지배가능경로를 이용한 오른쪽 우선 구문 분석[1]

김 창현°, 김 재훈, 서 정연
한국과학기술원 전자계산학과, CAIR

A Right-to-Left Parsing using Headable Path

Changhyun Kim, Jae-Hoon Kim, Jungyun Seo
Dept. of Computer Science and CAIR, KAIST

요 약

본 논문에서는 의존문법을 이용해 한국어와 같이 비교적 어순이 자유롭고, 지배성분 후위의 특성을 갖는 언어를 효율적으로 분석할 수 있는 오른쪽 우선 분석 방법을 제안한다. 지배가능경로를 이용하면 생성되는 의존 트리의 수를 줄일 수 있음을 보이며, 의존 관계의 검사를 위해서는 지배가능경로 상의 문장 성분만을 조사하면 됨을 보인다. 한국어에 적용한 실험 결과를 보이며, 각 방식에 따른 비교 기준은 생성되는 의존 트리의 갯수와 분석 수행 시간으로 한다. 이때 한국어 문장 성분 간의 의존관계는 품사 분류에 의한 기본적인 의존 관계만을 이용하며, 격틀이나 의미 속성 등 추가적인 제약 사항은 이용하지 않는다. 오른쪽 우선 구문 분석은 지배가능경로를 이용함으로써 의존 관계의 빠른 검색을 할 수 있었으며, 문장 지배 성분을 포함하지 않는 부분 의존 트리를 생성하지 않음으로써 생성되는 의존 트리의 수를 줄일 수 있었다.

1. 서론

본 논문에서는 의존문법을 이용한 자연어의 분석 시 지배가능경로(headable path)를 이용한 효율적 분석 방법론을 제안하고, 이를 한국어에 적용하여 기존의 의존문법을 이용한 분석 방법과 비교, 평가하고자 한다.

자연언어를 어순의 자유성 정도에 따라 분류해 보면 크게 영어, 불어, 중국어와 같이 비교적 어순이 고정된 언어, 독일어와 같이 어순의 자유가 영어보다는 많은 언어, 한국어나 러시아어, 라틴어 등과 같이 비교적 어순이 자유로운 언어로 구분할 수 있다[3, 7]. 그러나, 지금까지 대다수의 자연언어 분석에 이용되고 있는 구구조문법은 영어와 같이 비교적 어순이 고정된 언어를 위해 개발된 방법론이며, 한국어와 같이 어순이 비교적 자유로운 언어들의 특성을 제대로 반영하지 못한다는 문제를 가지고 있다. 최근들어 이를 극복하고자 의존문법을 많이 사용하고 있다[3, 6, 7, 8].

[1]본 연구는 한국통신의 장기기초과제 "대화체 기계번역에 관한 연구"의 일부임

프랑스의 Tesniere에 의해 시작된 의존문법(혹은 종속문법)은 문장 성분들 간의 의존관계(종속관계)를 기술하는 문법 이론으로써, 문장에 들어 있는 상이한 등급의 성분들 중에서 지배성분에는 어떤 것이 있으며, 또 이 성분에 결합되어 있는 종속 성분에는 어떤 것이 있는가를 기술한 문법이다[5]. 실제로 우리가 원하는 구문 분석을 수행하기 위해서는 의존문법에서 이용하는 품사 세분류에 따라 지배성분과 종속성분을 기술하는 이외에도 해당 언어, 해당 어휘의 구문적, 의미적 특성을 이용해야 한다.

본 논문에서는 의존문법을 이용한 구문분석 시, 지배성분 후위의 특징[2]을 이용해 문장의 오른쪽에서부터 분석을 시작하는 분석 알고리즘을 제안하고, 이 때 지배가능경로를 이용하여 효율적인 분석이 가능함을 보인다. 오른쪽 우선 분석방식의 효율성 검증에 그 목적이 있으므로 실험시에는 구문적, 의미적 특성을 이용하지 않고 단순히 품사의 세분류만을 이용하였다. 그 결과 생성되는 의존 트리의 수가 줄어들 수 있음을 보인다.

2. 의존 문법

의존 문법 혹은 종속 문법은 문장 성분들 간의 의존 관계(종속 관계)를 기술하는 문법 이론으로써, 문장에 들어 있는 상이한 등급의 성분들 중에서 지배 성분에는 어떤 것이 있으며, 또 이 지배 성분에 결합되어 있는 종속 성분에는 어떤 것이 있는가를 기술한 문법이다[5].

그러므로 의존 문법을 이용한 자연어 분석은 각 문장 성분들 간에 존재하는 의존 관계를 밝히고, 각 의존 관계에서의 지배 성분과 종속 성분을 밝히는 작업이라 할 수 있으며, 한국어와 같이 비교적 자유로운 어순을 갖는 언어의 분석에 적당하다.

의존 관계는 두 문장 성분 사이에 일어나는 이진 관계(binary relation)이며[9], 한 쪽이 다른 쪽을 지배(government, domination)하기 때문에 방향성을 갖는다. 지배하는 쪽을 지배 성분(governer, head)이라 하고, 지배받는 쪽을 종속 성분(dependent)이라 한다. 지배 성분 H와 종속 성분 D 간에 존재하는 의존관계는 'H ⇒ D'로 표기하고, '문장 성분 D는 문장 성분 H에 의존한다' 또는 '문장 성분 H는 문장 성분 D를 지배한다'라고 읽는다.

한국어를 포함한 알타이어족은 공통적으로 실질 형태소에 형식 형태소가 붙는 첨가적 성질을 띠고 있다. 이 때 문장 성분은 대개 형식 형태소에 의해 결정되므로 각 문장 성분 간의 올바른 의존 관계를 파악하기 위해서는 형식 형태소의 세분류가 이루어 져야 하며, 또한 품사 단위의 세분류가 이루어져야 한다. 표 1은 실험에 사용된 한국어의 품사 세분류를 이용한 의존 관계의 일부이다[3].

---

[2]수식받는 말이 수식하는 말의 뒤에 오는 현상. 이 때 수식하는 말을 종속 성분, 수식받는 말을 지배성분이라 하며, 이 특징은 알타이 어족의 공통적인 특성이다.

[3]의존관계 세분류는 [4]에 제안된 것을 이용하였으며, 완전한 의존관계 세분류표는 [4]를 참조하기 바람.

지배 성분	종속 성분	의존 관계
명사	명사, 대명사, 수사, 성상관형사, ...	수식
명사	성상부사	첨가
수사	대명사, 지시관형사, 관형격조사	수식
성상관형사	성상부사	첨가
...	...	...
수관형사	성상관형사	수식
수관형사	성상부사	첨가
용언	명사, 대명사, 수사, 격조사	격관계
용언	부사형어미	첨가

표 1: 의존 관계의 세분류 표의 일부

의존 문법을 이용한 분석 과정에서 생성되는 의존 트리는 반드시 최상위에 하나의 root node를 가지며, 주어진 문장의 모든 문장성분을 포함하는 의존 트리의 root node를 특히 문장 지배성분이라 한다.

3. 지배가능경로를 이용한 오른쪽 우선(right-to-left) 분석

본 장에서는 의존문법을 이용한 분석 방법으로 지배 가능 경로를 이용한 오른쪽 우선(right-to-left) 분석 방식을 살펴보고, 기존의 분석 방식과의 차이점을 예를 통해 살펴본다. 분석 방식의 비교를 위해 편의상 주어진 문장을 왼쪽에서부터 오른쪽으로 분석하는 기존의 방식을 왼쪽 우선(left-to-right) 분석 방식이라 한다.

3.1 오른쪽 우선 분석

[4] 문장 S는 각각의 문장 성분 $W_i(i \geq 1)$로 구성되며, 각 W_i는 i번째 문장 성분을 의미한다. 문장 S의 문장 성분이 n개일 때 $S = W_1W_2...W_n$ 이다.

특성 1) 지배 성분 후위의 원칙

[4]증명에 대한 자세한 내용은 [1]을 참조하기 바람

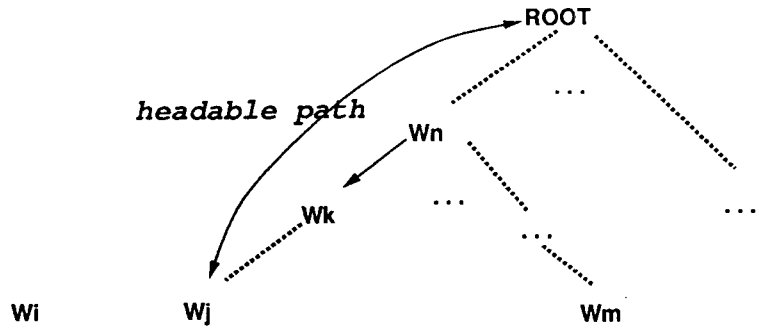

그림 1: 지배 가능 경로

한국어와 같이 알타어에 속하는 어족[2]에서는 문장 S 내의 두 문장 성분 W_i, W_j간에 의존 관계 $W_j \Rightarrow W_i$가 존재하면 반드시 $i < j$ 를 만족하며, 이를 **지배 성분 후위의 원칙**이라고 한다. 즉, 알타이어족에서는 항상 지배성분이 종속성분의 뒤쪽에 위치한다.

문장 $S = W_1 ... W_{n-1} W_n$에 대해 오른쪽 우선 분석 방식은 각 문장 성분 W_i에 대해 뒤에서부터 W_n, W_{n-1}...의 순서로 받아 들이며 분석을 수행한다.

정의 1) 문장 S $= W_1 W_2 ... W_n$에 대해 현재의 입력 문장 성분이 $W_i (i < n)$일 때 $j = i + 1$인 문장 성분 W_j가 존재한다. 이 때 문장성분 $W_j W_{j+1} ... W_n$을 모두 포함하는 각 의존트리에 대해 문장 성분 W_j에서 root까지의 경로를 '**문장 성분 W_i에 대한 지배가능경로**(headable path)'라 하고 지배가능경로 상에 존재하는 모든 문장 성분을 '**문장 성분 W_i에 대한 지배 가능 문장 성분**'이라고 한다.

그림 1에서 입력 문장 성분 W_i에 대한 지배가능경로는 문장 성분 W_j에서 ROOT까지의 경로가 된다.

[정리 1] 문장 S $= W_1 W_2 ... W_n$를 오른쪽 우선 분석 방식으로 분석할 경우, 현재의 입력 문장 성분 $W_i (i < n)$에 대한 지배 성분 $W_j (i < j)$는 문장성분 $W_{i+1} W_{i+2} ... W_n$을 모두 포함하는 의존트리의 지배가능경로 상에 존재해야 한다[1].

오른쪽 우선 분석 방식의 알고리즘은 그림 2과 같다.

1. 문장 $S = W_1...W_{n-1}W_n$의 각 문장 성분 W_i(i는 n에서부터 1까지)에 대해 단계 2, 3, 4를 수행한다.

2. 입력 문장 성분 W_i만으로 구성되는 의존 트리 T_i를 구성한다.

3. 문장 성분 $W_iW_{i+1}...W_n$을 모두 포함하는 각 의존트리 T에 대해

   3.1 의존 트리 T의 지배가능경로상에 있는 각 문장성분 W_j에 대해

   if (W_j와 W_i간에 의존 관계 $W_j \Rightarrow W_i$가 존재하면)

   then 의존 트리 T와 의존 관계 $W_j \Rightarrow W_i$를 포함하는 새로운 의존 트리 T_{new}를 생성한다.

4. if (문장 성분 $W_1W_2...W_n$을 모두 포함하는 의존 트리가 존재하면) 구문분석 성공

   else 구문분석 실패

그림 2: 지배가능경로를 이용한 오른쪽 우선 분석 방식의 알고리즘

오른쪽 우선 분석 방식에 의해 한국어 문장 "철수는(= W_1) 예쁜(= W_2) 꽃을(= W_3) 좋아한다(= W_4)" [5] 를 분석하는 과정은 그림 3과 같다.

첫번째 입력 문장 성분 W_4는 W_4 자신만을 포함하는 의존 트리 T_1을 구성한다. W_4와 의존관계를 구성할 의존트리가 없으므로 다시 새로운 입력 문장 성분 W_3을 받아 들이고, W_3으로 구성되는 의존 트리 T_2를 구성한다. 의존 트리 T_1의 지배가능경로 상에 있는 문장 성분 W_4와 문장 성분 W_3 사이에 의존 관계 '형용사 ⇒ 조사'가 존재하므로 의존트리 T_1과 문장성분 W_3을 포함하는 새로운 의존 트리 T_3을 구성한다. 다시 새로운 입력 문장 성분 W_2를 받아 들여 의존트리 T_4를 구성한다. 문장성분 W_3, W_4를 모두 포함하는 의존트리는 T_3이므로, T_3의 지배가능경로 상에 존재하는 문장 성분 W_3, W_4 각각에 대해 W_2와의 의존관계를 조사한다. 의존관계 '명사 ⇒ 관형형어미'가 존재하며, 의존관계 $W_3 \Rightarrow W_2$를 포함하는 새로운 의존트리 T_5가 구성된다. 다시 W_1을 받아 들이면서 위와 같은 과정을 반복하면 결국 전체 문장 성분 $W_1W_2W_3W_4$를 포함하는 의존 트리 T_7이 구성되며, 문장 "철수는 예쁜 꽃을 좋아한다"는 올바른 문장으로 인식된다.

---

[5] "철수는" : 명사+조사
"예쁜" : 형용사+관형형전성어미
"꽃을" : 명사+조사
"좋아한다" : 동사+어말어미
각 문장 성분간에 존재하는 의존 관계로는 '명사 ⇒ 관형형전성어미', '형용사 ⇒ 조사'가 있다.

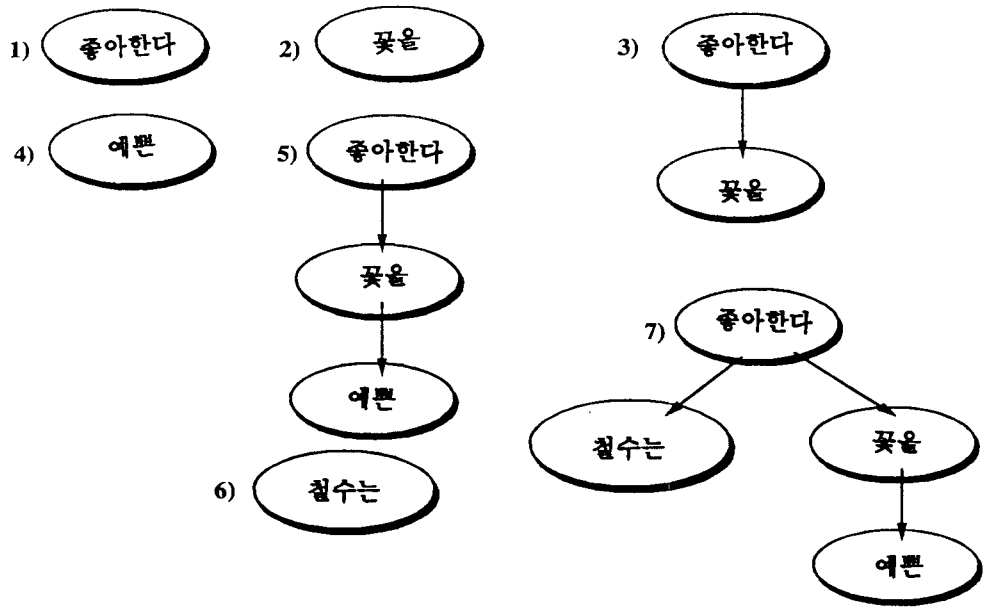

그림 3: 오른쪽 우선 분석 방식에 의한 의존 트리 생성 과정

3.2 비교

3.2.1 왼쪽우선분석

문장 S = $W_1W_2...W_n$에 대해 입력 문장 성분의 앞에서부터 W_1, W_2, ...의 순서로 받아들이며 분석한다. 즉, 새로운 입력 문장성분 $W_i(i > 1)$와 이제껏 생성된 모든 의존트리 중 문장 성분 W_{i-1}을 포함하는 모든 의존 트리와의 의존관계를 검사하게 되며, 의존관계가 존재하면 새로운 의존 트리를 구성한다.

한국어 문장 "철수는 예쁜 꽃을 좋아한다" 에 대한 분석 과정은 그림 4과 같다.

그림 4에서 첫번째 입력 문장 성분 W_1은 자신만으로 구성되는 의존 트리 T_1을 구성한다. W_1이 첫 의존트리이므로 새로운 입력 문장 성분 W_2를 받아 들이고, 의존 트리 T_2를 구성한다. 문장 성분 W_1을 포함하는 의존 트리 T_1에 대해 문장성분 W_1과 W_2 간의 의존관계가 존재하지 않으므로 다시 새로운 입력 문장 성분 W_3을 받아 들이고 의존 트리 T_3을 구성한다. 의존 트리 T_2의 문장성분 W_2와 의존 트리 T_3의 문장성분 W_3 간에 의존 관계 '명사 ⇒ 관형형전성어미'가 존재하므로 의존 트리 T_2와 T_3을 포함하는 새로운 의존 트리 T_4를 구성한다. 의존트리 T_3을 포함하는 더이상의 의존트리가 구성될 수 없으므로 새로이 구성된 의존 트리 T_4를 새로운 입력 의존 트리로 하여 의존 관계를 조사한다. 그러나, 역시 더 이상의 의존 트리가 구성될 수 없다. 다시 새로운 입력 문장 성분 W_4를 받아들여

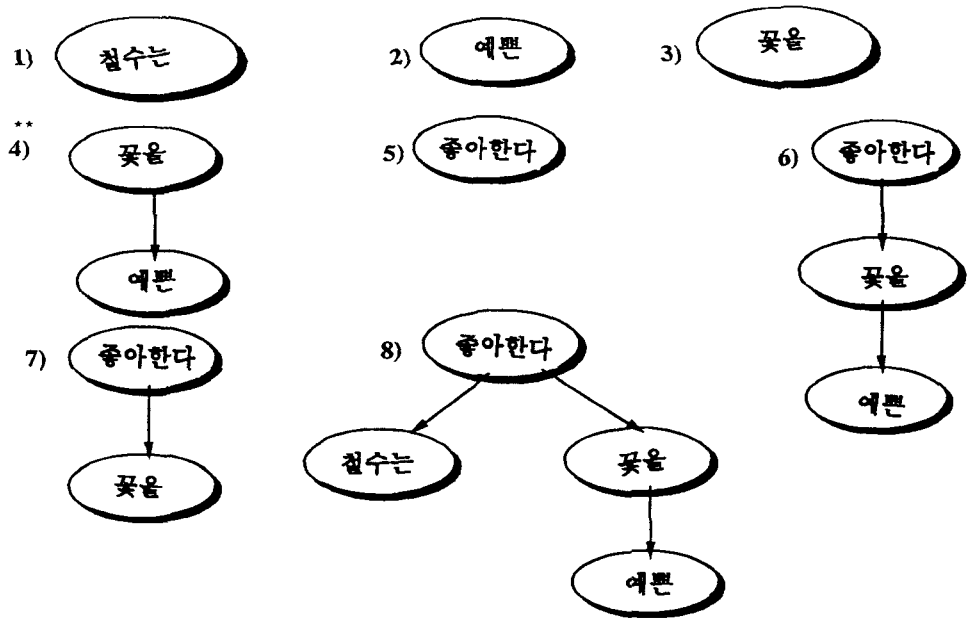

그림 4: 왼쪽 우선 분석 방식에 의한 의존 트리 생성 과정

위의 과정을 반복하면, 결국 전체 문장 성분 $W_1W_2W_3W_4$를 포함하는 의존 트리 T_8이 구성되고 문장 "철수는 예쁜 꽃을 좋아한다"는 문법적인 문장으로 인식된다.

3.2.2 특성 분석

그림 3과 4에서 보여진 바와 같이 문장 "철수는 예쁜 꽃을 좋아한다"에 대해 왼쪽 우선 분석 방식에 의해 생성된 의존 트리 T_4가 오른쪽 우선 분석 방식에서는 생성되지 않았다. 이와 같이 오른쪽 우선 분석 방식에서는 문장 $S = W_1W_2...W_n$의 문장 지배 성분 W_n을 포함하지 않는 의존 트리는 생성되지 않으며[5], 왼쪽 우선 분석에 의해 생성된 의존 트리 T_4와 같이 아직 지배성분이 결정되지 않은 문장 성분이 root가 되는 중간 과정의 의존트리를 생성하지 않음으로써 효율적인 분석을 수행할 수 있다.

[정리 2] 문장 $S = W_1...W_n$이 문법적인 문장일 경우, 문장 S에 대해 오른쪽 우선 분석 방식은 각 문장 성분 W_i로 구성되는 의존트리와 $W_iW_{i+1}...W_n$을 모두 포함하는 의존 트리만을 생성하며,이는 필요한 모든 구문 분석 과정을 거친다[1].

[5]물론, 각 문장성분 $W_1, W_2, ..., W_n$ 하나로 구성되는 의존트리는 생성된다.

문장 지배 성분을 제외한 해당 문장 내의 모든 문장 성분은 반드시 자신을 지배하는 지배 성분을 가지며, 올바른 지배성분을 찾기 위해서는 많은 정보가 요구된다. 이때, 지배 성분을 찾기 위해서는 각 의존 트리와의 의존 관계를 검사하여야 하며, 따라서 생성되는 의존 트리의 수를 줄이는 것은 구문분석의 속도에 중요한 영향을 미친다. 본 논문에서 제안하고 있는 오른쪽 우선 분석 방식은 지배 성분 후위의 원칙과 지배 가능 경로를 이용하여 생성되는 의존 트리의 수를 줄일 수 있을 뿐만 아니라, 지배 가능 경로를 이용한 빠른 의존 관계의 검색이 가능하다.

4. 실험 및 분석

본 장에서는 오른쪽 우선 분석 방식과 기존의 왼쪽 우선 분석 방식을 구현하고, 생성되는 의존트리 갯수의 비율과 분석 수행 시간에 따른 비율에 의해 평가를 하였다.

실험에 이용된 한국어 20문장은 길이별, 유형별로 구분하여 선택되었으며, 분석에 이용된 의존관계는 앞에서 살펴본 바와 같이 기본적인 의존관계만을 이용하였다.

그림 5의 빗금친 부분은, 각 입력 문장에 대해 왼쪽 우선 분석 방식으로 생성된 의존 트리 수를 100으로 놓고 오른쪽 우선 분석 방식에 의해 생성된 의존 트리 수를 빗금친 부분으로 나타내고 있다.

마찬가지로 그림 5의 검은 부분은, 왼쪽 우선 분석 방식의 경우 분석에 걸리는 시간을 100으로 할 때 오른쪽 우선 분석 방식에 의한 수행 시간을 의미한다. 그림 5를 살펴보면, 모든 입력 문장에 대해 오른쪽 우선 분석 방식이 더 빠름을 알 수 있다. 의존 문법을 이용한 구문 분석의 경우, 분석에 걸리는 시간은 생성되는 의존 트리의 수에 비례하며, 이는 분석 수행 속도의 대부분이 각 의존트리 간의 의존 관계를 검사하는데 소요되기 때문이다. 주어진 문장에서의 문장 지배 성분을 지배성분으로 갖지 않는 문장 성분이 있을 경우, 오른쪽 우선 분석 방식은 왼쪽 우선 분석 방식보다 더 적은 의존트리를 생성하며, 따라서 더 적은 의존 트리에 대해 의존 관계를 검사하므로 검사에 따른 시간이 줄어든다. 또한 오른쪽 우선 분석은 바로 이전 단계에서 생성된 의존 트리의 지배가능경로만을 따라 의존관계를 검사하게 되므로, 생성된 전체 의존 트리를 검색할 필요가 없으며 빠른 의존 관계의 검사가 가능하다.

5. 결론

본 논문에서는 한국어와 같이 비교적 어순이 자유롭고 지배성분 후위의 특성을 갖는 언어에 대해 의존 문법을 이용해 효율적으로 구문 분석을 수행하는 오른쪽 우선 분석 방법을 제안하였다. 오른쪽 우선 분석 방식은 의존 관계 검색 시에 지배가능경로를 이용하므로, 효율적인 의존 관계의 검색이 가능하며, 이로 인해 의존 관계의 조사에 걸리는 시간이 줄어들게 된다.

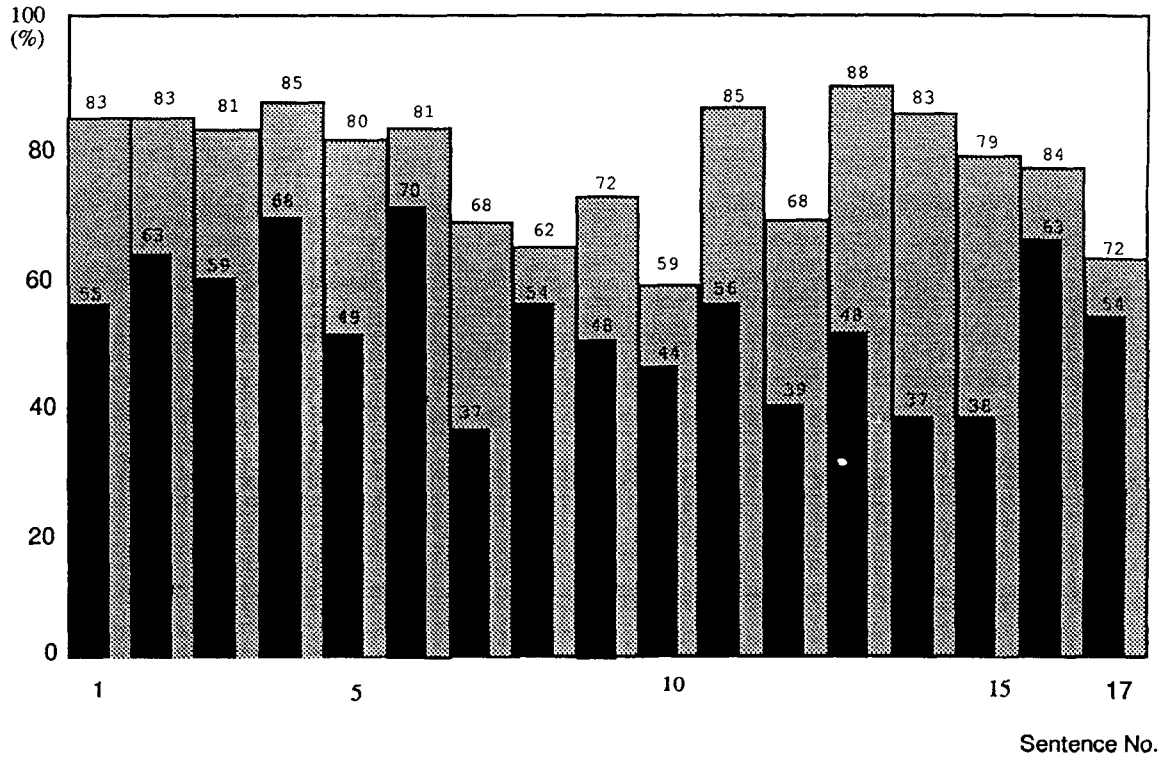

그림 5: 각 문장에 대한 의존트리 생성 수와 수행속도 비율

분석하는 방법론에 있어서 왼쪽 우선 분석 방식은 현재의 입력 문장 성분과 의존 관계를 형성하는 의존 트리의 탐색을 위해 이미 생성되어진 모든 의존트리를 대상으로 탐색을 한다. 이러한 탐색 방식은 의존트리의 수가 많아짐에 따라 탐색 시간도 커지게 되며, 이는 부분적인 인덱싱에 의해 어느 정도 줄일 수는 있다. 오른쪽 우선 분석 방식은 이미 생성된 의존 트리 중 이제까지 들어온 모든 문장성분을 가지는 의존 트리에 대해 지배가능경로만을 조사하므로 생성되는 의존트리에 대해 효과적인 의존 관계 검사가 가능하다.

생성되는 의존 트리에 있어, 오른쪽 우선 분석 방식은 2개 이상의 문장 성분을 포함하는 경우 항상 문장 지배 성분을 포함하는 의존 트리를 생성한다. 그러나, 왼쪽 우선 방식은 문장 지배 성분을 포함하지 않는 의존트리를 생성할 수 있으며, 이러한 의존 트리의 수 만큼 오른쪽 우선 분석 방식에 비해 더 많은 의존 트리를 생성하게 된다.

본 논문에서는 기본적인 의존 관계만을 가지고 실험을 하였다. 그러나, 실제의 구문분석에서는 가능한 한 애매성을 줄이고 올바른 의존 관계만을 형성하기 위하여 많은 제약들을 사용하며, 이러한 제약 사항들을 검사하는데에도 또한 구문분석의 많은 시간이 투자된다. 본 논문에서 제안한 오른쪽 우선 분

석 방법이 이러한 제약 사항들과 함께 수행될 경우 어떠한 수행 특성을 보일 지에 대해서는 좀 더 연구가 진행되어야 할 것이다. 그러나, 본 논문에서의 실험 결과로 보아 기존의 분석 방법보다는 좋은 결과를 가져올 것으로 기대된다.

참고 문헌

[1] 김창현, 한국어 구문 분석을 위한 오른쪽 우선 차트 파서, 한국과학기술원, 전산학과, 석사학위논문, 1993.

[2] 남기심,고영근, 표준 국어문법론, 탑출판사, 1985.

[3] 박용욱, 조혁규, 권혁철, "의존문법을 이용한 한국어 분석기의 구현", 한국정보과학회 봄 학술발표논문집, vol. 17, no. 1, pp. 191-194, 1990.

[4] 우승균, 구문관계를 이용한 한국어 구문분석, 한국과학기술원, 전산학과, 석사학위논문, 1992.

[5] 이점출, 의존문법개론, 한신문화사, 1991.

[6] 홍영국, 이종혁, 이근배, "의존문법에 기반을 둔 한국어 구문 분석기", 한국정보과학회 봄 학술발표논문집, vol. 20, no. 1, pp. 781-784, 1993.

[7] M.A. Covington, "Parsing Variable Word Order Languages with Unification-Based Dependency Grammar", *ACMC Research Report 01-0022, Univ. of Georgia*, 1988.

[8] FUKUMOTO Fumiyo, SANO Hiroshi, "A Framework For Restricted Dependency Grammar", SICONLP, pp. 11-16, 1990.

[9] I. A. Melčuk, *Dependency Syntax: Theory and Practice*, The State Univ. of New York Press, 1988.