복합 커널을 사용한 한국어 종속절의 의존관계 분석

김상수○ 박성배 이상조
경북대학교 컴퓨터공학과
{sskim, sbpark, sjlee}@sejong.knu.ac.kr

Analyzing Dependencies of Korean Subordinate Clauses using a Composite Kernel

Sang-Soo Kim○ Seong-Bae Park Sang-Jo Lee
Department of Computer Engineering
Kyungpook National University

1. 서론

절의 의존관계를 파악하는 일은 표면적으로 나타나는 정보만을 가지고 처리할 수 없고의미 정보 같은 추가적인정보가 필요할 것으로 판단하고 처리해왔다 그러나 최근 연구에서는 추가적인 정보를 사용하지 않고 여러가지 관점을 가지는 커널들을 결합하여 다양한 문제에 접근하고 있다본 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 선형 커널을 결합한 복합 커널을 제안하고 의존 관계를 잘 표현하는 다양한 인스턴스 공간을 제시한다. 그리고 제안된 커널과 인스턴스 공간을 활용한SVM을 사용하여 하위절의 의존관계를 파악하였고, 파스트리 및 자질 기반 방법보다 높은 성능을 발휘하는 것을 보인다

2. 절들 간의 의존 관계 파악을 위한 복합 커널

본 논문에서는 하위절의 의존관계 파악을 위해 파스트리 커널과 선형커널을 복합적으로 사용하는 복합 커널을 제안한다. 아래의 식과 같이 파스트리 커널과 선형 커널을 조합한 복합 커널(composite kernel)을 다음과 같이 구성하여 사용하였다.

$$K_C(R_1,R_2) = \alpha K_P(R_1,R_2) + (1-\alpha)K_L(R_1,R_2)$$
$$= \alpha K_P(tf_1,tf_2) + (1-\alpha)K_L(sf_1,sf_2)$$

여기서 K_P, K_L은 파스트리 커널 및 선형 커널을 의미하고, α는 계수(coefficient)이고 0.3을 사용하였다. R_1,R_2는 절의 의존관계 인스턴스를 의미하고, 파스트리 형태로 표현되어 있다. tf는 R_1, R_2에서 추출한 구조 정보(syntactic tree feature), sf는 정적인 자질(static feature)을 나타내고 있다

4. 한국어 종속절의 의존 관계 분석

본 논문에서는 절들의 의존 관계 분석 대상을 연결절이 다른 절과 의존관계가 성립되는 관계만을 대상으로 삼았다. 즉, 연결절만을 대상으로 삼았는데, 그 이유는 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 작업이기 때문이다. 본 논문에서 제안한 복합 커늘을 사용하기 위한 자질은 문법적 자질과 정적 자질로 나누어진다.

4.1 문법적 자질

문법적 자질은 절의 내부적 표현, 절의 외부적 표현, 부속절의 표현으로 나누어진다.

• 절의 내부적 표현

의존 대상이 되는 절을 어떻게 표현할 것인가를 나타낸다. 그림 1과 같이 총 4개의 계층에서 상위 3개의 계층 clause, phrase, POS layer만을 사용하였다

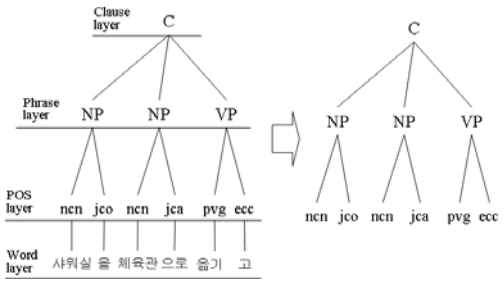

그림1. 절의 내부적 표현

• 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.

1) Path-enclosed Tree (PT)

- 트리로 표현된 2개의 절을 포함하고 가장 짧은 링크로 연결된 트리를 의미하고, 그림 2의 (a),(b),(c)을 말한다.

2) Flattened Path-enclosed Tree (FPT)

- PT에서 절을 연결하는 노드들 중에서 중심이 되는 노드 하나를 제외한 모든 노드를 제거한 트리이다. 그림 2의 (d),(e),(f)을 말한다.

• **부속절의 표현**

관형절을 어떻게 표현하는 것인가에 따라 다음과 같이 나누어 고려하였다.

1) Complete Tree (CT)

- 지배절 및 의존절에 각각 관형절을 사용하는 것이고, 그림 2의 (a)(d)가 있다.

2) Context-Sensitive Tree (CST)

- 의존절의 관형절을 사용하지 않고, 지배절의 관형절을 사용하는 것을 의미하고, 그림 2의 (b)(e)는 CST를 보여주고 있다.

3) Simple Tree (ST)

- 지배절 및 의존절에 나타나는 모든 관형절을 사용하지 않음을 의미한다. 그림 2의 (c)(f)는 ST의 예들 보여주고 있다.

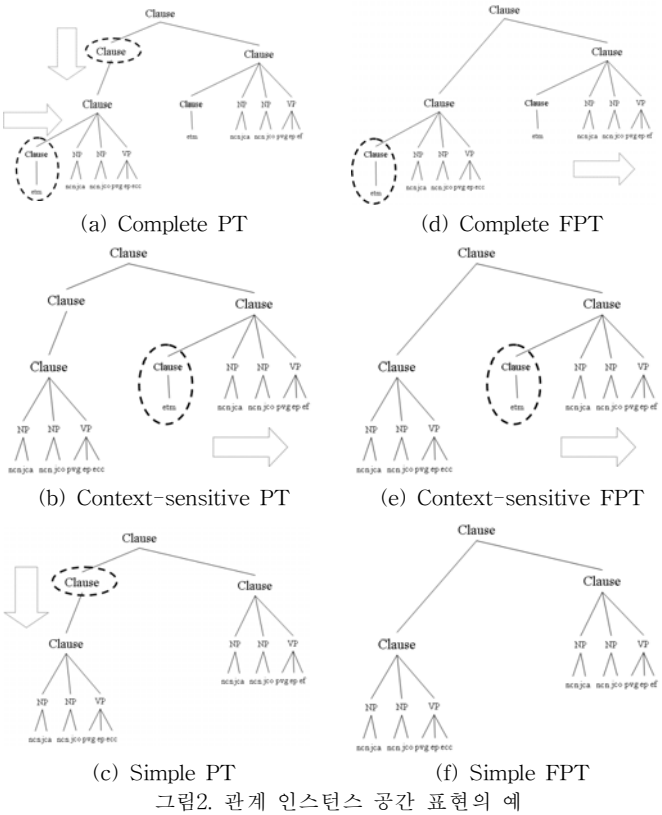

그림2. 관계 인스턴스 공간 표현의 예

4.2 절의 정적 자질(static feature)의 사용

정적 자질은 그림 3과 같이 의존 관계에 있는 절들의 단어 및 POS 테그를 추출하고, 두 절들 간의 거리를 추출하여 벡터로 표현하여 사용하였다.

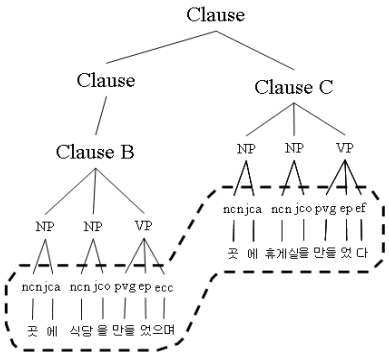

그림 3. 정적 자질의 추출

5. 실험 및 결론

본 논문에서 사용한 말뭉치는 STEP2000과제의 결과물인 구문구조 부착 말뭉치를 변형하여 만들었다. SVM은 SVM Light를 사용하였고, 사용된 파라메타는 λ=0.4, C=0.4을 사용하였다.

표 2에서는 FPT기반의 복합커널이 PT 기반의 복합커널보다 더 높은 성능을 보임을 보이고, 표 3에서는 FPT이고 context-sensitive인 경우 가장 높은 성능을 얻을 수 있었다. 그러나 complete인 경우와 값의 차가 너무 작아서 유의미한 결과를 가지는지 더 보강 연구가 필요할 것으로 보인다.

6. 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. 향후 연구에는 코퍼스 외부 자원(의미정보, 시소러스)을 활용하고, 다양한 자질 선택 및 커널의 개선을 통하여 성능을 개선하는 많은 연구가 있어야 할 것이다.

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

표 2. 외부적 연결 표현에 따른 성능 평가(단위:%)

관계 인스턴스 공간	인식률	절단위	문장단위
기준점(Base Line)	57.50	57.50	-
파스트리 커널[10]	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**	**78.14**	**76.62**

표 3. 내부절의 표현에 따른 성능 평가(단위:%)

관계 인스턴스 공간	인식률	절단위	문장단위
기준점(Base Line)	57.50	57.50	-
complete FPT	98.60	81.73	77.08
context-sensitive**FPT**	***98.50***	***82.12***	***77.55***
simpleFPT	91.41	78.14	76.62