문형 정보를 이용한 한국어 구문 분석

한용기*, 황이규**, 이용석**
* 군산전문대학 전자계산과
** 전북대학교 컴퓨터과학과

Korean Syntax Analysis Using Sentence Pattern Information

Yong-Gi Han*, Yi-Gyu Hwang**, Yong-Seok Lee**
* Dept. of Computer Science, Kunsan Junior College
** Dept. of Computer Science, Chonbuk National University

요약

대부분의 한국어 구문 분석은 용언과 명사구 사이의 하위범주화 정보를 이용하여 용언에 대한 명사구의 문법적 역할을 밝히는 방향으로 구문 분석을 시도하였다. 여기에 이용된 용언의 하위 범주화 정보가 단지 자릿수 서술어나 형용사, 자동사, 타동사 등으로 분류하는 수준이었기 때문에 구문 모호성이 많이 발생하고 틀린 문장이 구문적으로 옳기 때문에 옳은 문장으로 인식되는 경우가 발생하였다. 이러한 문제점을 해결하기 위하여 본 논문에서는 한국어의 용언에 따른 문장 형태(문형)를 세분류하고 문장에 필수적으로 나타나는 명사구(NP[case])와 수의적으로 나타나는 명사구(NP[case])를 분류하여 분석을 시도하였다. 확장된 PATR II로 문법을 기술하여 동적인 파싱을 쉽게 제어할 수 있도록 하였다. 문형 정보는 한국어의 기본 구조를 자연스럽게 표현할 수 있기 때문에 그 자체를 기계번역를 위한 한국어 문법으로 설정하는 것이 타당하다고 생각된다.

1. 서론

한국어에 있어서 구문 분석의 중심은 문장에서 명사구와 용언사이의 표층적 문법 관계를 밝히는 작업이다. 지금까지 대부분의 한국어 구문 분석에 있어서 문장의 문법적 관계가 한정되어 있었다. 즉, 주격이나 목적격을 표층적으로 나타나는 격 관계로 설정하고 나머지 명사구는 단지 보조적 명사구로 문장의 양태만을 나타내는 수의적 성분으로 분류하였다. 이러한 연구들은 대부분 5가지의 명사-동사간 문법 관계(subject(주격), object(목적어), location(-에, -에서 류의 처소격조사), instument(-로, -으로 류의 도구격조사), else(기타))를 설정하여 구문 분석을 시도하였다. 이러한 방법의 문제점은 주격, 목적격 등 5가지 문법적 관계 이외에도 특별한 격을 문장에 반드시 수반하는 용언이 많이 존재한다는 사실이다. 이러한 용언의 경우 나머지 격을 보조적인 의미로 파악하기 때문에 문장의 올바른 의미를 파악하기 힘들거나 모호성 발생의 원인이 되었다. 이러한 문제의 원인은 용언을 단순히 자리수 서술어나 형용사, 자동사, 타동사로 간략하게 분류함으로서 발생한다. 예를 들어 아래와 같은 문장을 살펴보자.

① 영희가 귀엽게 생겼다
② 영희가 생겼다*
③ 철수가 영희를 대표로 뽑았다
④ 철수가 영희를 대표에게 뽑았다*

①과 ②에서 '생겼다'는 자동사이기 때문에 주어만을 필수적 문장 성분으로 간주할 경우, ②도 구문적으로 옳은 문장으로 간주할 수 있다. 또한 ③과 ④에서도 '뽑았다'는 타동사이기 때문에 주

어와 목적어 성분만을 필수요소로 간주할 수 있으며 결국 ④의 문장도 구문적으로 옳은 문장이 된다. 그러나 ②와 ④는 옳은 문장으로 간주할 수 없으며, 이것은 결국 '생겼다'라는 용언은 '어떠어떠하게'라는 의미를 가지는 부사어를 문장의 필수 성분으로 가지게 되며, '뽑았다'라는 용언은 '무엇으로'라는 명사구를 하위범주화 요소로 요구함을 알 수 있다. 이러한 구분은 용언의 의미적 양태에 따른 분류가 아니라 구문적인 분류상에서도 가능하다.

본 논문에서는 앞에서 서술한 문제점을 극복하고 한국어 기계번역을 위한 문법체계에 적합한 방법으로 국어학에서 연구된 문형에 관한 내용을 소개하고 이를 바탕으로 자연스럽게 한국어 문장을 분석하는 방법론에 대해 기술하고자 한다. 이를 위해 용언 자체를 세분화하고 용언이 어떠한 격 체계를 필요로 하는지 분류하여 실제 문장을 분석하는 예를 보인다.

2. 한국어의 특징과 문형

문형이란 문장의 구조적 유형으로, 수많은 개개의 구체적인 문장을 그 구조적 형식의 공통성에 따라 공식화한 틀[강은국]로 정의 될 수 있다. 지금까지 한국어 문법을 문형을 토대로 분류한 방법중 대표적인 것을 살펴보면 다음과 같다.

1) 고영근 문형론

무엇이 무엇이다 [용언화한 체언]
무엇이 어떠하다 [형용사문]
무엇이 어찌한다 [자동사문]
무엇이 무엇을 어찌한다 [타동사문]

2) 영어식 문형 분류

NP[NOM] + IV	: 꽃이 핀다
NP[NOM] + ADJV	: 나뭇잎이 푸르다
NP[NOM] + NP-이다.	: 조국은 하나다
NP[NOM] + NP[COM] + CIV	: 물이 얼음으로 바뀐다
NP[NOM] + NP[ACC] + TV	: 꽃이 열매를 맺는다
NP[NOM] + NP[DAT] + NP[ACC] + DTV	: 철수가 영희에게 책을 준다
NP[NOM] + NP[ACC] + NP[COM] + CTV	: 우리는 영희를 대표로 뽑았다

VP : 동사구 - [영희를 밀었다] , IV : 자동사 - 사람이 [많이 다쳤다]
TV : 타동사 - 철수가 영희를 [밀었다] , DTV : 이중 타동사 - 사람이 [많이 다쳤다]
CIV : 불완전 자동사 - 조국은 하나[다] , CTV : 불완전 타동사 - 영희를 대표로 [뽑았다]

이러한 문형 분류는 대체적으로 4형에서 12형까지 분류하고 있다[강은국94]. 이렇게 문형을 분류하는 중점적인 기준은 근간 성분(주어,서술어,보어,목적어), 수식 성분(부사어, 관형사), 독립성분을 어떻게 파악하느냐의 관점에 따라 다양하다. 이러한 관점의 대부분은 영어식 분류법을 따르고 있는 것이 대부분이다. 또한 용언의 구조적 특성을 무시하고 단지 용언 자체의 구문적 특성을 서구적인 기준에 따라 용언을 분류하고자 하였다.

이러한 방법은 앞에서 언급한 것과 같이 ②와 ④의 예에서 발생하는 문제점을 가지고 있다. 이러한 문제점을 근본적으로 해결하기 위해서는 한국어 문장의 기본 구조를 용언의 특성에 따라 다시 분류해야만 한다는 점을 보여 주고 있다. 이러한 분류는 결국 용언이 어떠한 구문적 특성을 가지는 문장 요소를 필수적으로 동반하는지 광범위한 조사가 필요하다. 본 논문에서는 [강은국]의

41가지 기본 문형을 한국어 문법의 기본 구조로 설정하고 문장을 분석해 보았다.

41가지 기본 문형은 크게 아래와 같은 5가지로 분류할 수 있다.

㉠ 자동사 술어 기본문형
㉡ 타동사 술어 기본문형
㉢ 양면동사 술어 기본문형 (자,타동사 모두의 역할을 하는 동사)
㉣ 형용사 술어 기본문형
㉤ 명사 술어 기본문형

여기에서 문장의 기본 틀을 이루는 데 구분이 되는 표층 문법형태소는 다음과 같다.

이, 을, 에, 에서, 로, 와, 를 위해, 에 의해, 에 대해, 보다, 부사(AD), 라고

위의 11개의 표층격과 부사의 결합으로 41개의 기본적인 문장구조가 형성되는데 그 중 몇가지만 살펴보면 다음과 같다.

```
NP[이] + V
NP[이] + NP[을] + V
NP[이] + {NP[을]} + V          ;; {}은 생략 가능
NP[이] + A
NP[이] + N이다
```

<그림 2.1 > 가장 기본적인 문형

```
NP[이] + AD[게] + V
NP[이] + NP[에] + V
NP[이] + NP[이] + A
NP[이] + NP[보다] + A
NP[이] + NP[에서] + NP[을] + V
NP[이] + NP[와] + NP[에 대해] + NP[을] + V
NP[이] + NP[와] + NP[에 대해] + NP[을] + V
          . . .
```

<그림 2.2> 확장된 기본 문형

이러한 정보를 통해 술어를 중심으로 단어들이 가지는 구조적 특성으로 분류하면 구문적으로 가능하나 의미적으로 불가능한 구조를 제거할 수 있다. 아래 문장의 경우, 비록 "군다" 라는 용언이 자동사임을 알고 있기 때문에 이론적으로 주어와 용언[자동사]만이 필수적 성분으로서 문장을 이룬다고 말할 수는 없다. 왜냐하면 "철수가 군다"라는 문장은 옳지 않은 문장이기 때문이다. 즉 '군다'의 경우 반드시 필수적인 문장 요소로 "귀엽게"나 "성가시게" 같은 상태를 나타내는 부사가 존재해야 한다. 그러므로 "귀엽게"나 "성가시게"는 이러한 용언류에서는 필수적 문장 성분으로 생각할 수 있다.

```
철수가 성가시게 군다.          철수가 군다*
NP[가]  AD[게]  V
```

또한, 용언을 좀더 자세히 분류할 경우 구문적 반복 패턴이 존재한다. 예를 들면, "발원한다", "기원한다", "우러나온다" 와 같은 용언(자동사)의 경우를 살펴보면 다음과 같다.

사상은 실천에서 발원한다.
양서류는 어류에서 기원한다.
이런 감정은 완전히 민족적 자부심에서 우러나온다.

위의 예에서 "실천에서", "어류에서", "민족적 자부심에서" 등이 문장에서 생략된다면 자동사 문장에서 문장의 구조가 비록 올바를 지라도 문장의 의미는 전혀 통하지 않는다. 그러나 이런 용언류(시발 자동사라 칭하자)는 반드시 "어디에서부터" 라는 뜻을 나타내는 명사구가 "에서"류의 격조사를 매개로 반드시 나타나게 된다.

즉, 우리가 일반적으로 알고 있는 필수격보다 넓은 범주로의 필수격에 대한 확장의 필요성이 제기된다. 따라서 각 용언에 대한 필수격을 재정의한다. 또한 필수격의 대표적 격조사를 아래와 같이 정의한다.

이(subj, comp), 을(obj), 에(loc), 에서(sour), 로(tar), 와(with), 를 위해(for), 에 의해(by),
에 대해(to), 보다(than), 게(AD), 라고(call)

여기에서 주목할 점은 '에 대해', '를 위해', '에 의해' 도 하나의 필수격 조사로 포함시킨 곳이다.

우리는 현대화의 실현을 위해 싸운다.
철수는 선생님과 영희의 건강에 대해 견해를 나누었다.
그 운동회는 학생들에 의해 개최되었다.

이들을 하나의 격조사로 볼 경우, 문장의 분석이 좀 더 자연스러움을 알 수 있다. 이 조사와 용언의 결합을 필수격으로 포함시키지 않는 경우의 구문 분석을 위한 문법과 비교하여 보자.

그는 정의를 위해 싸운다.
NP[는] NP[를] Sen[여] Sen[ㄴ 다]
Sen -> Sen[suffix] Sen

그는 정의를 위해 싸운다.
NP[는] NP[를위해] Sen[ㄴ 다]
Sen -> NP[case] Sen[suffix]

<그림 2.3> '에의해', '에대해', '을위해'를 위한 구문 분석 문법

<그림 2.3>에서 보는 바와 같이 '위해 싸운다'를 용언의 연속으로 이해하기 위해서는 이를 위한 복합용언 처리 규칙이 필요하다. 그러나 국어학적 입장이 아닌 계산학적 입장에서 볼 때, 복잡성을 감소시키고 좀더 효율적인 분석을 위해 이러한 조사와 용언을 결합하여 하나의 필수격으로 설정하였다.

3. 조건단일화와 문형정보를 이용한 문장 분석

한국어 문법을 문형정보를 이용하여 표기하기 위해 문법 기술도구인 PATR II를 확장하여 사용하였다. PATR II는 다양한 문법 형식들을 간결하고 일관성 있게 기술할 수 있다[Shie86].

또한 한국어 문장을 파싱하는데 필요한 문법 기술을 보다 유연하게 하기 위하여 조건단일화 방법을 사용하였다[양승원95]. 영어의 경우, 기본적 문형 구조는 어순이 존재하기 때문에 NP의 역할이 위치에 따라 설정된다. 따라서 기본적 구조의 파악은 문맥자유 문법적 특성을 통해 충분히 파악이 가능한 계층구조를 형성한다. 그러나 한국어의 경우, 어순이 비교적 자유롭기 때문에 평면구조(flat)를 이루고 있어 평면 구조내의 명사구들이 용언에 대해 문법적 관계의 표현을 위해 문법형태소인 조사나 어미가 존재하므로 기본 문형의 틀로 CFG를 간주하고 각 NP[case]의 문법 형태소 정보의 파악을 위해 CSG적인 방법을 사용해야 한다. 조건단일화 방법은 이러한 문제를 효율적으로 해결할 수 있는 방법을 제공하고 있다.

조건 단일화 방법이란 문맥자유문법을 좀더 제약하여 자연어 문법을 표현할 수 있도록 문법 기술상에 유연성을 제공할 수 있도록 고안되었다. 이 방법은 단일화를 통해 각 어휘정보나 문법 정보들이 결합되는 경로를 동적인 제어를 통해 효율적으로 문법을 제어할 수 있도록 하였다.

아래는 문형 정보를 이용한 구문 분석틀로 어떠한 문법 형식이 필요한지를 보여주고 있다.

```
Sen2[suffix] -> NP[문법형태소]  Sen1[suffix]     ;; 단문 처리 규칙
        Sen1이 어떤 문형 범주에 포함되는가?
        NP의 문법형태소가 어떤 형태인가?
NP2[case] -> Sen[suffix] NP1[case]              ;; 내포문 처리 규칙
        Sen안에 모아진 문형들중  Sen cat가 요구하는 X[case]가 없을 때,
        NP1[case]는 Sen에 대해 X[case]관계를 가져야 한다.

Sen2 -> NP Sen1
         If Sen1 cat = V1
            If NP case = 이/가     Sen2 subj = NP,  Sen2 = Sen1
         If Sen1 cat = V2
            If NP case = 이/가     Sen2 subj = NP,  Sen2 = Sen1
            If NP case = 에        Sen2 loc = NP,   Sen2 = Sen1
         If Sen1 cat = V3
               . . .
```

<그림 3.1> 문형 정보를 이용한 한국어 기본 문법 틀

실제로 분석을 위해 쓰여진 문법은 아래와 같다.

```
(<s> <==> (<xp> <s>)
;N이 + V
;새가 운다
    (
        (*OR* (((x2 cat) =c v1)
               (((x1 p-part form) =c (*or* 이 가 은 는 에서 께서))
                 (x0 = x2)
                 ((x0 subj) = x1))))
;N이 + N에 + V
;그는 무명도에 정착했다.
               (((x2 cat) =c v2)
```

```
                (*or*
                 (((x1 p-part form) =c (*or* 이 가 은 는 에서 께서))
                   (x0 = x1)
                  ((x0 subj) = x1)))
                 (((x1 p-part form) =c (*or* 에 에게 께))
                   (x0 = x2)
                  ((x0 loc) = x1))))
;N이 + N로 + v
;그는 반장으로 선출되었다
            (((x2 cat) =c v3)
             (*or*
               (((x1 p-part form) =c (*or* 이 가 은 는 에서 께서))
                (x0 = x2)
                ((x0 subj) = x1))
               (((x1 p-part form) =c (*or* 로 으로))
                (x0 = x2)
                ((x0 tar) = x1))))
;N이 + N와 + V
;나는 영수와 다투었다
           (((x2 cat) =c v4)
            (*or*
             (((x1 p-part form) =c (*or* 이 가 은 는 에서 께서))
              (x0 = x2)
              ((x0 subj) = x1))
             (((x1 p-part form) =c (*or* 와 과))
              (x0 = x2)
              ((x0 with) = x1))))
;N이 + N에서 + V
;그는 학회에서 탈퇴하였다
          (((x2 cat) =c v5)
           (*or*
            (((x1 p-part form) =c (*or* 이 가 은 는 께서))
              (x0 = x2)
             ((x0 subj) = x1))
            (((x1 p-part form) =c (*or* 에서))
              (x0 = x2)
             ((x0 sour) = x1))))

               . . .
)))
```

<그림 3.2> 명사구-용언의 관계설정을 위한 문법

실제 문장이 어떠한 패턴을 가질 수 있는지 살펴보고, 아래 예문에 대해 문형 기반의 구문 분석을 행하고 평가해 보자.

예문) ㉮ [그는(subj)] [서점에서(loc)] [책을(obj)] 구입했다.
㉯ [나는(subj)] [철수와(with)] [마음이(comp)] 맞는다
㉰ [철수가(subj)] [화난 영희와(with)] [그 일에 대해(for)] [의견을(obj)] 이야기했다.

```
((subj (n-part ((form 나) (cat N)))
       (p-part ((form 는) (cat p))))
 (loc (n-part ((form 서점) (cat N)))
      (p-part ((form 에서) (cat p))))
 (obj (n-part ((form 책) (cat N)))
      (p-part ((form 을) (cat p))))
 (cat V15)
 (form 구입하다)
 (tense past)
 (mood dec))
```

<그림 3.3> 예문 ㉮의 구문 분석 결과

```
((subj (n-part ((form 나) (cat N)))
       (p-part ((form 는) (cat p))))
 (with (n-part ((form 철수) (cat N)))
       (p-part ((form 와) (cat p))))
 (comp (n-part ((form 마음) (cat N)))
       (p-part ((form 이) (cat p))))
 (cat V10)
 (form 맞다)
 (tense past)
 (mood dec))
```

<그림 3.4> 예문 ⑭의 구문 분석 결과

```
((subj (n-part ((form 철수) (cat N)))
       (p-part ((form 가) (cat p))))
 (with (n-part ((form 영희) (cat N)))
       (p-part ((form 와) (cat p)))
       (cat A1)
       (form 화나다)
       (tense present)
       (mood relation))
 (for  (n-part ((form 일) (cat N)))
       (p-part ((form 에대해) (cat p)))
       (modify (form 그) (cat DET)))
 (obj  (n-part ((form 의견) (cat N)))
       (p-part ((form 을) (cat p))))
 (cat V19)
 (form 이야기하다)
 (tense past)
 (mood dec))
```

<그림 3.5> 예문 ⑮의 구문 분석 결과

⑬와 ⑭의 단문 분석은 일반적으로 간단하다. 각 NP의 격이 어떠한지 형태소 분석 결과 나타나므로 쉽게 구문 분석 할 수 있다. ⑮의 경우, 복문 처리를 위한 규칙(NP2 -> S NP1)에서 S가 어떠한 격관계를 요구하는지 분석한 다음 아직 나타나지 않은 격관계로 NP1을 설정한다.

위의 문법 표현과 예문을 구문 분석한 결과에서 볼 수 있듯이 문형 표현을 위한 문법 체계는 비교적 간단하다. 이것은 용언 자체가 어떠한 문법 구조를 요구하는 지 이미 정보를 가지고 있기 때문이다. 따라서 문형 정보를 이용할 경우, 한국어 문법은 간결화 될 수 있다.

4. 결론

문형을 이용한 구문 분석을 통해 파싱 결과를 간단한 구문 트리로 나타낼 수 있으며 결과 자체가 구문적 모호성을 많이 배제하는 형태를 취할 수 있다. 문형 자체가 가지는 또다른 특징은 한국어 문법의 기본 구조를 자연스럽게 표현할 수 있다는 점이다. 또한 문형 자체의 분류가 의미론적 특성을 조금은 포함할 수 있기 때문에 용언 중심의 하위 범주화 사전이 완벽히 구비된다면 직접 의미 구조로의 사상이 가능하다. 그리고 문형을 위해 분류된 용언은 그 자체가 명사에 대한 제약 정보를 가지고 있기 때문에 모호성 해결을 위한 정보원으로 활용될 수 있다.

참고 문헌

[Shie86] Shieber, S. M., An Introduction to Unification-Based Approaches to Grammar, CSLI Lecture Notes, 1986.

[강은국94] 강은국, 조선어 문형 연구, 도서출판 박이정, 1994

[박인철95] 박인철, 배우정, 이용석, "한국어 개념 그래프 생성을 위한 의미 분석기의 설계," 한국정보과학회 춘계 발표논문집, 22권, 1호, 1995.

[양승원95] 양승원, 박영진, 이용석, "조건 단일화 기반 PATR II를 이용한 한국어 구문분석," 한국정보과학회 논문지, 22권, 5호, 1995.