결합범주문법과 구문분석[1]

조형준○ 박종철
한국과학기술원 전산학과 및 첨단정보기술 연구센터

Combinatory Categorial Grammar and Parsing

HyungJoon Cho Jong C. Park
Dept. of Computer Science and Advanced Information Technology Research Center (AITrc)
Korea Advanced Institute of Science and Technology

요 약

본 논문에서는 결합범주문법으로 한국어를 처리할 때 구문분석과정에서 복잡도를 높이는 역할을 하는 spurious ambiguity와 구조적 모호성이 있는 명사구 접속에 대해서 논한다. 통사적 처리와 의미적 처리가 동시에 수행되는 결합범주문법의 특징을 사용해서 spurious ambiguity로 인해 발생하는 복잡도를 줄이는 방안을 제시하고 접속항에서 접속의 중심이 되는 명사들 간의 공기유사도를 이용해서 접속항 선정에서 발생하는 복잡도와 오분석을 줄이는 방안을 제시한 뒤 이의 개선방안을 논의한다.

1 서 론

일반적으로 병렬문이란 두 개 이상의 구나 문장이 서로의 의미에 영향을 끼치지 않고 동등한 관계로 연결되어 있는 문장을 말한다. 병렬문에서는 일반적으로 평서문에서는 허용되지 않는 생략이 허용되고, 단어가 가지는 모호성 때문에 접속항 선정에 어려움을 유발한다. 접속항 선정에서 발생하는 문제점으로 인해 전체적인 시스템의 성능이 저하되거나 아예 문장을 처리하지 못하는 경우도 발생한다. 이런 문제점들을 해결하기 위해 구문분석의 전처리의 개념으로 접속항의 범위를 정해줌으로써 시스템의 성능을 향상시키고자하는 연구가 있었다[10, 3, 11].

결합범주문법 (combinatory categorial grammar[15, 16])은 단일화 (unification)에 기반한 어휘문법 (lexicalized grammar)이다. 결합범주문법은 병렬구문 (coordination), 추출 (extraction), 원거리 종속 관계 (long distance dependency)를 설명할 수 있다고 알려져 있고 최근에는 억양 (intonation), 어순뒤섞기 (scrambling), scope 등의 자연언어 현상에 대해서도 별도의 약정 (stipulation)이 없이 설명을 제공하는 것으로 알려져 있다[14, 9, 13]. 이와 같이 결합범주문법은 다른 문법들이 설명하지 못하는 언어현상들을 비교적 단순한 규칙을 사용하여 간결하게 설명할 수 있는 장점이 있는 반면 type raising과 composition으로 인해 시스템의 복잡도가 증가하는 문제도 있다. 이와 함께 구조적으로 모호한 명사구의 접속현상은 복잡도를 높일뿐만 아니라 잘못된 분석결과를 제시할 가능성도 증가한다.

본 논문에서는 병렬문에서 발생하는 문제점을 해결하기 위해 기존에 진행되었던 연구들과 함께 한국어 처리에 결합범주문법을 적용시켰을 때 구문분석과정에서의 복잡도를 높이는 spurious ambiguity와 명사구 접속의 구조적 모호성에 대해서 논하고 이의 처리방안을 제시한다.

[1]본 연구는 첨단정보기술 연구센터를 통하여 과학재단의 지원을 받았음.

2 관련 연구

본 절에서는 병렬문의 접속항 선정에서 발생하는 문제점을 해소하기 위해 기존에 진행되었던 연구들과 한국어 병렬문 처리를 위한 결합범주문법에 대해서 설명한다.

2.1 접속항 선정의 모호성 해소를 위한 방법들

접속항 선정의 모호성을 해소하기 위해 구문분석전의 전처리의 개념으로 고안된 방법들을 복합적인 정보를 이용한 방식, 공기유사성을 이용한 방식, 품사패턴을 이용한 방식으로 나누어서 설명한다. 여기에 제시된 방법들은 언어처리에 사용되는 문법이 자체적으로 접속항을 선정할 수 없는 경우를 위해 제안된 방법이다.

2.1.1 복합적인 정보를 사용하는 방법

복합적인 정보를 사용하는 방식에서는 병렬문을 이루는 접속항 사이에 형태적으로나 의미적으로 어느정도 유사성을 띄고 있다는 가정에서 출발한다. 이 때 유사성을 측정할 수 있는 요소로는 태거 (tagger)를 통해 나온 각 단어의 품사정보나 의미정보 또는 시스템에서 자체적으로 정의한 자질 (feature)들을 사용한다.

[8]의 방식은 간단한 알고리즘을 사용하면서 비교적 좋은 효율을 보인다. semi parser를 사용하여 입력문장을 이루고 있는 각각의 단어들의 의미정보와 형태정보를 할당한다. 할당된 정보를 사용해서 접속항을 선정하는 과정은 다음과 같다.

(a) 접속사 (and, or, but ...)가 나올 때까지 문장의 요소를 스택에 넣는다.

(b) 접속사가 나오면 우접속항 (post-conjunct)은 바로 그 뒤에 오는 구로 간주한다.

(c) 스택에 있는 요소들을 하나씩 꺼내서 각 요소들의 형태정보나 의미정보를 우접속항의 정보들과 비교한다. 형태정보와 의미정보가 일치할 경우의 유사도가 가장 높다고 간주되고 동일한 품사정보와 호환성이 있는 의미정보를 가지고 있을 경우, 동일한 품사정보를 가지고 있을 경우의 순으로 유사도가 정의된다. 이렇게 정의된 유사도를 기준으로 스택에서 있는 요소들 중 가장 유사도가 높은 것을 선택한다.

이 알고리즘은 비교적 단순하고 도메인에 독립적이면서도 약 81%라는 비교적 좋은 정확도를 나타낸다. 반면 두 개의 항으로 이루어진 병렬문의 접속항 선정에만 적용이 가능하고 접속항의 경계를 구별하지 못하고 접속항의 시작부분만 구별할 수 있다는 단점이 있다.

[11]의 방식 역시 영어 병렬문을 처리하기 위해서 제시되었고 접속항 선정시에 접속항의 대칭성에 기반해서 고안되었다. 대칭성을 측정하기 위해 접속패턴을 구(절) 대칭패턴, 단어 대칭패턴, 형태적 대칭패턴, 복합적인 대칭패턴의 4가지로 나누었다. 복합적인 형태를 제외한 나머지 3가지 병렬형태를 구분하기 위해서 구자질 (phrase feature), 단어자질 (word feature), 형태자질 (morphological feature)를 정의하고 각각의 단어에 각각의 자질을 할당한다. 이렇게 정의된 단어집합에 대해서 접속사 앞, 뒤의 모든 가능한 단어집합에 대해서 대칭성을 검사하고 가장 대칭적인 쌍을 찾아서 구문 분석기에 알려준다. 약 75%의 정확도를 나타낸다. 이 접근방법에서는 단어의 대칭성을 측정하기위해 구자질, 단어자질, 형태자질에 가중치를 주게 되는데 이때 가중치를 산정하는 방식과 접속항에서 생략현상이 발생했을 경우 이를 처리하는 방법이 문제가 될 수 있다.

2.1.2 공기 유사성을 사용한 방법

공기유사성을 사용한 방법에서는 한국어 명사구 병렬에서 나타나는 모호성을 해결하기 위해 제시된 방법이다[3]. 이 방법에서는 대량의 말뭉치로부터 명사와 동사간의 쌍을 찾고 각 명사가 특정 동사에 대해 어떤 격으로 쓰였는지에 대한 통계정보를 근거로 명사간의 유사도를 구한다. 두 명사 n_1과 n_2간의 공기유사성 $DSim(n_1, n_2)$은 다음과 같이 정의된다.

$$
\begin{aligned}
DSim(n_1,n_2) &\equiv \frac{2\cdot\sum_{g\in G}\mid CV_g(n_1,n_2)\mid}{\sum_{g\in G}\mid V_g(n_1)\mid+\sum_{g\in G}\mid V_g(n_2)\mid}\\
G &\equiv \{sbj, obj, loca, inst, modi\}\\
V_g(n) &\equiv \{v \mid v \text{ is a verb such that } f_g(n,v)\geq 1\}\\
\mid V_g(n)\mid &\equiv \sum_{v\in V_g(n)} f_g(n,v)\\
CV_g(n_1,n_2) &\equiv V_g(n_1)\cap V_g(n_2)\\
\mid CV_g(n_1,n_2)\mid &\equiv \sum_{v\in CV_g(n_1,n_2)} min\{f_g(n_1,n_2), f_g(n_2,v)\}
\end{aligned}
$$

즉 두 명사가 동일한 문법 관계로 동일한 동사와 함께 쓰이는 비율이 높을수록 이 두 명사의 유사도가 높아지게 되는 것이다.

공기유사성을 사용할 경우 (1)과 같이 수식어의 범위에 의해 생기는 모호성을 해결할 수 있고 (2)와 같이 의미적 유사성에 의해 해결될 수 있는 문제를 별도의 의미체계를 구축하지 않고도 해결할 수 있다.

(1) 프로그램의 공간 복잡도와 시간 복잡도의 정의

(2) 사과와 배를 담은 그릇

(3) 이 노드는 데이타 요소와 리스트(0.215)의 다음 원소(0.254)를 지시하는 포인터(0.204)를 포함한다.

하지만 여러가지 의미를 가진 명사에 대해서는 공기정보자체가 부정확해질 가능성이 있으며 말뭉치에 없는 데이타에 대해서는 대처할 방안이 없다는 단점과 함께 (3)과 같이 접속항을 이루는 명사 자체가 의미적으로 관계가 적을 경우 해결할 방안이 없다. 공기유사성 정보를 추출하기 위해 사용된 말뭉치에서 추출한 문장에 대해서 실험했을 때 정확도는 약 85.8%였고 적용도는 약 95.9%였다. 공기유사성을 사용한 방법은 WordNet과 같이 대용량의 지식베이스가 갖추어지지 않은 상태에서 의미체계에 대한 대안으로 적합하다.

2.1.3 품사패턴을 사용하는 방법

품사패턴을 사용하는 방법은 품사태깅된 말뭉치로부터 병렬문을 추출하고 좌접속항과 우접속항의 품사패턴을 학습시키고 학습된 품사패턴을 기반으로 접속항의 모호성을 해결하고자하는 접근방법이다[6].

(4) 나는 [검정 빵과 흰 빵]을 먹은 사람을 보았다.

(5) 파인애플이나 망고를 수입하는 나라

(4)에 해당되는 좌측 접속항의 품사패턴은 (N N CONJ:paa etm ncn jco)이고 우접속항의 품사패턴은 (ETM N:ncn ncn jcj)이다.[2] 하나의 병렬문에 여러개의 품사패턴을 적용할 수 있

[2](우접속항의 명사구형성정보:좌접속항의 품사열 패턴)

을 때 최장일치법을 적용시켜 가장 긴 길이의 패턴을 적용시켰다. 정확률은 약 71.6%였다. 품사패턴을 이용할 경우 (5)와 같이 접속항 선정에 구조적 모호성이 있을 경우 '파인애플'에 해당하는 접속항을 찾아내는데 문제가 있다.[3]

2.2 한국어 병렬문의 처리를 위한 결합범주문법

서술어 중심언어인 한국어에서는 서술어가 핵심성분이 되고, 주어나 목적이 등 나머지 성분은 주변성분이 된다. 주변성분들은 때에 따라서 탈락되기도 하지만 핵심 성분인 서술어는 탈락되지 않는다[2]. 그러나 한국어 병렬문에서는 (6), (7)과 같이 서술어가 생략되는 경우가 있다.

(6) [철수는 의사가], [영희는 선생님이] 되었다.

(7) [고추값이 오르면 고추를], [돼지고기값이 오르면 돼지고기를] 수입한다.

서술어의 생략이 발생했을 때 문제가 되는 것은 접속항이 되는 비정규문장성분 (nonstandard constituent) 즉 [철수가 의사가], [고추값이 오르면 고추를]의 통사범주와 의미를 표현하는 방법이 문제가 되고 결합범주문법에서는 type raising과 composition을 이용하여 접속항의 통사범주와 의미를 설명한다. 표 1은 한국어에서 발생하는 병렬문을 처리하기 위해 제시된 축약규칙이고[1], 그림 1은 제시된 축약규칙을 사용하여 주어진 문장을 처리하는 과정을 보인다.

결합범주문법에서 conjoin operation이 적용되기 위해서는 기본적으로 좌접속항과 우접속항의 문법 범주가 동일해야 한다. 이들 접속항을 이루고 있는 요소들은 일반적인 평서문에서 서술어라는 functor와 결합함으로써 문장을 구성한다. 하지만 좌접속항에서 서술어가 생략되었기 때문에 이들 범주로는 병렬문의 연산에 필요한 접속항을 구성할 수 없다. 이를 처리하기 위해서 결합범주문법에서는 type raising된 범주를 할당하고 이 범주들에 다시 composition 축약규칙을 적용시킴으로써 비정규문장성분의 통사범주와 의미를 유도한다. 하지만 결합범주문법의 특성인 spurious ambiguity와 병렬문 축약규칙으로 인해 구조적 모호성이 있는 명사구 접속현상을 처리할 때는 복잡도를 증가시키는 문제점이 존재한다.

3 Spurious Ambiguity와 명사구 접속

본 절에서는 결합범주문법에서 사용하는 축약규칙 중 하나인 composition이 가지고 있는 associative nature[17]로 인해 발생하는 spurious ambiguity와 이것이 구문분석기에 끼치는 영향에 대해 논의한 뒤 결합범주문법을 사용해서 구조적 모호성을 가지고 있는 명사구 접속 현상을 처리하고자 할 때 나타나는 현상에 대해서 설명한다.

[3]구조적 모호성이 발생했을 경우에는 접속항 후보의 의미나 공기정보를 사용해서 해결할 수 있다.

3.1 Spurious Ambiguity

결합범주문법이 병렬문을 처리할 수 있는 이유는 접속항으로 나타나는 비정규문장성분을 type raising이나 composition을 사용해서 하나의 문법범주로 축약할 수 있고 또 축약과정에서 비정규문장성분의 의미를 유도해 낼 수 있기 때문이다. (8)은 3개의 문장성분으로 이루어진 문장이다. (8)을 결합범주문법의 type raising과 composition에 의해 구문분석하는 과정에서 나타나는 구조는 (9), (10)와 같다. 이는 결합범주문법의 associative nature로 인해 나타나는 결과이다. 이와 같이 하나의 문장에서 여러 개의 구조가 유도되는 특성은 (11), (12)와 같이 병렬문을 처리할 때, 접속항들을 하나의 통사범주로 축약하는 것을 가능하게 한다.

(8) 철수가 사과를 먹었다.

(9) [[철수가 사과를] 먹었다.]

(10) [철수가 [사과를 먹었다.]]

(11) [[[철수가 사과를] , [영희가 딸기를]] 먹었다.]

(12) [철수가 [[손을 씻] 고 [밥을 먹었다.]]]

이렇게 주어진 통사범주에 대해서 여러가지 분석결과가 나오는 현상을 spurious ambiguity라고 한다. Spurious ambiguity에 의해 동일한 분석결과가 구문분석과정에서 반복해서 나타날 경우 시스템의 성능을 저하시킨다. 이런 현상은 결합범주문법으로 병렬문을 처리하는 과정에서 두드러 진다. (13)은 두개의 동사구가 연결된 병렬문이다[4].

(13) 철수는 사과를 먹고 영희는 딸기를 먹는다.

(13)을 구문분석하는 과정에서 나타나는 분석결과는 표 2와 같으며 문장의 길이가 길어질수록 가능한 분석결과의 개수가 급격히 증가한다.

· [[철수는	사과를]	먹]고	[[영희는	딸기를]	먹]는다.
· [[철수는	사과를]	먹]고	[영희는	[딸기를	먹]]는다.
· [철수는	[사과를	먹]]고	[[영희는	딸기를]	먹]는다.
· [철수는	[사과를	먹]]고	[영희는	[딸기를]	먹]]는다.

표 2: 병렬문에서의 spurious ambiguity

3.2 명사구 접속의 구조적 모호성

Spurious ambiguity가 단순히 구문분석과정에서 복잡도를 증가시킨다는 역할을 하는 반면 명사구 접속의 구조적 모호성은 구문분석과정에서 복잡도를 증가시키는 역할 뿐만 아니라 오분석을 행할 가능성도 증가시킨다. (14)는 결합범주문법으로 처리할 때 접속항 선정 과정에서 구조적 모호성을 보이는 문장이다.

[4]결합범주문법에서는 좌접속항 후보와 우접속항 후보의 통사범주가 동일할 때 병렬문 축약규칙을 적용시킨다.

규 칙					규칙이름	축약기호
X/Y		Y	$\longrightarrow$	X	Forward Application	$>$
Y		$X\backslash Y$	$\longrightarrow$	X	Backward Application	$<$
X	$conj$	X	$\longrightarrow$	X	Coordination	$<\phi^n>$
X/Y		Y/Z	$\longrightarrow$	X/Z	Forward Composition	$>B$
$Y\backslash Z$		$X\backslash Y$	$\longrightarrow$	$X\backslash Z$	Backward Composition	$<B$
X/Y		$Y\backslash Z$	$\longrightarrow$	$X\backslash Z$	Forward Crossed Composition	$>B_x$
X			$\longrightarrow$	$T/(T\backslash X)$	Forward Raising	$>T$
X			$\longrightarrow$	$T\backslash(T/X)$	Backward Raising	$<T$

표 1: 한국어 병렬문 처리를 위한 결합범주문법

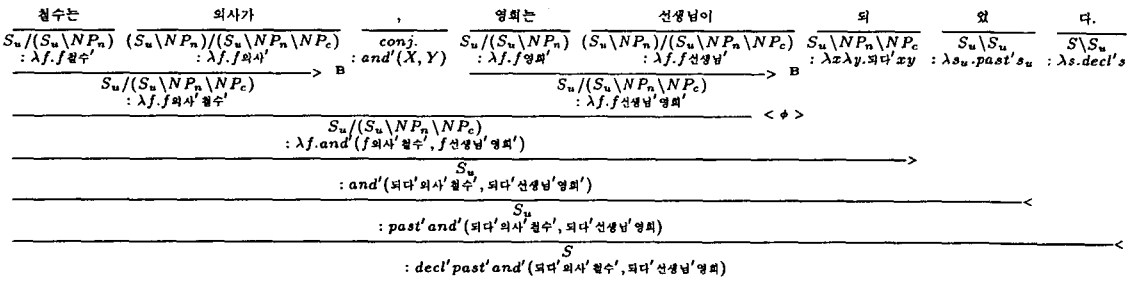

그림 1: 철수는 의사가, 영희는 선생님이 되었다.

(14) 정치의 장래나 농민들의 생활상의 문제와는 거리가 멀다.

표 3은 구문분석과정에서 나타날 수 있는 접속항 후보들을 나열한 것이다.

· [정치의	장래]나	[농민들]의	생활상의	문제
· [정치의	장래]나	[농민들의	생활상]의	문제
· [정치의	장래]나	[농민들의	생활상의	문제]
· 정치의	[장래]나	[농민들]의	생활상의	문제
· 정치의	[장래]나	[농민들의	생활상]의	문제
· 정치의	[장래]나	[농민들의	생활상의	문제]

표 3: 명사구 접속항 선정의 모호성

결합범주문법에서는 접속사의 좌우에 있는 접속항 후보들의 통사범주가 동일할 때 접속항축약규칙이 적용되기 때문에 표 3과 같은 분석결과가 발생한다.

3.3 말뭉치에서의 병렬문과 접속항

말뭉치를 분석한 결과를 토대로 병렬문이 언어현상에서 차지하는 비중을 알아보고 명사구 접속의 모호성을 해결하기위해 적용되는 구조적 유사성을 이용한 방법과 의미적 유사성을 이용한 방법의 차이를 논한다. 한국어에서 병렬문이 차지하는 비중을 알아보기 위해 KAIST 말뭉치를 분석했다. KAIST 품사태깅 말뭉치는 신문기사가 40,428 어절, 수필 41,666 어절, 교과서 50,208어절, 기술문서 2,729어절, 소설 40,498 어절, 모두 175,524 어절로 구성되어 있으며 총 17,123 문장으로 이루어져 있다[5]. 병렬을 이루는 형태를 조사한 결과는 표 4에 보인 것과 같다. 분석은 품사태깅된 말뭉치에서 기계적으로 조건에 부합되는 문장을 추출한 뒤, 병렬문의 여부를 수동으로 판별하였다.[5] 표 4를 통해 병렬문이 양적으로도 많이 나타나는 문형임을 알 수 있다[1].

표 5는 표 4에 나타난 병렬문 중 접속격 조사로 연결된 문장 중에서 무작위로 237문장을 추출해서 분석한 결과이다. '명사+명사'의 경우는 명사접속이 나타난 문장이 구조적 모호성을 가지지 않고 단순히 명사와 명사의 접속으로 이루어진 접속을 말한다. (15)의 경우 단순히 명사가 접속된 형태의 접속이다. '접속항 구조가 동일할 경우'[6] 는 좌접속항과 우접속항 전후에 모호성을 유발하는 구나 절이 존재하지만 접속항의 구조가 동일한 형태의 문장을 말한다. (16)이 이런 문장에 해당한다. '접속항의 구조가 동일하지 않을 경우'는 좌접속항과 우접속항 전후에 모호성을 유발하는 구나 절이 존재하고 접속항

[5]병렬문의 여부는 문장의 의미에 따라 판별된다. 따라서 품사태깅에만 의존해서는 병렬문의 여부를 판별할 수 없다.

[6]구조적으로 동일하다는 것은 접속항의 수식구조가 동일한 경우를 말한다.

- [수식어 명사] 접속사 [수식어 명사] 등

	신 문	수 필	교과서	기술문서	소 설	합 계
대등적 연결어미 (문장수/총문장수)	21.4% (728/3403)	33.7% (1092/3239)	20.1% (1101/5485)	40% (46/115)	44.5% (2174/4881)	30.0 % (5142/17123)
접속격 조사 (문장수/총문장수)	19.7% (671/3403)	10.7% (352/3239)	14.8% (812/5485)	20% (23/115)	6.6% (320/4881)	12.7% (2178/17123)
접속 부사 (문장수/총문장수)	2.7% (91/3403)	0.68% (22/3239)	1.1% (58/5485)	5.2% (6/115)	0.35% (17/4881)	1% (194/17123)
쉼표 (문장수/총문장수)	4.5% (154/3403)	3.9% (126/3239)	3.9% (216/5485)	18.3% (21/115)	0.88% (43/4881)	3.5% (605/17123)

표 4: 병렬문의 구성방법 및 발생빈도

의 형태적 구조가 동일하지 않는 문장이다. (17)은 '접속항의 구조가 동일하지 않을 경우'에 해당한다. 각 형태별 분석결과는 단순히 '명사+명사'의 경우가 38.4%, '접속항의 구조가 동일한 경우'가 46.8%, '접속항의 구조가 동일하지 않을 경우'가 14.8%로 명사구 접속에서 약 85.2%가 접속항의 구조가 동일한 것으로 나타났다. 하지만 구조적 유사성을 따져서 단순히 적용시킬 경우 (18)과 같은 문장을 (19)와 같이 분석할 가능성이 있으며 통사의 유사성이 아닌 의미적으로 연결된 병렬문은 오분석할 가능성이 있다.

	문장수/총문장수
명사+명사	91/237 (38.4%)
접속항의 구조가 동일할 경우	111/237 (46.8 %)
접속항의 구조가 동일하지 않을 경우	35/237 (14.8 %)

표 5: 명사구 접속의 구조적 분석

(15) 북한은 [핵무기 개발]과 [생화학무기 보유]에 대한 집착을 버리지 못하고 있다.

(16) [세계정세의 흐름]과 [주변환경의 변화]를 직시해야 할 것이다.

(17) 광운대 부정입학사건은 오늘날 [사학이 당면한 재정난]과 [대학경영자들의 도덕적 수준]을 말해 준다.

(18) 농민들의 [피땀]과 [희생]의 바탕 위에 농협은 면, 군, 도, 중앙에 이르기까지 정부 조직 다음으로 큰 막강한 조직으로 올라섰다.

(19) [농민들의 피땀]과 [희생의 바탕 위]에 농협은 면, 군, 도, 중앙에 이르기까지 정부 조직 다음으로 큰 막강한 조직으로 올라섰다.

구조적 유사도를 사용해서 분석한 방법이 표층정보를 바탕으로 많은 정보를 필요로하지 않는 방법인데 반해 의미적으로 명사구접속의 접속항을 찾고자하는 방법은 WordNet나 시소러스같이 대규모의 지식베이스를 필요로하는 방법이다. '의미적 유사도가 높은 경우'는 명사구접속이나 단순한 명사들간의 접속에서 접속항의 중심어가 의미적 유사도을 가지고 있는 문장이다. (20)은 접속을 이루는 중심어인 '절차'와 '과정'이 의미적 유사도을 가지고 있는 문장이고 (21)은 중심어의 의미적 유사도가 떨어지는 예문이다. 의미적 유사도를 적용시켜 접속항의 모호성을 해결할 때 문제가 되는 것은 형태는 동일하지만 문장에서 여러가지 의미로 쓰이는 단어들이다. (22)의 '교사'가 이런 단어에 해당한다. 분석결과는 표 6과 같다.[7]

	문장수/총문장수
의미적 유사도가 높은 경우	209/237 (88.1%)
의미적 유사도가 떨어지는 경우	28/237 (11.9%)

표 6: 명사구 접속의 의미적 분석

(20) 교육이란 목적보다 [절차]와 [과정]이 올바라야 비로소 그 기능을 다 할 수 있다.

(21) [의심가는 재산]과 [불투명한 신고]에 대해서는 기필코 대국민 해명과 설명이 있어야 한다.

(22) 오늘의 대학은 폭증하는 교육수요에 발 맞춰 [교사]나 [부대시설]을 많이 세워야한다.

표 6을 통해 접속항을 이루는 중심명사들이 의미가 접속항을 이루는 중요한 요소 중의 하나임을 알 수 있다.

4 구문 분석

본 절에서는 4.1, 4.2에서 논의한 문제점들이 실제 구문분석과정에서 어떻게 복잡도를 높이는지 알아보고 구문분석과정에서의 복잡도를 줄이는 방안에 대해서 논의한다.

[7] 명사간의 의미적 유사도는 사전을 참고로 수동으로 판단하였다.

4.1 의미구조를 이용한 spurious ambiguity의 개선

결합범주문법을 사용한 구문분석기는 CKY 알고리즘을 사용하여 구현되었다. 구문분석과정에서 spurious ambiguity가 문제가 되는 이유는 동일한 의미를 가진 분석을 여러 번 반복적으로 행함으로써 시스템의 성능에 악영향을 끼친다는데 있다. 이는 구문분석기의 셀 (cell)에 동일한 의미의 분석결과를 반복적으로 기록하려고 할 때 이를 방지함으로써 해결할 수 있으며 이 해결방안은 결합범주문법이 통사적 처리와 동시에 의미처리를 수행할 수 있다는 특징을 이용한 것이다.

(23) 철수는 사과를 먹고 영희는 딸기를 먹는다.

표 7은 (23)을 구문분석했을 때 나타나는 파싱테이블의 일부이다.[8] 표 7에 나타나는 각 단어에 대한 통사범주는 표 8과 같다.

단어	통사범주
철수가	$s/(s\backslash np_n)$
사과를	$(s/np_n)/(s\backslash np_n\backslash np_a)$
먹	$s\backslash np_n\backslash np_a$
고	conj.
영희가	$s/(s\backslash np_n)$
딸기를	$(s/np_n)/(s\backslash np_n\backslash np_a)$

표 8: 통사범주

셀 (C1,R3), (C5,R3)에는 동일한 의미를 지니지만 서로 다른 순서로 결합된 2개씩의 분석결과가 기록되어 있고 이로 인해 셀 (C1,R7)에는 동일한 의미를 지닌 4개의 분석결과가 나타나고 있다. 이런 현상은 각 셀에 들어있는 의미와 동일한 의미를 가진 분석결과가 동일한 셀에 기록되려고 할 때 이를 방지함으로써 해결할 수 있다. 이런 방법을 쓰면 셀 (C1,R3)와 셀 (C5,R3)에는 각각 하나씩의 분석결과가 기록되고 이로 인해 셀(C1,R7)에도 하나의 결과만 기록되게 된다. 이런 방식으로 긴문장에 대해서 동일한 의미를 가진 여러 개의 분석결과가 나오는 현상을 방지할 수 있다.

4.2 공기정보를 이용한 접속항의 구조적 모호성 해소

[3]에서 설명한 방법은 구문분석 전처리의 개념으로 접속항의 범위를 미리 선정해 주는 역할을 한다. 이에 비해 결합범주문법으로 구조적으로 모호한 명사구 병렬문을 처리할 때 나타나는 문제점은 가능한 모든 접속항 후보를 다 찾아내어 기록해 준다는데 있다. 접속항 후보들은 병렬문 축약규칙에 의해 선정되므로 통사적으로는 문제가 없지만 의미적으로 문제가 있는 항이 선정될 가능성이 있다. 필요없는 분석결과로 인해 구문분석 과정에서 발생하는 복잡도를 줄이는 측면에서 뿐만 아니라 구문분석의 정확도를 높이는 측면에서도 의미적으로 틀린 접속항 후보가 셀에 기록되는 것을 막을 필요가 있다. 4.3절의 분석결과에 의하면 명사구 접속항의 중심어 중에서 약 88.1%가 의미적으로 연관이 있으므로 접속항 후보들 중에서 접속항의 중심어의 유사도가 떨어지는 후보를 셀에 기록되는 것을 방지하는 것은 타당한 방법으로 생각된다. 이 방법을 사용하기 위해서는 WordNet이나 시소러스 같은 대규모 지식베이스가 필요하지만 현재 한국어에 관해 일반적으로 신뢰성이 높고 대용량 정보를 가지고 있다고 인정되는 지식베이스를 구하기 힘든 현실 때문에 대안으로 [3]에서 설명한 공기 유사성을 사용한 방법을 응용하였다. 알고리즘은 다음과 같다.

(a) 접속항 축약규칙이 적용될 때 접속항 후보의 통사범주를 살핀다.

(b) 명사나 명사구 접속을 나타내는 통사범주일 때는 (c)로 가고 아닐 경우에는 셀에 기록한다.

(c) 접속항 후보에서 접속의 대상이 되는 중심명사를 찾는다.

(d) 중심명사들의 공기 유사성을 계산한다.

(e) 이미 좌접속항의 중심명사에 병렬문 축약규칙이 적용된 셀이 있을 경우는 공기유사성을 비교한다.

(f) 새로 선정된 접속항 후보의 공기유사성이 높을 경우에는 셀에 기록하고 기존에 셀에 기록되어있던 후보를 삭제한다. 반대의 경우 기존에 기록되어 있던 후보는 그대로 두고 새로운 접속항 후보는 버린다.

(g) 새로운 접속항 후보가 셀에 기록되었을 경우, 접속항 리스트를 갱신한다.

명사간의 공기 유사 정보를 구하기 위해 KAIST에서 구축한 트리태깅된 말뭉치를 사용하였다[7]. 말뭉치는 총 31,000여 문장, 352,730 어절로 이루어져 있고 한 문장의 길이는 평균 11.35 단어이다. 사설, 경제학, 종교, 공상과학, 탐험, 일반소설, 역사 등의 영역을 담고 있다. 수동으로 태깅되었기 때문에 태깅의 정확도 면에서는 구문분석기를 사용해 태깅된 말뭉치에 비해 높은 정확도를 가지고 있다. 공기유사성 추출과정에서 공기유사도를 추출할 때 고려된 동사와의 격관계를 주격, 목적격, 부사격, 보격, 관형격으로 제한하였다. 추출된 공기유사도 정보는 공기유사도 사전의 형태로 저장되며 구문분석기가 명사구 접속이 일어난 병렬문을 처리할 때 사용된다.

[8] '먹는다'의 '는다'에는 시제와 상에 대한 정보가 들어있지만 설명을 간략히 하기 위해 '는다'를 제외한 나머지 문장에 대해서만 설명한다.

R7	s and'먹다'사과'철수,먹다'딸기'영희' [[철수가 사과를] 먹]고 [[영희가 딸기를] 먹] [[철수가 사과를] 먹]고 [영희가 [딸기를 먹]] [철수가 [사과를 먹]]고 [[영희가 딸기를] 먹] [철수가 [사과를 먹]]고 [영희가 [딸기를 먹]] (C1,R3)+(C5,R3)						
⋮	⋮	⋮	⋮	⋮	⋮	⋮	
R3	s 먹다'사과'철수' [[철수가 사과를]먹],[철수가 [사과를 먹]] (C1,R2)+(C3,R1),(C1,R1)+(C2,R2)				s 먹다'딸기'영희' [[영희가 딸기를]먹], [영희가 [딸기를 먹]] (C5,R2)+(C7,R1), (C5,R1)+(C6,R2)		
R2	$s/(s\backslash np_n\backslash np_a)$ λ f.사과'철수' [철수가 사과를] (C1,R1)+(C2,R1)	$s\backslash np_n$ λ y.먹다'사과'y [사과를 먹] (C2,R1)+(C3,R1)			$s/(s\backslash np_n\backslash np_a)$ λ f.딸기'영희' [영희가 딸기를] (C5,R1)+(C6,R1)	$s\backslash np_n$ λ y.먹다'딸기'y [딸기를 먹] (C6,R1)+(C7,R1)	
R1	철수가	사과를	먹	고	영희가	딸기를	먹
	C1	C2	C3	C4	C5	C6	C7

표 7: 구문분석기의 파싱테이블

4.3 실험 결과

트리 태깅된 말뭉치에서 추출한 명사의 수는 모두 15571개 였다. 이들을 모두 명사 유사도 사전에 저장하기에는 무리가 따르므로 말뭉치에서 9번 이상 발생한 명사 2141개에 대해서 명사 유사도 사전을 구축했고 나머지 명사들은 필요할 때 마다 동적으로 계산하는 방식을 취하였다. 실험은 명사 유사도 정보를 추출할 때 사용된 말뭉치에서 구조적 모호성을 가지는 명사구 접속이 일어난 문장(A) 87개와 유사도 정보를 추출할 때 사용되지 않은 말뭉치에서 구조적 모호성을 가지는 명사구 접속이 일어난 문장(B) 67개를 무작위로 선정하여 실시하였다.[9] 테스트 문장에서 명사구 접속항에 사용된 명사는 모두 466개이고 분포는 테스트 문장 A에서 294개, 테스트 문장 B에서 186개, 중복된 명사는 4개 이다. 테스트 문장 A에서 쓰인 단어들 중에서 12개 (4%)가 명사 유사도 사전에 나타나지 않았으며 테스트 문장 B에서 쓰인 단어들 중에서 25개 (13.44%)가 사전에 나타나지 않았다.

	적용도	정확도
A	71/78 (91.02%)	60/71 (84.5%)
B	56/67 (83.5%)	35/56 (62.5%)

표 9: 적용도와 정확도

각 실험 결과에 대한 적용도와 정확도는 표 9와 같다. 동일한 방법으로 학습 말뭉치에서 추출된 문장으로 실험된 [3]의 적용도 95.9%와 정확도 85.8%에 비해서 적용도와 정확도 면에서 떨어지는데 이는 [3]이 전산학이라는 한정된 도메인에 관한 대량의 말뭉치 (약 100만 어절)를 사용해서 명사간의 유사도를 추출한 반면 본 논문에서는 특정 도메인이 아닌 사설, 경제학, 종교, 공상과학 등 여러 도메인에 관한 비교적 소량 (약 35만 어절)의 말뭉치에서 명사간의 유사도 정보를 추출했기 때문인 것으로 추측된다. (24)는 분석에 성공한 예이고 (25)는 '술'과 '음식'이 '술'과 '재미'에 비해 비슷한 명사쌍 임에도 불구하고 분석에 실패한 예이다.

(24) 역사와 문화(0.181818)가 오랜 민족(0.101408)은 이래서 좋은 것이다.

(25) 술과 음식(0.060606)을 먹는 재미(0.20000)로 서로 다투어 모여들었습니다.

실패의 주요 원인은 공기 유사도 정보를 뽑을 때 사용된 말뭉치에서 해당 단어와 함께 쓰인 동사가 다양하지 못하거나 해당 단어 자체가 쓰이지 않는 경우도 있기 때문이다. 정확도를 높이기 위해 일반적으로 학습 말뭉치의 크기를 늘리거나 역치 (threshold)를 사용한다. 말뭉치의 크기를 늘리기 위해서는 구문분석된 말뭉치가 필요하게 되는데, 이를 수동으로 구축할 경우 시간과 비용에 대한 부담이 크며 자동으로 구축할 경우에는 구문분석의 정확도가 문제가 된다. 따라서 단순히 구문분석된 말뭉치만을 사용하는 방법으로는 정확도를 향상시키는데 한계가 있을 것으로 보인다. 역치 (threshold)를 사용할 경우에는

[9]테스트 문장 B의 도메인은 '겨울 이야기'라는 소설과 '교과서'이다.

정확도는 높아지지만 적용도가 떨어질 것으로 예상된다.

5 결 론

결합범주문법을 사용하여 언어현상을 처리할 때 나타나는 spurious ambiguity는 병렬문, 원거리 종속관계, quantifier floating과 같은 언어현상을 설명하기 위해서 필요한 현상임에도 불구하고 구문분석과정에서 복잡도를 증가시키는 문제점이 있다. 결합범주문법의 병렬문 축약규칙은 접속사 좌우의 통사범주가 동일하면 적용되기 때문에 구조적으로 모호한 명사구들이 접속된 경우를 처리할 때에는 복잡도가 증가하면서 오분석의 가능성도 증가하게 된다. Spurious ambiguity로 인해 복잡도가 높아지는 문제점은 구문분석과정에서 나타나는 의미부분을 사용하여 해결하였다. 이는 결합범주문법이 통사처리와 함께 의미처리를 동시에 행할 수 있다는 특징을 이용한 것이다. 구조적으로 모호한 명사구들이 접속된 병렬문을 처리할 때 나타나는 문제점을 해결하는 방안으로 접속항의 중심명사를 찾아 중심명사들 간의 공기유사도를 비교하는 방법을 제시하였다. 공기유사도는 트리태깅된 말뭉치로부터 추출되었고 추출된 정보는 공기유사도 정보사전의 형태로 저장되어 구문분석과정에서 접속의 대상이 되는 중심명사들간의 유사도를 비교할 때 사용된다. 공기 유사도 사전을 이용해서 실험을 한 결과 테스트 말뭉치에서 추출한 문장에 대해서는 적용도와 정확도가 학습 말뭉치에서 추출한 말뭉치에 비해서 떨어졌다. 이는 학습 말뭉치에서 사용된 용례의 부족에 기인하고 있으며 단순히 학습 말뭉치의 크기를 늘려서 정확도를 향상 시키는 데에는 한계가 있을 것으로 보인다. 용례의 부족에서 오는 문제점을 해결하기 위해 말뭉치에서 추출한 정보외에도 시소러스등과 같은 의미체계에서 추출한 정보들 이용하는 방안을 모색해야 할 것으로 생각된다.

References

[1] 조형준, 박종철. 한국어 병렬문의 통사, 의미, 문맥분석을 위한 결합범주문법. Submitted, 1999.

[2] 이관규. 국어 대등구성 연구, pages 53–136. 서광학술자료사, 1992.

[3] 양재형. 공기 유사성을 이용한 한국어 명사구 접속의 구조적 모호성 해결. *정보과학회논문지(B) 23(3)*, pages 311–321, 1995.

[4] 서정수. 국어문법, pages 1129–1178. 한양대학교 출판원, 1996.

[5] 김재훈. 오류-보정 기법을 이용한 어휘 모호성 해소. PhD thesis, 한국과학기술원, 1996.

[6] 박준식. 품사패턴을 이용한 한국어 병렬구문의 해석, 1998.

[7] 이공주. 언어 특성에 기반한 한국어의 확률적 구문 분석. PhD thesis, 한국과학기술원, 1998.

[8] Agarwal, R. and Boggess, L. A simple but useful approach to conjunct identification. In *Proceedings of 30th ACL*, pages 15–21, 1992.

[9] Berryl Hoffman. *The computational analysis of the syntax and interpretation of free word order in Turkish.* PhD thesis, University of Pennsylvania, 1995.

[10] Kurohashi, S and Nagao, M. A syntactic analysis method of long Japanese sentences based on detection of conjunctive structures. *Computational Linguistics 20(4)*, pages 507–534, 1994.

[11] Okumura, A and Muraki, K. Symmetric pattern matching analysis for English coordinate structures. In *Proceedings of 4th Conference Applied NLP*, pages 41–46, 1994.

[12] Jong C. Park. A unification-based semantic interpretation for coordinate constructs. In *Proceedings of 30th ACL*, pages 209–215, 1992.

[13] Jong C. Park. *A lexical theory of quantification in ambiguous query interpretation.* PhD thesis, University of Pennsylvania, 1996.

[14] Scott Prevost. *A semantics of contrast and information structure for specifying intonation in spoken language generation.* PhD thesis, University of Pennsylvania, 1995.

[15] Mark Steedman. Gapping as constituent coordination. *Linguistics and Philosophy 13*, pages 207–263, 1990.

[16] Mark Steedman. Surface structure and interpretation. *Linguistic Inquiry Monograph 30*, 1997.

[17] Mark Steedman. The syntactic process. Manuscript, 1999.