히는 연구에서 주로 있어왔다. 절을 인식하는 연구는 각각의 절의 시작과 끝 지점을 기계학습 방법을 사용하여 인식하는 문제로서 각각의 절의 시작과 끝을 인식을 통하여 절들의 의존 관계를 파악하려고 했다(Carreas, 2001; Molina, 2001; Lee. H-J, 2006). 문장의 의존구조에 관한 연구에서는 의존 관계를 파악하는 대상이 절이 아니라 하나의 단어 또는 어절 단위로 삼았다. 따라서 절들 사이의 의존 관계를 파악하기 보다는 문장 속에 속한 동사구의 의존 관계를 파악하는 방향으로 연구가 진행되었다(Uchimoto, 1999; Kudo, 2000; Kudo, 2002).

커널 함수(kernel method)는 각각의 개체의 고유한 표현(the original representation of objects)을 사용하면서, 두 개체들 사이의 유사성의 정도를 밝히는 능력이 좋다고 보고되고 있다. 이러한 능력으로 인하여 최근에는 커널 함수를 자연어처리 분야에서 많은 관심을 가지고 다양하게 적용되고 있다. 이 뿐만 아니라, 최근 연구에서는 한 문제에 여러가지 관점을 가지는 커널들을 다양한 방법으로 결합하여 정보의 손실을 줄이면서 많은 문제에 적용하고 있다(Zhang, 2006; Zelenko, 2003; Bunescu, 2005).

본 논문에서는 절들 간의 의존관계는 추가적인 정보 없이 구문 트리의 구문 구조 정보와 각 어휘 같은 정적 자질 정보만으로 파악할 수 있다고 가정했다. 이 가정에 따라 문법적 구문 정보를 잘 다루는 파스트리 커널과 정적 자질을 다루는 지수 커널을 결합한 복합 커널을 제안하고, 의존 관계를 잘 표현하는 다양한 인스턴스 공간을 제시한다. 그리고 제안된 커널과 인스턴스 공간을 활용하여 하위절의 의존관계를 파악하였고, 파스트리 및 자질 기반 방법보다 높은 성능을 발휘하는 것을 보인다.

절들 간의 의존관계 파악을 위한 복합 커널

절들간의 의존 관계 파악을 위한 확률 모델

본 연구에서는 절들 간의 의존관계를 이진 분류 문제로 정의하고, 이진 분류 문제를 가장 잘 다룬다고 알려진 SVM과 복합 커널을 이용하여 의존관계를 분석하였다. 한 문장에 속한 종속절들의 집합인 C는 $\{c_1, c_2, ..., c_n\}$으로 구성되고, 의존 관계의 패턴 집합 D는 $\{Dep(1), Dep(2), \cdots, Dep(n-1)\}$로 정의된다. 이때 $Dep(i)=j$는 절 c_i가 c_j와 의존관계에 있다는 것을 의미한다. 절 A가 절 C와 의존관계를 성립한다면 $Dep(A)=C$로 나타낼 수 있다. 이때 절 A를 의존절, 절 B를 지배절이라고 정의한다. 입력으로 C가 주어졌을 때 의존 관계 패턴인 D를 결정하는 문제는 P(D|C)가 최대로 되는 D를 찾는 문제로 볼 수 있고 이는 아래의 수식과 같이 정의할 수 있다

$$D_{best} = \arg max_D P(D|C)$$

만약 의존 관계 패턴인 D가 독립이라고 가정하면 P(D|C)는 아래의 식과 같아진다. $P(Dep(i)=j|f_{ij})$의 의미는 f_{ij}가 주어졌을 때 c_i와 c_j사이의 의존 관계가 존재하는 확률을 의미한다. 여기서 f_{ij}는 c_i와 c_j사이의 의존관계를 판단하는데 사용되는 다양한 언어적 및 표면적인 자질을 의미한다.

$$P(D|C) = \prod_{i=1}^{m-1} P(Dep(i) = j | f_{ij}), f_{ij} = f_1, ..., f_n \in R^n$$

$$P(Dep(i) = j | f'_{ij}) = \tanh\left(\sum_{k,l,f_{kl}} \alpha_{kl} y_{kl} K(f_{kl}, f'_{ij}) \right)$$

즉, SVM에서 나오는 마진의 값을 sigmoid 함수의 일종인 tanh로 취한 값이다. 이것은 의존관계의 확률을 나타낸다고 볼 수 있다(Kudo, 2000). 이렇게 구해진 확률을 바탕으로 전체 문장에 속한 절들의 의존관계를 다음의 제약(constraints)를 바탕으로 폭포수 파싱 모델(cascading parsing)을 사용하여 전체 문장에 속한 절들의 의존관계를 파악하였다(Kudo 2002).

1) 가장 마지막 절을 제외한 모든 의존절의 지배절은 항상 오른쪽에 나타난다.

2) 모든 의존절은 하나의 지배절만을 가지고, 지배절은 여러 개의 의존절을 가질 수 있다.

3) 절들 간의 의존관계는 교차(cross)가 발생하지 않는다.

파스 트리 커널과 다항 커널을 사용한 복합 커널

본 논문에서는 절들 간의 의존관계 분석을 위하여 문장의 구문정보와 어휘 등의 표면적인 정보만을 사용하였다. 그러나 이들은 하나의 문제를 다른 관점에서 다루고 있어서, 이 둘을 모두 다룰 수 있는 복합 커널(composite kernel)을 제안한다. 복합 커널은 문자의 구문 정보를 다루는 파스트리 커널과 어휘 등을 다루는 다항 커널로 구성된다. 두 커널의 조합은 아주 다양하게 만들어질 수 있으나, 본 논문에서는 구조를 잘 다루는 파스트리 커널과 다항 커널을 가장 대표적인 형태인 식(1),(2)와 같이 결합하였다. 여기서 다항 커널을 사용한 이유는 의존관계분석에서 벡터로 표현된 자질을 다루는데 다항 선형 커널보다 더 높은 성능을 보임을 보고되어서 본 논문에서는 선형커널보다 다항커널을 사용하였다.

$$\begin{aligned} K_c &= \alpha K_{tree}(R_1, R_2) + K_{poly}(R_1, R_2) \\ &= \alpha K_{tree}(tf_1, tf_2) + K_{poly}(vf_1, vf_2) \end{aligned} \qquad (1)$$

$$\begin{aligned} K_c &= \alpha K_{tree}(R_1, R_2) \times K_{poly}(R_1, R_2) \\ &= \alpha K_{tree}(\mathrm{tf}_1, \mathrm{tf}_2) \times K_{poly}(\mathrm{vf}_1, \mathrm{vf}_2) \end{aligned} \qquad (2)$$

여기서 K_{tree}, K_{poly}은 파스트리 커널 및 지수 커널을 의미하고, α는 계수(coefficient)이다. R_1,R_2는 절의 의존관계 인스턴스를 의미하고, 파스트리 형태로 표현되어 있다. tf는 R_1, R_2에서 추출한 구조 정보(syntactic tree feature), vf는 벡터로 표현되어진 정적인 자질(static feature)을 나타내고 있다. 본 논문에서는 이들 각각의 조합에 대하여 각각 실험을 수행하고 최적의 조합을 구하였다.

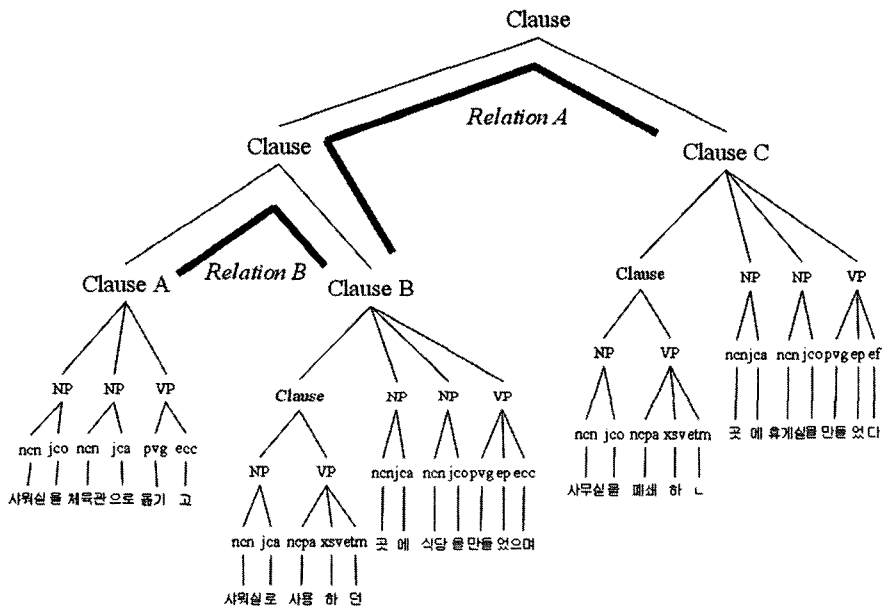
지수 커널은 SVM에서 사용하는 널리 사용된 커널이고. 벡터에서 나타날 수 있는 다양한 조합을 통해서 최적의 해를 구하는데 사용된다. 그리고 사용한 파스트

리 커널은 convolution 커널의 한 종류로 Collins(2001) 가 제안한 커널을 사용하였다. 파스트리 커널은 파스트리를 벡터로 표현하고, 이들의 내적을 계산함으로 두 파스트리간의 유사도를 측정한다.

절들 간의 의존관계 파악을 위한 자질 선택

절들 간의 의존관계 정의

한국어의 절들은 다른 절과 연결되는 연결절, 명사구를 수식하는 관형절 그리고



절 A	(C 샤워실을 체육관으로 옮기고)
절 B	(C (C 샤워실로 사용하던) 곳에 식당을 만들었으며)
절 C	(C (C 사무실을 폐쇄한) 곳에 휴게실을 만들었다.)

그림 1 절의 구성 및 의존관계 형성의 예

문장의 종료를 의미하는 종결절 등의 3종류로 나누어진다. 본 논문에서는 절들의 의존 관계 분석 대상을 연결절이 다른 절과 의존관계가 성립되는 관계만을 대상으로 삼았다. 그 이유는 관형절은 다음에 나타나는 명사구를 수식하고, 종결절은 항상 문장의 마지막에 등장함으로 의존 관계 파악에 용이하다. 그러나 연결절은 문장 속에서 나타나는 위치 등에 따라 의존 관계가 성립되기 보다는 어미의 변화와 문장 속에서 문맥에 따라 다양하게 의존 관계가 성립되어 의존관계 분석이 매우 어려운 작업이다. 그림 1은 한 문장에서 절의 구성과 절들 간의 의존관계의 예를 보여주고 있다. 이 문장에서는 모두 5개의 절을 가지고 있고, 이 들 중에 2개의 연결절, 즉 2개의 의존관계를 가지고 있다. 각 절의 구성을 살펴보면 절A와 절B가 하나의 의존관계를 형성하고 있고, 절A와 절B가 결합된 절이 절C와 결합되어있다. 그리고 절 B와 절C는 하위절로 관형절을 하나씩 포함하고 있다.

문법적 구성 자질(Syntactic Structure Feature)

본 연구에서는 절들 간의 의존관계 분석을 위하여 2개의 다른 시각에서 접근하였다. 하나는 절의 문법적 구성에 따른 접근이고, 나머지는 절을 구성하고 있는 어휘 및 거리에 따른 접근이다. 이중에서 절의 문법적 구성은 파스트리에서 자질을 추출하였다. 따라서 파스 트리에서 구조 정보를 추출하고, 표현하는 방법을 결정하는 일은 파스 트리를 사용하는데 있어서 아주 중요한 문제가 된다. 너무 많은 정보를 포함하게 되면 노이즈로 작용하여 성능을 저하시키고 너무 적은 정보를 가지고 처리하면 전반적인 성능의 저하를 불러온다. 이 문제는 절의 내부적 표현, 외부적 표현 그리고 부속절의 표현의 3개로 나누어 볼 수 있다.

- **절의 내부적 표현**

파스 트리에서 하나의 절은 그림 2와 같이 4개의 계층으로 구성되어 있다. 본 논문에서는 마지막 단어 계층을 제외한 절, 구, 테그(tag) 계층만을 사용하였다. 이유로는 word layer에 속하는 각 단어들의 데이터의 희소성의 발생 문제와 정적 자질에서 이 단어들을 사용을 들 수 있다. 그림 2의 예제는 그림1의 'Clause A'이고, 인스턴스 공간 표현에 사용된 각각의 계층을 보여주고 있다.

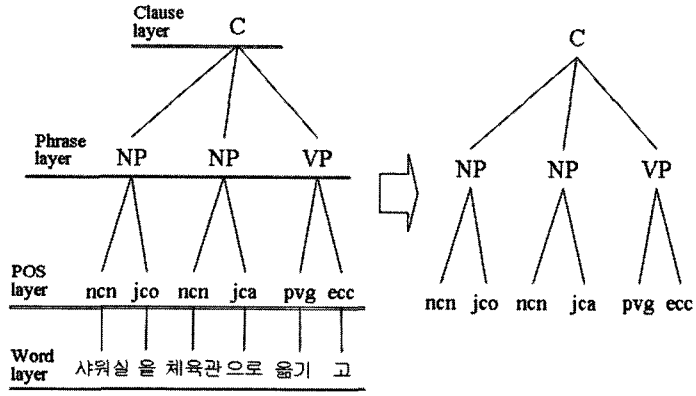
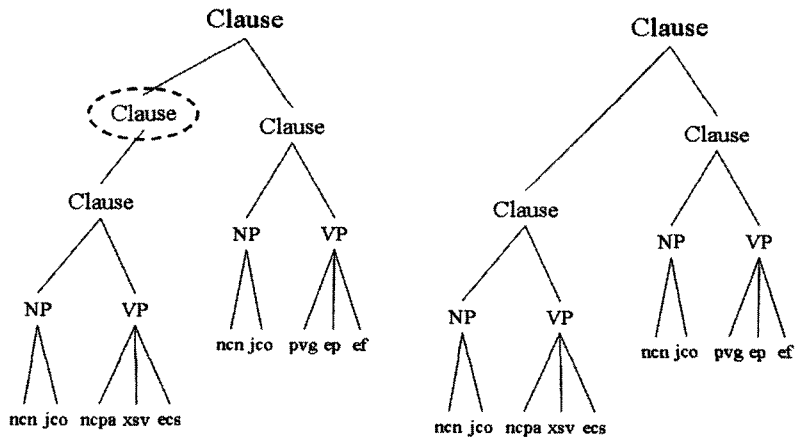


그림 2. 절의 내부적 표현

• **절의 외부적 표현**

절의 외부적 표현은 절들 간의 수직적인 관계를 표현하는 문제로 바꾸어 볼 수 있다. 즉, 두 절들 사이의 연결되는 단일 노드를 어떻게 표현할 것인가를 결정하는 것이다. 이 경우에는 모든 단일 노드를 사용하는 경우와 제거하는 경우로 나누어 각각의 경우에 대하여 각각 고려하였다(Zhang, 2006).



(a) Path-enclosed Tree (b) Flattened Path-enclosed Tree

그림 3. 절의 외부적 표현

(1) Path-enclosed Tree (PT)

- 트리로 표현된 2개의 절을 포함하고 가장 짧은 링크로 연결된 트리를 의미한다. 여기에서 단일 노드들을 모두 포함되어 있다. 그림 3-(a)를 말한다.

(2) Flattened Path-enclosed Tree (FPT)

- PT에서 절을 연결하는 노드들 중에서 중심이 되는 노드 하나를 제외한 모든 노드를 제거한 트리이다. 그림 3-(b)를 말한다.

• **부속절의 표현**

본 논문에서는 의존관계 대상을 연결절로 연결되는 절만을 대상으로 삼았다. 따라서 명사구를 수식하는 관형절은 처리대상에서 제외하였다. 그러나 하나의 연결절이 관형절을 포함하고 있을 때, 이들을 어떻게 표현하는 것인가에 따라 다음과 같이 나누어 고려하였다. 각각의 예는 그림 4의 (a),(b),(c)와 같다.

(1) Complete Tree (CT)

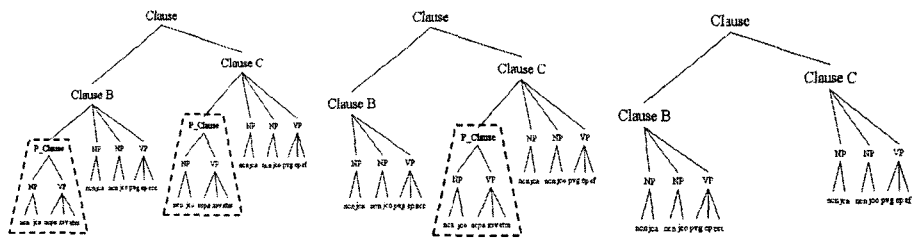
- 지배절 및 의존절에 각각 관형절을 사용 트리이다.

(2) Context-Sensitive Tree (CST)

- 지배절의 관형절만을 사용하는 것을 트리이다.

(3) Simple Tree (ST)

- 지배절 및 의존절에 나타나는 모든 관형절을 사용하는 트리이다.



(a) Complete FPT(CFPT) (b) Context-sensitive FPT(CSFTP) (c) Simple FPT(SFPT)

그림 4. 관계 인스턴스 공간 표현의 예

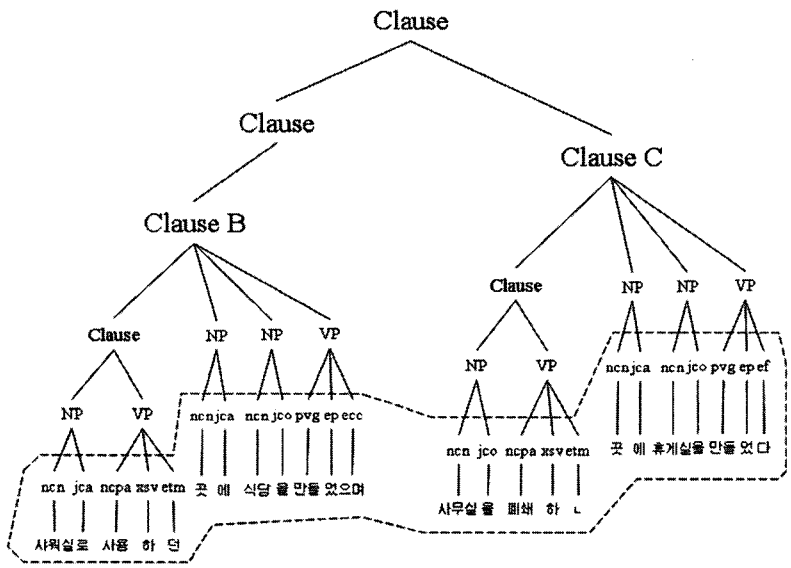


그림 5 정적 자질의 추출

절의 정적 자질(Static feature)의 사용

정적 자질은 그림 8과 같이 의존 관계에 있는 절들의 단어 및 POS 테그를 추출하고, 두 절들 간의 거리를 추출하여 벡터로 표현하여 사용하였다. 그림 8은 그림 5에서 절 B와 절 C의 의존 관계를 표현하고 있고, 이들 절간의 거리는 바로 이웃하고 있음으로 1로 나타낼 수 있다.

실험 및 분석

본 논문에서 사용한 말뭉치는 STEP2000과제의 결과물인 구문 구조 부착 말뭉치를 변형하여 만들었다. 전체 데이터에서 90%의 문장은 학습 데이터로, 나머지 10%는 실험 데이터로 사용하였다. 표 1은 사용한 코퍼스의 정보를 보여주고 있다.

본 논문에서는 파스 트리에서 표시되어 있는 절이 얼마나 정확하게 의존 관계대

표 1. 코퍼스의 구성 정보

분 류	학습데이터	실험데이터
문장의 수	6,240	694
모든 절의 수	24,226	2,650
관형절 및 종결절의 수	15,457	1,666
연결절의 수	8,769	984

상 절을 인식 및 판단하는 가를 측정하기 위해서 평가 측도로 정확도(accuracy)를 사용하였다. 절 단위는 한 문장에서 의존 관계 절을 정확하게 인식하는 것을 의미하고, 문장 단위는 전체 말뭉치에서 모든 의존관계가 올바르게 파악된 문장의 비율을 의미한다. SVM은 SVM Light를 사용하였고(Joachims, 2002) 사용된 파라메타는 파스트리 커널(λ=0.4), SVM(C=0.4)을 사용하였다. 다항널의 지수(dgree)는 2를 적용하였다.

복합 커널의 결합 계수에 따른 성능의 변화

두 개의 커널을 조합에 따른 성능의 변화를 살펴보았다. 절 외적 표현으로는 PT

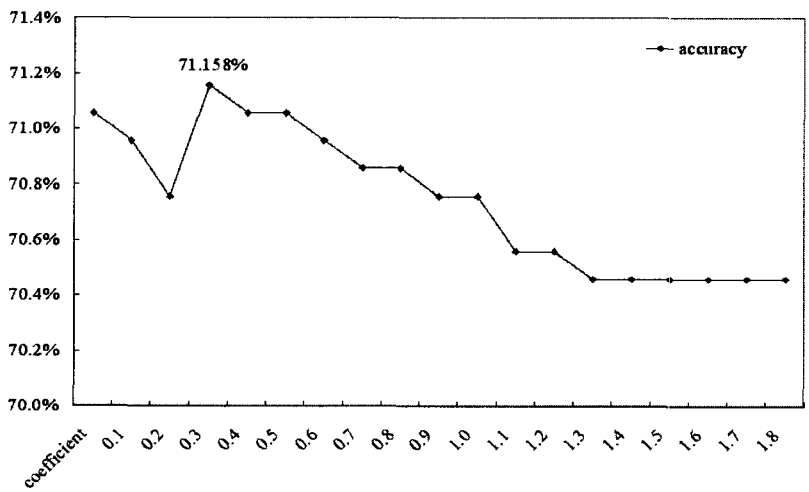


그림 6. 선형 조합에서 계수(coefficient) 변화에 따른 정확도

를 관현절의 표현은 ST를 사용하여 실험하였고, 사용한 커널의 조합은 식(1)을 사용하였다. 계수의 값을 0.1 만큼 변화하면서 실험을 수행하였다. 그 결과로는 값이 0.3일 때 가장 높은 결과를 얻었고, 전반적인 성능의 변화는 그림 6과 같다. 이는 절들 간의 의존관계 형성에 문법적 구조 정보가 약 30%, 어휘 및 거리정보가 약 70%의 영향을 미친다고 볼 수 있고, 본 논문에서 문법적 구조 정보를 사용하는 것이 유용하다는 것을 의미한다.

커널 조합 방법에 따른 성능 변화

식(1), (2)으로 표현된 커널 조합에 따른 성능을 실험해보았다. 먼저 절 외적 표현으로는 PT를 관현절의 표현은 ST를 사용하여 실험하였다. 실험의 대상은 의존관계 대상을 인식, 인식된 절들을 대상으로 의존 대상 절 하나를 선택했을 경우와 각각 하나의 절을 선택했을 때 문장 전체가 정확하게 인식되는 경우를 각각 측정하였다. 실험에서 기준점은 나타난 절의 바로 다음 절을 의존관계의 대상으로 삼았을 경우를 의미한다. 표 2의 결과에 따르면 복합 커널 B가 A보다 높은 성능을 보였다. 여기서 복합커널 A는 식(1)에 따른 커널의 조합이고 계수 값은 앞선 실험에서 밝혀진 0.3을 사용하였고, 복합커널 B는 식(2)에 따른 결과 값이다. 여기서 커널의 다양한 조합에 따라서, 그 값이 조금씩 변화하는 것을 알 수 있다. 따라서 두 개 이상의 커널을 결합하여 사용할 때, 다양한 실험을 통해서 최적의 조건을 찾아야 한다는 것을 알 수 있다.

표 2. 커널 조합에 따른 결과

관계 인스턴스 공간	절단위(Accuracy) %
기준점(Base Line)	57.50
복합 커널 A	71.15
복합 커널 B	74.14

복합 커널과 다른 모델과의 성능 평과

본 논문에서 제안한 복합 커널과 다른 모델과의 성능을 비교하였고, 그 결과는 표 3과 같다. 이 결과에 따르면, 기준점, 파스트리 커널만 사용한 것과 자질기반 SVM 보다 복합커널이 전반적으로 더 높은 성능을 보임을 알 수 있다. 이는 하나의 정보보다 복합 커널을 사용하여 보다 많은 정보를 사용함으로 보다 높은 성능을 얻을 수 있음을 보여준다. 여기서 자질 기반 SVM은 복합 커널에서 정적 자질만을 추출하여 사용한 값이다.

특히 FPT기반의 복합커널이 PT 기반의 복합커널보다 더 놓은 성능을 보였다. 이유로는 절들 간의 외부 연결에 단일 노드가 노이즈로 작용함을 알 수 있다.

인스턴스 표현 공간에 따른 성능 평가

마지막으로 인스턴스 표현 공간에 따른 실험의 결과는 표 4와 같다. 이 실험은

표 3. 절의 외부적 연결 표현에 따른 성능 평가(단위:%)

관계 인스턴스 공간	인식률	절단위	문장단위
기준점(Base Line)	57.50	57.50	·
파스트리 커널(Kim, S-S, 2002)	89.12	61.89	62.19
Feature_based SVM	56.46	70.05	61.03
복합커널(PT)	98.51	68.37	64.36
복합커널(FPT)	91.41	74.14	72.62

표 4. 내부절의 표현에 따른 성능 평가(단위:%)

관계 인스턴스 공간	인식률	절단위	문장단위
기준점(Base Line)	57.50	57.50	·
complete tree	***98.62***	***83.31***	***78.13***
context-sensitive tree	98.43	82.12	77.55
simple tree	91.34	74.14	72.62

절 외부 표현은 FPT를 사용하였고, 문장 단위 평가는 값은 폭포수 묶음 모델(cascading chunk)을 사용하였다. 이 결과 complete tree의 경우에 가장 높은 성능을 보임을 알 수 있다. 그러나 context-sensitive tree와 큰 값을 보여주고 있지는 않아서, 유의미한 결과를 가지는지 더 보강 연구가 필요할 것으로 보인다.

마지막으로 표 3와 4에서 문장단위의 성능이 전반적으로 60%이상의 성능을 보이는데, 이는 실험대상의 문장이 대체로 1-2개의 의존 관계를 가지는 문장이 많고, 입력대상이 절 단위로 인식된 상황이어서 높은 것으로 보인다.

결론 및 향후 연구

본 논문에서는 한국어 문장에서 문법적 구조 정보와 정적 정보를 사용할 수 있는 복합 커널을 제안하고, 이 커널에 맞는 적합한 의존 관계 인스턴스 공간을 제안하였다. 여기에 사용된 방법은 구조 정보를 담당하는 파스트리 커널과 정적 정보를 담당하는 지수 커널을 결합하였고, 최적화된 관계 인스턴스 설정을 통하여 SVM을 사용하여 절들 간의 의존관계를 파악하였다.

최적화된 관계 인스턴스 공간을 결정하기 위해서 절들 간의 연결을 PT와 FPT 계열로 나누었고, 내부절을 CT, CST, ST로 나누어 각각 실험하고, 그 성능을 보였다. 그리고 결정된 인스턴스 공간에서 복합 커널을 사용한 본 연구의 방법이 좋은 성능을 발휘함을 보였다. 향후 연구에는 코퍼스 외부 자원(의미정보)을 활용하고, 다양한 자질 선택 및 커널의 개선을 통하여 성능을 개선하는 많은 연구가 있어야 할 것이다.

참고문헌

서광진, (1993). 어절 사이의 의존관계를 이용한 한국어 구문 분석기, 한국 과학기술원 석사학위 논문, 1993.

Bunescu, R. C. & J. Mooney, R. (2005). A Shortest Path Dependency Kernel for Relation

Extraction, *In Proceedings of EMNLP'2005*, PP.724-731.

Carreras, X. & Marquez, L. (2001). Boosting Trees for Clause Splitting, *In Proceedings of CoNLL' 2001*, pp.73-75.

Chung, H. (2004). Statistical Korean Dependency Parsing Model based on the Surface Contextual Information, 고려대학교 박사학위 논문.

Collins, M. & Duffy, N. (2001). Convolution Kernels for Natural Language, *In Proceedings of NIPS' 2001*, pp.625-632.

Joachims, T. (2002). Learning to Classify Text Using Support Vector Machines, *Dissertation,* Kluwer.

Kim, S-S. Park, S-B. & Lee,S-J. (2007). Analyzing Depen- dencies of Korean Subordinate Clauses using parse tree kernels, *In Proceedings of CICLing 2007*, pp. 218 - .228.

Kudo, T. & Matsumoto, Y. (2000). Japanese Dependency Structure Analyisis Based on Support Vector Machines, *In Proceedings of joint Sigdata Conference on Empirical Methods in Natural Language Processing And Very Large Corpora,* pp.18-25.

Kudo, T. & Matsumoto, Y. (2002). Japanese Dependency Analyisis using Cascaded Chunking, *In Proceedings of the ACL'2002*, pp.1-7.

Lee, H.-J. Park, S.-B. Lee,S.-J. & Park, S.-Y (2006). Clause Boundary Recognition Using Support Vector Machines, *In Proceedings of PRICAI'2006*, pp. 505-514.

Molina, A. & Pla, F. (2001). Clause Detection using HMM, *In Proceedings of the CoNLL' 2001*, pp.73-75.

Uchimoto, K. Sekine, S. & Isahara, H. (1999). Japanese Dependency Structure Analysis Based on Maximum Entropy Models, *Proceedings of EACL'1999*, pp.196-203.

Zelenko, D. Aone, C. & Richardella, A. (2003). kernel methods for relation extraction, *Journal of Machine Learning Research*, no.2, pp.1083-1106.

Zhang, M. Zhang, T. Su, J. & Zhou, G.-D. (2006). A Comp- osite Kernel to Extract Relations between Entities with Both Flat and Structured Features, *In Proceedings of COLING- ACL'2006*, pp. 825-832.

1 차원고접수 : 2008. 3. 5
최종게재승인 : 2008. 3. 10

(*Abstract*)

Analyzing dependency of Korean subordinate clauses using a composit kernel

Sang-Soo Kim Seong-Bae Park Sang-Jo Lee Se Young Park

Department of Computer Engineering, Kyungpook National University

Analyzing of dependency relation among clauses is one of the most critical parts in parsing Korean sentences because it generates severe ambiguities. To get successful results of analyzing dependency relation, this task has been the target of various machine learning methods including SVM. Especially, kernel methods are usually used to analyze dependency relation and it is reported that they show high performance. This paper proposes an expression and a composit kernel for dependency analysis of Korean clauses. The proposed expression adopts a composite kernel to obtain the similarity among clauses. The composite kernel consists of a parse tree kernel and a liner kernel. A parse tree kernel is used for treating structure information and a liner kernel is applied for using lexical information. The proposed expression is defined as three types. One is a expression of layers in clause, another is relation expression between clause and the other is an expression of inner clause. The experiment is processed by two steps that first is a relation expression between clauses and the second is a expression of inner clauses. The experimental results show that the proposed expression achieves 83.31% of accuracy.

Keywords : Korean Clauses dependency, SVM, composit kernel, parse tree