파싱을 위한 선택 : 구문 형태소의 이용

황이규, 송연정, 이현영, 이용석
전북대학교 컴퓨터과학과 언어정보공학실
{yghwang, yjsong, hylee}@cs.chonbuk.ac.kr, yslee@moak.chonbuk.ac.kr

(Another Choice for Parsing : Using Syntactic Morpheme)

Y. G. Hwang, Y. J. Song, H. Y. Lee, Y. S. Lee
Dept. of Computer Science, Chonbuk National University

요 약

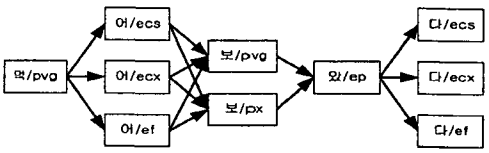
자연어 분석에서 발생하는 가장 큰 문제점은 분석의 각 단계에서 필요 이상의 모호성이 발생하는 것이다. 이러한 모호성은 각각의 분석 단계에서는 반드시 필요한 결과일 수 있지만 다음 단계의 관점에서는 불필요하게 과생성된 자료로 볼 수 있다. 특히 한국어 형태소 분석 단계는 주어진 문장에 대해 최소의 의미를 가지는 형태소로 분석하기 때문에 과생성된 결과를 많이 만들어 내는데, 이들 대부분이 보조용언이나 의존 명사를 포함하는 형태소열에서 발생한다. 품사 태깅된 코퍼스에서 높은 빈도를 나타내는 형태소들을 분석해 보면 주위의 형태소와 강한 결합 관계를 가지는 것을 발견할 수 있다. 이러한 형태소는 대부분 자립성이 없는 기능형태소로서, 개개의 형태소가 가지는 의미의 합으로 표현되기보다는 문장내에서 하나의 구문 단위로 표현될 수 있다. 본 논문에서는 이 형태소열을 구문 형태소로 정의하고, 필요한 경우 일반 형태소 해석의 결과를 구문 형태소 단위로 결합하고 이를 바탕으로 구문 해석을 하는 방법을 제안한다. 구문 형태소 단위를 이용하여 구문해석을 수행함으로써, 형태소 해석 결과의 축소를 통해 불필요한 구문 해석 결과를 배제할 수 있다.

1. 서론

한국어 분석에서 형태소 해석의 과생성 문제는 구문 해석 과정에서 많은 불필요한 연산을 수행하게 하며, 불필요한 구문 트리를 만들어 내기도 한다. 이러한 문제의 해결을 위해 형태소 해석 과정에서 간단한 형태소 배열 규칙을 이용하여 제약하거나 과생성된 형태소 해석 결과에서 하나의 품사열로 추정하기 위해 태깅을 도입하였다[1].

이는 형태소 해석 단계에서 어절을 의미 있는 최소의 단위로 분리하고, 다양한 문맥 지식이나 통계 지식을 이용하여 과생성된 형태소 분석 후보를 축소시키는 과정이다.

예를 들어 "먹어 보았다"라는 문장에 대한 해석 과정을 살펴보자. "먹어 보았다"란 문장을 최소의 의미를 가지는 형태소들로 분해하면 [그림 1]과 같다. 이는 가능한 최소의 의미로 문장을 분해한 후, 형태소 배열규칙을 이용하여 과생성된 형태소들을 부분적으로 제거한 것이다. 만일 형태소 배열 규칙을 적용하지 않는다면 더 많은 형태론적 모호성이 발생할 수 있다. [그림 1]에서 형태소 배열 규칙을 이용해도 많은 모호성이 아직 해결되지 않은 상태로 남아 있다.



[그림 1] "먹어 보았다"의 형태소 해석 결과

문장 해석의 최종 목표가 형태소 해석이 아니면, 입력 문장을 구문 해석이나 의미 해석 단계를 거쳐 주어진 문장에 대한 의미를 얻을 수 있다. "먹어 보았다"란 문장의 최종적인 해석 결과는 아래와 같은 형태를 보일 수 있다. 즉, "먹다"라는 어휘의 의미에 "시도", "과거", "평서"이라는 의미가 추가되었다.

먹/pvg(시도, 과거, 평서)

그런데, "시도"와 같은 양상 자질을 가지는 형태소는 특징이 있다. 양상 자질을 이루는 형태소는 문장에서 자주 출현하며, 여러 형태소들이 결합하여 하나의 양상 자질을 나타낸다. 이외에도 많은 형태소들이 서로 결합하여 문장에서 기능적, 구문적 단위를 형성할 수 있다. 이들은 특정 품사 배열과 특정 어휘 배열을 가지므로 쉽게 하나의 단위로 묶을 수 있으며, 형태소끼리 서로 강한 결합 관계를 형성하고 있다. 따라서 이들을 구문 해석 전에 효과적으로 묶을 수 있다면 형태소 모호성을 줄일 수 있으며, 구문 해석의 부담을 크게 줄일 수 있다.

본 논문에서는 형태소 해석 결과에 있는 모호성을 줄여 구문 해석의 부담을 줄이는 방안으로 여러 형태소가 결합하여 하나의 구문적 단위나 의미적 단위를 가질 수 있는 형태소들을 **구문 형태소**로 정의하고 한국어에 어떠한 형태소열이 구문적 형태소 단위를 이루고 있는지 살펴본다.

2장에서는 이와 구문 형태소와 관련된 국내 연구를 살펴 보고, 3장에서 어떠한 종류의 구문 형태소가 문장에서 나타나는지 살펴본다. 4장에서 이런 구문 형태소들의 결합과 이에 따른 문제점을 살펴본다. 5장에서 결론과 향후 연구 방안에 대해 토의한다.

2. 관련 연구

구문 형태소는 선행하는 형태소에 정보를 더하거나 그 자체로 하나의 문법 범주로 이용될 수 있다. 이것을 관용구의 관점에서 본 연구로 [2]이 있었다.

[2]에서는 현대 국어 관용구를 형태적, 통사적, 의미론적 결합관계의 특성에 따라 재분류하였다. 우리는 [2]에서 설명하는 '형태적 연어'의 처리에 중점을 두었다. 형태적 연어를 제외한 '숙어'나 '의미적 연어'는 생산성이 크지 않으며, 의미 해석의 관점에서 보아야 하기 때문에 이들은 본 논문에서 제외했다.

형태적 연어와 관련된 연구로 복수어 단위 분석 방법[3]이 있다. 복수어 단위정보는 단어자체에 관한 품사나 하위 범주화 정보뿐만 아니라 그 단어가 실제로 사용되는 문맥 특성을 지칭하는 것으로 말뭉치 사전을 기반으로 하여 구성된다[4]. 우리는 복수어 단위 분석 방법을 기반으로 하여 코퍼스에 나타나는 형태소열을 분류하여 하나의 구문이나 의미 단위로 간주할 수 있는 형태소열을 하나의 형태소로 간주하였다.

[5]에서는 어휘화된 배열 규칙과 포섭 관계를 이용하여 67%의 형태소 해석의 모호성을 축소하고 있다. 이는 다중 단어 개념을 품사 태깅 모델에 적용한 것인데, 다중 단어는 문장상에서 두 개 이상의 단어가 하나의 단어로서 역할을 담당하는 연어 속성을 가진 단어들을 말한다.

또한, 묶인말 정보를 이용하여 형태소 모호성을 해소한 예도 있는데, [6]에서는 관용어 중에서 통사적으로 강한 어순 제약을 보이며 다른 요소의 삽입 없이 공기하는 묶인말을 추출하여 사전을 구축하고 이에 어미-보조용언을 추가하여 품사 태깅 시스템에 이용하고 있다.

[3], [5]와 [6]에서는 형태소 모호성의 해소를 위해 어휘단계의 지식을 이용하고 있다. 이러한 연구들의 특징은 단순히 형태론적 모호성의 축소에 초점을 맞추었다. 본 논문에서는 구문 해석의 관점에서 어떠한 구문형태소들이 반드시 기능적 단위로 결합되어야 하며, 이 구문 형태소를 어떠한 기능을 문장에서 수행하고 있고, 따라서 이들을 어떠한 구문적 단위로 보고 문장을 해석해야 하는지에 초점을 맞추었다.

또 다른 연구로 부분적인 어절 결합 관계를 이용한 방법[7]이 있는데, 이 연구에서는 형태소 해석후 구문 해석 전에 어간 다음에 따르는 기능상의 어절들을 하나의 의미상 단위 어절로 결합시킨다. 이 연구에서는 파싱전에 중간 분석 단계에서 복합어구 생성시 여러 개의 내용어와 기능어간의 결합을 허용하고 있는데, 이를 효과적으로 수행하기 위해서는 형태소 지식이외에 구문적 지식이 필요하다. [7]의 경우, 구구조 문법을 이용하는 결합 단위로의 확장이 이루어진 반면, 본 논문에서는 기능적 단위의 형태소만을 결합 단위로 간주함으로써 형태소 해석과 구문 해석에서의 일관성을 유지하였다.

3. 구문 형태소

3.1 구문 형태소의 추출

본 논문에서는 여러 기능형태소가 결합하여 하나의 의미나 문법 단위를 이루어 선행하는 용언이나 체언과 결합하는 형태소의 나열을 구문형태소로 정의한다. 구문 형태소를 이루는 형태소열은 개개 형태소의 뜻이 결합하여 전혀 다른 뜻을 지닌다. 이런 형태소 개개를 분리하여 구문 해석이나 의미 해석에서 결합하는 방법은 많은 어려움이 있다. 본 논문에서는 이들을 구문 해석이나 의미해석 전에 하나의 단위로 결합하여 여러 응용에 효율적으로 이용될 수 있도록 구문형태소 단위의 결합을

시도하였다.

구문 형태소들을 찾아내기 위해 국어정보베이스[8]의 품사태깅된 코퍼스로부터 추출된 어휘화된 바이그램, 트라이그램, 4-gram을 대상으로 상위 빈도를 나타내는 형태소열 중 하나 이상의 어절에 걸쳐 있으며, 내용어가 포함되지 않고 기능어만으로 구성된 형태소열을 추출하였다. 아래의 예 중 "때문", "예정"과 같은 단어는 내용어이지만 출현 빈도가 높고 자립성이 약하기 때문에 포함시켰다. 이들을 대상으로 강한 결합성 여부와 내용어에 어떠한 기능적, 의미적 정보를 부가하고 있는지를 기준으로 분류하였다. [표 1]은 약 20만여 어절에서 얻어진 바이그램, 트라이그램, 4-gram 중 상위 20개를 보여주고 있다.

bigram	trigram	4-gram
고/ecx 있/px	고/ecx 있/px 다/ef	고/ecx 있/px 습니다/ef ./sf
에/jca 대하/pvg	고/ecx 있/px 습니다/ef	고/ecx 있/px 다/ef ./sf
지/ecx 않/px	하/xsv 고/ecx 있/px	하/xsv 고/ecx 있/px 다/ef
ㄴ/etm 것/nbn	적/xsn 이/jp ㄴ/etm	것/nbn 이/jp 다/ef ./sf
ㄹ/etm 것/nbn	에/jca 대하/pvg ㄴ/etm	하/xsv 기/etn 로/jca 하/pvg
는/etm 것/nbn	고/ecx 있/px 는/etm	기/etn 로/jca 하/pvg 있/ep
에/jca 따르/pvg	기/etn 로/jca 하/pvg	하/xsv 고/ecx 있/px 습니다/ef
기/etn 위하/pvg	ㄹ/etm 것/nbn 으로/jca	ㄹ/etm 것/nbn 으로/jca 보이/pvg
어/ecx 주/px	ㄴ/etm 것/nbn 으로/jca	하/xsv 고/ecx 있/px 는/etm
게/ecx 되/px	에/jca 따르/pvg 아/ecs	하/xsv ㄹ/etm 예정/ncn 이/jp
어/ecx 있/px	에/jca 대하/pvg 어/ecs	이/npd 에/jca 따르/pvg 아/ecs
어/ecx 지/px	ㄹ/etm 것/nbn 이/jp	ㄹ/etm 것/nbn 이/jp 라/ef
기/etn 때문/nbn	기/etn 위하/pvg 어/ecs	되/xsv 고/ecx 있/px 습니다/ef
어야/ecx 하/px	것/nbn 이/jp 다/ef	고/ecx 있/px 는/etm 것/nbn
어/ecx 오/px	ㄹ/etm 예정/ncn 이/jp	하/xsv 기/etn 위하/pvg 어/ecs
지/ecx 못하/px	하/xsv 기/etn 위하/pvg	는/etm 것/nbn 이/jp 다/ef
을/jco 위하/pvg	하/xsv ㄹ/etm 것/nbn	기/etn 때문/nbn 이/jp 다/ef
ㄹ/etm 예정/ncn	것/nbn 으로/jca 보이/pvg	기/etn 로/jca 하/pvg 있/ep
을/jco 통하/pvg	ㄴ/etm 것/nbn 이/jp	어/ecx 있/px 습니다/ef ./sf
어/ecx 보/px	기/etn 때문/nbn 에/jca	에/jca 대하/pvg 어서/ecs 는/jxc

[표1] 상위 빈도수의 인접한 형태소 열

이중 어휘를 제외하고 인접한 품사들을 다시 출현 빈도수로 정렬하여 인접한 품사들 사이의 강한 관계를 바탕으로 분류하여 구문 형태소들을 정리하였다.

3.2 양상 자질을 나타내는 구문 형태소

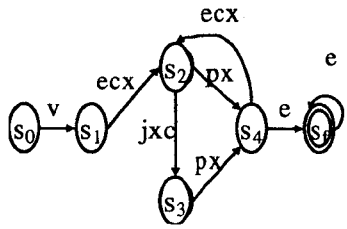
3.2.1 용언구에 나타나는 구문 형태소

한국어에서 용언의 뒤에 보조적 연결어미를 매개로 선행 용언에 속성을 첨가하는 또 다른 용언이 나타날 수 있는데 이를 보조용언이라고 한다. 이런 보조용언은 문법적 역할을 하기보다는 선행하는 용언에 단순히 양상 정보를 부가한다. 따라서 이런 보조용언을 용언의 한 부분으로 간주함으로써 형태론적 모호성을 줄일 수 있다.

여기에서, 용언에 후행하는 선어말 어미나 어말 어미도 구문 형태소와 통합할 수 있다. 선어말 어미가 나타내는 '시제', '높임', '공손' 등도 양상의 의미를 나타내며, 어말 어미가 나타내는 '평서', '감탄', '의문', '명령', '청유', '연결', '전성' 등도 용언에 의미를 첨가하는 것으로 간주할 수 있다. 따라서 구문 형태소의 기능을 확장하면 용언구를 통합적으로 인식할 수 있다.

Type 1 : <v> {<ecx> [jxc] <px>}+ <e>
예) "-어 버리-", "-고 있-"

Type 1형 구문 형태소 인식하기 위한 오토마타는 [그림 2]와 같다.



[그림 2] Type 1형 구문형태소 인식을 위한 오토마타

[표2]에서 실제 코퍼스에서 추출된 Type 1형 구문 형태소의 일부를 정리했다.

(-아/-어) 지다, (-게) 되다, (-게) 하다, (-게) 만들다, (-아/-어) 가다, (-아/-어) 오다, (-고) 있다, (-고) 계시다, (-아/-어) 내다, (-아/-어) 버리다, (-고) 나다, (-고) 말다, (-아/-어) 주다, (-아/-어) 드리다, (-아/-어) 두다, (-아/-어) 놓다, (-아/-어) 가지다, (-아/-어) 대다, (-지) 말다, (-지) 못하다, (-지) 않다, (-아/-어) 보다, (-아/-어) 보이다, (-어야) 하다, (-고) 싶다, (-아/-어) 있다, (-아/-어) 계시다, (-는가/-ㄴ가/-나) 보다

[표2] Type 1형 구문 형태소의 예

Type 1형 구문 형태소는 선행하는 용언의 어간과 통합되어 하나의 단위로 인식된다. 이러한 양상 정보는 자질의 형태로 표현된다. 예를 들어 "먹고 싶다"의 구문 해석 전 결과는 "먹/pvg[희망, 평서]"가 된다. '희망'의 양상 자질과 평서형 종결어미가 용언 "먹"과 결합하여 하나의 용언구를 이룬다.

3.2.2 의존명사에 나타나는 구문 형태소

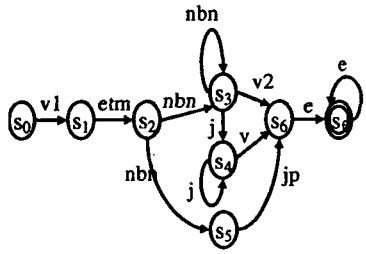
의존 명사가 후행하는 용언과 결합하여 선행하는 용언에 양상을 추가하는 기능을 수행하는 경우가 있다. 이런 형태소열은 구문 해석 트리에서 문장의 구조와 관련이 있는 것은 아니라 단지 선행하는 용언에 양상 자질이 더해진다.

Type 2 : <v1> <etm> <nbn>+ [<j>]* <v> <e>
예) "-을 수 있-", "-ㄴ 줄도 모르-"

Type 2의 변형으로 의존 명사에 서술격 조사가 결합한 type 2' 형태도 이 범주에 속한다.

Type 2' : <v1> <etm> <nbn> <jp> <e>
예) "-을 판이-", "-을 모양이-"

아래는 Type 2와 2'형 구문 형태소 인식하기 위한 오토마타를 표현한 것이다.



[그림 3] Type 2, 2'형 구문형태소 인식을 위한 오토마타

[표3]에서 실제 코퍼스에서 추출된 Type 2, 2'형 구문 형태소의 일부를 정리했다.

(-ㄴ/-는) 경우가 많다/있다/흔하다, (-ㄴ/-는) 바(가/도) 있다/없다, (-ㄴ/-는) 바에 따르다, (-ㄴ/-는) 셈이다, (-ㄴ/-는) 수가 많다/있다, (-ㄴ/-는) 적(도/은/이) 있다/없다, (-ㄴ/-는) 줄(도/은) 몰랐다/알았다, (-ㄴ/-은) 편이다, (-ㄹ/-을) 리(가/는) 없다, (-ㄹ/-을) 만(은) 하다, (-ㄹ/-을) 모양이다, (-ㄹ/-을) 바(는/를) 모릅니다/없습니다, (-ㄹ/-을) 뿐(만) 아니다, (-ㄹ/-을) 뿐이다, (-ㄹ/-을) 줄 모르다/알다, (-ㄹ/-을) 지 모르다/알다, (-ㄹ/-을) 지경에 이르다, (-ㄹ/-을) 지경이다, (-ㄹ/-을) 터이다

[표3] Type 2, 2'형 구문 형태소의 예

Type 2, 2'형 구문 형태소는 선행하는 용언의 어간과 통합되어 하나의 단위로 인식된다. 이러한 양상 정보도 Type 1형과 마찬가지로 자질의 형태로 표현된다. 예를 들어 "먹을 수 있다"의 구문 해석 전 결과는 "먹/pvg[가능, 평서]"가 될 것이다. 즉 '가능'의 양상 자질과 평서형 종결어미가 용언 "먹"과 결합하여 하나의 용언구를 이룬다.

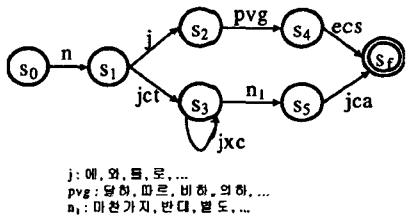
3.3 문법 범주를 나타내는 구문 형태소

의미 관점에서 볼 때, 조사가 의존명사에서 파생한 용언이나 일반 용언과 결합하여 문장 내에서 서술어로서의 기능을 하지 않고 하나의 조사로서 역할을 수행하는 경우가 있다. 이런 예를 영어의 표현으로 변환하여 보면 단지 하나의 전치사로 대응됨을 알 수 있다. 우리는 이러한 형태소열을 **의사 조사**로 정의하고 이를 구문 형태소의 범주에 포함시켰다. Type 1과 2, 2'형이 양상 자질의 형태로 표현되는데 반해 Type 3, 3'형은 문법 범주인 조사의 형태로 변환된다. 의사조사는 문장에서 의미 해석시 심층격으로 표현될 수 있다.

Type 3 : <n> <j> <v> <ecs>
예) "-에 관해", "-로 인해"

Type 3' : <n> <j> <n> <j>
예) "-와 반대로", "-와 별도로"

아래는 Type 3, 3'형 구문 형태소 인식하기 위한 오토마타를 표현한 것이다.



[그림 4] Type 3, 3'형 구문 형태소 인식을 위한 오토마타

[표4]에서 실제 코퍼스에서 추출된 Type 3, 3'형 구문 형태소의 일부를 정리했다.

(-에) 달해, (-에) 따라, (-에) 비해, (-에) 의해, (-에) 처해, (-에) 한해, (-에) 반해, (-와) 같이, (-와) 견주어, (-와) 관련하여, (-와) 달리, (-와) 함께, (-와) 더불어, (-를) 비롯해, (-를) 통해, (-를) 향해, (-를) 두고, (-를) 맞아, (-을) 가지고 (-로) 말미암아, (-로) 미루어, (-와) 마찬가지로, (-와) 반대로, (-와) 별도로, (-에) 있어,

[표4] Type 3, 3' 형 구문 형태소의 예

Type 3, 3'형 구문 형태소는 선행하는 체언과 통합되어 체언구로 인식될 수 있다. 이런 체언구에서 Type 3, 3'형은 의사 조사의 역할을 수행하고 있다. 예를 들어 "철수로 인해"의 구문 해석 전 결과는 "철수/nq+로인해/j[cause]"가 된다. 즉 "-로인해"가 하나의 문법 범주인 조사로 결합되어 체언인 "철수"와 함께 체언구를 이룬다.

4. 구문 형태소 단위의 결합

4.1 구문 형태소 생성

형태소 해석 후 구문 형태소 생성기는 구문 단위 형태소를 결합하여 구문 해석의 입력 단위로 이용한다. 이는 기존의 형태소 해석기와는 독립적인 관점에서 구문 형태소 생성 단계를 둔 것으로 이미 만들어진 다양한 형태소 해석기를 이용할 수 있다.

구문 형태소 생성기의 기능은 크게 두 가지 관점에서 정의된다. 첫째, 형태소 후처리기의 역할이다. 형태소 해석기가 생성해 내는 많은 모호성의 축소를 위해 다양한

형태소 제약 정보를 이용하는 방법[3, 6]이나 태거를 이용하는 방법[5, 9]과 마찬가지로 구문 형태소 단위의 인식을 통해 불필요한 모호성을 제거한다. 둘째, 구문 해석기의 전처리 과정으로서의 역할이다. 구문 해석의 효율을 위해 부분적인 어절 결합을 이용하는 방법[7]과 유사하게 우리는 기능적 단위의 형태소열을 구문 해석 전에 결합함으로써 불필요한 구문 해석 과정을 줄일 수 있다. 실제 시스템에서 나타나는 구문 형태소 자질 생성의 결과는 아래와 같은 형태를 가지고 있다.

```
"잡고 싶다"
   (@VP (form 잡고_싶다) (root 잡) (cat VP) (subcat pvg)
        (modal hope) (eform dec) )
"먹을 수 밖에 없다"
   (@VP (form 먹을_수_밖에_없다) (root 먹) (cat VV) (subcat pvg)
        (modal inevitable) (eform dec))
"컴퓨터에 대하여"
   (@NP (form 컴퓨터에_대하여) (root 컴퓨터) (cat NP)
        (subcat ncn) (jform to))
```

[표5] 구문 형태소 자질 생성 결과

4.2 구문 해석 단계에서의 모호성 축소

한국어는 교착어이기 때문에 어절 분절의 특성을 가지고 있다. 따라서 구문 해석의 전단계에서 오직 하나의 형태소열을 정확히 선택하는 것은 어렵다. 따라서 허용될 수 있는 수의 형태소 열들은 구문 해석기가 받아들여야만 한다. 즉, 태깅 단계에서 해결될 수 없지만 구문 해석이나 의미 해석에 의해 쉽게 해결될 수 있는 모호성도 많이 존재한다. 구문 형태소 생성기를 통과한 형태소 해석 결과도 일부분의 모호성을 포함하며, 이중 대부분은 구문 해석 과정에서만 모호성이 축소될 수 있는 것이다.

예를 들어, "나는" 이라는 어절에 대한 형태소 해석 결과중 "나/pvg+는/etm"과 "날/pvg+는/etm"과 같은 이형동품사의 경우 태깅 단계에서 적절한 후보를 선택하는 것은 주변 어절의 문맥 정보를 필요로 한다[9]. 그러나 이러한 주변 어절 문맥 정보를 구축하기 위해서는 대량의 품사 태그된 코퍼스가 필요하며, 문맥 정보를 유지하기도 쉽지 않다. 이러한 후보 중 올바른 결과는 구문 해석 과정중에서 쉽게 찾아질 수 있다. "나다"는 자동사이며, "날다"는 타동사이기 때문에 주어진 문장을 구문 해석하는 도중에 올바른 선택을 할 수 있다.

4.3 토의 사항

구문 형태소 단위의 형태소 인식으로 발생하는 문제점으로 용언과 이에 부가되는 양상 자질의 주체가 서로 다른 경우가 있다. 예를 들어,

"나는 영희가 밥을 먹게 했다"

라는 문장에서 '사역'의 양상을 나타내는 "-게 하"와 "먹다"의 주체가 다르다. 따라서 이들을 하나의 단위로 결합할 경우, 의미 해석에서 전혀 다른 결과를 만들어 낼 수 있다. 이러한 구문 형태소로 "-게 만들-", "-지 말-", "-는 줄(은/도/만) 모르-/알-", "-ㄹ 리(가/도) 없-", "-을 모양이-" 등이 있다. 이런 경우는 전체 구문 형태소에서 차지하는 비중이 작기 때문에 구문 형태소 단위의 인식으로 모호성 제거 후 의미 해석 단계에서 이를 해결하거나 구문 형태소 인식 단계에서 이런 형태소를 구문 형태소로 묶지 않도록 함으로써 해결할 수 있다.

5. 결론

구문 해석을 어렵게 하는 대표적인 원인은 형태소 해석의 과생성에 있다. 특히 한국어나 일본어와 같은 교착어의 특성을 가지는 언어일수록 이러한 현상이 빈번하게 발생하고 있다. 본 연구에서는 한국어에서 특히 형태소 과생성을 유발하는 복합 동사구와 의존 명사 등을 포함하는 어절에 대해 구문 형태소 단위의 처리를 제안하였다. 우리는 구문 형태소를 구문 해석의 기본 단위로 묶어주는 구문 형태소 생성기를 이용하여 구문 해석에 도움을 주는 방법을 보였다. 구문 형태소 생성기는 강한 결합을 가지는 기능 형태소들을 하나의 형태소로 간주하며, 이들 사이의 결합 관계를 이용하여 형태소 모호성을 축소할 수 있다. 또한 구문 해석의 입력 단위로 구문 형태소 단위를 사용함으로써 구문 해석 과정을 간략화 시킬 수 있다.

몇몇 구문 형태소의 경우, 하나의 양상 자질로 묶을 경우 의미 해석에서 문제가 발생하는 것을 볼 수 있다. 이는 발생 빈도가 낮지만 이에 대한 연구가 진행되어야 하며, 구문 형태소들을 좀 더 확장할 수 있도록 다양한 한국어 기능 어절에 대한 분석이 진행되어야 할 것이다.

참고문헌

[1] Eugene Charniak, Curtis Hendrickson, Neil Jacobson, Mike Perkowitz, "Equations for Part-Of-Speech Tagging," Proc. of the 11th National Conference on Artificial Intelligence (AAAI), pp.784-789, 1993

[2] 이희자, "현대 국어 관용구의 결합 관계 고찰", 제6회 한글 및 한국어 정보처리 학술대회, p. 333-352, 1994

[3] 강승식, 음절 정보와 복수어 단위 정보를 이용한 한국어 형태소 분석, 서울대학교 대학원 컴퓨터공학과 박사학위 논문, 1993

[4] S.Sato, "Example-Based Translation Approach," Proceedings of the International Workshop on

Fundamental Reasearch for the Future Generation of Natural Language Processing, pp.1-16, 1991.
[5] 김재훈, 오류-보정 기법을 이용한 어휘 모호성 해소, 한국과학기술원 전산학과 대학원 박사학위 논문, 1996
[6] 박혜준, 윤준태, 송만석, "말뭉치 품사 꼬리달기 시스템", 제21회 정보과학회 춘계 학술발표 논문집, pp. 829-832, 1994
[7] 김창제, 정천영, 김영훈, 서영훈, "부분적인 어절 결합을 이용한 효율적인 한국어 구문 분석기", 제22회 정보과학회 가을 학술발표 논문집, pp. 597-600, 1995
[8] KIBS : Korean Information Base System, http://kibs.kaist.ac.kr/
[9] 임희석, 언어 지식과 통계 정보를 이용한 한국어 품사 태깅 모델, 고려대학교 대학원 컴퓨터학과 박사학위 논문, 1997