대등접속구문과 미지격 명사구의 문법기능 결정

이용훈○ 김미영 이종혁
포항공과대학교 전자컴퓨터공학부 컴퓨터공학과, 첨단정보기술 연구센터
{yhlee95○, colorful, jhlee}@postech.ac.kr

Grammatical Role Determination of Unknown Cases in Korean Coordinate Structures

YongHun Lee○ Mi-Young Kim Jong-Hyeok Lee
Department of Computer Science and Engineering,
Division of Electrical and Computer Engineering,
Pohang University of Science and Technology,
and Advanced Information Technology Research Center(AITrc)

요 약

한국어의 정확한 구문분석을 위해서는, 격조사가 존재하지 않고, 보조사와 함께 쓰이거나 명사만으로 구성된 미지격 명사구들의 정확한 문법기능[1]을 파악하는 것이 중요하다. 또한 긴 문장의 효과적인 구문분석을 위해 대등접속구문을 파악하는 것 또한 중요한 과제이다. 본 논문에서는 위의 두 과제를 동시에 해결하는 방법을 제안하고자 한다. 즉, 한국어의 긴 문장의 대등접속구문을 파악하는 과정에서 미지격 명사구의 문법기능을 결정하고 이 문법기능정보를 이용하여 동시에 대등접속구문의 구간도 결정할 수 있는 방법을 제안한다.

1. 서 론

문장이 길어짐에 따라, 구문분석의 애매성은 급격히 증가하게 되므로, 긴 문장의 정확한 구문분석을 위해서는 대등접속구문의 파악이 중요하다. 한국어에서는 조사와 어미가 매우 발달되어 있어서, 대등접속의 인식이 비교적 쉽다. 즉 명사구 대등접속의 경우 "와/과, 하고, 나" 등의 접속조사를 보고 인식할 수 있고, 동사구 대등접속의 경우 "~고, ~며, ~든지" 등의 대등연결어미를 보고 인식할 수 있다. 하지만 이는 선접속부(pre-conjunction)의 마지막 머리어(head word)와 후접속부(post-conjunction)의 첫 시작 단어만을 판단할 수 있을 뿐 양쪽 끝의 범위는 알 수 없다.[1,2,3]

하지만 대등접속구문의 대등이라는 말에서 알 수 있듯이 선접속부와 후접속부가 대등적으로 연결되는 만큼 그 구조에 있어서도 병렬성이 보여진다. 따라서 이러한 특징을 사용하여 대등접속구문의 구간을 결정할 수 있다.

대등접속구문을 인식하기 위해서 사용할 수 있는 정보는 구조적인 유사성, 어휘적인 유사성, 의미적인 유사성 등이 있다. 하지만 구조적인 유사성을 파악하기 위해서는 선접속부와 후접속부 모두가 완전한 구문분석(full parsing)이 이루어져야 한다. 하지만 보조사가 있을 경우 완전한 구문분석을 할 수가 없다. 보조사는 격을 나타내는 격조사와는 달리 특별한 의미만 더해지는 역할을 한다. 따라서 보조사와 결합한 명사구는 문장에서 주어나 목적어, 부사어 등 다양한 문법기능을 할 수 있다. 이러한 문법기능의 중의성을 해소하기 위해서 용언의 필수격의 수와 종류를 기술한 하위범주화정보나 이에 의미적 제약까지 추가한 선택제약정보가 사용된다. 하지만 "철수는 영희도 매우 좋아한다."의 경우처럼 필수격의 선택제약정보가 서로 비슷하여 여러 문법기능으로 작용할 가능성이 있을 경우에는 선택제약정보만으로는 정확한 문법기능 결정이 어렵다.(여기서는 "철수는", "영희도" 둘 다 주어와 목적어가 될 수 있다.)

본 논문에서는 대등접속구문의 병렬성이라는 특징을 이용하여 대등접속구문을 인식하는 과정 중에 중의성이 있는 미지격 명사구의 문법기능과 대등접속구문의 구간을 동시에 결정할 수 있는 알고리즘을 제안하고 실험을 통하여 이 알고리즘이 유용함을 보인다.

2. 기존 연구

한국어의 대등접속구문에 관한 연구는 장재철[1]과 윤준태[2]를 들 수 있다. 윤준태[2]에서는 HPSG기반 구문분석방법을 사용하여 대등접속의 형태와 그 구간을 결정하는 방법을 제시하였다. 또한 장재철[1]은 구간분할 기반 구문분석의 중간단계에서 대등접속문을 이루는 선후접속부의 유사도를 측정하여 가장 높은 점수를 가진 구조를 대등접속구조로 결정하였다. 한국어와 비슷한 언어현상을 보이는 일본어의 대등접속구문의 결정에 관한 연구[3]에서는 태깅된 단어열의 유사성을 가지고 다이나믹 프로그래밍 기법을 사용하여 가장 높은 점수를 가지는 대등접속구문의 구간을 결정한다. 기존 연구들[1,2,3]을 종합해 볼 때, 대등접속구문에서 유용하게 사용되는 유사성 비교 판단기준은 크게 구조적인 유사성, 어휘적인 유사성, 품사정보의 유사성, 대응되는 단어간의 의미적인 유사성이 될 수 있을 것이다.

미지격 명사구의 문법기능 결정에 관한 연구도 여러 가지가 있다. 한국어 구문분석에 관한 여러 연구에서는 수작업으로 하위범주화정보를 구축하였고, 이러한 정보에 간단한 의미적 제약을 수작업으로 추가하여 미지격 명사구의 문법기능을 결정하는 데에 활용하였다. 하지만 수작업에 의한 작업의 어려움을 극복하기 위해 한국어나 일본어의 여러 연구[4,5,6,7]에서는 대량의 말뭉치로부터 통계적인 방법을 사용하여 주어와 목적어

1) 한국어와 같은 교착어에서는 격(case) 대신에 주어나 목적어와 같은 문법기능(grammatical function)을 용언의 항가로 본다. 본 논문에서는 격과 문법기능을 비슷한 의미로 사용할 것이다.

등에 대한 공기명사와 그 빈도정보를 추출하였다. 또한 통계적인 방법의 데이터부족현상(data sparseness)을 극복하기 위해 [4,6,7]에서는 이들 명사들을 일반화한 개념을 선택제약정보로 구축하였다. 하지만 이들 연구들은 모두 선택제약정보를 구축하는 방법에 초점이 맞추어져 있으며, 생략과 도치가 자주 일어나는 한국어의 긴 문장을 분석할 때, 이 정보를 어떻게 사용하는지에 대한 자세한 언급은 없다.

3. 제안사항

3.1 구문분석 과정

본 논문에서 사용하는 구문분석기는 의존문법을 기반으로 하고 있다. 구문분석의 전체적인 흐름(그림1)은 먼저 형태소분석(태깅)이 끝난 상태의 단어열에 대해 명사구와 동사구의 구묶음(Chunking)을 수행한다. 다음으로 본 논문에서 제안하는 미지격 명사구의 문법기능 결정과 대등접속구문의 결정을 시도한다. 이 과정이 끝나면 제일 오른쪽의 용언으로부터 앞쪽으로 문법기능이 결정된 명사구와 용언과의 의존관계를 설정한다. 그리고 마지막으로 부가어와 용언 간의 의존관계를 설정함으로써 구문분석을 마친다.

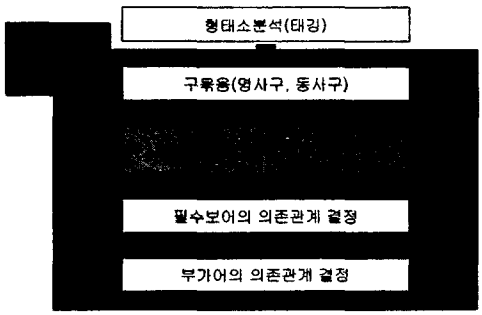

그림 1

3.2 대등접속구문의 결정

대등접속구문 결정의 전체적인 틀은 장재철[1]의 방법을 따르기로 한다. 장재철[1]에서는 대등접속 구조에 대한 점수를 다음 식(식1)으로 계산하고 있다. 각각의 후보에 대하여 점수를 계산한 다음 가장 높은 값을 가지는 것을 대등접속구조로 선택하게 된다.

$$score(T) = \sum I(arc) + structural\_sim(pre\_conj, post\_conj) + sim\_pos(pre\_head, post\_head) + sim\_word(pre\_head, post\_head)$$

I(arc) : 의존관계가 있는 명사구와 용언간의 상호정보값
structural_sim : 선접속부와 후접속부의 구조적 유사성
sim_pos : 선접속부 머리어의 품사와 후접속부 머리어의 품사 유사성
sim_word : 선접속부의 머리어와 후접속부 머리어의 단어유사성

식 1

위의 식에서 동사와 명사구 간의 상호정보(mutual information) 값은 올바른 문법기능을 가진 후보를 뽑아내기 위해 사용한다. 하지만 현재 본 연구실이 보유하고 있는 구문분석기는 이 값 대신에 이미 구축되어 있는 선택제약정보를 가지고 올바른 문법기능을 결정할 것이다. 하지만 서론에서도 말했듯이 이러한 선택제약정보만으로 올바른 문법기능 결정이 어려운 경우가 있는데 이 때에는 다음 절(3.2)에서 설명하는 방법으로 문법기능을 결정하게 된다. 미지격 명사구의 문법기능은 선택제약정보와 미지격 명사구 간의 개념유사도값을 사용하여 결정하고 나머지 유사도값은 장재철[1]에 제시한 그대로 사용하였다. 본 논문에서는 추가적으로 선접속부와 후접속부의 단어의 길이의 차이를 벌점으로 줌으로써 좀 더 정확한 범위를 결정하는데 사용하였다.(위의 점수에서 단어길이의 차만큼 빼준다.)

본 논문에서 사용하는 선택제약정보에는 필수격의 종류와 그 필수격으로 가능한 의미표지가 같이 저장되어 있다. 의미표지는 까도카와 시소러스의 형태를 따르며 그 구조는 3개의 계층으로 된 1110개의 의미표지에 각각 번호가 부여되어 있다. 각 개념간의 유사도는 다음 식(식2)으로 계산한다.

$$Csim(C_i, P_j) = \frac{2 \times level(MSCA(C_i, P_j))}{level(C_i) + level(P_j)} \times weight$$

C_i : 까도카와 시소러스 상의 의미코드
P_j : 선택제약정보의 의미패턴
$MSCA(C_i, P_j)$: 까도카와 시소러스 계층구조에서 두 개념간의 가장 가까운 상속노드
weight : 개념 C_i 가 개념 P_j 의 하위개념이면 1, 아니면 0.5
level(x) : x 개념의 까도카와 시소러스 상의 높이

식 2

3.2 대등접속구문과 미지격 명사구의 격할당

장재철[1]에서 제시한 식1에서 선접속부와 후접속부의 구조적 유사성은 각 접속부가 완전한 구문분석(full parsing)이 이루어진 상태에서 수행된다. 하지만 본 연구실의 구문분석기에서는 미지격 명사구의 격이 할당되지 않은 상태에서 대등접속문 인식을 시도한다. 따라서 대등접속구의 결정과 미지격 명사구의 문법기능 결정이 동시에 이루어질 수 있는 알고리즘을 제안한다.

먼저 알고리즘을 소개하기에 앞서 문법기능 결정에 있어 몇 가지 휴리스틱을 가정한다.

휴리스틱 1.

만약 미지격 명사구가 하나이고 비어있는 필수격의 수가 하나 이상일 경우, 또한 가능한 격의 수가 하나 이상일 경우에는 주어 < 목적어 < 필수부사어 순으로 우선순위를 둔다. 이는 주어, 목적어, 필수부사어 순으로 생략이 더 많이 일어난다고 생각되기 때문이다.

휴리스틱 2.

만약 미지격 명사구가 두 개 이상이고 각각의 명사구가 여러개의 격으로 쓰일 수 있을 경우 각각의 유사도의 합이 가장 큰 것을 선택한다. 예를 들면 "영수는 사과도 좋아한다." 의 경우 '영수는'은 주어와 목적어로 가능하지만 '사과도'는 목적어로만 가능하다. 즉 '사과도'는 유사도가 0이다. 따라서 (영수는:주어, 사과도:목적어)의 조합이 유사도의 합이 가장 크다. 반대로 두 명사구가 똑같이 주어와 목적어로 가능한 경우 '영수는 영희만 좋아한다'인 경우 (영수는:주어, 영희만:목적어), (영수는:목적어, 영희만:주어)의 유사도 합이 같다. 이럴 경우는 1의 원칙에 따라 주어, 목적어의 순서를 가진 첫번째 후보를 선택한다.

다음은 대등접속구의 결정과 미지격 명사구의 문법기능 결정을 동시에 수행할 수 있는 알고리즘이다.

1. 선접속부의 시작단어의 모든 후보를 선정한다.
   선접속부 머리어의 선택제약정보와 미지격 명사구 의미와의 유사도가 0보다 큰 격을 모두 선택한다.
   위의 휴리스틱을 적용하여 가장 큰 유사도의 합을 가지는 것을 선택한다.
2. 후접속부의 머리어의 모든 후보를 선정한다.(용언)
   머리어 후보의 앞에 나타나는 미지격 명사구와 유사도가 0보다 큰 격을 모두 선택한다.
   휴리스틱을 적용하여 가장 큰 유사도의 합을 가지는 것을 선택한다.(여러개의 조합 가능)
3. 1과 2에서 선택된 후보 각각에 대해서 (3.2)에서 제안한 식에 따라 점수를 계산한다.
4. 가장 큰 점수를 얻은 것을 대등접속구조로 선택한다.
   이때 그 내부에서 미지격 명사구에 할당되었던 문법기능을 미지격 명사구의 실제 문법기능으로 할당한다.

본 논문에서 제안하는 대등접속구문의 인식과정에서 미지격 명사구의 문법기능 결정이 어떻게 이루어지는지 실제로 미지격 명사구가 존재하는 대등접속문장을 분석해 보자.

그 섬에 사는 사람들은 유일신만 믿고 다른 신들은 믿지 않는다.

(a)그 섬에 사는 사람들은 유일신만 [믿고] | [다른] 신들은 믿지 않는다.
X타동사 X자동사
(b)그 섬에 사는 사람들은 유일신만 [믿고] | [다른 신들은 믿지 않는다.]
X타동사 SIO 타동사
(c)그 섬에 사는 사람들은 [유일신만 믿고] | [다른] 신들은 믿지 않는다.
SIO 타동사 X자동사
(d)그 섬에 사는 사람들은 [유일신만 믿고] | [다른 신들은 믿지 않는다.]
SIO 타동사 SIO 타동사
(e)그 섬에 사는 [사람들은 유일신만 믿고] | [다른] 신들은 믿지 않는다.
(S,O)|(O,S) 타동사 X자동사
(f)그 섬에 사는 [사람들은 유일신만 믿고] | [다른 신들은 믿지 않는다.]
(S,O)|(O,S) 타동사 SIO 타동사
(g)[그 섬에 사는 사람들은 유일신만 믿고] | [다른] 신들은 믿지 않는다.
(S,O)|(O,S) 타동사 X자동사
(h)[그 섬에 사는 사람들은 유일신만 믿고] | [다른 신들은 믿지 않는다.]
(S,O)|(O,S) 타동사 SIO 타동사

위의 문장은 선접속부 4가지, 후접속부 2가지 후보가 존재하므로 총 4×2=8가지 구조로 분석될 수 있다. 또한 선접속부의 '사람들은', '유일신만'과 후접속부 '신들은'은 모두 용언의 주격과 목적격의 선택제약정보를 만족하여 주어나 목적어로 사용될 수 있다.

후접속부 후보 [다른 신들은 믿지 않았다]에서 '신들은' 같은 경우 휴리스틱 1에 의해 주어와 목적어 중에 목적어로 선택된다. 마찬가지로 선접속부의 [유일신만 믿고]에서도 '유일신만'은 목적어로 선택된다. 하지만 [사람들은 유일신만 믿고]에서는 (주어, 목적어)와 (목적어, 주어) 모두가 가능하다. 이 경우 휴리스틱 2에 의해서 (주어, 목적어)가 선택된다.

위의 8가지 분석후보의 점수를 구하면 다음과 같다.

후보	a	b	c	d	e	f	g	h
점수	8	7	13	14	10	5	5	9

따라서 가장 점수가 큰 (d)가 대등접속구조로 선택되고, 미지격 명사구의 문법기능은 이미 점수를 계산할 때 가정한 (유일신만:목적어, 신들은:목적어)가 할당되게 된다. 이는 실제로 올바른 분석이며 이 알고리즘이 매우 유용하게 사용되었음을 나타낸다.

4. 실험 및 평가

본 논문에서 제안하는 알고리즘의 타당성을 입증하기 위해, Matec(형태소분석기평가대회) 99년도 문장 중, 대등문이 포함된 100개의 긴 문장 (문장당 평균 17.2어절)을 추출하여 실제 실험을 해 보았으며 결과는 다음과 같다.

	대등구문인식	보조사의 문법기능결정
대등구문 수 &보조사 수	119	90
올바른 분석수	97	71
정확율	81.5%	78.8%

대등구문의 인식은 그리 높진 않지만 괜찮은 결과가 나타났다. 하지만 보조사의 문법기능 결정 정확도는 조금 떨어지는 양상을 보였다. 이는 대등구문인식에 있어서 비교적 간단한 방법으로 유사도를 계산하였고 실제 대등구조에서 나타나는 구조적 유사도 계산에서 대응되는 문장성분간의 좀 더 정교한 유사도 측정이 이루어지지 않았기 때문이다. 또한 보조사의 문법기능 결정에 있어서도 용언의 선택제약정보에 의존하는 이상, 정교하지 않은 선택제약정보가 잘못된 문법기능 할당을 일으켰으며, 선택제약정보가 채워지 있지 않은 경우 올바른 문법기능 할당이 이루어지지 않았다.

5. 결론 및 향후 연구

본 논문에서는 긴 문장의 효과적인 구문분석을 위해 꼭 필요한 대등접속구문의 결정과, 문법적인 기능의 모호성을 가지는 미지격 명사구의 문법기능 결정의 두가지 중요문제를 동시에 해결하는 방법을 제안하였다. 앞으로의 연구에서는 대등접속구문을 인식하기 위한 좀 더 정교한 방법의 연구가 필요하고, 또한 대등접속구문이 아닌 내포문과 종속접속문에서의 구문분석에 대한 연구도 할 예정이다.

6. 감사의 글

본 연구는 첨단정보기술 연구센터를 통하여 과학재단의 지원을 받았습니다.

7. 참고문헌

[1]장재철, 박의규, 나동렬, "구간분할 기반 한국어 대등접속구문분석 기법", 한글 및 한국어 정보처리 학술대회, 2002
[2]윤준태, 송만석, "한국어의 대등접속구문 분석", 정보과학회 논문지(B) 제24권 제3호, 1997.3
[3]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, 1994
[4]이휘봉, 강인수, 이종혁, "개념패턴과 통계정보를 이용한 한국어 미지격의 구문관계 결정 방법", 한글 및 한국어 정보처리 학술대회, 1998
[5]양재형, 김영택, "통계 정보를 활용한 한국어 미지격 명사구의 문법기능 결정", 한국정보과학회논문지 제24권 제5호, 1994.5
[6]D. Kawahara, S. Kurohashi, "Fertilization of Case Frame Dictionary for Robus Japanese Case Anaysis", Computational Linguistics, 2002.8
[7]D. Kawahara, N. Kaji and S. Kurohashi, "Japanese Case Structure Analysis by Unsupervised Construction of a Case Frame Dictionary", Computational Linguistics, 2000.8