2001 년 한국정보처리학회 추계 학술발표논문집 제 8 권 제 2 호

어절 내부 의존관계를 고려한 확률 의존 문법 학습

최선화*, 박혁로,
전남대학교 전산학과
e-mail : shchoi,hrpark@dal.chonnam.ac.kr

Probabilistic Dependency Grammar Induction using Internal Dependency Relation in Words

Seon-Hwa Choi*, Hyuk-Ro Park
Dept. of Computer Science, Chonnam National University

요 약

본 논문에서는 코퍼스를 이용한 확률 의존문법 자동 생성 기술을 다룬다. 특히 의존 문법 생성을 위해 확률 재추정 알고리즘을 의존문법생성에 맞도록 변형하여 학습하였으며 정확한 문법 생성 및 희귀데이터(Data Sparseness)문제 해결을 위해서 구성요소의 대표 지배소들 간의 의존관계 만을 학습했던 기존 연구와는 달리 구성요소 내부의 의존관계까지 학습하는 방법을 제안한다. KAIST 의 트리 부착 코퍼스 31,086 문장에서 추출한 25,000 문장의 Tagged Corpus 을 가지고 한국어 확률 의존문법 학습을 시도 하였다. 그 결과 초기문법을 10.97% 에서 23.73% 까지 줄인 2,349 개의 정확한 문법을 얻을 수 있었다. 문법의 정확성을 실험 하기 위해 350 개의 실험문장을 Parsing 한 결과 69.61%의 파싱 정확도를 보였다. 이로서 구성요소 내부의 의존관계 학습으로 얻어진 의존문법이 더 정확했으며, 희귀데이터 문제 또한 극복할 수 있음을 알 수 있었다.

1. 서 론

구문분석은 문장의 구조를 밝힘으로서 문장의 명확한 의미를 포착하는데 도움을 준다. 효과적인 구문분석을 위해서는 해당 언어를 잘 기술하는 언어규칙의 집합인 문법이 필요하다. 이러한 문법의 습득을 위한 기본적인 정보원으로서, 텍스트 코퍼스가 대량으로 구축되어 이용되고 있다. 이 코퍼스로부터의 문법의 획득은 언어 전문가에 의해 수동적으로 이루어지거나 학습알고리즘을 통해 자동적으로 이루어진다.

언어전문가에 의한 수동적 지식의 획득은 작은 규모의, 제한된 분야에 대한 문법 구축으로는 비교적 성공 가능성이 있으며, 사람의 직관에 아주 가까운 정확한 문법을 얻을 수 있다. 그러나 기술하고자 하는 언어 현상의 규모가 커질수록, 분야의 제한이 없어질수록, 수동적 지식의 획득에는 어려움이 많다. 또한, 수동적 문법 구축은 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[1,2,4,6,10]. 이는 코퍼스에 기반하여 통계적 정보를 이용한 문법 학습, 혹은 문법 추론으로, 하나의 연구 분야를 이루고 있다. 이런 방식의 문법 학습은 여러 가지 장점을 갖는다. 지식의 획득 및 확장이 용이하고 습득된 통계정보로부터 문장 분석 결과의 적합성을 우선순위화 할 수 있는 등 모호성 처리가 자연스러우며, 잘못된 언어현상이나 비 자연스러운 문장에 대해서도 분석에 실패하지 않고 나름의 보유지식에 비추어 최적의 결과를 내주어주는 등의 장점이 있다.

이렇게 자동적으로 학습된 문법은 그 자체로서 구문분석의 중요한 요소이기도 하지만, 또한 부분적으로라도 구문 정보를 필요로 하는 여러 기타 자연언어처리 응용 시스템에 적용될 수 있다.

지금까지 대부분의 코퍼스 기반 문법 자동 학습은 구구조문법 형식의 문맥자유문법 학습에 치중되어 왔다[2,4,7]. 구구조문법은 어순이 비교적 고정적인 영어와 같은 언어의 문법을 작성하는 데에 효과적으로 적용되어 왔다. 이와는 상대적으로, 한국어나 일본어, 터키어, 러시아어와 같이 (부분적으로)자유 어순의 성격을 가진 언어의 문법 기술에는 의존문법이 더 효과적

일 수 있다. 의존문법은 문장 내의 임의의 두 단어 사이의 지배-피지배 관계를 정의함으로써 문법을 기술하므로, 구구조문법에 비해 단어의 발생 순서를 덜 제약적으로 표현할 수 있어서 빈번하게 일어나는 생략과 피지배소 단어들의 발생 순서 뒤바뀜에 효과적으로 대응할 수 있기 때문이다.

본 논문에서는 코퍼스를 이용한 의존문법의 통계적 자동학습을 목표로 한다. 확률 파라미터 값을 재추정하는 알고리즘으로는 인사이드-아웃사이드 알고리즘[8]이 널리 사용되고 있다. 이 재추정 알고리즘은 학습 코퍼스의 확률값이 최대가 되도록 반복적으로 규칙의 확률값을 재조정 함으로써 문법을 학습하게 되며 이 같은 각 반복과정에서 학습 코퍼스의 확률 값을 낮추지 않는 것이 보장된다. 이를 의존문법 학습에 적용하기 위해서는 의존문법에 적합한 형태로 변형되어야 한다[10].

본 논문에서는 의존문법의 학습에 적합하도록 변형시킨 인사이드-아웃사이드 알고리즘[10]을 사용한다. 의존구조의 기본요소로서 비 구성요소적 단위인 완결-링크와 완결-링크열을 정의하고 그에 대한 인사이드-아웃사이드 확률식을 정의한다. 또한 구성요소의 대표 지배소 들간의 의존관계 만을 학습했던 기존 학습방법[10]과는 달리 구성요소 내부의 의존관계까지 학습하고 실험을 통해 얼마나 효과적인 의존문법이 학습되는지 보인다.

2. 관련연구

2.1 구구조 문법 자동 학습

구구조 문법 자동학습은 비교사 학습과 교사 학습으로 나뉘어 볼 수 있다.

Lari 와 Young[4]과 Chen[7]은 구구조 문법의 비교사 학습을 위해 인사이드-아웃사이드 알고리즘을 활용하였으며 그렇게 해서 얻어진 문법을 언어모델링에 적용하였다. Lari 와 Young 은 일차적으로 제한된 갯수의 비단말노드와 단말노드, 즉 단어 집합을 가지고 촘스키 정규형의 모든 가능한 규칙을 나열함으로써 초기 문법 규칙을 얻는다. 그리고 나서 인사이드-아웃사이드 알고리즘을 이용하여 초기 문법 규칙들의 확률값을 재추정하여 0 에 가까운 확률값을 가지는 규칙들을 제거한다. 이렇게 얻어진 학습된 문법을 언어모델링에 적용한 결과 엔그램 모델보다 성능이 덜 효과적인 결과를 보였다고 보고 되었다.

Chen 은 그와는 달리, 두 단계의 탐색을 거쳐 문법을 학습한다. 먼저, 시작 비단말노드와 단말 노드들로 구성된 기본 문법에서 시작하여, 베이지안(Bayesian)체계 안에서 일종의 그리디(greedy) 휴리스틱 탐색을 적용하여 규칙을 첨가해 나감으로써 최적의 촘스키 정규형의 문법을 추출하고, 두번째 단계로 이 문법에 인사이드-아웃사이드 알고리즘을 적용하여 각 규칙의 확률값을 재추정한다. 이 경우 역시 언어 모델링에 적용되었는데 인공적으로 만들어진 코퍼스에 대해서는 엔그램 모델이나, Lari 와 Young 의 모델에 비해 좋은 성능을 보였으나, 자연적인 코퍼스에 대해서는 엔그램이 가장 좋은 성능을 보였다고 보고되었다.

Black 등[2]과 Pereira 와 Shabes [6]는 이러한 구문 정보가 부가된 코퍼스와 인사이드-아웃사이드 알고리즘을 이용하여 구구조 문법의 교사학습을 한 경우이다.

Black 등[2]등은 트리뱅크와 문법을 수동 구축하게 하고,이렇게 구축된 문법의 확률 파라미터 학습은 문법에 의해 성성된 파스와 트리뱅크의 파스와 비교해서 서로 일관성이 있는 경우에만 문법 규칙의 확률값을 재조정 하는 방식으로 인사이드-아웃사이드 알고리즘의 적용을 변형하였다. 이렇게 학습된 확률 문맥자유 문법은 실험결과 같은 도메인의 실험 문장들에 대해 최적해 파스의 경우 평균 73-75%의 정확도를 보였다.

Pereira 와 Schabes[6]는 부분적인 괄호 매겨진 코퍼스가 있을 때, 이를 이용하면 단순한 텍스트 코퍼스로부터 비교사 학습하는 것 보다 유용한 문법을 학습할 수 있다고 보고 인사이드-아웃사이드 알고리즘을 확장하여 부분적으로 괄호 매겨진 코퍼스로부터 문법을 학습할 수 있도록 하였다. 알고리즘 확장은 학습코퍼스의 괄호 매겨진 부분에 대해서는 문법에 의한 분석결과와 괄호 교차가 일어나지 않는 경우에만 고려되고 괄호 매겨지지 않은 부분에 대해서 원래의 비교사 알고리즘 처럼 작동되도록 하였다. 실험 결과, 괄호 매겨진 코퍼스로 학습한 문법은 90.36%의 파싱 정확도를, 괄호 매겨지지 않은 코퍼스로 학습한 문법은 37.35%의 파싱 정확도를 가진 문법이 자동 학습 되었다.

2.2 의존 문법 자동 학습

의존문법은 크게 두 가지로 나누어 볼 수 있다. 하나는 단어 순서를 명시하는 것과 단어 순서에 무관한 것이다. Carroll[1]은 인사이드-아웃사이드 알고리즘을 이용하여 단어순서를 명시하는 의존문법의 자동학습을 실험하였다. 그가 사용한 문법 규칙의 형태는 다음과 같다.

$$\overline{X} \rightarrow \alpha X \beta ,$$

X 는 단말노드이고, α 와 β 는 비단말노드들의 열로서 빈 열일 수 도 있다. Carroll 은 초기 문법으로 아무것도 없는 빈 상태에서 시작한다. 한 문장씩 차례로 분석에 필요한 규칙을 첨가하여 그렇게 수정된 문법을 반복적으로 인사이드-아웃사이드 알고리즘을 이용하여 확률값 재추정을 하는 방식으로 문법 학습이 이루어진다. 이 연구는 결국 의존문법 학습이라기 보다는 제한된 구구조 문법 형식을 빌린 의존문법을 학습함으로써 문법 탐색 공간을 줄이는 효과를 얻고자 한 시도라고 볼 수 있다.

어순이 가변적인 경우의 언어를 위해서는 단어순서와 무관한 의존문법이 더 적합할 수 있다. 단어 순서와 무관한 의존문법은 다음과 같은 형식으로 규칙을 기술한다.

$$x \rightarrow y(f)$$

여기에서 x 는 y 의 지배소, y 는 x 의 피지배소라고 말하고 그 의존관계의 기능적 역할은 f 로 표현된다.

이와 같이, 의존문법 규칙은 오직 단어와 단어 사이의 지배-피지배 관계와 각 관계의 기능적 역할을 표현할 뿐이다. 의존문법은 구구조문법과 달리, 문장을 구성요소(Constituents)들로 나누지 않고, 대신, 단어와 단어 사이를 연결하는 문법적인 관계를 구별함으로써 문장을 분석한다[5]. 이러한 형태의 의존문법의 확률 파라미터의 재추정을 위한 시도는 이승미[10]에 의해 시도 된바 있다. 의존관계 집합을 표현하는 단위요소로 완결-링크와 완결-링크열을 정의하고 이를 이용하여, 단어간 의존관계의 확률 값 학습을 시도하였다. 구성요소의 대표 지배소 들간의 의존관계만을 고려하여 학습한 결과 의존관계 정확도는 62.82%로 나타났다. 하지만, 이 방법은 초기 문법 3,080 개로 학습을 시작하여 학습 후 78.18%가 줄어든 627 개의 문법을 얻었다. 복잡한 구문구조를 표현하기에는 문법의 갯수가 너무도 적다. 즉, 희귀데이터(Data Sparseness) 문제로 인해 정확도가 떨어질 우려가 있는 것이다. 이것은 구성요소의 대표 지배소 들 간의 의존관계만을 학습한 결과라고 할 수 있다.

따라서, 본 논문에서는 이승미[10]의 방법을 기반으로 구성요소 내부 의존관계까지 학습하여 좀 더 정확한 의존문법을 생성하고 희귀데이터 문제를 해결하기 위한 시도를 하였다.

3. 확률 의존문법 학습 알고리즘

확률 의존문법 재추정 알고리즘은 인사이드-아웃사이드 알고리즘[8]을 의존문법에 적합하게 변형시킨 것[10]이다. 본 장에서는 완결-링크와 완결-링크열에 대한 인사이드-아웃사이드 확률을 정의하고, 이를 기반으로 확률 의존문법을 재추정 알고리즘을 기술한다.

3.1 완결-링크, 완결-링크열

단어간 의존관계의 확률값을 학습하는 것은 이 완결-링크, 완결-링크열의 인사이드/아웃사이드 확률값 계산을 통하여 이루어진다.

완결-링크

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

완결-링크열

- 같은 방향성을 가진 0 개 혹은 그 이상의 일련의 인접한 완결-링크들로 구성.

3.2 학습 알고리즘

인사이드 확률은 CYK 챠트의 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 설정한다: 학습 코퍼스에서 모든 가능한 단어쌍을 나열하고 의존관계의 확률값을 초기화 한다.
2. 학습 코퍼스의 초기 엔트로피를 계산한다.
3. 학습 코퍼스를 분석하여 각각의 의존관계의 발생빈도수를 재계산한다.
4. 재 계산된 발생빈도수에 의거하여 의존관계의 확률값을 새로 계산한다.

$$p_{new}(w_x \rightarrow w_y) = \frac{C(w_x \rightarrow w_y)}{\sum_z C(w_x \rightarrow w_z)} \qquad (3.1)$$

5. 수정된 의존문법을 이용하여 학습 코퍼스의 엔트로피를 새로 계산한다.
6. 이전 엔트로피 − 새 엔트로피 $> \varepsilon$ 이면, 3 에서 5 까지의 과정을 반복한다.

학습 코퍼스가 주어지면 초기 문법은 학습 코퍼스에서 발생한 모든 유일한 단어들의 쌍으로 초기화 될 수 있다. 이 쌍들은 잠정적인 지배-피지배 관계를 표현한다. 초기 확률 값은 무작위적으로 주어질 수 있다. 학습은 이 초기 문법으로 부터 시작한다.

4. 한국어 확률 의존문법 학습 실험

본 장에서는 확률 의존문법 학습 실험 결과로서 의존문법 학습 알고리즘의 수렴과 학습된 문법을 보인다. 또한 학습된 문법과 초기 문법의 크기 비교를 통해 학습 정도를 보이고 문법의 파싱 정확도를 실험한다. 이 알고리즘은 C 언어로 구현되었으며 실험은 Window2000 에서 시행되었다. 알고리즘은 KAIST 코퍼스의 트리 부착 코퍼스에서 추출한 한국어 문장 집합에 대해서 학습되고 실험되었다. 실험은 어절 내부 의존관계까지 다루었으며, 품사 단위에 대해 이루어졌다.

한국어는 부분 자유 어순 언어로서 지배소인 어휘가 항상 피지배소인 어휘의 오른쪽에 위치하는 제약을 가진다. 이 같은 제약 때문에 한국어 의존구조에서는 오직 S_l, L_l, 그리고 null, S_r 만이 고려 대상이 된다. 한국어에 대한 추상적 의존구조를 그림 4.1 에서 보인다.

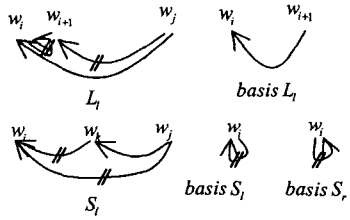

그림 4.1 : 한국어의 추상적 완결-링크와 완결-링크열

이승미[10] 연구는 각 어절의 기능적부분의 마지막 품사를 그 어절의 대표 품사로 보고 어절간 의존관계를 대표 품사간의 의존관계로 가정하였다. 본 논문에서는 어절 내부의 의존관계까지 고려하여 학습한다. 사용한 품사 집합, T 는 55 개의 품사로 구성되어 있다. 따라서 초기 문법은 T 의 모든 품사 쌍으로 구성되므로 3,080 개의 의존관계로 구성되었다. 초기 확률값은 치우침 없도록 모두 같은 값을 갖도록 설정되었다.

재추정 알고리즘의 실험은 KAIST 의 트리부착코퍼스 31,086 문장중 25,000 문장의 태그 부착 문장을 추출하여 학습문장으로 구성하였고, 350 문장의 태그 부

착 문장을 추출하여 실험문장으로 사용하였다.

표 4.1: 학습 코퍼스 엔트로피

반복	엔트로피
1	4.158206e+000
2	1.921697e+000
3	1.799868e+000
...	.....
104	1.618935e+000
105	1.618834e+000
106	1.618735e+000

표 4.1 에서 학습과정이 반복됨에 따라 학습 코퍼스의 엔트로피(bits/word)가 점차로 감소함을 보이고 있다.

표 4.2: 문법의 크기

	문법크기
초기문법	3,080
학습 후	2,742 (-10.97%)
Cut-off(freq. < 1.0)	2,349 (-23.73%)

표 4.2 에서 학습에 의한 문법의 크기 변화를 보인다. 초기 문법은 3,080 개의 의존관계로 구성되는데 학습한 결과 10.97%의 문법 크기 감소를 가져왔다. 확률값이 0 에 가까운 의존관계를 제거하여 걸렀을 경우 문법 크기와 문법의 효과에 미치는 영향을 보기 위해서 학습 후 빈도수가 1 보다 작은 의존관계는 모두 제거 하였다. 표 4.2 에서 보이듯이 23.73%의 감소를 가져왔다.

학습된 문법의 파싱 정확도를 실험하기 위하여 실험대상으로 350 문장 트리부착 코퍼스를 사용하였다. 실험 코퍼스 문장에 대해서, 학습된 문법을 이용하여 n-최적해 파서로 가장 좋은 점수의 파스만을 추출한 뒤 이를 트리뱅크 파스와 비교하였다. 파싱 정확도를 위한 비교 기준으로는 의존관계 정확도를 채택하였다. 의존관계 정확도는 분석결과에 있는 모든 의존관계에 대한 정확한 의존관계의 비율로 정의 된다.

$$\frac{\text{트리뱅크 의존관계와 일치하는 의존관계수}}{\text{분석결과의 의존관계 수}} * 100(\%)$$

실험대상인 자동 학습된 문법은 의존관계 정확도는 69.61%로 기존 연구보다 높게 나타남을 알 수 있다. 이 결과는 표 4.3 에 표시하였다.

표 4.3: 실험 집합 평가

	본 논문	이승미[10]
문장 갯수	350	409
평균 문장 길이 (단어)	13.2	11.4
문장 길이 범위	2-25	3-21
의존관계 정확도	69.61% (+6.79%)	62.82%

5. 결론

본 논문에서는 (부분) 자유어순 언어의 문법 자동 학습을 위해 확률 의존문법에 적용할 수 있는 변형된 확률 파라미터 재추정 알고리즘인 인사이드-아웃사이드 알고리즘을 사용하였다. 기존의 구성요소 대표 지배소간의 의존관계만을 고려했던 학습 방법은 복잡한 문장을 분석하기에 문법의 수가 충분하지 못하였고 희귀데이터문제를 해결할 수 없었다. 하지만 구성요소 내부 의존관계까지 고려하여 자동 학습된 문법은 2,349 개가 생성되었고, 학습 과정에서 보여지지 않은 실험문장에 대해서 평균 69.61%의 의존관계 정확도를 보여주었다. 이로서, 구성요소 대표 지배소간의 의존관계만을 학습한 방법보다 어절 내부 의존관계까지 학습하는 방법이 훨씬 더 정확한 문법을 생성해 주는 것을 알 수 있다.

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] E. Charniak. Statistical Language Learning. The MIT Press, 1993.

[4] 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.

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

[6] 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

[7] S.F.Chen. "Bayesian grammar induction for language modeling". In 33rd Annual Meeting of the Association for Computational Linguistics, pages 228-235, 1995

[8] F.Jelinek, J.D.Lafferty, and R.L.Mercer."Basic Methods of Probabilistic Context Free Grammars". Technical Report, IBM-T.J. Watson Research Center, 1990

[9] 이공주, "언어적 특성에 기반한 한국어의 확률적 구문분석", 한국과학기술원, 박사논문, 1998

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

[11] 최명석, "한국어 부분 구문 분석: 코퍼스로부터의 규칙 자동 추출",한국과학기술원, 석사논문, 1997