CRFs를 이용한 의존구조 구문 레이블링

정석원○, 최맹식, 김학수
강원대학교, IT대학, 컴퓨터정보통신공학전공
{nlpsw, nlpmschoi, nlpdrkim}@kangwon.ac.kr

Labeling Dependency Structures using CRFs

Seokwon Jeong○, Maengsik Choi, Harksoo Kim
Program of Computer and Communication Engineering,
College of Information Technology, Kangwon National University

요 약

본 논문에서는 의존구조 분석 결과로부터 구문 레이블을 생성하는 방법을 제안한다. 제안 시스템은 의존구조 분석 결과의 의존소-지배소 쌍에 대해 자질을 생성하고, 문장 단위로 CRFs를 이용하여 구문 레이블을 부착한다. 실험을 통해 90.8%의 정확도를 보였고, 구문 레이블이 없는 의존구조 시스템의 후처리로 사용 가능하다.

주제어: 구문 레이블, 의존 구조 레이블

1. 서론

구문 분석은 문장의 구조를 분석하는 것으로, 크게 구구조 분석과 의존구조 분석으로 나누어진다. 한국어의 경우, 영어와는 달리 비교적 자유로운 어순과 문장 구성 성분의 빈번한 생략 등으로 인해 의존구조 분석이 더 적합하다[1,2]. 의존구조 분석은 문장을 성분의 조합으로 간주하는 구구조 분석과는 달리, 문장을 구성하는 어절들의 의존관계를 파악하는 것으로, 각 어절의 지배소를 찾는 과정이다.

오진영 외[3]는 구문 레이블을 CRFs(Conditional Random Fields)를 이용하여 부착하고 의존구조 생성에 사용하였다. 그러나 일부 의존구조 분석 시스템은 구문 레이블을 사용하지 않으므로 구문 레이블이 생성되지 않는다[4].

본 논문에서는 의존구조 분석 결과로부터 구문 레이블을 생성하는 방법을 제안한다. 본 논문의 구성은 다음과 같다. 2장은 본 논문에서 제안하는 시스템을 설명하고 3장은 실험 및 분석, 4장에서 결론과 향후 연구 방향을 제시하며 끝을 맺는다.

2. 제안 시스템

본 논문에서 제안하는 시스템은 [그림 1]과 같은 의존구조 분석 결과를 입력으로 하여 구문 레이블을 부착하는 것이다.

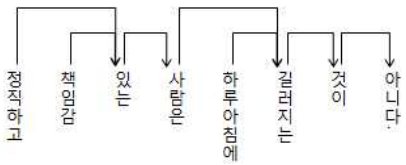

그림 1. 의존구조 분석 결과

[그림 1]의 문장으로부터 구문 레이블 부착 대상이 되는 의존소-지배소 쌍에 대하여 [표 1]의 자질을 추출한다.

표 1. CRFs 자질 예

자질	구문레이블
1:하 2:XSA 3:고 4:EC 5:있 6:VV 7:는 8:ETM 9:EC-ETM	기타
1:책임감 2:NNG 3:- 4:- 5:있 6:VV 7:는 8:ETM 9:NNG-ETM	주어
1:있 2:VV 3:는 4:ETM 5:사람 6:NNG 7:은 8:JX 9:ETM-JX	체언 수식어
1:사람 2:NNG 3:은 4:JX 5:지 6:VX 7:는 8:ETM 9:JX-ETM	주어
1:하루아침 2:NNG 3:에 4:JKB 5:지 6:VX 7:는 8:ETM 9:JKB-ETM	용언 수식어
1:지 2:VX 3:는 4:ETM 5:것 6:NNB 7:이 8:JKC 9:ETM-JKC	체언 수식어
1:것 2:NNB 3:이 4:JKC 5:아니 6:VCN 7:다 8:EF 9:JKC-SF	보어
1:아니 2:VCN 3:다 4:EF 5:아니 6:VCN 7:다 8:EF 9:SF-SF	기타

[표 1]에서 각각의 자질은 [표 2]와 같다. 의존구조는 어절 단위로 이루어지므로 어절을 내용어와 기능어로 분리하고 자질을 추출한다. 내용어는 21세기 세종계획 형

* 이 논문은 2013년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(2013R1A1A4A01005074) 또한 본 연구는 지식경제부 및 한국산업기술평가관리원의 산업융합원천기술개발사업(정보통신)의 일환으로 수행하였음. [10041678, 다중영역 정보서비스를 위한 대화형 개인 비서 소프트웨어 원천 기술 개발]

태소 태그에서 용언, 체언, 수식언, 독립언에 해당하며, 기능어는 관계언, 의존형태, 기호에 해당한다. 그리고 의존형태 중에서 접미사는 내용어에 포함된다.

표 2. 자질 집합

자질 번호	자질
1	의존소 내용어 어휘
2	의존소 내용어 품사
3	의존소 기능어 어휘
4	의존소 기능어 품사
5	지배소 내용어 어휘
6	지배소 내용어 품사
7	지배소 기능어 어휘
8	지배소 기능어 품사
9	의존소, 지배소의 마지막 품사

[표 1]과 같은 자질을 이용하여 CRFs를 모델을 생성한다. CRFs는 조건부 확률을 최대로 하는 무방향성 그래프 모델이다[5]. 입력열 $X = x_1 x_2 \ldots x_n$, 상태열 $T = t_1 t_2 \ldots t_n$가 주어지고, 가중치 $\Lambda = \{\lambda, \ldots\}$가 주어졌을 때, CRFs에서는 조건 확률로 식 (1)과 같이 정의된다.

$$P(T|X) = \frac{1}{Z(X)} exp(\sum_{i=1}^{n} \sum_{k} \lambda_k f_k(t_{i-1}, t_i, x, i)) \qquad (1)$$

여기서 $Z(X)$는 확률 값으로 만들어 주는 정규화 값이고, $f_k(t_{i-1}, t_i, x, i)$는 자질 함수이다. 또한 λ_k는 각 자질에 대한 가중치를 나타낸다. k는 k번째 자질이며, 자질 함수는 현재 시간 i에 대한 관측열 x, 상태 변이 $t_{i-1} \rightarrow t_i$에 대해서 전이의 향상을 측정할 수 있다. 매개변수들은 주어진 입력 열과 이에 대응하는 상태열에 대한 조건부 확률이 최대화하는 최대 우도(maximum likehood)에 의해서 추정된다[3]. 본 연구에서는 [6]의 CRFs를 사용하였다.

3. 실험 및 분석

3.1 실험환경

실험에서는 21세기 세종계획 구문 말뭉치를 사용하였다. 구구조 말뭉치를 의존구조 말뭉치로 변경하였고, 구문 레이블 부착 실험을 위해서 구문 분석 태그 중 기능태그만 사용하였다. 기능태그가 없는 어절은 ETC로 분류하였다. 전체 71,091문장(844,766어절) 중 구문 레이블을 학습하기 위해서 66,454문장(778,938어절)을 사용하였고 평가를 위해서 4,637문장(65,828어절)을 사용하였다.

실험 결과의 측정을 위하여 정확률을 사용하였다. 평가 척도는 식 (2)와 같다.

$$\text{정확률} = \frac{\text{시스템이 올바르게 추출한 정답 수}}{\text{시스템이 추출한 정답 수}} \qquad (2)$$

3.2 실험 및 분석

제안한 시스템의 성능 평가를 위해 구문 레이블별로 정확률을 측정하였다. 측정한 정확률은 [표 3]과 같다.

표 3. 성능

구문 레이블	정확률	빈도
주어	94.85	7,625
목적어	95.30	5,829
보어	79.61	1,373
체언 수식어	99.66	12,353
용언 수식어	93.32	7,704
접속어	90.69	1,278
독립어	75.00	32
기타	97.97	29,634
전체	90.80	65,828

[표 3]에서 전체 정확률은 90.80%이지만, 구문 레이블의 빈도에 따라 성능 차이가 크게 나타남을 알 수 있다. 특히 독립어의 빈도는 32, 정확률은 75%로 평균에 비해 매우 낮은 성능을 보였다.

4. 결론

본 논문에서는 의존구조 분석 결과로부터 구문 레이블을 부착하였다. 실험을 통해 한국어 문장에서 구문 레이블의 빈도에 따른 성능의 차이를 확인하였다.

향후 연구로서는 빈도가 낮은 구문 레이블의 성능을 향상시킬 수 있는 자질에 대한 추가 연구를 통해 전체 성능을 향상 할 수 있을 것이라 기대한다.

참고문헌

[1] J. Nivre, “An efficient algorithm for projective dependency parsing”, In Proc. of IWPT, pp.149-160, 2003.

[2] 김성용, 이공주, 최기선, “복합 레이블을 적용한 한국어 구문 규칙”, 한국정보과학회논문지: 소프트웨어및 응용, 제 31권, 제 2호, pp.235-244, 2004년 2월.

[3] 오진영, 차정원, “다단계 구단위화를 이용한 고속 한국어 의존구조 분석”, 한국시뮬레이션학회 논문지, Vol19, No.1, pp.103-111, 2010년 3월.

[4] 이용훈, 이종혁,. “SVM을 이용한 결정적 한국어 구문분석”, 한국어정보학 제10권 제2호, pp.7-14, 2008년 12월.

[5] J. Lafferty, A. McCallum, F. Pereira, “Conditional random fields: Probabilistic models for segmenting and labeling sequence data”, Proceedings. 18th International Conference on Machine Learning, pp.282-289, 2001.

[6] McCallum, Andrew Kachites., “MALLET: A Machine Learning for Language Toolkit”, http://mallet.cs.umass.edu. 2002.