구간분할 기반 한국어 대등접속 구문분석 기법

장재철[0], 박의규*, 나동렬*
연세대학교 컴퓨터과학과[0], 연세대학교 정보기술학부*
jcjang@cs.yonsei.ac.kr[0], {ekpark, dyra}@dragon.yonsei.ac.kr*

A Korean Conjunctive Structure Analysis based on Sentence Segmentation

Jae-Chul Jang[0], Eui-Kyu Park*, Dong-Yul Ra*
Dept. of Computer Science, Yonsei University[0]
Information Technology Division, Yonsei University*

요 약

본 논문에서는 한국어의 대등접속의 구문분석 문제를 다룬다. 이를 전체 문장의 구문분석 기법에 기반하도록 하여, 문장 전체에 대한 분석 결과를 가장 좋게 하는 대등접속 구조가 선정되도록 하였다. 본 기법의 특징으로는 구간 분할 기법에 기반하여 대등접속 구조 후보의 수의 감소가 가능하게 되기 때문에 구문분석의 안정화를 얻게 되었다. 또한 전체 구문분석기에서 한 부품으로 동작함으로써 전체 문장 구조가 올바른 대등구조를 선택할 수 있게 되어, 보다 전역적인 정보의 이용에 의한 분석이 되었다. 선접속부와 후접속부 간의 구조 및 어휘적 유사성, 평행연결의 이용 등은 본 기법의 또 다른 특징으로 볼 수 있다. 실험결과 정상적인 문장의 대등접속에 대한 분석에서 매우 효과적으로 동작함을 관찰하였다.

1. 서론

대등접속 문제는 자연어 이해에서 가장 어려운 문제 중의 하나로 인식되어 오고 있다. 대등접속은 자연어 처리에 가장 문제 거리인 중의성 해소 문제의 가장 대표적인 것이다. 문장이 대등접속을 가진 경우 문장의 구조에 대한 중의성의 수가 폭발적으로 증가한다. 대등접속은 명사에 의한 대등접속과 용언에 의한 대등접속 두 가지로 나누어 볼 수 있다. 전자의 예를 보자.

그 사람은 [[맛있는 빵] 과 [시원한 음료수]] 를 가게에서 샀다. (1)

(1)에서는 두 명사구 "맛있는 빵", "시원한 음료수"가 대등접속 조사 "과"에 의하여 연결된 경우이다. 후자의 경우로는,

그 선생은 [[아이들을 가르치고] [책을 쓰면서]] 한 시절을 보냈다 (2)

(2)에서 "아이들을 가르치고"와 "책을 쓰면서"가 대등접속 연결어미 "고"로 연결된 경우이다.

본 논문에서는 후자의 경우, 즉 두 용언에 의한 대등접속 문제를 다루고자 한다.

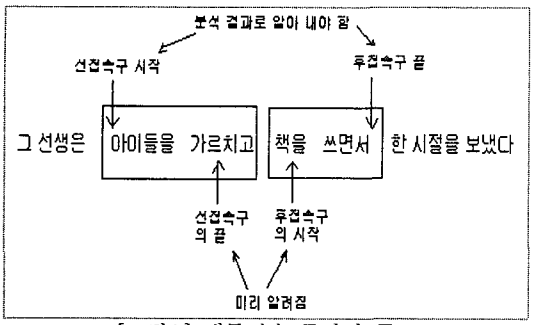
용언에 의한 대등접속을 분석하는데 있어서, 알아내야 할 주요 핵심은 먼저 대등접속이 있다는 것을 감지하는 작업과 어느 부분과 어느 부분이 대등하게 접속이 되는지 그 범위를 파악하는 일이다. 위 (2) 에서 보면 대등연결의 앞부분 "아이들을 가르치고"를 선접속부(pre-conjunctive structure), 뒷부분 "책을 쓰면서"를 후접속부(post-conjunctive structure)라 부른다. 선접속부의 가장 중심이 되는 어휘, 즉 머리어(head)는 대등접속 어미를 가진 마지막 용언 "가르치고" 이다. "쓰면서"는 후접속부의 머리어(post-head)이다. 여기에서 대등접속 구조의 분석에 들어가기 전에 이미 알려진 사항은 다음 두 가지 사항이다.

- 선접속부의 머리어
- 후접속부의 시작 단어

따라서 선접속부와 후접속부의 범위를 알아내기 위하여 추가적으로 알아야 할 사항은 다음 두 가지 사항이 되어야 할 것이다.

- 선접속부의 시작 단어

- 후접속부의 마지막 단어
(이것은 후접속부의 머리어가 되는 용언임)



[그림1] 대등접속 문장의 구조

대등접속의 구조 분석이 어려운 이유는 많은 중의성으로 인해 많은 수의 구문구조(parse tree)가 가능한 후보가 된다는 점이다. 이 중에서 올바른 구조를 고른다는 것은 어려운 문제이다. 예를 들어 다음 문장을 보자.

머리 위로 새들이 날고 배고픈 사자가 뛰는 양을 쳐다보았다. (3)

(3)에서 가능한 대등접속구조는 다음과 같다.

머리 위로 새들이 [날고] [배고픈] 사자가 뛰는 양을 쳐다보았다 (3a)

머리 위로 새들이 [날고] [배고픈 사자가 뛰는] 양을 쳐다보았다 (3b)

머리 위로 새들이 [날고] [배고픈 사자가 뛰는 양을 쳐다보았다] (3c)

머리 위로 [새들이 날고] [배고픈] 사자가 뛰는 양을 쳐다보았다 (3d)

머리 위로 [새들이 날고] [배고픈 사자가 뛰는] 양을 쳐다보았다 (3e)

머리 위로 [새들이 날고] [배고픈 사자가 뛰는 양을 쳐다보았다] (3f)

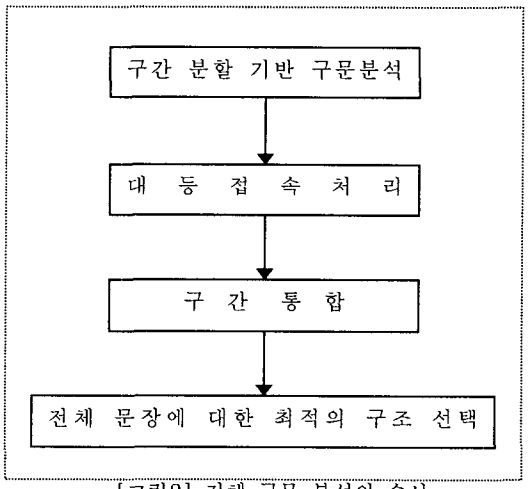
[머리 위로 새들이 날고] [배고픈] 사자가 뛰는 양을 쳐다보았다 (3g)

[머리 위로 새들이 날고] [배고픈 사자가 뛰는] 양을 쳐다보았다 (3h)

[머리 위로 새들이 날고] [배고픈 사자가 뛰는 양을 쳐다보았다] (3i)

영어의 대등접속 분석에 대한 연구와 마찬가지로 지금까지의 한국어 대등 분석에 대한 연구는 별로 많지 않다. 대표적인 연구로는 윤준태[4]의 것을 들 수 있다. 그러나 이 연구는 대등접속 분석 문제 만을 따로 떼어서 생각한 것이다. 즉, 구문분석기와의 관계가 명확하지 않기 때문에 전체적인 구문분석기의 동작과 대등접속 분석기와의 관계가 명확하지 않다. 일본어는 한국어와 유사한 구조의 언어이다. 일본어의 대등접속 분석에 대한 대표적인 연구로 나가오[2]의 연구가 있다. 이 기법에서는 아직 구문분석 되지 않은 두 부분 단어열의 유사성을 계산하여 가장 유사성이 높은 두 부분을 선 후 접속구로 선택하는 방법이다. 이 기법의 문제점은 아직 분석이 전혀 되지 않은 단어 열을 고려한다는 점에서 본 논문의 기법과 다르다.

본 논문에서의 대등접속 분석은 문장 전체에 대한 구문분석 과정 내에서 하나의 단계로 삽입되어 있으며, 다른 부분에 대한 구문분석과 밀접한 관련을 맺고 있다. 실질적으로 문장 안에서 대등접속 분석만을 따로 떼어서 생각하는 것은 적절치 않다. 문장의 일부인 대등접속 부분의 분석이 다른 부분의 분석에 영향을 미치지 않을 수 없기 때문이다.



[그림2] 전체 구문 분석의 순서

본 논문은 문장 전체에 대한 구문분석의 큰 틀 아래에서 특히 대등접속과 관련된 부분을 다루고 있다. 문장의 구문분석에 대한 큰 틀은 "구간 분할 기반 구문분석" 이라는 구호 아래 행하여 진다. 이 방법은 문장 전체를 여러 구간(segment)으로 나눈 다음 각 구간의 구문분석을 일단 수행하는 것을 기반으로 한다. 그 다음 각 구간의 구문분석 결과를 서로 연결하여 문장 전체에 대한 구문구조를 획득하도록 한다. 이에 대한 자세한 설명은 본 논문집의 다른 논문을 참조하기 바란다[6]. 본 논문에서는 평행 연결, 선접속부와

후접속부 사이의 구조 대응, 문장 전체의 관점에서 본 대등접속 구조의 정당성과 같은 사항을 고려하여 분석을 수행하도록 한다.

간단한 실험 결과에 따르면 본 논문에서 제안하는 기법이 앞으로 많은 발전 가능성이 있음을 보여 주었다.

2. 구간분할 기반 구문분석

자연어 구문분석에서 가장 문제가 되는 점은 중의성으로 인하여 가능한 구문구조의 수가 너무 많아진다는 것이다. 문장이 길어지면 그 수가 더욱 많아진다. 따라서 구문분석기(parser)가 이 중 올바른 구조를 골라 낸다는 것은 매우 어렵다. 다시 말하면 올바른 구조의 선택에 있어서, 잘못 고르는 경우가 많다는 것이다.

단계1 : 구간 분할 구문분석

따라서 우리는 문장을 여러 개의 조각(구간)으로 분할하고 각 구간 안에서는 모든 경우를 고려하는 완전한 구문분석(full parsing)을 수행한다. 구간의 구문분석에서는 CYK 구문분석 기법을 사용하여 모든 가능한 구조들을 구한 후 가장 최적으로 보이는 구조가 선택된다. 구간 분석 단계 후 각 구간을 나타내는 최적의 구조들이 구간마다 하나씩 결정된다. 기본적으로 문장은 각 용언을 구간 경계로 하여 나눈다. 그러나 용언 중에서 관형형 어미의 용언은 구간을 나누는 일에 이용되지 않는다. 이 기본적인 구간 나누기 원칙에 여러 가지의 예외적인 경우를 두어 효율적인 구간 분할이 되도록 할 수 있다.

단계 2 : 대등접속 분석

문장 내에 (용언에 대한) 대등접속이 있으면 특별히 이 단계를 거치게 된다. 이에 대한 자세한 사항은 다음 절 이하에서 설명한다.

단계 3 : 구간 통합 후처리

이 단계에서는 2가지의 주요 작업이 수행된다. 그 하나는 어느 구간의 용언에 대하여 필수격이 채워져 있지 않은 경우, 다른 구간의 명사구를 이용하여 필수격을 채워주려는 시도를 한다. 다른 하나는 어느 구간에서 용언에 연결되지 못한 명사구(dangling NP)가 존재하는 경우 이것을 다른 구간의 용언의 격으로 채워주려는 시도를 수행한다.

단계 4 : 전체 문장에 대한 최적의 구조 선택

이 단계에서는 문장에 대한 많은 구문분석 구조 중에서 최적의 것을 선택한다. 앞에서도 언급한 것처럼 구간마다 하나만의 구조가 구해진다. 따라서 문장 내에 대등접속이 없다면 단계3 이후에도 문장에 대한 구조는 단 하나이다. 그러나 대등접속이 존재하면 여러 가지의 대등연결 즉, 대등접속 구조가 제안된다. 그러면 각 대등분석 구조마다 따로 후처리가 수행되고 결국 각 대등분석 구조마다 하나의 전체 문장 구조가 생성된다. 그러면 이 단계에서는 문장에 대한 여러 구조들 중에서 최적의 것을 선택하는 것이다.

이와 같이 한국어 문장에 대하여 구간을 나누어 각 구간에 대하여 완전한 구문분석을 수행하여 각 구간에 대한 구문분석 결과를 후처리에서 합하여 주는 방법은 기존의 구문분석 기법과 비교할 때 장점 및 단점이 존재한다.

먼저 장점으로는 구문분석의 간략화를 들 수 있다. 다시 말하면 분석 과정에서 생성되는 구문 구조(parse tree)의 수를 크게 줄일 수 있다는 점이다. 그 이유는 문장의 길이가 길더라도 각 구간의 길이는 작게 된다. 그리고 각 구간에 대하여 완전한 구문분석을 수행함에 있어서, 구간 길이가 짧기 때문에 구간에 대한 구조의 수는 그리 크지 않게 된다.

단점으로는 문장 전체에 대하여 완전 분석을 하지 않기 때문에 완전히 맞는 즉 100% 맞는 구문 구조가 생성되지 못할 가능성이 있다는 점이다. 그러나 어차피 완전한 구문분석은 현재의 기술 수준 상 당분간은 불가능한 것으로 알려져 있다. 따라서 완전히 맞는 구조를 구하려다 분석이 실패하는 상황에 자주 처하는 것보다는, 조금 틀리더라도 모든 문장에 대하여 안정적으로 구조를 구할 수 있게 하는 것이 더 현명하다고 본 것이다.

3. 대등접속 구문분석 기법

3.1 대등접속 후보 선정

본 연구에서 이야기하는 대등접속 분석의 정도는 선행부와 후행부의 범위를 파악하는 것이 큰 부분이다. 물론 이것이 구문분석기 안에 들어있는 부품이므로 각 어절 간의 의존 관계도 밝혀진다.

- 후접속부가 속하는 구간 선정

후접속부의 범위는 선접속부의 마지막 단어인 대등접속 연결어미를 가진 용언(선접속부의 머리어)이 존재하는 구간 바로 다음 구간이라고 가정한다. 완전한 구문분석을 위해서는 선접속부 구간 이후의 모든 구간이 가능하다고 하여야 하나, 이는 분석의 복잡도를 크게 증가시키기 때문이다. (4)의 예문에서 이 현상을 살펴 보자.

그 남자가 자기의 아내를 구타하고 | 아들이 좋아하는 장난감을 부술 때 | 경찰이 도착했다. (4)

위 문장의 각 구간이 수직선 "|" 에 의해 표시되어 있다. 여기에서 선접속부의 머리어(또는 선머리어, pre-head)는 "구타하고"로 쉽게 파악되며 구간1의 마지막 단어이다. 따라서 후접속부는 구간2 안에 존재한다고 가정한다. 선접속부와 후접속부의 구분을 앞으로는 이와 같이 표기하도록 하겠다

• 후접속부 선정

이때 가능한 후접속부의 범위는 다음과 같이 두 가지가 된다:

그 남자가 자기의 아내를 구타하고] | [아들이 좋아하는] 장난감을 부술 때 | 경찰이 도착했다. (4a)

그 남자가 자기의 아내를 구타하고] | [아들이 좋아하는 장난감을 부술] 때 | 경찰이 도착했다. (4b)

후접속부의 머리어(후머리어, post-head)는 항상 후접속부의 마지막 단어가 된다. 후접속부가 속하는 구간 안의 모든 용언이 후머리어의 후보가 된다. (여기서 조금 더 세밀한 기법을 이용하면 후머리어의 후보수를 줄일 수 있다.)

이로서 다음과 같은 6가지의 대등접속 후보가 가능하고 이들 간의 경쟁을 통하여 가장 최상의 것을 선택하여야 한다.

그 남자가 자기의 아내를 [구타하고] | [아들이 좋아하는] 장난감을 부술 때 | 경찰이 도착했다. (4c)

그 남자가 [자기의 아내를 구타하고] | [아들이 좋아하는 장난감을 부술] 때 | 경찰이 도착했다. (4d)

[그 남자가 자기의 아내를 구타하고] | [아들이 좋아하는] 장난감을 부술 때 | 경찰이 도착했다. (4e)

그 남자가 자기의 아내를 [구타하고] | [아들이 좋아하는 장난감을 부술] 때 | 경찰이 도착했다. (4f)

그 남자가 [자기의 아내를 구타하고] | [아들이 좋아하는] 장난감을 부술 때 | 경찰이 도착했다. (4g)

[그 남자가 자기의 아내를 구타하고] | [아들이 좋아하는 장난감을 부술] 때 | 경찰이 도착했다. (4h)

• 선접속부 선정

선접속부의 마지막 단어는 대등접속 연결어미를 가진 단어로서 이미 결정되어 있다. 이 구간 내에 선접속부가 존재한다고 가정한다. 선접속부의 범위는 시작하는 단어만 알면 결정된다. 선접속부 시작 단어의 후보로는 먼저 선머리어 자체이다. 그 다음 가능한 후보로는 선머리어에 연결된 명사구가 시작하는 단어이다. 결국 (4)에서 선접속부의 시작은 다음과 같이 3가지 경우가 가능하다.

그 남자가 자기의 아내를 [구타하고] | 아들이 좋아하는 장난감을 부술 때 | 경찰이 도착했다. (4i)

그 남자가 [자기의 아내를 구타하고] | 아들이 좋아하는 장난감을 부술 때 | 경찰이 도착했다. (4j)

[그 남자가 자기의 아내를 구타하고] | 아들이 좋아하는 장난감을 부술 때 | 경찰이 도착했다. (4k)

위 예에서 "자기의 아내를"이 한 명사구를 이루어 선머리어 "구타하고"에 의존관계를 가지므로 "아내를"에서 선접속부가 시작한다고 보지 않고 "자기를"에서 시작한다고 본다. 관형어 "그"도 "남자가"의 일부로 하여 명사구 "그 남자가"를 이루므로 선접속부 시작 후보로 "남자가"는 되지 않고 "그"에서 시작할 수 있다고 본다.

선접속부 후보 1개와 후접속부 후보 1개를 선택하면 하나의 가능한 대등접속 후보가 만들어 지므로, 결국 위의 예에서 가능한 대등접속 구조의 모든 경우의 수는 2×3 = 6가지가 된다.

3.2 대등접속 구조의 결정

가능한 대등접속 구조가 여러 개가 나올 수 있으므로 이 중에서 가장 좋은 것을 선택하는 기준이 있어야 한다. 여기서는 각 대등접속을 포함하는 전체 문장의 분석 결과 즉, 전체 분석 구조(full parse tree)에 대하여 각각 점수를 계산하여 이 점수가 가장 높은 것을 선정하는 방식을 택한다. 이것은 바로 각 가능한 대등접속에 대하여 후처리를 수행하여 전체 문장에 대한 분석구조를 만든 다음 이 전체 분석구조에 대한 점수를 비교 기준으로 한다는 것이다.

김광백[6]에서도 설명한 것처럼 구문구조는 의존관계의 집합이므로 분석구조의 점수는 소속 의존관계의 점수를 합한 것으로 한다. 의존관계의 점수는 상호정보 I (N,P,V)를 사용한다[5,6]. 대등접속을 포함한 문장의 경우는 대등접속의 적합/부적합성을 반영하는 점수를 추가로 더한다. 이러한 적합/부적합의 여부는 선접속부와 후접속부 간의 구조의 유사성과 이들을 구성하는 어휘의 유사성으로 판단될 수 있다. 이것을 수식으로 표현하면 (식1)과 같이 된다.

$$score(T) = \sum_i I(arc_i) + conj(1,2) \qquad (식1)$$

(주) arc_i : 문장내의 한 의존관계
$score(T)$: 전체 분석구조의 점수

conj(1,2): 대등접속에 의한 점수

여기에서 $\sum_i I(arc_i)$ 에 대한 내용은 용언과 필수격 간의 관계로 이루어진 의미적 상호정보에 의한 점수이고, 이는 또한 김광백[6]에서 자세하게 논의된 바 있기 때문에 설명을 생략한다.

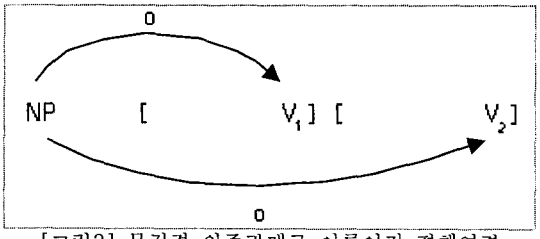
$$conj(1,2) = structural\_sim(1,2) + lexical\_sim(1,2) \quad (식2)$$

(주) *structural_sim(1,2)*: 선접속부와 후접속부의 구조 유사성에 의한 점수
lexical_sim(1,2): 선접속부와 후접속부를 구성하는 어휘의 유사성에 의한 점수

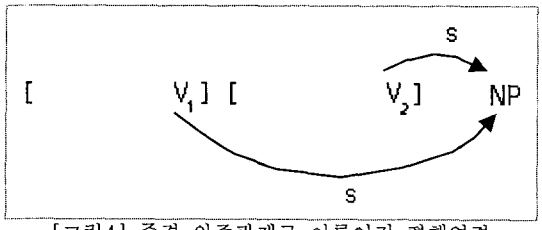
(식2)의 각 요소에 대한 설명은 이어지는 내용에서 하기로 한다.

3.3 평행 연결

본 논문에서는 대등접속의 효과적인 분석을 위하여 "평행연결"이라는 새로운 개념을 이용하는 기법을 제안한다. 평행연결이란 선접속부와 후접속부의 외부에 존재하는 명사구와 선머리어 또는 후머리어 중의 어느 하나와 의존관계가 있을 때 그 명사구와 나머지 다른 하나와도 같은 격으로 의존관계가 존재함을 말한다.



[그림3] 목적격 의존관계로 이루어진 평행연결



[그림4] 주격 의존관계로 이루어진 평행연결

위 그림에서 대등으로 접속되는 구간 외부에 있는 NP 가 선머리어 V_1 와 목적격(o)의 관계로 의존관계가 있다고 가정하자. 그렇다면 당연히 NP 는 후머리어와도 같은 목적격(o)의 관계로 의존관계가 있게 됨을 나타낸다. 이 그림과 유사하나 대칭인 상황이 다음 그림에 나타나 있다. NP가 후머리어 V_2와 주격관계(s)로 의존관계가 있다 하자. 그렇다면 당연히 NP와 선머리어 V_1과도 주격 관계로 연결되어 있다는 것을 나타낸다.

평행연결의 원리는 대등접속 분석 과정에서 중의성을 많이 제거하는 효과를 발휘한다. 구간에 대한 구문분석 후 대등접속 후보에 대하여 평행연결을 적용하여 추가되어야 하는 의존관계(Arc)를 파악한다. 이때 만약 평행 연결에 의하여 추가되는 의존관계가 그럴듯 하지 않은 것이라면, 즉 상호정보의 값이 낮은 것이라면, 해당 대등접속 후보는 가능성이 낮은 것으로 간주할 수 있다. 이 때 의존관계의 점수가 낮으므로 이것을 포함한 전체 분석구조의 점수도 낮아진다

3.4 선/후 접속부의 구조 및 어휘의 유사성

[표1] 구조 유사성에 의한 점수표

선접속부		후접속부		점수
s	자동사	s	자동사	10
s	자동사	s	타동사	4
s	타동사	s	자동사	4
s	타동사	s	타동사	6
o		o		10
×	자동사	×	자동사	10
×	타동사	×	자동사	4
×	자동사	×	타동사	4
×	타동사	×	타동사	6
s o		×	자동사	2
s o		s	자동사	8
s o		s	타동사	4
s o		o		2
s o		s o		10
s	자동사	s o		8
s	타동사	s o		4
o		s o		2
s o		s o		10
×	자동사	s o		2
×	자동사	o		8
×	타동사	o		3
o		×	자동사	8
o		×	타동사	3
o		s	자동사	1
o		s	타동사	0
s	자동사	s		1
s	타동사	s		0
s		×		1
×		s		1

대등접속이 보이는 또 다른 성질로는 선접속부와 후접속부의 구조 및 어휘의 유사성이다. 구조의 유사성이란 선접속부가 가진 격의 종류 및 수와 후접속부

가 가진 격의 종류 및 수가 유사할수록 그 대등접속의 가능성이 높다는 것이다. 여기서 특히 관련이 있는 것은 필수격들이다. 이에 기반하여 모든 가능한 구조 대응 관계 및 이에 대한 점수를 [표1]처럼 적용할 수 있다. 여기에서 부여된 점수들은 경험적인 값들로 실험에 의해 조정될 수 있다. 0점인 것은 구조 유사성이 전혀 없는 것이므로 대등접속 문장분석을 시도하지 않아도 무방하다. 이 표에서 얻는 점수가 바로 (식2)에서 보였던 structural_sim(1,2)이다.

또한, 두 접속부에 포함된 어휘의 유사성에서 가장 중요한 것은 선머리어와 후머리어의 품사가 같을수록 유사성이 높아진다는 것이다. 즉, 어느 한 쪽이 형용사이면 다른 쪽도 동사보다는 형용사일 가능성이 매우 높다는 것이다. 특히, 양쪽의 머리어가 동일 단어이면 대등접속의 가능성이 월등히 증가한다. 이런 현상과 관련된 점수가 (식2)에서 conj(1,2)의 요소로 작용했던 lexical_sim(1,2)이다. (여기에서 두 접속부의 머리어의 격을 채우는 명사구들 간의 어휘 유사성 비교는 현재로서 넣지 않았으나 추후 고려할 예정이다.)

$$\text{lexical\_sim}(1,2) = \text{sim\_pos}(V_1, V_2) + \text{sim\_word}(V_1, V_2) \quad \text{(식3)}$$

(주)$\text{sim\_pos}(V_1, V_2)$:품사가 같으면 5점, 다르면 0점을 부여
$\text{sim\_word}(V_1, V_2)$:단어가 같으면 5점, 다르면 0점을 부여

여기에서 부여되는 점수는 실험치에 의해 더 좋은 값으로 설정해 줄 수 있다.

3.5 분석 예

본 논문에서 제안하는 방법으로 대등접속을 가진 문장을 분석해 보고자 한다.

> 그 학생이 구입하고 | 사용한 컴퓨터를 친구에게 빌려 주었다 (5)

여기서 분석 가능한 모든 경우의 수는 네 가지로서 다음과 같다.

> [그 학생이 구입하고] | [사용한] 컴퓨터를 친구에게 빌려 주었다 : s | × = 1 (5a)
>
> [그 학생이 구입하고] | [사용한 컴퓨터를 친구에게 빌려 주었다] : s | o = 0 (5b)
>
> 그 학생이 [구입하고] | [사용한] 컴퓨터를 친구에게 빌려 주었다 : × | × = 6 (5c)
>
> 그 학생이 [구입하고] | [사용한 컴퓨터를 친구에게 빌려 주었다] : × | o = 3 (5d)

위의 분석 결과에서 구조 유사성 점수가 가장 큰 것은 (5c)이고, 실제로 올바른 분석 결과를 나타낸다. 이와 같이 대등접속구문의 분석에서 구조대응점수만으로도 몇 가지 예외적인 경우를 제외한 대부분의 대등하게 접속된 문장들의 대등성을 파악할 수 있다.

(6)은 (5)보다 더욱 복잡한 문장을 살펴 보기 위한 예이다. 이는 본 논문에서 제시한 평행연결 기법이 효과를 발휘한 예이다.

> 그 기와집 앞에는 마을 사람들이 아끼고 자랑하는 아름드리 느티나무가 당당한 모습으로 서 있었다. (6)

(6)에 대한 모든 경우의 분석들을 나열하면 다음과 같다. 이 때에 부사절이나 보조용언 등의 부가적인 부분은 김광백[6]에서와 같이 구문분석의 전처리 단계에서 제거되는 것으로 본다.

> 그 기와집 앞에는 마을 사람들이 [아끼고] | [자랑하는] 아름드리 느티나무가 당당한 모습으로 서 있었다 (6a)
>
> 그 기와집 앞에는 마을 사람들이 [아끼고] | [자랑하는 아름드리 느티나무가 당당한] 모습으로 서 있었다 (6b)
>
> 그 기와집 앞에는 마을 사람들이 [아끼고] | [자랑하는 아름드리 느티나무가 당당한 모습으로 서 있었다] (6c)
>
> 그 기와집 앞에는 [마을 사람들이 아끼고] | [자랑하는] 아름드리 느티나무가 당당한 모습으로 서 있었다 (6d)
>
> 그 기와집 앞에는 [마을 사람들이 아끼고] | [자랑하는 아름드리 느티나무가 당당한] 모습으로 서 있었다 (6e)
>
> 그 기와집 앞에는 [마을 사람들이 아끼고] | [자랑하는 아름드리 느티나무가 당당한 모습으로 서 있었다] (6f)

여기서 (6a),(6b),(6c)에서 "마을 사람들이" 는 선머리어와 후머리어에 주어 s의 관계로 평행연결된다. 이 경우 (6c)는 후접속부의 주어가 "아름드리 느티나무가" 로써 이미 존재하고, (6b)는 "당당한" 의 주어가 "모습" 으로 이미 존재하기 때문에 격충돌 현상이 일어나는 경우이다. 이 경우 대등구조 유사성을 따질 가치가 없으므로 후보에서 탈락되게 된다.

또, (6d)에서는 "자랑하는" 의 주어가 "느티나무" 이므로 평행연결에 의하여 역시 "아끼고" 의 주어도 "느티나무" 이어야 한다. 그러나 "마을 사람들이" 가 이미 "아끼고" 의 주격을 채우고 있으므로 격충돌 현상이 발생한다. 따라서 (6d)도 경쟁후보에서 제외되어도 무방하다.

따라서 (6e),(6f)는 평행연결 현상으로 인한 격충돌현상이 일어나지 않는다. [표2]는 각각의 경우에 대한 점수비교를 비교해 본 것으로 각 요소의 합계가 가장 큰 분석이 올바른 것으로 판단될 수 있다. (6e)

와 (6f)는 (6a)에는 있는 I(느티나무,o,아끼다)와 I(사람들,s,자랑하다)의 두 연결이 없다. 그러나 이 둘의 상호정보는 모두 양의 값을 가진다.

[표2] 각 문장의 점수 비교

	(6a)	(6e)	(6f)
상호정보합 $\sum_i I(arc_i)$	10	7	7
structural_sim(1,2)	6	1	4
sim_pos(V_1, V_2)	5	0	5
sim_word(V_1, V_2)	0	0	0
Score(T)	21	8	16

따라서 가장 높은 점수를 얻은 (6a)가 올바른 분석으로 선택된다. 이 분석은 실제로 올바른 분석이다.

4. 실험 결과

본 논문에서 제안한 기법의 타당성을 입증하기 위하여, 설명문, 신문기사, 교과서, 수필 등에서 추출한 100개의 문장에 대하여 대등접속 분석의 정확도를 측정하였다. 주어진 총 대등접속을 가진 문장의 개수에서 선행부 및 후행부의 범위를 정확히 알아낸 문장의 개수를 성능의 척도로 한다.

$$\text{정확율 } P = \frac{\text{범위가 바른 문장의 개수}}{\text{총 병렬접속 문장의 개수}} \quad \text{(식4)}$$

주어진 100개의 문장에 대한 실험치는 [표2]과 같다.

[표3] 대등 구문 분석 기법 실험의 정확율

	길이 ≤ 15	길이 > 15
문장 개수	58	42
바른 분석 수	49	32
정확도	84.5%	76.2%
평균정확도	81.0%	

오분석을 일으키는 가장 큰 원인으로 판명된 것은 생략현상이었다. 그 이유는 생략현상이 있으면 구조 유사성에 관한 정확한 판단이 어렵고 오히려 나쁜 점수를 주게 되기 때문이다. (7)은 이와 같은 생략현상이 초래하는 오분석을 보여준다.

기술 분야에 고석달 선생이 새로 충원되었고 일반 서무에는 사무직 김석현 씨를 배정해 주었다. (7)

(7)의 경우 역시 여러 경우의 수를 따져 볼 수 있지만 대등접속 분석시스템 내에서의 가장 우세한 분석으로는 다음과 같이 (7a)와 (7b)를 들 수 있다.

[기술 분야에 고석달 선생이 새로 충원되었고] | [일반 서무에는 사무직 김석현 씨를 배정해 주었다] (7a)

기술 분야에 고석달 선생이 새로 [충원되었고] | [일반 서무에는 사무직 김석현 씨를 배정해 주었다] (7b)

(7)의 올바른 분석은 (7a)이다. 그러나 후접속부에서 용언의 주어가 생략되었으므로, 본 시스템은 후접속부에 "고석달 선생이" 를 주어로 추가할 것을 시도한 후, 그 구조 유사성 점수를 증가시킴으로써 오분석을 초래한다. 즉, (7a)는 "주어 s + 자동사 | 목적어 o + 타동사" 의 구조이고, (7b)는 "자동사 | 목적어 o + 타동사" 에서 주어 s로서 "고석달 선생이" 가 선·후머리어에 각각 평행연결 되어야 하는 구조이다. 이들의 구조 유사성 점수를 [표1]의 점수로서 비교해 보면 (7a)와 (7b)는 각각 1점과 8점이다. (7b)는 평행연결에 의하여 (7a)에는 없는 연결 즉 "선생이" 가 "배정하다" 의 s 로 되는 연결을 더 가지게 된다. 이때 이 연결의 상호정보는 양의 값이다. 따라서 결과적으로 분석시스템은 (7b)가 올바른 분석인 것으로 판단하는데 이것은 잘못된 분석이다.

이러한 생략현상이 배제되지 않은 문장들까지 시스템이 분석해 내는 것은 이어지는 연구에서 극복해 내야 할 과제이다.

5. 결론

본 논문에서는 대등접속을 포함한 한국어 문장의 구문분석에 대하여 살펴 보았다. 여기서의 대등접속 분석 기법은 구간 분할에 기반한 구문분석이라는 문장 전체의 구문분석 전략에 맞도록 개발되었다. 본 기법의 가장 큰 장점으로는 가능한 대등접속 후보의 수가 기존의 기법보다 현저히 줄어든다는 점이다. 그 이유는 후머리어와 선접속부의 시작 단어를 특정한 구간 안에서만 탐색하기 때문이다. 또 다른 이유로는 이미 분석된 결과 즉 격 관계가 밝혀진 상태에서 분석이 이루어지므로 선접속부의 시작 단어 후보의 수가 대폭 감소된다.

본 기법은 구조의 유사성, 평행연결의 현상을 효과적으로 이용하므로 불필요한 대등접속 분석을 잘 제거할 수 있다. 본 기법이 가진 문제점은 선접속부와 후접속부가 특정 구간 내에 존재한다는 가정이다. 향후의 연구는, 이 가정을 없앨 경우 어느 정도 성능이

향상되며, 이 경우 필요한 계산의 정도가 어느 정도 증가하는지에 대한 고찰이 될 것이다. 또, 생략현상이 있는 문장의 경우 신뢰성 있는 상호정보를 이용함으로써 생략현상으로 인해 감소되는 구조유사성 점수를 보완할 수 있다. 따라서, 보다 신뢰성을 가진 상호정보의 개발이 과제로 남는다.

6. 참고 문헌

[1] R. Agarwal and L. Boggess, " A simple but useful approach to conjunct identification," in Proceedings of the 30th annual meeting of ACL, 1992.

[2] S. Kurohashi and M. Nagao, "A syntactic analysis method of long japanese sentences based on the detection of conjunctive structures," Computational Linguistics, V.20, No.4, pp.507-534, 1994.

[3] A. Okumura and K. Muraki, "Symmetric pattern matching analysis for English coordinate," in Proceedings of the 4th conference on Applied NLP, 1994.

[4] 윤준태, 송만석, "한국어의 대등접속 구문 분석," 정보과학회 논문지, 24권3호, pp.326-336, 1997.3.

[5] 엄미현,나동열, "한국어의 구조적 중의성 해소에 관한 연구," 연세대 대학원 석사학위논문, 1997.

[6] 김광백, 박의규, 나동열, 윤준태, "구간 분할 기반 한국어 구문분석," 한글 및 한국어처리 '2002 학술대회 논문집, 2002.10.

[7] 조형준, "한국어 병렬구문과 결합범주문법에서의 구문분석," 한국과학기술원 석사학위논문, 1999.12.