Chart와 단일화를 이용한 한국어 분석 기법

권 혁철, 채 영숙
(부산대학교 자연대 전자 계산학과)
윤 애선
(부산대학교 인문대 불어 불문학과)

Unification based Chart Parsing for Korean

Kwon, Hyuk-Chul, Chae, Young-Soog
(Pusan National University, Dep. of Computer Science)
Yun, Ae-Sun
(Pusan National University, Dep. of French Lang. & Literature)

요 약

이 논문은 상대적으로 어순이 자유로운 언어인 한국어의 특성을 반영하면서, 모든 가능한 문장 구조를 분석할 수 있는 한국어 분석 방법을 제시한다. 특히 구절 구조에 의한 통사 표현 기능을 하위 범주화와 단일화에 의해 보완하는 기법을 이용하면서, bottom-up과 left-right에 의해 분석이 가능한 단일 과정(one-path) 분석 기법을 이용하는 것이 본 논문의 특징이다. 그리고 하위 범주화와 서술어의 어미가 가진 양상 정보에 의존하여 한국어의 내포문 처리가 이루어져야함을 보여준다.

I. 서론

최근에 기계로 처리하기에 적합한 한국어 처리 방법의 연구가 이론적인 측면과 실용적인 측면에서 다양하게 연구되고 있다. 이론적인 측면의 연구는 언어학자들을 중심으로 이루어지고 있다. 언어학자 중심의 자연언어 처리 연구는 개별 언어의 특성보다 언어가 가지는 보편성에 연구의 촛점을 두어, 이론적 명확성과 단순성을 강조하는 것이 일반적이다. 그런데 개별 언어를 단순화된 이론에 적용시켜 기계로 분석하는 경우에 처리 속도가 비현실적으로 느리거나, 개별 언어가 가진 특성을 명확히 분석하지 못하는 경우가 발생할 수 있다.

실용적인 측면에서의 연구는 전산학자를 중심으로 이루어지고 있다. 전산학자의 일차적 관심은 실생활에 응용할 수 있는 시스템의 구성에 있다. 따라서 이들에 의해 개발된 자연언어 처리 시스템은 제한된 범위내에서만 처리가 가능하며, 시스템의 확장이 매우 어렵다. 따라서 현재 실용화된 자연언어 처리 시스템은 매우 제한된 범위에서만 작동되며, 보다 폭넓은 영역을 처리할 수 있는 시스템의 구현은 실패를 거듭하고 있다.

한편 언어학자들은 방대한 자료를 바탕으로 언어 현상을 관찰함으써 언어 현상에 대해 명확한 파악을 하고 있으며, 전산학자들은 인공지능 연구의 경험을 바탕으로 지적인 문제의 처리가 지식에 기반하여 이루어지며, 자연언어 처리도 성패가 지식의 처리 방법에 있음을 파악하고 있는 만큼, 이 두 접근이 잘 결합되어 연구된다면 보다 폭넓은 영역의 문제를 처리할 수 있는 자연언어 처리 시스템의 실용화가 가능하다고 보아진다.

80년대에 들어오면서 기계에 의한 한국어 처리 연구가 실용화의 측면에서 계속되고 있으며, 특히 최근에는 국가적 차원에서 이 분야에 대한 연구가 집중되고 있다. 그러나

한국어 자연언어 처리의 기본이 되는 한국어 분석 과정에 대한 기존 연구는 영어와 같이 한국어의 구조와 상이한 언어의 분석에 이용되는 기법을 사용하는 경우가 많다. 그러나 한국어의 특징에 대한 분석을 바탕으로 새로운 접근에 의한 한국어 분석 방법을 제시할 필요가 있다.

이 논문은 한국어의 특성에 적합한 한국어 구조 분석 기법을 제시하고 있다. 현재 각광을 받고 있는 단일화 문법을 염두에 두어 본 연구가 진행되었으나, 이 논문에서 제시되는 분석 기법은 특정 언어 이론에 바탕을 둔 것은 아니므로, 언어 이론과는 무관하다.

II. 한국어의 특징 및 분석 방향

한국어와 영어의 비교 분석을 바탕으로 한국어 분석의 방향에 대해 살펴보자.

한국어의 어순은 영어에 비해 상대적으로 자유롭다. 그러나 완전히 자유로운 것은 아니며, 술어가 제일 뒤에 나오고, 수식어는 수식되는 어구의 바로 앞에 나온다. 영어에서는 문장 성분의 순서가 격의 결정에 중요한 역할을 하지만, 한국어의 명사나 명사구는 한 개 이상의 조사를 가지는 것이 일반적이며, 이 조사가 명사나 명사구의 격을 결정한다. 그러나 조사가 모든 격을 결정하는 것만은 아니며, 문장의 하위 범주화가 격의 결정에 또 다른 역할을 한다.

(1) 철수가 영희를 좋아한다.
(2) 영희를 철수가 좋아한다.
(3) 철수는 영희를 좋아한다.
(4) 영희는 철수가 좋아한다.

위의 예문 1)과 2)는 한국어의 주어와 목적어의 순서가 바뀔 수 있음을 보여준다. 예문 3)과 4)는 조사만으로는 격의 결정이 불가능한 경우를 보여준다. 즉 주제화 조사(topicalizing postposition)의 격은 하위 범주화와 다른 문장 성분의 유무에 따라 결정됨을 알 수 있다.

한국어의 다른 특징으로는 내포문(embedded sentence)의 시작점을 나타내는 표시가 없는 점이다.

(5) 철수가 영희가 영수를 부르라고 말했다.
(6) 철수가 빨리 영수를 부르라고 말했다.

즉 문장 5)에서 내포문의 시작점은 "영희가"가 되며, 그 이유는 한 문장에는 주어가 한개 이상일 수 없기 때문이다. 그러나 문장 6)에서는 내포문의 시작점을 예측하기는 매우 어렵다.

이와 같은 특징에 따라 한국어를 ATN과 같은 top-down 분석 방법으로 구문 분석을 하는 것은 부적합하다[6]. 즉 ATN과 같은 top-down parsing 방법은 예측을 바탕으로 문장 분석이 이루어지지만, 한국어는 어순이 자유롭고, 내포문의 표시가 없으므로 예측에 의한 처리에 많은 무리가 따른다. 그러나 영어는 어순이 일정하며, 내포문의 시작점이 명확히 드러나므로 ATN에 의한 처리가 가능하며, 현재 영어의 분석에는 이론적 측면과 실용적 측면 모두에서 ATN이 효과적으로 사용되고 있다.

또 다른 한국어의 특징으로는 하위 범주화되는 성분이 문맥상 명확하면 생략이 가능하다는 것이다. 영어에서는 이 경우에 대용어(anaphora)를 사용하는 것이 일반적이다. 그런데 이와 같은 문장 성분의 생략은 통사 분석 단계에 많은 어려움을 주며, 분석 과정을 복잡하게 한다. 문장 6)은 부록에서 보는 바와 같이 3가지의 다른 구조의 문장으로 분석이 가능하다.

III. 한국어 분석 방법

이와 같은 한국어의 특성을 효과적으로 처리하기 위하여 본 연구는 다음과 같은 분석 방법을 이용함으로써 한국어 특성을 반영하고자 노력했다. 먼저 한국어 분석을 문맥자유 문법의 구절 구조(phrase structure)와 하위 범주화를 동시에 이용하여 진행했다. 한국어 분석에 구절 구조가 부적합한 것으로 주장되기도 하지만, 근본적으로 구절 구조를 완전히 무시한 자연언어 처리는 불가능하다.

그러나 영어와 다르게 한국어 분석에서는 구절 구조에 의한 문장 구조 표현의 역할이 상대적으로 약화될 수 밖에 없다. 구절 구조에 의한 통사 표현 기능을 다음과 같이 단순화 시켰다. 구현시에는 추가의 처리와 연계 관계와 처리 속도의 향상을 위해 X 통사론을 도입하는 등 다소 변형된 통사 규칙을 이용한다.

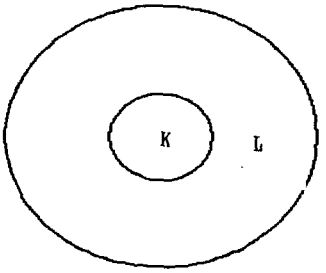
(a) S -> (PP | S | Adv)* Pred | S S

(b) Pred -> V | Adj | N copula

(c) PP -> NP post | S post

(d) NP -> N | S NP | PP NP | Art NP | S

위의 생성 규칙에서 PP는 후치사구(postpositional phrase)를 의미한다. 위의 생성 규칙에 의해 생성 가능한 한국어 문장은 실제 사용되는 한국어 문장을 훨씬 능가한다. 본 논문에서는 구절 구조에 의한 통사 표현을 계속적인 문장 구조 분석을 위한 출발점으로 설정했으며, 따라서 구절 구조는 한국어 문장의 구조와 특성을 보여주면서, 최소한 한국어의 올바른 모든 문장을 포함하는 정도의 생성 능력(power)을 가진 것으로 정의한다고 가정했다. 문맥 자유 문법으로는 문맥 의존적인 자연언어 문장의 처리가 불가능하지만, 올바른 한국어 문장에 비문이 추가된 문장은 문맥 자유 문법으로 처리가 가능하다. 즉 문맥 자유 문법으로 한국어를 능가하는 생성 능력을 가진 문법의 표현은 가능하다.



즉 L을 위의 문법에 의해 생성 혹은 분석 가능한 한국어 문장이라면, K는 우리가 올바른 한국어 문장이라고 인정하는 문장이다. L은 모든 가능한 문자열에서 많은 비문이 걸러진 형태로서, 언제나 K⊆L의 식을 만족해야 한다. 이 결과 L에 하위 범주화, 조사 및 어미의 기능 분석과 자질 분석을 바탕으로 다시 추가의 분석을 행하여 올바른 문장의 집합인 K를 분리한다. 이 결과는 본 연구가 이중 처리를 가정한 것으로 오해될 여지를 주지만, 다음과 같은 처리 방법을 이용함으로써 단일 처리로 분석을 행한다.

구절 구조 분석은 역우단 유도의 방법을 이용하며, 분석 가능한 모든 문장 구조를 생성한다[3,4]. shift/reduce conflict나 reduce/reduce conflict가 일어나는 경우에는 모든 가능한 path를 모두 찾으며, 이에 따라 다수의 parse tree가 생성될 수 있다. 그리고 구절 구조 규칙이 적용되어 축소(reduce)가 일어날 때마다 하위 범주화 규칙, 자질 분석, 조사 및 어미의 기능의 분석을 바탕으로 단일화(*unification*)에 의해 추가의 처리를 수행한다. 이 과정에서 구절 구조에 의해 받아들여진 비문이 제거되고, 문맥 의존적인 요소의 처리를 위해 분석된 구조가 재조정된다. 예를 들면 문장 7)과 같이 부사가 부사를 수식할 경우에 내부 구조를 분석하여 부사가 부사를 수식하는 구조로 문장 구조를 변환시킨다.

(7) 바닷가에 아주 자갈이 많이 있었다.

즉 본 접근 방식에서는 문맥 의존적인 요소는 경험적 규칙(heuristics)으로 해결한다.

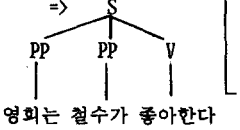
문장 4)를 예로써 분석 과정을 설명하면 다음과 같다.

(구절 구조 분석) (하위 범주화 분석)

(가) 영희는 => PP [TOPIC "영희"]

(나) 철수가 => PP [Subj "철수"]

(다) 좋아한다 => S [Sentence "좋아한다(Subj,obj)" [topic "영희"] [Subj "철수"]]



먼저 구절 구조 규칙에 의해 "영희는"이 PP가 되고, 조사의 기능 분석을 바탕으로 [TOPIC "영희"]를 유도한다. 같은 방법으로 "철수가"를 처리한다. 마지막으로 구절 구조 규칙을 이용하여 S를 생성하고, 앞의 분석과 하위

범주를 바탕으로 목적어가 주제어인 "영희"가 됨을 분석한다. 이를 위해 본 연구에서는 다음과 같은 가정을 했다.

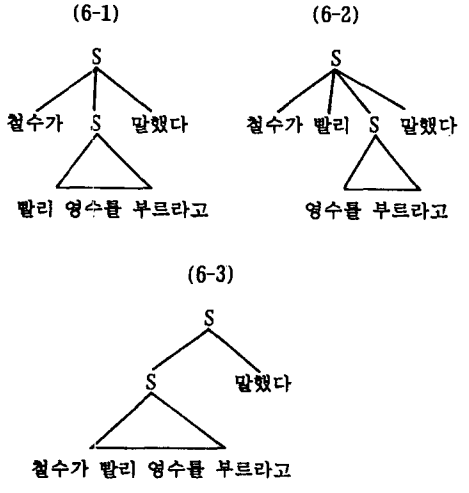
가정1) 하위 범주화되는 성분의 수와 주제화된 명사 수의 합은 서술어가 하위 범주화하는 성분의 수를 초과할 수 없다.

한편 중의적인 문장은 통사 분석 과정을 chart parsing을 변형한 방법을 이용함으로써 bottom-up, left-right의 순서에 의해, 단일 과정으로 한국어 분석을 수행한다[6].

본 논문에서 구절 구조를 상위에 두고 한국어 문장 분석을 진행한 이유는 다음과 같다.

먼저, 구절 구조에 의해 표현된 통사 구조가 다른 처리를 위한 출발점이 될 수 있기 때문이다. 즉 축소(reduce)하는 과정에서 하위 범주화, 조사 및 어미의 기능 분석, 자질 분석과 주제어 처리 및 생략된 성분의 처리를 수행함으로써 프로그래밍 작업과 프로그램의 개선을 쉽게 할 수 있다.

둘째, 어순이 자유로운 한국어의 특징을 반영하여, 필요에 따라 다른 문장 구조로 바꾸어 줄 수 있다. 문장 6)을 예로 설명하면 다음과 같다.



문장 6)은 3가지의 분석이 가능하다. 따라서 중의성이 많으므로 문장 6)은 바람직한 문장이라고 할 수 없다. 만약 문장 6)의 뜻이 6-1)이라면, 특수한 경우를 제외하면 한국어에서는 (PP|S|Adv)* Pred에서 (PP|S|Adv)*의 순서가 자유로우므로 "영수를 빨리 부르라고 철수가 말했다"로 문장 구조를 변환할 수 있으며, 이 결과가 훨씬 명확한 문장이 된다.

세번째로, 한 문장으로부터 분석 가능한 모든 통사 구조를 추출하는 과정을 쉽게 해 준다. 즉 본 논문에서는 chart에 기반한 방법을 이용하여 통사 분석을 진행하는데, 이 과정에 구절 구조에 의한 표현이 도움을 준다.

네번째, 문장이 일부 수정되는 경우에 재분석을 최소화하면서 처리할 수 있는 가능성을 부여할 수 있다.

앞에 제시된 한국어 분석 방법은 어휘 기능 문법(LFG)과 유사한 면이 있으나[5], 구절 구조에 의해 생성이 되는 문장이 통사적으로 비문인 한국어를 포함할 수 있으며, 하위 범주화된 요소가 생략될 수 있는 것으로 상정한 면 등에서 차이가 있다.

IV. 내포문 처리

내포문 처리에는 하위 범주화와 어미의 기능이 중요한 역할을 한다. 본 장에서는 완형 내포문의 처리를 예로써 내포문 처리에 대해 설명하겠다.

(8) 철수가 순희가 아름답다고 생각한다.
(9) 철수가 순희를 아름답다고 생각한다.

위의 문장 8)과 9)는 동일한 의미를 가진 문장이다. 이 문장 8)과 9)를 분석하기 위해서는 "생각하다"라는 동사가 두 가지의 다른 형태의 하위 범주화가 되어야 한다.

(8-1) 생각한다(Subj, S)
(9-1) 생각한다(Subj, Obj, S[-Subj]) {Obj -> Subj}

8-1)을 이용하여 8)을 분석하는 과정에서, 한 문장이 주어를 두 개 하위 범주화할 수 없으므로 "철수가"와 "순희가" 사이에 문장 경계가 있다고 가정할 수 있다. 이와 같은 문장 경계를 잘 설정하는 것은 처리 속도의 향상에 기여할 수 있다. 9-1)은 9)와 같이 "생각하다"가 목적어를 하위 범주화하는 경우에는 내포문의 주어가 생략되며, 주문의 목적어가 내포문의 주어가 된다는 것을 보여준다.

(10) 순희가 아름답다고 생각한다.

문장 10)은 8-1)과 9-1) 둘다로부터 분석이 가능하다. 8-1)에 의해 분석될 경우에는 생략된 요소가 추가되면 다음과 같이 다르게 해석될 수 있다.

(가) "생각한다"(X, "아름답다"("순희"))
(나) "생각한다"("순희", "아름답다"(X))

위의 가)와 나)의 X의 값은 문맥으로부터 채워져야하며, 이 경우에는 가)와 나)만으로는 생략된 성분을 채워 넣을 수 없다. 문맥으로부터 채워져야 하는 생략된 성분의 처리는 화용의 문제를 포함하므로 본 연구에서는 제외했다.

(다) 생각한다("순희",X,"아름답다"[-subj]) {obj->subj}

한편 9-1)에 의해 분석한 결과인 다)는 구조적으로 나)와 유사하다. 그 이유는 다)의 생략된 목적어인 X가 "아름답다"의 주어가 되기 때문이다.

(11) *철수가 순희를 순희가 아름답다고 생각한다.

한편 문장 11)은 비문인 문장이다. (11)이 비문인 이유는 "생각하다"가 목적어를 하위 범주화할 경우에는 내포문의 주어가 생략되어야 하는데 생략되지 않았기 때문이다. 이 문제는 단일화만으로는 처리가 불가능하다.

한국어는 영어와 다르게 어미가 가진 양상 정보(modality)에 따라 생략된 성분이 달라지는 경우도 있다.

(12) 철수는 영수를 가라고 설득했다.
(13) 철수가 가겠다고 설득했다.

위의 문장 12)에서 "가라고"의 주어는 "영수"이지만, 13)의 "가겠다고"의 주어는 "겠"이 주어의 의지를 반영하므로 "철수" 자신이 된다.

(14) 철수가 동생에게 밥을 먹여야겠다고 생각한다.
(15) 철수가 동생을 밥을 먹여야겠다고 생각한다.

문장 14)는 8-1)에 의해 하위 범주화되는 문장이며, 두 가지의 다른 분석이 가능하다. 내포문의 주어가 생략된 경우를 가정하면, 내포문 서술어의 어미에 "겠"이 있으므로 생략된 주어는 "철수"가 된다. 그런데 문장 15)는 약간 어색하지만 옳은 문장으로 보는 경우에, 9-1)에 의해 처리하면 문제가 발생하는 것처럼 보인다. 즉 이 문장의 주어는 "철수"이며, "동생"은 밥을 먹게 되는 대상이 되기 때문이다. 그러나 "먹여야"가 사역이므로, 주어가 간접 목적어가 되고, 주어가 "겠"에 의해 "철수"가 된다고 해석하면 무리가 없다.

V. 구현 및 결론

이 논문에서 제시된 한국어 분석 방법은 C와 LISP에 의해 현재 구현되어 확장되고 있다. bottom-up, left-right에 의해 구현되었으며, 모든 가능한 문장 구조를 모두 찾기 위해 chart를 변형한 방법을 이용하고 있다. 구현 방법과 구현 과정에 대한 구체적인 내용은 정보과학회 89년 가을 학술 논문 발표지에 게재할 예정이다[2].

본 연구에서 제시된 분석 방법은 실시간 처리가 가능하면서도 모든 가능한 문장 형태를 보여주며, 또한 문장이 부분적으로 수정된 경우에도 수정된 부분을 중심으로 최소한의 재분석을 통해 처리할 수 있는 강점이 있다. 그러나 처리 시간과 사용되는 기억 용량이 지나치게

소모되는 약점이 있으나, 한국어 문장에 대한 분석을 바탕으로 이 문제점을 개선하고 있다.

부록에 분석된 결과의 예가 있으며, 현재 분석된 결과를 표현하기 위한 효과적인 표현 방법과 문법을 기술하기 위한 언어의 설계 및 사전 구성 방법에 대해 계속적인 연구를 진행하고 있다.

참 고 문 헌

[1] 조 혁규, 장 명길, 권 혁철, "KPSG에 기반한 한국어 해석기의 구현", 정보과학회 '89봄 학술 발표 논문집, pp.229-232

[2] 조 혁규, 박 용욱, 권 혁철, "자연언어 구문 분석의 비결정성 처리에 관한 연구", 정보 과학회 '89 추계 학술 발표 논문지에 발간 예정

[3] 김 영택, 컴파일러 구성론, 회중당, 1988

[4] M. Tomita, Efficient Parsing for Natural Language, Kluwer Academic Pub., 1986

[5] P. Sells, Lectures on Contemporary Syntactic Theories, CSLI Lecture Notes, No. 3, CSLI, 1985

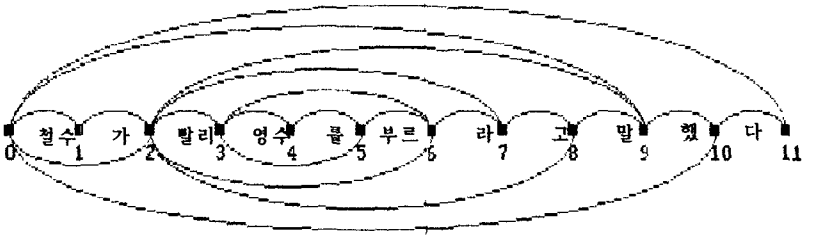
[6] M. Wiren, On Control Strategies and Incrementality in Unification-Based Chart Parsing, Thesis No. 140, Linkoping Univ., 1988

[7] P. T. Sato, "A Common Parsing Scheme for Left-and Right-Branching Languages", Colling, Vol. 14, pp. 20-30

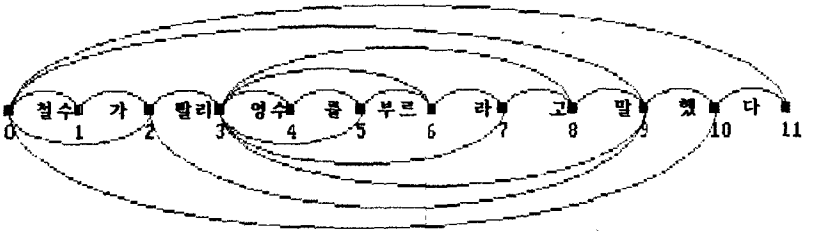
[8] M. King, Parsing Natural Language, Academic Press, 1983

부록

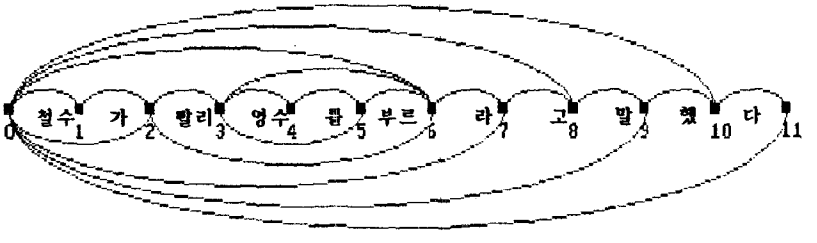
예문 " 철수가 빨리 영수를 부르라고 말했다."의 구문분석 결과



철수가 [빨리 영수를 부르라고] 말했다



철수가 빨리 [영수를 부르라고] 말했다



[철수가 빨리 영수를 부르라고] 말했다