한국어 확률 의존문법 학습

최선화[O], 박혁로
전남대학교 전산학과
{shchoi[O],hrpark}@dal.chonnam.ac.kr

Probabilistic Dependency Grammar Induction

SeonHwa Choi[0], Hyukro Park
Dept. of Computer Science, Chonnam National University

요 약

본 논문에서는 코퍼스를 이용한 확률 의존문법 자동 생성 기술을 다룬다. 의존문법 생성을 위해 구성성분의 기능어들 간의 의존관계를 학습했던 기존 연구와는 달리, 한국어 구성성분은 내용어와 기능어의 결합 형태로 구성되고 임의 구성성분 기능어와 임의 구성성분 내용어간의 의존관계가 의미가 있다는 사실을 반영한 의존문법 학습방법을 제안한다. KAIST의 트리 부착 코퍼스 31,086문장에서 추출한 30,600문장의 Tagged Corpus을 가지고 학습한 결과 초기문법을 64%까지 줄인 1,101개의 의존문법을 획득했고, 실험문장 486문장을 Parsing한 결과 73.81%의 Parsing 정확도를 보였다.

1. 서 론

문장의 통사 구조는 문장을 구성하는 단어들이 더 큰 의미를 이루기 위해 결합해가는 과정을 나타낸다. 구문 분석이란 이러한 문장의 통사 구조를 밝혀내는 과정이다. 효과적인 구문 분석을 위해서는 해당 언어를 잘 기술하는 언어 규칙의 집합인 문법이 필요하다. 문법획득은 언어 전문가에 의해 수동적으로 이루어지거나 학습알고리즘을 통해 자동적으로 이루어 진다.

언어전문가에 의한 수동적 지식의 획득은 사람의 직관에 아주 가까운 정확한 문법을 얻을 수 있지만, 기술하고자 하는 언어 현상의 규모가 커지고,분야의 제한이 없어질수록, 수동적 지식의 획득에는 어려움이 많다. 또한, 문법의 확장 및 관리가 힘든 전형적인 지식획득 병목현상을 나타내는 어려운 작업으로 알려져 있다. 이의 대안으로, 코퍼스로부터 자동적으로 문법을 학습하기 위한 연구가 많이 시도 되어 왔다[1,2,3,4,5]. 이런 방식의 문법 학습은 지식의 획득 및 확장이 용이하고 습득된 통계정보로부터 문장 분석 결과의 적합성을 우선 순위화 할 수 있는 등 모호성 처리가 자연스러우며, 비 자연스러운 문장에 대해서도 분석에 실패하지 않고 나름의 보유지식에 비추어 최적의 결과를 내어준다.

90 년대에 들어서 의존구조 분석에 기반하며 자유 어순 언어를 분석할 수 있는 방법[6]이 제안된 후에, 한국어 분석 분야에는 의존구조 분석이 주목을 받고 있다. 의존문법은 문장 내의 임의의 두 단어 사이의 지배-피지배 관계를 정의함으로써 문법을 기술하므로, 구 구조문법에 비해 단어의 발생 순서에 민감하지 않으므로 빈번하게 일어나는 생략과 피지배소 단어들의 발생 순서 뒤바뀜에 효과적으로 대응할 수 있기 때문이다.

본 논문에서는 한국어 구성성분은 내용어와 기능어의 결합 형태로 구성되고, 임의 구성성분 기능어와 임의의 구성성분 내용어간의 의존관계만이 구문적으로 의미가 있다는 사실을 반영한 의존문법 학습방법을 제안한다.

2. 한국어 의존 관계

2.1 의존 문법

의존문법이란 문장을 구성하는 구성요소와 또 하나의 구성요소 사이의 의존관계를 파악함으로써 문장을 분석한다[7]. 의존관계는 두 구성요소 사이에 존재하는데, 이중 한 구성요소는 지배소(governor)가 되며 다른 한 구성요소는 의존소(dependent)가 된다. 의존문법에 의한 문장의 분석결과는 문장내의 가능한 모든 의존관계들의 집합이다. 의존관계에 있는 두 구성요소 중 지배소는 의미의 중심이 되는 요소를 말하며, 의존소는 지배소가 갖는 의미를 보완해 주는 요소를 말한다.

한국어에 있어서 어절 단위의 구문요소 사이의 의존성은 한 요소의 기능어 형태소와 다른 요소의 내용어 형태소에 의해 결정될 수 있다.

"영희는 예쁜 인형을 가지고 있다." [예문 1]
"학교에 갈 수 있다" [예문 2]

[예문 1]에서 '인형을'은 기능어 형태소인 목적격 조사 '을'에 의해서 동사 '가지고'의 의존소가 되고, 내용어인 명사 '인형'에 의해서 관형어 '예쁜'의 지배소가 된다.[예문 2]는 두 구문요소 '학교에'와 '갈 수 있다'를 고려해 보면, '갈 수 있다' 는 기능어 형태소인 'ㄹ_수_있/다'와는 관계없이 내용어 형태소인 동사 '가'가 방향이나 목적지의 의존소를 요구하고 있고, '학교에'는 기능어 형태소인 '에'만으로도 목적지를 나타내는 것을 알 수 있으므로, 두 구문요소의 의존관계가 가능함을 알 수 있다.

2.2 의존문법 자동 학습

의존문법은 구 구조문법과 달리, 문장을 구성요소(Constituents)들로 나누지 않고, 대신, 단어와 단어 사이를 연결하는 문법적인 관계를 구별함으로써 문장을 분석한다[6]. 이승미[5]는 이런 형태의 의존문법 학습을 시도했다. 의존관계 집합을 표현하는 단위요소로 완결-링크와 완결-링

크열을 정의하고 이를 이용하여, 단어간 의존관계의 확률값을 학습하였다. 단어간의 기능어들의 의존관계만을 고려하여 학습한 결과 의존관계 정확도는 62.82%로 나타났다. 하지만, 한국어는 임의 성분을 중심으로 앞에 나오는 성분의 기능어와 뒤에 나오는 성분의 내용어 사이의 관계가 바로 문법적 관계를 나타낸다는 특성을 가지고 있다. 따라서, 그것을 학습하는 것이 기본이 될 것이다. 최선화[8]는 형태소 단위의 학습방법으로 이승미[5]의 단어 간의 의존관계만을 학습하는 것과 달리 한국어 단어 내에 존재하는 의존관계까지 학습하는 방법을 제안했다. 그런데, 한 어절을 구성하고 있는 형태소들 사이의 의존관계란 형태소들 사이의 접속관계이기 때문에, 형태소 분석이 끝난 상태에서 어절을 구성하고 있는 형태소들 사이의 의존관계를 고려하는 것은 무의미한 일이다.

따라서, 본 논문에서는 한국어 구성성분은 내용어와 기능어의 결합 형태로 구성되고, 임의 구성성분 기능어와 임의의 구성성분 내용어간의 의존관계만이 구문적으로 의미가 있다는 사실을 반영한 의존문법 학습방법을 제안한다.

3. 확률 의존문법 학습

3.1 완결-링크, 완결-링크열

의존관계로 표현되는 문장구조는 기본적으로 두 단어 사이의 의존관계 정의에서부터 시작해 일련의 부분 단어열에 대하여 구문구조의 조건을 만족하는 의존관계 집합을 찾아 확장해 나가는 과정에서 찾게 된다. 즉, 부분 단어열에 대한 의존관계 집합을 표현하는 단위요소에 대한 정의를 필요로 한다. 본 논문에서는 완결-링크와 완결-링크열을 이용하여 단어간 의존관계의 확률값 학습과 문장의 구문구조 파싱을 한다.

완결-링크와 완결-링크열은 다음과 같이 정의된다. 단어열 $W_{i,j}$[1]에 대해서 형성된 하나의 의존관계 집합은 다음의 조건들을 만족할 때 완결-링크로 정의된다.

- 배타적으로 $w_i \rightarrow w_j$ 혹은 $w_i \leftarrow w_j$ 가 존재
- $j-i$ 개 의존관계로 구성
- 내부 단어들은 그 단어열 안에 지배소 단어를 갖음
- 의존관계의 교차, 순환이 없음

완결-링크는 방향성을 가지는데 이는 가장 바깥 의존관계의 방향에 의해 결정된다. 즉, $W_{i,j}$ 에 대한 완결-링크의 가장 바깥 의존관계가 $w_i \rightarrow w_j$ 이면 우향이고 $w_i \leftarrow w_j$ 이면 좌향이다.

완결-링크열은 같은 방향성을 가진 0 개 혹은 그 이상의 일련의 인접한 완결-링크들로 구성된다. 즉, 최소단위 완결-링크열은 한 단어로 구성된 부분 단어열에 정의되는 0 개의 연속된 완결-링크이다. 완결-링크열 역시 방향성을 가지며, 구성요소인 완결-링크들의 방향성에 의해 결정된다. 만일 좌향 완결-링크들로 구성된 완결-링크열이라면 그것은 좌향 완결-링크열이다.

앞으로 완결-링크와 완결-링크열을 위해 다음의 표기법이 쓰인다. L 은 완결-링크를 의미하고, S 는 완결-링크열을 의미한다. 아랫첨자 r 과 l 은 방향성을 의미하여, r 은 우향을, l 은 좌향을 의미한다. 즉, L_r 은 우향 완결-링크, L_l 은 좌향 완결-링크, S_r 은 우향 완결-링크열 ,그리고 S_l 은 좌향 완결-링크열을 나타낸다.

완결-링크 및 완결-링크열이라 함은 그것이 형성되는 부분 단어열에 대해서는 의존관계의 설정이 모두 완결되어 있어서 그 안의 단어들에 대해서는 더 이상 의존관계의 형성을 위한 고려가 필요치 않다는 의미이다.

3.2 어절의 내용-기능어 품사 결정

의존문법을 학습하거나 의존문법을 적용하기 위해서는 어절 단위 구문요소의 내용어와 기능어 품사를 결정해야 한다. 한국어 어절은 여러 개의 형태소들이 결합할 수 있으므로, 이것들로부터 어절의 구문적 특성을 적절히 표현할 수 있는 대표 품사를 결정하는 일은 간단하지 않다. 본 절에서는 어절의 형태소 구성으로부터 자립어 및 기능어의 품사를 결정하는 원칙들을 나열한다.

- 어절의 가장 왼쪽 형태소의 품사가 내용어 품사
- 어절의 가장 오른쪽 형태소의 품사가 기능어 품사
- 형태소 하나로 구성된 어절은 그 자체가 내용어 및 기능어 품사
- 쉼표, 느낌표, 마침표, 특수기호 등은 문장의 구조에 영향을 주므로 기능어 품사
- 관형사나 수사가 어절의 맨 앞에 나오면 그 다음에 오는 체언이 내용어 품사
- 기능적으로 같은 역할을 하는 것은 단순화
  - 외국어 "f"는 명사로 대치
  - 상태성 명사+"하/xsm " 는 형용사
  - 동작성 명사+"하/xsv " 는 동사
- 보조사류는 구문적 기능을 결정하지 못하므로, 보조사가 어절 끝에 오면 그 앞 형태소를 기능어 품사
- 연속된 명사어 열은 마지막 형태소를 내용어 품사

3.3 의존문법 확률모델

한 문장의 확률은 그 문장이 갖는 모든 의존 구조, D 의 확률의 합이다. 또, 한 의존 구문구조, D의 확률은 그 안에 포함된 모든 의존관계의 확률의 곱으로 근사화 될 수 있다.

$$p(W_{1,n}) = \sum_D p(D, W_{1,n}) \cong \sum_D \prod_{w_i \rightarrow w_j \in D} p(w_i \rightarrow w_j)$$

여기에서 $1 \le i \le n+1$ (EOS : End Of String)이고 $1 \le j \le n$ 이다. 임의의 $p(x \rightarrow y)$ 는 다음과 같이 추정된다.

$$p(x \rightarrow y) = p(y \mid x) = \frac{C(x \rightarrow y)}{\sum_z C(x \rightarrow z)}$$

따라서, $\sum p(x \rightarrow y) = 1$ 이 된다. 여기에서 V 가 어휘집합을 표현한다고 할 때, , , , 이고 , 이다. 그러면, 임의의 문장 $W_{i,j}$ 의 확률은 완결-링크와 완결-링크열의 관점에서 표현하면 다음과 같다.

$$p(W_{1,n}) = \sum_D p(D, W_{1,n}) \cong \sum_{S_l(1,EOS)} p(S_l(1, EOS))$$

3.4 학습 알고리즘

인사이드 확률들은 CYK 차트의 관점에서 상향식(bottom-up)으로, 좌에서 우로 계산된다. 아웃사이드 확률들은 하향식으로, 우에서 좌로 계산된다. 이때, 미리 계산된 인사이드 확률값을 이용한다. 학습은 다음과 같이 진행된다.

1. 초기 의존문법을 설정
2. 학습 코퍼스의 초기 엔트로피 계산

[1] 는 각 문장내 i번째, j번째 단어를 가리키고, 는 문장내 i에서 j까지의 단어열을 가리킨다.

3. 의존관계의 발생빈도수 재계산
4. 재 계산된 발생빈도수에 의거하여 의존관계의 확률 값 계산

$$p_{new}(w_x \rightarrow w_y) = \frac{C(w_x \rightarrow w_y)}{\sum_z C(w_x \rightarrow w_z)}$$

5. 학습 코퍼스의 엔트로피를 재계산한다.
6. 이전 엔트로피 - 새 엔트로피 $>\varepsilon$ 이면, 3 에서 5 까지의 과정을 반복한다.

3 단계에서 5 단계까지의 반복은 모든 의존관계의 확률값이 안정되거나, 혹은 학습 코퍼스의 엔트로피 값이 최소값으로 수렴할 때 까지 계속된다.

4. 한국어 확률 의존문법 학습 실험

본 장에서는 확률 의존문법 학습 실험 결과로서 의존문법의 파싱 정확도를 실험한다. 알고리즘은 KAIST 코퍼스의 트리 부착 코퍼스에서 추출한 한국어 문장 집합에 대해서 학습되고 실험되었으며, 앞 어절의 기능어 형태소와 뒤 어절의 내용어 형태소간의 의존관계를 다루었다.

한국어는 부분 자유 어순 언어로서 중심어 후위원칙 제약을 가지므로 한국어 의존구조에서는 오직 S_l, L_r 그리고 null-S_r만이 고려 대상이 된다.

재추정 알고리즘의 실험은 KAIST 의 트리 부착코퍼스 31,086 문장 중 30,600 문장의 태그 부착 문장을 추출하여 학습문장으로 구성하였고, 486 문장의 태그 부착 문장을 추출하여 실험문장으로 사용하였다. 학습과정이 반복됨에 따라 학습 코퍼스의 엔트로피(bits/word)가 점차로 감소되면서 안정적인 결과를 얻을 수 있었다. 초기 문법은 3,080 개의 의존관계로 구성되는데 학습한 결과 빈도수가 1 보다 작은 의존관계는 모두 제거 한 후 1,101 개의 의존문법을 얻었다.

학습된 문법의 파싱 정확도를 실험하기 위하여 실험대상으로 486 문장 트리부착 코퍼스를 사용하였다. 실험 코퍼스 문장에 대해서, 학습된 문법을 이용하여 n-최적해 파서로 가장 좋은 점수의 파스만을 추출한 뒤 이를 트리 뱅크 파스와 비교하였다. 파싱 정확도를 위한 비교 기준으로는 의존관계 정확도를 채택하여 의존관계 정확도를 분석한 결과, 실험 대상인 자동 학습된 문법의 의존관계 정확도는 73.81%로 기존 연구 보다 높게 나타남을 알 수 있다. 이 결과는 표 1에 표시하였다.

표 1. 실험 집합 평가

	본 논문	이승미[5]	최선화[8]
문장 수	486	409	350
문장 길이범위	2-25	3-21	2-25
의존관계정확도	73.81%	62.82%	69.77%
본 논문과 비교		+10.99%	+4.05%

5. 결론

본 논문에서는 한국어 확률 의존문법 자동 학습을 시도하였다. 기존 연구의 어절 기능어들 간의 의존관계만을 고려하여 학습했던 방법은 한국어 의존 규칙의 특징을 반영하지 않은 방법으로 구조적인 정보를 전혀 추출하지 못하는 방법이다. 한국어에 있어서 어절 단위의 구문요소 사이의 의존성은 한 요소의 기능어 형태소와 다른 요소의 내용어 형태소에 의해 결정될 수 있다. 따라서 그것들을 학습해야 정확한 한국어 의존문법을 획득할 수 있다. 또한, 한 어절을 구성하고 있는 형태소들 사이의 의존관계란 형태소들 사이의 접속관계이기 때문에, 이미 형태소 분석이 끝난 상태에서 어절을 구성하고 있는 형태소들 사이의 의존관계를 고려하는 것은 무의미한 일이다. 본 논문에서의 기능어 형태소와 내용어 형태소간의 의존관계를 학습하여 만들어진 의존문법의 정확도는 평균 73.81%로 나타났다. 이로서, 한국어 의존관계의 특성을 반영하여 의존문법을 학습하는 방법이 올바른 방법임을 알 수 있다.

확률 의존문법 학습은 여러 방향으로 확장될 수 있다. 먼저, 어휘 의존문법 학습을 실험해 볼 수 있다. 본 논문에서는 품사 간 의존문법의 학습 실험에 그쳤지만, 사실 의존문법은 의존관계가 품사 단위 사이가 아니라 어휘 단위 사이에 정의될 때, 더 효과적이고 의미 있는 의존문법이 될 수 있다. 품사는 어휘의 특성에 관한 많은 정보를 일반화 과정을 통해 잃어버리기 때문이다. 두 번째로 초기 확률값이 학습 결과에 미치는 영향에 대한 고찰이 필요하다. 따라서 알고리즘의 변화가 없다면, 될 수 있는 한 초기문법이 효과적이면 학습 결과도 나아질 것이다. 미리 알고 있는 품사정보를 이용해서 수작업으로 약간의 품사 간 지배/피지배 관계에 대한 선호도를 주고, 그 정보를 품사가 부착된 학습 코퍼스에 적용하면 단순하게 학습 코퍼스의 모든 단어간 의존관계를 가정하는 것보다 더 나은 초기 의존문법을 구성할 수 있을 것이다.

참고문헌

[1] G.Carroll and E.Charniak. "Learning probabilistic dependency grammars from labeled text". In Working Notes, Fall Symposium Series, AAAI, pages 25-31, 1992

[2] E.Black, J.Lafferty, and S.Roukos. "Development and evaluation of a broad-coverage probabilistic grammar of English-language computer manuals". In 30th annual Meeting of the Association for computational Linguistics, pages 185-192, 1992

[3] K. Lari and S.J.Young. "The estimation of stochastic context-free grammars using the inside-outside algorithm". computer Speech and Language, 4:35-56, 1990.

[4] F.Pereira and Y.Schabes. "Inside-outside reestimation from partially bracketed corpora". In 30th Annual Meeting of the Association for Computational Linguistics, pages 128-135, 1992

[5] 이승미,"확률 의존 문법 학습",한국과학기술원,박사논문, 1998

[6] M.A.Covington. "A Dependency Parser for Variable-Word-Order Languages". Technical Report AI-1990-01, The University of Georgia, 1990.

[7] 홍영국, 권혁철, "단일화와 챠트를 이용한 한국어 구문분석 시스템의 구현," 1993 한국정보학회 봄 학술발표 논문집, pp. 781~784, 1993.

[8] 최선화, 박혁로, "어절 내부 의존관계를 고려한 확률 의존문법 학습," 2001 한국정보처리학회 추계학술발표 논문집, 8권 2호, pp. 781~784, 2001.