변환 규칙 학습기를 이용한 한국어 의존 구조 분석기*

이성욱, 서정연
서강대학교 전자계산학과 자연어처리 연구실

Dependency Structure Analysis System for Korean Using Automatically Acquired Transformation Rules

Song-Wook Lee, Jungyun Seo
Natural Language Processing Lab., Dept. of Computer Science, Sogang Univ.

요약

코퍼스 속의 언어적 규칙을 직접적으로 사용하여 한국어 의존 구조를 분석하기 위해, 본 한국어 의존 구조 분석기는 의존 구조가 나타나 있는 코퍼스로부터 변환 규칙 학습기로 규칙을 자동적으로 학습하고 그 규칙을 적용함으로써 한국어 의존 구조를 분석한다. 이를 위해 기존의 연구된 구구조 문법의 규칙 틀과는 다른 한국어 의존 구조에 맞는 규칙 틀을 연구하였고 또 의존 구조에서 발생할 수 있는 교차구조(Crossing structure)를 방지하는 연산을 고안하였다.

1. 서론

코퍼스를 이용하여 자연어 처리에서 발생하는 모호성을 해소하고자 하는 연구가 많이 이루어져 왔다. 특히 품사 태깅에 많이 이용되어왔는데 통계 자료 추출과 변환 규칙에 의한 품사 태거가 연구되었다. 언어 속의 문법을 추론하여 규칙을 자동적으로 학습하고 직접적으로 사용하는 방법인 변환 규칙 방법이 영어와 한국어의 품사 모호성해소에서 좋은 성능을 보였다[1,2,5,7]. 구문 분석에도 코퍼스의 통계정보를 사용하여 보다 정확한 결과를 내기 위한 방법이 많이 연구되었다. 또 근래에 품사 태깅에 사용한 변환 규칙 방법과 유사한 방법을 영어의 구문 분석에 사용하여 통계적인 방법인 inside-outside algorithm과 비교할 만한 성능을 보였다[3].

기존의 한국어 구문 분석은 많은 문법 규칙을 미리 정의해야 하고, 또한 그러한 규칙으로 분석을 했을 경우 여러 개의 결과가 나타나서 애매성을 해소하기 힘든 단점이 있다. 반면에 자동으로 학습된 변환 규칙을 이용하면 가장 적합하다고 생각되는 하나의 결과만을 찾아내어 효율적인 분석이 가능하고 학습된 변환 규칙 이외의 문법 규칙을 따로 정의할 필요가 없는 등 많은 장점을 가지고 있다. 그러나 영어의 구문 분석에 사용된 방법[2,3]은 구구조 문법에 기반 하였으므로 한국어의 의존 구조에 이 방법을 사용하기에는 적합하지 않다. 따라서 본 연구에서는 한국어의 의존 구조에 맞는 변환 규칙 틀을 연구하여 코퍼스에서 변환 규칙을 학습하였으며 변환 규칙을 적용할 때 발생할 수 있는 교차구조를 방지하기 위한 연산을 고안하였다. 본 연구에서는 Brill이 제안한 규칙 기반 학습 알고리즘[1]을 사용하여 구문 분석 규칙을 학습하며 학습된 규칙들을 이용하여 한국어 구문 분석에 이용한다.

2. 한국어 변환 규칙 학습기

변환 규칙 학습기를 설계할 때 고려할 사항은 다음과 같다.

초기 상태의 태거와 허용 가능한 변환 규칙 틀을 정의해야 한다.

변환된 코퍼스와 정답 코퍼스를 비교하여 변환 규칙을 선택하는 함수를 정의해야 한다.

변환 규칙 틀을 적용하기 위해 찾아보아야 하는 탐색의 범위를 결정해야 한다[1].

2.1 한국어 의존 관계

의존 관계는 어절의 형태에 따른 어절들 사이의 관계를 나타내는 것으로 의존 문법이론에 바탕을 두고

본 연구의 일부분은 한국 과학 재단의 특정 기초 연구인 "통계적 한국어 담화 분석"의 결과로 이루어졌습니다.

있다. 의존 문법은 구문단위 사이의 관계를 지배관계와 종속관계로 보아 문을 표현한다[5]. 그림 1은 "나는 밥을 먹었다"라는 문장의 의존 관계의 구조를 나타내는 트리이다. "먹었다"가 "나는", "밥을"을 지배하는 지배소이며 "나는","밥을"은 의존소이다.

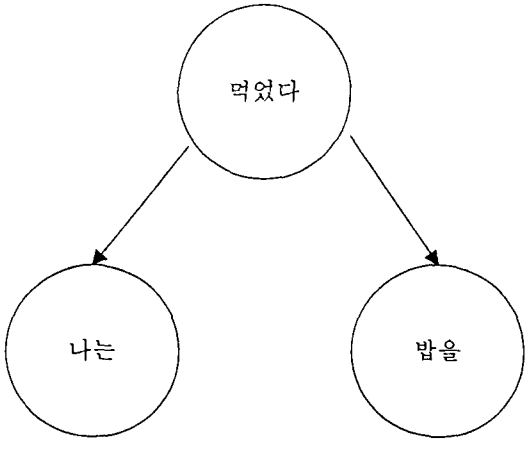


그림 1. 한국어 의존 구조

2.2 초기 파서

형태소 분석이 되어있는 코퍼스를 입력으로 받는다. 일단 모든 어절을 바로 오른 쪽에 있는 어절에 의존하게 한다. 차후에 이러한 의존 구조는 변환 규칙을 학습할 때마다 올바른 의존 구조와 유사하게 변환 된다. 그림2는 "나는 밥을 먹었다"라는 문장이 초기 파서에 의해 의존 트리로 표현된 예이다.

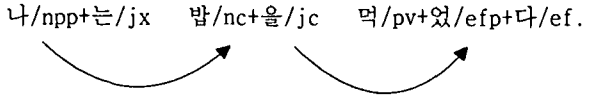


그림 2. 초기 파서에 의한 의존 구조

2.3 변환 규칙 틀

한국어의 특성상 모든 어절은 그 어절의 오른쪽에 있는 어절 중의 하나에 의존하므로 변환규칙 틀을 적용하는 범위는 현재 어절에서 오른 쪽의 어절들로만 제한적으로 검색하여 해당 규칙 틀을 적용한다.

(1-8) 현재 어절의 형태소가 X이면,
가장 가까운(먼) 형태소 Y를 포함한 어절에 의존한다.
가장 가까운(먼) 형태소 Y 또는 Z를 포함한 어절에 의존한다.
가장 가까운(먼) 형태소 Y 와 Z를 포함한 어절에 의존한다.
가장 가까운(먼) 형태소 Y 를 포함하며 Z는 포함하지 않은 어절에 의존한다.
(9-16) 현재 어절의 형태소가 W와 X를 포함하면
가장 가까운(먼) 형태소 Y를 포함한 어절에 의존한다.
가장 가까운(먼) 형태소 Y 또는 Z를 포함한 어절에 의존한다.
가장 가까운(먼) 형태소 Y 와 Z를 포함한 어절에 의존한다.
가장 가까운(먼) 형태소 Y 를 포함하며 Z는 포함하지 않은 어절에 의존한다.
(17) 현재 어절의 형태소가 X이고 오른쪽에 동일한 조건의 어절이 존재하면 그 어절이 의존하고 있지 않는 형태소 Y 또는 Z를 포함한 어절에 의존한다.

그림3은 변환규칙 틀 17을 의존 구조로 나타낸 것이다.

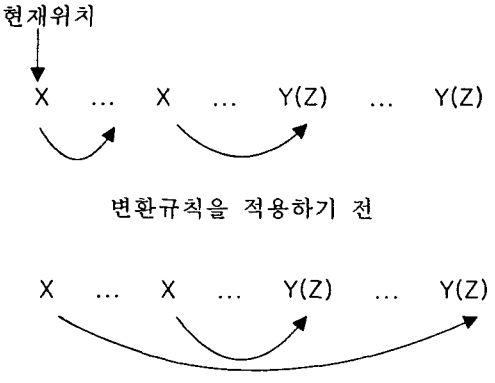


그림 3. 변환 규칙 틀(17)의 구조적 의미

위의 변환 규칙 틀을 적용하고 교차구조의 발생을 막기 위한 다음과 같은 연산을 수행한다.

변환 규칙을 수행했을 때, 의존 어절이 입력 문장의 i 번째 위치에 나타난 어절이라 하고 지배 어절이 j번째 어절이라 하면 다음을 만족하는 모든 어절들에 대해 다음의 연산을 수행한다.

p< i인 어절 p가 i<q<j 인 어절 q와 서로 의존 지배 관계에 있으면 어절 p의 지배소를 어절 q에서 어절 i로 바꾼다.

i<p<j인 어절 p와 q>j인 어절 q가 서로 의존 지배 관계에 있으면 어절 p의 지배소를 어절 q에서 어절 j로 바꾼다.

그림 4는 교차구조의 발생을 막는 연산 후의 구조적인 변화를 나타낸다.

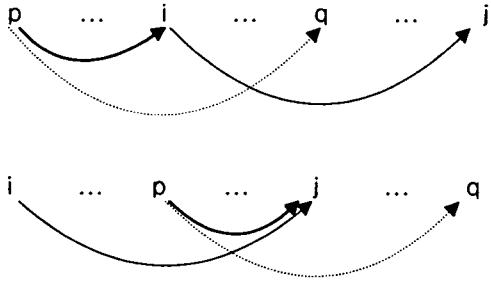
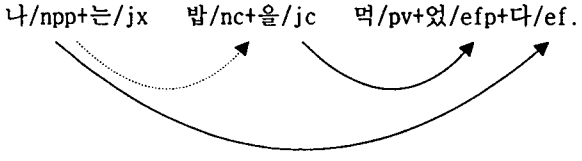


그림 4. 교차 구조를 방지하는 연산

예를 들어 만약 적용하는 변환규칙이 "현재 어절의 형식 형태소가 jx이면 오른쪽에서 가장 가까운 pv에 의존한다"이면 그림 2의 "나는 밥을 먹었다"는 다음과 같이 점선으로 표시된 의존관계를 삭제하고, 실선으로 표시된 관계를 새로 설정한다.



변환 후 교차구조가 없으면 변환 규칙 적용은 완료된다.

2.4 변환 규칙의 점수

변환 규칙을 적용하여 분석된 결과를 올바로 구문 분석된 코퍼스와 비교해서 일치되는 의존 지배 관계의 수를 점수로 사용한다.

예를 들어 위 그림의 변환은 의존 지배 관계가 올바르게 되었으므로 총 의존 지배 관계의 수인 2점이 된다.

2.5 변환 규칙의 학습

변환 규칙을 학습하는 과정은 다음과 같다.

Step 1 모든 입력 문장을 초기 파서로 그림 2와 같은 의존 구조를 갖게 한다. 이 때 초기 파서에 의한 점수를 계산한다. 이 때의 점수를 Threshold로 정의한다.

Step 2 모든 형태소에 대한 변환 규칙 틀을 현재 의존 구조의 입력 문장에 적용하여 그 중 가장 큰 점수를 갖는 변환 규칙을 찾는다. 이 때의 점수를 t 라 한다.

만약 t > Threshold 이면

학습된 변환 규칙 목록에 이 변환 규칙을 추가하고 Threshold는 t로 갱신된다.

만약 t < Threshold 이면

변환 규칙의 학습은 종료된다.

Step 3 Step2에서 학습된 변환 규칙을 입력 문장의 의존 구조에 적용하여 입력 문장의 의존 구조를 변환한다.

Step 4 Goto step 2

2.6 변환 규칙의 적용

학습된 변환 규칙들은 한국어의 의존 구조 분석에 사용된다. 먼저 입력 문장은 초기 파서에 의한 의존 구조를 갖게 되고 학습된 변환 규칙을 차례대로 적용하여 최종 분석된 의존 구조를 출력한다.

3. 실험 및 평가

실험에서 구문 분석된 동아일보 사설과 초등학교 교과서의 449문장, 5806어절을 사용하여 98개의 변환 규칙을 학습하였다. 형태소 분석의 태그는 KTS의 태그를 사용하였다[6]. 표1은 실험에서 학습된 변환 규칙의 예이다.

표1. 학습된 변환 규칙의 예

순위	의존소의 조건	지배소의 조건
1	Jc	pv 또는 xpv
2	Jx	pv 또는 s.
3	Ecs	pv 또는 xpv
4	Jca	pv 또는 xpv
5	A	pv 또는 xpv
6	Ecq	pv 또는 xpv
7	ajs	ecx 또는 ef
8	exa	pv 또는 xpv
9	nb	pv 또는 ecs
10	s,	ecx 또는 ef
11	ecq	pv와 ecs
12	jcm	nc 또는 nb
13	ad	pv 또는 pa
14	xa	pv 또는 xpv
15	ajs	xpv 없고 ecq
16	s,	nc 와 s.
17	npp	pv 와 s'
18	nb	jcp 와 ecq
19	jcp	ef 없고 jcp
20	ecq	pv와 px

실험에 사용하지 않은 초등학교 교과서 182문장, 1533어절에 학습된 98개의 변환 규칙을 적용하여 실험한 결과 의존 관계에서 84.45%의 정확률을 나타내었다.

실험 결과 학습된 규칙에서 문제점을 발견할 수 있었다. 변환 규칙 6, 11, 20과 같이 의존소의 조건이 같은 변환 규칙이 여러 개 학습되었는데 이러한 변환 규칙들을 적용하면 올바르게 의존된 어절이 다음 변환 규칙에 의해 잘못된 어절로 의존되어질 수 있다. 근본적으로 이러한 문제가 발생하는 이유는 변환 규칙을 적용하는 범위를 오른쪽에서 가장 가까운(먼)으로 한정시켰기 때문이다. 이 문제는 변환 규칙의 적용범위를 오른쪽에서 1번째, 2번째, ... 등으로 세분화하고 변환 규칙의 적용 규칙을 엄격하게 만들면 해결될 수 있을 것이다. 예를 들면 어휘 정보를 변환 규칙 틀에 포함하여 변환 규칙을 세분화하는 방법 등이 있다.

변환 규칙 학습 때 걸리는 시간도 문제점이다. 학습에 사용하는 어절 수를 N, 변환 규칙 틀의 수가 P이면 하나의 변환 규칙을 학습하는 데 걸리는 시간은 $O(N^P)$가 된다. 만약 한국어의 의존 관계를 의존소의 형식형태소와 지배소의 실질 형태소의 관계로만 가정한다면 변환 규칙 틀을 적용하기 위한 검색범위가 축소되어 학습시간을 단축시킬 수 있을 것이다. 학습시간을 단축시키기 위해서는 위와 같은 효율적인 변환 규칙 틀의 개발이 요구된다.

4. 결론 및 향후 과제

본 연구에서는 변환 규칙 학습기를 이용하여 코퍼스로부터 자동적으로 변환 규칙을 학습하였다. 학습된 변환 규칙이 한국어의 의존 구조의 특성과 유사하였다.

실험 결과, 잘못된 의존 구조를 만들어 낼 수 있는 문제점과 오랜 학습시간이 문제점으로 남아있다.

현재, 구문 분석된 코퍼스를 계속 늘여가고 있으며 더 나은 변환 규칙 틀과 어휘 정보를 사용한 변환 규칙 틀의 개발과 변환 적용 범위를 조절하여 N-best 결과를 찾아낼 수 있도록 하는 방법을 연구하고 있다.

Reference

[1] E. Brill. "A Simple Rule-Based Part of Speech Tagger," *In Proceedings of the 3rd Conf. on Applied Natural Language Processing*, pp 153-155, 1992

[2] E. Brill. "Automatic Grammar Induction and Parsing Free Text: A Transformation-based Approach," In *Proceedings, 31st Meeting of the Association of Computational Linguistics*, Columbus, OH. 1993a.

[3] E. Brill. "Transformation-based Error-driven Parsing," In *Proceedings, Third International Workshop on Parsing Technologies*, Tilburg, The Netherlands. 1993b

[4] Igor A. Mel'cuk, *Dependency Syntax: Theory and Practice*, State University of New York Press, 1987.

[5] 신상현, TAKTAG: 통계와 규칙에 기반한 혼합형 한국어 품사 태깅 시스템, 포함공대 전자 계산학과 석사학위 논문, 1996.

[6] 이상호, 서정연, 오영환, "KTS : 미등록어를 고려한 한국어 품사 태깅 시스템", 음성통신 및 신호처리 워크샵 논문집(제 SCAS-12권 1호), pp.195-199, 1995

[7] 임희석, 김진동, 임해창, "변형 규칙 기반 한국어 품사 태거의 개선", 제8회 한글 및 한국어 정보처리 학술 대회 논문집, pp.216-221, 1996.